







Abstracts of papers, posters and talks presented  
at the 2009 Joint RECOMB Satellite Conference on

---

# **REGULATORY GENOMICS - SYSTEMS BIOLOGY - DREAM4**

---

Wed Dec 2-Sun Dec 6, 2009  
MIT / Broad Institute / CSAIL

**Conference Chairs:**

**Manolis Kellis, MIT**

**Ziv Bar-Joseph, CMU**

**Andrea Califano, Columbia**

**Gustavo Stolovitzky, IBM**

## Conference Chairs:

Manolis Kellis ..... Associate Professor, MIT  
Ziv Bar-Joseph ..... Associate Professor, CMU  
Andrea Califano ..... Professor, Columbia University  
Gustavo Stolovitzky.....Systems Biology Group, IBM

## In Partnership With:

PLoS Computational Biology .....editor: Catherine Nancarrow  
PLoS ONE ..... editor: Peter Binfield

## Associate Editor, PLoS Computational Biology

Uwe Ohler

## Session Chairs:

### Regulatory Genomics:

Nitin Baliga  
Panayiotis (Takis) Benos  
Michael Brent  
Christina Leslie  
Uwe Ohler  
Ron Shamir

### DREAM4:

Fritz Roth  
Peter Sorger

### Systems Biology:

Diego di Bernardo  
James Collins  
Yuval Kluger  
Avi Ma'ayan  
Franziska Michor  
Roded Sharan  
Pablo Tamayo  
Dennis Vitkup

## Program Committee:

### Regulatory Genomics:

Manolis Kellis, chair  
Ziv Bar-Joseph, chair  
Orly Alter  
Nitin S Baliga  
Panayiotis (Takis)  
Benos  
Mathieu Blanchette  
Michael R. Brent  
Albert Erives  
Eleazar Eskin  
Ernest Fraenkel  
Nir Friedman  
Mikhail Gelfand  
Sridhar Hannenhalli  
Tim Hughes  
Uri Keich  
Christina Leslie

Hao Li

Eran Segal  
Ron Shamir  
Saurabh Sinha  
Mona Singh  
Christopher Workman

### Systems Biology and

### DREAM:

Andrea Califano, chair  
Gustavo Stolovitzky,  
chair  
M. Madan Babu  
Gary Bader  
Joel Bader  
Diego di Bernardo  
Jim Collins  
Joaquin Dopazo

Eleazar Eskin  
Igor Jurisica  
Pascal Kahlem  
Andre Levchenko  
Avi Ma'ayan  
Adam Margolin  
Satoru Miyano  
Dana Pe'er  
Theodore Perkins  
Timothy Ravasi  
Frederick Roth  
Michael Samoilov  
Roded Sharan  
Mona Singh  
Pavel Sumazin  
Denis Thieffry  
Ioannis Xenarios

## Local Organization and Student Volunteers:

Local organization  
Monica Concepcion  
Sally Lee  
Pia Handsom  
Karen Shirer

Matt Edwards  
Chuck Epstein  
Sudeshna Fisch  
Taran Gujral  
Christina Harview  
Eliana Hechter  
Yan Meng  
Michal Rabani  
Erroll Rueckert  
Mark Sevecka  
Tal Shay  
Eboney Smith  
Li Wang

CMU  
Anthony Gitter  
Peter Huggins  
Hai-Son Le  
Henry Lin  
Shan Zhong  
Guy Zinman

Local Volunteers:  
Loyal Goff  
Charlie Frogner  
Patrycja Missiuro  
Yang Ding  
Mary Addonizio  
Adam Callahan

Other  
Betty Chang  
Antonina Mitrofanova  
Yeison Rodriguez

## Cover art credit:

Yeast functional modules in amino-acid metabolism, Daniel Marbach  
Gene expression atlas for *Drosophila* Blastoderm. Fowlkes, DePace, Malik.  
Human membrane receptor interactions, Yanjun Qi, Judith Klein-Seetharaman.

Special thanks for booklet assembly to: Michelle Martin

## Online material:

All talks will be videotaped and shared freely online with presenter's permission.  
All papers accepted and published will be available freely online from PLoS.  
All videos of accepted papers will be linked to the papers through the PLoS site.  
Photos from the meeting will be made available through the conference website.  
All invited talks will be sync'ed with slides in an interactive site by the New York Academy of Sciences.

All these materials can be accessed at any time from:

<http://compbio.mit.edu/recombsat/>

For any inquiries, please contact the conference chairs at: [recombsat@mit.edu](mailto:recombsat@mit.edu)

## Industry and Government Sponsors:

The organizers wish to thank the following industrial and government sponsors for their generous support of this conference.

	<b>The NIH National Centers for Biomedical Computing and the MAGNet Center at Columbia University</b>
	<b>The New York Academy of Sciences</b>
	<b>IBM Research</b>

## Scientific Partners

The organizers also wish to thank the following journals for their invaluable support in publishing conference manuscripts.

	<b>PLOS Computational Biology</b>
	<b>PLOS ONE</b>

## Local Hosts and Organization

Many thanks also go to our local hosts for making this meeting possible.



**Massachusetts Institute of Technology**



**Broad Institute of MIT and Harvard**



**Computer Science and Artificial Intelligence Lab**

Wednesday December 2<sup>nd</sup>, 2009

Dear attendees,

Welcome to the 6th Annual RECOMB Satellite on Regulatory Genomics, the 5th Annual RECOMB Satellite on Systems Biology, and the 4th Annual DREAM reverse engineering challenges. The goal of this meeting is to bring together computational and experimental scientists in the area of regulatory genomics and systems biology, to discuss current research directions, latest findings, and establish new collaborations towards a systems-level understanding of activities in the cell. The next 5 days will consist of keynote presentations, oral presentations selected from submitted full length papers and 1-page abstracts, and posters presentations also selected from submitted abstracts. But most importantly, it's an opportunity to connect, discuss, exchange ideas, think together, and plan ahead.

For the second year in a row this meeting has been sold out, this year more than a month in advance. We have also received a record number of more than 50 full length papers and more than 250 abstracts, resulting in a scientific program of outmost quality. All submitted full length papers were fully reviewed by the program committee and 17 were accepted to the conference and by our partner journals, with 8 papers going to PLoS Computational Biology and 9 papers to PLoS ONE. These papers should appear online by the time the meeting starts, and will be ultimately linked together with the corresponding talks on the PLoS website, providing a new means for science dissemination. All abstracts were reviewed by the conference chairs and 16 additional session chairs, resulting in 50 abstracts selected for oral presentation, and most remaining abstracts for poster presentation.

Welcome to Boston, for what promises to be a very exciting meeting!

The conference chairs:

*Regulatory Genomics:*

Manolis Kellis (MIT), Ziv Bar-Joseph (CMU).

*Systems/DREAM:*

Andrea Califano (Columbia), Gustavo Stolovitzky (IBM).

# RECOMB Regulatory Genomics 2009

(★ Full length manuscript; ► Invited talk; Accepted abstract)

**Wednesday, Dec. 2, 2009**

---

*3pm Conference check-in open, Poster session I set-up*

*5pm Welcome Remarks*

5:15►	<u>Mark Biggin</u> : Evidence for Quantitative Transcription Networks ...	1
5:45	<u>Sarah E. Calvo</u> : Widespread translational repression.....	2
6pm★	<u>Hilal Kazan</u> : Learning binding preferences.....	3
6:15★	<u>Igor Ulitsky</u> : Towards prediction of MicroRNA function.....	4

*Light snacks – 6:30-6:45pm*

6:45	<u>Clifford A. Meyer</u> : Inferring key transcriptional regulators .....	5
7pm	<u>Jason Ernst</u> : Genome-wide discovery of chromatin states.....	6
7:15	<u>Mattia Pelizzola</u> : Human DNA methylomes at base resolution .....	7
7:30	<u>Leonid Mirny</u> : Different strategies for gene regulation.....	8
7:45►	<u>Bob Waterston</u> : Deciphering C. Elegans embryonic network.....	9

## **Regulatory Genomics Welcome Reception/Poster Session I**

**8:15pm-9:45pm**

*Hors d'oeuvres, snacks, refreshments, cash bar*

**Thursday, Dec. 3, 2009**

---

*Breakfast – 8am*

9am►	<u>Rick Young</u> : Programming cell state .....	10
9:30	<u>Jesse M. Gray</u> : Widespread RNA polymerase II recruitment.....	11
9:45★	<u>Yue Zhao</u> : Inferring binding energies .....	12
10am	<u>Guillaume Bourque</u> : Binding site turnover in stem cells .....	13

*Coffee / Snacks / Fruit Break – 10:15-10:45am*

10:45	<u>Michael Brodsky</u> : Identification and analysis of regulatory regions	14
11am	<u>Yang Ding</u> : Exact calculation of partition function .....	15
11:15★	<u>Sheng Zhong</u> : A analysis of transcription factor interactions.....	16
11:30	<u>Pouya Kheradpour</u> : Regulatory motifs associated with TF.....	17
11:45	<u>Quan Zhong</u> : Edgetic perturbation models of human .....	18

*Lunch Break / Networking Opportunities – 12-1pm*

1pm►	<u>Naama Barkai</u> : Evolution of nucleosome positioning .....	19
1:30	<u>Pieter Meysman</u> : Structural DNA for the prediction of binding.....	20

1:45	<u>Claes Wadelius</u> : Nucleosomes are positioned in exons.....	21
2pm	<u>Eugene Bolotin</u> : Identification of human HNF4 target genes.....	22

*Coffee / Snacks / Fruit Break – 2:15-2:45pm*  
*Poster set-up for Session II*

2:45	<u>Damian Wójtowicz</u> : Mapping of non-B DNA structures .....	23
3pm	<u>Elizabeth A. Rach</u> : The landscape of transcription initiation .....	24
3:15	<u>Julia Lasserre</u> : TSS detection.....	25
3:30	<u>Ron Shamir</u> : Signaling pathways analysis tool.....	26

***Regulatory Genomics Poster Session II – 3:45pm-5:15pm***  
*Hors d'oeuvres, snacks, refreshments*

5:15▶	<u>Nikolaus Rajewsky</u> : Post-transcriptional gene regulation.....	27
5:45	<u>Stein Aerts</u> : Regulatory network for retinal differentiation .....	28
6pm	<u>Raja Jothi</u> : A link between dynamics and network architecture ..	29
6:15	<u>Andrew J. Gentles</u> : Predicting histological transformation.....	30

*Break/light snacks – 6:30-6:45pm*

6:45★	<u>Ferhat Ay</u> : Analysis of Boolean regulatory networks.....	31
7pm★	<u>Lucia Marucci</u> : Turning genetic circuit into synthetic oscillator...	32
7:15	<u>Todd Wasson</u> : An ensemble model of competitive binding.....	33
7:30	<u>Jeremy Bellay</u> : Decomposition of genetic interaction networks...	34
7:45	<u>Justin Kinney</u> : Regulatory physics from DNA sequence data .....	35
8pm	<u>Erez Lieberman-Aidan</u> : Fractal model for chromatin dynamics...	36

***Dinner out on the town 8:15 – 9:45pm***  
 See recommended restaurants list and map

**Friday, Dec. 4, 2009**

---

*Breakfast – 8am*

9am▶	<u>Kevin White</u> : Transcriptional regulatory networks .....	37
9:30	<u>Zhi Xie</u> : Global Analysis of human protein-DNA Interactions .....	38
9:45★	<u>Manikandan Narayanan</u> : Simultaneous clustering .....	39
10am	<u>Andrei L. Turinsky</u> : Literature curation of protein interaction.....	40

***10:15 Welcome / DREAM Registration***

# *DREAM4 Reverse Engineering Challenges*

(★ Full length manuscript; ► Invited talk; Accepted abstract)

**Friday, Dec. 4, 2009**

---

10:45	<u>Saez-Rodriguez</u> : Discrete logic modeling to link pathway .....	41
11am	★ <u>Alexopoulos</u> : Identifying drug effects via pathway alterations ...	42
11:15	<u>Wagner</u> : Crosstalk among receptor tyrosine kinases .....	43
11:30	► <u>Garry Nolan</u> : Single cell signaling & pathology .....	44

*Lunch Break / Networking Opportunities – 12-1pm*

1pm	<u>Philip M. Kim</u> : Peptide recognition domain (PRD) .....	45
1:15	<u>Julio Saez-Rodriguez</u> : Challenge 3: Predictive signaling .....	46
1:30	<u>Robert J. Prill</u> : Challenges 1 and 3 overall results	

Best Prediction Teams 10 min Presentations

1:45	Peptide Recognition SH3 - PBIL: Kim, Hong, Chung
1:55	Peptide Recognition Kinase - Predikin: Ellis, Saunders, Kobe
2:05	Recognition PDZ - Chuck_Daly: Yanover, Zaslavsky, Bradley
2:15	Signaling Network - Giano4: DiCamillo, Corradin, Toffolo
2:25	Signaling Network - Team Steam: Schwacke

*Short Break/RG poster take-down – 2:35-2:45pm*

2:45	<u>Daniel Marbach</u> : Generating realistic benchmarks for gene .....	47
3pm	<u>Robert J. Prill</u> : Challenge 2 overall results	

In Silico Network Best Prediction 10 min Presentations:

3:15pm	Size 10 - <u>Küffner</u> , Erhard, Petri, Windhager, Zimmer
3:25pm	Size 100 - <u>Pinna</u> , Soranzo, de la Fuente
3:35pm	Size 100 - <u>Greenfield</u> , Madar, Ostrer, Bonneau
3:45pm	Size 100 multifactorial - <u>Irrthum</u> , Wehenkel, Geurts, Huynh-Thu

*Short Break/ SB registration open/ poster setp-up Systems Bio  
3:55-4:15pm*

4:15	► <u>Michael Yaffe</u> : Systems biology of DNA damage .....	48
4:45	★ <u>Qian</u> : Effective identification of conserved pathways .....	49
5pm	<u>Marbach</u> : Strengths and weaknesses of network inference .....	50
5:15	<u>Lefebvre</u> : A human B cell interactome .....	51
5:30	<u>Iorio</u> : identifying drug mode of action from gene expression .....	52

***DREAM Poster Session II - 8:15pm-9:45pm***

*Hors-d'oeuvre, heavier snacks, cash bar*

# RECOMB Systems Biology 2009

(★ Full length manuscript; ► Invited talk; Accepted abstract)

## Friday, Dec. 4, 2009

---

5:45 *SB Welcome Remarks*

6pm ► <u>Nevan J. Krogan</u> : Insights from interaction maps.....	53
6:30 ★ <u>Tomer Benyamini</u> : Metabolic flux balance analysis.....	54
6:45 ★ <u>Yongjin Park</u> : Dynamic networks.....	55

*Break/light snacks – 7pm-7:15*

7:15 ★ <u>Christina Chan</u> : A dynamic analysis of IRS-PKR signaling.....	56
7:30 <u>Doron Betel</u> : Comprehensive modeling of microRNA targets.....	57
7:45 ► <u>Franziska Michor</u> : The cell of origin of human cancers.....	58

### **Systems Biology: Welcome Reception/ Poster Session I- 8:15pm-9:45pm**

*Hors-d'oeuvre, heavier snacks, cash bar*

## Saturday, Dec. 5, 2009

---

*Breakfast – 8am*

9am ► <u>Jef D. Boeke</u> : Building <i>Saccharomyces cerevisiae</i> v2.0.....	59
9:30 <u>Diogo Camacho</u> : Decoding small RNA networks in bacteria.....	60
9:45 <u>Benjamin Logsdon</u> : Regulatory network reconstruction.....	61
10am ★ <u>Byung-Jun Yoon</u> : Accurate and reliable cancer classification....	62

*Coffee / Snacks / Fruit Break – 10:15-10:45am*

10:45 <u>Mark Brynildsen</u> : Metabolic strategies to enhance antibiotics ...	63
11am <u>Andrej Bugrim</u> : Role of growth factor signaling network.....	64
11:15 <u>Antti Larjo</u> : Simulating chemotactic and metabolic response ....	65
11:30 <u>Manway Liu</u> : Gene network analysis of diabetes.....	66
11:45 <u>Arjun Raj</u> : Variability in gene expression.....	67

*Lunch Break / Networking Opportunities – 12-1pm*

1pm ► <u>Ihor R. Lemischka</u> : Systems level approaches to stem cell fate..	68
1:30 <u>Dennis Vitkup</u> : Prediction of human disease genes.....	69
1:45 <u>Tomer Shlomi</u> : Predicting metabolic engineering KO Strategies	70
2pm ★ <u>Jun Zhu</u> : Dynamic changes in the blood transcriptional network.	71

*Coffee / Snacks / Fruit Break – 2:15-2:45pm*

2:45★	<u>Tal Peleg</u> : Network-free Inference of knockout effects.....	72
3pm	<u>Uri David Akavia</u> : A Bayesian framework to detect drivers .....	73
3:15	<u>Aedin Culhane</u> : Large scale analysis of stem cell expression .....	74
3:30	<u>Gang Fang</u> : Subspace differential coexpression analysis.....	75

**Systems Biology Poster Session II – 3:45pm-5:15pm**

*Hors-d'oeuvres, snacks, refreshments*

5:15▶	<u>Edward Marcotte</u> : Insights into evolution and disease .....	76
5:45	<u>Kenzie Maclsaac</u> : Condition specific master regulators.....	77
6pm	<u>Franck Rapaport</u> : Copy number alterations in cancer.....	78
6:15	<u>Guy Zinman</u> : New insights into cross-species conservation .....	79

*Light snacks – 6:30-6:45pm*

6:45★	<u>Theodore J. Perkins</u> : Structure of cellular networks.....	80
7pm★	<u>Önder Kartal</u> : Robustness as an evolutionary design principle..	81
7:15	<u>Jonathan Bieler</u> : Modeling 3D Flies .....	82
7:30▶	<u>John Reinitz</u> : Finding the rules by asking the right questions ....	83

**Museum Reception 8pm – 11pm**

*Warm Food, Cash Bar, Jazz Band, Wild Robots*

**Sunday, Dec. 6, 2009**

---

*Breakfast – 8am*

9am▶	<u>Walter Fontana</u> : Combinatorial complexity in systems biology ..	84
9:30	<u>Elhanan Borenstein</u> : Super-metabolism microbial communities .	85
9:45	<u>Niels Klitgord</u> : Predicting synthetic environments.....	86
10am	<u>Vebjorn Ljosa</u> : Large-scale learning of cellular phenotypes.....	87

*Coffee / Snacks / Fruit Break – 10:15-10:45am*

10:45	<u>Sarah Richardson</u> : Design of synthetic chromosomes .....	88
11:00	<u>Tamir Tuller</u> : Reconstructing ancestral gene content .....	89
11:30	Business meeting and announcement of next year's venue	
11:45	Closing remarks and Adjourn---Poster take-down	

# RECORD REGULATORY GENOMICS, SYSTEMS BIOLOGY, AND DREAMS

MIT / BROAD INSTITUTE  
DEC 3-6, 2009

compbio.mit.edu/recombstat

## CONFERENCE CHAIRS:

**MANOLIS KELLIS**  
**ZIV BAR-JOSEPH**  
**ANDREA CALIFANO**  
**GUSTAVO STOLOVITZKY**

Wednesday, Dec 2	
3pm	Conference check-in open. Poster session 1 set-up.
5pm	Welcome Remarks
5:15	5:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks
5:45	5:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
6:15	6:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
6:30	6:30 Break / Light meals
6:45	6:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
7pm	7pm <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
7:15	7:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
7:30	7:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
7:45	7:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
8pm	8pm <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
8:15-9:45p	Regulatory Genomics, Welcome Reception, Poster Session (Non-dreams, snacks, ref. refreshments, wine, cash bar)

Thursday, Dec 3	
8am	Breakfast
9am	9am <b>Rob Waterston:</b> Programming Cell Fate
9:30	9:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
9:45	9:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10am	10am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:15	10:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:45	10:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11am	11am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:15	11:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:30	11:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:45	11:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
12pm	12pm Lunch Break: Networking Opportunities
1pm	1pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:30	1:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:45	1:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2pm	2pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:15	2:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:45	2:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3pm	3pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:15	3:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:30	3:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:45	3:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning

Friday, Dec 4	
8am	Breakfast
9am	9am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
9:30	9:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
9:45	9:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10am	10am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:15	10:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:45	10:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11am	11am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:15	11:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:30	11:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:45	11:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
12pm	12pm Lunch Break: Networking Opportunities
1pm	1pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:30	1:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:45	1:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2pm	2pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:15	2:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:45	2:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3pm	3pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:15	3:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:30	3:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:45	3:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning

Saturday, Dec 5	
8am	Breakfast
9am	9am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
9:30	9:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
9:45	9:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10am	10am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:15	10:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:45	10:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11am	11am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:15	11:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:30	11:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:45	11:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
12pm	12pm Lunch Break: Networking Opportunities
1pm	1pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:30	1:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:45	1:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2pm	2pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:15	2:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:45	2:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3pm	3pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:15	3:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:30	3:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:45	3:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning

Sunday, Dec 6	
8am	Breakfast
9am	9am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
9:30	9:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
9:45	9:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10am	10am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:15	10:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
10:45	10:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11am	11am <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:15	11:15 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:30	11:30 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
11:45	11:45 <b>John D. Cherry:</b> Evidence for Quantitative Transcription Networks: Correlations and a Challenge
12pm	12pm Lunch Break: Networking Opportunities
1pm	1pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:30	1:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
1:45	1:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2pm	2pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:15	2:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
2:45	2:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3pm	3pm <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:15	3:15 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:30	3:30 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning
3:45	3:45 <b>Manolis Kellis:</b> Evolution of Nucleosome Positioning

## REGULATORY GENOMICS KEYNOTES

- BOB WATERSTON**
- RICK YOUNG**
- NIKOLAUS BARJAK**
- MIKOŁAJ BAJEWIKY**
- KEYVIN WHITE**

## DREAMS KEYNOTE SPEAKERS:

- GARRY NOLAN**
- MICHAEL YARFF**

## SYSTEMS BIOLOGY KEYNOTES

- NEVAN J. KRUGAN**
- FRANKISKA MICHOR**
- JEFF B. BOKEE**
- HOR R. LEMIS CHIKI**
- EDWARD MARCOTTE**
- JOHN REINITZ**
- WALTER FONTANA**

## PARTNER JOURNALS:



RG2: R. Shmueli, R. Milošević, Modeling and Recognition of Regulatory motifs and Inducers

RG1: U. Chirikov, Brent, Post-transcriptional regulation and regulatory RNAs



## Evidence for Quantitative Transcription Networks

Biggin, M.D.<sup>1</sup>, Kaplan, T.<sup>1</sup>, Aswani, A.<sup>1</sup>, Li, X.Y.<sup>1</sup>, Thomas S.<sup>2</sup>, Sabo, P.<sup>2</sup>, Brown, J.B.<sup>1</sup>, Boley, N.<sup>1</sup>, Atherton, J.<sup>1</sup>, Li, J.<sup>1</sup>, Davidson, S.M.<sup>1</sup>, Fisher, B.<sup>1</sup>, Hammonds, A.<sup>1</sup>, MacArthur, S.<sup>1</sup>, Fowlkes, C.C.<sup>1</sup>, Luengo Hendriks, C.L.<sup>1</sup>, Keränen, S.V.E.<sup>1</sup>, Hechmer, A.<sup>1</sup>, Simirenko, L.<sup>1</sup>, Malik, J.<sup>1</sup>, Knowles, D.W.<sup>1</sup>, Tomlin, C.<sup>1</sup>, Bickel, P.<sup>1</sup>, Stamatoyannopoulos, J.<sup>2</sup>, Celniker, S.<sup>1</sup>, Eisen, M.B.<sup>1</sup>

<sup>1</sup>Berkeley *Drosophila* Transcription Network Project, Lawrence Berkeley National Laboratory and UC Berkeley, Berkeley, California 94720, USA

<sup>2</sup>University of Washington, Department of Genome Sciences, Seattle, WA 98195, USA

The Berkeley *Drosophila* Transcription Network Project

(<http://bdtntp.lbl.gov/Fly-Net/>) is developing wet laboratory and computational/mathematical methods to allow predictive modeling of animal transcription networks, using the early *Drosophila* embryo as a test case. System wide data sets for *in vitro* and *in vivo* DNA binding, 3D cellular resolution protein and mRNA expression, transgenic promoter expression, and DNA accessibility in chromatin have been established and are being analyzed in conjunction with available comparative genomic DNA sequence information. Our work suggests that transcription networks have an unexpected structure in which functionally distinct transcription factors show a quantitative continuum of binding to highly overlapping sets of thousands of genomic regions. Highly bound genes include strongly regulated known and likely targets, moderately bound genes include unexpected targets whose transcription is regulated weakly, and poorly bound genes include thousands of non-transcribed genes and other likely non-functional targets. Quantitative differences in binding to common targets correlate with each factors known regulatory specificity, though these specificities appear to be more fuzzy and less distinct than commonly assumed. We propose that this Quantitative Transcription Network structure will be general to all metazoans and derives from the broad DNA recognition properties of animal transcription factors and the relatively high concentrations at which they are expressed in cells causing them to bind to highly overlapping sets of open chromatin regions. Predictive computational models of DNA occupancy *in vivo* and the regulation of target gene expression will also be presented.

Invited Talk

## Widespread translational repression by upstream open reading frames (uORFs): implications for human phenotypic variation and disease

Sarah E. Calvo<sup>1-3</sup>, David J. Pagliarini<sup>1-3</sup>, Vamsi K. Mootha<sup>1-3</sup>

<sup>1</sup>Broad Institute; <sup>2</sup>Center for Human Genetic Research, Massachusetts General Hospital;

<sup>3</sup>Department of Systems Biology, Harvard Medical School

Much phenotypic variation is due to changes in gene regulation rather than coding sequence. However, we are currently unable to interpret most forms of non-coding variation. Even within well-defined 5' UTR regions, which contain important post-transcriptional regulatory elements, we can predict few consequences of sequence variation. Here we focus on characterizing the widespread impact and variation of one 5' UTR regulatory element, the upstream open reading frame (uORF).

uORFs are common mRNA elements defined by a start codon in the 5' UTR that is out-of-frame with the main coding sequence. In individual studies, uORFs have been shown to reduce mRNA stability and protein translation. However no study to date has investigated their global effect on protein expression. We report that uORFs correlate with significantly reduced protein expression of the downstream ORF, based on analysis of 11,649 matched mRNA and protein measurements from four published mammalian studies. We demonstrate that uORFs typically reduce protein expression by 30-80%, with a modest 0-30% decrease in mRNA levels. We identify uORF-altering polymorphisms in 509 human genes and demonstrate that these variants can alter protein levels. Additionally, we provide support for pathogenicity of uORF-altering mutations in 5 human diseases. Lastly, we present preliminary evidence of translational regulation by uORFs during stress response. Together, our results suggest that uORFs influence the protein expression of thousands of mammalian genes and that variation in these elements can influence human phenotype and disease.

## Learning the sequence and structure binding preferences of RNA-binding proteins from noisy affinity data

Hilal Kazan<sup>1</sup>, Debashish Ray<sup>2</sup>, Esther T Chan<sup>3</sup>, Timothy R Hughes<sup>2,3,4</sup>, Quaid Morris<sup>1,2,3,4</sup>

<sup>1</sup>Department of Computer Science, University of Toronto, Toronto, Ontario, Canada;

<sup>2</sup>Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada;

<sup>3</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada;

<sup>4</sup>Donnelley Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada

RNA-binding domains are among the most common protein domains in metazoan genomes and many RNA-binding proteins (RBPs) bind RNA and control critical pathways of post-transcriptional regulation (PTR) of gene expression including mRNA splicing, export, stability, and translation. However, despite their ubiquity and the importance of their functional role, their binding preferences are largely unknown. To fill this gap, *in vitro* and *in vivo* binding data for an increasing number of RBPs have been recently published. To infer the RBP sequence and structural preferences from these data, novel RBP-specific motif finding methods are necessary. Here, we introduce a new motif-finding model, *RNAcontext*, which is specifically designed for discovering RBP binding specificities from noisy affinity data. Its key contribution is the incorporation of both sequence and structure features during the motif search. We evaluated our method on recently published *RNAcompete* binding affinity data and we show that it can discover the RBP structure and sequence-binding preferences with greater accuracy than a number of existing approaches. Additionally, we investigated the increase in predictive accuracy of our method due to the representation of structure context, and observed improvements in a number of cases. *RNAcontext* not only recovers known sequence and structure preferences of RBPs but also predicts novel secondary structure preferences for SF2/ASF which is consistent with its recently reported *in vivo* binding sites.

## Towards Computational Prediction of MicroRNA Function and Activity

Igor Ulitsky<sup>1</sup>, Louise Laurent<sup>2</sup>, Ron Shamir<sup>1</sup>

<sup>1</sup> Blavatnik School of Computer Science, Faculty of Exact Sciences, Tel-Aviv University, Israel

<sup>2</sup> Department of Center for Regenerative Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

While it has been established that microRNAs (miRNAs) play key roles throughout development and are dysregulated in many human pathologies, the specific processes and pathways regulated by individual miRNAs are mostly unknown. Here, we use computational target predictions in order to automatically infer the processes affected by human miRNAs. Our approach improves upon standard statistical tools by addressing specific characteristics of miRNA regulation. Our analysis is based on a novel compendium of experimentally verified miRNA-pathway and miRNA-process associations that we constructed, which can be a useful resource by itself. Our method also predicts novel miRNA-regulated pathways, refines the annotation of miRNAs for which only crude functions are known, and assigns differential functions to miRNAs with closely related sequences. Applying our approach to groups of co-expressed genes allows us to identify miRNAs and genomic miRNA clusters with functional importance in specific stages of early human development.

Full Length Paper

## Inferring Key Transcriptional Regulators using H3K4me2 Marked Nucleosome Occupancy Data

Clifford A. Meyer<sup>1</sup>, H. Hansen He<sup>1,3</sup>, Bo Jiang<sup>2</sup>, Jun Liu<sup>2</sup>, Myles Brown<sup>3</sup>, X. Shirley Liu<sup>1</sup>

<sup>1</sup>Dept of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115, USA; <sup>2</sup>Dept of Statistics, Harvard University, Cambridge, MA 02138, USA; <sup>3</sup>Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115, USA

ChIP-chip and ChIP-seq are widely used to study the genome-wide location and regulatory mechanism of transcription factors. Studies of regulatory mechanisms based on these technologies have been limited by their ability to target only a single transcription factor at a time and the availability of suitable antibodies. We describe an alternative approach to detect dynamic and genome-wide activities of multiple transcription factors driving a biological process using nucleosome-resolution ChIP-seq of histone H3 lysine 4 dimethylation (H3K4me2). From genome-wide profiles of H3K4me2 in the prostate cancer cell line LNCaP before and after androgen stimulation, we observe that a common pattern of change in nucleosome organization near androgen receptor binding sites is one in which a nucleosome in the immediate vicinity of the binding site is destabilized while the pair of nucleosomes flanking the site are stabilized. To measure the strength of this **nucleosome stabilization-destabilization** pattern we introduce an **NSD** score which we show to be highly correlated with androgen receptor binding.

We developed a novel motif analysis algorithm to identify DNA sequence motifs that are associated with high NSD scores. The probability of a transcription factor being bound between a pair of H3K4me2 marked nucleosomes is modeled using a DNA sequence component and an NSD score component. The DNA sequence component is computed from the factor's known position weight matrix while the NSD score component parameters are estimated using the EM algorithm. This analysis applied to the H3K4me2 ChIP-seq data in hormone stimulated LNCaP cells revealed key transcriptional regulators of the androgen response, AR, FoxA1, Oct1 and NKX3.1 at sites which were confirmed using q-PCR. In addition, we tested our enhancer nucleosome analysis on published histone ChIP-seq data relating to the heat shock response in yeast, human CD4<sup>+</sup> T-cell activation and human hematopoietic stem cell differentiation. For each study we found the key transcriptional regulators to be among the transcription factors most highly associated with the NSD-score.

We demonstrate that nucleosome resolution ChIP-seq of H3K4me2 followed by bioinformatic analyses is an effective way to detect stimulus dependent transcription factor activity and to study the effect of nucleosome occupancy in regulating transcription factor binding in metazoan systems.

## Genome-wide discovery and characterization of chromatin states from combinatorial patterns of epigenetic marks

Jason Ernst<sup>1,2</sup>, Pouya Kheradpour<sup>1,2</sup>, Tarjei S. Mikkelsen<sup>2</sup>, Peter V. Kharchenko<sup>3,4</sup>, Peter J. Park<sup>3,4</sup>, Gary H. Karpen<sup>5,6</sup>, Bradley E. Bernstein<sup>7,8</sup>, Manolis Kellis<sup>1,2</sup>

<sup>1</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology; <sup>2</sup>Broad Institute of MIT and Harvard; <sup>3</sup>Harvard Partners Center for Genetics and Genomics; <sup>4</sup>Children's Hospital Informatics Program; <sup>5</sup>Department of Genome and Computational Biology, Lawrence Berkeley National Lab; <sup>6</sup>Department of Molecular Cell Biology, University of California at Berkeley; <sup>7</sup>Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital; <sup>8</sup>Department of Pathology, Harvard Medical School

Nucleosomes are subject to a multitude of histone modifications, leading to the hypothesis of a histone code, capable of encoding a plethora of epigenetic information about cellular state. However a strictly combinatorial usage of these marks would imply thousands of possible distinct *codes*, while only a small number of distinct and biologically-meaningful combinations of marks seem to occur in practice, which we refer to as '*chromatin states*'. It is still unknown however what these chromatin states are, motivating the need for systematic methods to discover them and characterize them functionally in a complete genome.

To address these challenges, we have developed ChromHMM, a computational method that can use ChIP-seq and ChIP-chip datasets to discover *de novo* distinct chromatin states across a complete genome. The method is based on a Multivariate Hidden Markov Model, where each chromatin state is associated with a vector of emissions indicating the probabilities of each combination of chromatin marks at each location of the genome. By learning these states automatically across a complete genome, we have been able to uncover the most informative combinations of marks, and discover distinct classes of promoters, enhancers, and transcribed regions, independent of any previous annotations.

Applying our method to a published set of 41 chromatin marks in Human CD4<sup>+</sup>T cells, we inferred a set of 51 distinct chromatin signatures, which we have systematically characterized using a wide range of independent data. To our delight, these 51 states corresponded to distinct functions in the genome, including both active and repressed promoters and enhancers, 5' ends of genes, 3' ends of genes, exons and introns, transcribed and repressed regions, several distinct types of heterochromatin, repeat elements, and several other noteworthy states, even though no genomic annotations were available to the method during the learning phase.

We also applied our method to recently generated data for 9 chromatin marks across 8 ENCODE cell types, which allowed us to study the dynamics of chromatin state across different cell types, and the genomic sequence elements associated with changes in chromatin state.

Lastly, we applied our method to a set of 14 chromatin marks across two cell types in *Drosophila*, while enables us to gain a systematic and unbiased view of the chromatin landscape in both a key model organism and compare and contrast with the human chromatin landscape

## Human DNA methylomes at single-base resolution reveal widespread cell-specific epigenetic signatures

Ryan Lister<sup>1\*</sup>, Mattia Pelizzola<sup>1\*</sup>, Robert H. Downen<sup>1</sup>, R. David Hawkins<sup>2</sup>, Gary Hon<sup>2</sup>, Julian Tonti-Filippini<sup>3</sup>, Joseph R. Nery<sup>1</sup>, Leonard Lee<sup>2</sup>, Zhen Ye<sup>2</sup>, Que-Minh Ngo<sup>2</sup>, Lee Edsall<sup>2</sup>, Jessica Antosiewicz-Bourget<sup>4,5</sup>, Ron Stewart<sup>4,5</sup>, Victor Ruotti<sup>4,5</sup>, A. Harvey Millar<sup>3</sup>, James A. Thomson<sup>4,5,6,7</sup>, Bing Ren<sup>2,8</sup>, and Joseph R. Ecker<sup>1†</sup>, for the Roadmap Epigenomics Program

<sup>1</sup>Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. <sup>2</sup>Ludwig Institute for Cancer Research, University of California San Diego, La Jolla, CA 92093, USA. <sup>3</sup>ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, WA 6009, Australia. <sup>4</sup>Morgridge Institute for Research, Madison, WI 53707, USA. <sup>5</sup>Genome Center of Wisconsin, Madison, WI 53706, USA. <sup>6</sup>Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, WI 53715, USA. <sup>7</sup>Department of Anatomy, University of Wisconsin-Madison, Madison, WI 53706, USA. <sup>8</sup>Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA 92093, USA. \*These authors contributed equally to this work.

DNA cytosine methylation is a central epigenetic modification that plays essential roles in cellular processes including genome regulation, development and disease. Here we present the first genome-wide, single-base resolution maps of methylated cytosines in a mammalian genome, from both human embryonic stem cells and fetal fibroblasts, along with comparative analysis of mRNA and small RNA components of the transcriptome, several histone modifications, and sites of DNA-protein interaction for several key regulatory factors. Widespread differences were identified in the composition and patterning of cytosine methylation between the two genomes. Nearly one-quarter of all methylation identified in embryonic stem cells was in a non-CG context, suggesting that they may utilize different methylation mechanisms to affect gene regulation. Methylation in non-CG contexts showed enrichment in gene bodies and depletion in protein binding sites and enhancers. Non-CG methylation disappeared upon induced differentiation of the embryonic stem cells, and was restored in induced pluripotent stem cells. We identified hundreds of differentially methylated regions proximal to genes involved in pluripotency and differentiation, and widespread reduced methylation levels in fibroblasts associated with lower transcriptional activity. These reference epigenomes provide a foundation for future studies exploring this key epigenetic modification in human disease and development.

## Fundamentally different strategies of gene regulation in bacterial and eukaryotes

Leonid A. Mirny<sup>1,2</sup>, Zeba Wunderlich<sup>3</sup>

<sup>1</sup>Harvard-MIT Division of Health Sciences and Technology; <sup>2</sup>Department of Physics, MIT;

<sup>3</sup>Department of Systems Biology, Harvard Medical School.

The regulation of gene expression relies on the molecular process of transcription factors (TFs) binding to specific DNA sites. To bind its cognate site, a TF has to recognize it among  $\sim 10^6$  alternative (decoy) sites in bacteria or  $\sim 10^9$  sites in eukaryotes. Using information theory, we ask whether individual TFs possess enough information for such remarkably precise recognition. Our novel information-theoretical analysis of more than 950 recently characterized TF motifs gives two strikingly different answers for bacteria and eukaryotes. Bacterial TFs contain sufficient information to recognize cognate sites in their genomes. However, while longer eukaryotic genomes demand a TF to contain more information (i.e. be more specific), we find that eukaryotic TFs are much less specific than bacterial TFs and do not contain sufficient information to find a cognate site among  $10^9$  decoys. This information deficiency has a profound biological implication: the widespread binding of eukaryotic TFs to thousands of spurious sites on accessible DNA. This and other predictions are confirmed by our genomic and bioinformatics analysis of TF binding and are supported by a wide range of experimental results demonstrating extensive non-functional TF binding. We show that the apparent paradox of information deficiency in eukaryotes can be resolved if regulatory regions contain clusters of sites bound by several TFs. The discovered promiscuity of eukaryotic TFs can be advantageous for combinatorial regulation, modular use of TFs and more evolvable gene regulation. Our information-theoretical approach provides a new framework for the interpretation of functional genomics measurements and for understanding gene regulation in eukaryotes.

## Deciphering the *C. elegans* embryonic regulatory network

R.H. Waterston<sup>1</sup>, J.I. Murray<sup>1</sup>, T. Boyle<sup>1</sup>, M. Boeck<sup>1</sup>, Z. Zhao<sup>1</sup>, D. Mace<sup>1</sup>, L. Hillier<sup>1</sup>, V. Reinke<sup>2</sup>, D. Miller III<sup>3</sup>, P. Green<sup>1</sup>, M. MacCoss<sup>1</sup>, M. Gerstein<sup>4</sup>, M. Snyder<sup>5</sup>, A. Hyman<sup>6</sup>, S. Kim<sup>7</sup>

<sup>1</sup>*Department of Genome Sciences, University of Washington*, <sup>2</sup>*Department of Genetics, Yale University*, <sup>3</sup>*Department of Cell and Developmental Biology, Vanderbilt University* <sup>4</sup>*Department of Molecular Biophysics and Biochemistry, Yale University*, <sup>5</sup>*Department of Genetics, Stanford University*, <sup>6</sup>*Max Planck Institute of Molecular Cell Biology and Genetics*, <sup>7</sup>*Department of Developmental Biology, Stanford University*.

The *C. elegans* embryo is an excellent system in which to study the full regulatory network that directs embryonic development. The fixed cell lineage simplifies integration of results across different experiments; all the transcription factors are known through the complete genome sequence; and powerful genetic tools allows the perturbation of pathways to test hypotheses. Our lab has been pursuing three projects related to the overall goal of understanding the regulatory network. As part of the NHGRI modENCODE project, we are refining the annotation of the transcribed genome, using massively parallel sequencing and tiling arrays to assay the transcripts from different stages, conditions, tissues and cell types. Also with modENCODE we are generating strains expressing GFP-tagged transcription factors for use in ChIP-seq assays to define the transcription factor binding sites. Thirdly, we are using our recently developed technology to assay gene expression with single cell resolution continuously through embryogenesis to generate precise digital data detailing the cells in which the transcription factors are expressed. We have curated expression patterns for over 100 genes that are now publicly available. These data reveal a complex variety of patterns, with some suggesting that a combinatoric lineage code might be used to specify cell identity, others showing positional or tissue specificity and still others showing left-right asymmetry. These data sets are informing experiments to elucidate the embryonic regulatory network in specific cells and lineages.

Invited Talk

## Programming Cell State

Richard Young

*Whitehead Institute and MIT*

Discovering how transcriptional regulatory circuitry establishes and maintains gene expression programs in mammalian cells is important for understanding the control of cell state, the process of development and the mechanisms involved in cellular reprogramming. We are mapping the regulatory circuitry of embryonic stem (ES) cells and induced pluripotent stem (iPS) cells by investigating how key regulators control the gene expression program responsible for self-renewal and pluripotency. We have identified novel transcription factors, chromatin regulators, signaling components and noncoding RNAs that contribute to ES cell regulatory circuitry. Various genome-wide technologies have been used to map how these regulators contribute to control of genome expression. Advances in our knowledge of this regulatory circuitry have provided insights into the mechanisms by which somatic cells are reprogrammed into iPS and other cell types and have revealed how the controls of cell state can be manipulated to enhance reprogramming. These new advances and insights provide the foundation for further understanding developmental processes and are facilitating efforts to manipulate cell fates for regenerative medicine.

Invited Talk

## Widespread RNA polymerase II recruitment and transcription at enhancers during stimulus-dependent gene expression

Tae-Kyung Kim<sup>1\*</sup>, Martin Hemberg<sup>2\*</sup>, Jesse M. Gray<sup>1\*</sup>, David A. Harmin<sup>1,3</sup>, Scott Kuersten<sup>4</sup>, Allen M. Costa<sup>1</sup>, Kellie Barbara-Haley<sup>5</sup>, Eirene Markenscoff-Papadimitriou<sup>6</sup>, Gabriel Kreiman<sup>2</sup>, Michael E. Greenberg<sup>1</sup>

*\*These authors contributed equally to this work.*<sup>1</sup>Department of Neurobiology, Harvard Medical School. <sup>2</sup>Department of Ophthalmology, Children's Hospital Boston, and Center for Brain Science & Swartz Center for Theoretical Neuroscience, Harvard University. <sup>3</sup>Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology. <sup>4</sup>Life Technologies. <sup>5</sup>Molecular Genetics Core facility, Children's Hospital Boston. <sup>6</sup>Graduate Program in Neuroscience, University of California San Francisco.

During development and in mature organisms cells respond to changes in their environment in part through changes in gene expression. Extracellular factors including growth factors, hormones, and neurotransmitters activate programs of gene expression in a manner that is temporally and spatially controlled by the coordinated action of trans-acting transcription factors that bind to *cis*-acting DNA regulatory elements including enhancers, insulators, and promoters. Most studies of the mechanisms by which gene expression is induced in response to extracellular stimuli have focused on promoters, which lie adjacent to the site at which mRNA synthesis is initiated. In contrast, the mechanisms by which enhancers, which lie far away from the start site of mRNA synthesis, contribute to stimulus-dependent gene expression are not well characterized. We used genome-wide sequencing methods to study stimulus-dependent enhancer function in neurons. We identified over 20,000 neuronal activity-regulated enhancers that are bound by the general transcriptional co-activator CBP in an activity-dependent manner. A function of CBP at enhancers may be to recruit RNA polymerase II (RNAPII), as we also observed activity-regulated RNAPII binding to thousands of enhancers. Remarkably, RNAPII at enhancers transcribes bi-directionally a novel class of non-polyadenylated enhancer RNAs (eRNAs) within enhancer domains defined by the presence of histone H3 that is mono-methylated at lysine 4 (H3K4Me1). The level of eRNA expression at neuronal enhancers positively correlates with the level of mRNA synthesis at nearby genes, suggesting that eRNA synthesis occurs specifically at enhancers that are actively engaged in promoting mRNA synthesis. These findings reveal that a widespread mechanism of enhancer activation may involve RNAPII binding and eRNA synthesis.

## Inferring binding energies from selected binding sites

Yue Zhao<sup>1</sup>, David Granas<sup>1</sup>, Gary D. Stormo<sup>1</sup>

<sup>1</sup>*Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110*

We employ a biophysical model that accounts for the non-linear relationship between binding energy and the statistics of selected binding sites. The model includes the chemical potential of the transcription factor, non-specific binding affinity of the protein for DNA as well as sequence-specific parameters that may include non-independent contributions of bases to the interaction.

We obtain maximum likelihood estimates for all of the parameters and compare the results to standard probabilistic methods of parameter estimation. On simulated data, where the true energy model is known and samples are generated with a variety of parameter values, we show that our method returns much more accurate estimates of the true parameters and much better predictions of the selected binding site distributions. We also introduce a new high-throughput SELEX (HT-SELEX) procedure to determine the binding specificity of a transcription factor in which the initial randomized library and the selected sites are sequenced with next generation methods that return hundreds of thousands of sites. We show that after a single round of selection our method can estimate binding parameters that give very good fits to the selected site distributions, much better than standard motif identification algorithms.

Full Length Paper

## Extensive Binding Site Turnover in the Core Regulatory Network of Embryonic Stem Cells

Guillaume Bourque<sup>1</sup>, Galih Kunarso<sup>1</sup>, Justin Jeyakani<sup>1</sup>, Catalina Hwang<sup>1</sup>, Huck-Hui Ng<sup>2</sup>

<sup>1</sup>Computational & Mathematical Biology; <sup>2</sup>Stem Cell & Developmental Biology, Genome Institute of Singapore, 138672, Singapore.

Detection of novelty in genomic control elements is critical for understanding the wiring of transcriptional regulatory networks in their entirety. To systematically investigate the evolution of response elements in a mammalian gene regulatory network, we generated chromatin immunoprecipitation sequencing data sets for three important regulatory proteins in human embryonic stem (ES) cells (OCT4, NANOG and CTCF) and compared them with matching data sets obtained in mouse ES cells. Interestingly, and in contrast to CTCF, we found that the occupancy profiles of OCT4 and NANOG were drastically different with only ~5% of the regions most enriched in human being also enriched in mouse. Moreover, we established that a significant proportion of the innovation in control elements has been provided by transposable elements in both species. To assess the functional impact of these occupancy differences, we focused on genes with changes in expression levels after OCT4 depletion in human and mouse ES, respectively. By comparing the expression profiles, we identified a group of genes with conserved expression between species. However, even around those genes we found very few instances of conserved binding sites. Our results demonstrate the high rate of binding site turnover around key target genes of the mammalian ES cells regulatory network and implicate transposable elements as contributors to this evolutionary plasticity.

## Identification and analysis of cis-regulatory regions based on pattern generating potential

Majid Kazemian<sup>1\*</sup>, Charles Blatti<sup>1\*</sup>, Adam Richards<sup>2</sup>, Michael McCutchan<sup>3</sup>, Sudhir Kumar<sup>3</sup>, Scot Wolfe<sup>2</sup>, Saurabh Sinha<sup>1</sup>, Michael Brodsky<sup>2</sup>.

<sup>1</sup>University of Illinois at Urbana-Champaign; <sup>2</sup>University of Massachusetts Medical School; <sup>3</sup>Arizona State University; \*Equal contribution.

Computational analysis of cis-regulatory modules in a genome commonly relies on identifying statistically significant clusters of putative binding sites for transcription factors in a common regulatory network. We describe a new approach that uses both the expression pattern and DNA binding specificity of transcription factors to identify and characterize target modules that regulate patterned gene expression. First, transcription factor binding motif profiles are combined across multiple species using a statistical method that accommodates an irregular distribution of evolutionary distances. Next, a logistic regression model is created to relate these motif profiles and transcription factor expression patterns to target gene expression. The regression model is then used to scan the genome for segments with the potential to generate all or part of the patterned expression of a flanking gene. Finally, an *in silico* genetic analysis is used to infer edges in the transcriptional regulatory network depicting the direct contribution of individual factors to the activity of each module. Estimates of statistical significance are generated for each edge.

We apply this method to the experimentally well-characterized transcriptional regulatory network that controls anterior-posterior (A-P) patterning in the early *Drosophila* embryo. Cross-validation with *bona fide* modules indicates that using pattern generating potential for module identification gives greater specificity than purely motif enrichment-based approaches. We identify nearly 300 genomic regions with a statistically significant capacity to contribute to patterned gene expression during *Drosophila* embryo segmentation. These include a high frequency of apparently “redundant modules” driving similar expression patterns for the same gene. The network model generated with this method recovers the majority of factor-module edges identified by previous experimental studies and provides a detailed view of how protein-module interactions contribute to patterned gene expression. Surprisingly, we find that predictions of expression patterns, modules and network edges using evolutionarily-conserved motif profiles in this model is more specific than predictions using chromatin immunoprecipitation (ChIP) data. Direct comparison of ChIP data with gene expression patterns and with binding motif profiles reveals examples of apparent non-functional factor occupancy as well as lineage-specific, functional binding sites. We propose that measurement of pattern generating potential provides a general solution to integrate genome sequence, expression data and factor binding specificities to describe the complex transcriptional regulatory networks that function during metazoan development.



## A biophysical model for analysis of transcription factor interaction, binding site arrangement and regulatory target genes from genome-wide binding data

Xin He<sup>1</sup>, Chieh-Chun Chen<sup>2</sup>, Feng Hong<sup>3</sup>, Fang Fang<sup>4</sup>, Saurabh Sinha<sup>1</sup>, Huck-Hui Ng<sup>4</sup>, Sheng Zhong<sup>2,1,3,5</sup>

<sup>1</sup> Department of Computer Science, <sup>2</sup> Department of Bioengineering, <sup>3</sup> Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 61820.

<sup>4</sup> Gene Regulation Laboratory, Genome Institute of Singapore, Singapore, 138672.

How transcription factors (TFs) interact with cis-regulatory sequences and interact with each other is a fundamental, but not well understood, aspect of gene regulation. We present a computational method to address this question, relying on the established biophysical principles. This method, STAP (sequence to affinity prediction), can be applied to analyze large scale TF-DNA binding data, in the form of ChIP-chip or ChIP-seq, to discover cooperative interactions among TFs, infer the rules of interaction and predict new TF targets in sequences not covered by the experimental data. The distinctions between STAP and other statistical approaches for analyzing cis-regulatory sequences include the utility of physical principles and the treatment of the DNA binding data as quantitative representation of binding strengths. Applying this method to the ChIP-seq data of 12 TFs in mouse embryonic stem (ES) cells, we found that the strength of TF-DNA binding could be significantly modulated by cooperative interactions among TFs with adjacent binding sites. However, further analysis on five putatively interacting TF pairs suggests that such interactions may be relatively insensitive to the distance and orientation of binding sites. Testing a set of putative Nanog motifs, STAP showed that a novel Nanog motif could better explain the ChIP-seq data than previously published ones. We then experimentally tested and verified the new Nanog motif. A series of comparisons showed that STAP has more predictive power than several state-of-the-art methods for cis-regulatory sequence analysis. We took advantage of this power to study the evolution of TF-target relationship in *Drosophila*. By learning the TF-DNA interaction models from the ChIP-chip data of *D. melanogaster* (Mel) and applying them to the genome of *D. pseudoobscura* (Pse), we found that only about half of the sequences strongly bound by TFs in Mel have high binding affinities in Pse. We show that prediction of functional TF targets from ChIP-chip data can be improved by using the conservation of STAP predicted affinities as an additional filter.

Full Length Paper

## Regulatory motifs associated with TF binding and chromatin dynamics in *Drosophila* and Mammalian genomes

Pouya Kheradpour<sup>1</sup>, Christopher Bristow<sup>1</sup>, Jason Ernst<sup>1</sup>, Rachel Sealfon<sup>1</sup>, and Manolis Kellis<sup>1,2</sup>

<sup>1</sup>CSAIL of MIT and <sup>2</sup>Broad Institute of MIT and Harvard

Due to their central role in a number of processes, the global characterization of regulatory sequences is an important goal in molecular biology. While, experimental methods enable the identification of broad regions of regulatory importance, understanding the genomic basis of gene regulation requires identifying specific motif instances and the underlying grammatical structures necessary for their function. We previously demonstrated how a large number of genomes can be used to discover motif patterns and identify individual binding sites of motifs, while being robust to evolutionary turnover and sequencing or alignment errors (Xie 2005, Stark 2007, Kheradpour 2007).

We have extended our methods for use with the 29 sequenced placental mammals, which have much larger genome sizes than the 12 flies, higher repeat content, more highly varied nucleotide composition (from, e.g. CpG islands), and in many cases lower coverage sequencing (2X vs. 8-10X for the flies). We have used the larger number of sequenced vertebrates to study the trade-offs of using species at varying evolutionary distances, and we find that in many cases discovery power is optimal when analysis is restricted to the placental mammals, suggesting a high level of placental-specific regulatory elements.

We also applied these methods to chromatin modification datasets across developmental time points in the fly, and across several cell types in both human and fly in the context of the ENCODE and modENCODE projects in collaboration with the White, Karpen and Bernstein labs. We found motifs that show preferential enrichment and depletion patterns in specific stages and cell types, indicating factors that may play an important role in the establishment of chromatin modifications in each condition. We also find a general association between the expression of an activating factor and the enrichment of its corresponding motif in activating chromatin marks (but a corresponding depletion in repressive marks).

We further used motif occurrences to study the regulatory roles of transcription factors in developmental processes, using, in collaboration with the Celniker lab, the annotation of spatial and temporal gene expression patterns during *Drosophila* embryogenesis for 6000 genes including nearly all transcription factors. By following tissue lineages across stages, and determining the genes that change expression at a given stage, we are able to distinguish between activating and repressing factors in numerous cases.

Lastly, we have developed a number of motif instance signatures for transcription factor cooperation, and applied these to experimentally derived binding locations for a comprehensive set of transcription factors in fly. We find that over-representation of instances near one another, a mutual increase in conservation, and biases in the strands and ordering of their instances all correlate with interacting factors. Given pervasive cooperative binding between factors, we seek motif grammars that may lead to an improved understanding of the genomic basis for cooperative or antagonistic binding, and an enhanced ability to explain transcription factor binding and function at the systems level.

## Edgetic perturbation models of human inherited disorders

Quan Zhong, Marc Vidal

Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA

Cellular functions are mediated through complex systems of macromolecules and metabolites linked through biochemical and physical interactions, represented in interactome models as 'nodes' and 'edges', respectively. Better understanding of genotype-to-phenotype relationships in human disease will require modeling of how disease-causing mutations affect systems or interactome properties. Here we investigate how perturbations of interactome networks may differ between complete loss of gene products ('node removal') and interaction-specific or edge-specific ('edgetic') alterations. Global computational analyses of ~50 000 known causative mutations in human Mendelian disorders revealed clear separations of mutations probably corresponding to those of node removal versus edgetic perturbations. Experimental characterization of mutant alleles in various disorders identified diverse edgetic interaction profiles of mutant proteins, which correlated with distinct structural properties of disease proteins and disease mechanisms. Edgetic perturbations seem to confer distinct functional consequences from node removal because a large fraction of cases in which a single gene is linked to multiple disorders can be modeled by distinguishing edgetic network perturbations. Edgetic network perturbation models might improve both the understanding of dissemination of disease alleles in human populations and the development of molecular therapeutic strategies.

## Evolution of nucleosome positioning

Naama Barkai, Itay Tirosh

*Weizmann institute of Science, Rehovot, Israel*

Phenotypic diversity is largely driven by changes in gene regulation. Recent studies characterized extensive differences in the gene expression program of closely related species. In contrast, virtually nothing is known about the evolution of chromatin structure and how it influences the divergence of gene expression.

We analyzed the evolution of nucleosome positioning between closely related yeast species and in their inter-specific hybrid. I will discuss results of this analysis, and present evolutionary insights on the genetic determinants and regulatory function of nucleosome positioning.

## Use of structural DNA properties for the prediction of regulator binding sites with conditional random fields

Pieter Meysman<sup>1</sup>, Kristof Engelen<sup>1</sup>, Kris Laukens<sup>2</sup>, Thanh Hai Dang<sup>2</sup> and Kathleen Marchal<sup>1</sup>

<sup>1</sup>Department of Microbial and Molecular systems, K.U.Leuven, Kasteelpark Arenberg 20, B-3001 Leuven Heverlee, Belgium. <sup>2</sup>Intelligent Systems Laboratory, Department of Mathematics and Computer Science, Middelheimlaan 1, B-2020 Antwerpen, Belgium.

Molecular recognition of genomic target sites by regulator proteins is a vital process in the transcription regulation of genes in living cells. The types of physical interactions that contribute to the recognition of binding sites by a protein can roughly be divided into those enabling direct read-out and those that allow for indirect read-out. The former comprises base-specific recognition, such as stabilizing hydrogen bonds between regulator amino acids and a set of conserved bases in the genomic DNA sequence, while in the case of the latter variations within the DNA structure will be used as the basis for recognition. It is the direct form of recognition that is the focus of most current endeavors to model regulator binding sites, usually by modeling a conserved set of nucleotides, e.g. a position weight matrix (PWM). However by considering only a single recognition mechanism, these models overlook any information concerning binding site identity that can be derived from the use of indirect read-out by the regulator. It was therefore our goal to create a binding site model based on this second type of recognition which involves interactions between the regulator protein and the molecular structure of the DNA molecule.

The structural DNA properties of the binding sites, needed to construct the model, are derived from their nucleotide sequence using a number of higher-order value look-up functions, so-called structural scales, which are based on experimental data (e.g. X-ray crystallography of various DNA molecules). These structural properties were used as input data to train a model representing the common structural features shared by all known binding sites of a specified regulator. This was done using conditional random fields (CRF), a discriminative machine learning method. Two novel extension algorithms were included in the training of the models, namely an optimization method which allows the CRF to work with structural DNA properties, and a correction method which can compensate for any bias in the training set towards nucleotide conservation. Once trained, the models could be used to evaluate the likelihood of regulator binding for any given DNA sequence.

The performance of these models was demonstrated on data sets for 27 different *Escherichia coli* transcription factors and they show an overall improvement when compared to a standard PWM model and a previously proposed structural property model [1]. Further a set of novel binding site predictions, resulting from use of the trained models in a genome wide screening of *E. coli*, were validated using a microarray compendium of ca. 1500 arrays, as well as comparison with EcoCyc.

[1] Nikolajewa S, Pudimat R, Hiller M, Platzer M & Backofen R. (2007) BioBayesNet: a web server for feature extraction and Bayesian network modeling of biological sequence data. Nucl Acid Res 35: W688-W693.

## Nucleosomes are positioned in exons and have histone marks suggesting co-transcriptional splicing

Claes Wadelius<sup>1</sup>, Robin Andersson<sup>2</sup>, Stefan Enroth<sup>2</sup>, Alvaro Rada-Iglesias<sup>2</sup>, Francisco de la Vega<sup>3</sup>, Kevin McKernan<sup>4</sup>, Jan Komorowski<sup>2,5</sup>

<sup>1</sup> Dept Gen & Pathology, Uppsala Univ, Uppsala, Sweden; <sup>2</sup> Linnaeus Centre for Bioinformatics, Uppsala University, Sweden; <sup>3</sup> Life Technologies, Foster City, CA.; <sup>4</sup> Life Technologies, Beverly, MA; <sup>5</sup> Interdisciplinary Centre for Mathematical and Computer Modelling, Warsaw University, Poland

It is known that nucleosomes are well positioned over the first exon in active genes, with histone modifications reflecting the transcription rate. So far positioning of nucleosomes relative to other genomic features has not been determined and which histone modifications are located along a gene has only been partly analyzed.

We reanalyzed public nucleosome position data for man and *C. elegans*, histone modification data from man and mouse as well as gene and exon expression data from man to look for distinct patterns also in internal exons. We found one well-positioned nucleosome at internal exons with a signal clearly higher than at the TSS. The peaks are centered at +94 (human) and +101 (*C. elegans*) relative to the exon start meaning that the average 5' end of a nucleosome is positioned at +20 and +27 in man and worm, respectively. We found no positioned nucleosome in exons <50 bp, which comprise <5% of human exons, but in long exons there is a nucleosome positioned at the start and end. Nucleosomes were positioned at internal exons regardless of transcription level, in contrast to the situation at the TSS.

We systematically screened 38 histone modifications to see if the nucleosomes had distinct patterns related to gene and exon expression. The H3K36me3 signal was significantly higher in exons than in the following introns in highly expressed human and mouse genes,  $p < 10e-4$  for each exon-intron comparison. This applies from the third exon and onwards. On the other hand H3K27me2 and me3 were associated with gene silencing. In highly expressed genes, high exon usage was associated with high H3K36me3 and low H3K27me2 and the opposite was found in exons with low usage. These data suggest that H3K36me3 might facilitate exon inclusion during co-transcriptional splicing and that splicing is under epigenetic control. We have generated extremely deep data sets by sequencing nucleosomes, RNA and ChIP-DNA from HepG2 cells to further evaluate the findings. As expected we find that nucleosomes are positioned in exons also in these cells. Furthermore, we have mapped >5 million 50 bp reads to splice junctions and find numerous genes with alternative splice isoforms. We are currently analyzing how these events are related to histone modifications and signals from other nuclear proteins. Our results show that exons are functional units not only defined by their coding capacity but also by the way they are packaged in nucleosomes. The factors controlling nucleosome positioning at internal exons must be under strong evolutionary constraint given the strikingly similar pattern in man and worm, with a common ancestor around 1 billion years ago.

## Integrated Approach for the Identification of Human HNF4 $\alpha$ Target Genes Using Protein Binding Microarrays

Eugene Bolotin<sup>1,6</sup>, Hailing Liao<sup>4</sup>, Tuong Chi Ta<sup>2</sup>, Chuhu Yang<sup>1</sup>, Wendy Hwang-Verslues<sup>3</sup>, Jane R. Evans<sup>2</sup>, Tao Jiang<sup>5,6</sup>, and Frances Sladek<sup>2,6</sup>

<sup>1</sup>Genetics, Genomics and Bioinformatics; <sup>2</sup>Cell, Molecular, and Developmental Biology; and <sup>3</sup>Environmental Toxicology Graduate Programs; Departments of <sup>4</sup>Cell Biology and Neuroscience, <sup>5</sup>Computer Science and Engineering, <sup>6</sup>Institute for Integrated Genome Biology, University of California, Riverside, CA, 92521-0121 USA

Hepatocyte nuclear factor 4 alpha (HNF4 $\alpha$ ), a member of the nuclear receptor superfamily, is essential for liver function and linked to several diseases including diabetes, hemophilia, atherosclerosis and hepatitis. While many DNA response elements and target genes have been identified for HNF4 $\alpha$ , the complete repertoire of binding sites and target genes in the human genome is unknown. We adapt protein binding microarrays (PBMs) to examine the DNA binding characteristics of two HNF4 $\alpha$  species (rat and human) and isoforms (HNF4 $\alpha$ 2 and HNF4 $\alpha$ 8) in a high throughput fashion. We identified ~1,400 new binding sequences and used this dataset to successfully train a Support Vector Machine (SVM) model that predicts an additional ~10,000 unique HNF4 $\alpha$  binding sequences; we also identify new rules for HNF4 $\alpha$  DNA binding. We performed expression profiling of an HNF4 $\alpha$  RNAi knockdown in HepG2 cells and compared the results to a search of the promoters of all human genes with the PBM and SVM models, as well as published genome-wide location analysis. Using this integrated approach, we identified ~240 new direct HNF4 $\alpha$  human target genes, including new functional categories of genes not typically associated with HNF4 $\alpha$ , such as cell cycle, immune function, apoptosis, stress response and other cancer-related genes. In conclusion, we report the first use of PBMs with a full length native transcription factor in a crude nuclear extract, additionally we greatly expand the repertoire of HNF4 $\alpha$  binding sequences and target genes, thereby identifying new functions for HNF4 $\alpha$ . We also establish a web-based tool, HNF4 Motif Finder, that can be used to identify potential HNF4 $\alpha$  binding sites in any sequence.

## Genome-wide mapping and computational analysis of non-B DNA structures *in vivo*

Damian Wójtowicz<sup>1,5</sup>, Fedor Kouzine<sup>2</sup>, Arito Yamane<sup>3</sup>, Craig Benham<sup>4</sup>, Rafael Casellas<sup>3</sup>, David Levens<sup>2,\*</sup>, Teresa Przytycka<sup>1,\*</sup>

<sup>1</sup>Computational Biology Branch, NCBI/NIH; <sup>2</sup>Laboratory of Pathology, NCI/NIH; <sup>3</sup>Genomic Integrity and Immunity, NIAMS/NIH; Genome Center, UC Davis, CA; <sup>4</sup>University of Warsaw, Poland;

\*corresponding authors

The Watson-Crick structure is the natural state of DNA in a genome and it is known as B-DNA. However, DNA is a dynamic molecule that undergoes various deformations and adopts several alternative secondary structures such as single-stranded DNA, Z-DNA (left-handed), H-DNA (triplex), cruciform, or quadruplex. Although previous studies confirmed the existence of non-B DNA conformations at some particular sites and implicated their role in DNA transactions, e.g. the tight regulation of c-myc oncogene by FUSE element, it was not known how general such a regulation mechanism might be. Indeed, little is still known about the formation of non-B DNA conformations and their role on genomic scale.

As a result of concerted effort of experimental and computational groups we now obtained the first draft of genome-wide mapping of alternative DNA structures *in vivo*. Specifically, with a novel experimental technique developed in David Levens' group, extensive experimental analysis performed in David Levens' and Rafael Casellas' groups, and a computational analysis performed in Teresa Przytycka's group, we have been able to perform the first genome-wide analysis of occurrences of alternative DNA structures in living cells, their position with respect to several genomic marks such as transcription start site, transcription termination site, etc. The experimental method is based on the fact that non-B DNA structure contains region of single-stranded DNA whose ends can be isolated and sequenced using Solexa high-throughput sequencing technique. The experiment was performed on activated mouse B-cells under different conditions: untreated cells and cells treated with DRB (a drug inhibiting the RNA polymerase II).

This presentation focuses on the computational component of this research. Short reads obtained from sequencing were mapped to the mouse genome (mm9), and simple statistical method was applied to find non-B DNA regions across the genome for both treated and untreated mouse B-cells. From our analysis, it is clear that sequences forming alternative DNA structures are non-randomly distributed in the genome. We found an enrichment of non-B DNA conformations near transcription start sites indicative of their potential role in gene regulation. We have compared, for the first time on such scale, the experimental data and the genomic regions computationally predicted to have a high propensity to form non-B DNA conformations. We found that not only experimentally detected non-B DNA regions have a significant overlap with computationally predicted regions, but also various computationally classified non-B DNA conformations have different experimentally derived signatures. This offers not only strong evidence for the *in vivo* formation of the alternative structures like Z-DNA, quadruplexes, and SIDD sites but also provides the first look at their genome-wide landscape and possible role in gene regulation.

## The Landscape of Transcription Initiation in *Drosophila*: Linking Initiation Patterns to Distinct Core Promoter Motifs

Elizabeth A Rach<sup>1,2</sup>, David L Corcoran<sup>2</sup>, Ting Ni<sup>2,3</sup>, Shen Song<sup>2,3</sup>, Eric Spana<sup>4</sup>, Yuan Gao<sup>5</sup>, Jun Zhu<sup>2,3</sup>, Uwe Ohler<sup>2,6,7</sup>

<sup>1</sup>PhD Program for Computational Biology and Bioinformatics; <sup>2</sup>Institute for Genome Sciences & Policy; <sup>3</sup>Department of Cell Biology; <sup>4</sup>Department of Biology, Duke University, 101 Science Drive, Durham, NC 27708, USA; <sup>5</sup>Department of Computer Science, School of Engineering, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA; <sup>6</sup>Department of Biostatistics and Bioinformatics, Duke University School of Medicine, 2301 Erwin Road, Durham, NC 27710; <sup>7</sup>Department of Computer Science, Duke University, LSRC Building D101, 450 Research Drive, Durham, NC 27708, USA

Transcription initiation is a key component in the regulation of genes, and is mediated through interactions of transcription factors with the sequence immediately surrounding the transcription start site (TSS) in the core promoter. Recent years have seen an increase in 5' capping data used to map full-length mRNAs in mammals, which has shown the prevalence of dispersed patterns, rather than a single site, of transcription initiation. However, it was so far unknown if these observations extend to other model species with a simpler genomic organization.

In this work, we developed a novel Paired-End Analysis of capped Transcripts (PEAT) sequencing strategy using the Illumina paired-end sequencing platform. On a 0-24 hour *Drosophila melanogaster* embryo sample, we obtained over 17.5M paired reads characterizing the 5' ends of transcripts, which were mapped to the genome and grouped into TSS clusters for more than 4,000 genes. We defined three TSS cluster shapes based on their peak height and span: Narrow Peak, Broad Peak, and Weak Peak. A fourth group of clusters mapped downstream of the initiation region in the coding sequence.

We examined the composition of canonical motifs within the core promoter (+/-100bp) sequence surrounding the TSS as defined by the mode in each cluster. Results showed that Narrow Peak promoters contained high frequencies of the canonical TATA/INR/DPE/MTE motifs, while Weak Peak promoters had an enrichment in motifs which have so far been largely computationally characterized (Ohler 1/DRE/Ohler 6/Ohler 7). A combination of motifs from both sets was observed in Broad Peaked promoters. We validated the clusters against ChIP-chip data for the TATA-box Binding Protein (TBP), as well as the TBP-Related Factor 2 (TRF2) which has been linked to the DRE motif. 62% of Narrow Peak promoters were bound by TBP, and 95% of Weak Peak promoters were bound by TRF2. Broad Peak promoters had a mixture of both TBP and TRF2 binding.

The +/-100bp sequences surrounding coding clusters lacked all motifs, and showed minimal binding of either factor assayed by ChIP-chip. Such clusters may be examples of recently reported re-capped transcripts, and we validated examples that such clusters correspond to genuinely capped transcripts.

This study greatly extends our previous efforts to map TSSs utilized by *D.melanogaster* during embryogenesis by using a high-throughput 5' paired end mapping strategy (PEAT) to define TSS locations within three distinct promoter shapes at a high level of precision. The set of canonical core promoter motifs exclusively available in *D.melanogaster* allowed us to characterize the promoter shapes based on their regulatory sequence elements, and gain a deeper knowledge of the association of promoter shapes to different transcription factors.

## TSS detection helps unfold promoters' structure

Julia Lasserre<sup>1</sup>, Alexander Zien<sup>2</sup>, Martin Vingron<sup>1</sup>

<sup>1</sup>MPI for Molecular Genetics, Germany,<sup>2</sup>Fraunhofer FIRST now LIFE Biosystems, Germany

In spite of dramatic experimental progress, computational detection of human promoters is still a challenging problem, partly because of the need to discover new ones, but also because identifying promoters is linked to understanding their structure. Various methods have been used (probabilistic, generative, discriminative) but today, the state of the art in promoter recognition remains ARTS [1], which uses a SVM together with a powerful kernel entirely based on the sequence. Another interesting promoter detector is PROSOM [2] which uses energy profiles to cluster sequences, and where some of the clusters are actually found to be promoter-specific.

In this work, we want to take the best of both worlds. We first compute the energy profiles of our promoter sequences and build  $K=10$  clusters. For each cluster  $k$ , a different dataset is sampled from our database of promoters and background sequences according to pre-computed  $p(\text{cluster}=k | \text{sequence})$ , and an ARTS-kernel SVM [1] is trained. Interestingly, when studying the CpG content, we find that most clusters are almost pure ( $K_{\text{pure}}=8$ ) but some contain both CpG-poor and CpG-rich promoters ( $K_{\text{mixed}}=2$ ). For the latter type of clusters, two SVMs are trained (one for each promoter type), where only the positive data changes. Finally, for a new sequence, the outputs of these ( $K_{\text{pure}} + 2 \cdot K_{\text{mixed}}$ ) SVMs are combined to make one final prediction.

The advantages of this approach are two-fold. First of all, the clustering part allows different types of features to be combined. In ARTS [1], the energy profiles were already used but as part of the kernel itself. Unfortunately, for computational reasons, the kernel had to be kept linear and therefore the only combination of motif features with energy profiles was an OR. Our method provides an AND. Indeed, to be classified as promoter, a sequence should show a particular energy profile and the sequence motifs that are attached to this profile.

Secondly, we can now extract the important  $k$ -mers for each SVM using POIMS [3], which means we can assign a different set of  $k$ -mers that separate promoters from the background to each energy profile. This is of great help to understand the different types of promoters. In particular, it is interesting to have an approach that can separate the CpG-rich promoters (usually easier to find) from the CpG-poor.

In terms of detection, our method performs similarly to a single ARTS-kernel SVM (AUC of  $\sim 0.85$ ). The POIMS show peaks at different locations for each SVM, which means that the promoter signal is different in each cluster. The  $k$ -mers also tend to be different between CpG-poor promoters and CpG-rich promoters. Most of the important  $k$ -mers in each SVM are usually from the same bank (typically the TATA box, the initiator, CACAC, CCCCC, GGGGG, and others) but they are combined differently. These results can help draw a map of different types of promoters.

[1] S. Sonnenburg, A. Zien, G. Raetch: "ARTS: accurate recognition of transcription starts in human", ISMB (Supplement of Bioinformatics) 2006, p472-480

[2] T. Abeel, Y. Saeys, P. Rouze, Y. Van de Peer: "ProSOM", Bioinformatics 2008, 24(13):pi214-i31

[3] S. Sonnenburg, A. Zien, P. Philips, G Raetch: "POIMS: positional oligomer importance matrices-- understanding support vector machine-based signal detectors", Bioinformatics 2008, 24(13):pi6-i14

## SPIKE - Signaling Pathways Knowledge-base and Analysis Tool

Igor Ulitsky<sup>1</sup>, Dorit Sagir<sup>2</sup>, Eyal David<sup>2</sup>, Yaara Ber<sup>3</sup>, Arnon Paz<sup>2</sup>, Zippi Brownstein<sup>2</sup>, Guy Karlebach<sup>1</sup>, Rani Elkon<sup>4</sup>, Karen B. Avraham<sup>2</sup>, Adi Kimchi<sup>3</sup>, Yossi Shiloh<sup>2</sup>, Ron Shamir<sup>1</sup>

<sup>1</sup>*Blavatnik School of Computer Science, Faculty of Exact Sciences, Tel-Aviv University, Israel*

<sup>2</sup>*Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University*

<sup>3</sup>*Department of Molecular Genetics, Weizmann Institute of Science, Israel*

<sup>4</sup>*Division of Tumor Biology, The Netherlands Cancer Institute, Amsterdam, The Netherlands.*

Signaling networks that govern cellular physiology form an intricate web of tightly regulated interlocking processes. The DNA damage response (DDR), which encompasses various DNA repair mechanisms, cell cycle check-points, apoptotic and non-apoptotic forms of programmed cell death, as well as general stress responses, exemplifies this overwhelming complexity. Data on these networks are accumulating at an unprecedented pace, and hence, their assimilation, visualization and interpretation have become a major challenge in biological research.

To cope with this challenge, we are developing the SPIKE knowledge-base of signaling pathways. SPIKE contains three main software components: 1) A database (DB) of biological signaling pathways. Carefully curated information from the literature and data from large public sources constitute distinct tiers of the DB. 2) A visualization package that allows interactive graphic representations of regulatory interactions stored in the DB and superposition of functional genomic and proteomic data on the maps. 3) An algorithmic inference engine that analyzes the networks for novel functional interplays between network components.

SPIKE's database contains extensive, up-to-date, and highly curated data on 14 pathways, including cell-cycle regulation, DNA repair, apoptosis, autophagy, MAPK signaling, auditory pathways and more. It is continuously updated with new experimental results, while a team of experts oversees the expansion of each network map.

SPIKE is available at: <http://www.cs.tau.ac.il/~spike/>.

Development of SPIKE is supported in part by the European Union projects APOSYS, TRIREME and EuroHear and by the Israel Science Foundation.

## Post-transcriptional gene regulation by small RNAs and RNA binding proteins

Prof. Dr. Nikolaus Rajewsky

*Max-Delbrück-Center for Molecular Medicine, Berlin-Buch*

In recent years it has become apparent that a large fraction of all genes in animals is regulated post-transcriptionally by small RNAs.

Furthermore, animal genomes contain hundreds of genes with RNA binding domains. It is clear that many of these RNA binding proteins have important and specific functions in mRNA localization and stability as well as in regulating protein production. However, only recently technologies have become available to probe post-transcriptional regulatory networks on a genome-wide scale.

I will briefly review previous efforts to understand more about the function of microRNAs in post transcriptional gene regulation. I will then present ongoing work where we use high throughput quantitative proteomics, next generation sequencing and computational approaches to unravel the biological function of small RNAs and RNA binding proteins in well defined in-vivo systems such as *C. elegans*.

## A transcriptional regulatory network for retinal differentiation in *Drosophila*.

Stein Aerts<sup>1,2</sup>, Xiao-Jiang Quan<sup>1,2</sup>, Annelies Claeys<sup>1,2</sup>, Jiekun Yan<sup>1,2</sup>,  
and Bassem A. Hassan<sup>1,2</sup>.

<sup>1</sup>Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven, Belgium. <sup>2</sup>Center for Human Genetics, K.U. Leuven School of Medicine, Leuven, Belgium.

We have applied a hybrid computational and genetics pipeline to identify functional target genes (TG) and target enhancers of transcription factors (TF) in the gene regulatory network underlying retinal differentiation in *Drosophila*.

In the first step we performed TF perturbations by the GAL4-UAS system for Atonal and Senseless, followed by qRT-PCR and microarray experiments to yield sets of genes that are co-expressed directly or indirectly downstream of the TF.

In the second step we aimed to determine *direct* TF-TG relations by the computational prediction of TF motifs and their instances in putative target regions. To this end we developed a novel method called cisTargetX to predict motif clusters across the entire 12 *Drosophila* genomes. Next, it determines significant associations between motifs and subsets of co-expressed genes. We validated cisTargetX on gene sets from published microarray experiments (e.g., downstream of dorsal, Mef2, serpent, pointed, ovo, biniou, Su(H), and Eyeless) and found the correct motif and targets for nearly all tested TFs. We then applied cisTargetX to Atonal and Senseless downstream genes and were able to identify novel motifs together with a significant enrichment of predicted direct targets for both TFs.

In the third step we tested several predicted enhancers through extensive *in vivo* validations using newly designed GFP-reporter vectors that allow for Gateway cloning and for PhiC31-mediated genomic integration. This procedure identified three known and fifteen novel Atonal target enhancers out of 31 tested predictions. These results show that expression studies combined with computational predictions are a powerful tool for regulatory network discovery.

Surprisingly, most retinal Atonal targets are also regulated by Atonal in proneural factor in external sense organ precursors. This raises the question of how developmental diversity can result from a common regulatory program. Mapping the expression profiles of the identified regulatory sequences shows that each sense organ is marked by a unique combinatorial code of proneural target enhancer activation. Furthermore, all target genes identified are modulators of key signaling pathways like Notch, EGFR and Wnt. This suggests that subtle differences in the transcriptional modulation of cellular signaling underlie the developmental and evolutionary diversity of sensory organs. internal sensory organ precursors and by the Scute.

## Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture

Raja Jothi<sup>1,5,\*</sup>, S Balaji<sup>2,5,6</sup>, Arthur Wuster<sup>3</sup>, Joshua A Grochow<sup>4</sup>, Jörg Gsponer<sup>3</sup>, Teresa M Przytycka<sup>2</sup>, L Aravind<sup>2</sup>, M Madan Babu<sup>3,\*</sup>

<sup>1</sup>*Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA;* <sup>2</sup>*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA;* <sup>3</sup>*MRC Laboratory of Molecular Biology, Cambridge, UK;* <sup>4</sup>*Department of Computer Science, University of Chicago, Chicago, IL, USA;* <sup>5</sup>*These authors contributed equally to this work;* <sup>6</sup>*Present address: Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.*

\*Correspondence: [jothi@mail.nih.gov](mailto:jothi@mail.nih.gov) or [madanm@mrc-lmb.cam.ac.uk](mailto:madanm@mrc-lmb.cam.ac.uk)

In an effort to find answers that could explain differential cell-fate outcome in response to the same uniform stimulus, we explored the link between regulatory network architecture and the genome-scale dynamics of the underlying entities such as genes, mRNAs, and proteins [1]. By classifying DNA-binding TFs in the yeast regulatory network into three hierarchical groups/layers (top, core, and bottom) and integrating diverse genome-scale datasets, we found that at the protein level, the top-layer TFs (which trigger/initiate regulatory cascades) are relatively abundant, long-lived, and showed more cell-to-cell variability compared to the downstream (core- and bottom-layer) TFs. This and other results led us to conclude that the variability in expression of top-layer TFs might confer a selective advantage, as this may permit at least some members in a clonal cell population to initiate an effective response to fluctuating environments, whereas the tight regulation of the core- and bottom-layer TFs may minimize noise propagation and ensure fidelity in regulation. This result is critical to understanding phenotypic variability in fluctuating environments, where individuals of the same population exhibit different phenotypes in response to the same stimulus, e.g., fractional survival or cell-death in clonal cell populations upon drug treatment in certain diseases such as cancer. We propose that this dynamic variability in expression level of key regulatory proteins permits differential sampling (i.e., the survival network or the apoptotic network) of the same underlying regulatory network (governing all cells) by different members in a clonal population, which might result in divergent cell-fate outcomes among different individuals in an otherwise identical cell population.

Our findings have implications in developmental processes and cellular differentiation that involve populations of cells, and in synthetic biology experiments aimed at engineering gene regulatory circuits. In particular, the dynamics of TFs in terms of their abundance, half-life, and noise cannot be ignored as modulating these attributes could affect the outcome of a regulatory cascade. This study takes us one step closer towards a better understanding of how cells adapt to changing environments, how different phenotypic outcomes are mediated in clonal cell populations, and how mutations that disrupt the dynamics of key regulatory proteins may influence disease conditions.

[1] Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, Przytycka TM, Aravind L, Babu MM. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular Systems Biology* 5:294 (2009).

## A pluripotency signature predicts histological transformation and influences survival in follicular lymphoma patients

Andrew J. Gentles<sup>1</sup>, Ash A. Alizadeh<sup>2,3</sup>, Su-In Lee<sup>4</sup>, June. H. Myklebust<sup>3</sup>, Catherine M. Shachaf<sup>5</sup>, Babak Shahbaba<sup>6</sup>, Ron Levy<sup>3</sup>, Daphne Koller<sup>4</sup>, Sylvia K. Plevritis<sup>1</sup>.

<sup>1</sup>Department of Radiology, Stanford University; Department of Medicine (Divisions of <sup>2</sup>Hematology and <sup>3</sup>Oncology), Stanford University; <sup>4</sup>Department of Computer Science, Stanford University; <sup>5</sup>Department of Microbiology and Immunology, Stanford University; <sup>6</sup>Department of Statistics, University of California at Irvine.

Histological transformation of follicular (FL) to diffuse large B cell lymphoma (DLBCL-t) is associated with accelerated disease course and drastically worse outcome, yet the underlying mechanisms are poorly understood. We show that a network of gene transcriptional modules underlies histological transformation (HT). Central to the network hierarchy is a signature that is strikingly enriched for pluripotency-related genes. These genes are typically expressed in embryonic stem cells (ESC), including MYC and its direct targets. This core ESC-like program was independent of proliferation/cell-cycle and overlapped, but was distinct from, normal B-cell transcriptional programs. Furthermore, we show that the ESC program is correlated with transcriptional programs maintaining tumor phenotype in transgenic MYC-driven mouse models of lymphoma. Although our approach was to identify HT mechanisms, rather than to derive an optimal survival predictor, a model based on ESC/differentiation programs stratified patient outcomes in the training dataset as well as in an independent validation set. The model was also predictive of propensity of FL tumors to transform. Transformation was associated with an expression signature combining high expression of ESC transcriptional programs in combination with reduced TGF- $\beta$  signaling. Together, these findings suggest a central role for an ESC-like signature in the mechanism of HT and provide new clues for potential therapeutic targets.

## Scalable Steady State Analysis of Boolean Biological Regulatory Networks

Ferhat Ay<sup>1</sup>, Fei Xu<sup>1</sup>, Tamer Kahveci<sup>1</sup>

<sup>1</sup>*Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611;*

**Background:** Computing the long term behavior of regulatory and signaling networks is critical in understanding how biological functions take place in organisms. Steady states of these networks determine the activity levels of individual entities in the long run. Identifying all the steady states of these networks is difficult due to the state space explosion problem.

**Methodology:** In this paper, we propose a method for identifying all the steady states of Boolean regulatory and signaling networks accurately and efficiently. We build a mathematical model that allows pruning a large portion of the state space quickly without causing any false dismissals. For the remaining state space, which is typically very small compared to the whole state space, we develop a randomized traversal method that extracts the steady states. We estimate the number of steady states, and the expected behavior of individual genes and gene pairs in steady states in an online fashion. Also, we formulate a stopping criterion that terminates the traversal as soon as user supplied percentage of the results are returned with high confidence.

**Conclusions:** This method identifies the observed steady states of Boolean biological networks computationally. Our algorithm successfully reported the G1 phases of both budding and fission yeast cell cycles. Besides, the experiments suggest that this method is useful in identifying co-expressed genes as well. By analyzing the steady state profile of Hedgehog network, we were able to find the highly co-expressed gene pair **GL1-SMO** together with other such pairs.

**Availability:** Source code of this work is available at <http://bioinformatics.cise.ufl.edu/palSteady.html>

Full Length Paper

## How to turn a genetic circuit into a synthetic tunable oscillator, or a bistable switch.

Lucia Marucci<sup>1,2</sup>, David A. W. Barton<sup>3</sup>, Irene Cantone<sup>4</sup>, Maria Aurelia Ricci<sup>1</sup>, Maria Pia Cosma<sup>1</sup>, Stefania Santini<sup>2</sup>, Diego di Bernardo<sup>1,2</sup>, Mario di Bernardo<sup>2,3</sup>

<sup>1</sup>Telethon Institute of Genetics and Medicine (TIGEM), Naples, 80131, Italy; <sup>2</sup>Department of Computer and Systems Engineering, Federico II University, Naples, 80125, Italy; <sup>3</sup>Bristol Centre for Applied Nonlinear Mathematics, University of Bristol, Bristol, BS8 1TR, U.K.; <sup>4</sup>MRC Clinical Sciences Centre Faculty of Medicine, Imperial College London Hammersmith Hospital Campus Du Cane Road, London, W12 0NN, U. K.

Systems and Synthetic Biology use computational models of biological pathways in order to study *in silico* the behaviour of biological pathways. Mathematical models allow to verify biological hypotheses and to predict new possible dynamical behaviours. Here we use the tools of non-linear analysis to understand how to change the dynamics of the genes composing a novel synthetic network recently constructed in the yeast *Saccharomyces Cerevisiae* for In-vivo Reverse-engineering and Modelling Assessment (IRMA). Guided by previous theoretical results that link the dynamics of a biological network to its topological properties, through the use of simulation and continuation techniques, we found that the network can be easily turned into a robust and tunable synthetic oscillator, or a bistable switch. Our results provide guidelines to properly re-engineering *in vivo* natural and synthetic networks in order to tune their dynamics.

# An Ensemble Model of Competitive Multi-factor Binding of the Genome

Todd Wasson<sup>1</sup>, Alexander J. Hartemink<sup>1,2</sup>

<sup>1</sup>*Computational Biology and Bioinformatics*; <sup>2</sup>*Department of Computer Science, Duke University, Durham, NC 27708*

Hundreds of different factors adorn the eukaryotic genome, binding to it in large number. These DNA binding factors (DBFs) include nucleosomes, transcription factors (TFs), and other proteins and protein complexes, such as the origin recognition complex (ORC). DBFs compete with one another for binding along the genome, yet many current models of genome binding do not consider different types of DBFs together simultaneously. Additionally, binding is a stochastic process that results in a continuum of binding probabilities at any position along the genome, but many current models tend to consider positions as being either binding sites or not.

We present a model that allows a multitude of DBFs, each at different concentrations, to compete with one another for binding sites along the genome. The result is an ‘occupancy profile’, a probabilistic description of the DNA occupancy of each factor at each position. We implement our model efficiently as the software package COMPETE. We demonstrate genome-wide and at specific loci how modeling nucleosome binding alters TF binding, and vice versa, and illustrate how factor concentration influences binding occupancy. Binding cooperativity between nearby TFs arises implicitly via mutual competition with nucleosomes. Our method applies not only to TFs, but also recapitulates known occupancy profiles of a well-studied replication origin with and without ORC binding. Importantly, the sequence preferences our model takes as input are derived from in vitro experiments. This ensures that the calculated occupancy profiles are the result of the forces of competition represented explicitly in our model and the inherent sequence affinities of the constituent DBFs.

## An efficient and exhaustive approach for modular decomposition of global genetic interaction networks

Jeremy Bellay<sup>1</sup>, Gowtham Atluri<sup>1</sup>, Gaurav Pandey<sup>1</sup>, Michael Costanzo<sup>2</sup>, Anastasia Baryshnikova<sup>2</sup>, Brenda Andrews<sup>2</sup>, Vipin Kumar<sup>1</sup>, Charlie Boone<sup>2</sup>, Chad L. Myers<sup>1</sup>

<sup>1</sup>*Department of Computer Science, University of Minnesota;* <sup>2</sup>*Department of Molecular Genetics, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto.*

Genetic interactions (GI) provide a powerful perspective for analysis of the structure and organization of biological networks. Using Synthetic Genetic Array (SGA) technology, we have constructed the largest map of genetic interactions to date, consisting of quantitative interactions from ~5.4 million double mutants and covering ~30% of all possible double mutant combinations in yeast. This global view of genetic interactions in yeast provides unique insights into the inter-relations of various components and functions of the yeast cell, and can serve as a basis for the systematic characterization of the modular network structure.

We have developed a novel approach for the efficient and exhaustive discovery of GI modules from this quantitative interaction data. Our approach is based on classical algorithms from association analysis in data mining, but extends these approaches to leverage the quantitative information in genetic interaction data and discover coherent blocks (biclusters) of interactions. Due to theoretical guarantees provided by the association analysis framework, we are able to exhaustively find all coherent biclusters of negative or positive genetic interactions within a specified criterion. This guarantee cannot be provided by other current approaches, which employ approximations and/or heuristics to find such biclusters. As a result, this method produces orders of magnitude more modules than existing techniques, and covers several-fold more genes.

An exhaustive catalogue of modular structures from the yeast genetic interaction network has enabled us to explore fundamental organizational principles of the yeast genetic interaction network. Strikingly, we found that our biclusters suggest a surprisingly high degree of multi-functionality among yeast genes: many genes appear in dense clusters of interaction with a large number of highly diverse modules, suggesting these genes support a wide range of functional roles within the cell. This result suggests that the reuse of genes across multiple modules is the norm rather than the exception in biological systems, and we find several genes where the degree of multi-functionality is unexpectedly high (e.g. 10-15 distinct modules regarding mechanistic explanations for genetic interactions. While we do find evidence supporting the current “between-pathway” model for negative interactions and “within-pathway” model for positive genetic interactions, we find that these models fail to explain a large number of observed structures in the network. For example, the vast majority of positive interaction biclusters do not connect genes within pathways or protein complexes as expected, but appear instead to bridge across functionally distal sets of genes. We discuss these and other insights from the exhaustive modular decomposition of the global yeast GI network.)

Our exhaustive catalog of modular structures from the yeast genetic interaction network also revealed limitations in the current dogma.

## Regulatory physics from DNA sequence data

Justin B. Kinney<sup>1-3</sup>, Anand Murugan<sup>3</sup>, Curtis G. Callan, Jr.<sup>3</sup>, Edward C. Cox<sup>4</sup>

<sup>1</sup>Cold Spring Harbor Laboratory; <sup>2</sup>Lewis-Sigler Institute, Princeton University; <sup>3</sup>Department of Physics, Princeton University; <sup>4</sup>Department of Molecular Biology, Princeton University

The last decade has seen great advances in transcriptional genomics, but detailed studies of individual promoters and enhancers still rely on qualitative, low-resolution techniques. Here we introduce a simple mutational assay that uses ultra-high-throughput DNA sequencing to deeply probe the inner workings of a specific transcriptional regulatory sequence. A new analysis method, based on mutual information maximization, allows one to fit quantitative models of regulatory thermodynamics to these sequence data while making minimal assumptions about measurement noise and other experimental details. We applied these techniques to the well-studied *E. coli lac* promoter and demonstrated the resulting ability to (i) identify all regulatory protein binding sites within a targeted sequence, (ii) determine the precise sequence specificities and thus the identities of the proteins that bind these sites, and (iii) measure the quantitative *in vivo* interaction energies between these proteins in their native DNA-bound configuration. This general approach – using massive sequence data sets to precisely characterize specific molecular systems – should be useful for studying other sequence-function relationships in molecular biology.

## Consequences of the fractal globule model for chromatin dynamics in the nucleus

Lieberman-Aiden E.\*, Imakaev M.\*, van Berkum N.L., Lander E.S., Dekker J., and Mirny L.A.

A recent study (Lieberman-Aiden et al., Science, 2009) observed that the structure of the genome, on the scale of a few megabases, is consistent with a fractal globule conformation. The fractal globule is a dense, highly organized, and unknotted conformation of a polymer (e.g. chromatin fiber) into which a polymer can spontaneously fold. First proposed on theoretical grounds in 1988, this structure had never before been seen in experiment or simulations. Here, we discuss physical properties of the fractal globule and demonstrate that it can be an attractive model for DNA organization inside the cell as it provides a rapid access to condensed DNA and organizes the genome into discrete spatial sectors. We study how rapidly the globule folds and unfolds, forms loops, and how easily it can be unravelled in comparison to other condensed polymer states. We illustrate the possibility of local unravelling within a fractal globule and of local condensation, and discuss the consequences for our understanding of gene activation. We compute characteristic times for these processes and relate them to known response times derived from the literature. Biophysical modeling and theoretical analysis of the fractal globule will provide us with new insights into the nature of genomic compaction and gene regulation.

## **Integration and interpretation of massive datasets for diagnosis and treatment of disease**

Kevin White

The deluge of genomic and functional data on disease states and disease susceptibility poses a significant challenge: how do we integrate and interpret massive datasets in ways that inform us about the diagnosis and treatment of disease? Using breast cancer as an example, I will describe our recent efforts to define the transcriptional networks controlled by nuclear receptor proteins in breast cancer cells, to relate these biological network data to patient data, and to present our analyses in the context of recent genome sequencing results that reveal the genetic basis for different subtypes of breast cancer.

Invited Talk

## Global Analysis of Human Protein-DNA Interactions for Annotated and Unconventional DNA-Binding Proteins

Zhi Xie,<sup>2,#</sup> Shaohui Hu,<sup>1,4,#</sup> Akishi Onishi,<sup>3,4</sup> Xueping Yu,<sup>2</sup> Lizhi Jiang,<sup>3,4</sup> Jimmy Lin,<sup>5</sup> Hee-sool Rho,<sup>1,4</sup> Crystal Woodard,<sup>1,4</sup> Hong Wang,<sup>3,4</sup> Jun-Seop Jeong,<sup>1,4</sup> Shunyou Long,<sup>4</sup> Xiaofei He,<sup>1,4</sup> Herschel Wade<sup>6</sup>, Seth Blackshaw,<sup>3,4,\*</sup> Jiang Qian,<sup>2,\*</sup> Heng Zhu<sup>1,4,\*</sup>

<sup>1</sup>*Department of Pharmacology & Molecular Sciences,*

<sup>2</sup>*Department of Ophthalmology*

<sup>3</sup>*Department of Neuroscience,*

<sup>4</sup>*The HiT Center,*

<sup>5</sup>*Department of Cellular and Molecular Medicine,*

<sup>6</sup>*Department of Biophysics and Biophysical Chemistry,*

*Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA*

Protein-DNA interactions (PDIs) mediate a broad range of functions essential for cellular differentiation, function, and survival.

However, it is still a daunting task to comprehensively identify and profile sequence-specific PDIs in complex genomes. Here, we have used a combined bioinformatics and protein microarray-based strategy to systematically characterize the human protein-DNA interactome. We identified 17,718 PDIs between 460 DNA motifs predicted to regulate transcription and 4,191 human proteins of various functional classes. Among them, we recovered many known PDIs for transcription factors (TFs). We identified a large number of unanticipated PDIs for known TFs, as well as for previously uncharacterized TFs. Analysis of PDIs for these TFs revealed a complex landscape of DNA binding specificities in TF families. Surprisingly, we also found that over three hundred unconventional DNA-binding proteins (uDBPs) -- which include RNA binding proteins, mitochondrial proteins, and protein kinases - - showed sequence-specific PDIs. A number of newly identified PDIs have also been confirmed both in vitro and in vivo. Furthermore, one in-depth study of uDBPs, MAPK1, using combined in silico, in vitro and in vivo approaches, has revealed that MAPK1 acts as a transcriptional repressor of interferon-gamma response genes in human cells, suggesting an important biological role for such proteins.

## Simultaneous clustering of multiple gene expression and physical interaction datasets

Manikandan Narayanan<sup>1</sup>, Adrian Vetta<sup>2</sup>, Eric Schadt<sup>1</sup>, Jun Zhu<sup>1</sup>

<sup>1</sup>*Department of Genetics, Rosetta Inpharmatics (Merck), Seattle, WA, USA;* <sup>2</sup>*Department of Mathematics and Statistics, and School of Computer Science, McGill University, Montreal, QC, Canada*

We propose simultaneous clustering of multiple networks as a framework to integrate large-scale datasets on the interactions among and activities of cellular components. Specifically, we develop an algorithm that finds sets of genes that cluster well in multiple networks of interest, such as coexpression networks summarizing correlations among the expression profiles of genes and physical networks describing protein-protein and protein-DNA interactions among genes or gene-products. Our algorithm solves a well-defined problem of jointly clustering networks using techniques that permit certain theoretical guarantees on the quality of the detected clustering. These guarantees coupled with an effective scaling heuristic makes our method JointCluster an advance over earlier approaches. In systematic evaluation of JointCluster and some earlier approaches for combined analysis of the yeast physical network and two gene expression datasets under glucose and ethanol growth conditions, JointCluster discovers clusters that are more consistently enriched for various reference classes capturing different aspects of yeast biology. These robust clusters, which are supported across multiple genomic datasets and diverse reference classes, agree with known biology of yeast under these growth conditions, elucidate the genetic control of coordinated transcription, and enable functional predictions for a number of uncharacterized genes. Whereas many published studies have modeled the protein-protein and protein-DNA interactions within a single network for tractability, the flexible framework of JointCluster also enabled us to exploit the distinction between these two interaction types to greatly improve the biological enrichment of detected clusters.

Full Length Paper

## Literature Curation of Protein Interactions: Discrepancies Across Major Public Databases

Andrei L. Turinsky<sup>1</sup>, Brian Turner<sup>1</sup>, Emerson Cho<sup>1</sup>, Kyle Morrison<sup>1</sup>, Sabry Razick<sup>2</sup>, Ian Donaldson<sup>2</sup>, Shoshana J. Wodak<sup>1</sup>

<sup>1</sup>*Molecular Structure and Function Program, Hospital for Sick Children, Toronto, Canada;*

<sup>2</sup>*The Biotechnology Centre of Oslo, University of Oslo, Norway;* <sup>3</sup>*Department of Biochemistry and Department of Molecular Genetics, University of Toronto, Canada.*

Protein-protein interaction networks have become an important tool in biomedical research in recent years. Several resources around the world are devoted to the annotation of protein interactions from literature, thereby providing a valuable source of interactome data to the research community. However, the interpretation of the original publications is complicated by a number of challenges, which may result in discrepancies among annotations of the same publication.

We systematically investigated the consistency of protein-protein interactions across public resources. Our approach was to focus on PubMed publications that are annotated by two or more protein-interaction databases, and measure the similarity between such co-annotations. To enable this analysis, we consolidated annotations from 9 publicly available databases: BIND, BioGRID, CORUM, DIP, IntAct, HPRD, MINT, MPPI and MPact. After removing redundant records, the combined data represented a total of 272,119 interactions involving 70,474 proteins from 1348 organisms, based on the annotation of 44,159 publications. There were substantial degrees of overlap between the 9 source databases. 15,743 publications were annotated by two or more databases, providing us with the opportunity to assess discrepancies across databases.

Statistical analysis shows that whenever two databases annotate the same publication, their annotations share on average less than half of the interactions. Full agreement – i.e. both databases recording identical sets of protein-protein interactions – occurs in only 24% of the cases. On the other hand, in 41% of the cases, the annotated sets of interactions extracted from the same publication do not overlap at all, indicating that the two databases record all interactions described in the paper differently. The remaining 35% of the cases represent partial overlaps, distributed widely between full agreement and full disagreement.

The discrepancy for the sets of proteins annotated from the same paper is typically less pronounced, with severe disagreement occurring in only 14% of the cases. This indicates that annotators may agree on proteins but still disagree on the interactions they form with each other. Mammalian data stands out as having poor agreement, especially for interactions involving mouse and rat proteins.

We explored several factors that contributed to the discrepancies. One of the major factors is the inconsistent attribution of organisms to the protein-protein interactions described in the literature. We also examined factors such as the annotation of protein complexes, the handling of isoforms, the high- versus low-throughput studies, and the interaction-detection methods.

Our results provide quantitative evidence that alternative interpretations of the literature are common. Given the importance of the protein-protein interaction datasets, our work offers insights into how some of the discrepancies in annotations across public databases may be resolved in the future.

# Discrete logic modeling to link pathway maps and functional analysis of mammalian signal transduction

Julio Saez-Rodriguez<sup>1,2,3\*</sup>, Leonidas Alexopoulos<sup>1,2,3\*,#</sup>, Jonathan Epperlein<sup>1,2</sup>, Regina Samaga<sup>4</sup>, Douglas A. Lauffenburger<sup>1,3</sup>, Steffen Klamt<sup>4</sup> and Peter K. Sorger<sup>1,2,3</sup>

<sup>1</sup> Center for Cell Decision Processes

<sup>2</sup> Harvard Medical School, Department of Systems Biology, Boston, MA 02115, USA

<sup>3</sup> Massachusetts Institute of Technology, Department of Biological Engineering, Cambridge, MA 02139, USA

<sup>4</sup> Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, D-39106, Germany

# Current Address: Dept of Mechanical Engineering, National Technical University of Athens, 15780, Zografou, Greece

\* These two authors contributed equally to this work

Pathway maps inferred from high-throughput data or assembled automatically from the scientific literature are an important means to summarize diverse information on large numbers of proteins. However, as currently formulated, these maps cannot be used to predict the responses of biological pathways to specific stimuli in specific cell types. This is a critical limitation because we must capture the functional distinctions between cell types if we are to understand organogenesis, the pathophysiology of disease or design new drugs. Here we describe the development and testing of an efficient computational approach - implemented in *CellNetOptimizer* (CNO) software - for turning pathway maps into logical models that can be calibrated against experimental data. We use CNO to analyze a map comprising 85 proteins covering the immediate-early response of human cells to seven growth factors and inflammatory cytokines. A Boolean logic model derived from the map was trained against a set of 1150 protein state measurements obtained from transformed human hepatocytes (HepG2 cells). Remarkably, the data-trained model had significantly fewer interactions than the literature-derived map on which it was based, while predicting data absent from the training set with much higher accuracy. The reduction in complexity arose because the literature-based map made many false-positive predictions with respect to signaling in HepG2 cells. Our approach to logical modeling therefore links complex protein maps to input-output data on cells and represents a new means to assemble functional, cell-type specific models of mammalian signal transduction.

DREAM

# Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data

Alexander Mitsos<sup>1</sup>, Ioannis N. Melas<sup>4</sup>, Paraskeuas Siminelakis<sup>4</sup>, Aikaterini D. Chairakaki<sup>4</sup>, Julio Saez-Rodriguez<sup>2,3</sup>, Leonidas G. Alexopoulos<sup>1</sup>

<sup>1</sup>*Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA;* <sup>2</sup>*Department of Systems Biology, Harvard Medical School, Boston, MA;* <sup>3</sup>*Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA;* <sup>4</sup>*Department of Mechanical Engineering, National Technical University of Athens, Athens, Greece*

DREAM

Understanding the mechanisms of cell function and drug action is a major endeavor in the pharmaceutical industry. Drug effects are governed by the intrinsic properties of the drug (i.e., selectivity and potency) and the specific signaling transduction network of the host (i.e., normal vs. diseased cells). Here, we describe an unbiased, phosphoproteomic-based approach to identify drug effects by monitoring drug-induced topology alterations.

With the proposed method, drug effects are investigated under several conditions on a cell-type specific signaling network. First, starting with a generic pathway made of logical gates, we build a cell-type specific map by constraining it to fit 13 key phosphoprotein signals under 55 experimental cases. Fitting is performed via a formulation as an Integer Linear Program (ILP) and solution by standard ILP solvers; a procedure that drastically outperforms previous fitting schemes. Then, knowing the cell topology, we monitor the same key phosphoprotein signals under the presence of drug and cytokines and we re-optimize the specific map to reveal the drug-induced topology alterations.

To prove our case, we make a pathway map for the hepatocytic cell line HepG2 and we evaluate the effects of 4 drugs: 3 selective inhibitors for the Epidermal Growth Factor Receptor (EGFR) and a non selective drug. We confirm effects easily predictable from the drugs' main target (i.e. EGFR inhibitors blocks the EGFR pathway) but we also uncover unanticipated effects due to either drug promiscuity or the cell's specific topology. An interesting finding is that the selective EGFR inhibitor Gefitinib is able to inhibit signaling downstream the Interleukin-1alpha (IL-1a) pathway; an effect that cannot be extracted from binding affinity based approaches.

Our method represents an unbiased approach to identify drug effects on a small to medium size pathways and is scalable to larger topologies with any type of signaling perturbations (small molecules, RNAi etc). The method is a step towards a better picture of drug effects in pathways, the cornerstone in identifying the mechanisms of drug efficacy and toxicity.

# Crosstalk among receptor tyrosine kinases inferred from micro-Western phosphoproteomic arrays using Bayesian networks, ARACNe, and CLR

Joel P. Wagner<sup>1,2</sup>, Mark F. Ciaccio<sup>3,4</sup>, Chih-Pin Chuu<sup>3</sup>, Richard B. Jones<sup>3,4</sup>, Douglas A. Lauffenburger<sup>1,2</sup>

<sup>1</sup>Center for Cell Decision Processes, MIT; <sup>2</sup>Department of Biological Engineering, MIT; <sup>3</sup>Ben May Department for Cancer Research and the Institute for Genomics and Systems Biology, University of Chicago; <sup>4</sup>Committee on Cellular and Molecular Physiology, University of Chicago

Receptor tyrosine kinases (RTKs) are a subclass of cell surface receptors that possess an intrinsic kinase activity. Coactivation of multiple RTKs has been observed in tumor cells and been proposed as a reason for the limited success of clinical therapies directed at single RTKs.

In this work, 91 phosphorylation sites were measured at six time points (0, 1, 5, 15, 30, and 60 min.) following stimulation with five epidermal growth factor (EGF) concentrations (0, 2, 50, 100, and 200 ng/mL) in A431 epithelial cells using newly developed micro-Western arrays. Of the 91 phosphorylation sites, fifteen sites across ten RTKs and two from Src kinase were chosen for network modeling. Connectivities among the seventeen phosphorylation sites were inferred using a dynamic programming approach for exact Bayesian network inference, which proposed directed interactions, and two mutual information-based methods, ARACNe and CLR, which propose undirected interactions. The directionality of edges in the Bayesian network was constrained using equivalence class analysis, and the sign (positive vs. negative) of interactions was also estimated.

The three inference methods give significant agreement on the network topology. The ARACNe network represented a subnetwork of the Bayesian network, which represented a subnetwork of the CLR network. Edge weight thresholds for all three methods were established using data permutation studies.

Computationally, the results: [1] support previous reports suggesting that small amounts of experimental data collected from phenotypically diverse network states may enable feasible network inference; and [2] suggest that, using this data set, ARACNe may remove too many edges using the Data Processing Inequality, and CLR may include too many edges because of its mutual information normalization procedure.

Biologically, the results support: [1] coactivation of multiple RTKs following EGF stimulation; [2] experimentally observed heterodimerization between EGF receptor (EGFR) and platelet-derived growth factor receptor beta (PDGFRB), suggesting the two activate their substrates with 'AND gate' behavior; [3] a known role for Src in the activation of EGFR(Y845); [4] increased activity of hypo-glycosylated fibroblast growth factor receptor 1 (FGFR1) compared to the hyper-glycosylated form, consistent with experimental findings; and [5] decreased phosphorylation of PDGFRA(Y754), the only negatively regulated site in the resultant model, consistent with previous literature reports as a mechanism for turning off the MAPK cascade.

# Single Cell Signaling & Pathology in Autoimmunity & Cancer

Garry P. Nolan<sup>1</sup>,

<sup>1</sup> *Department of Microbiology & Immunology, Baxter Laboratory of Stem Cell Biology, Stanford University School of Medicine, Stanford, CA*

Intracellular assays of signaling systems has been limited by an inability to correlate functional subsets of cells in complex populations based on active kinase states or other nodal signaling junctions. Such correlations could be important to distinguish changes in signaling status that arise in rare cell subsets during functional activation or in disease manifestation. We have demonstrated the ability to simultaneously detect activated kinases and phosphoproteins in simultaneous pathways in subpopulations of complex cell populations by multi-parameter flow cytometric analysis, and now via high throughput mass spectrometry at the single cell level.

The focus of the presentation will be our mechanistic studies of Lupus and Rheumatoid Arthritis, as well as signaling of stem cells in Acute Myelogenous Leukemia stem. The detailed, correlated datasets generated via single cell phospho-flow allows for ready representation via automated signaling network determination using Bayesian analysis. Our pursuit of deep analysis of these datasets, and the limitations of cloud computing or standard multi-CPU systems, has stimulated our development of unique computational approach using field programmable gate arrays and GPU multiprocessor architectures in the development of a 'bioinformatics supercomputer' and associated algorithms.

By analyzing the immune system, or cancer, as individual cells we now observe structured network interactions within these tissues at a new level of clarity. I will present our initial generations of comprehensive network topology maps of signaling within, and between, primary immune subsets in normal and pathologic disease tissues in RA/Lupus and cancer. My emphasis will be on the application of these approaches directly to human samples in near-patient settings for the development of point-of-care mechanistic referencing of disease and drug action.

DREAM

# Peptide Recognition Domain (PRD) Specificity Prediction

Gary D. Bader<sup>1,2,3,4,5</sup> and Philip M. Kim<sup>1,2,3,4</sup>

<sup>1</sup> *Terrence Donnelly Centre for Biomolecular and Cellular Research*

<sup>2</sup> *Banting and Best Department of Biomedical Research*

<sup>3</sup> *Department of Computer Science*

<sup>4</sup> *Department of Molecular Genetics*

<sup>5</sup> *Samuel Lunenfeld Research Institute*

*University of Toronto, Toronto, Ontario M5S 3E1 Canada*

DREAM

Peptide recognition domains (PRD) mediate many important protein-protein interactions. They make up prominent modular protein domain families, and have expanded in recent evolutionary history, underlining their importance. PRDs bind short linear sequence motifs in other proteins. For example, SH3 domains typically recognize proline-rich motifs, PDZ domains recognize hydrophobic C-terminal tails, and protein kinases recognize short sequence regions around a phosphorylatable residue. Their binding specificity, in combination with contextually conferred specificity, determines the interaction partners for each PRD. This fact, together with the biological importance of their interaction makes specificity prediction an attractive problem. Because of its relatively simple nature, it occupies an area which is accessible to both biophysics/structure based approaches as well as machine learning methodologies. Part of the motivation of this challenge is an unbiased assessment of the advantages of each method and to stimulate the development of next-generation approaches that combine different elements of either field. The challenge as part of DREAM4 consists of predicting the binding specificity of given PRDs based on its sequence. The binding specificity is encoded in a position weight matrix (PWM) that describes the sequence space covered by the domains target peptides. While the PWM data model has well-known limitations, we chose it here as the most accepted and intuitive model. Moreover, it is the most natural representation of the results of the combinatorial peptide library experimental method. We specifically chose a mix of domain families (SH3, PDZ and protein kinases) with very different biochemical and biological properties. Moreover, the challenge includes naturally occurring and synthetically modified domains to detect the ability of different algorithms to assess the effects of evolutionary forces as well as biochemical realities. The performance of each method is tested by the similarity of the predicted PWM to the ones measured by phage display and combinatorial peptide library experiments.

## Challenge 3: Predictive Signaling Network Modeling

Robert Prill<sup>1</sup>, Leonidas Alexopoulos<sup>2,3,4#</sup>, Julio Saez-Rodriguez<sup>2,3</sup>, Douglas A. Lauffenburger<sup>2,4</sup>, Peter K. Sorger<sup>2,3,4</sup> and Gustavo Stolovitzky<sup>1</sup>

<sup>1</sup> IBM Research

<sup>2</sup> Center for Cell Decision Processes

<sup>3</sup> Harvard Medical School, Department of Systems Biology, Boston, MA 02115, USA

<sup>4</sup> Massachusetts Institute of Technology, Department of Biological Engineering, Cambridge, MA 02139, USA

# Current Address: Dept of Mechanical Engineering, National Technical University of Athens, 15780, Zografou, Greece

DREAM

Pathway maps summarizing literature knowledge are widespread and useful, but they do not allow prediction of pathway operation and do not usually include cell-type specific information. This is an important limitation because it is these differences between normal and diseased cells that are targeted for therapeutic intervention.

This challenge explores the extent to which our current knowledge of signaling pathways, collected from a variety of cell types, agrees with cell-type specific high-throughput experimental data. Specifically, we ask the challenge participants to create a cell-type specific model of signal transduction using the measured activity levels of signaling proteins in HepG2 cell lines upon perturbation with different combinations of extracellular ligands and small-molecule inhibitors [1]. The model, which can leverage prior information encoded in a generic signaling pathway provided in the challenge [2], should be biologically interpretable as a network, and capable of predicting the outcome of new experiments. The submissions will be scored by the prediction error in the test set and the parsimony of the submitted network.

### References:

[1] Alexopoulos L, Saez-Rodriguez J, Cosgrove B, Lauffenburger DA, Sorger PK. Networks reconstructed from cell response data reveal profound differences in signaling by Toll-like receptors and NF- $\kappa$ B in normal and transformed human hepatocytes. Submitted.

[2] Saez-Rodriguez J, Alexopoulos L, Epperlein J, Samaga R, Lauffenburger DA, Klamt S, Sorger PK. Discrete logic modeling to link protein signaling networks and functional analysis of mammalian signal transduction. Submitted.

# Generating realistic benchmarks for gene network inference: the DREAM4 *in silico* network challenge

Daniel Marbach<sup>1</sup>, Thomas Schaffter<sup>1</sup>, Dario Floreano<sup>1</sup>, Gustavo Stolovitzky<sup>2,3</sup>

<sup>1</sup>Laboratory of Intelligent Systems, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland; <sup>2</sup>IBM Computational Biology Center, Yorktown Heights, New York, USA; <sup>3</sup>Department of Biomedical Informatics, Columbia University, New York, USA.

Evaluating the performance of methods for inference of gene regulatory networks is difficult, because network predictions can in general not be systematically tested *in vivo* with the current technology. Consequently, *in silico* benchmarks based on simulated networks and expression data (see figure) are essential to assess the performance of network-inference methods. We have developed tools for the generation of biologically plausible benchmarks, which enable realistic *in silico* performance assessment. We have previously described a method for generating realistic network structures for the benchmarks by extracting modules from known gene networks of model organisms. Here, we introduce in addition a kinetic model that is more detailed and accurate than those used in other benchmarks proposed in the literature. It is based on a mechanistic, thermodynamical model of transcriptional regulation, and it includes both mRNA and protein dynamics. Furthermore, in addition to measurement noise, we also model internal noise in the dynamics of the gene networks. This framework allows inference methods to be tested *in silico* on networks with similar types of structural properties, regulatory dynamics, and noise as occur in biological gene networks.

Using this framework, we have generated benchmarks for a community-wide experiment for the fourth DREAM conference, the so-called “DREAM4 *in silico* network challenge”. In this presentation, I will introduce this challenge. I will describe the above-mentioned framework, discuss the different types of data that we provided for the participants, and summarize lessons from the previous edition of the challenge that we addressed in the design of this year’s benchmarks.

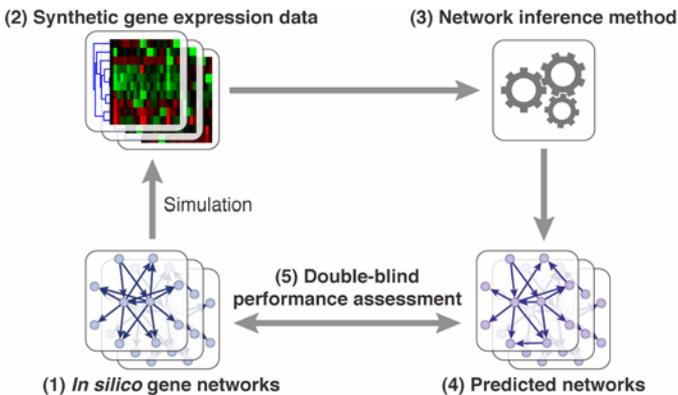


Figure: Design of the *in silico* network challenge. From a set of *in silico* networks, gene expression data is simulated. Participating teams are asked to predict the structure of the networks from this data. They are blind to the true structure of the *in silico* networks. We assess the submitted predictions blind to the inference methods that produced them, allowing for a double-blind evaluation.

DREAM

# Systems Biology of DNA Damage and Repair

Michael B. Yaffe<sup>1</sup>,

<sup>1</sup>*Depts of Biology and Biological Engineering, Koch Institute for Integrative Cancer Biology, Massachusetts Institute of Technology and Broad Institute, Cambridge, MA USA*

Post-translational modifications such as protein phosphorylation and acetylation, together with binding of the modified proteins to modular signaling domains (i.e. 14-3-3 proteins, WW domains, FHA domains, Polo-box domains, and BRCT domains) function together within signaling networks to control the cellular response to DNA damage. How signals emerging from these pathways are integrated and processed as a network is unclear. To address this, we have been developing systems models of signaling where kinase activities, protein phosphorylation, binding of substrates to phosphoserine/threonine binding domains, and cellular responses including cell cycle arrest and apoptosis are quantitatively measured at densely sampled points in time, and correlated mathematically using partial least squares regression/principal components analysis, or step-wise regression methods. We have used this to approach to map context-dependent signaling events that control the fate of individual cells after genotoxic stress.

In addition to the ATM/Chk2 and ATR/Chk1 pathways that contribute to these responses, we have identified a third DNA damage signaling pathway mediated by p38 MAPK activation of MAPKAP Kinase-2 that is required for p53-deficient, but not for p53-proficient, tumor cell survival after DNA damage. In response to cisplatin exposure, MAPKAP Kinase-2 is required for Cdc25A destabilization and activation of the G1 and intra-S checkpoints, while in response to doxorubicin or camptothecin treatment, MAPKAP Kinase-2 is required for targeting of Cdc25B to 14-3-3, and activation of the G2/M checkpoint. MAPKAP Kinase-2 depletion abolishes these cell cycle checkpoints in p53-defective tumor cells, sensitizes them to chemotherapy in culture, and induces dramatic tumor regression after low-dose chemotherapy in a murine xenograft tumor model. No such MAPKAP Kinase-2 dependence is seen in p53-proficient tumors. Chk1 responds to the same genotoxic stresses that activate MAPKAP Kinase-2, and Chk1 is activated normally in the MAPKAP Kinase-2-depleted cells that display aberrant checkpoint function. Thus, MAPKAP Kinase-2 and Chk1 function together as a molecular 'AND' gate that is required to integrate DNA damage signals to control cell cycle progression and prevent mitotic catastrophe within tumors. The biological basis for the 'AND' gate function of Chk1 and MAPKAP Kinase-2 is directly related to the subcellular context in which each kinase functions. ATM is second molecular target whose inhibition has been postulated to sensitize tumors to chemotherapy. Using in vitro cell culture models together with murine xenografts and E $\mu$ -myc driven lymphomas, we show that ATM inhibition can result in either pronounced tumor resistance or sensitization to DNA damaging chemotherapy, determined solely by the underlying state of the p53 pathway. These findings for MAPKAP Kinase-2 and ATM demonstrate how signals from the DNA damage network are re-routed in p53-defective tumor cells, and show how pathway- and network-focused diagnostics can be used to successfully predict therapeutic outcome in human cancer treatment.

DREAM

Invited Talk

# Effective identification of conserved pathways in biological networks using hidden Markov models

Xiaoning Qian<sup>1</sup> and Byung-Jun Yoon<sup>2</sup>

<sup>1</sup>*Department of Computer Science & Engineering, University of South Florida;*

<sup>2</sup>*Department of Electrical & Computer Engineering, Texas A&M University.*

The advent of various high-throughput experimental techniques for measuring molecular interactions has enabled the systematic study of biological interactions on a global scale. Since biological processes are carried out by elaborate collaborations of numerous molecules that give rise to a complex network of molecular interactions, comparative analysis of these biological networks can bring important insights into the functional organization and regulatory mechanisms of biological systems.

In this paper, we present an effective framework for identifying common interaction patterns in the biological networks of different organisms based on hidden Markov models (HMMs). Given two or more networks, our method efficiently finds the top  $k$  matching paths in the respective networks, where the matching paths may contain a flexible number of consecutive insertions and deletions. Based on several protein-protein interaction (PPI) networks obtained from the Database of Interacting Proteins (DIP) and other public databases, we demonstrate that our method is able to detect biologically significant pathways that are conserved across different organisms. Our algorithm has a polynomial complexity that grows linearly with the size of the aligned paths. This enables the search for very long paths with more than 10 nodes within a few minutes on a desktop computer.

DREAM

Full Length Paper

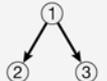
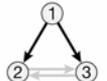
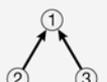
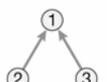
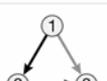
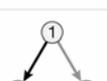
# Revealing strengths and weaknesses of methods for gene network inference

Daniel Marbach<sup>1</sup>, Robert J. Prill<sup>2</sup>, Thomas Schaffter<sup>1</sup>, Dario Floreano<sup>1</sup>, Gustavo Stolovitzky<sup>2,3</sup>

<sup>1</sup>Laboratory of Intelligent Systems, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland; <sup>2</sup>IBM Computational Biology Center, Yorktown Heights, New York, USA <sup>3</sup>Department of Biomedical Informatics, Columbia University, New York, USA.

DREAM

Spurred by advances in experimental technology, a plethora of methods for inference of gene regulatory networks has been developed. However, the strengths and weaknesses of these methods remain poorly understood. On the one hand, this can be explained by the difficulty of constructing adequate benchmarks, and on the other hand, by the lack of tools for a differentiated analysis of network predictions on such benchmarks. Here, we introduce a novel approach to analyze network predictions. In addition to assessing the overall accuracy, we evaluate the performance on different types of local connectivity patterns (motifs) of the networks. We have used this approach to assess the performance of 29 network-inference methods that have been applied independently by participating teams in a community-wide reverse engineering challenge (the DREAM3 *in silico* network challenge). Our results show that current inference methods are affected, to various degrees, by three types of systematic prediction errors: the *fan-out error*, the *fan-in error*, and the *cascade error* (see figure). In particular, all but the best-performing method failed to accurately infer multiple regulatory inputs (combinatorial regulation) of genes. In contrast to evaluation of overall accuracy, the network motif analysis reveals the strengths and weaknesses of inference methods, thereby suggesting possible directions for their improvement.

	(A) True structure	(B) Prediction confidence	(C) Systematic prediction error
Fan-out			<b>Fan-out error</b> Incorrect prediction of links between co-regulated nodes (co-regulation misinterpreted as interaction)
Fan-in			<b>Fan-in error</b> Reduced prediction confidence for multiple inputs (difficulties in predicting combinatorial regulation)
Cascade			<b>Cascade error</b> Incorrect prediction of "shortcuts" (indirect interaction misinterpreted as direct interaction)
FFL			Feed-forward loop (FFL) Same type of error as on fan-ins

**Median prediction confidence**

- High
- Medium
- Low
- No arrow: zero

Figure: (A) The true connectivity of the motifs. (B) As an example, we show how the motifs were predicted on average by one of the top-three methods of the

DREAM3 *in silico* challenge. This reveals three types of systematic prediction errors (C). The darkness of the links indicates their median prediction confidence.

## **A Human B Cell Interactome Identifies MYB and FOXM1 as Master Regulators of Proliferation in Germinal Centers**

Celine Lefebvre<sup>1</sup>, Presha Rajbhandari<sup>1</sup>, Mariano J. Alvarez<sup>1</sup>, Wei Keat Lim<sup>1</sup>, Mai Sato<sup>2</sup>, Kai Wang<sup>1</sup>, Pavel Sumazin<sup>1</sup>, Katia Basso<sup>2</sup>, Jean Gautier<sup>2</sup>, Riccardo Dalla-Favera<sup>2</sup>, and Andrea Califano<sup>1,2</sup>

<sup>1</sup>Center for Computational Biology and Bioinformatics, <sup>2</sup>Institute for Cancer Genetics, Columbia University, 1130 St Nicholas Avenue, New York, NY 10032, USA

By integrating data from high-throughput experimental assays and reverse-engineering algorithms, using a Bayesian framework, we have assembled and biochemically/functionally validated a cell-context-specific Human B Cell Interactome (HBCI), representing both transcriptional and post-translational interaction layers. We show that computational interrogation of the HBCI, using a novel Master Regulator Inference Algorithm (MaRInA), yields a complete repertoire of genes that control Germinal Center (GC) formation and maintenance, i.e. genes that individually or synergistically control GC-specific genetic programs. GCs, which represent the hallmark of antigen-mediated immune response, are structures where antigen-stimulated B cells proliferate, undergo somatic hypermutation of immunoglobulin genes, and are selected based on the production of high-affinity antibodies. GC B cells (centroblasts) derive from naïve B cells, from which they differ for the activation of genetic programs controlling cell proliferation, DNA metabolism and pro-apoptotic programs and for the repression of anti-apoptotic, cell cycle arrest, DNA repair, and signal transduction programs from cytokines and chemokines.

The Master Regulators inferred by our method recapitulate genes that were previously identified by targeted genetic and biochemical analyses. In addition they identify several new master regulators. In particular, the analysis revealed a hierarchical, transcriptional control module, which was extensively validated, both biochemically and functionally, where MYB and FOXM1 act as synergistic master regulators of GC-specific proliferative programs. Eighty percent of genes jointly regulated by these transcription factors are activated in the GC. These include genes encoding proteins in a previously uncharacterized complex that includes DNA pre-replication, replication, and mitosis control proteins, which we have experimentally validated by co-immunoprecipitation assays and confocal imaging. These results indicate that the HBCI analysis can be used for the identification of determinants of major human cell phenotypes and provides a paradigm of general applicability to normal and pathologic tissues.



## DRUG NETWORK (DRUNET): a new and powerful approach to identify drug mode of action from gene expression profiles

Francesco Iorio<sup>1,2</sup>, Roberta Bosotti<sup>3</sup>, Vincenzo Belcastro<sup>1</sup>, Nicola Brunetti<sup>1</sup>, Antonella Isacchi<sup>3</sup>, Diego di Bernardo<sup>1,4</sup>

<sup>1</sup>Telethon Institute of Genetics and Medicine (TIGEM); <sup>2</sup>University of Salerno; <sup>3</sup>Nerviano Medical Sciences; <sup>4</sup>University "Federico II" of Naples. - ITALY

DREAM

We exploited the notion of a network to design a novel tool able to predict the mode of action (MOA) of a drug starting from gene expression profiles, without any additional prior knowledge. Specifically, we constructed a "drug network" (DN) from gene expression profiles following drug treatments in human cell lines (The Connectivity Map database). In this DN, drugs are connected to each other into "communities" and "rich-clubs" able to capture similarities and differences in drug mode of action, and to predict mode of action of novel compounds.

At the heart of our approach is a novel definition of "distance" between two drugs. This is computed by combining gene expression profiles obtained with the same compound, but in different experimental settings, via an original rank-aggregation method, followed by a gene set enrichment analysis. The DN is then generated by considering each compound as a node, and adding a weighted edge between two compounds if their similarity distance is below a given significance threshold. By using a novel clustering based procedure, we identified 106 different communities in our DN, and we organized them into "rich-clubs", consisting of "community of communities". The whole topology of the resulting DN reflects hierarchy of similarities among the composing drugs, with drugs in the same community having a similar mode of action. We validated each community for which the MOA information on the drugs was available, and confirmed that for 60% of the communities; we correctly grouped together drugs that have similar MOA.

We then used the inferred DN to classify drugs not present in the original Connectivity Map. We generated, in house, gene expression data for 9 compounds, including well-known and novel *HSP90* inhibitors, *CDK* inhibitors and Topoisomerase inhibitors, for a total amount of 43 microarray hybridizations, and integrated them in the DN. These compounds were all correctly classified, since they were correctly placed in the right communities, and their closest neighbors had the correct MOA. Moreover, we identified and experimentally validated previously unrecognized similarities between *CDK* inhibitors and Topoisomerase inhibitors, by showing that the latter indirectly inhibit *CDKs* via p21.

Finally, we predicted and experimentally validated a novel MOA as an autophagia inducer of a well-known drug. Due to this new discovery, this drug might be tested for the treatment of a wide range of neurodegenerative disorders, thanks to its excellent safety profile.

In addition, we also developed a web-based interface to explore the DN and query it for classification of novel compounds.

Our approach represents a unique and robust method able to discover the MOA of novel drugs and to identify new effects of well-known drugs by using only gene expression data.

# Functional Insights from Protein-Protein and Genetic Interaction Maps

Nevan Krogan<sup>1,2</sup>

<sup>1</sup>*Cellular and Molecular Pharmacology, University of California, San Francisco;* <sup>2</sup>*California Institute for Quantitative Biomedical Sciences, University of California, San Francisco*

Pathways and complexes can be considered fundamental units of cell biology, but their relationship to each other is difficult to define. Comprehensive tagging and purification experiments have generated networks of interactions that represent most stable protein complexes. We describe this work in various organisms, including budding yeast and in infectious organisms like HIV and TB, and show how the analysis of pairwise epistatic relationships between genes complements the physical interaction data, and furthermore can be used to classify gene products into parallel and linear pathways.

# Metabolic Flux Balance Analysis with Context-dependant Biomass

Tomer Benyamini<sup>1</sup>, Ori Folger<sup>1</sup>, Eytan Ruppin<sup>1,2</sup>, Tomer Shlomi<sup>3</sup>

<sup>1</sup> *The Blavatnik School of Computer Science – Tel Aviv University;* <sup>2</sup> *The Sackler School of Medicine – Tel Aviv University;* <sup>3</sup> *Computer Science Department - Technion - Israel Institute of Technology.*

In recent years, flux balance analysis (FBA) has been very commonly used for metabolic network analysis, with a growing number of applications, including phenotype prediction, metabolic engineering, and studies of network evolution. Currently more than 30 genome-scale network models amenable for FBA analysis are available for various model organisms, industrially relevant organisms, pathogens and recently to human cellular metabolism, with major efforts ongoing to reconstructing additional models. FBA works by predicting a feasible flux distribution under a quasi steady-state assumption, satisfying stoichiometric mass-balance constraints as well as enzymatic directionality and capacity constraints. To account for cellular growth, FBA incorporates a pseudo growth reaction in the model which utilizes a few dozen metabolites required for biomass formation based on their experimentally measured ratios under some growth condition. However, considering the biomass demand of only a small number of metabolites and ignoring the fact that biomass composition significantly varies across growth conditions, FBA's flux distributions do not fully satisfy mass-balance constraint, which may lead to false predictions. Here, we present a novel method, Context-dependant Biomass FBA (CB-FBA), which addresses this problem by predicting a flux distribution that maximizes the production rate of a core biomass, while accounting for the growth-associated demand for the production of all intermediate metabolites produced in the process. Applying CB-FBA to predict metabolic phenotypes in a genome-scale metabolic network model of *E. coli* showed a significant improvement in gene essentiality and growth rate prediction performance over FBA.

# Dynamic networks from hierarchical Bayesian graph clustering

Yongjin Park<sup>1</sup>, Christopher Moore<sup>2,3</sup>, Joel S. Bader<sup>1,\*</sup>

<sup>1</sup>*Department of Biomedical Engineering and High-Throughput Biology Center, Johns Hopkins University, Baltimore, MD, USA;* <sup>2</sup>*Department of Computer Science and Department of Physics, University of New Mexico, Albuquerque, NM, USA* <sup>3</sup>*Sata Fe Institute, 1399 Hyde Park Road, Santa Fe, NM, USA*

\* E-mail: joel.bader@jhu.edu

Biological networks change dynamically as protein components are synthesized and degraded. Understanding the time-dependence and, in a multicellular organism, tissue-dependence of a network leads to insight beyond a view that collapses time-varying interactions into a single static map. Conventional algorithms are limited to analyzing evolving networks by reducing them to a series of unrelated snapshots. Here we introduce an approach that groups proteins according to shared interaction patterns through a dynamical hierarchical stochastic block model. Protein membership in a block is permitted to evolve as interaction patterns shift over time and space, representing the spatial organization of cell types in a multicellular organism. The spatiotemporal evolution of the protein components are inferred from transcript profiles, using Arabidopsis root development (5 tissues, 3 temporal stages) as an example. The new model requires essentially no parameter tuning, out-performs existing snapshot-based methods, identifies protein modules recruited to specific cell types and developmental stages, and could have broad application to social networks and other similar dynamic systems.

SysBio

Full Length Paper

# A dynamic analysis of IRS-PKR signaling in liver cells: a discrete modeling approach

Ming Wu<sup>1\*</sup>, Xuerui Yang<sup>2\*</sup>, [Christina Chan](#)<sup>123†</sup>

<sup>1</sup>Department of Computer Science and Engineering, <sup>2</sup>Department of Chemical Engineering and Material Science, <sup>3</sup>Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA

\*contributed equally, † corresponding author, email: [krischan@egr.msu.edu](mailto:krischan@egr.msu.edu)

A major challenge in systems biology is to develop a detailed dynamic understanding of the functions and behaviors in a particular cellular system, which depends on the elements and their inter-relationships in a specific network. Computational modeling plays an integral part in the study of network dynamics and uncovering the underlying mechanisms. Here we proposed a systematic approach that incorporates discrete dynamic modeling and experimental data to reconstruct a phenotype-specific network of cell signaling. A dynamic analysis of the insulin signaling system in liver cells provides a proof-of-concept application of the proposed methodology. Our group recently identified that double-stranded RNA-dependent protein kinase (PKR) plays an important role in the insulin signaling network. The dynamic behavior of the insulin signaling network is tuned by a variety of feedback pathways, many of which have the potential to cross-talk with PKR. Given the complexity of insulin signaling, it is inefficient to experimentally test all possible interactions in the network to determine which pathways are functioning in our cell system. Our discrete dynamic model provides an *in silico* model framework that integrates potential interactions and assesses the contributions of the various interactions on the dynamic behavior of the signaling network. Simulations with the model generated testable hypothesis on the response of the network upon perturbation, which were experimentally evaluated to identify the pathways that function in our particular liver cell system. The modeling in combination with the experimental results enhanced our understanding of the insulin signaling dynamics and aided in generating a context-specific signaling network.

# Comprehensive modeling of microRNA targets: predicting functional non-conserved and non-canonical sites

Doron Betel<sup>1</sup>, Anjali Koppal<sup>2</sup>, Phaedra Agius<sup>1</sup>, Chris Sander<sup>1</sup>, Christina Leslie<sup>1</sup>

<sup>1</sup>*Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York;* <sup>2</sup>*Department of Computer Science, Columbia University, New York*

Accurate prediction of microRNA targets is a challenging computational problem, impeded by incomplete biological knowledge and the scarcity of experimentally validated targets. The primary determinant for regulation, near-perfect base pairing in the seed region of the microRNA (positions 2-7), gives poor specificity as a prediction rule. In an effort to reduce false predictions, most computational methods restrict to perfect seed matches that are evolutionary conserved, despite experimental evidence that neither constraint holds in general. Here we present a new algorithm called mirSVR for predicting and ranking the efficiency of microRNA target sites by using supervised learning on mRNA expression changes from microRNA transfection experiments. We use support vector regression (SVR) to train on features of the predicted miRNA:mRNA duplexes as well as contextual features without restricting to perfect seed complementarity or filtering by conservation. In a large-scale evaluation on independent transfection and inhibition experiments, mirSVR significantly outperformed existing target prediction methods for predicting genes that are deregulated at the mRNA or protein levels.

mirSVR effectively broadens target prediction beyond the standard restrictions of perfect seeds and strict conservation without introducing a large number of spurious predictions. In particular, we found that mirSVR correctly identified functional but poorly conserved target sites, and that imposing a conservation filter resulted in a reduced rate of detection of true targets. mirSVR scores are calibrated to correlate linearly with the extent of downregulation and therefore enable accurate scoring of genes with multiple target sites by addition of individual site scores. Furthermore, the scores can be converted to an empirical probability of downregulation, which provides a meaningful guide for selecting a score cutoff. The model successfully predicted genes that are regulated by multiple endogenous microRNAs – rather than transfected microRNAs whose concentrations are above physiological levels – when analyzing targets bound to human Argonaute (AGO) proteins as identified by AGO immunoprecipitation. Finally, we tested the usefulness of including non-canonical sites in the model by evaluating performance on biochemically determined sites from recent PURE-CLIP experiments, 20% of which do not contain any perfect microRNA seed match. We found that mirSVR indeed correctly detected a significant number of these experimentally verified non-canonical sites.

# The cell of origin of human cancers

Franziska Michor<sup>1</sup>

<sup>1</sup>*Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York*

All cancers rely on cells that have properties of long-term self-renewal or “stemness” to maintain and propagate the tumor, but the cell of origin of most cancers is still unknown. Knowledge of the target of transformation is important for an understanding of the natural history of cancers and has therapeutic implications. We design stochastic mathematical models of the differentiation hierarchy of cells in tissues and study the evolutionary dynamics of cancer initiation. We consider different evolutionary pathways leading to cancer-initiating cells, and find mathematical evidence that a progenitor is the most likely cell of origin of hematopoietic and brain tumors. We also discuss experimental evidence supporting our findings. These results also have relevance to other tumor types arising in tissues that are organized as a differentiation hierarchy.

# Building *Saccharomyces cerevisiae* v2.0: The Synthetic Yeast Genome Project

Jessica S. Dymond<sup>1</sup>, Sarah Richardson<sup>1</sup>, Candice Coombes<sup>1</sup>, Lisa Scheifele<sup>1</sup>, Eric Cooper<sup>1</sup>, Joy Wu<sup>2</sup>, Derek Lindstrom<sup>3</sup>, Daniel Gottschling<sup>3</sup>, Srinivasan Chandrasegaran<sup>2</sup>, Joel Bader<sup>1</sup>, Jef D. Boeke<sup>1</sup>

<sup>1</sup> *The High Throughput Biology Center, Johns Hopkins University, School of Medicine, Baltimore, MD 21205;* <sup>2</sup> *Department of Environmental Health Sciences, Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD 21205;* <sup>3</sup> *Fred Hutchinson Cancer Research Center, Seattle, Washington 98109*

The Synthetic Yeast Genome Project (SYGP) is a consortium of researchers working to design, construct, and characterize a synthetic version of the *Saccharomyces cerevisiae* genome. Although other groups currently seek to construct synthetic prokaryotes, a sleek, easily manipulated eukaryotic genome featuring several useful modifications will allow more in-depth analysis of biological processes, as well as provide a chassis upon which novel systems may be developed. Incorporation of a user-controlled inducible evolution system in the synthetic yeast genome will provide a new mechanism through which rules dictating genome content and structure will be defined. Also, we are systematically deleting introns and repeat sequences such as retrotransposons from the genome, and relocating all tRNA genes, which are hotspots for genome instability. Construction of a 12 megabase genome is a formidable task; we have therefore developed an undergraduate synthetic biology course, Build-a-Genome, at Johns Hopkins University [1]. Build-a-Genome students perform gene synthesis in the context of the SYGP, constructing fragments of the designer synthetic yeast genome from oligos. Over four semesters, students have synthesized over 90% of chromosome III and current efforts are focusing on construction of these fragments into larger assemblies for eventual integrative replacement of the native yeast genome. Additionally, the first arm of a synthetic chromosome (IX) has been synthesized. Introduction of the synthetic chromosome arm, comprising approximately 1% of the genome and containing all designed changes, is phenotypically indistinguishable from the native genome and stably maintained over time. RNA profiles are normal in spite of loxP sites engineered into the 3' UTR of every nonessential gene. Further, the user-controlled evolution system permits large rearrangements and deletions, producing novel genotypes. Future work will focus on characterization of the inducible evolution system, as well as continued fabrication / incorporation of additional chromosomes.

[1] Dymond, J.S., Scheifele, L.Z., Richardson, S., Lee, P., Chandrasegaran, S., Bader, J.S., Boeke, J.D. 2009. Teaching synthetic biology, bioinformatics, and engineering to undergraduates: the interdisciplinary Build-a-Genome course. *Genetics*, 181: 13-21.

## Decoding small RNA networks in bacteria

Diogo M. Camacho<sup>1</sup>, Sheetal R. Modi<sup>1</sup>, Michael A. Kohanski<sup>1,2</sup> and James J. Collins<sup>1,2</sup>

<sup>1</sup>Howard Hughes Medical Institute and Department of Biomedical Engineering, Boston University; <sup>2</sup>Boston University School of Medicine, Boston University

Small, non-coding RNA molecules regulate a vast array of processes, from transcription to translation, in all kingdoms of life. In the bacterium *Escherichia coli*, small RNAs have been extensively studied and several approaches have been used for their identification and characterization of their molecular targets. However, these methods have mostly relied on sequence homology with other bacterial organisms, leading to a number of challenges and difficulties. Here we introduce a network approach to identify small RNA targets, using a compendium of gene expression arrays. With this approach, we identify and validate a large number of novel targets, as well as correctly identify known targets for several small RNA molecules. Building on the network results, we experimentally validated some of the predicted targets and introduce new insights into the functional roles of small RNAs in the bacterium. We will discuss how these small RNAs regulate specific processes in the cell and how they can be used in fine-tuning these processes. Lastly, we will discuss how network inference approaches can be utilized to identify novel small RNA molecules in bacteria.

# Regulatory network reconstruction from genome-wide gene expression and genotype data

Benjamin A. Logsdon<sup>1</sup>, Jason G. Mezey<sup>1,2</sup>

<sup>1</sup>*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca;*

<sup>2</sup>*Department of Genetic Medicine, Weill Cornell Medical College, NYC*

Genome-wide genotyping and expression profiling have been used to identify novel genetic and molecular variation underlying complex disease phenotypes. We present a scalable algorithm to infer the effects of genetic polymorphism on gene expression, as well as putative regulatory relationships among genes that define disease pathways, based on genome-wide patterns of genotype and gene expression variation in a population sample. The intuition behind the algorithm is to treat the effects of genetic polymorphisms on expression as “random genetic perturbations”, which we leverage to disentangle regulatory relationships among expression phenotypes. These regulatory relationships are represented using a directed graph, where a directed edge between two phenotypes represents a putative regulatory relationship, and directed edges from genotypes onto phenotypes represent putative genetic perturbations. Our representation allows for directed regulatory cycles among the phenotypes.

Our algorithm works by initially reconstructing an undirected graph, based on conditional relationships among gene expression phenotypes and genotypes, using a LASSO type penalization [1]. Based on the zero structure of this undirected graph among genotypes and phenotypes, the algorithm then deconstructs the undirected graph into a directed cyclic graph, with perturbations feeding in from genotypes. This is only possible when the effects of the perturbations are sparse, an expected consequence of the minimal pleiotropy assumed to exist in many genetic systems. We show that the deconstruction is equivalent to solving a non-linear programming problem, which we solve using an iterative mapping algorithm, Divide-and-Concur [2].

The deconstruction of the undirected graph into a directed cyclic graph is a novel approach for inferring the putative causal relationships underlying gene expression variation. Our approach also extracts more information than other methods proposed to identify causal relationships using random genetic perturbations by allowing for regulatory cycles among phenotypes. We demonstrate the applicability of our algorithm by inferring network structure from gene expression and genotype data collected for HapMap individuals [3].

[1] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Ser. B* **58**, 267-288 (1996).

[2] Gravel, S., Elser, V., Divide and concur: A general approach to constraint satisfaction. *Phys. Rev. E* **78**, 36706 (2008).

[3] Stranger, B., et al., Relative impact of nucleotide and copy number variation on gene expression phenotype. *Science* **315**, 848-853 (2007).

# Accurate and Reliable Cancer Classification Based on Probabilistic Inference of Pathway Activity

Junjie Su<sup>1</sup>, Byung-Jun Yoon<sup>1</sup>, Edward R. Dougherty<sup>1,2</sup>

<sup>1</sup>*Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128;* <sup>2</sup>*Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004.*

With the advent of high-throughput technologies for measuring genome-wide expression profiles, a large number of methods have been proposed for discovering diagnostic markers that can accurately discriminate between different classes of a disease. However, factors such as the small sample size of typical clinical data, the inherent noise in high-throughput measurements, and the heterogeneity across different samples, often make it difficult to find reliable gene markers. To overcome this problem, several studies have proposed the use of pathway-based markers, instead of individual gene markers, for building the classifier. Given a set of known pathways, these methods estimate the activity level of each pathway by summarizing the expression values of its member genes, and use the pathway activities for classification. It has been shown that pathway-based classifiers typically yield more reliable results compared to traditional gene-based classifiers. In this paper, we propose a new classification method based on probabilistic inference of pathway activities. For a given sample, we compute the log-likelihood ratio between different disease phenotypes based on the expression level of each gene. The activity of a given pathway is then inferred by combining the log-likelihood ratios of the constituent genes. We apply the proposed method to the classification of breast cancer metastasis, and show that it achieves higher accuracy and identifies more reproducible pathway markers compared to several existing pathway activity inference methods.

## Metabolic strategies to enhance antibiotics action

Mark P. Brynildsen<sup>1,2</sup>, Jonathan A. Winkler<sup>2,3</sup>, James J. Collins<sup>1,2,3</sup>

<sup>1</sup> Howard Hughes Medical Institute; <sup>2</sup> Department of Biomedical Engineering, Center for BioDynamics and Center for Advanced Biotechnology; <sup>3</sup> Program in Molecular Biology, Cell Biology and Biochemistry, Boston University, Boston, MA 02215.

With the ever-increasing incidence of antibiotic-resistant infections and a weak pipeline of new antibiotics, our antibiotic arsenal is in danger of becoming obsolete. To address this issue, both novel antibiotics and novel strategies to enhance the effectiveness of current antibiotics need to be developed. Recently, our lab discovered a common mechanism of cellular death induced by bactericidal antibiotics [1]. Three major classes of bactericidal antibiotics all stimulated production of the hydroxyl radical, which is a highly reactive oxygen species (ROS) generated as a by-product of aerobic respiration and Fenton chemistry. This phenomenon contributed to cell death and was independent of the primary drug-target interaction. Since hydroxyl radical production results from aberrant metabolic activity, we devised metabolic strategies to increase ROS production and potentiate currently available antibiotics. Our approach integrates ROS-generating reactome data and Flux Balance Analysis (FBA) in order to model opportunistic ROS production. With this method, we were able to predict how alterations to metabolism (e.g., enzymatic knockout) influence ROS production, and thereby, enhance or impair antibiotic action.

[1] Kohanski *et al* (2007) *Cell* Sep 7;130(5):797-810.

# Integrated analysis of genomic and proteomic data reveals key role of growth factor signaling network in prostate cancer cells

Andrej Bugrim<sup>1</sup>, Adaikkalam Vellaichamy<sup>3</sup>, Zoltan Dezso<sup>1</sup>, Arul Chinnaiyan<sup>2</sup> Arun Sreekumar<sup>2,4</sup>, Alexey Nesvizhskii<sup>2</sup>, Gilbert S. Omenn<sup>2</sup>

<sup>1</sup>GeneGo, Inc., 500 Renaissance Dr. St. Joseph MI, 49085; <sup>2</sup>Departments of Pathology, Internal Medicine, Human Genetics, School of Medicine, University of Michigan, Ann Arbor, MI, 48109; <sup>3</sup>Precision Proteomics, Institute for Genomic Biology, University of Illinois-Urbana/Champaign, IL, 61801; <sup>4</sup>Current address: Cancer Center, Medical College of Georgia, Augusta, GA.

Prostate cancer is one of the most commonly diagnosed cancers and the second leading cause of cancer-related death in North American men. While androgen withdrawal therapy is often effective initially, most cases progress to the much more aggressive androgen-independent phenotype. Despite significant research efforts, the mechanisms underlying tumor progression are poorly understood. Roles for several signaling pathways have been established, but not a systemic picture. In the present study we have investigated response of LNCaP prostate cancer cells to treatment with synthetic androgen (R1881), as a well-studied model system for prostate cancer progression. We have performed concurrent measurements of gene expression and protein levels following the treatment using microarrays and iTRAQ proteomics. Sets of up-regulated genes and proteins were analyzed using our novel concept of “topological significance”. This method combines high-throughput molecular data with the global network of protein interactions to identify nodes which occupy significant network positions with respect to differentially expressed genes or proteins. Our analysis identified the network of growth factor regulation of cell cycle as the main response module for androgen treatment in LNCaP cells. We show that the majority of signaling nodes in this network occupy significant positions with respect to the observed gene expression and proteomic profiles elicited by androgen stimulus. Our results further indicate that growth factor signaling probably represents a “second phase” response, not directly dependent on the initial androgen stimulus. We conclude that in prostate cancer cells the proliferative signals are likely to be transmitted from multiple growth factor receptors by a multitude of signaling pathways converging on several key regulators of cell proliferation such as c-Myc, Cyclin D and CREB1. Furthermore, these pathways are not isolated but constitute an interconnected network module containing many alternative routes from inputs to outputs. If the whole network is involved, a precisely formulated combination therapy may be required to fight the tumor growth effectively. Moreover, our results show that the predictions made from microarray and proteomics profiles converge on the same set of regulatory proteins despite significant paucity in the experimental data. Thus our approach could be instrumental in translating high-throughput datasets generated by vastly different technologies into consistent predictions of activity of underlying signaling pathways and key regulatory proteins.

## Simulating combined chemotactic and metabolic response using a visual formalism

Antti Larjo<sup>1</sup>, David Harel<sup>2</sup>, Olli Yli-Harja<sup>1</sup>, Hillel Kugler<sup>3</sup>

<sup>1</sup>Department of Signal Processing, Tampere University of Technology, Finland;

<sup>2</sup>Computational Biology Group, Microsoft Research, Cambridge, UK; <sup>3</sup>Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot, Israel;

<sup>3</sup>Computational Biology Group, Microsoft Research, Cambridge.

Some bacteria are able to perform movements towards attractants and/or away from toxins, a feature called chemotaxis. The signaling network responsible for the chemotactic behavior of *Escherichia coli* is particularly well known and it has also been the target of intensive modeling efforts.

Chemotactic bacteria exhibit different kinds of population behavior when subjected to an environment containing substances that can act as chemoattractants for the bacteria. Such population behavior arises from the interplay between chemotaxis and metabolism, in particular by the metabolic activity of the bacteria affecting the attractant concentration in the environment, which in turn influences the chemotactic response of the bacteria. If a model does not take into account the effects of bacteria on the environment and in particular with the attractant / toxin concentrations, it may not capture some of the phenomena related to population behavior, such as formation of the bands seen in chemotactic assays [1].

Our model for *E. coli* combines a precise model of chemotaxis with a comprehensive metabolic network model and allows the dynamic simulation of the combined system and its effect on the simulated environment having spatially varying attractant concentration.

The modeling approach we use is based on a visual language and is aimed at building fully executable models of complex biological systems. The language combines statecharts that capture discrete reactive system behavior at a higher level with other modeling methods, which can describe also the low-level behavior in detail.

In context of our chemotactic model the modeling approach allows us to combine different models for subsystems and to simulate populations of bacteria, as well as to model, e.g., divisions of bacteria. The chemotactic model we present is thus illustrative of the possibilities of our modeling approach but can also produce results that resemble more closely those obtained experimentally than in many previous models.

[1] J. Adler, "Chemotaxis in Bacteria," Science, Vol. 153, no. 3737, pp. 708 – 716.

## Gene Network Analysis of Diabetes Susceptibility Models

Manway Liu<sup>1</sup>, Marcelo A. Mori<sup>2</sup>, Olivier Bezy<sup>2</sup>, Katrine Almind<sup>2</sup>, Hagit Shapiro<sup>2</sup>, Simon Kasif<sup>1</sup>, and C. Ronald Kahn<sup>2</sup>

<sup>1</sup>*Department of Biomedical Engineering, Boston University, Boston Massachusetts, USA;*

<sup>2</sup>*Joslin Diabetes Center, Harvard Medical School, Boston, Massachusetts, USA*

Type 2 Diabetes is a global disease. It is often associated with risk factors such as obesity and is determined by interactions between environmental and genetic factors. While knowledge of the disease has greatly improved in the past decades, much of the underlying mechanisms remain unknown. Two of the factors confounding progress are the heterogeneity of the disease and its relatively weak effects at the level of gene transcript expression.

We have previously described an integrative, gene network analysis approach, called Gene Network Enrichment Analysis (GNEA), which was successfully applied in a meta-analysis of mouse models of the disease. In particular, GNEA identified dysregulated insulin signaling, as well as nuclear receptor signaling, across a spectrum of different gene knockouts, dietary conditions, and drug treatment models. The algorithm worked by searching for a locally optimal, cumulatively differentially expressed gene subnetwork in each experiment and then identifying the signaling pathways that were consistently over-represented across these subnetworks. By examining across multiple experiments, broad recurrent themes in the heterogeneous disease could be identified. The emphasis on cumulative differential expression, in addition, meant that signaling pathways with weakly differentially expressed genes could nonetheless be identified if the genes were robustly correlated with each other.

Despite the success in meta-analysis, GNEA had a number of statistical limitations that made it difficult to draw strong conclusions about the dysregulated signaling pathways in any individual experiment. Consequently, we have since substantially revised and improved GNEA such that it is capable of analyzing both individual experiments and across multiple experiments in a statistically robust manner.

The revised algorithm was recently applied to comparative analyses of the obesity and diabetes prone C57Bl/6 (B6) mouse vs. the obesity and diabetes resistant 129S6/SvEvTac (129) mouse to investigate the molecular mechanisms associated with diabetes susceptibility. Comparisons between the two strains were done at different ages and diets, with GNEA being applied to each comparison and across multiple comparisons. Interestingly, our analysis determined that immune gene sets were most significantly altered between the B6 and 129 mice even at an age when the two strains were otherwise indistinguishable by metabolic measurements. This phenomenon was confirmed by elevated expression of inflammatory markers, higher macrophage density and increased recruitment of lymphocytes to the adipose tissue of the B6 mice. Moreover, with increasing age or high fat diet, differences in adiposity and insulinemia between the strains became manifest, and the disparity in the inflammatory status of the adipose tissue became more pronounced. Taken together, the results suggest that differences in immune response in fat tissue may contribute to the predisposition to metabolic diseases.

# Variability in gene expression underlies incomplete penetrance in *C. elegans*: using single molecules to study the development of single cells

Arijun Rai<sup>1,2</sup>, Scott Rifkin<sup>3</sup>, Erik Andersen<sup>4</sup>, Mitch Guttman<sup>5,6</sup>, Ahmad Khalil<sup>7</sup>, John Rinn<sup>6,7</sup>, Alexander van Oudenaarden<sup>2,5</sup>

<sup>1</sup>University of Pennsylvania, Department of Bioengineering; <sup>2</sup>Massachusetts Institute of Technology, Departments of Physics; <sup>3</sup>University of California, San Diego, Department of Biology; <sup>4</sup>Princeton University, Lewis-Sigler Institute; <sup>5</sup>Massachusetts Institute of Technology, Department of Biology; <sup>6</sup>Broad Institute; <sup>7</sup>Harvard Medical School, Department of Pathology.

Phenotypic variation is ubiquitous in biology and is often traceable to underlying genetic and environmental variation. However, even genetically identical organisms in homogenous environments vary, suggesting that random processes may play an important role in generating phenotypic diversity; indeed, stochastic effects in gene expression can generate beneficial phenotypic variation in microbes. Few studies, however, have explored the impact of stochastic fluctuations in gene expression on phenotypic variation and cell fate decisions in multicellular organisms. In order to examine the consequences of gene expression variability in development, we explored intestinal specification in *C. elegans*, in which wild-type cell fate is invariant and controlled by a small transcriptional network. In contrast, cell fates in embryos with mutant *skn-1*, the first gene expressed in this network, are variable: while most mutant embryos fail to develop intestinal cells, some embryos nevertheless produce intestinal precursors. By counting transcripts in individual embryos, we show that mutations in *skn-1* result in large variability in the expression of the downstream gene *end-1*, arising partly from misregulation of chromatin remodeling. *end-1* expression is subsequently thresholded during a critical time window to produce an ON/OFF expression pattern of *elt-2*, the master regulator of intestinal differentiation. The loss of *skn-1* activity eliminates redundancy in the network, making *elt-2* activation particularly sensitive to variability in *end-1* expression. Although *end-3* can also activate *elt-2*, deleting *end-3* in wild-type animals results in variability in levels and timing of *elt-2* expression, suggesting that robust expression of the downstream target requires multiple transcriptional activators and also hinting at subtle differences in the roles of putatively redundant elements in the network. Thresholds and redundancy are common features of developmental networks, and our results show that mutations in such networks can expose otherwise buffered stochastic variability in gene expression, leading to pronounced phenotypic variation.

We also briefly discuss some results examining variability in the expression of new classes of non-coding RNAs and their implications for differentiation.

# Pursuing Pluripotency: Systems Level Quantitative Approaches to Understand Stem Cell Fate Decisions

Ihor R. Lemischka<sup>1</sup>

<sup>1</sup>*Professor of Gene and Cell Medicine and Director of the Black Family Stem Cell Institute, Mt. Sinai School of Medicine, New York, NY.*

Many years of genome-wide analyses of embryonic stem cells (ESCs) have provided extensive molecular “parts list” that collectively regulate the pluripotent state and mediate transitions in cell fate. The assembly of these parts into functional regulatory networks and linking such networks to measurable cell phenotypes is a major challenge. Similarly, efforts to elucidate regulatory network dynamics in concert with changes in cell fate are in their infancy. We have developed a short hairpin (sh) RNA-based genetic complementation strategy that allows us to precisely control dynamic expression of key pluripotency regulators such as Nanog, Esrrb and others. Temporal epigenomic, transcriptional, mRNA and proteomic analyses after down regulating a regulatory gene-product have provided an in-depth, multi-level view of an alteration in ESC fate. Computational analyses of our integrated data sets together with already published data sets have suggested regulatory network architectures and dynamics; in effect producing “movies” of changing ESC fates. Single cell studies are being pursued to explore the existence of “alternative” pluripotency “states” and to address issues such as the role of “noise” in cell fate decisions. Other efforts are focused on linking “generic” epigenetic regulatory machinery with sequence-specific transcriptional regulators in ESC and during the process of reprogramming adult cell to induced pluripotent stem cells (iPSCs). We have also embarked on efforts to generate patient-specific iPSC lines in order to model the etiology of complex genetic diseases and to understand the mechanisms of the reprogramming process. In addition, we are implementing synthetic biology approaches to engineer artificially controllable circuitry in both ESCs and iPSC. These efforts will impact on programming stem cells to distinct fates and will provide additional insights into regulatory circuit dynamics.

# Systems Biology Analysis and Prediction of Human Disease Genes

Dennis Vitkup<sup>1,2</sup>, Sarah Gilman<sup>1,2</sup>, Tzu-Lin Hsiao<sup>1,2</sup>,

<sup>1</sup>*Columbia University, Center for Computational Biology and Bioinformatics;* <sup>2</sup>*Department of Biomedical Informatics*

By analyzing, in parallel, large literature-derived and high-throughput experimental datasets we investigate genes harboring human inherited disease mutations in the context of human molecular networks. Our results demonstrate that network properties significantly influence the likelihood and phenotypic consequences of disease mutations. Genes with intermediate connectivities have the highest probability of harboring germ-line disease mutations, suggesting that disease genes tend to occupy an intermediate niche in terms of their physiological and cellular importance.

Our analysis also demonstrates that the functional compensation by close sequence homologs may play an important role in human genetic disease. Genes with a 90% sequence identity homolog are about 3 times less likely to harbor known disease mutations compared to genes with remote homologs. Moreover, close duplicates affect the phenotypic consequences of deleterious mutations by making a decrease in life expectancy significantly less likely. We also demonstrate that similarity of expression profiles across tissues significantly increases the likelihood of functional compensation by homologs.

We show that disease mutations are less likely to occur in essential genes compared with all human genes. Disease genes display significant functional clustering in the analyzed molecular network. For about one-third of known disorders with two or more associated genes we find physical clusters of genes with the same phenotype. These clusters are likely to represent disorder-specific functional modules and suggest a framework for identifying yet-undiscovered disease genes.

We develop a probabilistic method to project disease phenotypes from known disease genes to all genes in the human genome. We apply the developed method to investigate gene overlap between different disease phenotypes and analyze the results of several recent Genome Wide Association studies.

# Predicting Metabolic Engineering Knockout Strategies for Chemical Production: Accounting for Competing Pathways

Naama Tepper<sup>1</sup> and Tomer Shlomi<sup>1</sup>

<sup>1</sup>Department of Computer Science, Technion–IIT, Haifa 32000, Israel

**Metabolic engineering aims to use microbes as factories that can produce and degrade organic molecules for industrial and biomedical purposes.** In recent years metabolic engineering was successfully employed to produce various fuels and chemicals, and significant efforts are made for over-producing additional chemicals. Various computational methods are used to design genetic manipulations that can achieve metabolic engineering goals using constraint-based modeling of genome-scale metabolic networks. These methods aim to anticipate the effect of genetic manipulations on cellular metabolism, searching for specific manipulations that would lead to maximized production rate of chemicals of interest. For example, a commonly used method called OptKnock searches for sets of gene knockouts whose implementation should lead to the over-production of desired metabolites. However, **current methods do not account for the presence of competing pathways in a metabolic network that may diverge metabolic flux away from producing a required chemical**, resulting in lower (or even zero) chemical production rates in reality, leading to over-optimistic predictions regarding achievable chemical production rates.

**We developed a novel computational method called RobustKnock that predicts genetic manipulations that maximize the *guaranteed* production rate of chemicals of interest (given known constraints embedded in the model),** accounting for the presence of alternative pathways in the network. Specifically, this method extends OptKnock to pinpoint specific enzyme-catalyzed reactions that should be removed from a metabolic network, such that **the production of the desired product becomes an obligatory by-product of biomass formation** (i.e. required for cellular growth) due to stoichiometric mass-balance, thermodynamic, and flux capacity constraints. The predicted set of gene knockouts eliminates all competing pathways that may hinder the chemical's production rate, resulting in more robust predictions than those obtained with OptKnock. This is achieved by searching for a set of gene knockouts under which the minimal (guaranteed) production rate of a chemical of interest is maximized, instead of simply assuming that the maximized production rate would be achieved by chance as in OptKnock. The method is based on a bi-level max-min optimization problem that is efficiently solved via a transformation to a standard mixed-integer linear programming (MILP) problem. Applying RobustKnock to predict gene knockout strategies for the over-production of various chemicals in *E. coli* resulted in knockout strategies that are more robust than those achieved with OptKnock. Knockout strategies predicted by RobustKnock are shown to provide a guaranteed minimal chemical production rates that are close to the maximal theoretical rates predicted by OptKnock. A specific interesting prediction made by RobustKnock is regarding the production of Ethanol, which is a common target for metabolic engineering trials because of its potential usage as a bio-fuel. In this case, RobustKnock predicted a triple-knockout strategy that yields a minimal, guaranteed production rate that is very close to the maximal theoretical rate predicted by OptKnock, whereas the knockout strategy predicted by the latter may actually lead to zero production rate in reality.

# Characterizing Dynamic Changes in the Human Blood Transcriptional Network

Jun Zhu<sup>1</sup>, Yangqing Chen<sup>1</sup>, Amy Leonardson<sup>1</sup>, Kai Wang<sup>1</sup>, John R. Lamb<sup>1</sup>, Valur Emilsson<sup>2</sup>, Eric E. Schadt<sup>1</sup>

<sup>1</sup> Department of Genetics, Rosetta Inpharmatics, LLC, a wholly owned subsidiary of Merck & Co., Inc., 401 Terry Ave North, Seattle, WA 98109.; <sup>2</sup> Molecular Profiling Research Informatics Department, Merck Research Laboratories, 126 E. Lincoln Ave., Rahway NJ 07065-0900

Gene expression data generated systematically in a given system over multiple time points provides a source of perturbation that can be leveraged to infer causal relationships among genes explaining network changes. Previously, we showed that food intake has a large impact on blood gene expression pattern and that these responses, either in terms of gene expression level or gene-gene connectivity, are strongly associated with metabolic diseases. In this study, we explored which genes drive the changes of gene expression patterns in response to time and food intake. We applied the Granger causality test and the dynamic Bayesian network to gene expression data generated from blood samples collected at multiple time points during the course of a day. The simulation result shows that combining many short time series together is as powerful to infer Granger causality as using a single long time series. Using the Granger causality test, we identified genes that were supported as the most likely causal candidates for the coordinated temporal changes in the network. These results show that *PER1* is a key regulator of the blood transcriptional network, in which multiple biological processes are under circadian rhythm regulation. The fast and fed dynamic Bayesian networks showed that over 72% of dynamic connections are self links. Finally, we show that different processes such as inflammation and lipid metabolism, which are disconnected in the static network, become dynamically linked in response to food intake, which would suggest that increasing nutritional load leads to coordinate regulation of these biological processes. In conclusion, our results suggest that food intake has a profound impact on dynamic co-regulation of multiple biological processes, such as metabolism, immune response, apoptosis and circadian rhythm. The results could have broader implications on study designs of disease association or drug response in clinical trials.

# Network-free Inference of Knockout Effects in Yeast

Tal Peleg<sup>1,\*</sup>, Nir Yosef<sup>1,\*</sup>, Eytan Ruppin<sup>1,2</sup>, Roded Sharan<sup>1</sup>

<sup>1</sup>Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel;

<sup>2</sup>Affiliation School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel; *These authors contributed equally to this work*

Perturbation experiments, in which a certain gene is knocked out and the expression levels of other genes are observed, constitute a fundamental step in uncovering the intricate wiring diagrams in the living cell and elucidating the causal roles of genes in signaling and regulation.

Here we present a novel framework for analyzing large cohorts of gene knockout experiments and their genome-wide effects on expression levels. We devise clustering-like algorithms that identify groups of genes that behave similarly with respect to the knockout data, and utilize them to predict knockout effects and to annotate physical interactions between proteins as inhibiting or activating. In difference from previous approaches, our prediction approach does not depend on physical network information; the latter is used only for the annotation task. Consequently, it is both more efficient and of wider applicability than previous methods.

We evaluate our approach using a large scale collection of gene knockout experiments in yeast, comparing it to the state-of-the-art SPINE algorithm. In cross validation tests, our algorithm exhibits superior prediction accuracy, while at the same time increasing the coverage by over 25-fold. Significant coverage gains are obtained also in the annotation of the physical network.

# Conexic: A Bayesian framework to detect drivers and their function uncovers an endosomal signature in Melanoma

Uri David Akavia<sup>1,2</sup>, Oren Litvin<sup>1,2</sup>, Eyal Moses<sup>1</sup>, Yossi Tzur<sup>1</sup>, Dylan Kotliar<sup>1</sup>, Jessica Kim<sup>3,4</sup>, Patrick Oberholzer<sup>3,4</sup>, Levi A. Garraway<sup>3,4</sup>, Dana Pe'er<sup>1,2</sup>

<sup>1</sup>Department of Biological Sciences, Columbia University; <sup>2</sup>Center for Computational Biology and Bioinformatics, Columbia University; <sup>3</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School; <sup>4</sup>Broad Institute of Harvard and MIT

Genomics is revolutionizing our understanding of cancer biology. Tumor samples assayed for comprehensive chromosomal and gene expression data are accumulating at a staggering rate. A major challenge involves the development of analysis methods to uncover biological insights from these data, including the identification of the key mutations that drive cancer and how these events alter cellular regulation.

We have developed Conexic, a novel Bayesian Network-based framework to integrate chromosomal copy number and gene expression data to detect genetic alterations in tumors that drive proliferation, and to model how these alterations perturb normal cell growth/survival. The underlying assumption to our approach is that significantly recurring copy number change, coinciding with its ability to predict the expression patterns varying across tumors, strengthens the evidence of a gene's causative role in cancer. This method not only pinpoints specific regulators within an a large region of copy number variation, but can identify the effected targets and shed light on the way in which gene regulation is altered

We applied our Conexic framework to a melanoma dataset (Lin et al, Cancer Research, 2007) comprising 62 paired measurements of gene expression and copy number. In addition to confirming the role of known drivers in melanoma, our analysis suggests a number of novel drivers. Most strikingly, these point to a major role of protein trafficking and endosome biology in regulating this malignancy. These results have linked endosomal processing and sorting to adhesion and survival. Preliminary experimental validation supports several of these findings. Together, these results affirm the potential of Conexic to elaborate novel driver modules with biological and possibly therapeutic importance in melanoma and other cancers.



# Large Scale Gene Set Analysis of Stem Cell and Oncology Gene Expression Signatures

Daniel Gusenleitner<sup>1</sup>, Aedin C. Culhane<sup>1,4</sup>, Thomas Schwarzl<sup>1</sup>, Kermshlise C. Picard<sup>1</sup>, Shaita C. Picard<sup>1</sup>, Stefan Bentink<sup>1,4</sup>, Razvan Sultana<sup>1</sup>, Gerald Papenhausen<sup>1,2</sup>, Joe White<sup>1,4</sup>, Mick Correll<sup>1,2</sup>, John Quackenbush<sup>1,2,3,4</sup>

<sup>1</sup>*Biostatistics and Computational Biology*, <sup>2</sup>*Centre for Cancer Computational Biology*, <sup>3</sup>*Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA*. <sup>4</sup>*Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA*,

Experimentally derived gene expression signatures (EDGEs) reflect the transcriptional status of cells in a particular biological state. This distinguishes these signatures from other categorical collections of genes such as those in databases such as Gene Ontology, KEGG or BioCarta. The genes in each EDGE provide an indication to the biological pathways that are concurrently active in a cell state. Large scale analysis of the overlap and dependencies of these gene signatures may provide a means to deconvolve the inter-relationships between biological pathways, an essential component of systems biology research. **Here we describe an approach based on analysis of overlapping genes within EDGES designed to provide insight into the relationship between stem cells and cancer based on gene expression patterns.**

EDGEs (n=465) were manually extracted from published literature following a thorough search of PubMed using a defined set of cancer and stem cell gene signature search terms. These signatures were subject to rigorous quality control, mapped to the human genome in a consistent manner, and stored in a Gene Signature Database [1]. Gene set analysis was performed against a collection of cancer gene expression datasets (n=216) from 15 cancer types in the GeneChip Oncology Database GCOD [2]. Global analysis of EDGEs across hundreds of cancer datasets identified common features of stem cells and cancer, cancer-specific signatures, and tissue-specific profiles. Understanding these profiles and the relationships between them, promises to provide new avenues for identification of prognostic and therapeutic targets.

[1] GeneSigDB <http://compbio.dfci.harvard.edu/GeneSigDB>

[2] GCOD <http://compbio.dfci.harvard.edu/GCOD>

# Subspace Differential Coexpression Analysis for the Discovery of Disease-related Dysregulations

Gang Fang<sup>1</sup>, Rui Kuang<sup>1</sup>, Gaurav Pandey<sup>1</sup>, Michael Steinbach<sup>1</sup>, Chad L. Myers<sup>1</sup> and Vipin Kumar<sup>1</sup>

<sup>1</sup>*Department of Computer Science and Engineering, University of Minnesota, Minneapolis*

Differential expression (DE) analysis is widely used to identify over- or under-expressed genes that may be associated with diseases. Differential coexpression (DC) analysis adopts a complementary viewpoint, and targets the discovery of sets of genes that have different levels of coexpression across the two sample-classes, i.e., highly coexpressed in one class, but not in the other. Biologically, DC patterns may indicate the disruption of a regulatory mechanism possibly caused by the dysregulation of pathways or mutations of transcription factors.

Existing techniques for identifying DC patterns can be grouped into different categories based on: (i) the coexpression measures they use, (ii) the size of patterns they discover and (iii) the adopted algorithms. A common feature of all the existing approaches for DC analysis is that the coexpression of a set of genes is measured on all the samples in each of the two classes. Hence, they may miss patterns that only cover a subset of samples in each class, i.e., subspace patterns, due to the heterogeneity of the disease causes and subject population.

In this work, we extend differential coexpression analysis by defining a *subspace* DC pattern, i.e., a set of genes that are coexpressed in a relatively large fraction of samples in one class, but in a much smaller fraction of samples in the other class. We propose a general approach based upon the association analysis framework that allows an exhaustive yet efficient discovery of subspace DC patterns in a bottom-up manner. Within this framework, a family of biclustering algorithms, whose coexpression measures are antimonotonic, can be adapted into their corresponding differential versions for the direct discovery of DC patterns. Using a recently developed biclustering algorithm for illustration, we perform experiments on a combined lung cancer gene expression dataset. Permutation tests demonstrated the statistical significance of many subspace DC patterns (560 of size 2, 88 of size 3 and 7 of size 4), about 70% of which cannot be discovered if they are measured over all the samples in each of the classes. The biological relevance of the size-3 and size-4 patterns is evaluated via enrichment analysis with ten known cancer pathways and the molecular signature database. Interestingly, several subspace patterns have significant overlap with known cancer pathways (e.g. TNF $\alpha$ /NF $\kappa$ B and WNT signaling pathways). We also found several patterns that are enriched with the target gene sets of cancer-related microRNAs (e.g. miR-101) and cancer-related transcription factors (e.g. ATF2). Discovering them as subspace patterns allows further study of different causes of diseases and different demographics among subgroups of samples, which may potentially help personal diagnosis and treatment. By merging small patterns that overlaps with each other, bigger patterns (e.g. of size 10) are discovered, which correspond to densely connected gene/protein interaction subnetworks. We further organized all the discovered DC patterns into a single DC network, which provides a global view of these subnetworks and the overlaps among them.

[1] For details about this work and the source codes, please visit <http://vk.cs.umn.edu/SDC>

## Insights from proteomics into cellular evolution and surprising disease models

Edward M. Marcotte<sup>1,2</sup>, Kriston L. McGary<sup>1</sup>, Tae Joo Park<sup>1,3</sup>, Christine Vogel<sup>1</sup>, Jon Laurent<sup>1</sup>, John O. Woods<sup>1</sup>, Hye Ji Cha<sup>1</sup>, and John B. Wallingford<sup>1,3,4</sup>

<sup>1</sup>Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, <sup>2</sup>Department of Chemistry, and Biochemistry, <sup>3</sup>Department of Molecular Cell, and Developmental Biology, and <sup>4</sup>Howard Hughes Medical Institute, University of Texas at Austin, 2500 Speedway, Austin, TX 78712–1064, USA.

High-throughput protein function, expression, interaction, and localization assays are becoming widespread, producing thousands of systematically measured features of genes, their encoded proteins, and their mutational phenotypes. A central challenge is connecting this growing molecular- and cellular-level information with growing data on genetic variation to interpret the organismal consequences of this variation. Using data from quantitative shotgun proteomics, we describe the extent to which protein levels are constrained evolutionarily. These constraints relate to the tendency for proteins to operate not in isolation, but in pathways and complexes, and a systematic analysis of protein complexes reveals an intimate relationship between protein complexes and mutational phenotypes that is likely conserved across evolution. Because of these conserved relationships among protein expression, interactions, and mutational phenotypes, phenotypes can be mapped rationally between organisms, revealing new models of disease and candidate disease genes (the phenolog hypothesis), which we demonstrate for the case of new angiogenesis genes. Consideration of the cellular organization of proteins thus leads directly to an explanation of the organismal-level consequences of their perturbation.

# Novel methods for the discovery of condition specific master regulators of transcription

Kenzie D. Maclsaac<sup>1</sup>, Chris Ng<sup>1</sup>, Ernest Fraenkel<sup>1,2</sup>

<sup>1</sup>*Department of Biological Engineering, MIT;* <sup>2</sup>*Computer Science and Artificial Intelligence Laboratory, MIT*

Identifying the specific transcription factors that assemble at enhancers and drive tissue-specific and condition-specific transcription in mammals is an important and unsolved problem. Exhaustive genome-wide experimental strategies that profile almost all transcription-factor binding have been applied in yeast, but are not practical for mammalian systems where there are approximately 10-fold more regulatory proteins and a multitude of distinct tissues.

To address these limitations we have developed a joint experimental and computational strategy. We identified regulatory sites using either (1) genome-wide binding for coregulators (ChIP-Seq) or (2) genome wide DNase-hypersensitive data (DNase-Seq). By learning which DNA sequence motif features discriminate experimentally identified sequence regions from background sequence, we reveal the master regulators responsible for recruiting coregulators to their targets and altering chromatin structure.

The approach was carried out in mouse liver, cerebellum, and 3T3-L1 cells. Interestingly, the sequence motifs associated with recruitment of a coregulator protein varies across tissues. We further demonstrate that the transcription factors predicted to associate with these sites in vivo are indeed bound in ChIP experiments, and that regions bound by multiple transcription factors are more likely to recruit a coregulator. We tested several predictive models of coregulator recruitment and found that simple models, where individual motifs contribute independently to coregulator binding probability, generally perform as well or better than more complex models. We find no strong evidence of particular motif spacing or orientation constraints associated with coregulator binding

# Determining frequent patterns of copy number alterations in cancer

Franck Rapaport<sup>1</sup>, Christina Leslie<sup>1</sup>

<sup>1</sup>Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, NY

Cancer progression is often characterized by increasing genomic instability, giving rise to a complicated landscape of genomic alterations within an individual tumor and great diversity of these copy number aberrations (CNAs) across tumor samples. In recent years, array-based comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) arrays have been used to analyze the CNAs of tumor samples at a genomic scale and at progressively higher resolution. Individual CNAs may be as small as a few adjacent probes or as large as a whole chromosomes and may be difficult to detect above probe-level noise; moreover, it is unclear how to make sense out of diverse CNAs from hundreds of tumors.

Typical studies take a pipeline approach, starting with segmentation of individual copy number profiles, followed by a statistical procedure to call significant gains and losses, and ending with some analysis across samples to determine frequent CNAs or clusters with similar patterns of aberrations. The disadvantage of pipeline approaches, however, is that algorithmic choices and tuning parameters at each step may produce very different results, and mistakes or biases are propagated forward.

In this study, we propose a novel and mathematically robust method that avoids a pipeline approach and exploits probe-level correlations in aCGH data to discover subsets of samples that display common CNAs. Our method is related to recent work on maximum-margin clustering, which extends support vector machine-like optimization approaches to the problem of unsupervised learning (Figure 1).

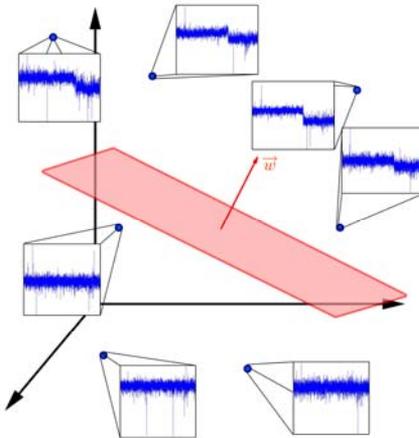


Figure 1: Our approach iteratively partitions aCGH samples into groups that share common CNAs. Each partition involves solving a large-margin optimization problem to determine both the vector  $w$ , which geometrically represents the normal vector to a hyperplane (shown in red) separating the samples, and the assignment of samples to groups. Constraints on  $w$  yield a sparse and piecewise constant solution, encouraging the partition to identify CNAs.

We tested our approach on a large cohort of glioblastoma aCGH samples from The Cancer Genome Atlas initiative and found that the results of our method were largely consistent with the original study in that they included almost all previously reported CNAs. However, we also found additional significant CNAs (including deletion of chromosome 21 and a deletion of a large part of the q arm of chromosome 19) missed by the original analysis but supported by earlier studies. Moreover, we were also able to find correlation between different CNAs, which are hard to investigate using classical methods of analysis.

# New insights into cross-species conservation of functional data

Guy E. Zinman\*<sup>1</sup>, Shan Zhong\*<sup>1</sup>, Ziv Bar-Joseph<sup>1</sup>

<sup>1</sup>Lane center for computational biology, Carnegie Mellon University, Pittsburgh PA, 15213.

\* These authors contributed equally to the work

Many biological processes operate in a similar manner across a large number of species (e.g., cell cycle), and often genes that participate in these processes share high sequence similarity. However, recent studies show that while the sequence information between close species is conserved, interaction data seem to differ substantially. Examples include co-expression studies (only 20% similarity across a large range of tissues between human and mouse), protein-protein interactions (less than 20% similarity between budding and fission yeast), protein-DNA interactions (less than 20% agreement for targets of the same TF in human and mouse) and genetic interactions (as low as 5% between budding and fission yeast).

This study tries to bridge between these two apparently contradictory observations: the high level similarity on the sequence and pathway levels, and the low conservation on the interaction level. To help explain this discrepancy we are checking the possibility of the existence of an intermediate level 'meta gene' that still retains most of the functional data.

We have carried out comprehensive analysis of the agreement of interaction datasets across three model organisms. The two yeast *S. cerevisiae* and *S. pombe*, and the nematode *C. elegans*. All datasets were converted to network representation that could easily be compared across species. Our results indicate that even though each of the individual interaction types is indeed not well conserved, their union is. Thus, a protein that participates in a specific biological process in one species, by interacting with other proteins that participate in the process, will show a similar behavior in another species. However, the type of interaction may change between the species, or that interaction may be mediated by a third protein in the other species. This is especially evident when focusing on specific biological processes like cell cycle or stress response, in which high functional conservation rates were found.



# Estimating the Stochastic Bifurcation Structure of Cellular Networks

Carl Song<sup>1,2</sup>, Vida Abedi<sup>2</sup>, Matt Scott<sup>3</sup>, Brian P. Ingalls<sup>3</sup>, Mads Kaern<sup>2</sup>, Theodore J. Perkins<sup>1,2</sup>,

<sup>1</sup>Ottawa Hospital Research Institute; <sup>2</sup>University of Ottawa; <sup>3</sup>University of Waterloo.

High throughput measurement of gene expression at the single cell level, combined with systematic perturbation of environmental or cellular variables, allows one to map out the steady state behaviors of cellular networks and their responses to varying conditions. In dynamical systems theory, this information is the subject of bifurcation analysis. Because cellular networks are inherently noisy, we generalize the traditional notion of a bifurcation diagram to define the *stochastic bifurcation structure* of a stochastic dynamical system. Moreover, we describe how statistical methods for density estimation, in particular, mixture density modeling and conditional mixture density modeling, can be employed to estimate the stochastic bifurcation structure of a system directly from empirical data. We apply these methods to single cell expression data measuring activity in the galactose network of *S. cerevisiae* at varying extracellular concentrations of galactose. Of the approaches tested, conditional density estimation and a somewhat nontraditional combination of mode detection and expectation-maximization appear most successful at accurately locating induced and noninduced subpopulations from noisy data, especially when one of those subpopulations is small. The proposed approach allows us to make several novel qualitative and quantitative observations about the switching behavior of the galactose network.

# Ground State Robustness as an Evolutionary Design Principle in Signaling Networks

Önder Kartal<sup>1,2</sup>, Oliver Ebenhöf<sup>1-3</sup>

<sup>1</sup>*Institute of Biochemistry and Biology, University of Potsdam, Potsdam/Golm, Germany;*

<sup>2</sup>*Max-Planck-Institute of Molecular Plant Physiology, Potsdam/Golm, Germany;*

<sup>3</sup>*Institute for Complex Systems, University of Aberdeen, Aberdeen, AB24 3UE, UK*

The ability of an organism to survive depends on its capability to adapt to external conditions. In addition to metabolic versatility and efficient replication, reliable signal transduction is essential. As signaling systems are under permanent evolutionary pressure one may assume that their structure reflects certain functional properties. However, despite promising theoretical studies in recent years, the selective forces which shape signaling network topologies in general remain unclear. Here, we propose prevention of autoactivation as one possible evolutionary design principle. A generic framework for continuous kinetic models is used to derive topological implications of demanding a dynamically stable ground state in signaling systems. To this end graph theoretical methods are applied. The index of the underlying digraph is shown to be a key topological property which determines the so-called kinetic ground state (or off-state) robustness. The kinetic robustness depends solely on the composition of the subdigraph with the strongly connected components, which comprise all positive feedbacks in the network. The component with the highest index in the feedback family is shown to dominate the kinetic robustness of the whole network, whereas relative size and girth of these motifs are emphasized as important determinants of the component index. Moreover, depending on topological features the maintenance of robustness differs when networks are faced with structural perturbations. This structural off-state robustness, defined as the average kinetic robustness of a network's neighborhood, turns out to be useful since some structural features are neutral towards kinetic robustness, but show up to be supporting against structural perturbations. Among these are a low connectivity, a high divergence and a low path sum. All results are tested against real signaling networks obtained from databases. The analysis suggests that ground state robustness may serve as a rationale for some structural peculiarities found in intracellular signaling networks.

# Modeling 3D Flies: reconstructing the drosophila segmentation network on the embryo geometry

Jonathan Bieler<sup>1</sup>, Christian Pozorini<sup>1</sup>, and Felix Naef<sup>1,2</sup>

<sup>1</sup>Computational Systems Biology Group, Ecole Polytechnique Federale de Lausanne.

<sup>2</sup>Swiss Institute of Bioinformatics (SIB).

The segmentation process in the early *Drosophila* embryo results from the dynamic establishment of patterned mRNA and protein profiles. The recent availability of spatio-temporal mRNA and protein expression atlases on the full 2D surface of the syncytium opens new possibilities for modeling this complex process. Until now, most models have assumed a one-dimensional geometry along a portion of the anterior-posterior axis, motivated by the nearly rotationally symmetrical observed patterns along this axis. While this approximation has been fruitful, the new data from the whole surface of the blastoderm justifies an extension of the models to the full geometry of the embryo.

Here, we develop a reaction diffusion model for the gap gene network on the curved surface of the blastoderm. We model the dynamics of both mRNA and protein of four trunk gap genes expression during the cleavage cycles 12, 13 and 14A: hunchback, Kruppel, giant and knirps. The model takes as a regulatory inputs the protein expression of the maternal bicoid and caudal gradients, plus the zygotic tailless and huckebein. The model is calibrated using non-linear optimization showing that the main features of spatio-temporal patterning on the whole embryo are well captured. However, anterior domains, e.g. those in the giant gene, are the most difficult to reproduce probably reflecting oversimplifying assumptions or missing genes in the network. A detailed analysis of hunchback suggests that it has concentration dependent activity. We implement this possibility and show that it leads to significant improvements. The model is further assessed by comparing predictions for gap gene mutants with experimental patterns showing satisfactory agreement. However, the current model is less successful at quantitatively predicting the shifts observed in bicoid dosage mutants. Covariance analysis around the optimal model identifies the stiff and soft directions in parameter space, showing e.g. that the regulation of Kruppel by the maternal gradient has to be tightly controlled. In conclusion, modeling patterning on the full egg captures and predicts both qualitative and quantitative aspects of early *drosophila* patterning, while uncovering important design properties of the regulatory network.

## Finding the Rules by Asking the Right Questions: Lessons From Non-Modular Behavior of the *eve* Promoter.

John Reinitz<sup>2</sup>, Hilde Janssens<sup>1,2</sup>, Ah-Ram Kim<sup>1</sup>, Carlos Martinez<sup>1</sup>, Maria Samsonova<sup>3</sup>, David H. Sharp<sup>4</sup>, and John Reinitz<sup>1</sup>

<sup>1</sup>Stony Brook University, Stony Brook, NY; <sup>2</sup>CRG, Barcelona, Spain; <sup>3</sup>State Polytechnic University, St. Petersburg, Russia; <sup>4</sup>Los Alamos National Laboratory, Los Alamos, NM.

The prediction of expression patterns from genomic sequence is an important unsolved problem in modern molecular genetics. Its solution requires an understanding of the transcriptional consequences of particular configurations of bound factors. An important aspect of the problem is to understand how modular enhancers arise from binding sites. We are currently using the *eve* gene of *Drosophila* as a testbed for finding the general rules by which sequence controls gene expression in metazoa. We believe that the most informative experimental materials for such studies are instances where the usual additive behavior of enhancers breaks down. Such instances can reveal underlying rules, but the complexity of the experimental phenomena require precise quantitative models for their interpretation. We consider two experimental situations in which modularity breaks down. In one case, a modular enhancer for stripe 2 fused to proximal sequences that do not drive any expression results in a fragment that expresses stripe 7, demonstrating nonmodular behavior. In another case, placing enhancers for stripes 2 and 3 adjacent to one another give rise to a novel expression pattern, an example of another type of nonadditive behavior. I will show how both types of nonadditive behavior can be understood using a quantitative model in conjunction with quantitative data from promoter-reporter constructs.

# Combinatorial Complexity in Systems Biology

Walter Fontana<sup>1</sup>

<sup>1</sup>*Systems Biology, Harvard Medical School*

The past decades have brought into view a staggering web of protein-protein interactions that collectively give rise to plastic, adaptive and coherent system behavior. What startles about these networks is their combinatorial complexity of post-translational modifications and assembly into noncovalent complexes. Curated protein-protein interaction maps derived from Y2H assays indicate extensive pleiotropy (meaning that a given protein often participates in many different complexes) and conflict (meaning that many proteins often compete for the same binding site). The dynamic phenomena associated with such combinatorial complexity have not been properly explored yet. Combinatorial complexity is linked to concurrency, which emphasizes the role of causal (in)dependence between interactions in shaping system dynamics. Systems marked by combinatorial complexity may exhibit assembly logjams and stochasticity in their composition. The latter is a phenomenon distinct from fluctuations in small networks and reflects the astronomic number of possibilities available to combinatorial systems. I will discuss a few themes emerging from glimpses at this vast and barely explored territory, while sketching the computational approach that made those glimpses possible. At the end of the day, combinatorial complexity may cause network "fluidity" and pervasive process interference, suggesting a view of the cell that is distinct from a wide-spread engineering metaphor based on sharply defined circuits.

This is joint work with (in alphabetical order): John Bachman [Harvard], Vincent Danos [Edinburgh], Eric Deeds [Harvard], Jerome Feret [INRIA, France], Russ Harmer [Harvard] and Jean Krivine [CNRS, France]. Thanks to the team at Plectix BioSystems ([www.cellucidate.com](http://www.cellucidate.com)) for assembling the software engines that enabled our research.

# Super-Metabolism and Functional Capacity in Microbial Communities

Elhanan Borenstein<sup>1,2</sup>, Marcus W. Feldman<sup>1</sup>

<sup>1</sup> *Department of Biology, Stanford University, Stanford, CA 94305-5020, USA;* <sup>2</sup> *Santa Fe Institute, Santa Fe, New Mexico 87501, USA.*

The vast majority of microbial species inhabit complex, diverse, and largely uncharted communities of varying sizes and structures. These communities play an essential role in human health, agriculture, and ecosystem dynamics. Previous studies demonstrated a strong association between the compositions of various microbial communities and their habitats. Clearly, natural selection acts not only upon each individual species in the community, but also upon the species composition of the community as a whole, making it an adaptive “super-organism”. Yet, our understanding of the contribution of community structure to functional capacity and the extent to which such contributions are universal is still lacking.

Here, we examine whether microbial communities are indeed universally endowed with a “super-metabolism” that allows them to carry out metabolic functions that are not within the reach of individual species or random collections of microbial species. Analyzing 953 fully sequenced microbial species and 50 metagenomic samples of various microbiomes, we find that many microbial communities have a significantly high number of novel enzymatic combinations in comparison to randomly assembled communities with similar complexities. Furthermore, enriched novel binary combinations cluster into higher-level unique metabolic modules, each of which typifies a set of microbiomes with similar habitats. Reconstructing the metabolic network of each community further reveals fundamental topological and functional properties that characterize these super-organisms.

Considering the incredible complexity of many microbiomes, the low coverage of metagenomic data, and our limited knowledge of most microbial species on earth, fully characterizing the relationship between community structure and function is a daunting task. The large-scale analysis presented here introduces a novel approach for studying community-level functional capacities that directly stem from community composition. Further analysis of this relationship can dramatically improve our understanding of microbial ecology, with potential applications for the treatment of microbiota related human diseases, environmental stewardship, and the design of synthetic communities.

# Predicting synthetic environments that induce microbial cross-feeding

Niels Klitgord<sup>1</sup>, Daniel Segrè<sup>1,2</sup>

<sup>1</sup>Program in Bioinformatics and Systems Biology, Boston University; <sup>2</sup>Department of Biology and Department of Biomedical Engineering, Boston University

Microbial communities are found ubiquitously on our planet and play important roles in human health, environmental dynamics and industrial applications. The symbiotic interactions underlying several aspects of the population and evolutionary dynamics of these communities are still poorly understood. In parallel to global approaches, such as metagenomic sequencing [1], several studies have approached this problem by exploring individual naturally occurring or artificially engineered interactions between pairs of organisms [2; 3]. Yet, no quantitative, systematic computational method exists for predicting or helping engineer interactions between any two microbes.

Here we propose a novel approach, based on stoichiometric models of metabolism, to identify artificial environments that induce mutualistic interactions between two given microbial species. Rather than modifying the microbes themselves, we aim at modifying the environment, and finding media compositions that sustain growth of two species only when simultaneously present. Our strategy is based on two major steps: *First*, we implement a procedure for automatically joining together the stoichiometric models for two species, embedding them into a common environment [4]. *Second*, we search the space of possible nutrient combinations for media that could not sustain growth of each species alone, but allow growth of both species simultaneously.

We applied our approach to three organism pairs of increasing complexity. The first is a simple toy model, in which one can arbitrarily pre-define expected mutualistic interactions. The second is a special case of the naturally occurring interactions between methanogenic archaea and hydrogen-producing microorganisms, which was recently analyzed in detail using flux balance models [2]. The third is an experimentally engineered synthetic biological system of two yeast strains that can grow only in the presence of each other, because each of them is unable to synthesize a specific essential metabolite.

In addition to recapitulating these known interactions, we use our approach to generate new experimentally testable predictions of environments that force interactions between other pairs of microorganisms. We envisage that these algorithms will make it possible to engineer novel metabolism-based interactions between pairs of organisms, paving the way for a new computationally-driven synthetic ecology discipline.

[1] Dinsdale EA *et al.* Functional metagenomic profiling of nine biomes. *Nature*. 2008 452(7187): 629-32.

[2] Stolyar S *et al.* Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol*. 2007 ;392.

[3] Shou W *et al.* Synthetic cooperation in engineered yeast populations. *Proc Natl Acad Sci U S A*. 2007 104(6): 1877-82.

[4] Klitgord N, Segrè D. The importance of compartmentalization in metabolic flux models: Yeast as an ecosystem of organelles. *Genome Informatics 2009*; Vol. 22 (in press)

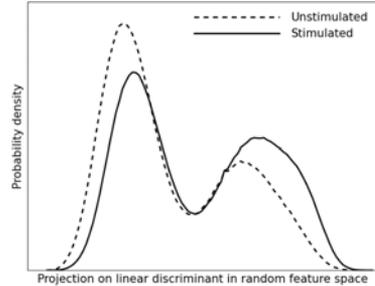
# Large-scale learning of cellular phenotypes from images

Vebjorn Ljosa<sup>1</sup>, Piyush B. Gupta<sup>1,2</sup>, Thouis R. Jones<sup>1</sup>, Eric S. Lander<sup>1,4</sup>, Anne E. Carpenter<sup>1</sup>

<sup>1</sup>Broad Institute of MIT and Harvard; <sup>2</sup>Dept. of Biology, MIT; <sup>3</sup>Whitehead Institute for Biomedical Research; <sup>4</sup>Dept. of Systems Biology, Harvard Medical School

Microscopy-based high-throughput screens can provide a broad view of biological responses and states at the resolution of single cells. Thousands of samples of cultured cells are perturbed by different chemicals or RNAi reagents. The samples are then stained and imaged, and samples that exhibit a phenotype of interest are chosen for further investigation.

Some phenotypes are readily identifiable in captured image data; for instance, mitotic arrest can be detected by measuring the intensity of a fluorescent marker for mitosis. Other phenotypes, while apparent upon visual inspection, are much harder to identify computationally. As an example, when signaling pathways related to cell migration are stimulated, T47D breast cancer cells take on a motile appearance, but this phenotype is not easily captured in a sparse set of measurements. Classifiers trained on hand-curated training sets can identify such phenotypes [1, 2]. We present a method that can learn to recognize



**Figure 1: Histogram of per-cell classifier scores for the unstimulated and stimulated replicates, showing the slight shift of cells from a nonmigratory to a migratory phenotype upon stimulation. This histogram is the basis for our nonparametric scoring method.**

phenotypes without requiring hand-labeled cells for training. Instead, a classifier is learned from larger portions of the experiment known to be enriched (if only slightly) by the phenotype of interest. As an example, we use an RNAi screen of T47D breast cancer cells [2]. The screen was performed in duplicate, and the second replicate was treated with a protein stimulant of cellular migration. As a result, a migratory phenotype putatively related to metastasis was slightly more prevalent in the stimulated replicate (~55% vs. ~45%). Such noisy training sets are unsuitable for most machine learning methods, but large-scale machine learning [3] allows us to overcome the noise by using huge training sets (in our case, the millions of cells found in each replicate). As a result, a classifier specific for the response of cells to the stimulant can be created without manual classification of cells (Figure 1). Our goal is not to classify individual cells, but to decide whether each *sample* is enriched for a phenotype. Because the number of cells per sample varies greatly, computing the fraction of motile-looking cells is insufficient to estimate the underlying probability of motility. We therefore use the empirical distribution of classifier scores (Figure 1) to give each cell a probabilistic (i.e., soft) label. We can then compute probability density functions of the proportion of motile-looking cells per sample and derive enrichment scores.

[1] T.R. Jones et al. (2009) "Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning," *Proc. Acad. Sci. USA*, **106**:1826–1831.

[2] P.B. Gupta et al., "Identification of novel effectors of ErbB2/3-mediated cell migration with high-throughput image-based screening," submitted.

[3] A. Rahimi and B. Recht (2008) "Random features for large-scale kernel machines," *Advances in Information Processing Systems (NIPS)*, **20**:1177–1184.

# Automated Design of Assemblable, Modular, Synthetic Chromosomes

Sarah M. Richardson<sup>1,2</sup>, Brian S. Olson<sup>3</sup>, Jef D. Boeke<sup>1,4</sup>, Amarda Shehu<sup>3</sup>, Joel S. Bader<sup>1,5</sup>

<sup>1</sup>High Throughput Biology Center, Johns Hopkins University School of Medicine;

<sup>2</sup>McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine; <sup>3</sup>Department of Computer Science, George Mason University; <sup>4</sup>Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine;

<sup>5</sup>Department of Biomedical Engineering, Johns Hopkins University School of Medicine

The goal of the *Saccharomyces cerevisiae* v2.0 project is the complete synthesis of a re-designed genome for baker's yeast. The resulting organism will permit systematic studies of eukaryotic chromosome structure that have been impossible to explore with traditional gene-at-a-time experiments. The efficiency of chemical synthesis of DNA does not yet permit direct synthesis of an entire chromosome, although it is now feasible to synthesize multi-kilobase pieces of DNA that can be combined into larger molecules. Our strategy for assembly involves the careful placement of restriction enzyme recognition sites at 10 kilobase intervals along the chromosome; each segment bounded by a restriction enzyme can be assembled from oligos and then joined to its neighbors by a digestion-ligation reaction. To date, designing a chromosome-sized sequence that can be thus assembled from smaller pieces has been accomplished manually by biological experts in a laborious fashion, with the caveat that the complexity of the problem precludes the human generation of an error-free, let alone optimal, solution. However, an optimal design may be obtained by framing the question as a formal optimization problem, which can be solved computationally, saving time and money. Our algorithm uses an efficient and highly parallelizable combination of suffix-tree indexing, dynamic programming, and dead-end elimination to compute an optimal configuration of restriction enzyme recognition sites. Our objective function takes into account the price of the restriction enzymes, the modification of genes annotated as essential, and the evenly spaced placement of restriction enzymes.

A recent human design of 90 kilobases assisted with available computational tools required over 40 man-hours of work. The design of the entire yeast genome would require nearly 3 years of this dedicated expert. In contrast, our implementation takes 2.5 minutes for the same 90 kb region, roughly a 1000x speed-up, and would only take 5 to 6 hours for the entire genome. Furthermore, the algorithm produces output that is superior to all of our expert-generated results, allowing us to quickly create several plans of action for inspection and evaluation – and perhaps concurrent synthesis. Our algorithm is not restricted to yeast. It takes as input any annotated sequence and a list of restriction enzymes and provides as output a minimally modified sequence with all changes annotated and a list of the oligonucleotides needed to construct it.

# Reconstructing Ancestral Gene Content by Co-Evolution<sup>&</sup>

Tamir Tuller<sup>1\*</sup>, Hadas Birin<sup>2\*</sup>, Uri Gophna<sup>3</sup>, Martin Kupiec<sup>3</sup> and Eytan Ruppin<sup>2,4</sup>

<sup>1</sup>*Mathematics and Computer Science & Molecular Genetics; Weizmann Institute of Science;* <sup>2</sup>*Blavatnik School of Computer Science;* <sup>3</sup>*Department of Molecular Microbiology and Biotechnology;* <sup>4</sup>*School of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel.*  
*\*These authors contributed equally to this work. &This work is under minor revision in Genome Research.*

Inferring the gene content of ancestral genomes is a fundamental challenge in molecular evolution. Due to the statistical nature of this problem, ancestral genomes inferred by the Maximum likelihood (ML) or the Maximum-Parsimony (MP) methods are prone to considerable error rates. In general, these errors are difficult to abolish by using longer genomic sequences or by analyzing more taxa. This study describes a new approach for improving ancestral genome reconstruction, the Ancestral Co-Evolver (ACE), which utilizes co-evolutionary information to improve the accuracy of such reconstructions over previous approaches. The principal idea is to reduce the potentially large solution space by choosing a single optimal (or near optimal) solution that is in accord with the co-evolutionary relationships between protein families. Simulation experiments, both on artificial and real biological data, show that ACE yields a marked decrease in error rate compared to ML or MP. Applied to a large dataset (95 organisms, 4873 protein families and 10,000 co-evolutionary relationships), some of the ancestral genomes reconstructed by ACE were remarkably different in their gene content from those reconstructed by ML or MP alone (more than 10% in some nodes). These reconstructions, while having almost similar likelihood/parsimony scores as those obtained with ML/MP, had markedly higher concordance with the co-evolutionary information. Specifically, when ACE was implemented to improve the results of ML, it added a large number of proteins to those encoded by LUCA (Last Universal Common Ancestor), most of them ribosomal proteins and components of the F<sub>0</sub>F<sub>1</sub> type ATP synthase/ATPases, complexes that are vital in most living organisms. Our analysis suggests that LUCA appears to have been bacterial-like and had a genome size similar to the genome sizes of many extant organisms.

## Regulatory Genomics Poster Session 1: Wed 8:15pm-9:45pm

(posters available for viewing Wed 3pm-Thu 2pm)

Agius	Learning compact models of DNA binding specificities for transcription factors from protein binding microarrays	119
Äijö	Learning gene regulatory networks with delayed ODEs and continuous-time expression representation	146
Akavia	Conexic: A Bayesian framework to detect drivers and their function uncovers an endosomal signature in Melanoma	90
Amzallag	Comparison of gene expression time courses between light entrained and temperature entrained drosophila flies reveal genes which peak two times a day	147
Arunachalam	Computational discovery of <i>Cis</i> -regulatory elements in multiple <i>Drosophila</i> species	120
Behrens	Studying the evolution of promoters: a waiting time problem	121
Betel	Comprehensive modeling of microRNA targets: predicting functional non-conserved and non-canonical sites	91
Biggin	Evidence for Quantitative Transcription Networks	148
Bolotin	Integrated Approach for the Identification of Human HNF4 $\alpha$ Target Genes Using Protein Binding Microarrays	122
Bristow	Exploring the CBP developmental time-course in <i>Drosophila</i>	92
Candeias	Temporal Dynamics of Regulatory Networks in <i>Drosophila melanogaster</i> Embryogenesis	149
Carson	Investigating Co-regulation Networks Using Generative Models	150
Carvalho	Applications of Centroid Estimation to Regulatory Genomics	123
Chang	The intersect of mRNA, microRNA and protein dynamics upon down-regulation of Nanog in mouse embryonic stem cells	93
Chun	Reverse Engineering of Gene Regulation Network from DREAM4 Data	151
Clarke	A missing ingredient in the Pho4 paradigm? Evidence for Pho4/Cbf1 binding site competition	124
Clote	RNA Structural Segmentation	94
Cook	Using ChIP seq to search for sequence determinants of binding of the <i>S. cerevisiae</i> transcription factor Sko1	125
Dabrowski	Effects of motif and CNS multiplicity on gene expression in subspaces of conserved eigensystems following stroke and seizures	126
Davis	Classification Trees Can Describe and Predict Conditional Transcription Factor Binding in vivo	152
Degner	Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data	114
Dojer	Identification of <i>cis</i> -regulatory modules in homologous sequences	153
Feng	Genome-wide survey of <i>D. melanogaster</i> insulator proteins binding preferences in genes	95
Frogner	Learning recurrent mRNA expression patterns from	154

	systematic analysis of <i>in-situ</i> images of Drosophila embryos	
Gitter	Backup in gene regulatory networks explains differences between binding and knockout results	127
Goff	Genomic relationship between small RNAs and histone modifications revealed by next-generation sequencing of human embryonic stem cells.	96
Greenfield	Inferring Topology and Dynamical Properties of Genome-wide Regulatory Networks	155
Gyenesei	Functional Inference from a Genome-Wide <i>in situ</i> Hybridization Atlas of the Mouse Embryo	156
Habib	Aromatase inhibition in a transcriptional network context	157
Hemberg	De novo detection of transcribed regions in mouse based on RNA-Seq	97
Huggins	Design of multiple hypothesis tests for microarray data	172
Jabbari	Novel thermodynamics-based algorithm for probe-specific position-dependent hybridization free energy	128
Jacobsen	Genes up-regulated after microRNA perturbation have significant over-representation of the ARE motif UAUUUAU in the 3'UTR	98
Ji	Genes as molecular machines: Microarray Evidence for structural genes regulating their own transcripts	99
Jungreis	A Computational Investigation of Widespread Stop Codon Readthrough in Drosophila	129
Kadri	Evolutionary role of microRNAs in developmental gene regulatory networks	100
Karlic	Towards a Histone Code for Transcription	101
Kartal	Ground State Robustness as an Evolutionary Design Principle in Signaling Networks	158
Konieczka	Evolution of the High Osmolarity Glycerol (HOG) stress response network across Ascomycota fungi	159
Kumar	The Msx1 Homeoprotein Recruits Histone Methyltransferase Activities to Control the Expression of Target Genes in the Developing Limb	102
Kural	Identification of Noncoding Motifs Under Selection in Coding Sequences	130
Laurila	Protein-protein interactions improve multiple transcription factor binding site prediction	131
Le	Distance functions for querying large, multi species, expression databases	115
Lee	Successful Enhancer Prediction from DNA Sequence	132
Li	Identifying motifs using GADEM with a starting	133
Lin	Modeling Idiopathic Pulmonary Fibrosis Disease Progression based on Gene and Protein Expression	116
Liu	Prediction of Polycomb target genes in mouse embryonic stem cells	178
Lorenz	Rapid Estimation of RNA Kinetics	103
Lu	A novel method to simulate genome-wide background noise and distinguish real binding sites from background noise	134
Maas	The RNA Editing Dataflow System (REDS) for the transcriptome-wide discovery of RNA modification sites	104

**Regulatory Genomics Poster Session 2: Thu 3:45pm-5:15pm**  
(posters available for viewing Thu 3pm-Fri 2pm)

MacIsaac	Novel methods for the discovery of condition specific master regulators of transcription	160
MacKenzie	Life After Comparative Genomics; Regulatory Systems, Homeostasis, Synergy, SNPs and Disease.	173
Mahony	Chromatin state dynamics and the acquisition of rostrocaudal positional identity during retinoidinduced neurogenesis	105
Majoros	Modeling the Evolution of Regulatory Elements by Simultaneous Detection and Alignment with PhyloPairHMMs	161
Marchal	De novo detection and qualification of regulatory motifs.	135
Martins	REGULATORY ELEMENT IDENTIFICATION WITH FUNCTIONAL GENOMIC COVARIATES	162
McGettigan	Identification of epigenetic changes in the brain of a rat model of schizophrenia using FAIRE-seq	106
Miller	microRNAs preferentially target dosage-sensitive genes	107
Missiuro	Predicting Genetic Interactions in <i>C. elegans</i> using Machine Learning	117
Morris	Using accessibility to predict RNA-binding protein targets	108
Nielsen	CATCHprofiles reveals nucleosome positioning of histone modifications	109
Novichkov	The automatic selection of TFBS score threshold in comparative genomics approach.	136
Piipari	Inference and validation of a large cis regulatory motif set using whole-genome <i>Saccharomyces</i> resequencing data	137
Polak	Large differences in transcription associated strand asymmetries of substitution patterns across metazoans	138
Pollard	Unraveling of an ancient regulatory pathway: RNAi insensitivity in the germline of <i>C. elegans</i>	163
Rautajoki	ESTOOLSDB – a comprehensive database for stem cell research	164
Ray	Discriminating functionality by kernel clustering k-mers in the regulatory genome	165
Regan	NF- $\kappa$ B and Forkhead – partners and opponents	175
Rieder	Spatial association of multiple coordinately expressed but functionally unrelated genes during cell differentiation	166
Robine	piRNA production in a <i>Drosophila</i> ovary cell line	110
Rossetti	Epigenetic silencing of a tumor suppressor network unmasks the dual face of master cell signals	111
Russo	Global Entrainment of Transcriptional Systems to Periodic Inputs	167
Sealfon	Supervised learning approaches to predicting enhancer regions and transcription factor binding sites in <i>D. melanogaster</i>	139
Skupsky	Integration site of the HIV promoter primarily modulates transcriptional burst size, rather than frequency	176
Srinivasan	Large-scale comparative analysis of RNA structures by TOPOFIT	112
Storms	The Effect of Orthology and Coregulation on Detecting	140

Regulatory Motifs		
Sugathan	Global DNase Hypersensitivity Mapping Reveals Growth Hormone (GH)-regulated Sex Differences in Mouse Liver Chromatin Structure	141
Taher	Function conservation in diverged noncoding elements	142
Tsai	The integrated pathway and proteomic resources for identifying the potential colorectal cancer biomarkers	118
Upadhyay	PPi module for visualization and analysis of protein-protein interfaces in Friend.	168
Vallania	SPLINTER: detection of rare regulatory variants using a large deviation theory approach	177
Vermeirssen	Composite network motifs in integrated metazoan gene regulatory networks	169
Waldman	TP53 cancerous mutations exhibit selection for translation efficiency	174
Wohlbach	Identification of genomic features novel to xylose-fermenting yeasts through comparative analyses of <i>Pichia stipitis</i> , <i>Candida tenuis</i> , and <i>Spathaspora passalidarum</i>	170
Won	Genome-wide prediction of transcription factor binding sites using chromatin modification	143
Xin	Epigenetic Profiling of Human Brain Development	113
Yan	Structural and Regulatory Evolution of Electrophysiological Systems	171
Zaslaver	Metazoan operons accelerate transcription and recovery rates	144
Zhou	Determinants of Transcription Factor Binding and Regulation	145

# Conexic: A Bayesian framework to detect drivers and their function uncovers an endosomal signature in Melanoma

Uri David Akavia<sup>1,2</sup>, Oren Litvin<sup>1,2</sup>, Eyal Moses<sup>1</sup>, Yossi Tzur<sup>1</sup>, Dylan Kotliar<sup>1</sup>, Jessica Kim<sup>3,4</sup>, Patrick Oberholzer<sup>3,4</sup>, Levi A. Garraway<sup>3,4</sup>, Dana Pe'er<sup>1,2</sup>

<sup>1</sup>Department of Biological Sciences, Columbia University; <sup>2</sup>Center for Computational Biology and Bioinformatics, Columbia University; <sup>3</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School; <sup>4</sup>Broad Institute of Harvard and MIT

Genomics is revolutionizing our understanding of cancer biology. Tumor samples assayed for comprehensive chromosomal and gene expression data are accumulating at a staggering rate. A major challenge involves the development of analysis methods to uncover biological insights from these data, including the identification of the key mutations that drive cancer and how these events alter cellular regulation.

We have developed Conexic, a novel Bayesian Network-based framework to integrate chromosomal copy number and gene expression data to detect genetic alterations in tumors that drive proliferation, and to model how these alterations perturb normal cell growth/survival. The underlying assumption to our approach is that significantly recurring copy number change, coinciding with its ability to predict the expression patterns varying across tumors, strengthens the evidence of a gene's causative role in cancer. This method not only pinpoints specific regulators within an a large region of copy number variation, but can identify the effected targets and shed light on the way in which gene regulation is altered

We applied our Conexic framework to a melanoma dataset (Lin et al, Cancer Research, 2007) comprising 62 paired measurements of gene expression and copy number. In addition to confirming the role of known drivers in melanoma, our analysis suggests a number of novel drivers. Most strikingly, these point to a major role of protein trafficking and endosome biology in regulating this malignancy. These results have linked endosomal processing and sorting to adhesion and survival. Preliminary experimental validation supports several of these findings. Together, these results affirm the potential of Conexic to elaborate novel driver modules with biological and possibly therapeutic importance in melanoma and other cancers.

RG Posters

# Comprehensive modeling of microRNA targets: predicting functional non-conserved and non-canonical sites

Doron Betel<sup>1</sup>, Anjali Koppal<sup>2</sup>, Phaedra Agius<sup>1</sup>, Chris Sander<sup>1</sup>, Christina Leslie<sup>1</sup>

<sup>1</sup>*Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York;* <sup>2</sup>*Department of Computer Science, Columbia University, New York*

Accurate prediction of microRNA targets is a challenging computational problem, impeded by incomplete biological knowledge and the scarcity of experimentally validated targets. The primary determinant for regulation, near-perfect base pairing in the seed region of the microRNA (positions 2-7), gives poor specificity as a prediction rule. In an effort to reduce false predictions, most computational methods restrict to perfect seed matches that are evolutionary conserved, despite experimental evidence that neither constraint holds in general. Here we present a new algorithm called mirSVR for predicting and ranking the efficiency of microRNA target sites by using supervised learning on mRNA expression changes from microRNA transfection experiments. We use support vector regression (SVR) to train on features of the predicted miRNA:mRNA duplexes as well as contextual features without restricting to perfect seed complementarity or filtering by conservation. In a large-scale evaluation on independent transfection and inhibition experiments, mirSVR significantly outperformed existing target prediction methods for predicting genes that are deregulated at the mRNA or protein levels.

mirSVR effectively broadens target prediction beyond the standard restrictions of perfect seeds and strict conservation without introducing a large number of spurious predictions. In particular, we found that mirSVR correctly identified functional but poorly conserved target sites, and that imposing a conservation filter resulted in a reduced rate of detection of true targets. mirSVR scores are calibrated to correlate linearly with the extent of downregulation and therefore enable accurate scoring of genes with multiple target sites by addition of individual site scores. Furthermore, the scores can be converted to an empirical probability of downregulation, which provides a meaningful guide for selecting a score cutoff. The model successfully predicted genes that are regulated by multiple endogenous microRNAs – rather than transfected microRNAs whose concentrations are above physiological levels – when analyzing targets bound to human Argonaute (AGO) proteins as identified by AGO immunoprecipitation. Finally, we tested the usefulness of including non-canonical sites in the model by evaluating performance on biochemically determined sites from recent PURE-CLIP experiments, 20% of which do not contain any perfect microRNA seed match. We found that mirSVR indeed correctly detected a significant number of these experimentally verified non-canonical sites.

# Learning expression primitives from systematic analysis of *in-situ* gene expression images of *Drosophila* embryos

Charlie Frogner<sup>1</sup>, Christopher A. Bristow<sup>2,3</sup>, Tom Morgan<sup>2</sup>, Lorenzo Rosasco<sup>1</sup>, Pouya Kheradpour<sup>2</sup>, Rachel Sealfon<sup>2</sup>, Tomaso Poggio<sup>1</sup>, Manolis Kellis<sup>2,3</sup>

<sup>1</sup>Center for Biological and Computational Learning, McGovern Institute, MIT, Cambridge, MA; <sup>2</sup>Department of Computer Science and Artificial Intelligence, MIT, Cambridge, MA; <sup>3</sup>Broad Institute, Cambridge, MA

Understanding the spatio-temporal control of gene expression during development is one of the major challenges in genomics. In order to dissect the regulatory constructs responsible for defining gene expression programs, we are combining image analysis of gene expression patterns in *Drosophila* embryos with sequence analysis of genomic regulatory regions, and functional genomic data.

Our initial dataset consists of ~75,000 images of *Drosophila* embryos from ~6,000 genes profiled at 5 stage ranges of embryogenesis. While previous research on *Drosophila* development has relied on human curation of these expression patterns, we have sought to apply computer vision approaches to analyze these data. We have developed an initial image processing pipeline to segment the embryo images and extract the pixels that correspond to mRNA staining, and coupled these to novel approaches for embryo segmentation that reliably deal with common issues, such as multiple overlapping embryos. We have then developed a new method for extracting detailed stain patterns from the images, based on a supervised learning approach trained on ubiquitously expressed genes. Together, these methods allow us to process over 90% of images, and can automatically extract mRNA expression patterns for many thousands of embryos, with minimal human inspection.

We used these algorithms to extract a robust representation of recurrent expression patterns in a systematic and unbiased way. We clustered the extracted mRNA stain patterns based on a spatial similarity metric to assemble groups of genes that show coherent expression patterns. Strikingly, these recover specific organ systems independently annotated by human curators, and in some cases suggest meaningful subdivisions of annotation terms, as well as new patterns that are not easily captured using existing annotation terms. The inferred clusters also show specific enrichments in known regulatory motifs associated with transcription factors involved in embryo development, and suggest specific regulatory connections to candidate regulators for these recurrent patterns.

While our initial results for early embryogenesis are very encouraging, there are several challenges going forward as we extend this approach to multiple time points and stages with more complex gene expression patterns. Overall, systematic image analysis of large-scale gene expression datasets, coupled with genome sequence analysis and large-scale functional datasets provides a general way to define common regulatory programs in animal genomes, by discovering genes that have coherent upstream regulation (e.g. transcription factor binding, chromatin marks, and motif instances) and coherent downstream expression patterns (based on *in-situ* image analysis). The approach presented here is scalable and robust, and should apply more generally to any species.

# The intersect of mRNA, microRNA and protein dynamics upon down-regulation of Nanog in mouse embryonic stem cells

Betty Chang<sup>1-3</sup>, David Braun<sup>3,5</sup>, Nektarios Paisios<sup>5</sup>, Yun Lu<sup>5</sup>, Ravi Sachidanandam<sup>4</sup>, Ihor R. Lemischka<sup>1-3</sup>

<sup>1</sup>Department of Gene and Cell Medicine; <sup>2</sup>Black Family Stem Cell Institute; <sup>3</sup>Graduate School of Biological Sciences; <sup>4</sup>Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029; <sup>5</sup>Department of Computer Science, Courant Institute of Mathematical Sciences, Graduate School of Arts and Sciences, New York University, New York, NY 10012

Nanog, one of the core transcription factors defining embryonic stem (ES) cells, when disrupted causes impaired self-renewal and cellular differentiation. The loss of Nanog results in dramatic changes in mRNA and protein expression. Chromatin immunoprecipitation studies have localized Nanog to thousands of genomic loci including the transcription start sites of over 70 microRNAs.

Here we monitor the dynamic changes in mRNA and nuclear proteins at 1, 3 and 5 days after depletion of Nanog expression by shRNA in mouse ES cells. This results in different classes of correlation between mRNA and protein expression levels. The first where mRNA and protein expression levels for the same gene change in parallel, either positively or negatively, the second where the mRNA and protein levels diverge from one another. In the latter class, where we see mRNA levels increasing or remaining stable and the analogous protein levels decrease, we attribute some of these discordances to microRNA (miRNA) activity. To examine these discordant relationships, we predict a set of miRNAs that may target these mRNA preventing protein translation or causing transcript degradation using available databases TargetScan, miRANDA, EIMMO and PITA. Of the 72 miRNA TSS bound by Nanog, 52 miRNAs are among our set predicted to play a role in regulating our discordant mRNA-protein pairs.

To further our understanding of the role of microRNAs, we perform deep sequencing of small RNAs of 18-30nts in length from mouse embryonic stem cells over our time course of Nanog depletion to identify dynamic changes in miRNA expression. We integrate these data together with mRNA and protein expression towards a comprehensive analysis of Nanog-centric regulation in the mouse ES cell.

## RNA Structural Segmentation

I. Dotu<sup>1,2</sup>, W.A. Lorenz<sup>1</sup>, P. Van Hentenryck<sup>2,3</sup>, P. Clote<sup>1</sup>

<sup>1</sup>Department of Biology, Boston College, Chestnut Hill, MA 02467; <sup>2</sup>Dynadec, One Richmond Square, Providence, RI 02906; <sup>3</sup>Department of Computer Science, Brown University, Box 1910, Providence, RI 02912.

We describe several dynamic programming segmentation algorithms to segment RNA secondary and tertiary structures into distinct domains. For this purpose, we consider fitness functions that variously depend on (i) base pairing probabilities in the Boltzmann low energy ensemble of structures, (ii) contact maps inferred from 3-dimensional structures, and (iii) Voronoi tessellation computed from 3-dimensional structures. Segmentation algorithms include a direct dynamic programming method, previously discovered by Bellman and by Finkelstein and Roytberg, as well as two novel algorithms – a parametric algorithm to compute the optimal segmentation into  $k$  classes, for each value  $k$ , and an algorithm that simultaneously computes the optimal segmentation of all subsegments.

Since many non-coding RNA gene finders scan the genome by a moving window method, reporting high-scoring windows, we apply structural segmentation to determine the most likely 5' and 3' boundaries of precursor microRNAs. When tested on all precursor microRNAs of length at most 100 nt from the Rfam database, benchmarking studies indicate that segmentation determines the 5' boundary with discrepancy (absolute value of difference between predicted and real boundaries) having mean  $-0.640$  (stdev 15.196) and the 3' boundary with discrepancy having mean  $-0.266$  (stdev 17.415). This yields a sensitivity of 0.911 and positive predictive value of 0.906 for determination of exact boundaries of precursor microRNAs within a window of approximately 900 nt. Additionally, by comparing the manual segmentation of Jaeger et al. with our optimal structural segmentation of 16S and 16S-like rRNA of *E. coli*, rat mitochondria, *Halobacterium volcanii*, and *Chlamydomonas reinhardtii* chloroplast into 4 segments, we establish the usefulness of (automated) structural segmentation in decomposing large RNA structures into distinct domains.

The journal version of this paper has been accepted for publication in Pacific Symposium on Biocomputing 2010.

## Genome-wide survey of *D.melanogaster* insulator proteins binding preferences in genes

Xin Feng<sup>1,2</sup>, Lincoln Stein<sup>2,3</sup>

<sup>1</sup>Department of biomedical engineering, Stony Brook University, NY, 11794, USA; <sup>2</sup>Cold Spring Harbor Lab, NY, 11724, USA; <sup>3</sup>Ontario Institute of Cancer Research, Toronto, Ontario, Canada M5G 0A3

Insulators are genomic elements that at the genetic level block the interaction between enhancers and promoters, and at the molecular level inhibit heterochromatin spread. The activity of insulator sites are mediated by one or more insulator binding proteins (IBPs). *Drosophila* has several such IBPs. Members of the modENCODE project have recently characterized the pattern of binding in *D.melanogaster* for six distinct IBPs. We have used these data to perform a genome-wide analysis of IBP binding sites with respect to protein-coding genes. We find that the IBPs CP190, GAF, BEAF and CTCF are enriched at the 5' ends of genes relative to the 3' ends. Binding is greatest immediately upstream of the TSS, consistent with their putative role in the promoter complex. In contrast, Su(Hw) demonstrates a uniform distribution across the gene. Much to our surprise, the bodies of genes are also enriched in IBP binding, and this binding is heavily biased towards the 5' side of exon/intron boundaries. This suggests that IBPs may be involved in the regulation of splicing as well. These results together provide more detailed insights into the IBP-gene relationships and support the roles of IBPs as chromatin structure organizers.

## Genomic relationship between small RNAs and histone modifications revealed by next-generation sequencing of human embryonic stem cells.

Loyal A. Goff<sup>1,2</sup>, Khalil, Ahmad<sup>2,3</sup>, Loewer, Sabine<sup>4,5</sup>, Swerdel, Mavis<sup>6</sup>, Hart, Ronald P.<sup>6</sup>, Rinn, John L.<sup>2,3</sup>, Kellis, Manolis<sup>1,2</sup>

<sup>1</sup>Computer Science & Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA; <sup>2</sup>Broad Institute, Cambridge, MA, USA; <sup>3</sup>Dept. of Pathology, Harvard Medical School, Boston, MA, USA; <sup>4</sup>Children's Hospital, Boston, MA, USA; <sup>5</sup>Dept. Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA; <sup>6</sup>Rutgers Stem Cell Research Center, Rutgers University, Piscataway, NJ, USA.

New sequencing technologies have enabled probing of the cellular state in ever increasing detail, revealing many new families of small RNAs and many new members of existing families, such as microRNAs, piRNAs, and rasiRNAs. However, the function of a large fraction of small RNAs found in such sequencing experiments remains unknown, as they do not fit in any of these functional categories. To understand potential new functions in chromatin regulation, we have used SOLiD sequencing of small RNAs (18-28 nucleotides) in several human ES cell lines in various stages of differentiation, and compared these to ChIP-Seq data for previously published hES histone modifications. The comparison revealed a striking relationship between the two.

We found that small RNAs with previously unknown functions align to the center of peaks for histone modifications, suggesting that endogenous small RNAs may nucleate epigenetic changes. Moreover, the association showed distinct relationships with separate marks, and was significantly reduced when comparing smRNAs from differentiated hES cells with histone modifications of undifferentiated cells, suggesting these relationships are likely to be cell-type specific and modification specific. While several recent studies have suggested a functional relationship between histone modifications and the RNAi pathway, this is the first genome-wide evidence of a likely global role for small RNAs in the establishment and maintenance of histone modification states. If smRNAs do in fact direct epigenetic modifications, this association may provide new understanding of epigenetic changes during differentiation, and possibly new therapeutic means for diseases associated with chromatin dysregulation such as cancer.

# De novo detection of transcribed regions in mouse based on RNA-Seq

Martin Hemberg<sup>1</sup>, Jesse M Gray<sup>2</sup>, Tae-Kyung Kim<sup>2</sup>, Michael E Greenberg<sup>2</sup>, Gabriel Kreiman<sup>1,3</sup>

<sup>1</sup>Department of Ophthalmology, Children's Hospital Boston; <sup>2</sup>Department of Neurobiology, Harvard Medical School; <sup>3</sup>Schwartz Center for Theoretical Neuroscience, Harvard University

By combining molecular biology experiments and computational methods researchers produced an extensive list of protein coding genes in multiple organisms. Technological advances, such as high-throughput sequencing, have made it possible to study transcription ('RNA-Seq') in an unbiased manner throughout the genome. Based on RNA-Seq and tiling microarray studies it has been suggested that most of the non-repetitive genome is transcribed. An RNA-Seq experiment produces millions of short (~35 bps) reads and by mapping them to the reference genome, one can determine the chromosomal locations of the reads. One of the drawback of this technique is that the RNA is fragmented and hence we have no knowledge of the length of individual transcripts found in the cell.

We present an algorithm based on Haar-wavelets for *de novo* transcript calling (HaTriC) in large mammalian genomes based on RNA-Seq data. The algorithm uses a multi-scale approach for detecting regions with sharp changes in read-densities. These break points determine a partitioning of the genome into segments of low or high read-densities. The distribution of the segment read-density is bimodal with one mode corresponding to high read densities and the other to background. The segments that belong to the mode with high read densities are labeled transcribed regions.

To test our algorithm, we apply it to a dataset collected from mouse cortical neurons. We employ a supervised learning approach and tune the parameters of the HaTriC algorithm by maximizing the number of transcribed regions that have a 1:1 correspondence to annotated genes for one of the chromosomes in the mouse genome. We show that the algorithm is robust when the optimal parameters are perturbed. Furthermore, we find that HaTriC is inclusive – the transcribed regions account for >90% of all reads. Using the optimized parameter setting, we apply HaTriC to the rest of the genome and we show that >60% of the genes with a density of at least 1 read/100 bps can be uniquely identified.

Since the HaTriC algorithm identifies transcripts regardless of their location with respect to annotated genes, it will also detect thousands of transcribed regions that are either non-coding or correspond to unannotated genes. We investigate these regions and we show that many of them correspond to classes of recently discovered non-coding RNAs such as divergent transcripts and long non-coding RNAs (lncRNAs). Using the available annotation of the mouse genome, we find that ~90% of the detected transcribed regions belong to known categories. Our novel algorithm provides a procedure for quantitative, rapid and automatic annotation of transcripts derived from the massive sequence data obtained in RNA-seq experiments.

# Genes up-regulated after microRNA perturbation have significant over-representation of the ARE motif UAUUUAU in the 3'UTR

Anders Jacobsen<sup>1,2</sup>, Jean Wen<sup>1</sup>, Debora S. Marks<sup>3</sup>, Anders Krogh<sup>1</sup>

<sup>1</sup> The Bioinformatics Centre, Department of biology, University of Copenhagen, Copenhagen, Denmark; <sup>2</sup> Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA; <sup>3</sup> Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA

MicroRNAs destabilize mRNAs by imperfect base-pairing in metazoan 3'UTRs. Perturbing a cell by over-expressing or inhibiting a miRNA is a commonly used strategy to identify direct regulatory targets of a miRNA. However, the direct miRNA targets only constitute a fraction of all expression changes following a miRNA perturbation. We investigated whether 3'UTR sequences *other than mi-croRNA binding sites* correlated with changes in expression after small RNA perturbations.

To investigate this exhaustively we analyze expression profiles from 18 perturbations of 15 different miRNAs in 4 different cell lines. Using a non-parametric correlation statistic, we analyzed all sequences (words) of length 5-7 nucleotides for over-representation in 3'UTRs of up or down-regulated genes following miRNA over-expression or inhibition respectively. Word enrichment was averaged over all experiments to produce a list of words that were enriched in up or down-regulated genes independently of the sequence of the perturbed miRNA.

The most significantly enriched sequence in up-regulated genes was the heptanucleotide UAUUUAU – a core motif of AU-rich elements (AREs) which affect mRNA turnover. Our results suggest a novel association between miRNA and ARE-mediated mRNA regulation.

RG Posters

# Genes as molecular machines: Microarray Evidence for structural genes regulating their own transcripts

Sungchul Ji, Andrew Davidson, and Julie Bianchini

Department of Pharmacology and Toxicology, Rutgers University, Piscataway, NJ 08855

It is generally believed that transcript levels inside the cell are controlled by regulatory genes, some of which encode proteins that participate in transcription and/or transcript degradation processes. We here report DNA microarray evidence that structural genes also play a significant role in regulating their own transcript levels, most likely acting as *molecular machines*. The microarray data utilized in the present analysis were measured by Garcia-Martinez et al. (*Mol. Cell* 15:303-313 (2004)) at 6 time points (0, 5, 120, 360, 450 & 850 minutes) after switching glucose to galactose. Each one of the ~6000 intracellular RNA trajectories carries two kinds of information – i) the name of the gene encoding the RNA under consideration, and ii) the time-dependent intracellular concentration of the RNA. The former can be described in an N-dimensional sequence (or *genotype*) space, where a point represents an N nucleotide-long RNA molecule, and the latter in the 6-dimensional concentration (or *phenotype*) space, wherein a point represents the kinetic trajectory of an RNA molecule measured over the 6 time points. Thus, for any pair of RNA molecules, it is possible to calculate i) the *genotypic similarity* as the degree of the overlap between the pair of nucleotide sequences (using the CrustalW2 program on line), and ii) the *phenotypic distance* as the Euclidean distance between the corresponding two points in the 6-dimensional concentration space. When the phenotypic distances of a set of all possible RNA pairs (numbering  $(n^2 - n)/2$ , where n is the number of RNA molecules belonging to a given metabolic group) were plotted against the associated genotypic similarities, most points were found to lie below a straight line with a negative slope whose magnitude depended on the functions of RNA molecules (e.g., -24.1 and -2.6 for the glycolytic and oxidative phosphorylative RNA molecules, respectively). A greater absolute slope in the *phenotypic distance vs. genotypic similarity* (PDGS) plot indicates greater variations in phenotypes for a given degree of genotypic similarity and hence a smaller effect of structural genes on the intracellular concentrations of their transcripts. The absolute inverse of the slope of the PDGS plots therefore can be utilized as a quantitative measure of the *self-regulatory power of structural genes* (SRPSG):

$$\text{SRPSG} = (|\text{Slope of PDGS plot}|)^{-1}.$$

For a given group of RNA molecules with a common function (e.g., glycolysis), two PDGS plots can be generated by dividing the RNA trajectories into two phases – the early phase from 0 to 120 minutes and the late phase from 360 to 850 minutes after the glucose-galactose shift. Out of the 10 functional groups examined so far, 6 groups showed a steeper slope in the early phase compared to the late phase in the PDGS plots, indicating that the *self-regulatory power of structural genes* was lower during the energy-poor early phase as compared to the energy-replenished late phase. These observations support the notion that metabolic regulation requires both *genetic information* and *free energy* carried by *conformons*, sequence-specific *conformational strains* embedded in and generating mechanical forces on biopolymers acting as molecular machines, including *structural genes*. Single-molecule experiments with the actomyosin system have demonstrated that *conformons* are necessary and sufficient to cause goal-directed molecular motions (S. Ji, “Free energy and information contents of *Conformons* in proteins and DNA”, *BioSystems* 54:107-130 (2000); A. Ishijima et al, “Simultaneous measurement of chemical and mechanical reaction”, *Cell* 70: 161-171 (1998); Y. Ishii and T. Yanagida, “How single molecule detection measures the dynamics of life, *HFSP Journal* 1(1):15-29 (2007)).

# Evolutionary role of microRNAs in developmental gene regulatory networks

Sabah Kadri<sup>1,2</sup>, Veronica Hinman<sup>1,3</sup>, Panagiotis V. Benos<sup>4</sup>

<sup>1</sup>Lane Center for Computational Biology, Carnegie Mellon University; <sup>2</sup>Joint CMU-Pitt PhD Program in Computational Biology, Carnegie Mellon University; <sup>3</sup>Department of Biological Sciences, Carnegie Mellon University; <sup>4</sup>Department of Computational Biology, University of Pittsburgh

MicroRNAs (miRNAs) are ~22 nucleotide long non-coding single-stranded RNA molecules that play very important roles in post-transcriptional regulation of genes in diverse taxa, from plants to vertebrates and invertebrates. We are interested in the role of miRNA genes in the evolution of development. Towards that goal, we have studied miRNA expression during development of sea urchin and sea star embryos, using a combination of computational and experimental approaches.

The echinoderms, sea urchin (*Cl. echinoidea*) and sea star (*Cl. asteroidea*) are excellent model organisms as a great deal is known in these echinoderms of how transcription factor interaction networks have evolved and their effects in development. Although the adult morphologies are strikingly different in these two species, many aspects of their early development are highly conserved. The sea urchin, *S.purpuratus* is the only echinoderm with a sequenced genome. miRBase [4] contains miRNAs from the adult sea urchin, but nothing is known about the role of miRNAs in the early developmental stages of these organisms.

Initially, we used simple homology-based methods to identify conserved miRNAs in the sea urchin genome. HHMMiR [1], a probabilistic model that predicts miRNA genes without evolutionary constraint was used to predict novel miRNA-containing hairpins in the sea urchin transcriptome data. In order to obtain a set of miRNAs in certain developmental stages, Solexa sequencing was carried out on embryonic samples from both organisms. Initial analysis of this data found 34 families of highly conserved miRNAs and 5 echinoderm-specific miRNAs. Reads were mapped to the sea urchin genome and HHMMiR was used to predict potential novel miRNAs from the extracted hairpins.

A small subset of miRNAs (including the highly conserved *miR-10* and *miR-31*) was selected for further experimental validation and miR-target interactions. We used miRanda [2] and RNAhybrid [3] to make target predictions in the sea urchin. Whole mount in situ hybridizations were used to study spatial expression patterns in developing embryos. Due to the well-characterized gene networks in these organisms, this information was useful to enhance the target recognition pipeline.

[1] Kadri,S et al. 2009 BMC Bioinformatics. 2009 Jan 30;10 Suppl 1:S35.

[2] Stark, A et al. 2003 PLoS Biol 1, E60

[3] Rehmsmeier, M et al. 2004 RNA 10, 1507-1517

[4] Griffiths-Jones S., 2006 Nucleic Acids Res 34(Database issue):D140–D144

## Towards a Histone Code for Transcription

Rosa Karlic<sup>1,2</sup>, Ho-Ryun Chung<sup>1</sup>, Julia Lasserre<sup>1</sup>, Kristian Vlahovicek<sup>2</sup>, Martin Vingron<sup>1</sup>

<sup>1</sup>Max-Planck-Institut für molekulare Genetik, Department of Computational Molecular Biology, Ihnestraße 73, 14195 Berlin (Germany); <sup>2</sup>Bioinformatics Group, Division of Biology, Faculty of Science, Zagreb University, Horvatovac 102a, 10000 Zagreb (Croatia);

Histones are frequently decorated with covalent modifications, which are tightly linked to gene regulation. These modifications are thought to constitute a “Histone Code”, which is read out by proteins to bring about specific downstream effects, supported by the finding that individual modifications can be associated with transcriptional activation or repression. However, in general very little is known about the relationship between histone modifications and the transcriptional process.

Using recently published genome wide localization data of 38 histone modifications and one histone variant measured in human CD4+ T-cells we address two major questions: 1) What is the nature of the relationship between histone modifications and transcription? It is possible that the levels of modifications have to exceed a certain threshold in order to promote or repress transcription, thereby encoding the on/off status of a gene. Alternatively, the levels of modifications could encode the actual level of gene expression. 2) Which histone modifications are involved in distinct steps during the transcription cycle?

We derived quantitative models to predict the abundance of transcripts from a small number of modifications and relate them to the transcription cycle. We show that our models faithfully capture the measured expression levels of genes, suggesting that the levels of modifications are quantitatively related to gene expression. Given the good agreement between modeled and measured expression levels, we proceed to analyze these models to infer the relationships between modifications and the steps leading from PIC formation and Pol II recruitment to transcription initiation and transcription elongation. A key finding for our analysis is that RNA Polymerase II (Pol II) is much more abundant in expressed high CpG content promoters (HCPs) than in low CpG content promoters (LCPs). We argue that HCPs are regulated at the transition to elongation, while steps preceding initiation are rate limiting in LCPs. Quantitative models involving H3K4me3 and H3K79me1 are the most predictive of the expression levels in LCPs, while HCPs require H3K27ac and H4K20me1. We propose a preliminary “Histone Code of Transcription”, where H3K4me3 is involved in Pol II recruitment and/or initiation, the combinatorial action of H3K27ac and H4K20me1 leads to the transition to elongation, and finally H3K79me1 and H4K20me1 signal the transition to an elongating Pol II.

# The Msx1 Homeoprotein Recruits Histone Methyltransferase Activities to Control the Expression of Target Genes in the Developing Limb

Roshan M Kumar<sup>1</sup>, Jingqiang Wang<sup>2</sup>, Celia D Keim<sup>2</sup>, Raphael Margueron<sup>3</sup>, Danny Reinberg<sup>3,4</sup>, Cory Abate-Shen<sup>2</sup>, Richard A Young<sup>1,5</sup>

<sup>1</sup>Whitehead Institute for Biomedical Research, Cambridge MA; <sup>2</sup>Departments of Urology and Pathology & Cell Biology, Herbert Irving Comprehensive Cancer Center, Columbia University, College of Physicians and Surgeons, New York, NY; <sup>3</sup>Department of Biochemistry, New York University Medical School, New York, NY; <sup>4</sup>HHMI; <sup>5</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA.

Homeoproteins play central roles in directing gene expression programs during development, although the precise molecular mechanisms by which they act to pattern gene expression remain largely undefined. The Msx1 homeoprotein is expressed in undifferentiated progenitors of many tissues during mouse embryogenesis and is a key regulator of the differentiation of such tissues, including the developing limb, where it acts as a transcriptional repressor and negative regulator of differentiation. We have now identified *bona fide* target genes for Msx1 that are regulated in both cultured myoblast cells and mid-gestation embryos in the muscle and mesenchymal components of the limb. Using genome wide approaches, we have identified *MyoD*, *Myf5*, *HoxA9*, and *Six1* among the genes that are both bound and repressed by Msx1 in myoblast cells and in embryonic limbs. Msx1 binding at genomic regulatory regions is associated with an increase in repressive chromatin methylation marks, including H3K27me3 and H3K9me2. Accordingly, we found that Msx1 interacts with the histone methyltransferases Ezh2 and G9a and is able to recruit these activities to target genes. We propose that Msx1 regulates myoblast differentiation via its ability to control the histone methylation status of target genes and thereby repress their expression during limb development *in vivo*.

## Rapid Estimation of RNA Kinetics

W. Andrew Lorenz<sup>1</sup>, Peter Clote<sup>1</sup>

<sup>1</sup>Department of Biology, Boston College, Chestnut Hill, MA 02467.

The kinetics RNA folding has been experimentally shown to play a critical regulatory role in controlling plasmid copy number in *E. coli* and other bacteria (*hoksok* system), and is believed to play important roles in a number of cellular processes. *De novo* design of RNA and DNA conformational switches, once considered to be a futuristic dream of synthetic biology, is now a commercial reality; by designing such a switch, UCSB group Plaxco, Heeger and two students have developed a portable cocaine sensor.

Estimating the kinetics of RNA secondary structure folding requires substantial computational time for repeated Markov process event driven simulations (Gillespie's algorithm), as done in Vienna RNA Package *kinfold* by Flamm et al. To be of practical use in RNA molecular design, it is of paramount importance to develop a rapid and accurate computation of folding kinetics. This is our contribution in this article.

In our method, a small subset of less than 100 low energy locally optimal structures is generated, by using a new, highly nontrivial computation of the Boltzmann partition function of all locally optimal structures. A network is defined from these structures and their intersections, where the intersection of two RNA secondary structures is defined to be the secondary structure whose base pairs belong to both structures. The expected folding time along this (tiny) network is shown to be very close to that for full kinetics. Moreover, rather than simulating a Markov process along the network, we develop a novel numerically robust, linear time direct computation of the expected folding time between any two network nodes.

To demonstrate the utility of our method, we apply our algorithm to optimize the kinetics of switching between two metastable states (low energy secondary structures) for a targeted bistable switch.

# The RNA Editing Dataflow System (REDS) for the transcriptome-wide discovery of RNA modification sites

Stefan Maas<sup>1</sup>, Christina P. Godfried Sie<sup>1</sup>, Dylan E. Dupuis<sup>1</sup>, Ivan Stoev<sup>2</sup> and Daniel Lopresti<sup>2</sup>

<sup>1</sup>Department of Biological Sciences, Lehigh University; <sup>2</sup>Department of Computer Science and Engineering, Lehigh University.

RNA editing by adenosine deamination, catalyzed by the adenosine deaminases acting on RNA (ADARs), is a posttranscriptional mechanism for the regulation of gene expression and particularly widespread in mammals. A-to-I RNA editing generates transcriptome and proteome diversity and also regulates important functional properties of neurotransmitter receptor genes in the central nervous system by changing single codons in pre-mRNA.

We have previously identified wide-spread editing of non-coding transcripts [1]. In contrast, almost all currently known cases of A-to-I RNA editing that affect protein-coding sequences have been discovered serendipitously. However, since it is expected that many more such recoding editing sites exist and to understand the overall importance of RNA editing in gene regulation, it is crucial to map RNA editing sites in a systematic way.

Here we present the RNA Editing Dataflow System (REDS), a computational pipeline that allows us to predict A-to-I RNA editing sites (as well as other types of RNA modifications) in any genome for which genomic and expression databases are available. We show that a high percentage of the predicted target sites are likely bona fide editing events and go on to experimentally validate novel recoding events in three vertebrate species.

Apart from the identification of novel editing sites, our analysis provides insights on the overall landscape of RNA editing, insights on why certain sequences and RNA folds are more prone to undergo RNA editing, and with REDS provide a computational tool that should foster progress in the discovery of RNA modification well beyond this study.

[1] Athanasiadis, A., Rich, A., and Maas, S. 2004: Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biology*, 2 (12), e391, 1-15, Epub 2004 Nov 9.

# Chromatin state dynamics and the acquisition of rostrocaudal positional identity during retinoid-induced neurogenesis

Shaun Mahony<sup>1\*</sup>, Christopher C. Reeder<sup>1\*</sup>, Esteban Mazzoni<sup>2\*</sup>, Scott McCuine<sup>3</sup>, Thomas A. Jessell<sup>2,4</sup>, Richard A. Young<sup>3</sup>, Hynek Wichterle<sup>2</sup>, David K. Gifford<sup>1</sup>

<sup>1</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA; <sup>2</sup>Departments of Pathology, Neurology, and Neuroscience, Center for Motor Neuron Biology and Disease, Columbia University Medical Center, New York, NY; <sup>3</sup>Whitehead Institute for Biomedical Research, Cambridge MA; <sup>4</sup>Howard Hughes Medical Institute, Kavli Institute for Brain Science, Departments of Biochemistry and Molecular Biophysics, Columbia University, New York, NY.

\*Joint contributors

During motor neuron development, a retinoic acid gradient orients cell identity and promotes the cellular transition from pluripotency to neurogenesis. Retinoic acid signaling is also widely used to stimulate differentiation in *in vitro* models of neurogenesis. The retinoid-induced regulatory cascade leading to neuronal identity is facilitated by the remodeling/resolution of histone modifications at many developmental regulator genes from a transcriptionally poised state to transcriptionally permissive or repressive states. Little is known about the dynamics of such chromatin state transitions during differentiation or how these transitions temporally relate to gene transcription. Likewise, it is unclear what direct role retinoid signaling plays to initiate the process.

Using an *in vitro* model of motor neuron development, we characterize the retinoid response during early neurogenesis and the subsequent genomic events that lead to the differentiation of post-mitotic motor neurons. ChIP-Seq analysis of retinoic acid receptor (RAR) binding reveals that the initial directly mediated retinoid response focuses on the establishment of rostrocaudal positional identity via the induction of Hox genes and cofactors. We temporally profiled the localization of three histone modifications, H3K4me3, H3K27me3, and H3K79me2 using ChIP-Chip and integrated these data with a matched gene expression time-series to examine the chromatin state dynamics of developmental regulators downstream of the retinoid response.

Chromatin state analyses of our time-series data suggest complex temporal relationships between histone modifications and gene expression. For example, RAR binding in the Hox clusters corresponds with the rapid, non-progressive, clearance of Polycomb group proteins and H3K27me3 over domains spanning all Hox genes that will be expressed during differentiation. The same anterior Hox genes, however, display sequential expression profiles corresponding with their order in the genome. Thus, the temporal dynamics of Hox gene expression are not directly controlled by the methyltransferase activity of Polycomb- and Trithorax-group proteins.

A number of studies have noted that genes displaying 'bivalent' histone modifications in embryonic stem cells will typically display resolved histone modifications exclusively associated with transcriptional activation or repression in post-mitotic cells. In our analyses, the aforementioned example and others suggest more complex relationships between chromatin state and gene expression during differentiation, and these associations are not always temporally coupled.

# Identification of epigenetic changes in the brain of a rat model of schizophrenia using FAIRE-seq

Paul A. McGettigan<sup>1</sup>, Niamh C. O'Sullivan<sup>1</sup>, Brendan J. Loftus<sup>1</sup>, Keith J. Murphy<sup>1</sup>

<sup>1</sup>University College Dublin, Belfield, Dublin 4, Ireland

Schizophrenia is a serious condition that affects approximately 1% of the population over the age of 18. It manifests as disorganized behaviour and thinking, with positive symptoms including psychosis, negative symptoms such as withdrawal, and cognitive deficits. The age of onset is usually in young adulthood, (15-25 years in males – later in females), and this developmental emergence is a key characteristic of the disease. Other features include decreased forebrain activity (hypofrontality), and a decrease in the startle reflex as measured by prepulse inhibition (PPI).

Epidemiological research has established that multiple genetic and environmental factors are likely to contribute to schizophrenia. It is hypothesized that these factors interact at critical periods in early development to create a defective epigenetic state within certain brain structures. This causes dysregulation of connectivity and functionality, resulting in onset of disease symptomatology. The DNA regions that are most likely affected under this hypothesis are the functional regulatory elements that control the activation of genes in different cell types and at different stages of development.

The FAIRE-seq technique (Formaldehyde Assisted Identification of Regulatory Elements) can identify changes in chromatin accessibility (regions of open chromatin). These regions may be depleted of nucleosomes or more loosely bound to nucleosomes thus enabling protein-DNA binding. The regions identified by this protocol can include promoters, enhancers and insulators.

We applied FAIRE-seq to the rat isolation rearing model (animal model for PPI). We looked at differences between social and isolation reared rats in two distinct brain regions which are both linked to schizophrenia (hippocampus and prefrontal cortex (PFC)). This allowed us to identify tissue specific as well as model-specific differences in chromatin accessibility.

Rats from the same litter were separated into 2 groups and raised either socially or in isolation. At post natal day 80 brain tissue was extracted. 3 biological samples were generated per group (hippocampus social, PFC social, hippocampus isolation, PFC isolation) and prepared for Illumina sequencing.

We were able to identify 29415 total peaks. Of these 1532 peaks showed tissue specific differences and 14 peaks exhibited model-specific differences. The most significant peaks between isolation and social reared rats were several that were present in regions near genes associated with B cells of the immune system. These peaks were present in both tissues. Surprisingly the peaks were present in social rats and absent (or weaker) in isolation reared rats, this is despite the lack of a plausible B-cell population in the brain. While in line with the accumulating evidence implicating the immune system in schizophrenia, this result (validation still in progress) poses several questions as to the identity of the cells which may be expressing these immunological genes and what their function is in the brain.

## microRNAs preferentially target dosage-sensitive genes

Martin L. Miller<sup>1</sup>, Debora S. Marks<sup>2</sup>

<sup>1</sup>*Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA;* <sup>2</sup>*Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA.*

Surprisingly, most genes are not considered “dosage sensitive” and expression values may fluctuate physiologically across cells with no known pathological effects. On the other hand, groups of genes are known to be particularly sensitive to overexpression and dysregulation may have harmful consequences, for example upregulation of oncogenes. MicroRNAs can tune the expression levels of genes, but paradoxically such regulation levels are relatively small compared to the natural fluctuation of gene expression.

It is known that dosage-sensitive genes are tightly regulated at multiple stages, including by post-transcriptional microRNA regulation. If this is the case, we may expect to see dosage-sensitive genes enriched for microRNA regulation and that this phenomena should be evolutionarily conserved.

To test this hypothesis we investigated the relationship between the number of microRNA targets in a gene and its likelihood of being a dosage-sensitive.

Across worm, fly and mammals, we find that dosage-sensitive genes are more targeted than the rest of the genome, an observation that was not found in for example essential (lethal) genes.

As a consequence, we speculate that microRNA-based perturbations will have different effects on dosage-sensitive genes compared to other genes. Preliminary observations support this as we find that oncogenes are less down-regulated than other target genes after small RNA transfection, and further, inhibiting microRNAs with antagomirs up-regulates oncogenes with microRNA targets less than expected.

## Using accessibility to predict RNA-binding protein targets

Xiao Li<sup>1,4</sup>, Gerald Quon<sup>2,4</sup>, Howard Lipshitz<sup>1</sup>, Quaid Morris<sup>1-4</sup>

<sup>1</sup>Department of Molecular Genetics; <sup>2</sup>Department of Computer Science; <sup>3</sup>Banting and Best Department of Medical Research; <sup>4</sup>Donnelley Centre for Cellular and Biomolecular Research, University of Toronto, 160 College St, Toronto, Ontario M5S 1E3, Canada

Many RNA-binding proteins (RBPs) bind single-stranded RNA (ssRNA) in a sequence-specific manner. However, the inferred sequence preferences alone for these proteins are not sufficient to distinguish known targets from other potential targets that contain the same sequence motifs. Here we investigated whether mRNA secondary structure could help distinguish targets from other non-target mRNAs that contain a match to the RBP sequence binding preferences. Target site accessibility is a computationally-derived measure of the probability that a region of an RNA will be unpaired based on the ensemble of structures that the full-length transcript can assume. Target site accessibility predicts the binding of miRNA and siRNAs, suggesting that the computational models used to estimate accessibility are sufficiently accurate to be useful for *in vivo* prediction. However, it is not clear whether this result generalizes to RBPs because these ncRNAs compete for the same binding interface as mRNA secondary structure, whereas RBPs can bind RNA through a variety of interfaces. Furthermore, unlike miRNA and siRNAs, many RBPs bind mRNA in the nucleus where mRNA sequence-based predictions of mRNA secondary structure may be less accurate.

We assessed the impact of target site accessibility by attempting to predict whether transcript would co-purify in a RIP-chip assay based on the transcript sequence and the RBP consensus sequence for a diverse set of RBPs containing five different RNA-binding domains and with subcellular functions in both the cytoplasm and nucleus. We found the added predictive value of target site accessibility is substantial. Of the 13 RBPs, target site accessibility provided a significant difference in predictive accuracy for 11 of them. Importantly, considering target site accessibility never decreased accuracy. This increase in predictive power was caused neither by differential placement of consensus sites in bound transcripts, nor by the AU-richness of the flanking region. Our results suggest that target site accessibility is a general cis-regulatory mechanism and that it should be always considered when attempting to predict targets of trans-acting regulators.

## CATCHprofiles reveals nucleosome positioning of histone modifications

Fiona G. Nielsen<sup>1,2</sup>, Henk Stunnenberg<sup>2</sup>, Martijn Huynen<sup>1,2</sup>

<sup>1</sup>*Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Geert Grooteplein 26-28, 6525 GA Nijmegen, The Netherlands;*

<sup>2</sup>*Molecular Biology, Nijmegen Centre for Molecular Life Sciences, Faculty of Science, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands*

We have developed the tool CATCHprofiles to cluster and analyse ChIP profile patterns at high resolution. CATCHprofiles implements an unsupervised hierarchical clustering algorithm with an exhaustive all-pairs comparison at each clustering step. In the comparison, CATCH simultaneously aligns the ChIP signal profiles, such that the profiles are repeatedly clustered by their most informative alignment. We clustered the ChIP-seq profiles of 39 histone modifications [1] over 20340 RefSeq transcription start sites (TSSs) genome-wide.

Our results show that the profile patterns of active promoters vary not only in the presence of specific modifications, but also in the relative positioning of these modifications within the TSS. The individual cluster patterns show how certain modifications are highly correlated at all genomic positions while some locations of highly constrained nucleosomes show distinctly different patterns of histone modifications.

The high resolution patterns obtainable by CATCHprofiles allow for an expansion of the current epigenetic analyses standard from binary presence/absence of marks at the TSS to the level of genome-wide patterns of nucleosome resolution. This increase in pattern resolution is an important prerequisite to gain insight into the dynamics of chromatin reorganisation factors and nucleosome positioning.

[1] Wang *et al.* 2008 *Nature genetics* 40 (7) p. 897-903

## piRNA production in a *Drosophila* ovary cell line

Nicolas Robine<sup>1</sup>, Nelson C. Lau<sup>2</sup>, Eric C. Lai<sup>1</sup>

<sup>1</sup>Department of Developmental Biology, Sloan-Kettering Institute, New York, New York 10065, USA,<sup>2</sup> Brandeis University, Waltham, MA USA.

Piwi proteins, a subclass of Argonaute-family proteins, carry 24–30-nt Piwi-interacting RNAs (piRNAs) that mediate gonadal defense against transposable elements (TEs). We deep-sequenced the small RNAs produced in the *Drosophila* ovary somatic sheet (OSS) cell line and found that it expresses miRNAs, endogenous small interfering RNAs (endo-siRNAs), and piRNAs in abundance. In contrast to intact gonads, which contain mixtures of germline and somatic cell types that express different Piwi-class proteins, OSS cells are a homogenous somatic cell population that expresses only PIWI and primary piRNAs. Detailed examination of its TE-derived piRNAs and endo-siRNAs revealed aspects of TE defense that do not rely upon ping-pong amplification. In particular, we provide evidence that a subset of piRNA master clusters, including flamenco, are specifically expressed in OSS and ovarian follicle cells. These data indicate that the restriction of certain TEs in somatic gonadal cells is largely mediated by a primary piRNA pathway.

# Epigenetic silencing of a tumor suppressor network unmasks the dual face of master cell signals

Stefano Rossetti<sup>1</sup>, Nicoletta Sacchi<sup>1</sup>

<sup>1</sup>*Cancer Genetics Program, Roswell Park Cancer Institute, Buffalo, NY.*

Epigenetic silencing of tumor suppressor genes is common in breast cancer cells. We found that an aberrant signaling of retinoic acid (RA) via the RA receptor alpha (RARA) results in the concerted epigenetic silencing of a tumor suppressor gene network downstream of RARA. This network includes the RA receptor beta 2 (RARβ2), which mediates RA growth-inhibitory action, and TGFBR2, the main receptor of transforming growth factor beta (TGFB). Unexpectedly, we observed that both RA and TGF beta signals, which have anticancer effects in normal cells, exacerbate the tumor phenotypic features of cancer cells that underwent loss of RARβ2 and TGFBR2 tumor suppressor activities. Apparently, as a consequence of epigenetic silencing of canonical receptors, master signals such as RA and TGFB, exploit alternate targets to promote, rather than inhibit, tumorigenesis.

This work was partially supported by the National Cancer Institute grant NCI R01-CA127614-01 (NS).

# Large-scale comparative analysis of RNA structures by TOPOFIT

Preethi Srinivasan<sup>1</sup>, Tripti Kulkarni<sup>1</sup> and Valentin Ilyin<sup>1</sup>

<sup>1</sup>*Biology Department, Boston College, Chestnut Hill, MA*

The central dogma of biology presented RNA as the fundamental unit in genetic translational mechanism. We now understand that RNA molecules serve diverse structural, catalytic and regulatory function in eukaryotic cells. However, in spite of significant advancements in our understanding of the structure of RNA and its functions, there is no one collated archive of RNA structures and their relationships with one another. Structure comparison between RNA chains can provide new insights into structural neighbors of different RNA strands and possible commonality between the genes responsible for different diseases.

We present a large-scale comparative analysis of known RNA structures as structural alignments based on the TOPOFIT method (<http://topofit.ilyinlab.org>). The comparative analysis of RNA structures retrieves a number of common structural motifs and some correlations with their sequences. This knowledge can be applied towards prediction of 3D structure of RNA from its sequence. The TOPOFIT method was previously used also to create TOPOFIT DB, a database of protein structure alignments. It is based on the discovery of a saturation point on the alignment curve (topomax point) which presents an ability to objectively identify a border between common and variable parts in a protein structural family and produce accurate structural comparisons. The TOPOFIT method has been adapted to RNA structures so as to facilitate large-scale comparison of nucleic acid structures and study the functional annotations that follow. This can possibly aid in answering future challenges such as structure determination of higher order alignments of two adjacent G-quadruple and folding topology of RNA G-quadruplexes. Significant contribution can be made towards the Structural research on the sensing domains of Riboswitches in their ligand bound states. Further unique insights can be obtained from the structural perspective on RNA architecture and small molecule recognition, and from the functional perspective on RNA-mediated gene regulation through transcription termination, translational inhibition, and RNA splicing.

## Epigenetic Profiling of Human Brain Development

Yurong Xin<sup>1</sup>, Anne O'Donnell<sup>2</sup>, Benjamin Chanrion<sup>1</sup>, Yongchao Ge<sup>3</sup>, Fatemeh Haghghi<sup>1</sup>

<sup>1</sup>Department of Psychiatry; <sup>2</sup>Department of Genetics and Development, Columbia University, New York, NY; <sup>3</sup>Department of Neurology, Mt. Sinai School of Medicine, New York, NY.

We have developed a cost-effective, unbiased, whole-genome methylation profiling technique that can assay the methylation state of more than 80% of CpG sites in human genome. Using our methodology, which couples advances in next-generation sequencing with enzymatic fractionation of DNA by methylation state, we are mapping the methylation at high coverage for samples drawn from our postmortem brain collection. We focused on the prefrontal cortex (PFC) due to converging evidence from neuroimaging and functional studies implicating this region in a number of neurodevelopmental disorders as autism, depression, and schizophrenia. Secondly, we also examined the auditory cortex because some disorders such as schizophrenia include defects in sensory perception and processing. Although these regions are of interest in brain-based disorders, our initial investigations are focused on characterization of the base-line methylation profiles within normal brains. With these data we are for the first time able to explore DNA methylation profiles within two distinct brain regions with differing neurodevelopmental trajectories; the evolutionarily conserved auditory temporal cortex developing early as compared to the PFC which undergoes maturation well into early adulthood. Preliminary results reveal that DNA methylation is relatively more conserved in the auditory cortex than the PFC. These data suggest that DNA methylation together with other epigenetic factors, is essential in mediating global morphological and/or functional changes, such as during neuronal differentiation and development, or in pathophysiological states such as neurodevelopmental diseases.

## Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.

Jacob F. Degner<sup>1,3</sup>, John C. Marioni<sup>1</sup>, Athma A. Pai<sup>1</sup>, Joseph K. Pickrell<sup>1</sup>, Everlyne Nkadori<sup>1,2</sup>, Yoav Gilad<sup>1</sup> and Jonathan K. Pritchard<sup>1,2</sup>

<sup>1</sup>Department of Human Genetics, <sup>2</sup>Howard Hughes Medical Institute, and <sup>3</sup>Committee on Genetics, Genomics and Systems Biology, University of Chicago, 920 E. 58th St., CLSC 507, Chicago, IL 60637.

### ABSTRACT

Next-generation sequencing has become an important tool for genome-wide quantification of DNA and RNA. However, a major technical hurdle lies in the need to map short sequence reads back to their correct locations in a reference genome. Here we investigate the impact of SNP variation on the reliability of read-mapping in the context of detecting allele-specific expression (ASE). We generated sixteen million 35 bp reads from mRNA of each of two HapMap Yoruba individuals. When we mapped these reads to the human genome we found that, at heterozygous SNPs, there was a significant bias towards higher mapping rates of the allele in the reference sequence, compared to the alternative allele. Masking known SNP positions in the genome sequence eliminated the reference bias but, surprisingly, did not lead to more reliable results overall. We find that even after masking, 5-10% of SNPs still have an inherent bias towards more effective mapping of one allele. Filtering out inherently biased SNPs removes 40% of the top signals of ASE. The remaining SNPs showing ASE are enriched in genes previously known to harbor cis-regulatory variation or known to show uniparental imprinting. Our results have implications for a variety of applications involving detection of alternate alleles from short-read sequence data.

## Distance functions for querying large, multi species, expression databases

Hai-Son Le<sup>1</sup>, Ziv Bar-Joseph<sup>1 2</sup>

<sup>1</sup>*Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA;* <sup>2</sup>*Computer Science Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*

Advances in sequencing technology have led to a remarkable growth in the size of sequence databases allowing researchers to study new genes by utilizing knowledge about their homologs in other species. Expression databases, including the Gene Expression Omnibus (GEO) and ArrayExpress have also grown exponentially over the last decade and now hold hundreds of thousands of arrays from multiple species. However, while several methods exist for finding co-expressed genes in the *same* species as a query gene looking at co-expression of homologs or arbitrary genes in *other* species is challenging. Unlike sequence, which is static, expression is dynamic and changes between tissues, conditions and time. Thus, to carry out cross species analysis using these databases we need methods that can match experiments in one species with experiments in another species.

To facilitate queries in large databases we developed a new method for identifying such similar experiments in different species. Instead of relying on the description of the experiments, which rarely provides a match, we develop a method to determine the similarity of expression profiles by introducing a new distance function and utilizing a group of known orthologs. Our method uses a training data of known similar pairs to learn the parameters for distance functions between pairs of experiments based on the rank of orthologous genes overcoming problems related to difference in noise and platforms between species. We show that the function we learn outperforms simpler rank comparison methods that have been used in the past for single species analysis.

We used our method to compare millions of array pairs from mouse and human expression experiments. The resulting matches highlight conditions and diseases that are activating similar pathways in both species and can also hint at diseases where these pathways seem to differ.

# Modeling Idiopathic Pulmonary Fibrosis Disease Progression based on Gene and Protein Expression

Tien-ho Lin<sup>1</sup>, Jose D Herazo<sup>2</sup>, Kazuhisa Konishi<sup>2</sup>, Naftali Kaminski<sup>2</sup>, Ziv Bar-Joseph<sup>1</sup>

<sup>1</sup>*School of Computer Science, Carnegie Mellon University;* <sup>2</sup>*Simmons Center for Interstitial Lung Disease, University of Pittsburgh Medical School.*

Idiopathic pulmonary fibrosis (IPF) is a progressive fibrotic interstitial lung disease without a known cause and without established cure. Recently several high-throughput gene and protein expression analysis on IPF patients have been conducted to identify differentially expressed genes and proteins. However most of the analyzing tools are designed for static (snapshot) datasets.

We have proposed a method for classification of time series gene expression that takes temporal information into account, and showed improvement on classifying multiple sclerosis (MS) patients. Our method can both classify the time series expression datasets and account for individual differences in progression rates. Hidden Markov models (HMMs) is used to represent the expression profiles of the two classes. Using such a HMM we align a patient's time series gene expression to a common profile. Conceptually, a hidden state in our HMM correspond to a phase in the treatment response.

For biomarker discovery, we propose a backward stepwise feature selection method that utilizes the alignment to the HMM profiles, termed *HMM-RFE*. In the MS dataset, this feature selection method has been shown to improve classification accuracy and identify genes relevant to MS. In the IPF dataset, the selected genes can be further examined by more experiments to find out the causal factors of different disease progression outcome.

We collected time series expression of 29,807 genes and 13 proteins of 20 IPF patients for 3 to 7 visits, spanning over 2 years. The accuracy of predicting disease outcome improves with more time points and achieved 95% based on leave-one-out cross validation. Using the expression of only 13 proteins in 2 time points, our method still has 85% accuracy. For comparison we also classify the data by linear SVM that does not consider temporal ordering. For all three data sources (gene expression, protein expression, and the combined gene and protein expression), HMM outperforms SVM across all time points, indicating the importance of temporal information. The model not only predicts a patient's outcome, but also infers the disease progression of an individual and potentially shed light on the mechanism of IPF progression.

# Predicting Genetic Interactions in *C. elegans* using Machine Learning

Patrycja Vasilyev Missiuro<sup>1,2</sup>, Hui Ge<sup>2</sup>, Tommi S. Jaakkola<sup>1</sup>

<sup>1</sup>Computer Science and Artificial Intelligence Lab (CSAIL), MIT, Cambridge, MA;

<sup>2</sup>Whitehead Institute, MIT, Cambridge, MA

Our main objective is the discovery of genetic interactions based on sparse and incomplete information. We develop a set of machine learning techniques to investigate and predict gene properties across a variety of *Caenorhabditis elegans* datasets.

First, we show how Bayesian sets method can be applied to gain intuition as to which datasets are the most relevant for predicting genetic interactions. In order to directly apply this method to microarray data, we extend Bayesian sets to handle continuous variables. Using Bayesian sets, we show that genetically interacting genes tend to share phenotypes but are not necessarily co-localized.

One of the major difficulties in dealing with biological data is the problem of incomplete datasets. We describe a novel application of collaborative filtering (CF) in order to predict missing values in the biological datasets. We adapt the factorization-based and the neighborhood-aware CF<sup>1</sup> to deal with a mixture of continuous and discrete entries. We use collaborative filtering to input missing values, assess how much information relevant to genetic interactions is present, and, finally, to predict genetic interactions. We also show how CF can reduce input dimensionality.

Using collaborative filtering we fill in the missing entries in the input data describing genes. Since the input matrix is no longer sparse, we are able to use Support Vector Machines to predict genetic interactions. We find that SVM with a nonlinear *rbf* kernel has greater predictive power over CF.

Overall, our approaches achieve substantially better performance than previous attempts at predicting genetic interactions. We emphasize the features of the studied datasets and explain our findings from a biological perspective. We hope that our work possesses an independent biological significance by helping one gain new insights into *C. elegans* biology: specific genes orchestrating developmental and regulatory pathways, response to stress, etc.

[1] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. Proceedings of the 13<sup>th</sup> SIGKDD international conference on Knowledge Discovery and Data Mining, pp 95-104, 2007.

# The integrated pathway and proteomic resources for identifying the potential colorectal cancer biomarkers

Kun-Nan Tsai<sup>1,2</sup>, Guang-Wu Chen<sup>1,2</sup>, Kuei-Tien Chen<sup>3</sup>, and Err-Cheng Chen<sup>3</sup>

<sup>1</sup>Research Center for Emerging Viral Infections, <sup>2</sup>Department of Computer Science & Information Engineering, Chang Gung University, Kwei-Shan, Taoyuan, Taiwan.

<sup>3</sup>Institute of Medical Biotechnology and Laboratory Science, Chang Gung University, 259 Wen-Hwa 1st Road, Kwei-Shan, Tao-Yuan, 333, Taiwan

## Abstract

Quickly finding out the potential colorectal cancer biomarkers so as to investigate further the significance, distribution, clustering of the potential biomarkers can offer the important information of protein expression; benefit the follow-up study on human colorectal cancer insight. But there is still want of a set of available analytical methods at present, so it is quite essential and even urgent to develop the annotated method. Here, we present a potential biomarkers knowledge-based approach included the following steps: (1) profiling the significant protein-protein interaction through pathway database, (2) utilizing Fisher's exact test to analyze pathways information and to rank the significant pathways, (3) identifying related proteins as the potential biomarkers for human colorectal cancer by referring to genomic, transcriptomic, proteomic, functional and disease related information from GeneCard, (4) annotating potential biomarkers in the pathways through density, quality and cluster. Our results illustrate the influence of different stages of human colorectal cancer on the protein expressions, at least its involvement in cell proliferation, cell motility, and cell survival, etc. According to GeneCard information, all meaningful biological information will contribute to find the novel potential biomarkers for different stages of human colorectal cancer. In addition, by our biological experimental result, it has verified that ADAM10 protein was the common biomarkers for different stages of human colorectal cancer. These are evidences that prove the practicability of our approach. With this information we hope it will make a contribution to the further study on the prevention and cure of human colorectal cancer.

# Learning compact models of DNA binding specificities for transcription factors from protein binding microarrays

Phaedra Agius<sup>1</sup>, Aaron Arvey<sup>1</sup>, William Chang<sup>1</sup>, William Stafford Noble<sup>2</sup>, Christina Leslie<sup>1</sup>

<sup>1</sup>Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, NY,

<sup>2</sup>Department of Genome Sciences, University of Washington, Seattle, WA.

Accurately modeling the DNA sequence preferences of transcription factors (TFs), and using these models to predict *in vivo* genomic binding sites for TFs, are key pieces in deciphering the regulatory code. For years, these efforts have been frustrated by the inadequacy of TF binding site motif models which match large numbers of sites and produce an unreliable list of potential target genes. Recently, protein binding microarray (PBM) experiments have emerged as a new source of high-resolution data on *in vitro* TF binding specificities. PBMs measure the binding of a fluorescently tagged TF to a carefully designed set of ~44K double-stranded DNA probes. How best to use this data to represent the TF's binding preferences is an open question. So far, PBM data has been analyzed via rank statistics on probe intensities from the TF binding experiment, so that individual sequence patterns (e.g. 8-mers or longer gapped patterns) are assigned enrichment scores (E-scores). This representation is informative but unwieldy because every TF is assigned a list of thousands of scored sequence patterns. Here we apply supervised learning methods to PBM data to learn compact and statistically more powerful models of *in vitro* TF binding preferences. These models can be readily used to scan intergenic regions for predicting *in vivo* binding sites. We use a novel 1<sup>st</sup> order Markov mismatch string kernel to represent probe sequence similarities, and we use support vector regression (SVR) to learn the mapping from probe sequences to PBM binding intensities (Figure 1a). Using a large data set of 33 yeast and 114 mouse TFs for which PBM data for two independent probe designs is available, we show that our SVR models can better predict probe intensity than the current E-score method (Figure 1b). Moreover, by using SVRs to score yeast intergenic regions, we are better able to predict *in vivo* occupancy as measured by ChIP chip experiments than a previous occupancy scoring method based on E-scores or PSSM-based prediction. This flexible SVR framework can be broadly used in place of PSSMs to improve modeling of regulatory sequences.

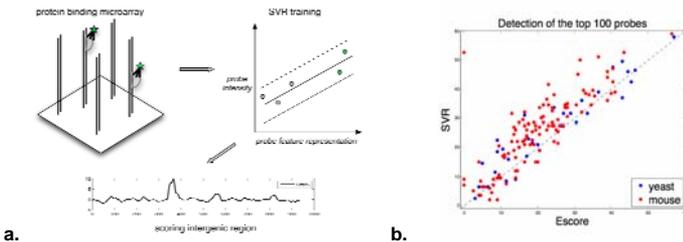


Figure 1. a. PBM data for a TF is used to train an SVR model that predicts probe intensity from sequence. The trained SVR can scan intergenic regions to predict TF occupancy. b. SVR models outperform E-scores for detection of top 100 TF-bound probes within top 100 predictions when testing on an independent probe set.

# Computational discovery of *Cis*-regulatory elements in multiple *Drosophila* species

Manonmani Arunachalam<sup>1</sup>, Karthik Jayasurya<sup>2</sup>, Uwe Ohler<sup>2,\*</sup> and Pavel Tomančák<sup>1,\*</sup>

<sup>1</sup>Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany

<sup>2</sup>Institute for Genome Sciences & Policy, Duke University, Durham NC, USA

Gene regulation lies at the heart of most biological processes and transcription factors are the key molecules that control tissues specific gene expression. In higher eukaryotes transcription factors control gene expression by binding regulatory DNA segments called *cis*-regulatory modules (CRMs). The increasing number of sequenced genomes of multicellular eukaryotes along with high-throughput methods such as whole genome expression data allows for systematic characterization of the CRMs that control gene expression. We here present two approaches that utilize such data to identify *cis*-regulatory elements and modules.

A first step towards understanding gene regulation is the identification of the regulatory elements present in the genome. In the first approach, we take advantage of the large database of spatio-temporal patterns of gene expression in *D. melanogaster* embryogenesis to identify sets of developmentally co-expressed genes. We developed a computational method that identifies DNA binding sites for transcription factors from families of co-regulated genes that are expressed during *Drosophila* embryo development. This method discovers over-represented motifs in a set of co-regulated genes using the exhaustive motif enumeration technique. Clustering the predicted motifs identifies the CRMs, which assist in translating a combinatorial code of TF inputs into a specific gene expression output. After searching the whole genome, predicted CRMs were verified experimentally by establishing expression patterns of the genes that are associated with these CRMs.

It is well known that the gene expression is substantially controlled through CRMs and those key regulatory sequences are conserved in related species. The conservation of CRMs can be studied by comparing the related genomes, and alignment methods are widely used computational tools for comparing the sequences. However, in distantly related species, alignments of non-coding CRM sequences become non-reliable. To identify orthologous CRMs in distantly related species we developed a non-alignment based method based on word frequencies, where the given sequences are compared using Poisson based metric. When starting with a set of CRMs involved in *Drosophila* development, we show here that our non-alignment method has better discriminative power than conservation scores based on alignments, and successfully detects similar CRMs in distantly related species (*D. ananassae*, *D. pseudoobscura*, *D. willistoni*, *D. mojavensis*, *D. virilis*, *D. grimshawi*). This method proved efficient in discriminating functional CRMs from known non-functional candidates with similar binding site sequence occurrences.

# Studying the evolution of promoters: a waiting time problem

Sarah Behrens<sup>1</sup>, Martin Vingron<sup>1</sup>

<sup>1</sup>Computational Molecular Biology, Max Planck Institute for Molecular Genetics,

Inhnestr. 63-73, 14195 Berlin

While the evolution of coding DNA sequences has been intensively studied over the past decades, the evolution and structure of regulatory DNA sequences still remain poorly understood. However, there is growing body of experimental evidence that promoter sequences are highly dynamic and that significant changes in gene regulation can occur on a microevolutionary time scale.

In order to give a probabilistic explanation for the rapidness of cis-regulatory evolution, we have addressed the following question: how long do we have to wait until a given transcription factor (TF) binding site (a given k-mer or a set of k-mers) emerges at random through the evolutionary process of single nucleotide mutations?

Using a Markovian model of sequence evolution, we can exactly compute the expected waiting time until a TF binding site is newly created in a promoter sequence of a given length. The evolutionary rates of nucleotide substitution are estimated from a multiple species alignment (*Homo sapiens*, *Pan troglodytes* and *Macaca mulatta*). Since the CpG methylation deamination process (CG->TG and CG-> CA) is the predominant evolutionary substitution process, we have also incorporated these neighbor dependent substitution rates into our model.

As a result, we obtain expected waiting times for every k-mer,  $3 \leq k \leq 10$ . Therefore, we can identify TF binding sites which can be easily generated during evolution and those which are not very "convenient" to "wait for". For example, "CCCTG" is the fastest emerging 5-mer with an expected waiting time of 82 million years (Myrs) to appear in one promoter of length 1 kb and approximately 4,000 years to occur in at least one of all the human promoters, while "ATATA" is the slowest emerging 5-mer (338 Myrs for one promoter; 17,000 years for appearance in at least one of all the human promoters). For 10-mers, the average expected waiting time is 96 billion years for one promoter and around 5 Myr for all promoters - suggesting that in terms of time, it is more favorable to create several short TF binding sites instead of one long TF binding site.

Our results indicate that new TF binding sites can indeed appear on a small evolutionary time scale and that the CpG methylation deamination process probably is one of the driving forces in generating new TF binding sites. Our approach of calculating waiting times for TF binding sites in dependency of their length and composition sheds new light on the process of TF binding site emergence and therefore extends the previous knowledge about the dynamics of promoter sequence evolution.

# Integrated Approach for the Identification of Human HNF4 $\alpha$ Target Genes Using Protein Binding Microarrays

Eugene Bolotin<sup>1,6</sup>, Hailing Liao<sup>4</sup>, Tuong Chi Ta<sup>2</sup>, Chuhu Yang<sup>1</sup>, Wendy Hwang-Verslues<sup>3</sup>, Jane R. Evans<sup>2</sup>, Tao Jiang<sup>5,6</sup>, and Frances Sladek<sup>2,6</sup>

<sup>1</sup>Genetics, Genomics and Bioinformatics; <sup>2</sup>Cell, Molecular, and Developmental Biology; and <sup>3</sup>Environmental Toxicology Graduate Programs; Departments of <sup>4</sup>Cell Biology and Neuroscience, <sup>5</sup>Computer Science and Engineering, <sup>6</sup>Institute for Integrated Genome Biology, University of California, Riverside, CA, 92521-0121 USA

Hepatocyte nuclear factor 4 alpha (HNF4 $\alpha$ ), a member of the nuclear receptor superfamily, is essential for liver function and linked to several diseases including diabetes, hemophilia, atherosclerosis and hepatitis. While many DNA response elements and target genes have been identified for HNF4 $\alpha$ , the complete repertoire of binding sites and target genes in the human genome is unknown. We adapt protein binding microarrays (PBMs) to examine the DNA binding characteristics of two HNF4 $\alpha$  species (rat and human) and isoforms (HNF4 $\alpha$ 2 and HNF4 $\alpha$ 8) in a high throughput fashion. We identified ~1,400 new binding sequences and used this dataset to successfully train a Support Vector Machine (SVM) model that predicts an additional ~10,000 unique HNF4 $\alpha$  binding sequences; we also identify new rules for HNF4 $\alpha$  DNA binding. We performed expression profiling of an HNF4 $\alpha$  RNAi knockdown in HepG2 cells and compared the results to a search of the promoters of all human genes with the PBM and SVM models, as well as published genome-wide location analysis. Using this integrated approach, we identified ~240 new direct HNF4 $\alpha$  human target genes, including new functional categories of genes not typically associated with HNF4 $\alpha$ , such as cell cycle, immune function, apoptosis, stress response and other cancer-related genes. In conclusion, we report the first use of PBMs with a full length native transcription factor in a crude nuclear extract, additionally we greatly expand the repertoire of HNF4 $\alpha$  binding sequences and target genes, thereby identifying new functions for HNF4 $\alpha$ . We also establish a web-based tool, HNF4 Motif Finder, that can be used to identify potential HNF4 $\alpha$  binding sites in any sequence.

## Applications of Centroid Estimation to Regulatory Genomics

Luis E. Carvalho<sup>1</sup>, William A. Thompson<sup>2</sup>, Lee A. Newberg<sup>3,4</sup>, Charles E. Lawrence<sup>2</sup>

<sup>1</sup>*Dept. of Mathematics and Statistics, Boston University, Boston, MA 02215;* <sup>2</sup>*Center for Computational Molecular Biology and the Division of Applied Mathematics, Brown University, Providence, RI 02912;* <sup>3</sup>*Dept. of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180;* <sup>4</sup>*Wadsworth Center, NYS Dept. of Health, Albany, NY 12201.*

Many problems in regulatory genomics, such as the identification of conserved regulatory motifs and the prediction of regulatory gene interactions, can be seen as statistical inference problems. The traditional approach for this class of problems has been to myopically select the most likely solution, the maximum likelihood or MAP estimators. However, these estimators ignore contributions from all other solutions in the ensemble. Though this approach has the appeal of plausibility and theoretical justifications based on good properties of the solution as more data is observed, we argue that there is no principled reason for these estimators to work well. First, although a wealth of genomic and comparative data is available when addressing these problems, the solution space is very large and often much larger than the size of the data and thus the desired asymptotic behavior of the most likely solution is not valid anymore. Moreover, as an estimator, the most likely solution can be shown, under the principles of statistical decision theory, to minimize the risk of a very restrictive loss function, the zero-one, all-or-nothing loss.

We present an alternative estimator, the centroid estimator, and discuss some applications of this new contender to the problems mentioned above, prediction of conserved regulatory motifs given sequence and phylogenetic data and inference of regulatory networks. The centroid estimator minimizes a more general loss function that compares ensemble components and can thus be shown to yield more reliable and accurate solutions. Perhaps more importantly, the centroid estimate can be obtained with no worse computational complexity than required for maximum likelihood estimation. The literature already contains three applications in which centroid estimators have been shown to improve the prediction of ground truth standards: RNA secondary structure prediction; structural prediction by homology; and motif finding. In each of these three cases exactly the same statistical model was employed for both centroid and MAP estimators. Here we illustrate the performance of these estimators with transcription regulation examples and offer directions for future research and improvements.

## A missing ingredient in the Pho4 paradigm? Evidence for Pho4/Cbf1 binding site competition

Wee Siong Goh<sup>1</sup>, Jonathan Aow<sup>1</sup>, Jin Quan Run<sup>1</sup>, Yuriy Orlov, Geoffrey Lim<sup>1</sup>, Neil D. Clarke<sup>1</sup>

<sup>1</sup>*Genome Institute of Singapore, Singapore*

The yeast Pho4 regulon is a model system for studying the interdependencies of nucleosome occupancy and transcription factor (TF) binding, and their effects of such binding on gene expression. Of particular interest are experiments and computational models that involve variants of the Pho5 promoter.[1-3] These studies have suggested that differences in nucleosome occupancy at TF binding site lead to differences in how a binding site contributes to the maximal induction of the promoter, on the one hand, and how a site contributes to the responsiveness of the promoter to sub-saturating levels of TF, on the other.

To explore related issues, we have constructed 35 promoter variants of the Pho4-target gene VTC4, fused to GFP. All of the promoters contain either one or two Pho4 binding sites, and all of the binding sites are palindromes with perfect matches to the core CACGTG recognition site. The sites vary at flanking positions that extend the palindromic half-sites by two nucleotides, and which confer more subtle effects on binding affinity. Under fully inducing condition the expression levels from these promoters are remarkably well correlated with in vitro measurements of the free energy of binding of Pho4 to variant binding sites. However, at high phosphate concentrations, expression differences are uncorrelated with predicted Pho4 binding, but are correlated instead with predicted binding by Cbf1. Cbf1, which has been implicated in chromatin remodeling, has the same CACGTG core specificity as Pho4, but different specificity for flanking bases. Interestingly, differences in the expression pattern of Pho4 target genes that have previously been attributed to differences in the exposure or occlusion of Pho4 binding sites are also associated with differences in the relative affinities of the promoters for Pho4 or Cbf1. We are validating these conclusions with ChIP experiments. We have also extended these analysis genome-wide, using phosphate-dependent gene expression data, nucleosome occupancy maps, and predicted binding for both Pho4 and Cbf1.

Our results suggest that descriptions of Pho4-mediated gene regulation that lack consideration for Cbf1 may be neglecting an important component of the system and that the chromatin state of the promoter, which affects Pho4 binding, is partly a consequence of binding by Cbf1.

1. Raveh-Sadka T, Levo M, Segal E (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* 19: 1480-1496.
2. Lam FH, Steger DJ, O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453: 246-250.
3. Kim HD, O'Shea EK (2008) A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* 15: 1192-1198.

# Using ChIP seq to search for sequence determinants of binding of the *S. cerevisiae* transcription factor Sko1

Kristen E. Cook<sup>1,2</sup>, Erin K. O'Shea<sup>1,2,3</sup>

<sup>1</sup>Department of Molecular and Cellular Biology, Harvard University, ; <sup>2</sup>Harvard University Faculty of Arts and Sciences Center for Systems Biology; <sup>3</sup>Howard Hughes Medical Institute.

In *Saccharomyces cerevisiae*, the ATF/CREB family transcription factor Sko1 activates ~ 50 genes in response to salt stress. For a subset of these genes, Sko1 is bound to the promoter both before and after stress. At these genes, Sko1 acts as a repressor under favorable growth conditions, and then switches to an activator in response to high salt. At another class of genes, Sko1 is recruited to promoters in response to stress. These two binding modes cannot be explained by known Sko1 binding motifs. While a consensus binding site for Sko1 has been reported, this site is not sufficient to predict Sko1 binding, or to differentiate classes of promoters with different binding behaviors. Our approach uses chromatin immunoprecipitation followed by sequencing to precisely identify the location of Sko1 binding sites that are occupied in response to stress, to a resolution of < 200 base pairs (width of binding peak at half maximum). We will then use these regions to constrain motif searches.

## Effects of motif and CNS multiplicity on gene expression in subspaces of conserved eigensystems following stroke and seizures

Michał Dabrowski<sup>1</sup>, Norbert Dojer<sup>2</sup>, Malgorzata Zawadzka<sup>1</sup>, Jakub Mieczkowski<sup>1</sup>, Bożena Kamińska<sup>1</sup>

<sup>1</sup>Nencki Institute, Laboratory of Transcription Regulation, Warsaw; <sup>2</sup>Institute of Informatics, University of Warsaw, Poland

Last year, we reported two SVD modes (eigensystems) with eigenarrays conserved between the datasets from gene expression profiling of rat brain following either stroke (in the MCAO model) or kainate-induced seizures. We reported that these two conserved modes separate concurrent genome-wide effects of biological processes of inflammation/apoptosis (mode 2) and synaptic activity (mode 3) on gene expression. We also reported identification of the motif binding transcription factor AP1 as associated with up-regulation of expression in the subspace of the mode 2, and of several motifs, including the motifs binding Creb and Egr, as associated with gene up-regulation the subspace of mode 3.

We now complement these previous findings by demonstrating a mode 2 and MCAO-specific antagonistic effect of the motif binding protein Satb1 on the AP1-driven up-regulation of gene expression. Motif binding Satb1 on its own was associated with down-regulation of gene expression in the subspace of mode 2. The effect of Satb1-binding motif on gene log-expression was linearly dependent on the count of this motif in all the putative regulatory regions of each gene. Satb1 is the most characterized nuclear matrix associated region (MAR) binding protein, involved in regulation of apoptosis and inflammation. Our results suggest a role of chromatin conformation in regulation of the response to the ubiquitous transcriptional regulator AP1.

Among the previously identified transcription factors regulating mode 3, we report a highly significant linear effect of the count of the motif binding Creb on gene log-expression, following both MCAO and the kainate-induced seizures. Interestingly, we find that another simpler variable, namely the count of CNSs (conserved non-coding sequences) per gene, is also linearly proportional to gene log-expression in the subspace of neuronal-activity specific mode 3. This effect, however, is dependent on the CNSs' content of Creb-binding motifs. These findings suggest that the more numerous CNSs that had been reported before for neuron-specific genes may reflect not only a more complex regulation, but also a need for their strong activation (via Creb) in response to rapidly changing synaptic activity.

## Backup in gene regulatory networks explains differences between binding and knockout results

Anthony Gitter<sup>1</sup>, Zehava Siegfried<sup>2</sup>, Michael Klutstein<sup>2</sup>, Oriol Fornes<sup>3</sup>, Baldo Oliva<sup>4</sup>, Itamar Simon<sup>2</sup>, Ziv Bar-Joseph<sup>1,5</sup>

<sup>1</sup>Computer Science Department, Carnegie Mellon University; <sup>2</sup>Department of Molecular Biology, Hebrew University Medical School; <sup>3</sup>Municipal Institute for Medical Research (IMIM-Hospital del Mar); <sup>4</sup>Department of Experimental Sciences and Health, Pompeu Fabra University; <sup>5</sup>Machine Learning Department, Carnegie Mellon University

The complementarity of gene expression and protein-DNA interaction data led to several successful models of biological systems. However, recent studies in multiple species raise doubts about the relationship between these two types of data. These studies show that the overwhelming majority of genes bound by a particular transcription factor are not affected when that factor is knocked out. Here we show that this surprising result can be partially explained by considering the broader cellular context in which transcription factors operate. Factors whose functions are not backed up by redundant paralogs show a fourfold increase in the agreement between their bound targets and the expressions levels of those targets. In addition, we demonstrate that incorporating protein interaction networks provides physical explanations for knockout effects.

Building on work we presented at the 5th Annual RECOMB Satellite on Regulatory Genomics, we predict several pairs of transcription factors we believe can compensate for each other's deletion. New double knockout experiments and additional double knockout data from literature directly support our conclusions. We find that single knockouts of these transcription factors affect the expression levels of their bound genes weakly, if at all, whereas double knockouts of both a factor and its redundant partner lead to significant differential expression.

Our results highlight the robustness provided by redundant transcription factors and indicate that in the context of diverse cellular systems, binding is still largely functional.

# Novel thermodynamics-based algorithm for probe-specific position-dependent hybridization free energy

Hosna Jabbari<sup>1</sup>, Peter Clote<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of British Columbia <sup>2</sup> Department of Biology, Boston College.

DNA microarray technology plays an important role in the identification of both transcription and microRNA networks, providing new and fundamental insights in molecular and cell biology. Transcription factor binding sites have been identified in vivo by ChIP-chip, a method where DNA-binding proteins are cross-linked to the DNA, subsequently fragmented, separated by an antibody reaction, and the protein-binding DNA regions are identified by hybridization to a tiling array. While it has long been known that cellular conditions (starvation, cold shock, heat shock, disease, etc.) are reflected in the profile of transcription levels of certain protein-coding genes, only recently has the regulatory network of microRNAs begun to be explored. The future seems poised for further discoveries concerning the microRNA regulatory network, especially since it has been reported that microRNA gene chips in fact appear to capture more pathogenomic tumor information than do protein-coding gene chips such as GeneChips. More generally, high-throughput bio-technologies such as gene expression arrays, single nucleotide polymorphism (SNP) arrays, etc. herald a new era of pharmacogenomics and toxicogenomics.

Despite technological advances, DNA microarray technology is still limited, in that messenger RNA concentrations cannot be inferred due overwhelmingly to cross-hybridization noise caused by non-specific binding (NSB). Statistics-based algorithms, such as dChip, MAS 5.0, RMA, GC-RMA, etc. attempt to estimate NSB by fitting data. In contrast there is increasing evidence that physical modeling of microarray hybridization yields information not available otherwise; this is evidenced by the now widely accepted Langmuir adsorption model to explain probe-target saturation at high concentration levels. We present a thermodynamics-based algorithm that computes the free energy of cross-hybridized structures causing non-specific binding, singly the most important source of error in current microarray data analysis. The novelty of our approach consists of merging the following two methods.

1. Inclusion of position-dependent weights for nucleotides in the chip probe involved in hybridization to account for dependencies of binding strengths on distance from surface.
2. Dynamic programming algorithms to compute the minimum free energy and Boltzmann partition function for the hybridization of DNA-RNA complexes.

Although many thermodynamic-based algorithms such as Unafold, DINAMelt, Vienna RNA package, and Nupack exist that calculate RNA minimum free energy or hybridization, our intended application cannot be inferred from any of them individually or combined. The partition function in the above mentioned tools are calculated in solution, whereas in microarrays, cDNA probes are attached to a glass surface, that affects hybridization to the target RNA. In our proposed algorithm, we calculate the Boltzmann partition function for hybridization of cDNA probe and the target RNA, when the cDNA sequence is concatenated to a linker and the target RNA sequence. Our method uses experimentally determined thermodynamic parameters for DNA-RNA stacked pair interactions. In addition, we include position-dependent weights within the multiple recursions of calculating partition function, to account for the position of the probe nucleotide from the glass surface, when that nucleotide is hydrogen-bonded either to another probe nucleotide (probe partial secondary structure) or to target or non-target RNA. Our dynamic programming algorithm runs in  $O(n^3)$  time and  $O(n^2)$  space (where  $n$  is the length of the total sequence including probe, linker and target), matching the state-of-the-art algorithms for predicting RNA secondary structure or hybridization. At this point, we are not considering pseudoknotted structures that can form inter or intra-molecularly, to accelerate the running time of our algorithm. However, extension of our algorithm to include important pseudoknotted structures is a future plan.

# A Computational Investigation of Widespread Stop Codon Readthrough in *Drosophila*

Irwin Jungreis<sup>1</sup>, Clara S. Chan<sup>2</sup>, Michael F. Lin<sup>1,3</sup>, Manolis Kellis<sup>1,3</sup>

<sup>1</sup>MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA; <sup>2</sup>MIT Biology Department, Cambridge, Massachusetts 02139, USA; <sup>3</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA.

In a 2007 paper in *Genome Research*, Lin et al used a comparative genomics approach to determine protein coding regions of the *Drosophila* genome. One of their findings was that for 149 genes it appears that the open reading frame following the stop codon is protein coding, suggesting that translational stop codon readthrough is occurring. While stop codon readthrough is known to occur for isolated genes in various species, it was not previously known to be this widespread in any species.

We have used various computational approaches to investigate this widespread stop codon readthrough in *Drosophila*.

By refining the techniques used in the 2007 paper, we have expanded the list of *Drosophila* genes that display this phenomenon to over 300. We have found additional evidence that these post-stop codon regions are protein coding, and ruled out alternative explanations including alternative splicing and dicistronic translation.

We have found several examples that appear to be double readthrough, suggesting that readthrough is regulated rather than an unregulated low level of ribosomal leakage.

We have investigated various properties that distinguish readthrough candidates from other genes. They are generally longer and have longer 3'- and 5'-UTR. Although there are many readthrough candidates with each of the three stop codons, the stop codon is unusually well conserved, whichever one it is. The distribution of stop codons and the subsequent base is almost the reverse of the distribution among background genes. We have also found that several bases before the stop codon are significantly conserved.

We have looked for stop codon readthrough using these techniques in other species. We have found that several of the readthrough candidates in *Drosophila* appear to also exhibit readthrough in other insect species. We have also found isolated examples that appear to be stop codon readthrough in mammals and also in nematodes.

It was previously known that one example of stop codon readthrough in *Drosophila*, *hdc*, has a stem loop following the stop codon that causes readthrough. We have found that the energy of this stem loop is significantly conserved and found some other readthrough candidate genes that also have conserved stem loops following the stop codon.

Finally, we consider the likely amino acid sequence produced when the readthrough genes are translated.

# Identification of Noncoding Motifs Under Selection in Coding Sequences

Deniz Kural<sup>1</sup>, Yang Ding<sup>1</sup>, Jiantao Wu<sup>1</sup>, Alicia Korpi<sup>1</sup>, Jeffrey H. Chuang<sup>1</sup>

<sup>1</sup>*Boston College Department of Biology, Chestnut Hill, MA 02467, USA*

Coding nucleotide sequences contain myriad functions independent of their encoded protein sequences. We present an algorithm (COMIT) to detect functional non-coding motifs in coding regions using sequence conservation, explicitly separating nucleotide from amino acid effects. COMIT agrees with diverse experimental datasets, including splicing enhancers, silencers, replication motifs, and microRNA targets, and predicts many novel functional motifs. Intriguingly, COMIT scores are well-correlated to scores uncalibrated for amino acids, suggesting that nucleotide motifs often override peptide-level constraints.

# Protein-protein interactions improve multiple transcription factor binding site prediction

Kirsti Laurila<sup>1</sup>, Heidi Pukkila<sup>1</sup>, Olli Yli-Harja<sup>1</sup>, Harri Lähdesmäki<sup>1,2</sup>

<sup>1</sup>*Department of Signal Processing, Tampere University of Technology, Finland;*

<sup>2</sup>*Department of Information and Computer Science, Helsinki University of Technology, Finland.*

An important step in gene expression processes is transcriptional regulation. This regulation is largely controlled by transcription factors binding to DNA sites, chromatin structure and other epigenetic factors. As the identification of transcription factor binding sites with experimental methods can be very laborious, prediction methods are needed. Several prediction algorithms and programs have been developed but most of them either predict binding sites for a single factor at a time or they rely on prior knowledge of co-operating transcription factors [1]. Here, we present a probabilistic model that predicts binding of several transcription factors simultaneously with the help of transcription factor binding specificity models and protein-protein interactions [2].

Our method tries to mimic the situation in the nucleus by including in prediction several transcription factors and thus explicitly modeling the competition between the factors. As factors are known to interact with each other, we also add the prior knowledge of existing interactions to the model.

Our results show that our protein-protein interactions guided method performs better than the method without interactions or predictions where individual binding prediction results of separate TFs have been combined. The number of false positives is reduced remarkably compared with the individual predictions. Moreover, binding sites that were unpredictable with other methods could be identified with our protein-protein interaction guided method.

We have also studied how additional data sources, such as evolutionary conservation and nucleosome locations, affect the predictions.

[1] Hannenhalli . (2008), Eukaryotic transcription factor binding sites – modeling and integrative search methods. *Bioinformatics*, 24:1325-31.

[2] Laurila K, Yli-Harja O, Lähdesmäki H. (2009) A protein-protein interaction guided method for competitive transcription factor binding improves target predictions, *Nucleic Acids Res.*

# Successful Enhancer Prediction from DNA Sequence

Dongwon Lee<sup>1</sup>, Rachel Karchin<sup>1,2</sup>, Michael Beer<sup>1,3</sup>

<sup>1</sup>*Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA;*

<sup>2</sup>*Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA;*

<sup>3</sup>*McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA*

Identifying the sequence elements which determine enhancer function is necessary to fully understand how enhancers direct the spatial and temporal regulation of gene expression. However, the systematic identification of enhancers has been limited due to the fact that they are often distant from the genes they regulate, leading to large amounts of potentially regulatory sequence. Visel et. al recently introduced a new experimental method for identifying enhancers with tissue specific activity in the mouse genome. [1] Several thousands of p300 bound DNA elements (an accurate marker of enhancer activity) were collected by ChIP-seq technology, and their spatial and temporal activities were validated by transgenic mouse enhancer assay. This dataset provides an unprecedented opportunity to study sequence features of tissue specific enhancers.

Here, we show that sequence features in the experimentally identified enhancer set are enough to accurately (around 94% of area under ROC curve with 5-fold cross-validation) discriminate them from the random genomic regions by a support vector machine (SVM) classifier. We also show that the most predictive sequence elements (those with large SVM weights) are related to known transcription factor binding sites important for tissue specific development. E-box elements (CANNTG), TAAT cores, and TCF binding sites (ACAAAG) for example, have large SVM weights in the forebrain enhancers, consistent with previously reported experimental evidence. Finally, we show that these elements have positional constraints within the enhancers, and that they are more likely to be evolutionarily conserved than less predictive elements in the enhancers. Our findings provide new insights into the general structure of mammalian enhancers.

[1] Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 854-858 (2009).

## Identifying motifs using GADEM with a starting

Leping Li<sup>1</sup>, Gordon Robertson<sup>2</sup>, Brad G. Hoffman<sup>3</sup>, Pamela A Hoodless<sup>4,5</sup>, Steven JM Jones<sup>2,5</sup>

<sup>1</sup>Biostatistics Branch, National Institute of Environmental Health Sciences, RTP, NC 27709; <sup>2</sup>BC Cancer Agency Genome Sciences Centre, Suite 100, 570 West 7th Ave, Vancouver, BC, Canada V5Z 4S6; <sup>3</sup>Department of Cancer Endocrinology, BC Cancer Agency, 675 West 10th Ave, Vancouver, BC, Canada V5Z 1L3; <sup>4</sup>Terry Fox Laboratory, British Columbia Cancer Agency and <sup>5</sup>Department of Medical Genetics, University of British Columbia, Vancouver, B.C., Canada V6T 1Z3.

Previously, we reported GADEM, an efficient *de novo* motif discovery tool for large-scale genomic sequence data. We present an updated version, v1.3, that has substantial improvements and additions. We added a 'seeded' analysis in which a user-specified position weight matrix (PWM) is the starting PWM model. Seeded analyses are at least 10x faster and perhaps more accurate than the already scalable 'unseeded' analyses, and can identify short and less abundant motifs, and variants of dominant motifs. No existing tools are very effective for short/degenerate and/or less abundant motifs in large datasets. In addition, we propose an approach for estimating the number of binding sites in the data, include non-uniform priors for motif location that take advantage of the high spatial resolution of ChIP-seq data, and support higher-order Markov background models. Finally, output now reports each motif's fold enrichment in input data vs. background/random sequence data. These changes substantially enhance GADEM's functionality and efficiency for motif discovery in large-scale genomic data.

We tested the new version on four additional ChIP-seq datasets from adult mice (Hnf4a and FoxA2 in liver, and Foxa2 and Pdx1 in pancreas islets) (Hoffman et al., submitted). Runs targeted 400-bp peak core sequences, and each dataset contained between ~7 and 13 thousand FDR-thresholded peaks. For Hnf4a, seeded and unseeded runs returned similar dominant Hnf4a-like motifs in 74.4% vs. 71.2% of peak cores respectively, despite the seeded analysis being ~13 times faster than unseeded; both seeded and unseeded analyses also identified variant Hnf4a-like motifs. For Foxa2, seeded runs returned similar FoxA2 motifs in both tissues. For Pdx1, whose motif is short, unseeded analysis failed to identify a satisfactory motif in Pdx1 ChIP-seq sequences, while seeded analysis identified both a Pdx1 and a Pdx1:Pbx1 dimer motif in 45.1% and 47.6% of peak cores respectively. These results highlight that GADEM v1.3 offers increases in speed, stability and (possibly) sensitivity. In our hands, it is effective for motif discovery with the large enriched region sets typical of mammalian ChIP-seq data.

Availability: The updated software in C source is available at [www.niehs.nih.gov/research/resources/software/gadem/](http://www.niehs.nih.gov/research/resources/software/gadem/).

Contact: [li3@niehs.nih.gov](mailto:li3@niehs.nih.gov).

## A novel method to simulate genome-wide background noise and distinguish real binding sites from background noise

Hong Lu<sup>1,2</sup> and Volker P. Brendel<sup>1,2</sup>

<sup>1</sup>Dept. of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, 50011, U.S.A.; <sup>2</sup>Dept. of Statistics, Iowa State University, Ames, IA, 50011, U.S.A

The recognition and prediction of *cis*-regulatory elements affecting gene expression remains a challenging problem for computational biology. Transcription factor binding sites (TFBS) typically exhibit some sequence variability at each position over a short (5-15 nucleotide) segment and often function in flexible distance relative to the transcription start site (TSS). These features intrinsically lead to high false-positive rates of prediction in practice. Here, we present a novel method for genome-wide detection of TFBS using multivariate data structures. In preliminary studies, three factors were considered in the model: site position-specific mononucleotide preferences, site position-specific dinucleotide preferences, and motif distance relative to the TSS. Training of models was based on known protein sequences that include a particular motif (positive set) and randomized sequences as a negative set. A novel aspect of this study concerns the generation of the randomized sequences: non-motif containing promoter sequences were simulated by recombination and reorganization of promoter sequences not known to contain the motif. The randomization procedure preserves mono- and dinucleotide composition but breaks higher order composition and spatial motif preferences. The three factors were combined into a prediction rule using logistic regression. Current work extends the model by including factors that reflect DNA deformability characteristic of DNA-protein interaction sites. For several well-characterized motifs, results indicate better prediction accuracy compared to existing motif prediction programs.

## De novo detection and qualification of regulatory motifs.

Kathleen Marchal<sup>1</sup>, Marleen Claeys<sup>1</sup>, Valerie Storms<sup>1</sup>, Kathleen Marchal<sup>1</sup>

<sup>1</sup> *CMPG-bioi, Department of Microbial and Molecular Systems KULeuven, Kasteelpark Arenberg 20, B-3001 Leuven, Belgium*

Identification of regulatory motifs is a prerequisite in the process of deciphering the regulatory network that controls gene expression. Computational methods have already proven to be an efficient tool. Continuous development of these methods for improved performance and broader applications is a must.

Regulatory motifs are characterized as conserved sites in a non-functional background. In motif detection methods, conservation is typically quantified by overrepresentation, or by an evolutionary relation to a common ancestor. We developed an algorithm PhyloMotifSampler that searches motifs in both spaces of conservation simultaneously. PhyloMotifSampler uses a probabilistic approach, a Gibbs Sampling strategy, an explicit model for motif evolution, and it scores potential regulatory motifs with an integrated dual-space conservation score. PhyloMotifSampler does not need sequences to be prealigned which makes it attractive for datasets where such alignment is unreliable.

De novo motif detection programs typically report a list of candidate regulatory motifs. Recent approaches aim to distinguish correct from incorrectly predicted motifs by assigning a posterior probability to each individual motif to be a true positive. We developed an algorithm FuzzyClustering that goes one step further and clusters not the individuals but pairs of candidate motifs jointly tracked during Gibbs Sampling, increasing the probability of the qualified motif to be part of a functional regulatory network. The clustering method is based on eigenwaarde computation and is an advanced integration of graph based clustering with motif detection. FuzzyClustering can qualify both regulatory motifs as well as regulatory modules. It can be integrated with one particular motif detection program or used independently as a postprocessing for the results of many different motif detection programs.

On a synthetic dataset of 10 genes with each 5 orthologous sequences in star topology and phylogenetic distances of 0.7 to root, the performance (as percentage of correctly predicted motif instances) was stepwards improved from 67% (MotifSampler, classic Gibbs Sampling search based on overrepresentation only) to 86% (PhyloMotifSampler) to 95% (PhyloMotifSampler + FuzzyClustering).

Conclusively, the two-dimensional picture (PhyloMotifSampler) and the motif qualification (FuzzyClustering) enables to study the evolutionary loss and the degree of conservation of regulatory motifs in a dual space.

# The automatic selection of TFBS score threshold in comparative genomics approach.

E.D. Stavrovskaya<sup>1,2</sup>, D.A. Rodionov<sup>3</sup>, A.A. Mironov<sup>1,2</sup>, I. Dubchak<sup>4</sup>, P. S. Novichkov<sup>4</sup>

<sup>1</sup>Department of Bioengineering and Bioinformatics, Moscow State University, Leninskiy Gory 1-73, Moscow, 119992, Russia; <sup>2</sup>Institute for Information Transmission Problems, Bol'shoi Karetnyi per. 19, Moscow, 127994, Russia; <sup>3</sup>Burnham Institute for Medical Research, La Jolla, CA 92037; <sup>4</sup>Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Reconstruction of transcriptional regulatory networks is one of the major challenges facing the bioinformatics community in view of constantly growing number of complete genomes. The comparative genomics approach has been successfully used for the analysis of the transcriptional regulation of many metabolic systems in various bacterial taxa. The key step in this approach is selection of an optimal site score threshold for the search of potential transcription factor binding sites (TFBS), given a position weight matrix. Here we demonstrate that this problem is tightly coupled with a problem of discovering the optimal content of regulon and suggest an approach to solve both problems simultaneously.

First, considering  $S^*$  being a potential optimal threshold, we calculate regulatory potential  $Z$  for each orthologous group (OG). A particular OG is described by two parameters: the number of orthologous genes  $N_i$ , and the length of upstream region  $L_i$  averaged by all genes in an OG. Applying the TFBS profile to upstream regions of all orthologs in a group, the number of genes  $K_i$  having a potential TFBS with score  $s \geq S^*$  can be calculated. We define the regulatory potential in terms of probability to observe the number of orthologous genes in a group, having at least one high scoring site, being  $K_i$  or greater, given that upstream regions are random sequences.

$$Z_i = -\log(P(k \geq K_i | N_i, L_i, S^*))$$

At the second step we utilize "Bernoulli Estimator" (BE) routine[1] which assumes that input values are a mixture from two distributions representing the noise and the signal. Only distribution describing the noise is required for automatic inference of the optimal threshold distinguishing the signal from the noise. Applying BE to regulatory potentials calculated for all OGs given  $S^*$ , the most probable content of a regulon can be automatically identified. At the same time, the probability to observe such regulon, given that upstream regions are random sequences (BE probability), is calculated. Considering all possible values of  $S^*$ , the optimal threshold delivering the minimum to BE probability, can be obtained.

The approach was tested on 7 *Shewanella* genomes using position-specific weight matrix of SOS response regulator LexA. The clear deep minimum of BE probability is achieved at the score threshold value 5.4, and selects 7 OGs as potential members of LexA regulon. The manual analysis of LexA regulon in *Shewanella* genomes reveals 15 OGs being the members of LexA regulon. The comparison of the seven found orthologous groups with results of manual analysis shows that all of them are true positives.

[1] Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci. 2004 Feb;13(2):443-56.

# Inference and validation of a large *cis* regulatory motif set using whole-genome *Saccharomyces* resequencing data

Matias M.T. Piipari<sup>1</sup>, Thomas A. Down<sup>2</sup>, Tim J.P. Hubbard<sup>1</sup>

<sup>1</sup> Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire (UK)

<sup>2</sup> Wellcome Trust/Cancer Research UK Gurdon Institute, Cambridge, Cambridgeshire (UK)

*Saccharomyces cerevisiae* is a good model organism for investigating eukaryotic *cis*-regulatory elements given its compact genome and the unique resources and datasets available for its study. These include whole-genome resequencing data for many of its known strains, large *in vivo* and *in vitro* transcription factor binding and gene expression data sets. We conducted a computational motif discovery study to infer a close to complete core promoter regulatory motif dictionary of the *S. cerevisiae* genome. 200 motifs were inferred in a single simultaneous experiment from a megabase of *S. cerevisiae* non-coding sequence (2000 x 500bp sequence windows upstream of genes) using NestedMICA.

The inferred motifs were analyzed in the context of the *Saccharomyces* Genome Resequencing Project dataset that provides an alignment of and SNP calls for over 80 *Saccharomyces* strains. We find evidence for selective constraint for a large proportion of the motifs: over 2/3 of them show SNP and insertion/deletion rates that are lower than the un-annotated non-coding regions of the *S. cerevisiae* genome, and a subset of 22 motifs show less variation than even protein coding sequence. We will present the correlation between motif hit score and the conservation of the motif amongst yeast strains, and between motif information content and its conservation. The association of motif hit conservation rate with distance from transcription start sites will also be discussed.

The NestedMICA motif set was also compared with reference motifs to identify known regulatory motifs from the set. Reference motifs included a recently published set of protein binding microarray based transcription factor motifs, motifs derived from the genome-wide *S. cerevisiae* ChIP-chip dataset, as well as TRANSFAC and JASPAR databases. Significant reciprocal matches are found for 81 predicted motifs, however 10 of the 22 putative novel regulatory motifs with low variation rates have no known close matches. This strongly suggests that the NestedMICA set contains additional functional motifs. The *S. cerevisiae* motifs were also compared with a NestedMICA derived *S. pombe* motif set to probe the overlap of regulatory motifs between these distantly related yeasts.

Our work demonstrates successful use of genome-scale NestedMICA motif inference in finding potential regulatory features of a eukaryotic genome. We provide strong evidence that our computationally inferred motifs are functional using independent lines of evidence. The analysis software developed for this study is publicly available. This includes an easy to use motif set visualization and analysis tool iMotifs (<http://www.github.com/mz2/imotifs>) as well as an analysis pipeline used for identifying significant motif matches and computing summary statistics from the re-sequencing dataset. These tools can be readily applied to other model organism re-sequencing and motif datasets to help improve our understanding of eukaryotic *cis*-regulatory elements.

# Large differences in transcription associated strand asymmetries of substitution patterns across metazoans

Paz Polak<sup>1</sup>, Robert Querfurth<sup>1</sup> and Peter F. Arndt<sup>1</sup>

<sup>1</sup>Max Planck Institute for Molecular Genetics, Ihnestr. 63 - 73, 14195 Berlin

Unraveling the evolutionary forces and variations of neutral substitution patterns among taxa and along genes (and their promoters) is a major prerequisite for detecting selection within these regions. In human, it has been shown that the interplay of transcription processes with repair introduces several biases in mutational patterns; in particular it invokes strand specific mutations. These strand asymmetries introduce strand specific distribution of words that can be explained primarily by biases in mutational processes rather by selection. Furthermore, the degree and the direction of these biases might vary between different lineages; a difference that further impacts tests that infer lineage specific adaptations. Therefore, in order to have a correct background substitution model in promoter regions and transcribed regions, one has to study mutation patterns associated with transcription across animals.

We establish a comparative analysis of three or more genomes that correctly handles effects due to the non-stationarity and irreversibility of the nucleotide substitution process. This way we are able to quantify the 12 different rates for exchanges of one nucleotide by another as well as the rate of the neighbor dependent CpG deamination process. Moreover, our maximum likelihood method estimates these rates along each internal branch of a given phylogeny tree.

Here, we used the multiple alignments of closely related species to quantify the nucleotide substitution rates around the 5' end of genes in 6 clades: primates, rodents, Laurasiatheria, fishes, flies and worms. Our analysis is focused on the intronic and intergenic regions and reveals extraordinary rich patterns of substitution asymmetries across species; in the majority of taxa at least one substitution type shows significant asymmetry in intronic sequences but none was common for all taxa. For example, in mammalian introns, the rates of A→G transition are greater from T→C rates along the non-template strand, whereas in cold-blooded species fishes, flies and worms we detected the reverse trend. We propose that the asymmetric patterns in transcribed regions are results of transcription associated mutagenic processes and transcription coupled repair, which both seem to evolve in a taxon related manner.

In mammals, the levels of asymmetries are stronger within genes starting within CpG islands (CGI) than in genes lacking this property. Interestingly, in intergenic (putative) promoter regions of CGI-related genes there are also strand asymmetries which are opposite to the one in intronic regions; we hypothesize that these patterns are evidence that in all mammals a significant number of CGIs are origins for antisense and sense transcription. Finally, we show that expression levels in the human/mouse germline and the degree of strand asymmetries in transcripts are correlated, which demonstrates that certain phenotypes drive mutations in mammalia.

# Supervised learning approaches to predicting enhancer regions and transcription factor binding sites in *D. melanogaster*

Rachel Sealfon<sup>1,2</sup>, Chris Bristow<sup>1,2</sup>, Pouya Kheradpour<sup>1,2</sup>, Manolis Kellis<sup>1,2</sup>

<sup>1</sup>Computer Science and Artificial Intelligence Laboratory, MIT; <sup>2</sup>Broad Institute

New experimental datasets, such as genome-wide profiling of chromatin marks and transcription factor binding, offer the potential for gaining insight into the combinatorial code of gene regulation. Previous computational work has focused on motif identification and some initial work on identifying combinations of factors that drive specific patterns of expression. However, the relative roles played by enhancers, transcription factors, chromatin marks, and individual motifs in driving gene expression remain unclear. Towards this goal, we used a machine learning approach to study the predictive power of experimental and sequence-based combinations of features in the context of both enhancer prediction and prediction of transcription factor binding sites.

Although general and sequence-specific transcriptional regulators can in theory bind anywhere in the genome, numerous studies suggest that their binding is in fact clustered within relatively small regions, known as enhancers or cis-regulatory modules. These are portions of DNA that interact with transcription factors to regulate a modular portion of the spatiotemporal expression pattern of a gene. Enhancers have been defined experimentally by their ability to drive tissue-specific gene expression in transgenic studies, and computationally predicted by their increased sequence conservation in multiple related species, and also by their abundance of regulatory motif instances.

We developed an integrative approach to enhancer prediction that leverages the wealth of available experimental data on chromatin marks and transcription factor binding. Using a supervised learning framework, we identified combinations of ChIP-chip protein binding data, chromatin-marks, chromatin-associated factors, and sequence conservation features that are characteristic of the experimentally validated enhancers in the REDFly database. We found that while chromatin marks alone had low predictive power, including chromatin mark features as well as transcription factor binding features dramatically improved the power of our classifier. The improvement in classifier performance using combinations of types of features relative to any individual feature type suggests that each class of functional elements plays distinct yet necessary roles in defining enhancer regions in the cell.

We have also applied supervised learning methods for predicting transcription factor binding locations based on combinations of regulatory motifs. For each experiment in a compendium of ChIP-chip studies, we constructed a classifier to distinguish between regions bound by the given factor and regions bound by any other factor. For each factor, we compared the performance of the known motif for the factor, the most enriched motif reported by motif discovery algorithms, enrichment of the top 5 most enriched motifs, and depletion of the 5 least enriched motifs. While the results differ across factors, we found that combinations of features typically outperformed individual motifs, and predictive power increased when depleted motifs were included as features. This result suggests that binding of an individual transcription factor at a given site may be highly dependent on the local combination of bound factors, which provide both synergistic and antagonistic influences.

# The Effect of Orthology and Coregulation on Detecting Regulatory Motifs

Valerie Storms<sup>1</sup>, Marleen Claeys<sup>1</sup>, Aminael Sanchez<sup>1,2</sup>, Bart De Moor<sup>3</sup>, Annemieke Verstuyf<sup>4</sup>, Kathleen Marchal<sup>1</sup>

<sup>1</sup>*CMPG, Department of Microbial and Molecular Systems, K.U.Leuven, Leuven 3001, Belgium;* <sup>2</sup>*Laboratory of Molecular Biology, Institute of Plant Biotechnology, Central University 'Marta Abreu' of Las Villas (UCLV), Santa Clara 54830, Cuba;* <sup>3</sup>*Department of Electrical Engineering ESAT-SCD, K.U.Leuven, Leuven 3001, Belgium;* <sup>4</sup>*Departement Experimentele Geneeskunde, LEGENDO, K.U.Leuven, Leuven 3000, Belgium*

Despite the enormous collection of tools available for the discovery of transcription factor binding sites (motifs), the detection of motifs remains a challenging problem. Initially motif discovery was performed on a set of genes, known to be coregulated and thus containing binding sites for the same transcription factor. But with the growing number of sequenced genomes, detecting motifs through 'phylogenetic footprinting' became feasible and the next generation of motif discovery tools has therefore integrated the use of orthology in addition to the coregulation information. While these tools treat the orthologous sequences as independently, and thus ignore the underlying phylogeny that describes their relatedness, the most advanced motif discovery tools explicitly incorporate these phylogenetic relations by means of an evolutionary model.

Despite the potential of comparative data in addition to coregulated data, so far no independent study has evaluated the extent of information contained within either the coregulation or the orthologous space, and the conditions under which complementing both spaces becomes useful. In this study we performed such analysis by 1) designing appropriate datasets (real and synthetic) and 2) applying two of the more advanced probabilistic methods for motif discovery: Phylogibbs<sup>[1]</sup> and Phylogenetic sampler<sup>[2]</sup>, which both model the evolutionary dependency in the orthologous space. To set the base line performance we included MEME<sup>[3]</sup>, as a representative of algorithms that cannot explicitly incorporate phylogenetic relations.

Our results show that the success rate of combining coregulation and orthology information depends on the complex relation between the algorithm and the dataset. The performed tests illustrate that the nature of the used algorithm is crucial in determining how to exploit multiple species data in the best way to improve motif discovery performance. The results of this unbiased comparison also point out the strengths and weaknesses of current implementations, information which is useful for both developers and users.

<sup>[1]</sup>Siddharthan R, Siggia ED, van Nimwegen E (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 1: e67; <sup>[2]</sup>Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, et al. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics* 23: 1718-1727; <sup>[3]</sup>Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2: 28-36.

# Global DNase Hypersensitivity Mapping Reveals Growth Hormone (GH)-regulated Sex Differences in Mouse Liver Chromatin Structure

Aarathi Sugathan<sup>1</sup>, Guoyu Ling<sup>1</sup>, Tali Mazor<sup>2</sup>, Ernest Fraenkel<sup>2</sup>, David J. Waxman<sup>1</sup>.

<sup>1</sup>Department of Biology, Boston University, Boston, MA; <sup>2</sup>Department of Biological Engineering, Massachusetts Inst. of Technology, Cambridge, MA

Global expression profiles have identified >1,000 genes that are sex-differentially expressed in mouse liver, impacting numerous physiological and pathophysiological processes. Sex-dependent plasma GH patterns regulate these genes, with male-specific genes suppressed and female-specific genes induced in livers of male mice given GH by continuous infusion (female-like GH pattern). Changes in chromatin structure are a hallmark of epigenetic regulation and developmental plasticity, and can be characterized at a genome-wide level by DNase I hypersensitivity (DH) site analysis in combination with high-throughput sequencing. Using this strategy, we investigated the hypothesis that sex differences in local chromatin structure characterize sex-specific genes. Global DH maps were generated for male and female mouse liver, and for livers of male mice given a 7-day GH infusion. High-resolution DH maps based on 55 million sequence tags of DNase I-released fragments identified ~43,000 high confidence DH sites, which were ~600 nt wide (mean value), corresponding to the depletion of ~3 nucleosomes. To compare DH sites between biological samples, samples were scaled by the number of sequence tags in putative peak regions for each pairwise comparison. In this way, 1,284 DH sites were identified as robustly sex-specific. Initial analysis localized ~80% of sex-independent DH sites to the coding region or within 5 kb of known genes or ESTs, while sex-specific DH sites were more likely to be intergenic, with ~40% found >5 kb from known genes. Many sex-specific genes were associated with DH sites showing the same sex-specificity. Moreover, sex-specific DH sites were >10-fold enriched for sites being nearest to a corresponding sex-specific gene, as compared to sex-independent DH sites. However, 6-10% of the sex-specific DH sites were >500 kb from the nearest known sex-specific gene. Continuous GH infusion of males abolished liver sex-specificity by suppressing 83% of male-specific DH sites and by inducing 26% of female-specific DH sites. In many cases, sequencing density was sufficiently high to visualize putative transcription factor binding sites as footprints within DH regions. In some cases extended sex-specific DH regions up to 100 kb in length were observed. Thus, sex-specific gene expression is associated with sex-differences in chromatin structure, with sex-specific DH sites being common in mouse liver, responsive to GH treatment, and likely to contain regulatory sequences that mediate the sex-specific actions of GH on liver gene expression. Supported in part by NIH grant DK33765 (DJW).

# Function conservation in diverged noncoding elements

Leila Taher<sup>1</sup>, David McGaughey<sup>2</sup>, Andrew McCallion<sup>2</sup>, Ivan Ovcharenko<sup>1</sup>

<sup>1</sup> Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA. <sup>2</sup> McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins Medical Institute, Baltimore, United States of America

Transcriptional regulatory elements show a highly modular organization, and consist of a variable number of degenerate binding sites for several transcription factors. Mutations in regulatory elements outside of active binding sites or those that do not deactivate transcription factor binding are likely to have little or no impact on the function of these elements. Due to the circumscribed impact of mutations on the function of regulatory elements, regulatory sequences often diverge extensively while retaining their ancestral functions. Therefore, cross-species sequence comparison – the primary method to identify regulatory elements in metazoan genomes – often has limitations in detecting functionally conserved elements that lack sufficient sequence similarity.

This work explores the structure of regulatory information encoded in the distribution of TFBS in absence of sequence similarity. We modeled noncoding sequences as arrangements of transcription factor binding sites (TFBSs), and compared noncoding regions using alignments between sequences of TFBSs instead of nucleotide alignments. Additionally, our alignment model contemplates evolutionary events in regulatory sequences, e.g., matches, mismatches, and duplications of TFBSs. To train our model, we developed a strategy that reconstructs the TFBS structure of diverged sequences based on phylogenetic relationships among groups of species. Thereby, we constructed a set of 2,500 noncoding elements that have diverged between human and zebrafish but are likely to share the same ancestry. Our method correctly detects ancestral identity for over 50% of these elements embedded into 50kb stretches of background DNA. We evaluated the significance of the alignment scores by comparing true orthologs to sequences with similar GC-content, and enforced an almost zero false positive rate of predictions. Applying our method to a selected set of 3,000 human loci, we predicted approximately 400 pairs of sequences that are very likely to share a common ancestor and have preserved their function despite not being conserved between human and zebrafish.

We validated predicted zebrafish elements with reporter-gene assays in transgenic zebrafish, observing *in vivo* enhancer activity in 86% of elements (6/7). Moreover, the human and zebrafish putative functional orthologs directed highly overlapping tissue-specific expression patterns.

The present study constitutes the first genome-wide computational attempt to establish the ancestral identity of groups of diverged noncoding elements and verify the preservation of their function.

## Genome-wide prediction of transcription factor binding sites using chromatin modification

Kyoung-Jae Won<sup>1</sup>, Bing-Ren<sup>2</sup>, Wei Wang<sup>1</sup>

<sup>1</sup>University of California, San Diego, Dept. Chem & Biochem, 9500 Gilman Drive, La Jolla, CA 92093, USA <sup>2</sup>Ludwig Institute for Cancer Research and Department of Cellular and Molecular Medicine, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA

Identification of target loci of transcription factors (TFs) in a specific tissue or at a specific developmental stage is crucial for understanding transcriptional regulation in eukaryotes. However, genome-wide prediction of TF binding sites in eukaryotes has suffered from limited information to find functional elements in a genome.

We present here an integrated approach, called Chromia, that integrates discrete information (sequence) and continuous information (chromatin modification information) to tackle this task. Chromia is composed of a hidden Markov model (HMM) that integrates information from multiple sources. Using 8 histone signatures and a motif score as input, Chromia captures the characteristic patterns of a TF binding motif occurrences and the histone modification signature associated with regulatory elements such as promoters and enhancers.

We demonstrated its usefulness on genome-wide predictions of target loci of 13 TFs in the mouse embryonic stem (mES) cell. Using the independent ChIP-seq analyses of these bindings as the gold standard, we showed that the performance of our HMM model was significantly better than many other computational methods. Particularly, Chromia predicts genome-wide binding sites in enhancers. In the cross-validation test Chromia showed superior performance to the baseline TF identifier using histone signature, showing the advantages of using integrated approach based on HMMs for this problem. Also, Chromia identified more binding targets of TFs of the genes affected by the RNAi experiments than ChIP-seq experiments.

The encouraging results of this study suggest Chromia as a novel approach of predicting condition specific TF binding sites at a genomic scale using epigenetic information.

# Metazoan operons accelerate transcription and recovery rates

Alon Zaslaver<sup>1</sup>, L. Ryan Baugh<sup>2</sup>, Paul Sternberg<sup>1</sup>

<sup>1</sup> Howard Hughes Medical Institute and California Institute of Technology, Division of Biology, 1200 E. California Blvd., Pasadena, California 91125; <sup>2</sup> Department of Biology and Center for Systems Biology, Duke University, Durham, North Carolina, USA.

Existing theories efficiently explain why operons are advantageous in prokaryotes, but their emergence in metazoans is still an enigma. We present a combination of genomic meta-analysis, experiment and theory to explain how operons could be adaptive during metazoan evolution. Focusing first on nematodes, we show that operon genes, typically consisted of growth genes, are significantly up-regulated during recovery from multiple growth-arrested states, and that this expression pattern is anti-correlated to the expression pattern of non-operon genes. In addition, we find that transcriptional resources are initially limited during arrest recovery, and that recovering animals are extremely sensitive to any additional limitation in transcriptional resources. By clustering growth genes into operons, fewer promoters compete for limited transcriptional machinery, effectively increasing the concentration of transcriptional resources and accelerating growth during recovery. A simple mathematical model of transcription dynamics reveals how a moderate increase in transcriptional resources can lead to a substantial enhancement in transcription rate and recovery. We find evidence for this design principle in different nematodes as well as in the chordate *C. intestinalis*. As recovery from a growth arrested state into a fast growing state is a physiological feature shared by many metazoans, operons could evolve as an evolutionary solution to facilitate these processes.

## Determinants of Transcription Factor Binding and Regulation

Xu Zhou<sup>1</sup>, Erin O'Shea<sup>1,2</sup>

<sup>1</sup>Department of Molecular and Cellular Biology; <sup>2</sup>Howard Hughes Medical Institute, Harvard University Faculty of Arts and Sciences Center for Systems Biology, Department of Chemistry and Chemical Biology, Northwest Labs, 52 Oxford Street, Cambridge, MA, 02138, USA.

Specific gene expression programs play an important role in regulating many essential biological processes. This regulation is conducted by transcription factors (TFs) whose binding typically requires recognition of a specific DNA binding motif. However, high affinity sequence motifs by themselves are not sufficient to predict binding events, even for a simple organism such as *Saccharomyces cerevisiae*. To complicate matters, binding of a TF is necessary but not sufficient for transcriptional regulation. Despite a myriad of studies on transcription mechanisms, it is still unclear what determines the genomic locations to which a TF is bound and what determines whether this binding is functional (i.e., influences the transcription of a gene). Answering these questions will help us to understand how specific transcriptional regulation is achieved.

The phosphate responsive pathway (*PHO*) in *S. cerevisiae* has been widely used as a model system to study transcriptional activation and nucleosome remodeling. The transcriptional activation of the *PHO* pathway is regulated by two TFs, Pho2 and Pho4. When cells are grown in phosphate (Pi) limited media, the expression profile of a set of genes is specifically regulated, and appear to have diverse dependence on Pho2 and Pho4. Here we use *PHO* pathway as a model system to explore the determinant factors for TF binding and its regulation. We hypothesize that the sequence of a DNA binding motif, the presence or absence of cofactors, and chromatin structure determine TF binding and its influence on gene expression.

We first identified genome-wide *in vivo* binding sites for both Pho2 and Pho4 using modified biotin-tagging chromatin immunoprecipitation followed by high throughput parallel sequencing. Second, we determined genome-wide nucleosome positions and occupancy under both Pi rich and limited conditions with the same sequencing technique. Third, we used expression epistasis analysis to quantify the role of Pho2 and Pho4 in regulating gene expression. We will integrate the information regarding TF binding motif and nucleosome structure to explain TF binding, as well as the information of TF binding and nucleosome structure to explain transcriptional regulation.

# Learning gene regulatory networks with delayed ODEs and continuous-time expression representation

Tarmo Äijö<sup>1</sup>, Harri Lähdesmäki<sup>1,2</sup>

<sup>1</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>2</sup>Department of Information and Computer Science, Helsinki University of Technology, Helsinki, Finland

Regulation of gene expression is fundamental to the operation of a cell. Revealing the structure and dynamics of a gene regulatory network (GRN) from gene expression measurements is of great interest and represents a considerably challenging computational problem. In addition to the difficulties arising from stochasticity in the underlying biological phenomena and considerable measurement noise, many of the existing methods have fundamental limitations, such as difficulties to cope with non-uniform measurement intervals and delays. For example, methods such as Bayesian networks are based on the assumption of equally spaced measurements. On the other hand, ordinary differential equation (ODE) based methods with the first-order approximation of the derivative might have difficulties with long gaps between measurements. The current knowledge on biology suggests that the time delays are significant in gene regulation. In this work, we propose solutions for these problems and analyze their effectiveness.

Here we extend our previously published method [1] by incorporating delays into analysis and by replacing the first-order order approximation of the derivative with an analytic estimate from a continuous expression profile, obtained from a Gaussian process (GP) representation. Delays in gene expression modeling are naturally incorporated into the Bayesian framework. The aforementioned method is based on the use of Bayesian analysis with ODEs and non-parametric GP modeling for the transcriptional level regulation. The main differences between our method and the existing ODE based methods are non-parametric modeling of molecular kinetics and Bayesian analysis.

For the validation we use a recently published *in vivo* data set [2]. The validation is done based on the structure learning and predictive capabilities of the method. The obtained results demonstrate that our method provides more accurate network structure learning and it is able to predict the dynamics of the studied GRN. Our results also show that the model extensions mentioned above have a positive effect on the prediction and structure inference compared to the results of existing methods. In particular, our results demonstrate the important role of delays in gene regulation modeling.

In addition, we will shortly review our approach to the predictive signaling network modeling challenge (DREAM 4, Challenge 3) in the poster. Our approach to this challenge is quite similar to the method presented in this abstract.

[1] T. Äijö and H. Lähdesmäki. (2009) Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, in press.

[2] I. Cantone, et al. (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1), 172 – 181.

## Comparison of gene expression time courses between light entrained and temperature entrained *drosophila* flies reveal genes which peak two times a day.

Arnaud Amzallag<sup>1,2</sup>, Herman Wijnen<sup>3</sup>, Félix Naef<sup>1,2</sup>

<sup>1</sup>IBI, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; <sup>2</sup> Swiss Institute of Bioinformatics, Lausanne, Switzerland; <sup>3</sup>University of Virginia, Charlottesville, VA, USA

The circadian clock is an endogenous and autonomous biological clock, but it is virtually always entrained (*i.e.* synchronized) by environmental clues (called *zeitgebers*) such as daylight and temperature variations. In order to probe the importance of the nature of the *zeitgeber* in the genome wide circadian response, we compare microarray data measured every four hours during two days in *drosophila* heads, which were entrained either exclusively by light or exclusively by temperature, in wild type and clock mutant flies. Using simple correlation estimates, we find that apart from a group of known circadian genes, several other genes seem to respond with a coherent phase to the different *zeitgebers*: notably, some genes seem to peak two times a day in both conditions, a pattern that was not described in the fly model before. We confirm this tendency with independent data from another team (for the light/dark entrainment). The group of genes span several diverse gene ontologies.

The “two peaks per day” pattern could be due to antiphase circadian responses between different cell types in the head or to transcription factors with two peaks per day. To address this additional question, we try to infer the time-of-the-day dependent activity of DNA binding sites in the promoters of the genes of interest, through regression of the gene expressions on TRANSFAC motif counts in the promoters of interest.

## Evidence for Quantitative Transcription Networks

Biggin, M.D.<sup>1</sup>, Kaplan, T.<sup>1</sup>, Aswani, A.<sup>1</sup>, Li, X.Y.<sup>1</sup>, Thomas S.<sup>2</sup>, Sabo, P.<sup>2</sup>, Brown, J.B.<sup>1</sup>, Boley, N.<sup>1</sup>, Atherton, J.<sup>1</sup>, Li, J.<sup>1</sup>, Davidson, S.M.<sup>1</sup>, Fisher, B.<sup>1</sup>, Hammonds, A.<sup>1</sup>, MacArthur, S.<sup>1</sup>, Fowlkes, C.C.<sup>1</sup>, Luengo Hendriks, C.L.<sup>1</sup>, Keränen, S.V.E.<sup>1</sup>, Hechmer, A.<sup>1</sup>, Simirenko, L.<sup>1</sup>, Malik, J.<sup>1</sup>, Knowles, D.W.<sup>1</sup>, Tomlin, C.<sup>1</sup>, Bickel, P.<sup>1</sup>, Stamatoyannopoulos, J.<sup>2</sup>, Celniker, S.<sup>1</sup>, Eisen, M.B.<sup>1</sup>

<sup>1</sup>Berkeley *Drosophila* Transcription Network Project, Lawrence Berkeley National Laboratory and UC Berkeley, Berkeley, California 94720, USA

<sup>2</sup>University of Washington, Department of Genome Sciences, Seattle, WA 98195, USA

The Berkeley *Drosophila* Transcription Network Project (<http://bdtnp.lbl.gov/Fly-Net/>) is developing wet laboratory and computational/mathematical methods to allow predictive modeling of animal transcription networks, using the early *Drosophila* embryo as a test case. System wide data sets for *in vitro* and *in vivo* DNA binding, 3D cellular resolution protein and mRNA expression, transgenic promoter expression, and DNA accessibility in chromatin have been established and are being analyzed in conjunction with available comparative genomic DNA sequence information. Our work suggests that transcription networks have an unexpected structure in which functionally distinct transcription factors show a quantitative continuum of binding to highly overlapping sets of thousands of genomic regions. Highly bound genes include strongly regulated known and likely targets, moderately bound genes include unexpected targets whose transcription is regulated weakly, and poorly bound genes include thousands of non-transcribed genes and other likely non-functional targets. Quantitative differences in binding to common targets correlate with each factors known regulatory specificity, though these specificities appear to be more fuzzy and less distinct than commonly assumed. We propose that this Quantitative Transcription Network structure will be general to all metazoans and derives from the broad DNA recognition properties of animal transcription factors and the relatively high concentrations at which they are expressed in cells causing them to bind to highly overlapping sets of open chromatin regions. Predictive computational models of DNA occupancy *in vivo* and the regulation of target gene expression will also be presented.

RG Posters

# Temporal Dynamics of Regulatory Networks in *Drosophila melanogaster* Embryogenesis

Rogério Candeias<sup>1-3</sup>, Manolis Kellis<sup>2,3</sup>

<sup>1</sup>PhD Program in Computational Biology of Instituto Gulbenkian de Ciência, Oeiras, Portugal; <sup>2</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; <sup>3</sup>Broad Institute, Cambridge, Massachusetts 02141, USA.

Dynamic modeling of regulatory networks that control gene expression requires temporal information on the activation/repression latencies of regulator and target pairs, which have been experimentally inaccessible at the genome scale. We developed a new discretization method using multi-step functions, to systematically infer latency information for individual edges of a large-scale regulatory network from whole-genome time-course expression profiles. Our method has wider applicability and shows increased accuracy relative to previous approaches such as Pearson or Spearman correlation. It also exhibits good predictive power of expression co-localization for regulator/target pairs benchmarked against the ImaGO annotation for *Drosophila melanogaster*.

Application of this method to *D. melanogaster* led to several new insights on its network dynamics. First, the measured delays are significantly longer than expected by chance, and are specific to *D. melanogaster*. They are not found in yeast, suggesting that they are likely relevant to animal genomes and developmental processes. Second, we found that regulator binding site multiplicity on the target promoter region is related to an increased latency, which is consistent with a slower activation associated with protein accumulation. Third, regulators of the same functional category were more likely to show similar delay distributions, suggesting different time-scales may be at play for different biological processes. Fourth, the combinatorial sum of multiple regulators is able to better explain the target expression profile than expected. Lastly, network motifs such as transcription cascades and feed-forward loops showed characteristic time delay distributions, suggesting both connectivity and dynamics contribute to the function of networks motifs.

Overall, temporal information has the potential to fundamentally change the way we think about gene regulatory networks, and the dynamic network of 21,231 temporally-decorated edges provided here enables the study of information flow and developmental dynamics at the systems-level.

# Investigating Co-regulation Networks Using Generative Models

Matthew B. Carson<sup>1</sup>, Nitin Bhardwaj<sup>2</sup>, Hui Lu<sup>1</sup>

<sup>1</sup>*Bioinformatics Program, University of Illinois at Chicago;* <sup>2</sup>*Program in Computational Biology and Bioinformatics, Yale University*

Proteins, often referred to as the ‘workhorses of the cell’, are produced through the process of gene expression, during which an organism turns its genetic code (DNA) into functional units. Regulation of this expression process increases the versatility of an organism, allows for adaptation to the environment, and increases the efficiency with which resources such as sugars are metabolized by controlling when and in what quantities RNA molecules and proteins are produced. Many diseases are related to failures in one or more components of this system. Examining regulation helps us to understand how an organism evolves and develops, and how malfunctions may break down this process. Much of the control of gene expression is believed to occur by the cell’s adjustment of transcription initiation frequency. This level of control is carried out by transcription factors (TFs), and transcription factor networks (TFNs) can be used to describe the interactions between these transcription factors and their target genes.

In this work we use generative networks to model the creation of TFNs during evolution in order to understand how these networks form and develop. In particular, we examine how the number of TF partners (those that regulate common genes) scales with the number of genes a TF regulates. It has been observed that in several model organisms the degree distribution of this partnership network appears to follow an exponential saturation curve. The co-regulatory network of our generative model shows a trend similar to that of the model organisms. We show that through various modifications to our model we are able to identify the necessary conditions for this observed saturation curve characteristic. This indicates that the saturation curve seen in these co-regulatory networks could be a product of evolutionary development, during which regulators gain and lose interactions with targets over time.

RG Posters

# Reverse Engineering of Gene Regulation Network from DREAM4 Data

Hyonho Chun<sup>1</sup>, Minghua Deng<sup>1,5</sup>, Jia Kang<sup>3</sup>, Haisu Ma<sup>4</sup>, Xianghua Zhang<sup>1,6</sup>, Hongyu Zhao<sup>1,2</sup>

<sup>1</sup>Department of Epidemiology and Public Health, Yale University; <sup>2</sup>Department of Genetics, Yale University; <sup>3</sup>Program in Translational Informatics, Yale University; <sup>4</sup>Program in Computational Biology, Yale University; <sup>5</sup>School of Mathematical Sciences and Center for Theoretical Biology, Peking University; <sup>6</sup>Department of Electronic Science and Technology, University of Science and Technology of China;

## Background:

The objective of in-silico network challenge is to infer gene regulation networks from the provided simulated steady-state and time series data. In order to reasonably reconstruct relationships among genes in the presence of noise, one critical task is to reliably estimate the mean and variance associated with the unobserved true wild type expression level for each gene in order to identify genes with different expression levels across different experimental conditions. One common assumption made in parameter estimation based on high-dimensional genomics data is the sparsity of regulatory signals. However, based on our empirical observations in the provided datasets, the sparsity assumption may not appropriately characterize the nature of the underlying network. And for this reason, using knock-out and knock-down data (which are often perceived as the most informative datasets in inferring network connectivity) to obtain mean and variance estimates may lead to incorrect inference of the network structure.

## Method:

We develop a novel method to estimate mean and variance associated with the wild type expression using time-series data; utilizing this information, network structure can then be subsequently deduced from the knockout and knockdown data. We apply our method to the sub-challenge 2 (InSilico\_Size10) dataset.

## Results:

Our method yields networks that are capable of well characterizing the pattern of variation across different datasets. Furthermore, in datasets where the underlying networks are conspicuously non-sparse, we expect our method to outperform existing approaches that are built upon the sparsity assumption.

# Classification Trees Can Describe and Predict Conditional Transcription Factor Binding *in vivo*

Matthew Davis<sup>1</sup>, Michael Eisen<sup>1-4</sup>

<sup>1</sup> Department of Molecular and Cell Biology, University of California, Berkeley, CA; <sup>2</sup> Howard Hughes Medical Institute, University of California, Berkeley, CA; <sup>3</sup> Genomics Division, Ernest Orlando Lawrence Berkeley National Lab, Berkeley, CA. <sup>4</sup> California Institute for Quantitative Biosciences, University of California, Berkeley, CA

Transcription factors recognize signals specified by nucleotide sequences to bind and regulate the expression of genes. While several biochemical methods have identified recognition motifs for these signals, simple combinatorial models of sequence signal fail to predict the binding of transcription factors to their target nucleic acids *in vivo*. A piecewise model accounting for the variation in binding by multiple factors, chromatin accessibility, and multiple binding site target profiles can account for a significant amount of the variance observed. Using the model system of anterior-posterior pattern specification in *Drosophila melanogaster* embryos, we have collected genome wide transcription factor binding data for many of the factors known to regulate gene expression in this system. Using principal components analysis, we have previously shown both factor-dependent and factor-independent components of combinatorial binding events across the entire genome. To learn rules that describe the conditional interactions between factor binding events in a given region of DNA, we have employed a classification tree algorithm to partition all regions of bound DNA in our dataset. The dependencies learned by this method have isolated contiguous stretches of nucleotides bound by different combinations of factors. The results are sets of binary rules for predicting the binding strength of one factor, given the binding of the other factors. The predictive power of these rules are dependent on the DNA being accessible to transcription factors. These contiguous regions of bound DNA contain enrichments of various oligonucleotide motifs that may represent genuine binding sites. While the generation of this descriptive model has provided insights into the combinatorial regulation of one system in one species, we are extending this method to assess the evolution and robustness of these rule sets by comparison to other *Drosophila* species.

## Identification of *cis*-regulatory modules in homologous sequences

Norbert Dojer

*University of Warsaw*

The process of transcriptional regulation is facilitated by proteins called transcription factors which bind to DNA sequences to help or prevent the initiation of transcription by RNA polymerase. DNA fragments with binding sites for a group of commonly acting transcription factors, driving complex expression patterns, are often referred to as *cis*-regulatory modules (CRMs).

We provide a novel method for identification of CRMs in DNA sequences based on phylogenetic footprinting and using a database of motifs of transcription factor binding sites. As opposite to classical methods we assume evolutionary conservation of transcription factor binding sites rather than of DNA sequence itself. Moreover, we do not impose the necessity of strict preservation of binding sites order across the species. The quality of our method is validated on a set of experimentally verified mammalian CRMs.

# Learning recurrent mRNA expression patterns from systematic analysis of *in-situ* images of *Drosophila* embryos

Charlie Frogner<sup>1</sup>, Christopher A. Bristow<sup>2,3</sup>, Tom Morgan<sup>2</sup>, Lorenzo Rosasco<sup>1</sup>, Pouya Kheradpour<sup>2</sup>, Rachel Sealfon<sup>2</sup>, Tomaso Poggio<sup>1</sup>, Manolis Kellis<sup>2,3</sup>  
<sup>1</sup>Center for Biological and Computational Learning, McGovern Institute, MIT, Cambridge, MA; <sup>2</sup>Department of Computer Science and Artificial Intelligence, MIT, Cambridge, MA; <sup>3</sup>Broad Institute, Cambridge, MA

Understanding the spatio-temporal control of gene expression during development is one of the major challenges in genomics. In order to dissect the regulatory constructs responsible for defining gene expression programs, we are combining image analysis of gene expression patterns in *Drosophila* embryos with sequence analysis of genomic regulatory regions and functional genomic data. Our initial dataset consists of ~75,000 images of *Drosophila* embryos from ~6,000 genes profiled at 5 stage ranges of embryogenesis. Whereas previous research on *Drosophila* development has relied on human curation of these expression patterns, we have sought to apply computer vision approaches to analyze the data. We have developed an image processing pipeline to segment the embryo images and extract the pixels that correspond to mRNA staining, incorporating novel approaches to embryo segmentation that reliably deal with common issues, such as multiple overlapping embryos. We have also devised a new method for extracting detailed stain patterns from the images, based on a supervised learning approach trained on ubiquitously expressed genes. Together, these methods allow us to process over 80% of the images, and can automatically extract mRNA expression patterns for many thousands of embryos, with minimal human inspection.

We used these algorithms to extract a robust representation of recurrent expression patterns in a systematic and unbiased way. We then clustered the extracted mRNA stain patterns according to spatial similarity, to assemble groups of genes that show coherent expression patterns. Strikingly, these clusters recover specific organ systems independently annotated by human curators, and in some cases suggest meaningful subdivisions of annotation terms, as well as new patterns that are not easily captured by existing annotation terms. The inferred clusters also show specific enrichments in known regulatory motifs, as well as binding data associated with transcription factors involved in embryo development, and suggest specific regulatory connections to candidate regulators for these recurrent patterns.

While our initial results for early embryogenesis are very encouraging, there are several challenges going forward as we extend this approach to multiple time points and stages with more complex gene expression patterns.

Systematic image analysis of large-scale gene expression datasets, coupled with genome sequence analysis and large-scale functional data, provides a general way to define common regulatory programs in animal genomes, by discovering genes that have coherent upstream regulation (e.g. transcription factor binding, chromatin marks, and motif instances) and coherent downstream expression patterns (based on *in-situ* image analysis). The approach presented here is scalable and robust, and should apply more generally to any species.

# Inferring Topology and Dynamical Properties of Genome-wide Regulatory Networks

Alex Greenfield<sup>2</sup>, Aviv Madar<sup>1</sup>, Harry Ostrer<sup>4</sup>, Eric Vanden-Eijnden<sup>3</sup>, Richard Bonneau<sup>1,3</sup>

<sup>1</sup>Center for Genomics and Systems Biology, New York University, <sup>2</sup>Computational Biology Program, New York University, <sup>3</sup>Courant Institute of Mathematical Sciences, New York University, <sup>4</sup>Human Genetics, New York University School of Medicine

Learning and characterizing regulatory networks, responsible for the remarkable ability of organisms to adapt to changing environment is a key problem in modern biology with applications spanning bioengineering, drug development, and many other biological fields. Detailed regulatory networks (RNs) can be modeled as a system of ordinary differential equations (ODEs), describing the rate of change in mRNA concentrations as a function of relevant predictors (e.g. transcription factors). We have recently described a network reconstruction algorithm, Inferelator-1, which infers regulatory influences for genes and gene clusters. The typical input is: 1) a microarray compendium composed of time-series and equilibrium measurements, and 2) prior information, such as a list of considered predictors. The output is a sparse dynamical model for each gene or gene cluster, i.e. an ODE describing the rate of change in transcription as a function of just a few predictors. We have shown that RNs learned using Inferelator-1, at least for small genomes (e.g. *halobacterium*), could explain observed mRNA measurements (training-set), as well as predict unobserved mRNA measurements (a large test-set). Inferelator-1, however; 1) approximates the predictors level to be constant throughout a time interval, which becomes a crude approximation as the time-interval length increases, 2) solves the system of ODEs as an uncoupled system which is not a realistic model for the underlying RN, and 3) produces only one set of ODEs, describing a single RN without an associated confidence estimate. These limitations diminish our ability to model more complex systems over long time intervals.

Here, we use resampling to create an ensemble of datasets as input for Inferelator-1. Thus, producing an ensemble of RNs as output. Using this ensemble we derive empirical distributions for many putative regulatory interactions together with their corresponding kinetic-parameters. To take advantage of these empirical distributions we have developed Inferelator-2—a Bayesian, dynamical modeling approach. Inferelator-2 searches for regulatory interactions and kinetic-parameters, defining a set of ODEs, then maximizes the probability of this ODE set given the observations (the posterior). In the core of Inferelator-2 is an importance sampling Markov Chain Monte Carlo algorithm, designed to efficiently sample the posterior. Importantly, to compare the merit of considered RNs, Inferelator-2 solves the system of ODEs as a coupled system, providing a more realistic model for the underlying RN.

Initial results, using synthetic and real datasets, suggest that our resampling approach together with the new Inferelator-2 method significantly boost our ability to correctly learn topology and to model dynamics over longer time scales. These changes to our framework provide an essential step toward learning more complex RNs, such as mammalian RNs, and over longer time scales, such as the time scales required to model cell differentiation.

## Functional Inference from a Genome-Wide in situ Hybridization Atlas of the Mouse Embryo

Attila Gyenesei<sup>1,2</sup>, Mei Sze<sup>1</sup>, Colin Semple<sup>1</sup>, Duncan Davidson<sup>1</sup>, Richard Baldock<sup>1</sup>.

<sup>1</sup>MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK; <sup>2</sup>Turku Centre for Biotechnology, BioCity, 5<sup>th</sup> floor, 20521, Turku, Finland.

Genome-wide in situ hybridization (ISH) datasets allow the identification of distinct patterns of sub-structural and cell type-specific expression below the level of a tissue type, which provides a basis for the inference of functional interactions, and can reveal exquisitely detailed patterns even for relatively lowly expressed genes. The EU FP6 consortium EurExpress has developed a genome-wide transcriptome atlas database for mouse embryo ([www.eurexpress.org](http://www.eurexpress.org)), which contains expression data of more than 18,000 genes by RNA ISH on sagittal sections from E14.5 wild type murine embryos. The group at the MRC Human Genetics Unit was responsible for the database development, informatics infrastructure and high-throughput data analysis.

In this study we investigated the spatial gene expression at a common developmental stage and down to a cellular level in the developing mouse. We used these data to identify coexpressed gene clusters, demonstrated their statistical and biological significance, and compared these results with coexpression defined using genome-wide, publicly available microarray data.

During the analysis we applied both well-known hierarchical clustering and our unique biclustering method to reveal the gene clusters. Biclustering was able to identify those genes that were coregulated not only for the whole but subsets of the EurExpress data. Moreover, it revealed genes participating in more than one gene network. Annotation enrichment was calculated for each discovered cluster using hypergeometric distribution. The functional annotation types used in this study were gene ontology, InterPro conserved domain identifiers, mammalian phenotype ontology terms, and cytogenetic band as a proxy for genomics position. The significance of enrichment across all clusters was determined using permutation strategy with false discovery rate calculation.

The significant enrichment of functional annotation demonstrated the statistical and biological significance of coexpressed EurExpress clusters. Interestingly, most of the clusters significantly enriched for functional annotation terms were found to contain significantly enriched terms from more than one annotation type offering fundamental insights into various pathways.

We also show that EURExpress coexpression clusters can successfully be used to infer novel functional relationships between genes at various levels. Based on these results, we believe that EURExpress data will be a potent tool in uncovering many more novel functional associations relevant to development and disease.

## Aromatase inhibition in a transcriptional network context

Tanwir Habib<sup>1</sup>, Edward J Perkins<sup>2</sup>, Daniel Villeneuve<sup>3</sup>, Gerald Ankleby<sup>3</sup>, David Bencic<sup>4</sup>, Nancy Denslow<sup>5</sup>, Li Liu<sup>6</sup>, Natàlia Garcia-Reyero<sup>7,8</sup>

<sup>1</sup>BTS, Vicksburg, MS, USA; <sup>2</sup>Environmental Laboratories, US Army Corps of Engineers, Halls Ferry Road, Vicksburg, MS, USA; <sup>3</sup>U.S. Environmental Protection Agency, ORD, NHEERL, MED, Duluth, MN, USA; <sup>4</sup>U.S. Environmental Protection Agency, Cincinnati, OH, USA; <sup>5</sup>Department of Physiological Sciences and Center for Environmental and Human Toxicology, University of Florida, Gainesville, FL, USA; <sup>6</sup>ICBR, University of Florida, Gainesville, FL, US; <sup>7,8</sup>Department of Chemistry, Jackson State University, Jackson, MS, USA

A variety of chemicals in the environment have the potential to inhibit aromatase, an enzyme critical to estrogen synthesis. We examined the responses of female fathead minnow ovaries (FHM, *Pimephales promelas*) to a model aromatase inhibitor, fadrozole, using a transcriptional network inference approach. Fish were exposed for 8 days to 0, or 30mg/L fadrozole and samples and then left in clean water for 8 more days. Samples were analyzed for significant changes in the gene expression with a 15,000 probe FHM microarray. The top 1674 significantly changed genes based upon 1.5-fold change and  $P < 0.05$  across all the time points, including some additional genes relevant to the Hypothalamus-Pituitary-Gonadal (HPG) axis as well as sex steroid levels, were chosen for network modeling. In order to gain biological understanding of the significantly expressed genes, we also analyzed the functional annotations. Some of the gene overrepresented ontology annotations were lipid, fatty acid and steroid metabolism, signal transduction, oxidoreductase, kinases, localization, cell signalling, and calcium ion transport. StAR-related lipid transfer was the most highly connected gene in the network model. Key HPG genes such as chorionic gonadotropin beta, low density lipoprotein, steroidogenic acute regulatory protein, cytochrome P450 family members, and estrogen receptor were found significantly expressed with the fadrozole exposure and were present in the steroidogenic network obtained from the source network.

Our results showed that the inferred network was extremely successful in detecting HPG axes interactions. Some of these interactions that were previously known included gonadotropin-releasing hormone receptor and its interaction with G-proteins, adenylate cyclase, and gonadotropin. The interaction network also suggested the role of calcium in association with cAMP in the stimulation of steroidogenesis in the gonads.

# Ground State Robustness as an Evolutionary Design Principle in Signaling Networks

Önder Kartal<sup>1,2</sup>, Oliver Ebenhöf<sup>1-3</sup>

<sup>1</sup>*Institute of Biochemistry and Biology, University of Potsdam, Potsdam/Golm, Germany;*

<sup>2</sup>*Max-Planck-Institute of Molecular Plant Physiology, Potsdam/Golm, Germany;*

<sup>3</sup>*Institute for Complex Systems, University of Aberdeen, Aberdeen, AB24 3UE, UK*

The ability of an organism to survive depends on its capability to adapt to external conditions. In addition to metabolic versatility and efficient replication, reliable signal transduction is essential. As signaling systems are under permanent evolutionary pressure one may assume that their structure reflects certain functional properties. However, despite promising theoretical studies in recent years, the selective forces which shape signaling network topologies in general remain unclear. Here, we propose prevention of autoactivation as one possible evolutionary design principle. A generic framework for continuous kinetic models is used to derive topological implications of demanding a dynamically stable ground state in signaling systems. To this end graph theoretical methods are applied. The index of the underlying digraph is shown to be a key topological property which determines the so-called kinetic ground state (or off-state) robustness. The kinetic robustness depends solely on the composition of the subdigraph with the strongly connected components, which comprise all positive feedbacks in the network. The component with the highest index in the feedback family is shown to dominate the kinetic robustness of the whole network, whereas relative size and girth of these motifs are emphasized as important determinants of the component index. Moreover, depending on topological features the maintenance of robustness differs when networks are faced with structural perturbations. This structural off-state robustness, defined as the average kinetic robustness of a network's neighborhood, turns out to be useful since some structural features are neutral towards kinetic robustness, but show up to be supporting against structural perturbations. Among these are a low connectivity, a high divergence and a low path sum. All results are tested against real signaling networks obtained from databases. The analysis suggests that ground state robustness may serve as a rationale for some structural peculiarities found in intracellular signaling networks.

[Original Full Length Paper]

## Evolution of the High Osmolarity Glycerol (HOG) stress response network across *Ascomycota* fungi

Jay H. Konieczka<sup>1,3</sup>, Michelle Chan<sup>3,4</sup>, Amanda Socha<sup>3</sup>, Ilan Wapinski<sup>3,5</sup>, Mark Styczynski<sup>3</sup>, Courtney French<sup>3</sup>, Jenna Pfiffner<sup>3</sup>, Dawn A. Thompson<sup>3</sup>, Aviv Regev<sup>3,4,6</sup>, and Erin K. O'Shea<sup>1,3,6</sup>

<sup>1</sup>FAS Center for Systems Biology and <sup>2</sup>Dept. of Molecular & Cellular Biology, Harvard University, Cambridge, MA; <sup>3</sup>The Broad Institute, Cambridge, MA; <sup>4</sup>Dept. of Biology, Massachusetts Institute of Technology, Cambridge, MA; <sup>5</sup>Dept. of Systems Biology, Harvard Medical School, Boston, MA; <sup>6</sup>Howard Hughes Medical Institute.

Divergence in gene regulatory networks plays a major role in the evolution of every kingdom of life. While comparative studies of system evolution have been remarkably effective in identifying the functional genomic elements and tracing major evolutionary events, most studies have relied on sequence data alone. There have been few studies done in yeasts combining sequence and functional genomic data, and they underscore the power of comparative functional genomics in analyzing the function and evolution of molecular systems. Early results have been promising, but are limited in phylogenetic or biological scope, and have had no choice but to employ ad hoc approaches to the study of system evolution.

We aim to systematically analyze the structure and evolution of the gene regulatory network underlying the High Osmolarity Glycerol (HOG) in response to osmotic stress in a tractable subset of *Ascomycota* fungi. The species in this study span more than 300 million years of evolution and include the model organisms *S. cerevisiae* and *S. pombe*, as well as the human fungal pathogens *C. albicans* and *C. glabrata*. The HOG regulatory network is controlled by a highly conserved MAP-Kinase (Hog1, p38 in humans), which in budding yeast evokes a stress response to facilitate survival in challenging osmotic conditions and mediates general stress response in fission yeast.

We presented each species in our study with low, medium, and high osmotic stress and measured genomic expression response profiles over 80 minutes. From these data, we identify conserved and derived gene regulatory modules (sets of co-regulated genes) in the evolution of the osmotic stress response (OSR). In the coming months, we will identify the Hog1-activated OSR genes in each species by measuring global gene expression differences between wild-type and strains lacking Hog1. In addition to identifying the conserved and divergent components of the Hog1 OSR, this will enable more finely detailed mapping of the *cis*-elements responsible for similarities and differences in gene regulation among species. Species- and clade-specific components will provide insights into lifestyle requirements and inform understanding of the ecologies and/or histories of those species. Understanding gained from systematically identifying gene regulatory changes across broad evolutionary space will provide significant insight into the forces shaping gene regulatory programs.

# Novel methods for the discovery of condition specific master regulators of transcription

Kenzie D. Maclsaac<sup>1</sup>, Chris Ng<sup>1</sup>, Ernest Fraenkel<sup>1,2</sup>

<sup>1</sup>*Department of Biological Engineering, MIT;* <sup>2</sup>*Computer Science and Artificial Intelligence Laboratory, MIT*

Identifying the specific transcription factors that assemble at enhancers and drive tissue-specific and condition-specific transcription in mammals is an important and unsolved problem. Exhaustive genome-wide experimental strategies that profile almost all transcription-factor binding have been applied in yeast, but are not practical for mammalian systems where there are approximately 10-fold more regulatory proteins and a multitude of distinct tissues.

To address these limitations we have developed a joint experimental and computational strategy. We identified regulatory sites using either (1) genome-wide binding for coregulators (ChIP-Seq) or (2) genome wide DNase-hypersensitive data (DNase-Seq). By learning which DNA sequence motif features discriminate experimentally identified sequence regions from background sequence, we reveal the master regulators responsible for recruiting coregulators to their targets and altering chromatin structure.

The approach was carried out in mouse liver, cerebellum, and 3T3-L1 cells. Interestingly, the sequence motifs associated with recruitment of a coregulator protein varies across tissues. We further demonstrate that the transcription factors predicted to associate with these sites in vivo are indeed bound in ChIP experiments, and that regions bound by multiple transcription factors are more likely to recruit a coregulator. We tested several predictive models of coregulator recruitment and found that simple models, where individual motifs contribute independently to coregulator binding probability, generally perform as well or better than more complex models. We find no strong evidence of particular motif spacing or orientation constraints associated with coregulator binding

# Modeling the Evolution of Regulatory Elements by Simultaneous Detection and Alignment with PhyloPairHMMs

Majoros WH<sup>1</sup>, Ohler U<sup>1,2</sup>

<sup>1</sup>*Institute for Genome Sciences & Policy, Duke University* ; <sup>2</sup>*Department of Biostatistics & Bioinformatics, Duke University*

Detection of regulatory elements based on pre-aligned cis-regulatory modules (CRMs) can be strongly affected by the quality of the pre-computed alignments. Using a model such as a Pair Hidden Markov Model (PHMM) permits the annotation of regulatory elements to proceed simultaneously with the alignment of orthologous CRMs, thereby potentially improving annotation accuracy via relaxation of the search space. However, generalizing PHMMs for use with more than two species is difficult, due to increased memory requirements and computational complexity. By expanding the computation of emission probabilities in a standard PHMM so as to perform efficient inference over a Bayesian network, we arrive at a model which can be used for progressive alignment and annotation of arbitrarily many orthologous sequences. These PhyloPairHMMs can incorporate sub-states for multiple transcription factors, allowing us to very easily model combinatorial control in a fully probabilistic framework. Finally, by seamlessly incorporating a stochastic gain/loss mechanism into the states of the PhyloPairHMM, we are able to adequately model, and capture, patterns of evolutionary gain and loss of regulatory elements. Results on a well-known benchmark set of *Drosophila* CRMs show that our program identifies binding sites with greater accuracy than current systems.

# REGULATORY ELEMENT IDENTIFICATION WITH FUNCTIONAL GENOMIC COVARIATES

Andre L. Martins<sup>1</sup>, Adam Siepel<sup>1</sup>

<sup>1</sup>*Biological Statistics and Computational Biology, Cornell University*

Despite the many benefits of genome sequencing, raw sequence data has been of limited use for the central task of comprehensively identifying and characterizing the sequences that control gene regulation in the human genome. New high-throughput functional genomic assays—such as ChIP-Seq, bisulphite sequencing, Methyl-Seq, DNase-Seq and FAIRE-Seq—promise to provide rich layers of additional information about gene regulation that can be mapped onto the genome sequence. However, each of these assays has certain limitations, and, in principle, much more can be learned about gene regulation by considering them in combination than can be learned by considering them separately. Toward this end, we have developed a method for probabilistic sequence analysis that considers diverse functional genomic data types as covariates, allowing flexible integration of primary sequences and functional data.

In our approach, data integration is accomplished by combining logistic regression (LR) with a hidden Markov model (HMM) for DNA sequences. The combined LR-HMM model is a highly general conditional-generative probabilistic graphical model that defines a probability distribution over annotations and sequences conditional on an arbitrarily complex set of covariates. This structure makes it suitable for unsupervised as well as supervised learning (which is important because training data is sparse), yet avoids elaborate modeling of the possibly complex interdependencies among functional genomic data sets. It also provides some information about the relative importance of each covariate. The model can be easily extended to consider comparative genomic data through a phylo-HMM, and to consider nonlinear dependencies through kernel logistic regression.

We show results, on both real and simulated data, demonstrating that the LR-HMM method can improve cis-regulatory binding site identification significantly, as compared with more naive methods. Furthermore, we explore the usefulness of the method in predicting cell type specific binding site identification, taking advantage of new genome-wide assays produced by the second phase of the ENCODE project. Our results suggest that rigorous modeling of functional genomic covariates, in some cases, may allow accurate identification of binding sites with a reduced number of (still-expensive) functional genomic experiments.

# Unraveling of an ancient regulatory pathway: RNAi insensitivity in the germline of *C. elegans*

Daniel A. Pollard<sup>1,2</sup>, Maxwell J. Kramer<sup>1,2</sup>, Matthew V. Rockman<sup>1,2</sup>

<sup>1</sup>Center for Genomics & Systems Biology, New York University; <sup>2</sup>Department of Biology, New York University

Eukaryotes utilize the microRNA and RNA interference (RNAi) pathways to regulate gene expression post-transcriptionally. microRNAs regulate endogenous gene expression and play critical roles during development. The natural role for the RNAi pathway is less well understood but is believed to function in protecting the germline from mobile elements and exogenous RNA. The core molecules in these regulatory pathways are well characterized and well conserved across eukaryotes however the components and deployment of the pathways vary substantially both within and across species. The round worm *Caenorhabditis elegans* has played a central role in the elucidation of these pathways and yet the commonly used Hawaiian natural isolate is completely insensitive to RNAi in the germline. We sought to expand the characterization of the loss of RNAi sensitivity in *C. elegans* to identify new genes involved in the pathway and better understand the genetics underlying the loss of this pathway in the population.

Using quantitative trait locus mapping techniques we first examined the genetics underlying the difference in germline RNAi sensitivity between the fully sensitive N2 lab strain and the fully insensitive Hawaiian isolate. Although the insensitivity does not segregate as a single locus Mendelian trait (implying complexity), mapping revealed only one large-effect QTL, centered over the argonaut gene *ppw-1* (previously implicated in this insensitivity). We next surveyed germline RNAi sensitivity in 41 natural isolates representing all known *C. elegans* haplotypes. Germline RNAi insensitivity is common and geographically widespread in natural populations, with nearly one in four natural isolates showing insensitivity. Surprisingly, association mapping with the natural isolates implicated no significant genomic regions, suggesting a high level of genetic complexity or heterogeneity. To test if the RNAi insensitivity in the population has heterogeneous causes we performed complementation tests between the insensitive natural isolates and a *ppw-1* null strain. The *ppw-1* null both complemented and failed to complement natural isolates, suggesting germline RNAi insensitivity may have been gained multiple times through separate mechanisms. We are currently performing QTL mapping of RNAi sensitivity in the *ppw-1* complementing natural isolates.

We conclude that germline RNAi insensitivity is a widespread and complex trait in *C. elegans* with heterogeneous and potentially novel underlying molecular mechanisms. We propose that the losses of the RNAi pathway response in the germline may be the result of relaxed selection due to infrequent outcrossing and small effective population size in *C. elegans*. These results suggest that we have caught an ancient and highly conserved regulatory pathway in the process of unraveling and falling apart.

## ESTOOLSDB – a comprehensive database for stem cell research

Kirsi Rautajoki<sup>1</sup>, Lingjia Kong<sup>1</sup>, Kalle Leinonen<sup>1</sup>, Riikka Lund<sup>2,3</sup>, Paul Gokhale<sup>3</sup>, Janne Seppälä<sup>1</sup>, Heidi Pukkila<sup>1</sup>, Jarno Mäkelä<sup>1</sup>, Lauri Tuomisto<sup>1</sup>, Laura Järvenpää<sup>1</sup>, Lauri Hahne<sup>1</sup>, Olli Yli-Harja<sup>1</sup>, Reija Autio<sup>1</sup>

1) Department of Signal Processing, Tampere University of Technology, Tampere, Finland, 2) Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland, 3) Centre for Stem Cell Biology and the Department of Biomedical Science, University of Sheffield, Sheffield, UK

Genome wide studies of gene expression, genetic variations and gene regulation produce a vast amount of information on cellular responses e.g. to different extracellular stimuli or during the disease progression. The acquired data is usually generated in the context of a certain project aiming at answering the questions and hypothesis at hand, and it is analyzed and interpreted within this framework. However, the data can powerfully provide novel insights into many aspects of science, particularly when combined with other previously acquired and published results [1]. Our ESTOOLS database consisting experimental data of more than 900 samples, including pluripotent stem cells, different multipotent progenitors as well differentiated cells from a variety of tissues, is an ambitious endeavor to accomplish this goal. The database includes gene and exon expression data, gene copy number data, SNP data as well as ChIP-chip and ChIP-seq data. In addition to the data gathered from public sources referring peer-reviewed articles, some unpublished data has been included. The data within the database is mainly downloaded in the raw format and each of the samples has been pre-processed and normalized utilizing similar methods. Thus, the comparison of the data values from different sources within the same array platform is straightforward. With various advanced data analysis options different data values can be fused and mined together. A lot of attention has been paid for the careful and comprehensive biological annotation of all the samples. This detailed information can be used to search e.g. responses in gene expression levels to certain differentiation pathways, or epigenetic and transcriptional characteristics of stem cells in comparison with many other cell types. The database provides various tools to browse, retrieve, visualize, and analyze the data. For example, ESTOOLS database includes robust clustering methods, tools for detecting differentially or similarly expressed genes and tools for finding the enriched Gene Ontologies and KEGG pathways of the resulted gene lists.

The focus of the database is on stem cell biology, including the maintenance and differentiation of these pluripotent cells. However, also other sample types, such as cancer cells, fibroblasts, and tissue fractions have been included to enable comparison and identification of the typical features of stem cells. The purpose of the database is to provide high quality, continuously updated stem cell data for the research community, and it is available under an open source license.

[1] Kilpinen, S., Autio, R., Ojala, K., Iljin, K., Bucher, E., Sara, H., Pisto, T., Saarela, M., Skotheim, R.I., Björkman, M., Mpindi, J.-P., Haapa-Paananen, S., Vainio, P., Edgren, H., Wolf, M., Astola, J., Nees, M., Hautaniemi, S., Kallioniemi, O. (2008) Systematic bioinformatic analysis of expression levels of 17330 human genes across 9783 samples from 175 types of healthy and pathological tissues./ *Genome Biology* /2008, 9:R139.

# Discriminating functionality by kernel clustering k-mers in the regulatory genome

Pradipta Ray<sup>1</sup>, Le Song<sup>1</sup>, Eric P. Xing<sup>1</sup>

<sup>1</sup>School of Computer Science, Carnegie Mellon University

**Motivation:** Identifying positions and function of individual transcription factor binding sites (TFBS) in the regulatory genome is a challenging problem in genomics. Degeneracy in nucleotides and presence of numerous noisy copies of the binding sites make supervised TFBS prediction difficult, with high false positive rates. Further, the underlying organization of TFBSs into spatially contiguous clusters or *cis*-regulatory modules (CRMs) add to the complexity of prediction. **Methods:** In this work, we study how much discriminative power regarding functionality is present solely in a particular regulatory subsequence without knowledge of its position in the CRM. We split up the *cis*-regulatory regions and its flanking regions into k-mers and proceed to cluster the k-mers based on the number of their occurrences inside the CRMs. This is in the spirit of work done by [1], which analyzes a similarity matrix of k-mers in order to identify clusters corresponding to graph cuts which may correspond to functional elements. Our method uses a kernel k-means algorithm and applies a Gaussian RBF kernel on the k-mer occurrences. The advantage of our method is that we do not need to explicitly specify a restricted similarity measure across k-mers in order to cluster; the RBF kernel retains the ability to find clusters which may have non-linear boundaries in the input space (ie. the occurrence count).

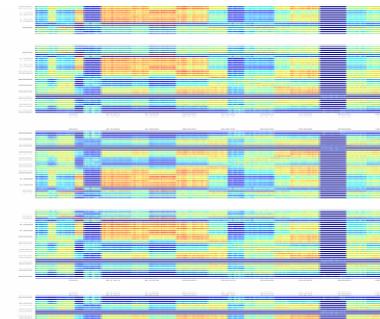
**Results:** We apply our method to 33 early developmental CRMs and flanking regions of 500 bp around them in *Drosophila melanogaster* curated from the NUS CRM database, which contain 512 5-mers (with reverse complementary k-mers treated identically). We set the number of clusters to 20, and compare with randomly generated clusters of the same sizes. We compare the number of k-mers located within a fixed threshold (10 bp) of TFBSs and within CRMs in each cluster for the kernel clustering we generate as opposed to random clusterings. We find that we manage to cluster together k-mers in the vicinity of binding sites and *cis*-regulatory modules better than random clusterings with p-values of 0.001 or below, implying that purely on the basis of sequence specificity and frequency of k-mers, we can achieve a significant level of discrimination with respect to segregating binding vs non-binding sites, or regulatory vs non-regulatory (flanking) regions. Shown in the figure is

the heat map of the kernel matrix for the 512 5-mers based on the results of the kernel clustering. This work is complementary to [2], which shows enrichment by clustering generated on evolutionary parameters of each sliding window in *Drosophila* regulatory genome.

## References

[1] Fratkin E, Naughton BT, Brutlag DL, Batzoglou S (2006) MotifCut: Regulatory motifs finding with maximum density subgraphs. *Bioinformatics* 22: e150–157

[2] Ray, P.; and Xing, E. P.; Analysis of Co-evolution in *Drosophila* regulatory



genome; RECOMB Regulatory Genomics Satellite 2008, Boston.

# Spatial association of multiple coordinately expressed but functionally unrelated genes during cell differentiation

Dietmar Rieder<sup>1</sup>, Marcel Scheideler<sup>1</sup>, Gernot Stocker<sup>1</sup>, Maria Fischer<sup>1</sup>, Waltraud G. Müller<sup>2</sup>, James G. McNally<sup>2</sup>, Zlatko Trajanoski<sup>1</sup>

<sup>1</sup>Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria; <sup>2</sup>Laboratory of Receptor Biology and Gene Expression, National Cancer Institute, National Institutes of Health, Bethesda, MD/USA

The three-dimensional (3D) organization of the human genome is an important but poorly understood aspect of gene regulation. Transcriptionally active genes were often found to be positioned in the nuclear interior whereas both silent and active genes were reported to be associated with the nuclear periphery. Besides the positioning of genes at certain nuclear landmarks, a spatial association for functionally related active genes is indicated by the nucleolar clustering of rDNA genes or the spatial aggregation of tRNA genes. Furthermore, spatially close encounters of selected active genes sharing the same RNA-Polymerase II (POL-II) transcription factory or SC35-enriched splicing speckles were observed during erythroid differentiation. However it remains unclear what causes genes to associate and whether a spatial association of coordinately expressed genes can be found in general or if it applies just to some selected and functionally related examples. In the present study we tested the hypothesis that coordinated expression of multiple genes in the regulatory situation of cell differentiation is related to a specific 3D organization of the genome. We used a human stem cell model and generated gene expression profiles for the transition from uncommitted stem cells to mature adipocytes. Based on our microarray data analysis we selected a cluster of seven coordinately expressed but functionally unrelated genes that were located on five different chromosomes. We then developed and applied a combinatorial multiplex FISH procedure, which enabled us to simultaneously study the spatial organization of multiple genes in single cells by high-throughput 3D microscopy. The computational analysis of the DNA-FISH signals revealed that several genes were spatially associated, but there was no evidence that they form a single cluster in the nuclear space. However, we frequently observed smaller non-randomly organized sub-clusters which mostly consisted of two or three members. When examining the pairwise association we found three gene pairs exhibiting a remarkable high frequency of close spatial association (13-18%, random < 1%). We then examined the positioning of the individual chromosomes to find a possible explanation why some genes are more often associated than others, but we found that the chromosome positioning alone does not sufficiently account for individual gene associations. To investigate if our co-expressed gene pairs share the same mRNA production environment, we performed triple color immuno-RNA-FISH experiments using intron probes in combination with either a POL-II or a SC35 antibody. We could show that the studied genes sometimes shared the same POL-II site but more often than that, they were associated with the same SC35-enriched splicing speckle. This indicates that active genes generally can be brought together at RNA processing sites and that this process is not intrinsic to groups of closely related genes.

Our approach of combining computational and experimental methods provides new insight on the spatial organization of coordinately expressed but functionally unrelated genes and the results suggest that non-random positioning of genes in the nucleus plays an important role in gene regulation.

# Global Entrainment of Transcriptional Systems to Periodic Inputs

Giovanni Russo<sup>1</sup>, Mario di Bernardo<sup>1</sup>, Eduardo D. Sontag<sup>2</sup>

<sup>1</sup>*Department of Systems and Computer Engineering University of Naples Federico II, Italy;*

<sup>2</sup>*Dept. of Mathematics, Rutgers University, USA*

The activities of all living organisms are governed by complex sets of finely regulated biochemical reactions.

Often, entrainment to certain external forcing signals helps control the timing and sequencing of reactions. For example, human activities are clearly regulated by the day-night cycle.

A pressing open problem is to understand the onset of entrainment and under what conditions it can be ensured in the presence of uncertainties, noise and environmental variations. We address the problem of providing firm mathematical conditions for transcriptional systems and other biological networks to be globally entrained to external periodic inputs.

From a mathematical viewpoint, the problem of formally showing that entrainment takes place in biology is known to be very difficult, since the model is in general nonlinear. For such systems, driving the system by an external periodic signal, does not guarantee the system response to be also a periodic solution. Our main result shows that, even if the transcriptional modules are modeled by nonlinear ODEs, they can be entrained by any positive periodic signal. Surprisingly, such a property is always preserved if the system parameters are varied: that is, entrainment is obtained independently of the particular biochemical conditions. Furthermore, we prove that combinations of the above transcriptional module, like cascades, or single-input-multi-output interconnections, also show the same property of being able to be entrained by positive periodic inputs, independently of the system parameters. Finally, we show how to engineer novel synthetic biochemical systems that exhibit entrainment.

From a dynamical system/theoretical viewpoint, the key idea is to consider entrainment as a property of all system trajectories (i.e. they converge towards each other), instead that studying the stability of some invariant set (typical of Lyapunov-based techniques).

# PPI module for visualization and analysis of protein-protein interfaces in Friend.

Amit Upadhyay, Valentin A. Ilyin.

*Department of Biology, Boston College. Chestnut Hill, MA*

The interactions among proteins are largely responsible for the complexity of biological systems. Identification of interactions between proteins will help in elucidation of protein networks providing insights into biological processes which in turn finds application in drug designing, protein engineering and developing scoring functions for docking. Studying the interfaces between interacting proteins will provide greater understanding of principles governing the binding of proteins. A number of methods have been described for studying protein interfaces. But none of the methods could be effectively used in predicting protein-protein recognition sites.

We describe an application library for visualization and analysis of protein-protein interfaces based on Voronoi-Delaunay tessellation (VDT), in the Friend software (<http://ilyinlab.org/friend>, an integrated analytical application designed for simultaneous analysis and visualization of multiple structures and sequences of proteins and/or DNA/RNA). The presented PPI method only using VDT is more objective as compared to those based on changes in solvent accessible surface area ( $\Delta$ SASA) and various radial cutoffs that lead to ambiguity. The library takes structural data files in PDB format as input and enables visualization of the chain-level interfaces as well as a detailed qualitative and quantitative analysis of the interfaces. There is no restriction with respect to the number of chains present in the PDB files and also interactions involving heteroatom's can be considered unlike most methods. This library can therefore be extended to study protein-ligand interactions as well.

We compared results of interface analysis by the PPI-Friend with other methods on a number protein complexes consisting of two chains. The protein structures were solvated in order to identify interactions mediated through water. The study revealed that the number of indirect interactions was very large which is often not considered in most methods. Another observation was that polarity of interface was found to be much higher as compared to other methods.

The library can be used to carry out large-scale statistical analysis of protein interfaces in order to determine the significance of these indirect interactions in prediction of protein recognition sites. The data obtained by this study can further be used in training machine-learning based approaches for protein docking.

## Composite network motifs in integrated metazoan gene regulatory networks

Vanessa Vermeirssen<sup>1</sup>, Tom Michoel<sup>1</sup>, Yves Van de Peer<sup>1</sup>

<sup>1</sup>Laboratory for Bioinformatics and Evolutionary Genomics, VIB Department of Plant Systems Biology, Ghent University, Belgium

Differential gene expression is a tightly controlled process that governs development, function and pathology of metazoan organisms. Several molecular interactions, e.g. protein-DNA interactions between transcription factors and target genes, protein-protein interactions between transcription factors, closely work together in order to establish proper gene expression in space and time. Biological networks have mainly focused on the relationships between one or two types of molecular interactions.

In order to get a systems level understanding of how different molecular interactions interrelate to form a coordinated response in gene regulation, we studied composite network motifs in integrated networks containing protein-protein, transcription regulatory, protein-DNA, miRNA-mRNA, sequence homology and genetic interactions of the worm *C. elegans*. Through a computationally efficient and mathematically rigorous method, we identified dense clusters of several composite network motifs in this integrated *C. elegans* network.

We discuss the biological function of these composite network motifs in the context of eukaryotic gene regulation. We conclude that composite network motif clustering is a useful data integration method to unravel the topological organization of gene regulation in metazoan organisms.

# Identification of genomic features novel to xylose-fermenting yeasts through comparative analyses of *Pichia stipitis*, *Candida tenuis*, and *Spathaspora passalidarum*

Dana J. Wohlbach<sup>1,2</sup>, Thomas W. Jeffries<sup>2,3</sup>, Alan Kuo<sup>4</sup>, Igor V. Grigoriev<sup>4</sup>, Kerrie W. Barry<sup>4</sup>, Audrey P. Gasch<sup>1,2</sup>

<sup>1</sup> Department of Genetics, University of Wisconsin, Madison; <sup>2</sup> Great Lakes Bioenergy Research Center, Madison, Wisconsin; <sup>3</sup> Department of Bacteriology, University of Wisconsin, Madison; <sup>4</sup> DOE Joint Genome Institute, Walnut Creek, California

Efficient fermentation of cellulosic feedstocks is an essential step in the production of ethanol from plant materials. The six-carbon sugar glucose and the five-carbon sugar xylose are the two most abundant monomeric carbohydrates found in hemicellulose. Although *Saccharomyces cerevisiae*, the yeast most commonly utilized for fermentation during ethanol production, is able to ferment glucose, it is unable to ferment xylose. However, several Ascomycete yeasts that both ferment and assimilate xylose have been identified including *Pichia stipitis*, whose genome has recently been sequenced.

To elucidate the genetic features that underlie the ability to ferment xylose, we performed whole-genome sequencing with the Joint Genome Institute (JGI) on two novel xylose-fermenting yeasts, *Candida tenuis* and *Spathaspora passalidarum*, and performed comparative genomic analyses between the xylose-fermenting yeasts *P. stipitis*, *C. tenuis*, *Sp. passalidarum* and other closely related non-xylose-fermenting yeasts, including *S. cerevisiae*. Here we present analysis of the genome sequences, including phylogenetic reconstruction and an examination of CUG codon usage in these yeasts. Additionally, mapping of xylose growth and fermentation phenotypes onto ortholog groups allowed us to identify several genes unique to xylose-fermenting species. We also present a comparative analysis of gene expression across species in response to glucose or xylose. We anticipate that the genomic features identified through our analysis may be candidates for engineering more efficient xylose production in *S. cerevisiae*.

# Structural and Regulatory Evolution of Electrophysiological Systems

Qinghong Yan<sup>1</sup>, Barbara Rosati<sup>2</sup> and David McKinnon<sup>1</sup>

<sup>1</sup>Program in Neuroscience, Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, NY 11794, U.S.A.; <sup>2</sup>Department of Physiology and Biophysics, Stony Brook University, Stony Brook, NY 11794, U.S.A.

There are two primary ways by which biological systems can evolve, either by changes in the sequence of protein coding regions or by changes in gene regulatory function. There is considerable evidence showing that regulatory evolution is the predominant mechanism underlying the evolution of developmental systems. In contrast, for physiological systems, studies have largely focused on the evolution of individual proteins and it has been concluded by some researchers that structural evolution predominates in these systems. A limitation of these prior studies has been the failure to consider the function of the system as a whole, and in particular the role of computational function in physiological systems. My studies have shown that evolution of the computational function of electrophysiological systems relies predominantly on regulatory evolution, similar to what is seen in developmental systems.

Like most physiological systems, electrophysiological systems must solve two sets of problems, one primarily physical and the other primarily computational. The physical tasks that ion channels perform are relatively simple, requiring the specific transport of common cellular ions across the cell membrane in response to one or more stimuli. The computational tasks range from simple, such as producing a timing cycle in cardiac tissue, to complex, for networks of neurons in the cortex. I have shown that the evolution of the computational function of cardiac electrophysiology is predominantly achieved by regulatory evolution. In particular I have shown that the scaling of action potential duration and morphology with body weight is mediated almost entirely by changes in the expression of multiple ion channel and transporter genes. I have shown for two simple traits, which contribute to this overall scaling, that changes in cis-regulatory function underlies their evolution. There is no significant change in the function of the proteins that mediate electrical activity in this system.

It is concluded that the relative balance between physical and computational tasks will vary considerably among different physiological systems and that this will be an important factor in determining whether a particular system evolves primarily by regulatory or structural evolution. For systems that can call upon a sufficiently complex set of pre-existing structural components, many of the challenges to computational function that arise during the course of evolution may be most readily solved by regulatory evolution. In contrast, purely physical tasks are more likely to require novel structural solutions.

# Design of multiple hypothesis tests for microarray data

Peter Huggins<sup>1</sup>, Ziv Bar-Joseph<sup>1</sup>

<sup>1</sup>Carnegie Mellon University

Multiple hypothesis testing (MHT) is an important statistical topic in genomics, e.g. microarray analysis and identification of differentially expressed genes (DEGs). Popular software such as SAM identifies DEGs while controlling a type I error such as FDR.

“Proportion-like” measures of Type I error (e.g. FDR) have several desirable properties: 1) Less stringent than family-wise error rates, providing more power, 2) Weak DEGs can be identified as differentially expressed if there is a large proportion of DEGs, 3) Preserved under concatenation: Combining DEG lists with  $\leq X\%$  errors gives  $\leq X\%$  errors in the aggregate.

These properties suggest the possibility of improving power in microarray analysis, simply by partitioning the set of genes *a priori* into subsets, and running existing microarray MHT software (e.g. SAM) on each subset separately. We call this *partition-based MHT*. Obviously a randomly chosen partition cannot be expected to improve power. But partition-based MHT can gain power if the partition is well-chosen (e.g. using biological side-knowledge), so that DEGs are disproportionately concentrated in particular subsets. Under the same mild statistical assumptions already proposed for microarray analysis, partition-based MHT provides the same FDR control as the underlying MHT procedure.

We demonstrate on real microarray data that partition-based MHT can indeed improve power in practice. Our strategy is simple and can be applied using any existing FDR-controlling microarray analysis software such as SAM. Simply partition the genes into a few subsets (as few as two subsets), run SAM on each gene subset separately [under a common FDR control], and concatenate the DEGs found. One way to partition genes is to split into two subsets: one subset containing *a priori* suspected DEGs (e.g. based on expert knowledge), and one subset containing the remaining genes.

In this way, expert knowledge can be easily brought to bear to improve statistical power in microarray analysis, by exploiting suspected structure in the alternative to the joint null. We emphasize, however, this does *not* mean partition-based MHT is Bayesian. Every partition-based MHT is a valid multiple hypothesis test in the classical (frequentist) sense. Still one has the freedom to choose which test to perform, much like choosing a one-sided versus two-sided test in dimension one. One interpretation of our results is that for high dimensional biological data such as microarray (and perhaps SNP association), careful design of rejection regions can have a large impact on statistical power.

## Life After Comparative Genomics; Regulatory Systems, Homeostasis, Synergy, SNPs and Disease.

Alasdair MacKenzie, Lynne Shanley, Scott Davidson, Marissa Lear, John Barrow, Gemma Halliday and Ben Wen Qing Hing.

*Gene Regulatory Systems Group, School of Medical Sciences, Institute of Medical Sciences, University of Aberdeen, Aberdeen, AB25 2ZD, Scotland, UK. E-mail- mbi167@abdn.ac.uk .*

Thanks to the efforts of different Genome Sequencing programs and the development of powerful computer algorithms, the use of comparative genomics for the identification of important gene regulatory regions is now routinely used. We believe that the next stage in increasing our understanding of the role of gene regulation in health and the development of disease must be to understand the regulatory systems (ligand-receptor, signal transduction and protein-DNA interactions) that modulate the activity of these regulatory regions<sup>1,2</sup>. Another critical step forward must be to understand how separate regulatory regions interact with each other to modulate gene expression and how these interactions may differ between individuals as a result of regulatory polymorphisms.

We used comparative genomics to understand the cell specific mechanisms that regulate neurogenic inflammation, a process associated with chronic diseases such as arthritis, inflammatory bowel disease and asthma. We found that the promoter of the TAC1 gene, that encodes substance-P (SP); an important player in neurogenic inflammation, only responds to the inflammatory stimulus; capsaicin (chilli extract), in sensory neurones in the presence of a highly conserved and remote enhancer (214kb from TAC1). Culturing of transgenic dorsal root ganglion explants and primary sensory neurones in the presence of different cell signalling agonists and antagonists demonstrated that synergy was required between this enhancer and the TAC1 promoter to allow a response to the p38MAPK pathway that modulates aspects of the inflammatory response. These novel observations have important implications for the understanding of the mechanisms underlying the development of chronic pain. We present further data demonstrating the role of a second remote TAC1 enhancer that can suppress the glucocorticoid receptor (GR) induced activity of the TAC1 promoter in the amygdala<sup>3</sup>. Intriguingly this second enhancer is also controlled by GR suggesting GR mediated homeostasis at the TAC1 locus. Because SP has anxiogenic properties when expressed in the amygdala this result has important implications for understanding the mechanisms underlying susceptibility to anxiety and chronic depression. We are currently exploring the effects of several common human SNPs on the regulatory activity of these sequences.

Finally, we have established collaborations with psychiatric geneticists at the Institute of Psychiatry in London who have carried out GWAS analysis on polymorphisms within large sample groups of patients suffering from major depressive disorder. We are currently in the process of studying the strongest “hits” from these analyses that, intriguingly, are frequently contained within highly conserved non-coding sequences

1. Miller KA, et al. (2007) *Dev Biol* 311(2):665-678.
2. Miller KA, et al. (2008) *Dev Biol* 317(2):686-694.
3. Davidson S, et al (2006) *Mol Psychiatry* 11(4): 410-421.

# TP53 cancerous mutations exhibit selection for translation efficiency<sup>§</sup>

Yedaël Y. Waldman<sup>1,2</sup>, Tamir Tuller<sup>\*1-3</sup>, Roded Sharan<sup>1</sup>, Eytan Ruppin<sup>1,3</sup>

<sup>1</sup>Blavatnik School of Computer Science; <sup>2</sup>Department of Molecular Microbiology and Biotechnology; <sup>3</sup>School of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel. \*These authors contributed equally to this work.

The tumor suppressor gene TP53 is known to be a key regulator in cancer, and more than half of human cancers exhibit mutations in this gene. Being a known tumor suppressor, one would expect cancerous mutations would decrease the levels of TP53 either by diminishing protein synthesis or by producing a truncated product. However, more than 75% of TP53 cancerous alterations are missense point mutations that lead to the synthesis of a stable full-length protein that in many cases is expressed at higher levels than the wild type p53. These intriguing results can be partially explained by recent evidence showing that these point mutations not only disrupt p53 function as a tumor suppressor but also possess gain of function (GOF) and dominant negative effects (DNE) on p53 wild type copies by binding to the latter, thus making the mutated tumor suppressor an *oncogene*. This hence brings about the possibility that TP53 mutations may be under selection in cancerous cells for increasing defected p53 oncogenic activities. One such potentially likely mechanism examined here is increasing the mutated protein levels via higher translation efficiency (TE). Here we perform the first large scale analysis of TE in human cancer mutated TP53 variants, analyzing 17,851 point mutations reported in various tumors collected from 2081 different studies. We identify a significant increase in TE of 1.4 fold that is correlated with the frequency of TP53 mutations (P-value=4.71×10<sup>-5</sup>). Furthermore, mutations with known oncogenic effects significantly increase their TE compared to other TP53 mutations (P-values 0.0424 and 2×10<sup>-5</sup>, for DNE and GOF mutations, respectively). Further analysis shows that TE may have influence both on selecting the location of the mutation and on its outcome: codons with lower TE show stronger selection toward non-synonymous mutations (P-value=4.2×10<sup>-3</sup>) and, for each codon, frequent mutations show stronger increase in TE as compared to less frequent mutations (P-value=4.31×10<sup>-4</sup>). In addition to significant differences in TE between tumors from different origins, we find that TP53 mutations show significantly higher TE increase in progressive vs. primary tumors (TE increase fold 1.50 vs. 1.42, P-value=3×10<sup>-5</sup>). This result, further supporting the role of TE selection in these mutations, gives rise to the possibility that TE analysis of existing TP53 mutations may be another additional indicator that should be considered in predicting the outcome of the cancer. Finally, an analysis of both chromosomal aberrations and point mutations in TP53 in NCI-60 cancerous cell lines points to an interesting co-adaptation between TP53 point mutations and aberrations in the tRNA pool, increasing the overall TP53 TE (TE increase fold 1.825 vs. 1.640, P-value=0.03).

Taken together, these results show that TE plays an important role in the selection of TP53 cancerous mutations. As more mutational data accumulates on a large scale for other oncogenes in the future, a similar analysis should be performed to study the general scope of TE changes in cancer development.

<sup>§</sup> This work has been accepted for journal publication in *Cancer Research*. The full journal version may be found at: [http://www.cs.tau.ac.il/~tamirtul/P53\\_Website/P53\\_TE.html](http://www.cs.tau.ac.il/~tamirtul/P53_Website/P53_TE.html)

## NF-kB and Forkhead – partners and opponents

Erzsebet Ravasz Regan, Md. Ruhul Abid, Marianne Grant, William C. Aird

*Department of Medicine and Center for Vascular Biology Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston MA*

In mammalian cells, FOXO transcription factors are known to be activators of death and cell cycle arrest promoting genes. However, we and others have recently shown that FOXO proteins induce the expression of genes involved in endothelial health. Indeed, despite the fact that FOXOs are ubiquitously expressed, mice that are null for these transcription factors display a vascular-specific phenotype. The novel health promoting role of FOXO involves the cooperative activity of other positive acting factors, including p50 and p65 NF-kB.

Here we hypothesized that *in silico* methods could be leveraged to accurately predict the type(s) of FOXO – NF-kB co-regulation, and thus guide our *in vitro* experiments. In order to capture the logic by which FOXO and NF-kB factors regulate their target genes, we used a large public compendium of human microarrays (GEO database). We have designed a simple gradient-based algorithm to assign two-input Boolean gates to genes that show combinatorial regulation by FOXO (FOXO1, FOXO3 and FOXO4) and p50 and p65 NF-kB (such as cooperation via an AND gate or one factor inhibiting activation by the other via an AND NOT gate). The three FOXO factors, paired up with the two NF-kB proteins, are predicted to control a large number of genes via AND gates. In order to narrow our predictions to direct FOXO - NF-kB target genes, we performed a search of transcription factor binding sites on the promoters of all genes represented in the microarray compendium. We found a 30-bp sequence containing a FOXO – NF-kB binding site pair with 9 conserved bps in between, repeated almost unchanged on a large number of genes. There was a statistically significant overlap between genes with predicted AND gates and genes that have the 30-bp binding pair sequence on their promoter, observed for all 6 pairings of FOXO1, FOXO3 and FOXO4 with p50 and p65. This finding is consistent with activating co-binding on target promoters. Surprisingly, however, there was also a statistically significant overlap between genes with predicted NF-kB AND NOT FOXO1 gates and genes that have the 30-bp binding pair sequence on their promoter. This indicates inhibition of NF-kB driven transcription by FOXO1. Most of these genes display FOXO4 AND NF-kB gates, suggesting that FOXO1 inhibits an otherwise positive interaction between FOXO4 and NF-kB.

Genes predicted to be activated by most combinations of FOXO and NF-kB factors were enriched in pathways consistent with known roles of FOXO and NF-kB, including cancer, apoptosis, VEGF, MAPK, p38, Notch and TNFR1 signaling, leukocyte transendothelial migration, regulation of actin cytoskeleton, and natural killer cell mediated cytotoxicity. However, the gene group predicted to be co-activated by a FOXO4 – NF-kB pair but inhibited by FOXO1 was linked to subcellular localization and protein transport, previously unknown functions of FOXO proteins. Experimental validation of our predictions is ongoing. Based on real-time-PCR results from endothelial cells in which FOXO1 and NF-kB were up and down-regulated together (in all four combinations) we present some preliminary evidence for the existence of target genes controlled through NF-kB AND NOT FOXO1 logic.

# Integration site of the HIV promoter primarily modulates transcriptional burst size, rather than frequency

Ron Skupsky<sup>1</sup>, John Burnett<sup>2</sup>, Jonathan Foley<sup>2</sup>, David Schaffer<sup>1,3</sup>, Adam Arkin<sup>1,2,4</sup>

<sup>1</sup>California Institute for Quantitative Biology, <sup>2</sup>Department of Bioengineering, <sup>3</sup>Department of Chemical Engineering and Helen Wills Neuroscience Institute, UC Berkeley; <sup>4</sup>Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA

Mammalian cellular expression patterns are regulated by factors specific to each gene in concert with its surrounding genomic environment. HIV is a retrovirus that infects human immune cells, whose RNA genome is reverse-transcribed to DNA and integrated semi-randomly into the host-cell genome upon infection, and provides a natural system to dissect the contributions of genomic environment to gene-expression regulation. Previously, we have shown that expression heterogeneities, and their modulation by specific host factors at viral integration sites, are key determinants of HIV-infected cell fate and a possible source of latent infection, which represents the most significant obstacle to complete eradication of the virus in patients. Here, we assess the integration-context dependence of initial expression heterogeneities, that may exist in-vivo shortly after infection but before significant viral expression has commenced, using diverse single integrations of an HIV-promoter/GFP-reporter cassette in cultured Jurkat T-cells. Systematically fitting a model of stochastic gene expression to our single-cell fluorescence measurements suggests that transcript production in bursts accounts for the wide, highly skewed, expression patterns, that we observe across single-integration, clonal populations. We find that transcriptional burst size (the predicted average number of transcripts produced with each gene-activation event) is the primary systematic covariate over viral integration sites, while burst frequencies are scattered about a typical value of several per cell-division time with little correlation to integration-clone expression mean. This pattern of modulation approximately preserves the relative widths of single-gene expression distributions across productive and repressive viral integrations, allows the virus to sample a particularly 'noisy' range of basal expression patterns that may act as an initial determinant of infected-cell fate, and implicates genomic environment as an important control parameter for expression heterogeneity that may be exploited by integrating viruses.

# SPLINTER: detection of rare regulatory variants using a large deviation theory approach

Francesco LM Vallania<sup>1</sup>, Todd E Druley<sup>1,2</sup>, Ingrid Borecky<sup>1</sup>, Robi D Mitra<sup>1</sup>

<sup>1</sup>Center for GenomeSciences, Washington University School of Medicine, St. Louis, Missouri, USA; <sup>2</sup>Division of Pediatric Hematology and Oncology, Department of Pediatrics, Washington University School of Medicine, St. Louis, Missouri, USA.

Regulatory variants represent an important class of mutations involved in common diseases. Regulatory variants can affect gene expression levels by mutating transcription factor binding sites and methylation sensitive loci in non-coding regions (*cis*-regulatory variants) as well as affecting protein-coding regions of transcription factors (*trans*-regulatory variants). Detection of regulatory variants (in particular rare variants) is an important problem in post-genomic biology due to costs and time-requirements associated to traditional sequencing approaches for the analyzed sample sizes. Pooled-DNA sequencing approaches have been shown to enable fast and accurate detection of rare genomic variants at a fraction of time and cost of traditional methods. Existing computational approaches however have been limited to detection of mutations in single samples. In particular, detection of rare insertions and deletions (IN/DELS) remains a difficult and important computational challenge. To address this deficiency, we have developed a new information theory-based algorithm called SPLINTER (Short Variant Prediction by Large deviation Inference and Non-linear True frequency Estimation by Recursion), which detects and quantifies short IN/DELS as well as single nucleotide substitutions (SNPs) in pooled-DNA samples. Sequencing reads are generated from pooled-DNA sequencing experiments and mapped back to their reference sequence by employing a new semi-local alignment strategy. SPLINTER is able to analyze aligned reads using a second-order error model without making any prior assumption on the variant distributions. To determine the accuracy of SPLINTER, we sequenced synthetic pools of DNA molecules containing various combinations of up to 15 known sequence variants. We were able to detect rare 1~2 base pair long IN/DELS and single nucleotide substitutions with 100% specificity and sensitivity in pools of 500 alleles and with 100% sensitivity and ~99.9% specificity in pools of 1000 alleles. We could detect 4bp IN/DELS resulting with 100% sensitivity and 100% specificity in pools as large as 50 alleles and with 100% sensitivity and 98.8% specificity in pools of 1000 alleles. SPLINTER was also able to accurately estimate the abundance of each variant in a pool. We next applied our approach to sequence 14 loci in 974 individuals, grouped into 8 pools ranging from 94 to 150 individuals (188 to 300 haploid genomes) per pool. We identified both novel and previously known variants with significant dbSNP enrichments between 72 and 100%. These results indicate that SPLINTER can accurately detect rare variants in large pools, providing a novel and sensitive method that will allow for significant progress in the discovery of new regulatory variants. Given the non-biased nature of SPLINTER, as no prior knowledge is incorporated in any step of the detection and estimate of the variant frequency, an additional promising application is detection of induced mutations in *in vitro* evolution experiments.

# Prediction of Polycomb target genes in mouse embryonic stem cells

Yingchun Liu<sup>1,2</sup>, Zhen Shao<sup>1,2</sup>, Guocheng Yuan<sup>1,3</sup>

<sup>1</sup> Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115; <sup>2</sup> These authors contributed equally; <sup>3</sup> Corresponding author: [gcyuan@jimmy.harvard.edu](mailto:gcyuan@jimmy.harvard.edu)

## Abstract

Polycomb group (*PcG*) mediated gene silencing plays an important role in metazoan development. A fundamental question is how *PcG* targets only specific genes. We have developed a computational approach to predict *PcG* target genes based on sequence-specific transcription factors (TFs) binding and motif information. Our model is able to predict *PcG* targets in mouse embryonic stem cells (ESCs) with good accuracy even with a reduced version containing only five TFs (*Zf5*, *Tcfcp2l1*, *CTCF*, *E2f1*, *Myc*). We found that TF-based predictions are significantly more accurate than simple correlation with the CpG dinucleotide density and sequence conservation, two features that have been previously implicated in *PcG* recruitment. Interestingly, although our model is trained using ESC-specific data, it is predictive of intrinsic lineage-dependent target plasticity. Although *PcG* is conserved and targets genes of similar functions in mouse and in *Drosophila*, different motif signatures are associated with the *PcG* target sequences. Our results provide new insights into the evolution of gene regulatory network for animal development.

## Author Summary

Gene activities in eukaryotic cells are regulated by the concerted action of transcription factors and epigenetic factors. The epigenetic regulation of Polycomb group (*PcG*) proteins plays an important role in the maintenance of pluripotency in embryonic stem cells and the control of cell differentiation. In mammals, *PcG* specifically targets thousands of genes, many of which are developmental regulators. A fundamental question is how such target specificity is established. We have developed a computational model to predict *PcG* target genes in mouse embryonic stem cells by combining transcription factor binding and motif information. Our model has good prediction accuracy even for a vastly reduced version based on only five transcription factors. Furthermore, our model is able to detect intrinsic differences among *PcG* targets in terms of lineage-related target plasticity, although the constancy of the DNA sequence precludes it from predicting lineage-specific changes. By comparing the *PcG*-associated motif signatures between mouse and *Drosophila*, we find major differences even at highly conserved loci, suggesting major evolutionary divergence of the *PcG* targeting mechanism.



## **DREAM Poster Session 1: Thu 3:45pm-5:15pm**

(posters available for viewing Thu 3pm-Fri 2pm)

Chang	In silico prediction for regulation of transcription factors on their shared target genes - The relevant molecular pathways for promoter use	179
Chun	Reverse Engineering of Gene Regulation Network from DREAM4 Data	181
Cosgrove	Accounting for dependency within mRNA expression compendia enables accurate FDR-based edge selection in transcriptional regulatory network inference	182
Ellis	Predikin: Combining Structure and Sequence Information to Predict Phosphorylation Sites.	183
Haynes	Utilizing Global Constraints in Regulatory Network Inference	184
Hou	Integration of discrete and continuous, comparative and reconstruction-based dynamic modeling into gene network inference	185
Hurley	Combining network inference algorithms reveals insights into cancer cell biology	187
Joshi	Diverse aspects of posttranscriptional regulatory network analysis	188
King	Structure-Based Prediction of Protein-Peptide Specificity in Rosetta: DREAM4 Challenge 1	198
Lambeck	Network inference by considering multiple objectives: Insights from in vivo data captured from a synthetic network	180
Meyer	Meta-Analysis in Transcriptional Network Inference	190

## **DREAM Poster Session 2: Fri 8:15pm-9:45pm**

(posters available for viewing Fri 3pm-Sat 2pm)

Reimand	Generalised linear models of transcriptional regulation to predict process-specific factors in <i>Saccharomyces cerevisiae</i>	191
Shen	Hybrid modeling and robustness analysis of <i>Caulobacter</i> cell cycle regulatory network	192
Taylor	Optimizing GO-based similarity for stroke pathologies using networks reverse-engineered from gene expression data	193
Vescio	Automatic Generation of In-Silico Biological Networks: a Cellular Automata Approach.	194
Yu	Inferring Master Regulators of Glucocorticoid-Resistance in T-ALL	195
Zhang	Bayesian Learning and Optimization Approaches to Learning Gene Regulatory Networks	196
Zhang	A probabilistic phylogenetic model to improve regulatory network inference	197

# ***In silico* prediction for regulation of transcription factors on their shared target genes - The relevant molecular pathways for promoter use**

Li-Yun Chang<sup>1\*</sup>, Li-Yu D Liu<sup>2\*</sup>, Wen-Hung Kuo<sup>3</sup>, Hsiao-Lin Hwa<sup>1</sup>, Ming-Kwang Shyu<sup>1</sup>, King-Jen Chang<sup>3</sup>, Fon-Jou Hsieh<sup>1</sup>

1. Department of Obstetrics and Gynecology, College of Medicine, 2. Department of Agronomy, Biometry Division, 3. Department of Surgery, College of Medicine, National Taiwan University, Taipei, Taiwan, R.O.C.\*These authors contributed equally to this work.

**Introduction:** Our research findings (Liu et al. 2009; Chang et al. unpublished data, 2009) suggested that the subtypic phenotype of group IE like breast cancer can be demonstrated by estrogen receptor  $\alpha$  (ER $\alpha$ ) transcriptional regulatory network, at least in part. We thus hypothesized that the ongoing transcriptional pathways within this network may provide insights into differences of both ER $\alpha$  specific and ER $\alpha$  shared transcriptional regulatory activities on ER $\alpha$  target genes in both group IE like and group IIE like breast cancers. To test this hypothesis, we aimed at building up a network with clusters of gene sets based on their relevant nature of promoter use by co-expressed transcription factors (TFs) of interest and filling the multivariate space of the previously established network (Liu et al. 2009).

**Results:** In this study, we utilized a statistical measure of multivariate association, i.e. coefficient of intrinsic dependence (CID) for measuring the association between co-existing variables and a variable of interest, being capable for prediction on promoter use of functional TFs to their shared target genes. We typically designed data partition via hierarchical clustering before the subsequence of CID measurement. This step is to ensure the co-expressed TFs of interest to be grouped for studying functional transcriptional activities of interest. We ran a bivariate CID ( $N = 2$ ) on a dataset of thirty-seven clinical gene expression arrays (37A), in which *ESR1/E2F1*, *ESR1/GATA3* were two TFs of interest, respectively. Both E2F1 and GATA3, two TFs, also known to be the primary ER $\alpha$  target genes were selected for demonstrating the features of this method. As a result, two sets of gene pools were identified. We concluded the statistically selected genes to be regulated by ER $\alpha$  and E2F1, ER $\alpha$  and GATA3 via four mechanisms, respectively, based on their features of multivariate associations and those of univariate associations. Biologically, those regulatory events involving cross-talk between two TFs were divided into four distinct mechanisms due to partially dependent on (a) two independent transcriptional activities (Mechanism 1); (b) a TF dominant regulation (Mechanism 2 and Mechanism 3); and (c) no individual transcriptional activity (Mechanism 4).

**Conclusions:** We demonstrated two ER  $\alpha$  involved promoter use pathways that were operated by ER  $\alpha$  and a TF on the promoter of their shared target gene in a given population, respectively. They can be statistically identified by combined methods of bivariate CID, univariate CID and Galton-Pearson's correlation coefficient and are part of network motifs. Each target gene expression may be co-regulated by many TFs ( $N > 2$ ) simultaneously and/or sequentially. A simplified model of biologically pre-programmed gene expression patterns for the ER  $\alpha$  involved co-regulatory networks predicted by multivariate CID was proposed.

# Network inference by considering multiple objectives: Insights from in vivo data captured from a synthetic network

Sandro Lambeck<sup>1</sup>, Andreas Dräger<sup>2</sup>, Reinhard Guthke<sup>1</sup>

<sup>1</sup>Leibniz Institute for Natural Product Research and Infection Biology – Hans Knoell Institute, Beutenbergstr. 11a, D-07745 Jena, Germany

<sup>2</sup>Center for Bioinformatics, University of Tuebingen, Sand 14, D-72076 Tuebingen, Germany

**Background:** The reconstruction of gene regulatory networks from data is an important issue in molecular systems biology. Obtained models can be used to predict the response towards unseen perturbations and to gain insights into the architecture of the biological system under study.

**Methods:** We address the problem of finding gene regulatory networks for various data sets derived from the same underlying network but challenged to different stimuli. To this end, we propose to adapt solutions of systems of non-linear difference equations by including prior knowledge into a multiple objective optimization, facilitated by the application of an evolutionary algorithm. Furthermore, we compare the quality of results of two non-linear transfer functions, namely the logistic and double logistic function. The application is driven by data obtained from a recently published synthetic gene regulatory network in *S. cerevisiae*, being especially designed for the assessment of inference methods.

**Results:** Inferred networks were in high accordance with the designed structure and indicated the favor of a multi-objective approach. We showed benefits from including all available data and constraints such as sparseness and advantages from the use of the non-linear double logistic transfer function. In summary, quality of fit and assessment of performance indicates that multi-objective optimization can be used to integrate information from multiple datasets corresponding to the same underlying network.

# Reverse Engineering of Gene Regulation Network from DREAM4 Data

Hyonho Chun<sup>1</sup>, Minghua Deng<sup>1,5</sup>, Jia Kang<sup>3</sup>, Haisu Ma<sup>4</sup>, Xianghua Zhang<sup>1,6</sup>, Hongyu Zhao<sup>1,2</sup>

<sup>1</sup>Department of Epidemiology and Public Health, Yale University; <sup>2</sup>Department of Genetics, Yale University; <sup>3</sup>Program in Translational Informatics, Yale University; <sup>4</sup>Program in Computational Biology, Yale University; <sup>5</sup>School of Mathematical Sciences and Center for Theoretical Biology, Peking University; <sup>6</sup>Department of Electronic Science and Technology, University of Science and Technology of China;

## Background:

The objective of in-silico network challenge is to infer gene regulation networks from the provided simulated steady-state and time series data. In order to reasonably reconstruct relationships among genes in the presence of noise, one critical task is to reliably estimate the mean and variance associated with the unobserved true wild type expression level for each gene in order to identify genes with different expression levels across different experimental conditions. One common assumption made in parameter estimation based on high-dimensional genomics data is the sparsity of regulatory signals. However, based on our empirical observations in the provided datasets, the sparsity assumption may not appropriately characterize the nature of the underlying network. And for this reason, using knock-out and knock-down data (which are often perceived as the most informative datasets in inferring network connectivity) to obtain mean and variance estimates may lead to incorrect inference of the network structure.

## Method:

We develop a novel method to estimate mean and variance associated with the wild type expression using time-series data; utilizing this information, network structure can then be subsequently deduced from the knockout and knockdown data. We apply our method to the sub-challenge 2 (InSilico\_Size10) dataset.

## Results:

Our method yields networks that are capable of well characterizing the pattern of variation across different datasets. Furthermore, in datasets where the underlying networks are conspicuously non-sparse, we expect our method to outperform existing approaches that are built upon the sparsity assumption.

# Accounting for dependency within mRNA expression compendia enables accurate FDR-based edge selection in transcriptional regulatory network inference

Elissa J. Cosgrove<sup>1</sup>, Timothy S. Gardner<sup>1,3</sup>, Eric D. Kolaczyk<sup>2</sup>

<sup>1</sup>Departments of Biomedical Engineering and <sup>2</sup>Mathematics and Statistics, Boston University, Boston, MA 02215; <sup>3</sup>Amyris Biotechnologies, Emeryville, CA 94608.

Transcriptional regulatory network inference (TRNI) from large compendia of DNA microarrays has become a fundamental approach for predicting transcription factor (TF)-gene interactions at the genome-wide level. While many TRNI methods rely on user-defined or truth set-based thresholds for determining the network, correlation-based methods can theoretically utilize established tests for statistical significance to select a threshold at a desired level of prediction accuracy via control of the false discovery rate (FDR). However, recent work demonstrates that the very gene-gene correlations in question can induce what effectively amounts to a dependency among microarray experiments, or, equivalently, a reduction in sample size. This effective dependency invalidates the assumption of independent and identically distributed (i.i.d.) experiments upon which the statistical tests are based. In this work, we characterize this effective dependency in an *E. coli* microarray compendium and explore its consequences in correlation-based TRNI.

A test for i.i.d. experiments in the *E. coli* compendium was rejected, with visually evident structure based on groups of experiments from the same publication. Accordingly, we observed a substantial reduction in the effective sample size,  $n_{eff}=14.7$ , for a compendium with  $n=376$  experiments, and found that  $n_{eff}$  of select subsets of the data actually exceeded  $n_{eff}$  of the full compendium. Consistent with the latter result, we observed improved performance in TRNI using subsets of the data compared to results using the full compendium. Finally, using the set of known *E. coli* genetic regulatory interactions from RegulonDB, we demonstrated that false discovery rates (FDR) derived from  $n_{eff}$ -adjusted p-values matched FDR based on the RegulonDB truth set, while use of p-values based on  $n$  vastly over-estimated the number of true edges. These results support utilization of  $n_{eff}$  as a potent descriptor of microarray compendia and highlight a straightforward correlation-based method for TRNI with demonstrated meaningful statistical testing for significant edges, readily applicable to microarray compendia from any species, even when a truth set is not available.

# Predikin: Combining Structure and Sequence Information to Predict Phosphorylation Sites.

Jonathan J Ellis<sup>1</sup>, Neil Saunders<sup>1</sup> and Boštjan Kobe<sup>1</sup>

<sup>1</sup>*School of Chemistry and Molecular Biosciences, University of Queensland, Australia*

Phosphorylation of serine, threonine and tyrosine residues is a ubiquitous process in cellular regulation. It has been estimated that up to 2% of protein encoding genes in the human genome encode for enzymes that are responsible for protein phosphorylation, and that 30-50% of human proteins are phosphorylated. The experimental matching of specific kinases to the phosphorylation sites they regulate is a costly and laborious process. Computational methods can play an important role in reducing both the cost and time of this process.

Predikin is an algorithm that is able to prediction protein kinase binding specificity. It combines structural and sequence data that is compiled into a knowledge base from which it is able to make predictions about, potentially completely uncharacterised, kinases. However, no structural information is required for predictions; that is, predictions may be made based on just the kinases and substrate amino acid sequences.

Substrate Determining Residues (SDRs) are located within the kinase sequence. The pre-compiled knowledge-base is then queried to retrieve all kinases that have similar SDRs. For each kinase retrieved, the algorithm identifies the specific site(s) phosphorylated by that kinase where know using a confidence level provided by the user (e.g., experimentally verified, by similarity, probably etc.). The heptapeptide around each phosphorylation site identified in this manner is used to build a position specific weight matrix, and this matrix is used to score potential phosphorylation sites.

The predictions of Predikin compare favourably with other tools for protein kinases prediction, while Predikin's ability to create weight matrices "on-the-fly" means it is able to make predictions on a much wider set of sequence than some other methods.

Predikin is able to scan a kinase (or set of kinases) against a database of potential phosphoproteins to identify likely phosphorylation sites, or scan a phosphoprotein against a database of kinases to identify the most likely kinase for a specific phosphorylation site.

Predikin is able to identify kinases from their sequence. This means that whole proteomes can be scanned identifying potential kinases and predicting the specific sites they phosphorylate. This has been done for the *S. cerevisiae* proteome. This analysis has suggested functions for previously uncharacterised proteins, and has firmly establish the usefulness of Predikin as a tool that can identify links between kinases and phosphoproteins for further experimental testing.

# Utilizing Global Constraints in Regulatory Network Inference

Brian C. Haynes<sup>1,2</sup>, Michael R. Brent<sup>1,2</sup>

<sup>1</sup> Department of Computer Science, Washington University, St Louis; <sup>2</sup> Center for Genome Sciences, Washington University, St Louis

Recently, there has been a major focus on gaining a mechanistic understanding of gene regulation by using gene expression data to infer networks of regulatory interactions. However, previous efforts have neglected a key source of information inherent in any data set that contains genetic perturbations; namely, which gene was perturbed and which genes were differentially expressed in response to that perturbation. Surprisingly, genetic perturbations have generally been treated as generic examples of the possible cellular states, without regard to the particular gene perturbed in each sample. In the work described here, we show that using the identity of the perturbed gene and the genes that respond to it leads to major improvements in the accuracy of network inference. Specifically, the genes that respond to a perturbation of Gene X must be reachable from the node for Gene X via a directed path in the influence graph. This poses a global constraint on the influence graph – a constraint whose satisfaction depends on the entire graph rather than a local piece of the graph. To represent this constraint we construct a reachability graph which can be thought of as the transitive closure of the influence graph. In this work we present a new Bayesian inference method, N-sieve, for inferring the underlying structure of a transcriptional network using gene expression data. N-sieve combines a probabilistic, local measure of the fit between predictions and observations with a prior that favors structures based on reachability. We incorporate this knowledge into the model using a structural prior based on a Laplace distribution centered at perfect agreement with the transitive closure graph derived from gene perturbation experiments. We evaluate our method and compare it to other network inference methods on synthetic networks generated by GRENDL [1] and on the transcriptional network of *Escherichia coli* using gene expression data from the M3D compendium [2].

The use of this prior on a global topological feature yields substantial accuracy benefits by both evaluations. In the synthetic benchmark, this structural prior overcomes the effects of technical noise to allow for robust network inference from noisy data. For reconstructing the *E. coli* transcriptional network, we see a 25% improvement in accuracy over the next best method. We also note that the gains in accuracy produced by the local goodness of fit approach combined with the global structure constraints are greater than either one in isolation. This is exciting because it reveals that tremendous gains in network reconstruction accuracy can be achieved by finding an approximate solution to a decidedly simpler problem – identification of the network's transitive closure. This suggests that research effort should also be invested in developing algorithms that accurately reconstruct the transitive closure matrix of the interaction graph from gene expression data to apply better constraints to the original problem of inferring the influence network representing direct regulatory interactions.

[1] Haynes, B.C., Brent, M.R. (2009) Benchmarking regulatory network reconstruction with GRENDL. *Bioinformatics*, 25, 801-807.

[2] Faith, J.J. et al. (2007) Many microbe microarrays database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, 34 (Database issue), D866-D870.

# Integration of discrete and continuous, comparative and reconstruction-based dynamic modeling into gene network inference

Ping Hou<sup>1</sup>, Zhengyu Ouyang<sup>1</sup>, Yang Zhang<sup>1</sup>, Haizhou Wang<sup>1</sup>, Mingzhou (Joe) Song<sup>1</sup>

<sup>1</sup>*Department of Computer Science, New Mexico State University*

With the development of microarray technology which could provide time course and perturbation data sets. We get a chance to develop algorithms to figure out the insights of the Gene Regulatory Networks (GRNs) for biologists. It is not an easy job to show the performance of the algorithms biologically.

The Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges create a platform to evaluate current reverse engineering in bioinformatics research fields. We participate in the DREAM 4 challenge 2, which is “In Silico Network Challenge”. The challenge provided knockouts, knockdowns, time course and multifactorial noised data. Such data set types can obtain from the real biological experiment design.

The perturbed time course data set enable us to apply our new developed comparative modeling algorithm to detect the gene interactions. Traditional reconstruction-based method has limited power especially when we have limited sample size. Using comparative as a complementary can detect subtle changes under different experimental conditions even when the data size is small. At the same time while discrete logical network is better for explaining the qualitative properties, the continuous dynamical system model has an advantage over capturing the qualitative characteristics. As well as our reconstruction algorithm, we cannot only reconstruct the topology but also the dynamical system models which have ability to produce prediction under various conditions, such as double knockouts.

For DREAM 4 Challenge 2, we predicted the gene regulation networks of sub-challenges “InSilico\_Size10” and “InSilico\_Size100”. We used two models: generalized Logical networks (GLNs) and dynamical system models (DSMs). The entire modeling process involved several modeling modes of the two models, as we will explain below.

The first step is the use of comparative modeling for GLN and DSM to take advantage of the time series data given under different conditions. For the second half of time series, we observed, in some trajectories, the expression levels of gene vary dramatically compared to those in the first half. For some instance, in perturbing condition for the first half time series, gene expressions are very steady in distinct perturbation levels compared to the steady state expression, then change immediately in normal condition for the second half time series to steady state expression levels and maintain it. This discrete property can be detected by comparing different GLN models reconstructed based on the two half time series. On the contrary, for some other instance, gene expressions of the second half time series recover gradually from distinct expressions levels, while the data of the first half are steady in perturbation levels or only distribute normally around the steady state level. Similarly we compare different DSM models reconstructed based on the two half time series to capture this continuous trait. So we applied comparative modeling in both GLN and DSM to explore such dy-

dynamic properties. We separated the time course data into two categories as “with perturbation” and “without perturbation”, then use both to reconstruct the models and compare with each other to find the common part. As we assume, the common part is the inherent regulate relationship of original networks, but the different part is caused by the perturbations. We extract the common part as regulatory relationship candidate for our further analysis.

Furthermore, we deal with the knockout, knockdown, and wild-type steady state data. We applied GLN model reconstruction with zero<sup>th</sup> Markov order. In data of knockout and knockdowns, when a particular gene is inhibited, expression levels of some other genes would be extremely distinct high or low in the inhibited steady state compared to the normal steady state. As we assume that when the parent gene has been knocked out or knocked down, the expression levels of child gene will change accordingly. Based on this hypothesis, we combine the three types of steady state data for knockouts, knockdowns and wild type. Then we built GLN models from the steady state data we combined. By reconstruct the GLN models with zero<sup>th</sup> Markov order, we can analysis the steady state data to find the inherent regulate relationship of the network. We also use these regulate relationship as candidate for our future analysis.

Actually, the GRNs of this challenge using Ordinary Differential Equations (ODEs) to describe the transcription and translation of GRNs. Our final goal is to discover the mechanism GRNs work by predicting the differential equations that embed in GRNs. In the end, we used DSM modeling to capture the dynamics in the time course data. In the previous two steps, we collected some regulatory relationship candidates. Based on this regulatory relationship topology, we used these regulation genes as parent candidates to reconstruct the gene regulation networks using DSM modeling. By using these regulatory relationships as candidates, we exclude majority possible topology and focus on more credible topology of GRNs. It makes our prediction more precise. After this step, we can recover differential equations that fit reasonably well the DREAM4 data.

# Combining network inference algorithms reveals insights into cancer cell biology

Daniel Hurley<sup>1,2</sup>, Edmund J. Crampin<sup>1</sup>, Cristin G. Print<sup>2</sup>

<sup>1</sup>Auckland Bioengineering Institute, University of Auckland; <sup>2</sup>Department of Molecular Medicine and Pathology, Faculty of Medicine and Health Sciences, University of Auckland;

Gene network inference algorithms draw on a wide range of theoretical bases, and they use transcriptomic, proteomic and other data in a wide range of different ways. However, this variety of approaches has made comparison of different methods and replication of results difficult.

Beginning with a series of simple conceptual divisions that can be used to classify the stages of network inference, we have created a technology-independent computational framework to support network inference activity. We have implemented a number of published network inference algorithms into the framework, including methods using mutual information (MI), Bayesian networks, and dynamical systems approaches. Processing and visualisation components are implemented in the framework as individual separate modules, and components can be scripted to run in series. This enables one dataset to be rapidly used by several different inference algorithms to create a series of networks, increasing the productivity of network inference analysis.

To compare networks generated by different approaches, novel visualisation methods are required. Our group has developed a series of visualisation tools within the framework that can be used to compare networks in a simple parent-child format, and we will demonstrate these on networks generated by common published inference algorithms.

A common evaluation approach for gene networks is to measure the ability of an inference algorithm to recreate a canonical ('gold-standard') network given specific data. Typically, networks are compared edge-by-edge and a binary classification of true-positive or false-positive is made for each edge in the inferred network, generating sensitivity-specificity or precision-recall statistics. For a more detailed comparison, we suggest that comparing the shortest paths between nodes will show differences that are lost in direct edge-by-edge comparison. For every edge in a gold-standard network, we compute the shortest path between the corresponding vertices in the candidate network, and the results are evaluated for significance using a resampling approach. We use this technique to show differences between networks that would not otherwise be significantly differentiated by direct precision-recall statistics.

Using this set of tools, we have analysed microarray data in cancer cells, both novel data from experimental cell lines and existing published clinical datasets. We will present the comparative performance of algorithms on this dataset, as well as some new insights gained by the combination of multiple techniques.

## Diverse aspects of posttranscriptional regulatory network analysis.

Anagha joshi<sup>1,2</sup>, Yves Van de Peer<sup>1,2</sup>, Tom Michael<sup>1,2</sup>

<sup>1</sup>*Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium;*

<sup>2</sup>*Department of Molecular Genetics, UGent, Technologiepark 927, 9052 Gent, Belgium*

Gene expression is regulated at diverse levels in each cell during its life stages and in response to its environmental changes. Firstly DNA is transcribed to mRNA controlled by transcription factors. Due to advent of microarray technology, it is feasible now to built transcriptional regulatory network in diverse species. Different methods have been developed to extract regulatory information not only from expression data but also integrating protein-DNA interaction, protein-protein interactions or phylogenetic similarity. Thus the steps and regulatory programs that govern gene expression at this level are reasonably well known, much less is known about the organization of the later steps in the gene expression program. As soon as mRNA is formed, it is dynamically associated with RNA binding proteins (rbps) which defines the cellular localization, lifetime and translation rate of specific RNA transcript. Hundreds of RBPs are encoded in the eukaryotic genome, but because few have been studied in detail and few of their mRNA targets are known, the nature and extent of an RBP-mediated posttranscriptional program has been obscure.

Since the number of RBPs encoded in eukaryotic genomes approaches that of transcription factors, it has been suggested that the regulatory program that controls the posttranscriptional fate of mRNAs their localization, translation, and survival may prove to be nearly as diverse and complex as the regulation of transcription itself. The technology has made it feasible to generate large scale data at posttranscriptional level. There is a need to develop bioinformatics methods to build posttranscriptional regulatory network integrating diverse data sources. In this paper we demonstrate that though many methods are developed to infer regulatory network at transcriptional level, they can be used to unravel different aspects of regulation at posttranscriptional level.

We make an attempt to unravel various aspects of posttranscriptional regulation by integrating data incrementally from different sources. We generated a small transcriptional and posttranscriptional expression compendium under six stress conditions and concluded that the transcriptional and translational profiles are correlated in severe stress conditions in contrast to mild stress conditions where the response is observed mostly at posttranscriptional level. We used this data to build regulatory networks at transcriptional and posttranscriptional levels. The clusters obtained from posttranscriptional profiles are functionally more coherent. The regulatory network obtained from posttranscriptional profiles is equally informative about the regulation at transcriptional level as transcriptional profiles moreover it also provides information about posttranscrip-

tional regulation thus we suggest that conditional sampling at posttranscriptional level should be preferred than at transcriptional. We then integrated expression data with known RNA binding protein targets for functional characterization of RNA binding proteins. We further build an integrated network by adding known transcription factor target information. Finally we find three overrepresented network motifs in integrated networks including a novel 'regulatory loop' specific to posttranscriptional regulations.

# Meta-Analysis in Transcriptional Network Inference

Patrick E. Meyer<sup>1,\*</sup>, Benjamin Haibe-Kains<sup>1,2,\*</sup>, Gianluca Bontempi<sup>1</sup>

<sup>1</sup>Machine Learning Group, Université Libre de Bruxelles; <sup>2</sup>Functional Genomics Unit, Jules Bordet Institute

The sequencing of the human genome and the advent of high-throughput "omics" technologies, such as transcriptomics through gene expression profiling, have enabled scientists to study thousands of genes in parallel. Using these technologies, early biological studies showed that it is generally not individual genes but rather biological pathways and networks that drive an organism's response and the development of its particular phenotype.

Many biological networks have been recently shown to be involved in carcinogenesis, including signal transduction networks to transcriptional regulatory networks, among others. In order to better understand organisms and the manner in which they play out their genetic programs, various approaches for network inference have been proposed in the literature to help us comprehend not only the structure of the networks that exist, but also the interactions between oncogenes. In this paper we will focus on methods based on information theory, namely ARACNE, CLR and MRNET.

While high-throughput technologies have delivered vast quantities of data, those datasets, generally analyzed in isolation, have not yet led to models able to infer robust transcriptional networks. One of the main problems lies in the limited amount of samples within each of those datasets, making difficult the detection of functional relationships. To circumvent this difficulty, several datasets can be collected and analyzed together with the hope to improve estimation of functional relationships. However, different laboratories may use different microarray platforms and preprocessing procedures, hence producing heterogeneous datasets where noise, range and mean of each gene expression differ. Two approaches have been proposed in the literature to infer transcriptional networks from heterogeneous datasets: (i) the pooling of datasets at the gene expressions level to infer a single transcriptional network, and (ii) the pooling of networks inferred from each dataset separately in order to generate a "consensus" network. Here we propose a novel methodology working at an intermediate level that is the estimation of the mutual information matrix (MIM) on which the above-named information-theoretic inference methods rely. Indeed, once the MIM is estimated from each dataset separately, they can be pooled in a meta-analytical framework to get an overall estimate, enabling the inference of the global transcriptional network. This approach is theoretically justified by the averaging estimators theory. These three pooling methods are statistically compared on artificial datasets where the true underlying network is known, what enables a thorough performance assessment. We then apply the best method to a set of six real breast cancer microarray datasets (including more than 1000 patients) in order to discover new transcriptional interactions, potentially revealing novel therapeutic targets or prognostic genes.

---

\* The authors contributed equally to the present work.

# Generalised linear models of transcriptional regulation to predict process-specific factors in *Saccharomyces cerevisiae*

Jüri Reimand<sup>1,2</sup>, Juan M. Vaquerizas<sup>2</sup>, Annabel E. Todd<sup>2</sup>, Jaak Vilo<sup>1</sup>, Nicholas M. Luscombe<sup>2,3</sup>

<sup>1</sup>University of Tartu, Institute of Computer Science, Liivi 2, Tartu, Estonia; <sup>2</sup>EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK; <sup>3</sup>EMBL-Heidelberg Gene Expression Unit, Meyerhofstrasse 1, Heidelberg D-69117, Germany.

Transcriptional regulation is a complex and often poorly characterised process. While some regulatory circuits have been described in great detail in model organisms such as the yeast *S.cerevisiae*, the regulation of many key processes is still far from clear. Primary high-throughput evidence of transcription factor (TF) activity comes from perturbation experiments, protein-DNA interactions (ChIP-chip and ChIP-seq) and computationally predicted binding sites (TFBS).

Here, we combine a carefully reprocessed microarray compendium of 269 TF knockout mutants with ChIP-chip and TFBS data to infer process-specific transcriptional regulators. The core of our method is a multinomial generalised linear model that fits discrete variables in a probabilistic manner and provides means to include only high-confidence values (i.e. statistically significant binding and differential expression).

We benchmarked our model on cell cycle (CC) data of *S.cerevisiae*, using phase annotations (M, G1, S, G2) of cell cycle regulated genes as the predicted response variable. Our model is successful in recovering the core players of this system, as 7 of the 9 key CC TFs are present among top-10 predictors. Next, we aimed to detect regulators to less characterised cellular processes such as the yeast stationary phase (SP). When inferring TFs to genes that are active during different SP stages (diauxic shift, SP maintenance, SP exit), the model reveals attractive candidates with previous knowledge of related function, e.g. alternative carbon source regulation, stress response, response to TOR and PKA pathways, cell cycle regulation, etc. In conclusion, generalised linear models of transcriptional regulation allow the integration of high-throughput data to reconstruct process-specific regulatory systems, and may be applied to select candidates for experimental validation.

## Hybrid modeling and robustness analysis of *Caulobacter cell cycle regulatory network*

Xiling Shen<sup>1</sup>, Lucy Shapiro<sup>2</sup>, David Dill<sup>4</sup>, Mark Horowitz<sup>3</sup>, Harley McAdams<sup>2</sup>

<sup>1</sup>Faculty of School of Electrical and Computer Engineering, Cornell University, Ithaca, NY; <sup>2</sup>Developmental Biology, Stanford University, CA; <sup>3</sup>Department of Electrical Engineering, Stanford University, CA; <sup>4</sup>Department of Computer Science, Stanford University, CA.

Regulatory networks are often complex and hard to decipher. Deleting seemingly essential components such as regulatory genes does not always generate noticeable phenotypes, suggesting redundancy in such networks. On the other hand, modification to components thought to be redundant could lead to genetic defects or loss of evolutionary competitiveness, sometimes only in subpopulations or specific environment. We adopted a novel approach to studying regulatory networks by conducting experiments that target at specific hypothesis generated by analytical tools adapted from control theory and electrical circuit analysis. To demonstrate the power of this top-down systematic approach, we have chosen to model and analyze the *Caulobacter* cell cycle regulatory network.

*Caulobacter crescentus* is a model organism for studying bacterial asymmetrical division. The motile swarmer cell sheds its flagella and grows a stalk before replicating its chromosome. Master regulator proteins such as CtrA and DnaA are expressed at different times along cell cycle progression, driving or blocking various cell processes. Besides transcription, epigenetic control, phosphorylation, protein localization, and sRNA regulate the activities of the regulators as well, synchronizing them to cell cycle progressions and responding to environmental variations. To investigate this network, we first model the *Caulobacter* regulatory network as a hybrid control system, which enables system-level simulation by integrating detailed kinetic equations for known reactions with abstract phenomenological models for yet uncharacterized reactions. We then applied timing and signal analysis tools from engineering, e.g. formal verification, to examine the robustness of the modeled network's operation under diverse perturbations and environmental conditions in order to identify components that were missing in the computational model. These conditions were replicated *in vivo*, and, by comparing the *in vivo* phenotypes with the *in silico* predictions, we were able to discover previously unknown mechanisms as well as understand the subtle functional roles of these seemingly redundant mechanisms. For example, we discovered that in bacterial cell cycle regulation, basal expression from methylation-regulated promoters serve as a contingent reset apparatus for failed chromosome replication. Several "hidden" promoters of master regulators are only activated when the regulatory circuitry enters unwanted states. Additional pathways were found to exist to tightly couple chromosome replication to cytokinesis. Our analysis also showed the existence of a negative feedback loop to prevent over-initiation of DNA replication, which led to the discovery of the role of HdaA in repressing the replication initiation factor DnaA. The *in silico* model was subsequently updated with these new findings, replacing the abstract phenomenological models with detailed kinetics, which allows a new round of engineering analysis to identify even more targets. We are currently trying to model and analyze the role of phosphorylation and protein localization in the asymmetric division of *Caulobacter* cells.

## Optimizing GO-based similarity for stroke pathologies using networks reverse-engineered from gene expression data

Antonio Sanfilippo<sup>1</sup>, Ronald C. Taylor<sup>1</sup>, Jason E. McDermott<sup>1</sup>, Nat Beagley<sup>1</sup>, Bob Baddeley<sup>1</sup>, Rick Riensche<sup>1</sup>, Gene Roseberry<sup>1</sup> and Banu Gopalan<sup>2</sup>

<sup>1</sup>*Pacific Northwest National Laboratory, Richland, WA*

<sup>2</sup>*Genomic Medicine Institute, Cleveland Clinic, Cleveland, OH*

Inference of transcriptional regulatory networks (as well as intelligent automated annotation of such) forms a core subarea of systems biology. Cross-Ontological Analytics (XOA, Sanfilippo et al., 2007) is a new algorithm that can link genes into networks using aggregated semantic similarities between Gene Ontology (GO) annotations found for those genes in the GO database. XOA performs this task by “translating” associative links across the biological process, molecular function and cellular component GO sub-ontologies into a hierarchical links within a single sub-ontology, so that all GO similarities can be computed as intra-ontological relationships and therefore yield commensurable scores. Thus, XOA can take as input a gene pair and establish the presence and strength of a link (edge) between the gene pair in terms of a single score. The resulting network formed by such edges provides new and useful information as to functional and possible regulatory relationships between the genes. As described in Sanfilippo et al. (2009), the basic XOA algorithm can be specialized to a specific network using simulated annealing (SA) to select the set of GO annotations over which XOA will operate to establish links across gene pairs. Using the TGFB and other signaling pathways from the NetPath database, we have shown that such supervised learning of a modified set of GO annotations improves on the scoring used in the basic XOA algorithm, filtering out XOA scores which lead to false positives. In this talk/poster, we show that such SA-based optimization of XOA can be generalized for use on any set of genes that have adequate expression data available, with the network to be used for XOA training being dynamically derived from the expression data rather than from literature or public databases. Here, the reference regulatory network is inferred using the Context Likelihood of Relatedness (CLR) algorithm (Faith et al, 2007). This CLR-based network is built from edges having very high CLR scores, with each edge selected having a score which passes a cutoff on the set of all gene pair-wise scores calculated using data from a set of microarrays measuring gene expression in a mouse model of stroke. The XOA algorithm is trained via SA using this CLR network, resulting in a separate (but overlapping) XOA network between the set of genes used in the CLR-built network. Using our results, we describe interesting stroke-related biological findings in the network returned by CLR and in the network found by the (trained) XOA algorithm, e.g., pointing out new subnetworks found by XOA that are missed by CLR at the given cutoff. We also identify important genes in the network using topological analysis of the integrated CLR-XOA networks. We show that combined use of the XOA algorithm with a network inference algorithm operating on gene expression data can produce a more informative, expanded joint network, and that such results can be produced with a minimal amount of additional work compared to use of the gene expression based inference algorithm by itself.

XOA web site: <http://xoa.pnl.gov>; corresponding author: ronald.taylor@pnl.gov

# Automatic Generation of In-Silico Biological Networks: a Cellular Automata Approach.

Basilio VESCIO<sup>1</sup>, Carlo Cosentino<sup>1</sup>, Francesco Amato<sup>1</sup>

<sup>1</sup>Department of Clinical and Experimental Medicine, "Magna Graecia" University of Catanzaro, Italy

The increasing number of novel theoretical and numerical reverse-engineering tools developed in the field of systems biology requires more and more quantitative data and system-level knowledge for their assessment. On the other hand, while biotechnologies have greatly evolved during the last decade, the time and cost required for experimental measurements, especially in the case of time-series data, are still high. Moreover, the topologies of most real biological networks are not completely known.

Several approaches have been proposed so far in order to generate network models with realistic topology properties and dynamics. Over the last decade, various properties have been noticed in real biological network topologies, such as small world, scale free and clustering properties, and techniques have been devised in order to reproduce some or all of them for the generation of biologically plausible models.

In our work, we try to see the problem from a holistic point of view, regarding topological properties as emergent properties that arise from complexity, showing how the latter can generate such properties, rather than trying to replicate them through a custom building algorithm.

Cellular Automata (CAs) are discrete models studied in computability theory, mathematics, theoretical biology and microstructure modeling. Moreover, they are able to describe complexity in natural systems, at different levels of organization. We use Life-like cellular automata, based on Conway's famous Game of Life, to reproduce chaotic behaviours and self-organization, leading, under certain energetic and entropic conditions, to the generation of patterns of interaction exhibiting small world, scale free properties and motifs commonly found in real biological networks.

By adding dynamics to the generated network structures, this approach can be used to build *in silico* models and experiments for the performance assessment of reverse-engineering methods on gene networks. Besides, it may help to gain new knowledge on the evolutionary and self-organizing mechanisms of biological systems.

# Inferring Master Regulators of Glucocorticoid-Resistance in T-ALL

Jiyang Yu<sup>1,2</sup>, Wei Keat Lim<sup>1,2</sup>, Giusy Della Gatta<sup>3</sup>, Adolfo Ferrando<sup>3</sup>, Andrea Califano<sup>1-3</sup>

<sup>1</sup>: Department of Biomedical Informatics; <sup>2</sup>: Center for Computational Biology & Bioinformatics; <sup>3</sup>: Institute of Cancer Genetics and Herbert Irving Comprehensive Cancer Center, Columbia University, New York, New York, USA.

**Abstract:** Due to their ability to induce apoptosis in immature lymphoid cells, Glucocorticoids (GC) constitute an important pharmacological option in the treatment of T-cell acute lymphoblastic leukemia (T-ALL). GC-resistance in T-ALL patients is thus an important clinical problem and a major contributor to therapeutic failure [1]. We tested whether interrogation of molecular interaction networks could help identify Master Regulator of GC-resistance in T-ALL, which could then be used as pharmacological targets to re-sensitize cells to GC treatment. This hypothesis is supported by the recent discovery that inhibition of NOTCH1 can increase GC-sensitivity in T-ALL.

We developed a transcriptional interaction network for T-ALL, by analyzing a large datasets of 228 T-ALL gene expression profiles (GEP) using the ARACNe reverse-engineering algorithm [2]. We then developed a GC-resistance signature by comparing 10 GC-resistant and 22 GC-sensitive cell lines. Specifically, we built a Probit regression model with Wald test-statistic to score the association of each gene with one of two phenotypes (GC-resistant, GC-sensitive); this model significantly outperformed t-like statistics, Logit model, and related statistical tests. Finally, we used Gene Set Enrichment Analysis (GSEA) with Efron's novel "maxmean" statistic [3] to discover TFs whose activated (repressed) targets were highly enriched in GC-resistance (GC-sensitivity) signature genes (FDR < 0.05).

Computational predictions were then integrated with results from genome-wide, pooled shRNA screens [4] using the GC-resistant cell line (CUTLL1) treated with either GC (dexamethasone) or vehicle only (DMSO). Among the inferred Master Regulators, we selected those whose shRNA-mediated silencing resulted in a significant GC-sensitivity increase ( $p < 0.10$ ). Seven significant MRs were identified by this procedure, of which two (ZNF148, EWSR1) were previously reported to associate with reduced chemotherapeutic efficacy in the literature [5-6]. Interestingly, four of the seven are zinc finger proteins (ZNF) suggesting that the ZNF TF family might control key genes responsible for GC-resistance in T-ALL. We are currently extending the approach to identify post-transcriptional and post-translational acting proteins.

[1] Pui, et al, New England Journal of Med, 2004

[2] Margolin, et al, Nature Protoc, 2006

[3] Efron & Tibshirani, 2006

[4] Silva, et al, Nature Genetics, 2005

[5] Okada, et al, Anticancer Research, 2006

[6] Howard, et al, Cancer Research, 2001

## Bayesian Learning and Optimization Approaches to Learning Gene Regulatory Networks

Chaoyang Zhang<sup>1</sup>, Peng Li<sup>2</sup>, Ping Gong<sup>3</sup>, Youping Deng<sup>3</sup>, Edward Perkins<sup>4</sup>

<sup>1</sup>*School of Computing, University of Southern Mississippi*; <sup>2</sup>*Laboratory of Molecular Immunology, National Heart, Lung and Blood Institute, National Institute of Health*; <sup>3</sup>*SpecPro Inc.*; <sup>4</sup>*Environmental Laboratory, U.S. Army Engineer Research and Development Center*

Modeling and reconstruction of biological networks is a challenging inverse problem because of its nonlinearity, high dimensionality, non-uniqueness, sparse and noisy data, and significant computational cost. Several approaches have been proposed for reverse engineering of biological networks. They include Probabilistic Boolean Network (PBN), Bayesian Dynamic Bayesian Network (DBN), Ordinary Differential Equations (ODE), Partial Differential Equations (PDE). PBN and DBN are two widely used approaches. Their performance have been evaluated and compared using yeast cell cycle datasets in our previous work, and the results show that in most comparison cases, DBN outperforms PBN in term of accuracy. PBN can monitor the dynamic behavior in complicated systems based on large amounts of gene expression data, but the accuracy is relatively low. The DBN is better suited for characterizing time series gene expression data than its static version but it is inefficient or even infeasible to infer large biological networks from time series data. To better model the dynamic system of gene regulation and improve accuracy, we proposed and developed a new framework that integrates a state-space forward model and a Bayesian learning and optimization inverse model for gene network reconstruction. In the inverse model, Kalman filter, Kalman smoother and Expectation-Maximization algorithm are used to learn the gene regulation relationships. This framework takes advantage of deterministic algorithms while capturing the statistical nature in gene regulation. Theoretically, it can overcome several limitations in the existing approaches and significantly improve the accuracy and speed. The methods and algorithms have been implemented in Matlab and evaluated by both synthetic and experimental time series gene expression data. The synthetic datasets were generated using GeneSim and GeneNetWeaver. The performance of the new approach was also evaluated using DREAM3 datasets and compared with those we obtained by PBN and DBN in 2008. The evaluation results in terms of precision and recall show that the performances of DBN and the new approach are compatible, and both approaches outperform PBN. The new approach is much more computationally efficient than DBN and the speedup can be up to 10, depending on specific network topology. In addition, the new approach was applied to earthworm (*Eisenia fetida*) time series gene expression data collected in the Environmental Laboratory, U.S. Army Engineer Research and Development Center. Preliminary result of the neurotransmission networks reconstructed from experimental data shows that the new approach can identify hub genes that could not be found by DBN. This approach is being applied to DREAM4 dataset and we expect to obtain some promising results in learning biological networks.

# A probabilistic phylogenetic model to improve regulatory network inference

Xiuwei Zhang<sup>1,2</sup>

<sup>1</sup>Laboratory for Computational Biology and Bioinformatics, EPFL (Ecole Polytechnique Fédérale de Lausanne), Lausanne, Switzerland; <sup>2</sup>Swiss Institute of Bioinformatics.

The determination of transcriptional regulatory networks remains a difficult problem in biology, while the computational methods to reconstruct these networks still suffer from poor prediction. In part, this is due to the limited data and knowledge that we have about any single organism. We develop a probabilistic phylogenetic model which improves the prediction of the regulatory networks for a family of organisms using the phylogenetic relationships among these organisms.

The main part of this phylogenetic model uses the structure of the phylogeny of this family, where each node represents the true regulatory network for corresponding modern or ancestral organism, and each edge represents the evolution of the regulatory network from the parent organism to the child organism. We then add one node for each modern organism, representing the observed regulatory network for this organism, which is a noisy network that we can obtain by existing inference methods, and draw an edge from the true network to its noisy network. With the noisy networks of all modern organisms as input to this model, we can efficiently infer the true networks by our refinement algorithm, which is a dynamic programming algorithm to maximize the likelihood of the whole graph.

Our phylogenetic model is very flexible and its refinement algorithm can be adapted to work with various network evolutionary models. We give algorithms for two different network evolutionary models: one considers only the gains and losses of regulatory connections during evolution, and the other is an extension of this basic model, which also allows duplications and losses of genes. With the extended network evolutionary model, the input networks are allowed to have different gene contents for different organisms; thus our refinement model can handle more complex data automatically.

To test the performance of our refinement model, we use various standard network inference algorithms to infer regulatory networks for the chosen family of organisms from their gene-expression data, then use our refinement model to refine the inferred networks, and compare networks after and before refinement. We plot ROC (receiver-operator characteristic) curves for each algorithm, and the results on both simulated and biological data show significant improvement of our refinement model over the standard network inference algorithms with both network evolutionary models. We also compare our method with previous work which uses phylogenetic information to improve network inference, and the ROC curves of our method consistently dominate those of previous work under comparable conditions. Our refinement algorithms also allow one to specify confidence values for the entries of the input noisy networks, which adjust the impact of phylogenetic information on regulatory connections. Networks refined with appropriate confidence values are shown to have even better accuracy. Besides the refined networks, the ancestral networks inferred by our refinement algorithms from real biological data can provide insights on the mechanism of regulatory network evolution. We are currently extending this phylogenetic model to integrate the evolution of transcriptional factor binding sites, which is better studied than the evolution of regulatory networks.

# Structure-Based Prediction of Protein-Peptide Specificity in Rosetta: DREAM4 Challenge 1

Chris King<sup>1,2</sup>, Phil Bradley<sup>2</sup>

<sup>1</sup>University of Washington; <sup>2</sup>Fred Hutchinson Cancer Research Center

The small number of genes in the human genome is able to produce organisms of such astounding complexity due in large part to the signaling networks created by protein-protein interactions. Such interactions are often mediated by peptide recognition domains (PRDs), in which a given protein binds specifically to one or more short, linear amino acid sequences. The ability to predict these interactions based solely on the sequence of the PRD is the first step toward prediction of entire signaling networks using only genomic data. The peptide sequence specificity of a PRD is uniquely determined by that protein's three-dimensional structure. Presented here is progress towards structure-based prediction of PRD-peptide specificity as part of the Rosetta molecular modeling package, benchmarked by participation in the DREAM4 Challenge.

Our method requires two inputs: a structural model of the target PRD, and one or more structures of homologous PRDs bound to their peptide ligands. Homologous structures are used to construct a set of geometric constraints that are then applied to a flexible-backbone protein design algorithm that generates an ensemble of peptide ligands for the target PRD. A subset of peptides from low-energy models is then compiled into a position-weight matrix (PWM) for that PRD. We present promising results from a number of different benchmark tests of increasing difficulty, including the DREAM4 Challenge 1 target set, and evaluate the major hurdles and strategies for increasing the accuracy of the algorithm in cases where high-resolution structural modeling becomes difficult. By using structural modeling rather than sequence-based statistical inference, it is possible to predict binding specificity for PRD families with very little experimental data, and for individual proteins that deviate substantially from canonical binding modes. Additionally, a structure-based approach offers the unique opportunity for rational design of specific peptide inhibitors for medical therapeutics and new PRD modules for synthetic biological applications.

## Systems Biology Poster Session 1: Fri 8:15pm-9:45pm

(posters available for viewing Fri 3pm-Sat 2pm)

Aho	Gene expression profile of human adipose stem cells cultured in allogeneic human serum and fetal bovine serum	246
Aho	Reconstruction and Validation of RefRec: a Global Model for the Yeast Molecular Interaction Network	266
Aid	DNA motif discovery approach adapted to ChIP-chip and ChIP-Seq data.	267
Al-Akwaa	SSBBN: Gene Regulatory Network Construction using Spectral Subtraction Denoising, Biclustering and Bayesian Network	268
Apri	How to analyze the robustness of biological models with oscillatory behavior?	223
Ay	Scalable Steady State Analysis of Boolean Biological Regulatory Networks	269
Baryshnikova	The Genetic Landscape of a Cell	224
Belcastro	CENTRO: A CoExpression NeTwoRk Omnibus for gene function and pathway discovery	247
Bonneau	Comparative analysis of genomics data collections: Multi-species Integrative Biclustering	249
Carmel	A universal nonmonotonic relationship between gene compactness and expression level in multicellular eukaryotes	225
Carson	Investigating Co-regulation Networks Using Generative Models	226
Chang	The intersect of mRNA, microRNA and protein dynamics upon down-regulation of Nanog in mouse embryonic stem cells	199
Chen	Empirical mode decomposition for time-series gene expression data de-noising and clustering	227
Clark	Characterizing Artificial Chemistries	228
Dabrowski	Effects of motif and CNS multiplicity on gene expression in subspaces of conserved eigensystems following stroke and seizures	200
Dalkic	Distinct topological changes of the different cancer types	250
Degner	Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.	229
Deshpande	A scalable algorithm for discovering conserved active subnetworks across species	270
Devay	Host factors involved in HCV replication	201
Di Bernardo	Mathematical modeling of RNA interference	230
Dotu	Computing folding pathways between RNA secondary structures	231
Erkkilä	Incorporating spatial information of heterogeneous cell populations into Bayesian mRNA- and miRNA-expression analysis	251
Gentles	A pluripotency signature predicts histological transformation and influences survival in follicular lymphoma patients	202
Goel	Dynamic Flux Estimation – A novel framework for metabolic pathway analysis	271

Guan	Sampling Bayesian Network with Fast Mixing MCMC	232
Gyenesei	Functional Inference from a Genome-Wide in situ Hybridization Atlas of the Mouse Embryo	252
Habib	Aromatase inhibition in a transcriptional network context	203
Ilyin	Functional annotation with TOPOFIT-DB including non-sequential relations	233
Jang	Simulation-based Perturbation Studies: Genome-Wide Cause and Effect predictions of mRNA Expression under Perturbation	234
Keränen	On computational analysis of quantitative, 3D spatial expression in <i>Drosophila</i> blastoderm	253
Kim	When Two Plus Two Doesn't Equal Four: Modeling Non-Additive, Non-Modular Enhancer Behavior in the <i>Drosophila melanogaster</i> <i>eve</i> Promoter	235
Kivinen	Selection of an optimal set of blood biomarker proteins	254
Komorowski	Local molecular interdependency networks underlying HIV-1 resistance to reverse transcriptase inhibitors	204
Konieczka	Evolution of the High Osmolarity Glycerol (HOG) stress response network across Ascomycota fungi	205
Lachmann	GATE: Grid Analysis for Time-Series Expression	255
Lahesmaa	SATB1 dictates expression of multiple genes including IL5 human involved in T helper cell differentiation	206
Lambeck	Reconstruction of a dynamic regulatory map from murine liver regeneration data	272
Lee	Evolvability of the expression pattern of the <i>Drosophila</i> gap-gene system	256
Leiserson	Inferring Fault Tolerance from E-MAP Data	273
Li	Human Cancer Proteome Variation Database and Mutated Peptides Identification in Shotgun Proteomics	257
Lin	Modeling Idiopathic Pulmonary Fibrosis Disease Progression based on Gene and Protein Expression	258
Liu	A tri-partite clustering analysis on microRNA, gene and disease model	259
Maas	The RNA Editing Dataflow System (REDS) for the transcriptome-wide discovery of RNA modification sites	274

**Systems Biology Poster Session 2: Sat 3:45pm-5:15pm**  
 (posters available for viewing Sat 3pm-Sun 12pm)

Mackenzie	Life After Comparative Genomics; Regulatory Systems, Homeostasis, Synergy, SNPs and Disease.	260
Mar	Identifying Cell Lineage-Specific Gene Expression Modules	261
Mayo	Hierarchical Model of Gas Exchange within the Acinar Airways of the Human Lung	236
Mazloom	Linking MicroRNA and mRNA Co-Expressed Clusters to Regulatory Networks in Cancer	275
Michaut	Exploring the Monochromatic Landscape in Yeast using Genetic Interactions and Known Pathways	276
Michoel	Enrichment and aggregation of topological motifs in integrated interaction networks	277
Miller	microRNAs preferentially target dosage-sensitive genes	278
Missiuro	Predicting Genetic Interactions in <i>C. elegans</i> using Machine Learning	279
Molina	Studying transcription bursts from modeling high temporal resolution gene expression dynamics with multi-layered Hidden Markov Models	237
Molinelli	MODELS FROM EXPERIMENTS: COMBINATORIAL DRUG PERTURBATIONS OF CANCER CELLS	207
Morine	Combined inter-organ transcriptomic and metabolic analysis reveals altered nutrient handling and novel nutrient-sensitive biomarkers of the metabolic syndrome	208
Ng	Identifying uncharacterized genes and functional networks in human autophagy	280
Nir	Data Integration for High-throughput Morphological and Transcriptional Genetic Screens	262
Pandey	An Association Analysis Approach to Biclustering	263
Pando	Adaptation of a synthetic gene circuit through diverse evolutionary paths	238
Parnell	Network analysis of gene-diet interactions for obesity	209
Pique-Regi	Genome-wide identification of Transcription Factor Binding Sites using DNase-seq footprints and other annotations	281
Pollard	Unraveling of an ancient regulatory pathway: RNAi insensitivity in the germline of <i>C. elegans</i>	282
Pu	Prediction of Chromatin Modification (CM)-related Functional Domains and Genes in Human	283
Qian	Effective identification of conserved pathways in biological networks using hidden Markov models	284
Rossetti	Epigenetic silencing of a tumor suppressor network unmasks the dual face of master cell signals	210
Roy	Learning probabilistic networks of condition-specific response: Digging deep in yeast stationary phase	285
Sahoo	Discovery of a branchpoint between B cell and T cell development using MiDReG	211
Sales	Bayesian Nonparametric Clustering of Temporal Gene Expression: A Case Study of Dendritic Cells Modulation of the Immune Response	212
Schwank	Organ growth control – from classical genetics to a systems level approach	213

Selvarajoo	Evidence for simple governing rules in complex biological networks	239
Turinsky	Literature Curation of Protein Interactions: Discrepancies Across Major Public Databases	240
Upadhyay	PPI module for visualization and analysis of protein-protein interfaces in Friend.	286
Vacic	Estimating significance of CNV-pathway associations in schizophrenia	214
Van Mourik	Continuous-time modeling of cell fate determination in Arabidopsis flowers	215
VanderSluis	Redundancy and asymmetric divergence of paralogs from genome-wide analysis of genetic interactions	287
Vandin	Identification of Significantly Mutated Pathways in Cancer	216
Venner	Networks of Evolutionary Template Matches for Prediction of Enzymatic Function	217
Vermeirssen	Composite network motifs in integrated metazoan gene regulatory networks	241
Wadelius	Nucleosomes are positioned in exons and have histone marks suggesting co-transcriptional splicing	288
Waldman	TP53 cancerous mutations exhibit selection for translation efficiency	218
Wang	Reconstructing Gene Cooperation Network of Lipotoxicity by Synergy Analysis	219
Wang	Bottom-up Engineering of Synthetic Gene Networks	242
Wetmur	Stochastic modeling for loss of imprinted mRNA expression with transcriptional pulsing	243
White	Microbial interaction web inference using metagenomic data	289
Wohlbach	Identification of genomic features novel to xylose-fermenting yeasts through comparative analyses of <i>Pichia stipitis</i> , <i>Candida tenuis</i> , and <i>Spathaspora passalidarum</i>	264
Wojtowicz	Genome-wide mapping and computational analysis of non-B DNA structures in vivo	244
Wooten	Network Based Analysis Identifies AXIN1/PDIA2 and Endoglin Haplotypes Associated with Bicuspid Aortic Valve in a European Cohort	220
Wu	A dynamic analysis of the integrated control in the hepatic insulin signaling	221
Wu	Phosphopeptide-based signatures accurately predict the response of NSCLC cell lines to tyrosine kinase inhibitors	265
Xie	Global Analysis of Human Protein-DNA Interactions for Annotated and Unconventional DNA-Binding Proteins	290
Yeang	An integrated analysis of molecular aberrations in cancer	222
Ylipää	Association of genetic features with pathways using multiple high-throughput data	291
Zaslaver	Metazoan operons accelerate transcription and recovery rates	245
Zheng	Computational Modeling of Crosstalk in Cancer Signaling Networks	292

# The intersect of mRNA, microRNA and protein dynamics upon down-regulation of Nanog in mouse embryonic stem cells

Betty Chang<sup>1-3</sup>, David Braun<sup>3,5</sup>, Nektarios Paisios<sup>5</sup>, Yun Lu<sup>5</sup>, Ravi Sachidanandam<sup>4</sup>, Ihor R. Lemischka<sup>1-3</sup>

<sup>1</sup>Department of Gene and Cell Medicine; <sup>2</sup>Black Family Stem Cell Institute; <sup>3</sup>Graduate School of Biological Sciences; <sup>4</sup>Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029; <sup>5</sup>Department of Computer Science, Courant Institute of Mathematical Sciences, Graduate School of Arts and Sciences, New York University, New York, NY 10012

Nanog, one of the core transcription factors defining embryonic stem (ES) cells, when disrupted causes impaired self-renewal and cellular differentiation. The loss of Nanog results in dramatic changes in mRNA and protein expression. Chromatin immunoprecipitation studies have localized Nanog to thousands of genomic loci including the transcription start sites of over 70 microRNAs.

Here we monitor the dynamic changes in mRNA and nuclear proteins at 1, 3 and 5 days after depletion of Nanog expression by shRNA in mouse ES cells. This results in different classes of correlation between mRNA and protein expression levels. The first where mRNA and protein expression levels for the same gene change in parallel, either positively or negatively, the second where the mRNA and protein levels diverge from one another. In the latter class, where we see mRNA levels increasing or remaining stable and the analogous protein levels decrease, we attribute some of these discordances to microRNA (miRNA) activity. To examine these discordant relationships, we predict a set of miRNAs that may target these mRNA preventing protein translation or causing transcript degradation using available databases TargetScan, miRANDA, EIMMo and PITA. Of the 72 miRNA TSS bound by Nanog, 52 miRNAs are among our set predicted to play a role in regulating our discordant mRNA-protein pairs.

To further our understanding of the role of microRNAs, we perform deep sequencing of small RNAs of 18-30nts in length from mouse embryonic stem cells over our time course of Nanog depletion to identify dynamic changes in miRNA expression. We integrate these data together with mRNA and protein expression towards a comprehensive analysis of Nanog-centric regulation in the mouse ES cell.

## Effects of motif and CNS multiplicity on gene expression in subspaces of conserved eigensystems following stroke and seizures

Michał Dabrowski<sup>1</sup>, Norbert Dojer<sup>2</sup>, Malgorzata Zawadzka<sup>1</sup>, Jakub Mieczkowski<sup>1</sup>, Bożena Kamińska<sup>1</sup>

<sup>1</sup>Nencki Institute, Laboratory of Transcription Regulation, Warsaw; <sup>2</sup>Institute of Informatics, University of Warsaw, Poland

Last year, we reported two SVD modes (eigensystems) with eigenarrays conserved between the datasets from gene expression profiling of rat brain following either stroke (in the MCAO model) or kainate-induced seizures. We reported that these two conserved modes separate concurrent genome-wide effects of biological processes of inflammation/apoptosis (mode 2) and synaptic activity (mode 3) on gene expression. We also reported identification of the motif binding transcription factor AP1 as associated with up-regulation of expression in the subspace of the mode 2, and of several motifs, including the motifs binding Creb and Egr, as associated with gene up-regulation the subspace of mode 3.

We now complement these previous findings by demonstrating a mode 2 and MCAO-specific antagonistic effect of the motif binding protein Satb1 on the AP1-driven up-regulation of gene expression. Motif binding Satb1 on its own was associated with down-regulation of gene expression in the subspace of mode 2. The effect of Satb1-binding motif on gene log-expression was linearly dependent on the count of this motif in all the putative regulatory regions of each gene. Satb1 is the most characterized nuclear matrix associated region (MAR) binding protein, involved in regulation of apoptosis and inflammation. Our results suggest a role of chromatin conformation in regulation of the response to the ubiquitous transcriptional regulator AP1.

Among the previously identified transcription factors regulating mode 3, we report a highly significant linear effect of the count of the motif binding Creb on gene log-expression, following both MCAO and the kainate-induced seizures. Interestingly, we find that another simpler variable, namely the count of CNSs (conserved non-coding sequences) per gene, is also linearly proportional to gene log-expression in the subspace of neuronal-activity specific mode 3. This effect, however, is dependent on the CNSs' content of Creb-binding motifs. These findings suggest that the more numerous CNSs that had been reported before for neuron-specific genes may reflect not only a more complex regulation, but also a need for their strong activation (via Creb) in response to rapidly changing synaptic activity.

## Host factors involved in HCV replication

Piroska Devay<sup>1</sup>, Brigitte Wiedmann<sup>1</sup>, Alex Gaither<sup>2</sup>

Novartis Institute of Biomedical Research, <sup>1</sup>Infectious Disease, <sup>2</sup>Developmental and Molecular Pathways

The current therapy for HCV is combination of interferon-alpha and ribavirin. This treatment is not equally effective against all viral genotypes and it is often poorly tolerated due to side effects. A widely pursued strategy is to inhibit viral proteins essential for replication. Alternatively, targeting host proteins required for viral propagation can be an equally potent strategy to treat HCV infection.

To identify host factors necessary for viral replication several gene candidates identified in an siRNA screen performed at Novartis in a subgenomic HCV replicon model system and genes reported in the literature in similar screens (1, 2, 3) were analyzed and compared to host factors identified by siRNA screens in infectious virus model systems (4). The dataset generated was also compared to a list of potentially important host proteins reported by de Chassey et al (2008) obtained in a yeast-two-hybrid system (5).

The analysis of data obtained in subgenomic HCV replicon model systems has revealed that pathways related to apoptosis and cell survival have a key role in viral replication. Interestingly, among host factors required for the replication/propagation of infectious virus developmental processes and signaling pathways key to the immune response were the best represented.

[1] Tai, WA, Benita, Y, Peng, LF, Kim, SS, Sakamoto, N, Xavier, RJ and Chung, RT (2009): A functional genomic screen identifies cellular cofactors of Hepatitis C replication. *Cell Host and Microbe* 5: 298-307

[2] Ng, TI, Pilot-Matias, T, He, Y et al. (2007): Identification of host genes involved in Hepatitis C virus replication by small interfering RNA technology. *Hepatology*, 45 (6):1413-1421

[3] Borawski J, Troke P, Puyang X, Gibaja V, Zhao S, Mickanin C, Leighton-Davies J, Wilson CJ, Myer V, Cornellataracido I, Baryza J, Tallarico J, Joberty G, Bantscheff M, Schirle M, Bouwmeester T, Mathy JE, Lin K, Compton T, Labow M, Wiedmann B, Gaither LA. Class III phosphatidylinositol 4-kinase alpha and beta are novel host factor regulators of hepatitis C virus replication. *J Virol*. 2009 Oct;83(19):10058-74.

[4] Li, Q, Brass, AL, Ng, A, Hu, Z et al (2009): A genome-wide genetic screen for host factors required for Hepatitis C virus propagation. *Proc. Natl. Acad. Sci.*, Aug 27

[5] De Chassey, B, Navaratti, V, Tafforeau, L et al. (2008): Hepatitis C virus infection protein network. *Molecular Systems Biology*, 4 (230): 1-12

# A pluripotency signature predicts histological transformation and influences survival in follicular lymphoma patients

Andrew J. Gentles<sup>1</sup>, Ash A. Alizadeh<sup>2,3</sup>, Su-In Lee<sup>4</sup>, June. H. Myklebust<sup>3</sup>, Catherine M. Shachaf<sup>5</sup>, Babak Shahbaba<sup>6</sup>, Ron Levy<sup>3</sup>, Daphne Koller<sup>4</sup>, Sylvia K. Plevritis<sup>1</sup>.

<sup>1</sup>Department of Radiology, Stanford University; Department of Medicine (Divisions of

<sup>2</sup>Hematology and <sup>3</sup>Oncology), Stanford University; <sup>4</sup>Department of Computer Science, Stanford University; <sup>5</sup>Department of Microbiology and Immunology, Stanford University;

<sup>6</sup>Department of Statistics, University of California at Irvine.

Histological transformation of follicular (FL) to diffuse large B cell lymphoma (DLBCL-t) is associated with accelerated disease course and drastically worse outcome, yet the underlying mechanisms are poorly understood. We show that a network of gene transcriptional modules underlies histological transformation (HT). Central to the network hierarchy is a signature that is strikingly enriched for pluripotency-related genes. These genes are typically expressed in embryonic stem cells (ESC), including MYC and its direct targets. This core ESC-like program was independent of proliferation/cell-cycle and overlapped, but was distinct from, normal B-cell transcriptional programs. Furthermore, we show that the ESC program is correlated with transcriptional programs maintaining tumor phenotype in transgenic MYC-driven mouse models of lymphoma. Although our approach was to identify HT mechanisms, rather than to derive an optimal survival predictor, a model based on ESC/differentiation programs stratified patient outcomes in the training dataset as well as in an independent validation set. The model was also predictive of propensity of FL tumors to transform. Transformation was associated with an expression signature combining high expression of ESC transcriptional programs in combination with reduced TGF- $\beta$  signaling. Together, these findings suggest a central role for an ESC-like signature in the mechanism of HT and provide new clues for potential therapeutic targets.

## Aromatase inhibition in a transcriptional network context

Tanwir Habib<sup>1</sup>, Edward J Perkins<sup>2</sup>, Daniel Villeneuve<sup>3</sup>, Gerald Ankleky<sup>3</sup>, David Bencic<sup>4</sup>, Nancy Denslow<sup>5</sup>, Li Liu<sup>6</sup>, Natàlia Garcia-Reyero<sup>7S</sup>

<sup>1</sup>BTS, Vicksburg, MS, USA; <sup>2</sup>Environmental Laboratories, US Army Corps of Engineers, Halls Ferry Road, Vicksburg, MS, USA; <sup>3</sup>U.S. Environmental Protection Agency, ORD, NHEERL, MED, Duluth, MN, USA; <sup>4</sup>U.S. Environmental Protection Agency, Cincinnati, OH, USA; <sup>5</sup>Department of Physiological Sciences and Center for Environmental and Human Toxicology, University of Florida, Gainesville, FL, USA; <sup>6</sup>ICBR, University of Florida, Gainesville, FL, US; <sup>7S</sup>Department of Chemistry, Jackson State University, Jackson, MS, USA

A variety of chemicals in the environment have the potential to inhibit aromatase, an enzyme critical to estrogen synthesis. We examined the responses of female fathead minnow ovaries (FHM, *Pimephales promelas*) to a model aromatase inhibitor, fadrozole, using a transcriptional network inference approach. Fish were exposed for 8 days to 0, or 30mg/L fadrozole and samples and then left in clean water for 8 more days. Samples were analyzed for significant changes in the gene expression with a 15,000 probe FHM microarray. The top 1674 significantly changed genes based upon 1.5-fold change and  $P < 0.05$  across all the time points, including some additional genes relevant to the Hypothalamus-Pituitary-Gonadal (HPG) axis as well as sex steroid levels, were chosen for network modeling. In order to gain biological understanding of the significantly expressed genes, we also analyzed the functional annotations. Some of the gene overrepresented ontology annotations were lipid, fatty acid and steroid metabolism, signal transduction, oxidoreductase, kinases, localization, cell signalling, and calcium ion transport. StAR-related lipid transfer was the most highly connected gene in the network model. Key HPG genes such as chorionic gonadotropin beta, low density lipoprotein, steroidogenic acute regulatory protein, cytochrome P450 family members, and estrogen receptor were found significantly expressed with the fadrozole exposure and were present in the steroidogenic network obtained from the source network.

Our results showed that the inferred network was extremely successful in detecting HPG axes interactions. Some of these interactions that were previously known included gonadotropin-releasing hormone receptor and its interaction with G-proteins, adenylate cyclase, and gonadotropin. The interaction network also suggested the role of calcium in association with cAMP in the stimulation of steroidogenesis in the gonads.

## Local molecular interdependency networks underlying HIV-1 resistance to reverse transcriptase inhibitors

*Marcin Kierczak*<sup>1\*</sup>, *Michał Damiński*<sup>2\*</sup>, *Jacek Koronacki*<sup>2</sup>, *Jan Komorowski*<sup>1,3</sup>

1 The Linnaeus Centre for Bioinformatics, Uppsala University and The Swedish University of Agricultural Sciences, Uppsala, Sweden.

2 Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland.

3 Interdisciplinary Centre for Mathematical and Computer Modeling, Warsaw University, Warszawa, Poland.

\* Authors contributed equally.

HIV-1 resistome is a complex system of interdependent mutations. In order to uncover molecular interdependency networks that form HIV-1 reverse transcriptase (RT) resistome, we applied a Monte Carlo-based approach to feature selection and interdependency discovery. By considering mutation-induced changes in the physicochemical properties of mutating amino acids, we were able to reveal local interdependency networks that underlie resistance to six anti-viral drugs. We selected significant properties ( $p$ -value  $\leq 0.05$ ) and analyzed the networks of the 20% strongest interdependencies between them. The topology of each network was validated by mapping it onto the 3D structure of RT and by relating the findings to the published knowledge. Local interdependency networks are easy to read since they usually count less than 10-15 nodes.

Rapid emergence of drug resistant HIV-1 mutants is the major cause of many treatment failures. A number of individual drug resistance mutations is known today but the way they interact to create resistance often remains an open question. So far this question could be answered in an experimental way only. Our method shows how a systems biology approach can help guiding experimental treatments towards fewer choices and deeper understanding of the interactome. The method is applicable to a wide range of similar problems in the domain of proteomics.

## Evolution of the High Osmolarity Glycerol (HOG) stress response network across *Ascomycota* fungi

Jay H. Konieczka<sup>1,3</sup>, Michelle Chan<sup>3,4</sup>, Amanda Socha<sup>3</sup>, Ilan Wapinski<sup>3,5</sup>, Mark Styczynski<sup>3</sup>, Courtney French<sup>3</sup>, Jenna Pfiffner<sup>3</sup>, Dawn A. Thompson<sup>3</sup>, Aviv Regev<sup>3,4,6</sup>, and Erin K. O'Shea<sup>1,3,6</sup>

<sup>1</sup>FAS Center for Systems Biology and <sup>2</sup>Dept. of Molecular & Cellular Biology, Harvard University, Cambridge, MA; <sup>3</sup>The Broad Institute, Cambridge, MA; <sup>4</sup>Dept. of Biology, Massachusetts Institute of Technology, Cambridge, MA; <sup>5</sup>Dept. of Systems Biology, Harvard Medical School, Boston, MA; <sup>6</sup>Howard Hughes Medical Institute.

Divergence in gene regulatory networks plays a major role in the evolution of every kingdom of life. While comparative studies of system evolution have been remarkably effective in identifying the functional genomic elements and tracing major evolutionary events, most studies have relied on sequence data alone. There have been few studies done in yeasts combining sequence and functional genomic data, and they underscore the power of comparative functional genomics in analyzing the function and evolution of molecular systems. Early results have been promising, but are limited in phylogenetic or biological scope, and have had no choice but to employ ad hoc approaches to the study of system evolution.

We aim to systematically analyze the structure and evolution of the gene regulatory network underlying the High Osmolarity Glycerol (HOG) in response to osmotic stress in a tractable subset of *Ascomycota* fungi. The species in this study span more than 300 million years of evolution and include the model organisms *S. cerevisiae* and *S. pombe*, as well as the human fungal pathogens *C. albicans* and *C. glabrata*. The HOG regulatory network is controlled by a highly conserved MAP-Kinase (Hog1, p38 in humans), which in budding yeast evokes a stress response to facilitate survival in challenging osmotic conditions and mediates general stress response in fission yeast.

We presented each species in our study with low, medium, and high osmotic stress and measured genomic expression response profiles over 80 minutes. From these data, we identify conserved and derived gene regulatory modules (sets of co-regulated genes) in the evolution of the osmotic stress response (OSR). In the coming months, we will identify the Hog1-activated OSR genes in each species by measuring global gene expression differences between wild-type and strains lacking Hog1. In addition to identifying the conserved and divergent components of the Hog1 OSR, this will enable more finely detailed mapping of the *cis*-elements responsible for similarities and differences in gene regulation among species. Species- and clade-specific components will provide insights into lifestyle requirements and inform understanding of the ecologies and/or histories of those species. Understanding gained from systematically identifying gene regulatory changes across broad evolutionary space will provide significant insight into the forces shaping gene regulatory programs.

## **SATB1 dictates expression of multiple genes including IL-5 involved in human T helper cell differentiation**

Helena Ahlfors<sup>1,2</sup>, Amita Limaye<sup>3</sup>, Laura L. Elo<sup>1,4</sup>, Soile Tuomela<sup>1,5</sup>, Mithila Burette<sup>3</sup>, Dimple Notani<sup>3</sup>, Kamal Gottimukkala<sup>3</sup>, Tero Aittokallio<sup>1,4</sup>, Omid Rasool<sup>1</sup>, Sanjeev Galande<sup>3,\*</sup> & Riitta Lahesmaa<sup>1,6,\*</sup>

<sup>1</sup>Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland

<sup>2</sup>National Graduate School in Informational and Structural Biology, Finland

<sup>3</sup>National Centre for Cell Science, Ganeshkhind, Pune 411007, India

<sup>4</sup>Department of Mathematics, University of Turku, Turku, Finland

<sup>5</sup>Turku Graduate School for Biomedical Sciences, Turku, Finland

<sup>6</sup>Immune Disease Institute, Harvard Medical School, Boston, U.S.

\* These authors contributed equally to this work.

Special AT-rich binding protein 1 (SATB1) is a global chromatin organizer and a transcription factor induced by interleukin-4 (IL-4) during the early T helper 2 (Th2) cell differentiation. IL-5, predominantly produced by Th2 cells, plays a key role in the development of eosinophilia in asthma. Here we show that SATB1 regulates multiple IL-4 target genes involved in Th cell polarization or function. We demonstrate that during the early Th2 cell differentiation SATB1 represses IL-5 expression and this inhibition is mediated by direct binding of SATB1 to the IL-5 promoter and recruitment of the HDAC1 corepressor. Furthermore, SATB1 knockdown induced upregulation of IL-5 is partly counteracted by simultaneous downregulation of GATA3 in polarizing Th2 cells, suggesting a competitive SATB1/GATA3 mediated regulatory mechanism for control of IL-5 transcription. Our studies provide new mechanistic insights into the stringent regulation of IL-5 expression during human Th2 cell differentiation.

# MODELS FROM EXPERIMENTS: COMBINATORIAL DRUG PERTURBATIONS OF CANCER CELLS

Sven Nelander<sup>2</sup>, Evan Molinelli<sup>1</sup>, Weiqing Wang<sup>1</sup>, Peter Gennemark<sup>2</sup> Poorvi Kaushtik<sup>1</sup>, Nicholas Gauthier<sup>1</sup>, Martin Lee Miller<sup>1</sup>, Chris Sander<sup>1</sup>

<sup>1</sup> Computational Biology Center, Memorial Sloan-Kettering Cancer Center, <sup>2</sup> Department of Mathematical Sciences, University of Gothenburg

We present a novel method for deriving network models from molecular profiles of perturbed cellular systems. The network models aim to predict quantitative outcomes of combinatorial perturbations, such as drug pair treatments or multiple genetic alterations. Mathematically, we represent the system by a set of nodes, representing molecular concentrations or cellular processes, a perturbation vector and an interaction matrix. After perturbation, the system evolves in time according to differential equations with built-in nonlinearity, similar to Hopfield networks, capable of representing epistasis and saturation effects. For a particular set of experiments, we derive the interaction matrix by minimizing a composite error function, aiming at accuracy of prediction and simplicity of network structure. To evaluate the predictive potential of the method, we performed 21 drug pair treatment experiments in a human breast cancer cell line (MCF7) with observation of phospho-proteins and cell cycle markers. The best-derived network model rediscovered known interactions and contained interesting predictions. Possible applications include the discovery of regulatory interactions, the design of targeted combination therapies and the engineering of molecular biological networks. This algorithm is being applied to cell lines related to Hepatocarcinoma and Glioblastoma. We present results and performance analysis from networks of various sizes.

# Combined inter-organ transcriptomic and metabolic analysis reveals altered nutrient handling and novel nutrient-sensitive biomarkers of the metabolic syndrome

Melissa J Morine<sup>1</sup>, Jolene Mc Monagle<sup>1</sup>, Sinead Toomey<sup>1</sup>, Clare Reynolds<sup>1</sup>, Aiveen Maaron<sup>2</sup>, Aidan Moloney<sup>2</sup>, Claire Gormley<sup>3</sup>, Peadar O'Gaora<sup>4</sup>, Helen M Roche<sup>1</sup>

<sup>1</sup>Nutrigenomics Research Group, School of Public Health, Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland. <sup>2</sup>Teagasc, Ashdown Food Research Centre, Dunsany, Co Meath. <sup>3</sup>School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Ireland. <sup>4</sup>Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland.

Nutritional systems biology (NSB) is an ambitious field, wherein the molecular-level effects of dietary habits are studied across multiple tissues and on multiple levels of organization. Metabolic syndrome (MetS) – defined as a combination of metabolic abnormalities that increase risk of diabetes and cardiovascular disease – is a fundamental study system in NSB. Although diet is not listed as a risk factor, the onset of the MetS is at least partly controlled by dietary habits acting as metabolic stressors, including energy dense, high-fat diets that promote obesity and insulin resistance. In this study, we have assessed the effects of conjugated linoleic acid (CLA) – a polyunsaturated fatty acid with anti-diabetic and anti-inflammatory properties – on gene expression profiles in liver, skeletal muscle and adipose tissue, as well as on plasma markers of the MetS, in genetically obese mice. At the end of a 28-day dietary intervention, results from plasma markers indicated a significant effect of CLA supplementation on insulin sensitivity and lipoprotein profile: plasma glucose, triglycerides and non-esterified fatty acids were all significantly reduced in mice fed the high-CLA diet ( $\alpha=0.05$ ). Because of the strong link between insulin resistance and carbohydrate handling, we applied a modified gene set enrichment analysis (GSEA) to assess the effects of CLA supplementation on carbohydrate metabolic pathways. The generic procedure in GSEA involves defining functionally cohesive gene sets (e.g., metabolic pathways), and looking for coordinated up- or down-regulation of these sets. However, in many metabolic pathways, simple up- or down-regulation is neither biologically relevant nor feasible. With a simple computational adjustment, we assessed both single- and bi-directional pathway changes and observed widespread alterations to carbohydrate metabolism. Of the 15 pathways assessed, 14 were significantly changed in liver (6 down-, 8 bi-directionally regulated;  $\alpha=0.05$ ), 9 were changed adipose (1 up-, 2 down-, 6 bi-directional;  $\alpha=0.05$ ), and 5 in muscle (1 down-, 4 bi-directional;  $\alpha=0.05$ ). These results illustrate prevalence of bi-directional regulation in metabolic pathways, and show dramatic changes to carbohydrate flux in these insulin-sensitive tissues. Finally, in a novel integration of metabolic and transcriptomic data, we applied canonical correlation analysis to assess covariance relationships between plasma markers and gene expression in liver. While typical single-gene expression analyses rank gene expression changes based on fold change or p-value, these ranking criteria are not necessarily relevant to health or clinical outcomes. Results from CCA summarize the molecular effects of CLA supplementation in a way that is both information-rich and clinically relevant, and highlight a number of novel diet-sensitive biomarkers of metabolic syndrome.

## Network analysis of gene-diet interactions for obesity

Laurence D. Parnell<sup>1</sup>, Yu-Chi Lee<sup>1</sup>, Chao-Qiang Lai<sup>1</sup>, Jose M. Ordovas<sup>1</sup>

<sup>1</sup>*Nutrition & Genomics Laboratory, JM-USDA Human Nutrition Research Center on Aging at Tufts University, Boston, MA USA.*

Genome-wide association studies (GWAS) for clinical measures of metabolic diseases such as cardiovascular disease, type 2 diabetes and dyslipidemia have uncovered many genetic variants associating with disease phenotypes. Often, those associations fail to replicate in other populations, thought largely due to genetic and environmental differences. While at least two GWAS are currently underway to specifically examine gene-diet interactions, a substantial number of gene-diet interactions have been described from approaches focused on specific genes. In a gene-diet interaction, the allele associating with a phenotype does so only when a dietary factor passes a given threshold.

We have constructed a database of over 300 such interactions pertaining to obesity and blood lipids that previously existed only in disparate research reports. Placing these triangular interactions (gene/genetic variant, phenotypic measure of obesity or dyslipidemia, and dietary factor modulating the association) together into a network has revealed some startling connections. For example, there may be deeper relationships than previously thought between blood levels of triglycerides and body mass index (an obesity measure) due to sharing of their most influential dietary interactor across several genetic variants. These and other results will be presented in detail.

The derived network allows us to traverse from traditional biochemical pathways to specific variants of genes encoding pathway constituents and metabolic fluxes to disease phenotypes and dietary/environmental factors. This will facilitate the management of genetic risk of disease onset and progression by specific dietary intervention.

# Epigenetic silencing of a tumor suppressor network unmasks the dual face of master cell signals

Stefano Rossetti<sup>1</sup>, Nicoletta Sacchi<sup>1</sup>

<sup>1</sup>*Cancer Genetics Program, Roswell Park Cancer Institute, Buffalo, NY.*

Epigenetic silencing of tumor suppressor genes is common in breast cancer cells. We found that an aberrant signaling of retinoic acid (RA) via the RA receptor alpha (RARA) results in the concerted epigenetic silencing of a tumor suppressor gene network downstream of RARA. This network includes the RA receptor beta 2 (RARβ2), which mediates RA growth-inhibitory action, and TGFBR2, the main receptor of transforming growth factor beta (TGFB). Unexpectedly, we observed that both RA and TGF beta signals, which have anticancer effects in normal cells, exacerbate the tumor phenotypic features of cancer cells that underwent loss of RARβ2 and TGFBR2 tumor suppressor activities. Apparently, as a consequence of epigenetic silencing of canonical receptors, master signals such as RA and TGFB, exploit alternate targets to promote, rather than inhibit, tumorigenesis.

This work was partially supported by the National Cancer Institute grant NCI R01-CA127614-01 (NS).

## Discovery of a branchpoint between B cell and T cell development using MiDReG

Debashis Sahoo<sup>1</sup>, Matthew A. Inlay<sup>1</sup>, Deepta Bhattacharya<sup>2</sup>, Thomas Serwold<sup>1</sup>, Jun Seita<sup>1</sup>, Holger Karsunky<sup>3</sup>, Sylvia K. Plevritis<sup>4</sup>, Irving L. Weissman<sup>1</sup>, David L. Dill<sup>5</sup>

<sup>1</sup>Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA; <sup>2</sup>Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110; <sup>3</sup>Cellerant Therapeutics, San Carlos, CA, 94070, USA; <sup>4</sup>Department of Radiology, Stanford University, CA, 94305, USA; <sup>5</sup>Department of Computer Science, Stanford University, CA, 94305, USA

MiDReG (Mining Developmentally Regulated Genes) is a general method to predict functionally important genes in a developmental pathway using a database of publicly available gene expression datasets. MiDReG bases its predictions on the gene expression patterns between the initial and terminal stages of the differentiation pathway, coupled with "if-then" rules (Boolean implications) mined from large-scale microarray databases. We used MiDReG to predict genes in B cell development using logical combinations of genes that marks hematopoietic stem cell and the mature B cell in the mouse datasets. One of the predicted genes Ly6d was studied extensively because it splits the common lymphoid progenitors (CLP) into two populations. The Ly6d- population (ALP) retains full lymphoid potential, whereas the Ly6d+ population (BLP) behaves as a B cell committed progenitor. These data demonstrate the predictive power of MiDReG in a given developmental pathway.

The database of publicly available gene expression datasets contains diverse microarray samples. We apply BooleanNet to discover invariant gene expression relationships between genes simply, as the "if-then" rules (Boolean implications). We extended this approach to test Boolean implication between complex logical combinations of genes using a concept called "virtual gene". A virtual gene is an artificial gene that is created using arbitrary arithmetic and predefined operations on the gene expression values for multiple genes. MiDReG uses a mutually exclusive Boolean implication (VA high => VB low) that holds for a given developmental pathway where VA and VB define the early precursors and differentiated conditions respectively. For B cell development we used "Kit high AND Mpl high" as VA that marks hematopoietic stem cell stage and "Cd19 high AND Cd3e low" as VB that marks mature B cell stage. MiDReG predicted genes X that satisfy "VA high => X low" and "VB high => X high". Genes encoding surface proteins were identified using GO-terms "membrane" and "membrane fraction" and filtered using commercially available antibodies. We considered 4 genes, Cd34, Cd27, Il1r1, and Ly6d, as antibodies to these proteins were readily available, and examined their surface expression during B cell development.

Ly6d expression split common lymphoid progenitors (CLP) into two different populations. We showed that Ly6d+ subset differentiates into B cell only whereas Ly6d- subset differentiates into B cell, T cell, dendritic cell, and NK cell. We examined both *in vitro* and *in vivo* conditions and observed that Ly6d- populations developmentally precede Ly6d+ population. These demonstrate the ability of MiDReG to predict a novel branchpoint in B cell and T cell development. Previously we demonstrated that MiDReG predicts functionally important genes in B cell development. Therefore, MiDReG opens the possibility of understanding less well-characterized developmental pathway.

# Bayesian Nonparametric Clustering of Temporal Gene Expression: A Case Study of Dendritic Cells Modulation of the Immune Response

Ana Paula Sales<sup>1</sup>, Feng Feng<sup>1</sup>, Thomas B. Kepler<sup>1,2</sup>

<sup>1</sup>Duke University Laboratory of Computational Immunology, Duke University, Durham, NC, 27705; <sup>2</sup>Department of Biostatistics & Bioinformatics and Immunology, Duke University, Durham, NC 27705

Dendritic cells play a central role in the immune system by sensing common pathogen patterns and activating different arms of the immune system, tailoring the immune response to the particular type of pathogen. This modulation is done primarily by the expression and release of cytokines, molecules that serve as messengers among immune cells. The molecular mechanisms by which dendritic cells translate the different stimuli into the expression of different groups of cytokines remain largely unknown.

We have developed a nonparametric Bayesian mixture model to cluster gene expression dynamics, and have used it to analyze the expression profiles of cytokines released by dendritic cells stimulated with DNA containing unmethylated CpG regions, which are commonly present in microbial genomes. In our method, the temporal expression patterns of genes are governed by underlying Gaussian processes, which are then clustered using a Dirichlet process. The covariance function of the Gaussian process accounts for the overall increase in variance in the gene expression caused by the loss of cell population synchronization over time, providing a natural representation of the dynamics of gene expression. In addition, this model overcomes several other issues present in the majority of clustering methods. For example, the model does not require the pre-specification of the number of clusters; instead it estimates this from the data. Additionally, because all of the inference is based on posterior probabilities, the model is able to provide confidence measures that account for both the uncertainty in the resulting clusters as well as the uncertainty about the number of clusters. Furthermore, our model can easily accommodate missing data and any number of replicates, and it accounts for the nonstationarity inherent in gene expression time series.

Our cluster analysis reveals that the regulation of transcription of cytokines in dendritic cells in response to CpG proceeds in several waves, with both up- and down-regulation of groups of cytokines. In particular, we identify ten clusters of tightly co-expressed cytokines, where many of them are implicated in related functions and pathways. Together the individual clusters and the group of clusters as a whole provide insight into the dynamics of cytokine expression in response to CpG stimulation.

# Organ growth control – from classical genetics to a systems level approach

Gerald Schwank, Tinri Aegerter-Wilmsen, Simon Restrepo, Konrad Basler

*Institut für Molekularbiologie, Winterthurerstr. 190, CH-8057 Zürich, Switzerland*

The *Drosophila* wing is one of the best studied organs and well suited for a systems biology approach. It originates from a group of stem cells, and grows into a pear-shaped two dimensional cell-layer of approximately 50 000. At the end of larval development patterning and growth of the organ are completed.

In the last two decades, developmental and genetic studies have revealed key mechanisms of how patterning is established. Based on these findings models were established and experimentally validated. The same developmental and cell biological toolkit used to study patterning however could not reveal how growth and the correct final size of the organ is established. Why did these traditional approaches, which focus on single genes and outputs of regulatory pathways not lead to validated models of organ growth? One reason is that tissue growth is a combination of several cell- biological processes, namely cell proliferation, cell growth, apoptosis, and the assembly of an extra cellular matrix. Another reason is that multiple signaling pathways interact to modulate and influence growth of the wing. Thus, a transition, away from the “single gene study“ to a systems level understanding is necessary. This requires expertise from different fields such as physics and computer science to generate and interpret quantitative datasets, which will serve as input for multi-scale modeling.

Here, I will present models which try to explain how the establishment of the main axes - anterior/posterior (AP), dorso/ventral (DV), and proximal/distal (PD) - regulates organ growth. In particular I will focus on the Dpp morphogen gradient, a key player in establishing the AP axis. The data I will present here demonstrates that Dpp controls organ growth by promoting cell proliferation and inhibiting cell death. I will show that Dpp regulates growth as it does patterning: not directly by activating its target genes, but via the transcriptional repressor Brinker (brk). Uncoupling of the Dpp and Brk system, and spatial and temporal manipulation of both factors further shows that Dpp regulates growth via absolute levels, and not as previously hypothesized via graded levels. In other words, the amount of Dpp and not differences of Dpp between adjacent cells drives growth. We integrated our data into a model, which explains the riddle how a gradient of a growth factor can lead to uniform growth within an organ by an additional growth-input from a second independent system. In further mathematical model, also presented here, we assume that differences of physical parameters like mechanical forces within the organ-epithelium could potentially provide such a second system. To summarize, we believe that multiple growth-inputs, which can be either biological or biomechanical factors, together lead to uniform growth of organs. To ultimately test this hypothesis in future, we will have to manipulate systematically multiple pathways at the same time, and also look at multiple read-outs in a quantitative fashion in order to generate multidimensional models. I will outline bottlenecks and technical limitations which currently still constrict us in achieving that goal, and try to provide possible solutions tackling that problem.

## Estimating significance of CNV-pathway associations in schizophrenia

Vladimir Vacic<sup>1</sup>, Shane E. McCarthy<sup>1</sup>, Seungtae Yoon<sup>1</sup>, Dheeraj Malhotra<sup>1</sup>, Vladimir Makarov<sup>1</sup>, Lilia M. Iakoucheva<sup>2</sup>, Jonathan Sebat<sup>1</sup>

<sup>1</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, <sup>2</sup>The Rockefeller University, New York, NY

Recent findings from genome-wide association and copy number variation (CNV) studies suggest that the etiology of schizophrenia includes cumulative effects of a large number of functionally diverse genes. Systems biology approaches offer insights into pathology of this complex disease which go beyond identifying individual risk alleles. Based on the premise that molecular events which disrupt one or more components of a cellular module may affect the functionality of module as a whole, these methods aim to identify common functional units that contain the risk genes. Pathway-based approaches further the understanding of the mechanisms behind the genesis of schizophrenia, and provide more robust, potentially more reproducible results. Also, individually rare risk factors unable to pass the statistical significance threshold may become significant on a test which considers a combination of events that converge on a common pathway.

In this study we examine pathway associations of groups of rare CNVs identified from a combined sample of 3,325 schizophrenia patients and 3,542 controls. We propose a novel method for estimating the significance of disease-pathway associations, tailored to the intricacies of CNV-gene and CNV-pathway relationships. CNV-gene events can be stratified into several disjoint categories based on how a CNV affects the gene structure: whether the whole gene is contained within the CNV, or only some but not all exons are affected, or the CNV is contained within an intron, or it affects the promoter region. Distinguishing between these categories is important because they are likely to have different functional effects, ranging from disrupting protein structure to changing gene dosage. Another CNV analysis-specific concern is that individual CNV-gene events may not be independent due to the possibility that one CNV observed in one sample may affect multiple neighboring genes, which may lead to accounting for the effect of one structural variation more than once. Our method addresses these issues by iteratively incorporating most significant gene-wise events from any category, and adjusting the  $p$ -values of as of yet unselected genes taking into consideration the repeated occurrence of CNVs which have previously been accounted for. The statistical significance of associations is empirically estimated using a *null* distribution obtained by randomly permuting case and control label.

The pathways we identify are consistent with findings from the previous reports, such as ephrin receptor signaling, ionotropic glutamate receptor pathway signaling, metabolism of amino acids, axonal guidance and ERK-MAPK signaling.

## Continuous-time modeling of cell fate determination in *Arabidopsis* flowers

S. van Mourik, A.D.J. van Dijk, M. de Gee, G.H. Immink, K. Kaufmann, G.C. Angenent, R.C.H.J. van Ham and J. Molenaar

*Plant Sciences Group, Wageningen University and Research Center, The Netherlands*

The genetic control of floral organ specification is currently being investigated by various approaches, both experimentally and via simulations. Most research on this topic is based on various boolean or related methods, but so far a quantitative, continuous-time approach is missing. We propose an ODE model that describes the gene expression dynamics of a representative gene regulatory network in the model plant *Arabidopsis thaliana*. The network consists of six genes that regulate each other's expression dynamics via dimer complexes. The model incorporates transcription regulation via Michaelis-Menten kinetics, decay, dimer formation, trigger mechanisms that lead to cell differentiation, and a mass balance.

The unknown parameters in the model are estimated using a novel model identification procedure. The model is validated by simulation studies of known mutant plants. Finally, simulation experiments are carried out to predict the effects of a new type of mutation, in which interaction strengths of dimers are mutated, that has yet not been applied to *Arabidopsis*.

On a qualitative level the parameter identification was satisfactory for parameters with realistic values. The mutant simulation experiment showed mostly realistic results. Only one mutant (out of five) could not be fully reproduced correctly. This confirms a reasonable predictive power with respect to genetic mutations. The model predictions for the new mutation type showed clear organ alterations. In further model development, experimental validation of these mutants could provide a valuable tool for justification or falsification of model components and of the values of the identified parameters.

# Identification of Significantly Mutated Pathways in Cancer

Fabio Vandin<sup>1</sup>, Eli Upfal<sup>2</sup>, Benjamin J. Raphael<sup>2,3</sup>

<sup>1</sup>Department of Information Engineering, University of Padova, Italy. <sup>2</sup>Department of Computer Science; <sup>3</sup>Center for Computational Molecular Biology, Brown University, Providence, RI.

Cancer is driven by somatic mutations that alter cellular signaling and regulatory pathways. Recent cancer genome sequencing studies have shown that functional mutations that drive cancer development are distributed across a large number of genes, with many mutations observed at low or moderate frequencies. This mutational heterogeneity complicates efforts to distinguish functional mutations from sporadic, passenger mutations. It is hypothesized that somatic mutations target a relatively small number of pathways. Thus, standard practice in cancer sequencing studies is to assess whether genes that are mutated at sufficiently high frequency significantly overlap known cancer pathways. However, restricting attention to both highly mutated genes and known pathways limits the ability to identify novel cancer genes. An additional source of information about gene and protein interactions is large-scale interaction networks, such as the Human Protein Reference Database (HPRD) and STRING. These resources incorporate both well-annotated pathways and interactions derived from high-throughput experiments, automated literature mining, cross-species comparisons, and other computational predictions. *De novo* identification of mutated subnetworks in these large interaction networks is a daunting challenge. The naïve approach of examining all subnetworks of a particular size is problematic. First, the enumeration of all such subnetworks is prohibitive for subnetworks of a reasonable size. Second, the extremely large number of hypotheses that are tested makes it difficult to achieve statistical significance. We introduce a new method to identify mutated subnetworks in a large interaction network and to rigorously assess their statistical significance. Our method relies on two key features. First, we use a network flow model (related to a diffusion kernel or random walk on a graph) to determine a local neighborhood of “influence” for each mutated gene in the network, and we build a reduced network from these neighborhoods. Second, we derive a novel statistical test to compute the false discovery rate (FDR) associated with the number of identified subnetworks. We tested our approach on the HPRD network and somatic mutation data from two recently published studies: (i) 601 genes in 91 glioblastoma multiforme (GBM) patients from The Cancer Genome Atlas (TCGA) project; (ii) 623 genes in 188 lung adenocarcinoma patients sequenced during the Tumor Sequencing Project (TSP). In the GBM data, we find significant subnetworks with 19 genes and 22 genes. The genes in both of these subnetworks are enriched for members of pathways known to be important for GBM: the p53 pathway ( $p < 4 \times 10^{-3}$ ) and the RTK/RAS/PI(3)K pathway ( $p < 4 \times 10^{-6}$ ), respectively. In the lung data, we find four significantly mutated pathways with 7 or more genes including the p53 pathway, the RTK/RAS/PI(3)K pathway, the ephrin receptor gene family, and the Notch signaling pathway. Mutated genes in the later pathway include the Notch receptor, the ligand Jagged, and the Mastermind transcriptional co-activator. While Notch signaling is deregulated in a number of cancers, it was not reported as mutated in the TSP publication. These results demonstrate that our method successfully recovers known pathways and identifies new target pathways from a large unannotated interaction network.

## Networks of Evolutionary Template Matches for Prediction of Enzymatic Function

Eric Venner<sup>1,2,3</sup>, A. Martin. Lisewski<sup>1</sup>, R. Matthew. Ward<sup>1,2</sup>, Serkan Erdin<sup>1,3</sup>, Olivier Lichtarge<sup>1,2,4,5</sup>

<sup>1</sup> *Departments of Molecular and Human Genetics,* <sup>2</sup> *Program in Structural and Computational Biology and Molecular,* <sup>3</sup> *W. M. Keck Center for Interdisciplinary Bioscience Training, Houston, TX 77005,* <sup>4</sup> *Department of Pharmacology and* <sup>5</sup> *Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.*

Each year, hundreds of protein structures are solved and submitted to the Protein Data Bank (PDB), a large proportion of which have no known function. Previously, we developed Evolutionary Trace Annotation (ETA) to address this problem. ETA creates function-associated 3D templates from identifies 3D clusters of functionally important amino acids on the surface of the surface of a protein of interest and matches them to clusters found on proteins within a library of known structures. Matches may be either from the protein of interest to the library, from the library to the protein of interest or both.

ETA has been found to be highly accurate in large scale controls; however, we believe there is room for improvement because the method suffers from several limitations. First, predictions lack confidence values, making analysis of the coverage/accuracy tradeoff impossible and combination with other methods or datasets cumbersome. Second, ETA does not make use of all available sequence and structure information; matches to proteins with unknown function are disregarded.

To address these issues we have turned to a network-based analysis of ETA matches with the expectation that matches to un-annotated proteins will allow non-local information in the network to act as a source for additional predictive information. To perform this analysis we construct a network in which proteins are represented by nodes and ETA matches are represented by edges. We then label the network with the known functional information and let that information diffuse through the network, resulting in a measure of influence of a particular functional label has over a particular node. Competition of many functional labels naturally leads to both a prediction and a confidence value for that prediction. Our results address ETA's limitations and demonstrate improvement in both accuracy and coverage over ETA on enzymatic functional predictions.

## TP53 cancerous mutations exhibit selection for translation efficiency<sup>§</sup>

Yedaël Y. Waldman<sup>1,2</sup>, Tamir Tuller<sup>\*1-3</sup>, Roded Sharan<sup>1</sup>, Eytan Ruppin<sup>1,3</sup>

<sup>1</sup>Blavatnik School of Computer Science; <sup>2</sup>Department of Molecular Microbiology and Biotechnology; <sup>3</sup>School of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel. \*These authors contributed equally to this work.

The tumor suppressor gene TP53 is known to be a key regulator in cancer, and more than half of human cancers exhibit mutations in this gene. Being a known tumor suppressor, one would expect cancerous mutations would decrease the levels of TP53 either by diminishing protein synthesis or by producing a truncated product. However, more than 75% of TP53 cancerous alterations are missense point mutations that lead to the synthesis of a stable full-length protein that in many cases is expressed at higher levels than the wild type p53. These intriguing results can be partially explained by recent evidence showing that these point mutations not only disrupt p53 function as a tumor suppressor but also possess gain of function (GOF) and dominant negative effects (DNE) on p53 wild type copies by binding to the latter, thus making the mutated tumor suppressor an *oncogene*. This hence brings about the possibility that TP53 mutations may be under selection in cancerous cells for increasing defected p53 oncogenic activities. One such potentially likely mechanism examined here is increasing the mutated protein levels via higher translation efficiency (TE).

Here we perform the first large scale analysis of TE in human cancer mutated TP53 variants, analyzing 17,851 point mutations reported in various tumors collected from 2081 different studies. We identify a significant increase in TE of 1.4 fold that is correlated with the frequency of TP53 mutations (P-value= $4.71 \times 10^{-5}$ ). Furthermore, mutations with known oncogenic effects significantly increase their TE compared to other TP53 mutations (P-values 0.0424 and  $2 \times 10^{-5}$ , for DNE and GOF mutations, respectively). Further analysis shows that TE may have influence both on selecting the location of the mutation and on its outcome: codons with lower TE show stronger selection toward non-synonymous mutations (P-value= $4.2 \times 10^{-3}$ ) and, for each codon, frequent mutations show stronger increase in TE as compared to less frequent mutations (P-value= $4.31 \times 10^{-4}$ ).

In addition to significant differences in TE between tumors from different origins, we find that TP53 mutations show significantly higher TE increase in progressive vs. primary tumors (TE increase fold 1.50 vs. 1.42, P-value= $3 \times 10^{-5}$ ). This result, further supporting the role of TE selection in these mutations, gives rise to the possibility that TE analysis of existing TP53 mutations may be another additional indicator that should be considered in predicting the outcome of the cancer. Finally, an analysis of both chromosomal aberrations and point mutations in TP53 in NCI-60 cancerous cell lines points to an interesting co-adaptation between TP53 point mutations and aberrations in the tRNA pool, increasing the overall TP53 TE (TE increase fold 1.825 vs. 1.640, P-value=0.03).

Taken together, these results show that TE plays an important role in the selection of TP53 cancerous mutations. As more mutational data accumulates on a large scale for other oncogenes in the future, a similar analysis should be performed to study the general scope of TE changes in cancer development.

<sup>§</sup> This work has been accepted for journal publication in *Cancer Research*. The full journal version may be found at: [http://www.cs.tau.ac.il/~tamirtul/P53\\_Website/P53\\_TE.html](http://www.cs.tau.ac.il/~tamirtul/P53_Website/P53_TE.html).

## Reconstructing Gene Cooperation Network of Lipotoxicity by Synergy Analysis

Xuwei Wang<sup>1</sup>, Xuerui Yang<sup>1</sup>, Aritro Nath<sup>4</sup>, Linxia Zhang<sup>1</sup>, Christina Chan<sup>1-5</sup>

<sup>1</sup>Department of Chemical Engineering and Material Science, Michigan State University, East Lansing, MI 48824, USA; <sup>2</sup>Department of Computer Science and Engineering,

Michigan State University, East Lansing, MI 48824, USA; <sup>3</sup>Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA; <sup>4</sup>Genetics program, Michigan State University, East Lansing, MI 48824, USA; <sup>5</sup>Center for Systems Biology, Michigan State University, East Lansing, MI 48824

Co-operation between multiple genes plays an important role in regulating complex cellular activities and the ensuing phenotypes that are observed. Identifying such cooperative interactions between genes and phenotype, complements the previous efforts to identify novel target genes by inferring gene interactions without phenotype information, or identifying single genes relevant to a phenotype.

As proof-of-concept we present an integrative strategy to construct a phenotype-specific gene cooperation network of lipotoxicity induced by saturated free fatty acids (FFAs), which contributes to the pathogenesis of a diverse array of obesity-related diseases. Considering the multiple intracellular metabolic pathways modulated by saturated FFAs that are involved in the induction of lipotoxicity, we first select a subset of relevant genes by integrating the gene expression and metabolite profiles. This strategy reflects, in part, the “multi-level” regulatory characteristic of cellular activities, and thereby is one approach that can be applied to aid in the selection of genes that are involved in the observed phenotype and reconstruction of a phenotype-specific gene network. With the pool of genes selected, we then adopt an information-theoretic concept of synergy to quantify the cooperative effects of genes on lipotoxicity and construct a cooperative gene network. Our method constructs a “phenotype-specific” gene network of cooperative gene interactions, both synergistic and antagonistic, that contribute to the induction of a phenotype.

Subsequent analyses confirmed that the method indeed captured gene pairs that provided synergistic information on the phenotypes. The two genes in the identified pairs discriminated the phenotypes much better than either of them alone, and exhibited phenotype-specific correlation patterns. Moreover, network analyses were performed to extract information at different levels, i.e., hub genes, enriched pathways and gene-pathway associations. These analyses recovered known mechanisms, as well as revealed novel target genes involved in the induction of lipotoxicity for further investigation.

# Network Based Analysis Identifies AXIN1/PDIA2 and Endoglin Haplotypes Associated with Bicuspid Aortic Valve in a European Cohort

Eric C Wooten<sup>a</sup>, Lakshmanan K Iyer<sup>a</sup>, Maria Claudia Montefusco<sup>a</sup>, Alyson Kelley Hedgepeth<sup>a</sup>, Douglas D Payne<sup>b</sup>, David E Housman<sup>c</sup>, Michael E Mendelsohn<sup>a</sup>, and Gordon S Huggins<sup>a</sup>

<sup>a</sup>MCRI Center for Translational Genomics, Molecular Cardiology Research Institute, and <sup>b</sup>Cardiothoracic Surgery Division, Tufts Medical Center, Boston MA; <sup>c</sup>Massachusetts Institute of Technology, Cambridge MA.

Bicuspid Aortic Valve (BAV) is a highly heritable congenital heart defect. The low frequency of BAV (1% of general population) limits our ability to perform genome-wide association studies. We present the application of four prioritization techniques, reducing the multiple-testing penalty by restricting analysis to SNPs relevant to BAV in a genome-wide SNP dataset from a cohort of 68 BAV probands and 830 control subjects. Two knowledge-based approaches, CANDID and STRING, were used to systematically identify BAV genes, and their SNPs, from the published literature, microarray expression studies, and a genome scan. We additionally tested Functionally Interpolating SNPs (fitSNPs) present on the array; the fourth prioritization consisted of SNPs selected by Random Forests, a machine learning approach. These approaches reduced the multiple testing penalty by lowering the fraction of the genome probed to 0.19% of the total, while increasing the likelihood of studying SNPs within relevant BAV genes and pathways. Three loci were identified by CANDID, STRING, and fitSNPs. A haplotype within the AXIN1-PDIA2 locus ( $p$ -value of  $2.926 \times 10^{-06}$ ) and a haplotype within the Endoglin gene ( $p$ -value of  $5.881 \times 10^{-04}$ ) were found to be strongly associated with BAV. The Random Forests approach identified a SNP on chromosome 3 in association with BAV ( $p$ -value  $5.061 \times 10^{-06}$ ). The results presented here support an important role for genetic variants in BAV and provide support for additional studies in well-powered cohorts. Further, these studies demonstrate that leveraging existing expression and genomic data can identify biologically relevant genes and pathways associated with a congenital heart defect.

# A dynamic analysis of the integrated control in the hepatic insulin signaling

Ming Wu<sup>1</sup>, Christina Chan<sup>1,2,3†</sup>

<sup>1</sup>Department of Computer Science and Engineering; <sup>2</sup>Department of Chemical Engineering and Material Science; <sup>3</sup>Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA; † corresponding author, krischan@egr.msu.edu

Hepatic insulin signaling plays a central role in glucose and lipid metabolism. Dysregulation of insulin response in liver is believed to associate with the pathogenesis of type 2 diabetes. The signal transduction begins with the insulin receptors on the cell membrane, and the intracellular signaling process is mediated by insulin receptor substrates (IRS), which activate many downstream effectors to regulate glucose and lipid homeostasis. Previous studies observed that the different insulin receptor substrates, IRS1 and IRS2, may function distinctly under different conditions (i.e., fasting or feeding) in response to different levels of insulin stimulation, but the underlying mechanisms of such differential regulation is unclear. Among the numerous regulators involved in the IRS1 and IRS2 pathways, we recently identified a novel player, the double-stranded RNA-dependent protein kinase (PKR), which is affected by insulin and involved in regulating insulin signaling through feedback to both IRS1 and IRS2. Our previous research discovered that IRS1 and IRS2 are differentially regulated by PKR.

Given the complexity in insulin signaling, we applied systems modeling to analyze the dynamic behaviors underlying the molecular network. Although there are many models on the physiological insulin–glucose dynamics, mathematical frameworks that can describe the interactions and regulations at the molecular level are scarce. Our group developed an IRS1-PKR regulatory network model in liver cells using a novel discrete dynamic modeling approach. However, an integrated model of both IRS1 and IRS2 regulation is required for a comprehensive understanding of the hepatic insulin responses under different levels of insulin stimulation.

Here we present a dynamic model that integrates current information on the IRS1 and IRS2 mediated insulin-signaling processes. The model extends our previous discrete model of insulin response in liver cells through IRS1 and PKR mediated feedbacks, to include the transcriptional regulation of IRS2 in the insulin signaling as well as the processes induced by IRS2. We apply a theoretical dynamic analysis, which provides a detailed mathematical explanation of the differential regulation of IRS1 and IRS2 under different conditions. Further, based on the model, we propose an underlying mechanism, involving PKR, for the “switch-like behavior” during fasting and after re-feeding. Our modeling study provides insights into the key components and essential feedbacks that may contribute to insulin resistance under different conditions.

To the best of our knowledge this is the first attempt at an integrated mechanistic modeling of the insulin pathways. The study provides a dynamic model of hepatic insulin signaling, connecting molecular mechanisms and the physiological phenomenon, to provide novel insights into the pathogenesis and treatment of type 2 diabetes.

# An integrated analysis of molecular aberrations in cancer

Chen-Hsiang Yeang<sup>1</sup>

<sup>1</sup>*Institute of Statistical Science, Academia Sinica*

Cancer is a complex disease where various types of molecular aberrations drive the development and progression of malignancies. Large-scale, simultaneous screenings of multiple types of molecular aberrations (e.g., mutations, copy number variations, DNA methylations, gene expressions) become increasingly important in the prognosis and study of cancer. Consequently, a computational model integrating multiple types of information is essential for the analysis of the comprehensive data. In this work we propose an integrated modeling framework to identify the statistical and causal relations of various molecular aberrations and gene expressions in cancer. To reduce spurious associations among the massive number of probed features, we sequentially apply three layers of logistic regression models with increasing complexity and uncertainty regarding the possible mechanisms connecting molecular aberrations and gene expressions. Layer 1 models associate gene/microRNA expressions with the molecular aberrations on the same loci. Layer 2 models associate expressions with the aberrations on different loci but have known mechanistic links. Layer 3 models associate expressions with nonlocal aberrations which have unknown mechanistic links. We apply the layered models to the integrated datasets of NCI-60 cancer cell lines and discover/reaffirm the following prominent links: (1) Protein expressions are generally consistent with mRNA expressions. (2) Several gene expressions are modulated by composite local aberrations. For instance, CDKN2A expressions are repressed by either frame-shift mutations or DNA methylations. (3) Amplification of chromosome 6q in leukemia elevates the expression of MYB, and the downstream targets of MYB on other chromosomes are up-regulated accordingly. (4) Amplification of chromosome 3p and hypo-methylation of PAX3 together elevate MITF expression in melanoma, which up-regulates the downstream targets of MITF. (5) Mutations of TP53 are negatively associated with its direct target genes but are positively associated with its related microRNAs. (6) Some co-expressed genes are not associated with observed aberrations but share consensus motifs on their promoters. (7) Hyper-sensitivity of many compounds in leukemia is likely attributed to the high activities of cell division and growth in leukemia samples. These validated analysis results justify the utility of the layered models for the incoming flow of cancer genomic data.

# How to analyze the robustness of biological models with oscillatory behavior?

Mochamad Apri<sup>1,2</sup>, Jaap Molenaar<sup>1,2</sup>, Maarten de Gee<sup>1</sup>, G.A.K van Voon<sup>1</sup>

<sup>1</sup>*Biometris, Department of Mathematical and Statistical Methods, Wageningen University, Wageningen, The Netherlands;* <sup>2</sup>*Netherlands Consortium for Systems Biology, P.O. Box 94215, 1090GE Amsterdam, The Netherlands*

It is remarkable but well-known that many biological systems such as complex molecular networks are often robust under vastly different conditions. Although these systems might experience strong internal or external perturbations, e.g., through environmental changes or noise, their important functions, such as oscillatory behavior, are preserved. Thus, robustness is an essential feature of biological systems and any mathematical model describing their behavior should also have this property. This implies the need of an efficient tool to analyze the robustness of such models.

In this talk, we concern with parametric robustness of biological models that show oscillatory behavior. Several studies have been devoted to this problem. However, they only considered the Hopf bifurcation as an indication of the robustness, that is when the periodic oscillations turn into steady-state equilibria, whereas other types of bifurcation that might occur to the periodic solution, such as period doubling which leads to a chaotic behavior, were not considered. Furthermore, to guarantee the existence of periodic solution, they give one specific value for the allowed maximum perturbation for all parameters. This implies that all parameters are considered to have the same sensitivity to the perturbation, which is not necessarily true since the model might be more sensitive to some parameters than to the others.

Here, we present an efficient method to fast analyze the parametric robustness of biological models that show oscillatory behavior. The method is based on Floquet theory and a continuation method. Starting from the nominal parameter set, we construct in an efficient way an estimation of a region in parameter space in which oscillatory behavior exists. Using this method, all bifurcations that might occur to the models with periodic behavior can be detected. They are for examples Hopf, period doubling, or Neimark-Sacker bifurcations. Additionally, indication to a global bifurcation can also be captured. As extra information, we obtain for free the period of the solution and the amplitude of the species that are involved in the model. The method is especially attractive for models with high dimensional parameter space and is also applicable if the robustness region is far from being regular shape. As examples, the method is applied to the Rosenzweig-MacArthur model that consists of 3 state variables and 2 parameters to illustrate the ideas and to the Laub-Loomis model which is a high dimensional system that consists of 7 state variables and 14 parameters.

## The Genetic Landscape of a Cell

Anastasia Baryshnikova<sup>1,2\*</sup>, Michael Costanzo<sup>1,2\*</sup>, Jeremy Bellay<sup>3</sup>, Yungil Kim<sup>3</sup>, Huiming Ding<sup>1,2</sup>, Judice L.Y. Koh<sup>1,2</sup>, Kiana Toufighi<sup>1,2</sup>, Gary D. Bader<sup>1,2</sup>, Chad L. Myers<sup>3#</sup>, Charles Boone<sup>1,2#</sup>

<sup>1</sup>*Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto ON, Canada M5S 3E1.*

<sup>2</sup>*Department of Molecular Genetics, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto ON, Canada M5S 3E1.*

<sup>3</sup>*Department of Computer Science & Engineering, University of Minnesota-Twin Cities, Minneapolis MN, U.S.A. 55455*

\* These authors contributed equally to this work

Genetic interactions define double mutant combinations that deviate from the expected phenotype. We assembled the first genome-scale genetic interaction map, examining 5.4 million gene-gene pairs for synthetic genetic interactions and generating quantitative profiles for ~75% of all genes in the budding yeast *Saccharomyces cerevisiae*. Correlation of genetic interaction profiles reveals a functional map of the cell in which genes of similar bioprocesses cluster together in coherent subsets. Specific patterns of genetic interactions predict novel gene functions, enabling us to identify novel components of an amino acid permease sorting pathway and a novel gene involved in ER/Golgi secretion. We also identified more complex regulatory connections that only emerge from a genome-scale network. In particular, we identified a select subset of yeast polarity genes that appear to be regulated by the Elongator/Urmylation tRNA modification pathway and predict that similar regulation of a subset of microtubule and cytoskeletal organization genes underlies a human disorder, familial dysautonomia. Genetic interaction profiles also highlight cross-connections between bioprocesses, mapping a global view of cellular pleiotropy, and genetic interaction degree correlated with a number of different gene attributes, enabling the prediction of genetic network hubs in other systems. Finally, the genetic landscape provides a key for interpretation of chemical-genetic interactions and drug target identification.

# A universal nonmonotonic relationship between gene compactness and expression level in multicellular eukaryotes

Liran Carmel<sup>1,2</sup>, Eugene V. Koonin<sup>2</sup>

<sup>1</sup>Department of Genetics, the Alexander Silberman Institute of Life Sciences, the Hebrew University of Jerusalem, Jerusalem 91904, Israel; <sup>2</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Analysis of gene architecture and expression levels of four organisms, *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*, reveals a surprising, nonmonotonic, universal relationship between expression level and gene compactness. With increasing expression level, the genes tend at first to become longer but, from a certain level of expression, they become more and more compact, resulting in an approximate bell-shaped dependence. Compactness of highly expressed genes has been reported in the past, but the relative compactness of genes with low expression level is unexpected. There are two leading hypotheses to explain the compactness of highly expressed genes. The selection hypothesis predicts that gene compactness is predominantly driven by the level of expression whereas the genomic design hypothesis predicts that expression breadth across tissues is the driving force. We observed that the connection between gene expression breadth in humans and gene compactness to be significantly weaker than the connection between expression level and compactness, a result that is compatible with the selection hypothesis but not the genome design hypothesis. The initial gene elongation with increasing expression level could be explained, at least in part, by accumulation of regulatory elements enhancing expression, in particular, in introns. This explanation is compatible with the observed positive correlation between intron density and expression level of a gene. Conversely, the trend toward increasing compactness for highly expressed genes could be caused by selection for minimization of energy and time expenditure during transcription and splicing, and for increased fidelity of transcription, splicing and/or translation that is likely to be particularly critical for highly expressed genes. Regardless of the exact nature of the forces that shape the gene architecture, we present evidence that, at least, in animals, coding and noncoding parts of genes show similar architectonic trends.

# Investigating Co-regulation Networks Using Generative Models

Matthew B. Carson<sup>1</sup>, Nitin Bhardwaj<sup>2</sup>, Hui Lu<sup>1</sup>

<sup>1</sup>*Bioinformatics Program, University of Illinois at Chicago;* <sup>2</sup>*Program in Computational Biology and Bioinformatics, Yale University*

Proteins, often referred to as the ‘workhorses of the cell’, are produced through the process of gene expression, during which an organism turns its genetic code (DNA) into functional units. Regulation of this expression process increases the versatility of an organism, allows for adaptation to the environment, and increases the efficiency with which resources such as sugars are metabolized by controlling when and in what quantities RNA molecules and proteins are produced. Many diseases are related to failures in one or more components of this system. Examining regulation helps us to understand how an organism evolves and develops, and how malfunctions may break down this process. Much of the control of gene expression is believed to occur by the cell’s adjustment of transcription initiation frequency. This level of control is carried out by transcription factors (TFs), and transcription factor networks (TFNs) can be used to describe the interactions between these transcription factors and their target genes.

In this work we use generative networks to model the creation of TFNs during evolution in order to understand how these networks form and develop. In particular, we examine how the number of TF partners (those that regulate common genes) scales with the number of genes a TF regulates. It has been observed that in several model organisms the degree distribution of this partnership network appears to follow an exponential saturation curve. The co-regulatory network of our generative model shows a trend similar to that of the model organisms. We show that through various modifications to our model we are able to identify the necessary conditions for this observed saturation curve characteristic. This indicates that the saturation curve seen in these co-regulatory networks could be a product of evolutionary development, during which regulators gain and lose interactions with targets over time.

# Empirical mode decomposition for time-series gene expression data de-noising and clustering

Chang-Ray Chen<sup>1</sup>, Cheng-Wei Chang<sup>1</sup>, Ian C. Hsu<sup>1</sup>

<sup>1</sup>*Department of Biomedical Engineering and Environmental Sciences, National Tsing Hua University, Hsinchu 30013, Taiwan.*

The advance of high-throughput techniques is promising and has urged interdisciplinary efforts to tackle the challenges of unrevealing the complexity of biological systems. The genome-wide expression data for cell cycle study is a typical example. However, gene expression data is noisy. Moreover, time-series gene expression data is even more complicated because of the computational challenges of noise reduction, data modeling, and pattern recognition. One of the most studied time-series gene expression cases is the yeast cell cycle data. Many algorithms have been introduced in this area for different applications, such as identifying cell cycle-regulated genes by Fourier-based methods; normalizing and sorting periodic gene expression patterns by singular value decomposition (SVD); identifying differentially expressed genes by using aligned continuous curves; and estimating missing time points by interpolation methods. In this study, we focused on identifying more unknown yeast cell cycle-regulated genes by a novel method, empirical mode decomposition (EMD).

The idea of EMD was conceived and developed since 1998. It has been intensively used in many fields of science and engineering for nonlinear and non-stationary time series analysis. EMD can be used to decompose a signal into a series of so-called intrinsic mode function (IMF) from the highest to the lowest frequencies. EMD is empirical, adaptive, and data-driven for extracting various trends from time series data as well as for acting as a noise filter.

In this study, we developed an EMD approach for processing and modeling three yeast cell cycle datasets of synchronization experiments. The results demonstrate that by removal of decomposed high frequency IMFs the periodic gene expression patterns are significantly improved as evaluated by gene-gene correlations. Furthermore, to evaluate whether EMD captures intrinsic gene expression patterns effectively, we compared de-noised data with the original data after applying SVD or principal component analysis (PCA). The de-noised data also showed significantly improved gene sorting in representation of sequential transcriptional responses to synchronization. These findings suggest that EMD is particularly suited to noise reduction and intrinsic pattern detection for time-series gene expression data. We are currently in the process of optimizing sorting parameters for gene clustering by EMD, as its ability for signal decomposition and noise reduction is powerful for identifying unknown cell cycle-regulated genes.

## Characterizing Artificial Chemistries

Edward Clark<sup>1,2</sup>, Simon Hickinbotham<sup>1,2</sup>, Susan Stepney<sup>1,2</sup>, Tim Clarke<sup>1,3</sup>, Peter Young<sup>1,4</sup>

<sup>1</sup>York Centre for Complex Systems Analysis; <sup>2</sup>University of York Department of Computer Science; <sup>3</sup>University of York Department of Electronics; <sup>4</sup>University of York Department of Biology.

Stochastic Simulations (SS) of biological systems traditionally rely on a pre-defined set of reactions. Artificial Chemistries (ACs) offer potential advantages in that as well as implementing the defined set of reactions, ACs also allow evolution to occur in an unprescribed manner that depends only on the encoding of the molecules in the simulation and the properties of the AC.

ACs are not often used to construct SS of biological systems as they have been considered “not complex enough”, “too sticky” or “too exclusive”. Here we move beyond such qualitative judgments to develop a quantitative characterization of the properties of chemistries, allowing analysis and comparison between ACs and between an AC and real chemistry. If the appropriate properties of amino acid chemistry can be measured, it may be possible to construct an artificial chemistry with similar properties. This approach is likely to provide an AC that is far more computationally tractable than a true model of amino acid chemistry.

If we have two sets of “chemicals” (artificial or otherwise), the properties of the chemistry can be obtained by measuring which chemicals from set 1 bind to which chemicals in set 2. We have constructed mathematically meaningful definitions of degeneracy and redundancy allowing such properties to be quantified.

The equations that relate the properties of the chemistry are directly useful for theorists and are currently being applied in the area of immunology to describe the properties of the chemistry between paratopes and epitopes. The method has proven to be a useful tool for investigating the validity of competing theories.

One application of this work would be to facilitate the generation of even richer artificial network data, allowing multi-component and multi-step molecular interactions between molecules and the genome.

Quantitative characterization of ACs could also provide the basis for SS of evolving cellular control systems. For example, when genes are added to a bacterial strain to produce a pharmaceutical component, they are sometimes quickly evolved out by the organism. This is a problem that can be examined by allowing evolution of an AC network, allowing the design, testing and discovery of control motifs that are more robust than those currently used.

We provide the ability to characterize the properties of an artificial chemistry. This is the foundation for development of an artificial chemistry that has properties similar to those observed in nature while still being computationally tractable. An AC with these properties could provide a great deal of insight and experience for synthetic biologists designing genetic control systems used to harness a rapidly evolving organism as a means of production.

## Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.

Jacob F. Degner<sup>1,3</sup>, John C. Marioni<sup>1</sup>, Athma A. Pai<sup>1</sup>, Joseph K. Pickrell<sup>1</sup>, Everlyne Nkadori<sup>1,2</sup>, Yoav Gilad<sup>1</sup> and Jonathan K. Pritchard<sup>1,2</sup>

<sup>1</sup>Department of Human Genetics, <sup>2</sup>Howard Hughes Medical Institute, and <sup>3</sup>Committee on Genetics, Genomics and Systems Biology, University of Chicago, 920 E. 58th St., CLSC 507, Chicago, IL 60637.

### ABSTRACT

Next-generation sequencing has become an important tool for genome-wide quantification of DNA and RNA. However, a major technical hurdle lies in the need to map short sequence reads back to their correct locations in a reference genome. Here we investigate the impact of SNP variation on the reliability of read-mapping in the context of detecting allele-specific expression (ASE). We generated sixteen million 35 bp reads from mRNA of each of two HapMap Yoruba individuals. When we mapped these reads to the human genome we found that, at heterozygous SNPs, there was a significant bias towards higher mapping rates of the allele in the reference sequence, compared to the alternative allele. Masking known SNP positions in the genome sequence eliminated the reference bias but, surprisingly, did not lead to more reliable results overall. We find that even after masking, 5-10% of SNPs still have an inherent bias towards more effective mapping of one allele. Filtering out inherently biased SNPs removes 40% of the top signals of ASE. The remaining SNPs showing ASE are enriched in genes previously known to harbor cis-regulatory variation or known to show uniparental imprinting. Our results have implications for a variety of applications involving detection of alternate alleles from short-read sequence data.

## Mathematical modeling of RNA interference

Giulia Cuccato<sup>1</sup>, Athanasios Polynikis<sup>2</sup>, Velia Siciliano<sup>1</sup>, Mario di Bernardo<sup>2,3</sup>,  
Diego di Bernardo<sup>1,3</sup>

<sup>1</sup>Telethon Institute of Genetics and Medicine (TIGEM), Naples, Italy; <sup>2</sup> Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom; <sup>3</sup> Department of Systems and Computer Science, University of Naples Federico II, Naples, Italy

RNA interference (RNAi) is a recently discovered molecular process involved in post-transcriptional regulation of RNA. RNAi is routinely used in molecular biology as a tool to understand the function of genes. Despite its widespread use, there is still the lack of a validated mathematical model that can efficiently capture the RNAi phenomenon. A mathematical model of RNAi is fundamental in systems and synthetic biology in order to carry out *in silico* investigations and give specific guidelines for *in-vivo* experiments, as well as, inform the design of synthetic biological circuits involving RNAi.

We have performed *in vivo* measurements of mRNA levels of a transgene (TTR) stably expressed in the human embryonic kidney (HEK) cell line, by quantitative real-time PCR for a large range of concentrations of siRNA oligomers directed against the TTR coding sequence (from 0 pmol to 200 pmol at 20 pmol intervals). The experiments clearly indicate the impact of RNA interference on the mRNA levels of the targeted gene. We have also analysed a set of experimental data performed by Kim et al [1], which measure the effects of RNA interference on the protein levels of the targeted gene. The last study also measures the effects of RNAi associated to the length of nucleotides of different synthetic siRNAs.

Searching the literature, we found that different models are proposed to model RNAi. These models differ significantly; for example they give different qualitatively and quantitative predictions for gene regulatory networks involving RNAi interactions between the genes. Here we study the effectiveness of these models and compare them against the *in vivo* experimental data, both on mRNA and protein levels. We show that a simple Michaelis-Menten kinetics is the most efficient way to model RNAi for both experimental data sets. One significant feature of this model is its ability to predict saturation effects of RNA silencing, which confirms recent experimental evidences that the siRNA-programmed RISC is a classical Michaelis-Menten enzyme in the presence of ATP [2].

We have also associated the parameters of the model with their underlying biological role. For example, a parameter of the model is associated to the length of the nucleotide sequence of the siRNA oligomer complementary to the target mRNA. The model confirms experimental evidences that RNA interference efficiency increases with the length of the nucleotide sequence.

We conclude that the Michaelis-Menten model has a simple mathematical form, amenable to analytical investigations and a small set of parameters with an intuitive physical meaning, that make it a unique and reliable mathematical tool. The three parameters of this model have a well defined biological meaning, and their values can be easily tuned to accommodate for different efficiencies of RNAi.

[1] Kim DH, Behlke M, Rose S, Rossi J (2005) Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nat Biotech* 23: 222-226.

[2] Haley B, Zamore P (2004) Kinetic analysis of the RNAi enzyme complex. *Nat Struct & Mol Biol* 11:7.

## Computing folding pathways between RNA secondary structures

I. Dotu<sup>1</sup>, W.A. Lorenz<sup>1</sup>, P. Van Hentenryck<sup>2</sup>, P. Clote<sup>1</sup>

<sup>1</sup>Department of Biology, Boston College, Chestnut Hill, MA 02467; <sup>2</sup>Department of Computer Science, Brown University, Box 1910, Providence, RI 02912.

Determining an optimal pathway between two nodes in a network or graph is an important problem with many applications in computational and systems biology. Depending on the notion of “optimality”, the solution can be simple or intractable. If “optimal” means the shortest path in a weighted graph, then Dijkstra’s well-known shortest path algorithm provides an elegant and computationally efficient solution. If “optimal” means that the maximum energy over all nodes in the path is least possible, then the problem is NP-complete, as recently shown by J. Manuch, C. Thachuk, L. Stacho and A. Condon, in a paper presented at the 15th International Meeting on DNA Computing and Molecular Programming, June 8-11, 2009. In this paper, we apply a technique from combinatorial optimization to provide a nearly optimal folding pathway between two RNA secondary structures. Our technique can easily be applied to analogous problems in systems biology.

Given an RNA sequence and two designated secondary structures A,B, we describe a new algorithm that computes a nearly optimal folding pathway from A to B. The algorithm, *RNA<sub>tabu</sub>path*, employs a tabu semi-greedy heuristic, known to be an effective search strategy in combinatorial optimization. Folding pathways, sometimes called routes or trajectories, are computed by *RNA<sub>tabu</sub>path* in a fraction of the time required by the barriers program of Vienna RNA Package. We benchmark *RNA<sub>tabu</sub>path* with other algorithms to compute low energy folding pathways between experimentally known structures of several conformational switches. The ***RNA<sub>pathfinder</sub>*** web server, source code for algorithms to compute and analyze pathways, and supplementary data are available at

<http://bioinformatics.bc.edu/clotelab/RNApathfinder>.

# Sampling Bayesian Network with Fast Mixing MCMC

Yongtao Guan

*University of Chicago.*

The Bayesian network is a class of statistical models where a joint distribution of quantities of interests is represented by a directed acyclic graph (DAG). The Bayesian network plays an important role in causal inference of complex systems such as gene regulatory networks. A full Bayesian analysis often requires one to sample from the posterior distribution of DAGs to take into account of model uncertainty. Despite significant recent progress, the computational aspect of network inference remains an open challenge. The main difficulty comes from the multi-modality nature of the posterior distribution on DAGs. While order-graph based sampling approaches are effective for modest amount of nodes (Friedman and Koller [1], Ellis and Wong [2]), they have difficulties to scale up because that, when evaluating the likelihood of a node order, one has to sum over all order-compatible DAGs, whose number grows exponentially with the number of nodes. In addition, it is often desirable to directly sample DAGs because it is easier to specify a prior for DAG than for node order.

In this talk I will describe a novel algorithm that can sample DAGs effectively. First I will introduce a generic Markov chain Monte Carlo sampling algorithm (Guan and Stephens [3]) that is fast mixing on a multi-modal distribution in the sense that the spectral gap of the chain is polynomially (as oppose to exponentially) small. Then I will describe a representation of DAG that is upper-triangular adjacency matrix and discuss the convenience of such a representation for local modifications of a DAG, in particular, the loop-checking for edge-reversal move. Finally I will demonstrate the effectiveness of the algorithm via simulation studies.

[1] N Friedman and D Koller. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, 50, 2003.

[2] B Ellis and WH Wong. Learning Causal Bayesian Network Structures From Experimental Data. *JASA*, 103, 2008.

[3] Y Guan and M Stephens. Small World MCMC With Tempering: Ergodicity, Spectral Gap and Applications. In Preparation.

## Functional annotation with TOPOFIT-DB including non-sequential relations

Valentin A. ILYIN

*Biology Department, Boston College, Chestnut Hill, MA*

Vast amount of data in structural biology was commemorated by the rapid increase in the number of protein structures solved. Many of new proteins have no annotated functions, and the function is yet to be discovered. In addition to common efforts, in recent years a Structural Genomics Initiative, aimed at solving structures with unknown functions and having low sequence similarity to known structures, has been established. The initiative has resulted in thousands of protein structures being deposited to PDB, with the increasing amount of deposited structures every year. Unfortunately, a significant number of those structures have little or no biological/biochemical information, including absence of functional annotation. Therefore, precise and reliable computational methods are needed to functionally annotate and analyze vast amounts of available structural data. Here we present our attempt to address this problem.

We present functional annotation by accurate structural similarity with TOPOFIT-DB, for detailed comparative analysis and functional annotation of protein structures. The annotation relies on our database of structural alignments, TOPOFIT-DB (<http://topofit.ilyinlab.org>), containing over 170 million of structural alignments (as of July 2009), and also a TOPOFIT one-2-all search server for comparing a user submitted structure against all known structures in the PDB. Comparative analysis of an unknown structure from the PDB can be done on the fly by retrieving data from TOPOFIT-DB database or, in case of a new structure, by calculation of structural alignments using TOPOFIT method (it usually takes from 1 to 3 hours depending on the protein size). The results page displays a list of structural neighbors for the protein of interest along with links to the GO, SCOP, EC, and other annotation servers. The distinctive feature of the TOPOFIT-DB is that it contains non-sequential alignments including all cases found by TOPOFIT from circular permutation to complex and completely reverse alignments. As it was found recently, non-sequential relations between proteins occur very often. Thus, the ability to include the non-sequential alignments in the data will allow more comprehensive comparative analysis and functional annotation of a protein structure. For each alignment the user has an opportunity to visualize the corresponding alignment plot, as well as to highlight the residues of interest on the plot. Structural superimposition corresponding to the alignments can be viewed in our integrated analytical front-end application. The Friend software has the TOPOFIT method integrated and is capable of reproducing and visualizing alignments stored in the database (<http://ilyinlab.org/friend>).

TOPOFIT-DB is updated on weekly basis during next week after PDB update.

# Simulation-based Perturbation Studies: Genome-Wide Cause and Effect predictions of mRNA Expression under Perturbation

In Sock Jang<sup>1,2</sup>, Andrea Califano<sup>2,4</sup>

<sup>1</sup>Department of Electrical Engineering, Columbia University; <sup>2</sup>Center for Computational Biology and Bioinformatics; <sup>3</sup>Department of Biomedical Informatics, Columbia University; <sup>4</sup>Institute for Cancer Genetics, Columbia University

While reverse engineering has been relatively successful in reconstructing the connectivity of regulatory networks, use of these models to predict cellular behavior is still in its infancy. For instance, there are no experimentally validated methods that can predict the genome wide effect of a molecular perturbation in a cell, such as for instance silencing/ectopic expression of a specific gene or stimulation with a receptor-specific ligand.

We introduce a probabilistic method that can successfully predict the global transcriptional profile of a human B cell, after a specific genetic perturbation, using a regulatory model inferred by the ARACNe and MINDy algorithms. ARACNe is used to infer pairwise transcriptional interactions, while MINDy is used to infer the post-translational modulation of transcription factor activity by co-factors and signaling proteins. We apply several regression models to the Markov blanket of each gene in the predicted molecular interaction network (i.e. LARS, quadratic programming) and we use a Bayesian approach to iteratively predict how the perturbation effect will propagate through the network. We validated our predictions using experimental perturbations in human B cells, produced by lentivirus mediated shRNA silencing of specific transcriptional and post-translational regulators, including the transcription factors BHLHB2, MEF2B, FOXM1, and MYB and the post-translational modulators STK38 and HDAC1. In each case, we show that the method is able to predict relative changes in gene expression that correlate with those measured by microarray gene expression profiling following silencing of the corresponding genes. Only the relative differential expression is considered, rather than absolute expression levels. This prevents artificial correlation effects from genes that are not affected by the perturbation.

# When Two Plus Two Doesn't Equal Four: Modeling Non-Additive, Non-Modular Enhancer Behavior in the *Drosophila melanogaster eve* Promoter

Ah-Ram Kim<sup>1</sup>, John Ionides<sup>2</sup>, John Reinitz<sup>1</sup>

<sup>1</sup>Department of Applied Mathematics and Statistics, and Center for Developmental Genetics, Stony Brook University, Stony Brook, NY; <sup>2</sup>Department of Computational biology, Center for Advanced Studies, St. Petersburg State Polytechnic University, St. Petersburg 195251, Russia

The prediction of expression patterns from genomic sequence is a major research challenge in modern molecular genetics. Its solution requires an understanding of the transcriptional consequences for particular configurations of bound factors. It is our conjecture that the most informative experimental materials for such studies are instances where the usual additive, modular behavior of enhancers breaks down. Such instances can reveal underlying rules, but the complexity of the experimental phenomena requires precise quantitative models for their interpretation. We are currently developing a generalized and predictive model of quantitative, time-resolved mRNA expression at cellular resolution. We have used this model to understand non-additive, non-modular behavior of two enhancer fusions of the -3.8 to -3.3 kb (stripe 3 enhancer) and the -1.6 to -1.1 kb (stripe 2 enhancer) regions of gene *eve* with and without spacer sequence between them. One of our key findings is that one third of stripe 3 enhancer in a fusion construct is recruited as a functional part of stripe 2 by Bcd and Hb mediated coactivation and Bcd mediated cooperative binding. Furthermore, analysis of our model shows that spacer DNA between two enhancers is functionally involved in generating this novel behavior, which contradicts the classical picture of the regulatory regions of metazoan genes.

# Hierarchical Model of Gas Exchange within the Acinar Airways of the Human Lung

Michael L. Mayo<sup>1,3</sup>, Stefan Gheorghiu<sup>2</sup>, Peter Pfeifer<sup>3</sup>

<sup>1</sup>*Environmental Laboratory, Engineer Research and Development Center, US Army Corps of Engineers, Vicksburg, MS, 39180;* <sup>2</sup>*Center for Complexity Studies, Aleea Parva 5, Bucharest 061942, Romania;* <sup>3</sup>*Department of Physics, University of Missouri, Columbia, MO, 65211.*

The acinar airways lie at the periphery of the human lung and are responsible for the transfer of oxygen from air to the blood during respiration. Within these airways, oxygen is transported almost exclusively by diffusion; as a result, it is screened from accessing most of the available membranes responsible for the gas exchange, a mechanism called *diffusional screening*. These screening effects have been shown to support a heterogeneous concentration gradient across the entire irregular surface of the alveolar membranes in the lung. Previous efforts to identify the relationship between the acinar structure and its function have been restricted to numerical studies of the gas exchange across model pre-fractal surfaces (prototypical irregular surfaces).

Here, we develop an *exactly solvable* diffusion-reaction model of the gas exchange across the entire branching network of acinar airways within the human lung. We derive simple equations for both the oxygen current delivered to the red blood cells within the pulmonary arteries, as well as develop an efficiency factor for the entire airway network. We demonstrate that the oxygen current is insensitive to “changes” in the surface permeability across a wide range of permeabilities. While such fault tolerance has been observed in other treatments of the lung, it is obtained here as a fully analytical result.

# Studying transcription bursts from modeling high temporal resolution gene expression dynamics with multi-layered Hidden Markov Models

Nacho Molina<sup>1</sup>, David Suter<sup>2</sup>, Ueli Schibler<sup>2</sup>, and Felix Naef<sup>1</sup>

<sup>1</sup>*Computational Systems Biology Group, Ecole Polytechnique Federale de Lausanne, Switzerland.* <sup>2</sup>*Department of Molecular Biology and NCCR Frontiers in Genetics, Sciences III, University of Geneva, Switzerland.*

The intrinsic stochasticity in the dynamics of mRNA and protein expression has important consequences on gene regulation and on non-genetic cell-to-cell variability. Recently experimental work in prokaryotes and eukaryotes relying on single cell resolution time lapse imaging has enabled a quantitative analysis and modeling of the stochastic processes underlying observed fluctuations.

Here we develop an algorithm to deconvolve time traces from single mammalian fibroblast cells that exploit a novel shorted-lived bioluminescence reporter providing unprecedented temporal resolution. One aim of this project is to investigate the bursting nature of transcription in general, and how it may be involved in the control of circadian or ultradian gene expression. To this end we analyze the promoter of a classical circadian transcription factor, BMAL1, as well as non-circadian regulatory sequences. In this work, we focus on the analysis of such signals using stochastic models that describe the three main processes of gene expression: gene activation, transcription and translation. Previous studies have mostly focused on describing the variability across populations and identifying the different sources of noise. Instead, we aim to reconstruct the temporal sequence of gene activity, mRNA and protein states from individual time traces. For this we developed a 3-layered Hidden Markov Model to describe gene activation, mRNA synthesis and protein translation. Deriving analytical approximations for the transition probabilities, we implemented decoding and estimation algorithms that enable us both to infer instantaneous gene activity status, mRNA, and protein copy number. Moreover the same method is used to learn the activation, synthesis and degradation rates defining the stochastic model, as well as to compute the uncertainty of the inferred trajectories.

We have tested the implementation by extensive simulations using the Gillespie algorithm. Importantly, we have successfully applied the method to parse real recordings indicating that rapid bursting at timescales of tens of minutes may be an intrinsic property of the transcription process in mammalian cells. It remains an open question as to how the bursting relates to fundamental processes such as transcription factor binding dynamics or chromatin remodeling. Our new algorithms will enable us to tackle these important issues in a principled and quantitative fashion.

# Adaptation of a synthetic gene circuit through diverse evolutionary paths

Bernardo Pando<sup>1</sup>, Alexander van Oudenaarden<sup>1,2</sup>

<sup>1</sup>*Department of Physics, Massachusetts Institute of Technology;* <sup>2</sup>*Department of Biology, Massachusetts Institute of Technology*

Point mutations and gene duplications are two of the major mechanisms that drive genome evolution. How these processes determine the dynamics of adaptation is poorly understood. Here we explore the relative contributions of these two mechanisms to the adaptation dynamics of a synthetic gene circuit in the budding yeast *Saccharomyces cerevisiae* using an experimental evolution approach. In this circuit a synthetic transcriptional activator whose DNA binding affinity can be increased by adding an extracellular inducer, drives the expression of an essential gene. In the absence of inducer, the growth rate is significantly lower than the wildtype rate, nevertheless yeast populations consisting of the order of  $10^7$  cells rapidly and irreversibly adapt and approach wildtype growth rates after only a few days. We find that a narrow spectrum of point mutations in the transcriptional activator explains this recovery and that many mutants recover by inverting the logic of the transcriptional activator. In the presence of low inducer concentration we find a larger adaptation rate and in addition to a similar mutation spectrum we find several gene duplication events of the transcriptional activator, providing a partial explanation for the faster adaptation dynamics. Our work suggests that the apparent mutation rate is determined by a combination of the point mutation and gene duplication rates and that their relative contributions are strongly dependent on the coupling between genes in the network. This provides a starting point for unraveling the relative contributions of point mutations and gene duplications during the adaptation dynamics of endogenous gene networks.

## Evidence for simple governing rules in complex biological networks

Kumar Selvarajoo, Masaru Tomita and Masa Tsuchiya

*Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan.*

The search for governing rules or laws of self-organization is crucial for understanding the behavior of any complex system. We focus on the complex innate immune signaling networks. The Toll-like receptors (TLRs) found on host cells, are defensive detectors of invading pathogens. Upon identification of foreign intruders, they produce intracellular signals to induce proinflammatory response which eventually leads to the elimination of the infection. The fundamental behavior of TLR networks still remains largely uncovered despite multitudes of experimentations over the last decade. Here we studied the dynamics of two distinct signal transduction pathways of TLR -3 and -4, in response to double stranded RNA (poly (I:C)) and Gram-negative bacteria, respectively. A computational model based on perturbation-response approach was built and analyzed with experiments from wildtype and several genetic knock-out murine macrophages. Using small perturbation on kinetic evolution equations resulting in first-order terms with fewer parameters and the law of mass conservation, our model predicts, in TLR3 signaling, the existence of missing intermediary steps between extracellular poly (I:C) stimulation and intracellular TLR3 binding, and the presence of a novel MAP kinase pathway. In TLR4 signaling, our model suggests novel signaling intermediates and crosstalk mechanisms. Furthermore, we show the enhanced activation of alternative pathways when molecules at pathway junctions are removed is due *Signaling Flux Redistribution* or *SFR*, which stems from the law of mass-conservation. This phenomenon was experimentally validated in MyD88 and TRAF6 knock-out murine macrophages. Through our work we show evidence for the existence of simple physical rules governing complex biological networks.

## Literature Curation of Protein Interactions: Discrepancies Across Major Public Databases

Andrei L. Turinsky<sup>1</sup>, Brian Turner<sup>1</sup>, Emerson Cho<sup>1</sup>, Kyle Morrison<sup>1</sup>, Sabry Razick<sup>2</sup>, Ian Donaldson<sup>2</sup>, Shoshana J. Wodak<sup>1</sup>

<sup>1</sup>Molecular Structure and Function Program, Hospital for Sick Children, Toronto, Canada; <sup>2</sup>The Biotechnology Centre of Oslo, University of Oslo, Norway; <sup>3</sup>Department of Biochemistry and Department of Molecular Genetics, University of Toronto, Canada.

Protein-protein interaction networks have become an important tool in biomedical research in recent years. Several resources around the world are devoted to the annotation of protein interactions from literature, thereby providing a valuable source of interactome data to the research community. However, the interpretation of the original publications is complicated by a number of challenges, which may result in discrepancies among annotations of the same publication.

We systematically investigated the consistency of protein-protein interactions across public resources. Our approach was to focus on PubMed publications that are annotated by two or more protein-interaction databases, and measure the similarity between such co-annotations. To enable this analysis, we consolidated annotations from 9 publicly available databases: BIND, BioGRID, CORUM, DIP, IntAct, HPRD, MINT, MPPI and MPact. After removing redundant records, the combined data represented a total of 272,119 interactions involving 70,474 proteins from 1348 organisms, based on the annotation of 44,159 publications. There were substantial degrees of overlap between the 9 source databases. 15,743 publications were annotated by two or more databases, providing us with the opportunity to assess discrepancies across databases.

Statistical analysis shows that whenever two databases annotate the same publication, their annotations share on average less than half of the interactions. Full agreement – i.e. both databases recording identical sets of protein-protein interactions – occurs in only 24% of the cases. On the other hand, in 41% of the cases, the annotated sets of interactions extracted from the same publication do not overlap at all, indicating that the two databases record all interactions described in the paper differently. The remaining 35% of the cases represent partial overlaps, distributed widely between full agreement and full disagreement.

The discrepancy for the sets of proteins annotated from the same paper is typically less pronounced, with severe disagreement occurring in only 14% of the cases. This indicates that annotators may agree on proteins but still disagree on the interactions they form with each other. Mammalian data stands out as having poor agreement, especially for interactions involving mouse and rat proteins.

We explored several factors that contributed to the discrepancies. One of the major factors is the inconsistent attribution of organisms to the protein-protein interactions described in the literature. We also examined factors such as the annotation of protein complexes, the handling of isoforms, the high- versus low-throughput studies, and the interaction-detection methods.

Our results provide quantitative evidence that alternative interpretations of the literature are common. Given the importance of the protein-protein interaction datasets, our work offers insights into how some of the discrepancies in annotations across public databases may be resolved in the future.

# Composite network motifs in integrated metazoan gene regulatory networks

Vanessa Vermeirssen<sup>1</sup>, Tom Michoel<sup>1</sup>, Yves Van de Peer<sup>1</sup>

<sup>1</sup>Laboratory for Bioinformatics and Evolutionary Genomics, VIB Department of Plant Systems Biology, Ghent University, Belgium

Differential gene expression is a tightly controlled process that governs development, function and pathology of metazoan organisms. Several molecular interactions, e.g. protein-DNA interactions between transcription factors and target genes, protein-protein interactions between transcription factors, closely work together in order to establish proper gene expression in space and time. Biological networks have mainly focused on the relationships between one or two types of molecular interactions.

In order to get a systems level understanding of how different molecular interactions interrelate to form a coordinated response in gene regulation, we studied composite network motifs in integrated networks containing protein-protein, transcription regulatory, protein-DNA, miRNA-mRNA, sequence homology and genetic interactions of the worm *C. elegans*. Through a computationally efficient and mathematically rigorous method, we identified dense clusters of several composite network motifs in this integrated *C. elegans* network.

We discuss the biological function of these composite network motifs in the context of eukaryotic gene regulation. We conclude that composite network motif clustering is a useful data integration method to unravel the topological organization of gene regulation in metazoan organisms.

## Bottom-up Engineering of Synthetic Gene Networks

Xiao Wang, Tom Ellis, and James J. Collins

*Howard Hughes Medical Institute, Department of Biomedical Engineering, Center for Biodynamics and Center for Advanced Biotechnology, Boston University, Boston, MA 02215*

Synthetic gene networks can be constructed from bottom up with desired properties. However, constructing predictable gene networks with desired functions remains a challenge. It is because of the lack of well-characterized components and unpredictability of the assembled networks. The need for extensive, iterative characterization and retrofitting for optimization drastically slows down the construction of desired synthetic gene networks. Here we present a diversity-based approach that combines promoter library synthesis and mathematical modeling to quickly construct gene networks with desired properties (1). In this approach, promoters with random strength diversities are synthesized and characterized in parallel. When coupled with mathematical modeling to simulate the network at a whole system level, promoters that are optimal for the intended functions can be selected before the actual network assembly, without the need for post-hoc modifications. This approach will first be demonstrated in yeast by constructing negative feedforward loop networks. Then the method will be used to produce a synthetic gene network that acts as a timer, tunable by component choice. We utilize this network to control the timing of yeast flocculation phenotype, to illustrate a practical application of our approach. The construction of a synthetic cellular counter (2) will also be presented to lead to the discussion of future directions for library-modeling based approaches in synthesis of more complicated gene networks.

1. Ellis, T.\* , X. Wang\*, and J. J. Collins. 2009. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature biotechnology* 27:465-471.
2. Friedland, A. E.\* , T. K. Lu\*, X. Wang, D. Shi, G. Church, and J. J. Collins. 2009. Synthetic gene networks that count. *Science* 324:1199-1202.

(\* equally contributed)

## Stochastic modeling for loss of imprinted mRNA expression with transcriptional pulsing

James G. Wetmur<sup>1</sup>, Andreas I. Diplas<sup>2</sup>, Jia Chen<sup>2</sup> and Jianzhong Hu<sup>1</sup>

<sup>1</sup>Departments of Microbiology and <sup>2</sup>Preventive Medicine, Mount Sinai School of Medicine, New York, NY 10029, USA

Imprinted genes are expressed from a single maternal or paternal allele. Loss of imprinting (LOI) of the maternally imprinted *PLAGL1* has been examined in a human trophoblast cell line following treatment with 5-aza-2'-deoxycytidine (AZA). We have developed a stochastic model for LOI taking the following observations into account. (1) mRNA expression varied significantly from cell-to cell with a substantial fraction of cells showing no significant expression. This variation in expression was not due to cell cycle. (2) Those cells showing LOI exhibited biallelic expression which varied significantly in expressed allele frequency from cell to cell but was centered about equal expression from both alleles.

With transcriptional pulsing, each allele of a gene (A1 and A2) a time point  $t_i$  can be either in an active state, during which transcription is very efficient, or an inactive state, in which transcription is hindered. The synthesized mRNA is proposed to follow an exponential decay. The amount of total mRNA from each allele at time  $t_i$  is given by:

$$A1 = \left[ \sum_{s=0}^{s=2} P(t, \lambda_1) e^{-\frac{t-s}{k}} \rho_e \right] \rho_m$$

$$A2 = \left[ \sum_{s=0}^{s=2} P(t, \lambda_2) e^{-\frac{t-s}{k}} \rho_e \right] \rho_m$$

$p(t, \lambda)$  is a hypergeometric random function determining the transcription state at time  $t$ : 1 = active or 0 = inactive with the pulsing frequency  $\lambda$ .  $\rho_e$  is a Gaussian random variable from 0% to 100% for the transcription efficiency of each pulse.  $k$  is the time constant for mRNA exponential decay.  $\rho_m$  is a Gaussian random variable to estimate the measurement error. Each analysis of A1 and A2 represents a single cell measurement. All analyses were performed with sufficient steps such that the distributions of A1 and A2 were independent of  $t_i$ . For a wide range of parameters, the stochastic transcriptional pulsing model fits the experimental data for LOI occurring as an all-or-none process.

## Genome-wide mapping and computational analysis of non-B DNA structures *in vivo*

Damian Wójtowicz<sup>1,5</sup>, Fedor Kouzine<sup>2</sup>, Arito Yamane<sup>3</sup>, Craig Benham<sup>4</sup>, Rafael Casellas<sup>3</sup>, David Levens<sup>2,\*</sup>, Teresa Przytycka<sup>1,\*</sup>

<sup>1</sup>Computational Biology Branch, NCBI/NIH; <sup>2</sup>Laboratory of Pathology, NCI/NIH; <sup>3</sup>Genomic Integrity and Immunity, NIAMSD/NIH; Genome Center, UC Davis, CA; <sup>5</sup>University of Warsaw, Poland;

\*corresponding authors

The Watson-Crick structure is the natural state of DNA in a genome and it is known as B-DNA. However, DNA is a dynamic molecule that undergoes various deformations and adopts several alternative secondary structures such as single-stranded DNA, Z-DNA (left-handed), H-DNA (triplex), cruciform, or quadruplex. Although previous studies confirmed the existence of non-B DNA conformations at some particular sites and implicated their role in DNA transactions, e.g. the tight regulation of c-myc oncogene by FUSE element, it was not known how general such a regulation mechanism might be. Indeed, little is still known about the formation of non-B DNA conformations and their role on genomic scale.

As a result of concerted effort of experimental and computational groups we now obtained the first draft of genome-wide mapping of alternative DNA structures *in vivo*. Specifically, with a novel experimental technique developed in David Levens' group, extensive experimental analysis performed in David Levens' and Rafael Casellas' groups, and a computational analysis performed in Teresa Przytycka's group, we have been able to perform the first genome-wide analysis of occurrences of alternative DNA structures in living cells, their position with respect to several genomic marks such as transcription start site, transcription termination site, etc. The experimental method is based on the fact that non-B DNA structure contains region of single-stranded DNA whose ends can be isolated and sequenced using Solexa high-throughput sequencing technique. The experiment was performed on activated mouse B-cells under different conditions: untreated cells and cells treated with DRB (a drug inhibiting the RNA polymerase II).

This presentation focuses on the computational component of this research. Short reads obtained from sequencing were mapped to the mouse genome (mm9), and simple statistical method was applied to find non-B DNA regions across the genome for both treated and untreated mouse B-cells. From our analysis, it is clear that sequences forming alternative DNA structures are non-randomly distributed in the genome. We found an enrichment of non-B DNA conformations near transcription start sites indicative of their potential role in gene regulation. We have compared, for the first time on such scale, the experimental data and the genomic regions computationally predicted to have a high propensity to form non-B DNA conformations. We found that not only experimentally detected non-B DNA regions have a significant overlap with computationally predicted regions, but also various computationally classified non-B DNA conformations have different experimentally derived signatures. This offers not only strong evidence for the *in vivo* formation of the alternative structures like Z-DNA, quadruplexes, and SIDD sites but also provides the first look at their genome-wide landscape and possible role in gene regulation.

# Metazoan operons accelerate transcription and recovery rates

Alon Zaslaver<sup>1</sup>, L. Ryan Baugh<sup>2</sup>, Paul Sternberg<sup>1</sup>

<sup>1</sup> Howard Hughes Medical Institute and California Institute of Technology, Division of Biology, 1200 E. California Blvd., Pasadena, California 91125; <sup>2</sup> Department of Biology and Center for Systems Biology, Duke University, Durham, North Carolina, USA.

Existing theories efficiently explain why operons are advantageous in prokaryotes, but their emergence in metazoans is still an enigma. We present a combination of genomic meta-analysis, experiment and theory to explain how operons could be adaptive during metazoan evolution. Focusing first on nematodes, we show that operon genes, typically consisted of growth genes, are significantly up-regulated during recovery from multiple growth-arrested states, and that this expression pattern is anti-correlated to the expression pattern of non-operon genes. In addition, we find that transcriptional resources are initially limited during arrest recovery, and that recovering animals are extremely sensitive to any additional limitation in transcriptional resources. By clustering growth genes into operons, fewer promoters compete for limited transcriptional machinery, effectively increasing the concentration of transcriptional resources and accelerating growth during recovery. A simple mathematical model of transcription dynamics reveals how a moderate increase in transcriptional resources can lead to a substantial enhancement in transcription rate and recovery. We find evidence for this design principle in different nematodes as well as in the chordate *C. intestinalis*. As recovery from a growth arrested state into a fast growing state is a physiological feature shared by many metazoans, operons could evolve as an evolutionary solution to facilitate these processes.

## Gene expression profile of human adipose stem cells cultured in allogeneic human serum and fetal bovine serum

Bettina Lindroos<sup>1</sup>, Kaisa-Leena Aho<sup>2</sup>, Hannu Kuokkanen<sup>3</sup>, Sari Rätty<sup>4</sup>, Heini Huhtala<sup>5</sup>, Riina Lemponen<sup>5</sup>, Olli Yli-Harja<sup>2</sup>, Riitta Suuronen<sup>1,6,7</sup>, Susanna Miettinen<sup>1</sup>

<sup>1</sup>University of Tampere and Tampere University Hospital, Regene – Institute for Regenerative Medicine, Finland; <sup>2</sup>Department of Signal Processing, Tampere University of Technology, Finland; <sup>3</sup>Department of Plastic Surgery, Tampere University Hospital, Finland, <sup>4</sup>Department of Gastroenterology and Alimentary Tract Surgery, Tampere University Hospital, Finland, <sup>5</sup>Tampere School of Public Health, University of Tampere, Finland, <sup>6</sup>Department of Eye, Ear and Oral Diseases, Tampere University Hospital, Finland, <sup>7</sup>Department of Biomedical Engineering, Tampere University of Technology, Finland

Human adipose stem cells (ASCs) can be differentiated *in vitro* to various cell lineages, such as cartilage, bone and muscle, and are thereby suitable for human tissue engineering applications [1,2]. Before differentiation, ASCs are typically cultured in medium containing fetal bovine serum (FBS). To obtain cells free of animal-derived culture reagents for safe clinical use, FBS could be replaced by human serum (HS).

To explore how the serum affects ASCs, we compared the gene expression of undifferentiated human ASCs cultured in either allogeneic HS (alloHS) or FBS using DNA microarrays. The microarray data involving a paired 5 vs 5 experimental setting was normalized using the robust multi-array average (RMA) algorithm in R. Hierarchical clustering was performed using 1 – Pearson correlation as distance measure and the average linkage method. The differentially expressed genes were found using the Linear Models for Microarray Data (LIMMA) algorithm in R. The *P* values were adjusted for multiple testing using the Benjamini & Hochberg's method to control the false discovery rate (FDR) and a threshold of FDR<0.05 was applied. The differentially expressed genes were further analyzed to find enriched (*P*<0.05) associations to Gene Ontology (GO) Biological Processes and KEGG pathways. In the enrichment analysis, DAVID Functional Annotation Tool implementing a modified Fisher's exact test was used. The samples clustered according to the serum supplement rather than according to the donor suggesting that the sera have a systematic effect on the ASCs. Altogether 1281 genes were found to be differentially expressed between the conditions. Of these, 844 genes were overexpressed and 437 genes were underexpressed in alloHS compared to FBS. Many of the most enriched GO terms were associated with cell cycle. Also genes of signaling pathways involved in cell cycle regulation and cell differentiation were significantly enriched. Interestingly, the expression changes on a particular signaling pathway were coherent with simultaneous studies on the differentiation capacity of human ASCs in alloHS and FBS. These results show that ASCs respond to different culture media through regulation of gene expression. Understanding the effects of different culture media on the ASCs is central in optimizing their controlled and clinically safe culture and differentiation methods.

[1] Zuk, P.A. et al. Human adipose tissue is a source of multipotent stem cells. *Mol Biol Cell* 13(12), 4279-95, 2002.

[2] Mesimäki, K. et al. Novel maxillary reconstruction with ectopic bone formation by GMP adipose stem cells. *Int J Oral Maxillofac Surg* 38(3):201-9, 2009.

# CENTRO: A CoExpression NeTwoRk Omnibus for gene function and pathway discovery

Vicenzo Belcastro<sup>1,4</sup>, Velia Siciliano<sup>1,4</sup>, Francesco Gregoretti<sup>2</sup>, Francesco Iorio<sup>1</sup>, Gennaro Oliva<sup>2</sup>, Diego di Bernardo<sup>1,3</sup>

<sup>1</sup>Telethon Institute of Genetics and Medicine, Via P. Castellino, Naples, Italy; <sup>2</sup>Institute of High Performance Computing and Networking ICAR-CNR, Naples, Italy; <sup>3</sup>Dipartimento di Informatica e sistemistica. Università degli studi di Napoli Federico II; <sup>4</sup>The Open University, PO Box 197, MK7 6BJ, Milton Keynes, United Kingdom;

Co-expression networks, where two genes are connected if they are co-expressed according to a given similarity measure, have been applied to mammalian systems but only to specific cell-types or tissues, thus limiting the number of experimental data points to the order of hundreds, and hence the number of genes for which there is sufficient statistical evidence of co-expression. Here we scaled up this process to all the available expression data, solving along the way the problems of handling, normalizing and computing a similarity for this massive dataset. We reverse-engineered a co-expression network for *Homo Sapiens* (*Mus Musculus*) from a set of 20,255 (8895) hybridizations. The human (mouse) network is characterized by a set of 22283 (45101) nodes and a set of 4`817`629 (14`641`095) edges, where the edge is weighted by the Mutual Information (MI) similarity between the two genes.

Our parallel-computing based algorithm consists of a novel normalization step that yields a single dataset containing comparable expression values, followed by the computation, for each pair of transcripts, of their Mutual Information (MI). MI was computed for over 200 millions and 1 billion of transcript pairs in human and mouse, respectively.

Validation of the resulting networks is non-trivial. We first compared our human inferred interactions against four published protein-protein, metabolic and kinase-substrate interaction networks. 14,691 probesets over 22,283 of the human network have at least one interaction, for a total of 277,945 interactions. To validate mouse predictions we first mapped mouse genes over their human orthologous and than we used the same Golden Standard set of interactions.

By sorting the recovered interactions according to their weights, we obtain a ROC curve which goes up to 90% of correct predictions in human and 100% in mouse; more than 60% of the 15,000 highest human scoring edges are correct; a random prediction would have inferred not more than 0.0007%. We then checked via a Y2H technique that we correctly predicted a novel protein complex involved in spindle check point in human cells. We then checked whether we could assign functions to genes via a "guilty-by-association" technique, i.e. for each gene, its predicted function is the one that is most common among its neighbors. We show that we can correctly assign gene function more that 40% of the time; moreover, we experimentally validated two gene co-localization predictions in mitochondrion. We also applied a novel community finding algorithm to the human network and show that the more that 60% communities we obtained are enriched for gene with similar functions. We verified that genes involved in the same molecular pathways lie in the same communities in the recovered network and that communities are conserved across species. Here we show that massive expression datasets, including a variety of different tissues and cell types, if properly analyzed are able to shed light on gene regulation and gene function. Our human

and mouse network can be explored by our CENTRO online tool which also annotated all known metabolic, transcription and protein interactions. It is therefore a unique tool to help molecular biology and molecular medicine research.

## Comparative analysis of genomics data collections: Multi-species Integrative Biclustering

Peter Waltman<sup>1,3\*</sup>, Thadeous Kacmarczyk<sup>2\*</sup>, Ashley R Bate<sup>2</sup>,  
Patrick Eichenberger<sup>2\*</sup>, Richard Bonneau<sup>1,2,3\*</sup>

1 – Computer Science Department, Courant Institute for Mathematical Sciences, New York University,

2 – Center for Genomics and Systems Biology, Biology Department, New York University,

3 – Computational Biology Program, New York University

Extensive data from high-throughput experimental technologies has recently enabled new methods to learn and model the complex regulatory networks that organisms use to respond to their ever-changing environment. A key challenge in the analysis of genomics data is the identification of modules of genes with similar or identical regulatory controls, a non-trivial problem due to the complexity of regulatory networks. Recent works that compare functional genomics data sets for closely related species suggest that many co-regulated modules are conserved (in whole or in part) across several species. This suggests that comparative analysis of functional genomics data-sets could prove powerful in accurately identifying biologically relevant conserved co-regulated groups. Our approach to this problem integrates collections of data-types across multiple species to learn regulatory modules conserved across multiple species. We describe an extension of the cMonkey algorithm (a biclustering algorithm that integrates multiple systems biology data-types to estimate condition dependent modules) that allows for the simultaneous biclustering of data compendia spanning multiple species. For each of the conserved co-regulated modules detected we also identify the experimental conditions (within the single species data-sets supporting the module) the experimental condition over which the module is relevant. We identify conserved modules of orthologous genes, yielding evolutionary insights into the formation and surprising high degree of conservation of regulatory modules. The method also provides a framework that allows insights from well-studied organisms to be used to aid the analysis of related but less well studied organisms. We present results from the multi-species biclustering of a Gram positive group containing *Bacillus subtilis*, *Bacillus anthracis*, and *Listeria monocytogenes*.

## Distinct topological changes of the different cancer types

Ertugrul Dalkic<sup>1-3</sup>, Christina Chan<sup>1-3</sup>

<sup>1</sup> Cell and Molecular Biology Program, Michigan State University, East Lansing, MI, USA; <sup>2</sup> Cellular and Molecular Biology Lab, Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, MI, USA; <sup>3</sup> Center for Systems Biology, Michigan State University, East Lansing, MI, USA.

Recent research on protein-protein interaction and gene regulatory interaction networks suggests increased connectivity plays a role in cancer. We analyzed condition-specific networks associated with the different cancer types by integrating different types of -omics data. To date, integration of -omics data has mostly been performed with the aim of minimizing the errors of the individual sources of data. However, some studies have shown that integration of genome-scale expression and interaction data can be useful for specific network generation and analysis and this approach can help in the comparison of specific networks obtained from different types of tumors.

We integrated genome-wide expression and protein-protein interaction data to reveal sample- or normal/tumor-specific protein-protein interaction networks and analyzed the differences of these specific networks obtained from paired normal and tumor samples for different cancer types. We generated sample-specific protein-protein interaction networks, by including only the highly expressed genes among all the genes in a sample, which could be normal or tumor. The networks generated, suggested that the sample-specific networks of tumors are more connected than normal for only colorectal cancer but not for lung, breast, prostate, head and neck, and stomach cancers. Differential expression was previously used to analyze the network level properties of cancer. Therefore, we also generated normal- and tumor-specific protein-protein interaction networks, by including the differentially expressed genes across all samples, from paired normal and tumor data for the different cancers. We compared the sample-specific networks which do not depend on differential expression but on high level of expression with normal/tumor-specific networks which depend on differential expression. The networks obtained, differentially expressed (upregulated or downregulated) normal/tumor-specific and highly expressed sample-specific networks, shared scale-free topology and similar enrichment in biological processes. The normal/tumor-specific networks showed the same trend for all cancers, where the upregulated networks are more connected for tumors than the normal-specific networks, confirming previous reports. Since the upregulated networks are more connected in tumor than normal for all cancers including colorectal cancer, we compared the overlap between the highly expressed sample-specific networks for each cancer type with their upregulated tumor-specific networks. Indeed, colorectal cancer has the highest overlap in their genes between the highly expressed tumor sample-specific and the upregulated tumor-specific networks, in other words, colorectal cancer has more genes which are highly expressed and upregulated. Our results show a distinct topological characteristic for colorectal cancer. We obtained a list of genes which are both upregulated and highly expressed in colorectal tumor samples, and removal of which renders colorectal cancer topological results similar to the other cancers. Analysis of these genes provides insight and potential factors that may play a role, specifically, in the development of colorectal cancer.

# Incorporating spatial information of heterogeneous cell populations into Bayesian mRNA- and miRNA-expression analysis

Timo Erkkilä<sup>1,2</sup>, Pekka Ruusuvaori<sup>1,2</sup>, Ilya Shmulevich<sup>1,2</sup>, Harri Lähdesmäki<sup>1</sup>

<sup>1</sup>*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland;*

<sup>2</sup>*Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103-8904, United States*

Microdissection of tissue samples is preferred when samples contain significant proportions of cell types not wanted to be measured in a microarray experiment. However, performing manual purification on each tissue sample can be very time-consuming, and reduced yield of mRNA or miRNA may consequently become a bottleneck in array hybridization. On the other hand, biological studies often concentrate on finding differential expression between cell types being spatially connected, e.g., cancer, stromal, and epithelial cells, so that whatever becomes discarded from dissected samples may contain, in fact, information relevant to the whole study.

Several authors have already addressed this problem of sample heterogeneity by proposing statistical models for expression profiles extracted from mixed cell populations. With such approaches no manual dissection is necessary; the idea is to computationally reverse-engineer the cell type specific expression profiles from the mixture profiles.

We follow the footsteps of previous authors by proposing a linear model called “Dsection” for the same reverse-engineering problem, but which is built fully Bayesian. Not only is our model capable of incorporating prior knowledge of proportions of cell types being, say, extracted from digital images of H&E stained tissues in an automatic manner, but the model also allows for taking multiple biological conditions, e.g., treatments, into account simultaneously. This makes it possible to assess scores for differential expressions between any tissue-condition pairs by utilizing Markov Chain samples from the posterior distributions, and by simulations we show that that our score, called “D-score”, outperforms assessments based on simple fold-change of expressions.

We show with simulated and real data that by incorporating prior information on model parameters one is able to obtain more accurate estimates for cell type specific expression profiles than without such information. Furthermore, prior densities can be tuned to reflect the quality of any incorporated, additional information in a natural way.

# Functional Inference from a Genome-Wide in situ Hybridization Atlas of the Mouse Embryo

Attila Gyenesei<sup>1,2</sup>, Mei Sze<sup>1</sup>, Colin Semple<sup>1</sup>, Duncan Davidson<sup>1</sup>, Richard Baldock<sup>1</sup>.

<sup>1</sup>MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK; <sup>2</sup>Turku Centre for Biotechnology, BioCity, 5<sup>th</sup> floor, 20521, Turku, Finland.

Genome-wide in situ hybridization (ISH) datasets allow the identification of distinct patterns of sub-structural and cell type-specific expression below the level of a tissue type, which provides a basis for the inference of functional interactions, and can reveal exquisitely detailed patterns even for relatively lowly expressed genes. The EU FP6 consortium EurExpress has developed a genome-wide transcriptome atlas database for mouse embryo ([www.eurexpress.org](http://www.eurexpress.org)), which contains expression data of more than 18,000 genes by RNA ISH on sagittal sections from E14.5 wild type murine embryos. The group at the MRC Human Genetics Unit was responsible for the database development, informatics infrastructure and high-throughput data analysis.

In this study we investigated the spatial gene expression at a common developmental stage and down to a cellular level in the developing mouse. We used these data to identify coexpressed gene clusters, demonstrated their statistical and biological significance, and compared these results with coexpression defined using genome-wide, publicly available microarray data.

During the analysis we applied both well-known hierarchical clustering and our unique biclustering method to reveal the gene clusters. Biclustering was able to identify those genes that were coregulated not only for the whole but subsets of the EurExpress data. Moreover, it revealed genes participating in more than one gene network. Annotation enrichment was calculated for each discovered cluster using hypergeometric distribution. The functional annotation types used in this study were gene ontology, InterPro conserved domain identifiers, mammalian phenotype ontology terms, and cytogenetic band as a proxy for genomics position. The significance of enrichment across all clusters was determined using permutation strategy with false discovery rate calculation.

The significant enrichment of functional annotation demonstrated the statistical and biological significance of coexpressed EurExpress clusters. Interestingly, most of the clusters significantly enriched for functional annotation terms were found to contain significantly enriched terms from more than one annotation type offering fundamental insights into various pathways.

We also show that EURExpress coexpression clusters can successfully be used to infer novel functional relationships between genes at various levels. Based on these results, we believe that EURExpress data will be a potent tool in uncovering many more novel functional associations relevant to development and disease.

## On computational analysis of quantitative, 3D spatial expression in *Drosophila* blastoderm

Soile V. E. Keränen<sup>1</sup>, Angela DePace<sup>2</sup>, Ann Hammonds<sup>1</sup>, Bill Fisher<sup>1</sup>, Oliver Rübél<sup>1</sup>, Gunther Weber<sup>1</sup>, Clara Henriquez<sup>1</sup>, Charless Fowlkes<sup>3</sup>, Cris L. Luengo Hendriks<sup>4</sup>, E. Wes Bethel<sup>1</sup>, Hans Hagen<sup>5</sup>, Bernd Hamann<sup>6</sup>, Jitendra Malik<sup>7</sup>, Sue Celniker<sup>1</sup>, David W. Knowles<sup>1</sup>, Michael B. Eisen<sup>7</sup>, Mark D. Biggin<sup>1</sup>

<sup>1</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA; <sup>2</sup>Harvard Medical School, Boston, MA, USA; <sup>3</sup>UC Irvine, Irvine, CA, USA; <sup>4</sup>Uppsala University, Uppsala, Sweden; <sup>5</sup>University of Kaiserslautern, Kaiserslautern, Germany; <sup>6</sup>UC Davis, Davis, CA, USA; <sup>7</sup>UC Berkeley, Berkeley, CA, USA

The development of species specific morphologies results from complex, quantitative action of gene expression networks. The analysis of such networks requires computationally analyzable, cellular resolution datasets of spatial gene expression. The Berkeley *Drosophila* Transcription Network Project (BDTNP) is developing cellular resolution, quantitative expression maps for *Drosophila* embryos (<http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp?w=summary>). For blastoderm stage embryos, we now have data for the protein and mRNA expression of over 100 genes, 7 transcription factor mutant strains, 23 transgenic promoter constructs, and 3 *Drosophila* species. Computational analyses of this spatial and temporal expression and morphology data have revealed previously unseen changes in anterior-posterior and dorsal-ventral pattern formation that are interconnected with each other as well as quantifiable morphological features. With our new more accurate methods, we have detected far more changes in gene expression and morphology between closely related *Drosophila* species than previously realized, raising the possibility that speciation may involve mostly subtle quantitative changes in the regulation of a large proportion of genes and morphogenetic processes. We can also show quantitative differences in the output of a *cis*-regulatory element and its putative native target pattern in anterior-posterior or dorsal-ventral direction, suggesting that development of sequence based models of gene expression regulation might benefit from quantitative comparisons of actual outputs of the regulatory sequences and their native expression patterns. Because our numerical PointCloud atlas data is tabulated in a text file, aside from data comparisons, it can quite readily be adapted for computational modeling of the expression data. We have modified our visualization tool PointCloudXplore to enable writing of analysis and modeling tools within its context, helping to make the data analysis more interactive.

## Selection of an optimal set of blood biomarker proteins

Virpi Kivinen<sup>1</sup>, Matti Nykter<sup>1,2</sup>, Dan Martin<sup>2</sup>, Olli Yli-Harja<sup>1</sup>, Ilya Shmulevich<sup>2</sup>

<sup>1</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>2</sup>Institute for Systems Biology, Seattle, USA.

As all measurement technologies have limitations in capacity, it is important to know which measurements are most informative. This is the case, for example, in measuring protein concentrations in blood samples. With current technologies, it is possible to measure hundreds of proteins in one measurement in a fast and cheap way [1]. This potentially allows accurate determination of an individual's state of health by measuring a well defined set of biomarker proteins. We assume that state of health can be quantified by measuring the activity of each biological pathway. Here, a 'pathway' refers to a collection of biologically related proteins, e.g., canonical pathways, tissue/organ specific proteins, functional annotations, etc.

The selection of an optimal biomarker protein set can be formulated as an optimization problem: given a fixed number of proteins to be measured, cover as many pathways as possible with highly specific biomarker proteins. By covered, we mean that at least one protein that belongs to the pathway is measured. By specific, we mean that the protein abundance is proportional to the activity of the pathway.

We represent the proteins and pathways as a bipartite graph with proteins on the left side and the pathways on the right, with connections from proteins to the pathways in which they are members. To approximate the pathway relevance to the system, we hypothesize that pathways with a high number of proteins are the most important. Similarly, we assume that the specificity of the protein is proportional to the number of pathways of which it is a member. Thus, our goal is to cover as many pathways as possible with minimum, or fixed, number of highly specific proteins. The optimal solution for the objective function is obtained with a modified genetic algorithm.

Mouse gene expression data was used to show that the proposed approach preferentially selects biologically relevant pathways. Gene Set Enrichment Analysis (GSEA) was performed on expression data sets from 61 different mouse tissues [2], with curated gene sets from the GSEA website. As a result, a set of globally most enriched pathways was obtained. The protein sets selected by our algorithm cover more pathways than sets of randomly selected proteins. On average, enriched pathways are covered more efficiently than other pathways, reflecting the importance of the selected proteins.

[1] Fan, R., Vermesh, O., Srivastava, A., Yen, B., Qin, L., Ahmad, H., Kwong, G., Liu, C.-C., Gould, J., Hood, L., and Heath, J. 2008. Integrated barcode chips for rapid, multiplexed analysis of proteins in microliter quantities of blood. *Nat. Biotechnol.*, vol. 26, no. 12, pp. 1373 – 1378.

[2] Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M., Walker, J., and Hogenesch, J. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 16, pp. 6062 – 6067.

## GATE: Grid Analysis for Time-Series Expression

Alexander Lachmann<sup>1\*</sup>, Ben D. Macarthur<sup>1,2\*</sup>, Ihor R. Lemischka<sup>2</sup>, Avi Ma'ayan<sup>1</sup>

<sup>1</sup>*Department of Pharmacology and Systems Therapeutics, Systems Biology Center New York (SBCNY);* <sup>2</sup>*Department of Gene and Cell Medicine, Black Family Stem Cell Institute; Mount Sinai School of Medicine, One Gustave Levy Place, New York, NY 10029;*  
*\* contributed equally*

GATE (Grid Analysis of Time-Series Expression) is an integrated computational software platform for the analysis and visualization of high-dimensional bio-molecular time-series. GATE uses a correlation-based clustering algorithm to arrange molecular time-series on a two-dimensional hexagonal array and dynamically colors individual hexagons according to the expression level of the molecular component to which they are assigned, to create animated movies of systems-level molecular regulatory dynamics. In order to infer potential regulatory control mechanisms from patterns of correlation, GATE allows interactive interrogation of movies against a wide variety of prior knowledge datasets. GATE movies can be paused and are interactive, allowing users to reconstruct networks and perform functional enrichment analyses. Movies created with GATE can be exported in Flash format and inserted directly into PDF manuscript files as interactive figures.

# Evolvability of the expression pattern of the *Drosophila* gap-gene system

Ho-Joon Lee<sup>1</sup>, Denis Thieffry<sup>2</sup>, Eugene Shakhnovich<sup>3</sup>

<sup>1</sup>Department of Systems Biology, Harvard Medical School, Boston, USA; <sup>2</sup>Department of Biology, Aix-Marseille University, Marseille, France; <sup>3</sup>Department of Chemistry and Chemical Biology, Harvard University, Cambridge, USA

The gap-gene system and its role in early *Drosophila* embryogenesis has been extensively studied and modeled using different approaches. Here we show that the wild-type expression pattern or phenotype of the *Drosophila* gap-gene developmental system has an evolutionary implication. We apply evolutionary algorithms to a population of gap-gene expression patterns using a previous logical formulation [1]. The expression patterns in the *in silico* evolution are evaluated using the Manhattan distance as a fitness function with respect to a target expression pattern. Our target patterns are taken from the wild-type, random and two experimentally observed mutant patterns for comparison. Each pattern in the population is selected according to the fitness-dependent Poisson distribution. The simple truncation selection is also used. Mutations occur to an expression level of a randomly selected gap gene. We also introduce a population fitness index (PFI) to quantify the fitness of a population in terms of the fittest patterns at each generation time. We further use hierarchical clustering to examine the resulting population spectra. Our results show that a population of gap-gene expression patterns tends to take longer to reach the wild-type pattern than random or mutant patterns with similar PFIs. This slow evolution is in contrast to a previous study where scale-free networks were found to evolve faster than random networks [2]. We hypothesize that the evolvability of gap-gene expression patterns observed here is due to constraints resulting from conserved core processes of the *Drosophila* body plan [3].

## References

- [1] Sanchez and Thieffry, Journal of Theoretical Biology 2001
- [2] Oikonomou and Cluzel, Nature Physics 2006
- [3] Kirschner and Gerhart, PNAS 1998

# Human Cancer Proteome Variation Database and Mutated Peptides Identification in Shotgun Proteomics

Jing Li<sup>1</sup>, Zeqiang Ma<sup>1</sup>, Robbert J. C. Slebos<sup>2</sup>, David L. Tabb<sup>1,3</sup>, Daniel C. Liebler<sup>1-3</sup>, Bing Zhang<sup>1\*</sup>

<sup>1</sup> Department of Biomedical Informatics, Vanderbilt University School of Medicine; <sup>2</sup> Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine; <sup>3</sup> Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37232.

\* Correspondence should be addressed to B.Z (bing.zhang@vanderbilt.edu).

Human cancer is a disease that involves DNA and protein sequence alterations. Therefore identification and annotation of the mutations in genes, especially in proteins involved in the oncogenesis and tumor progression are crucial for understanding cancer biology as well as cancer prevention, diagnosis, and therapeutics. We have developed a human Cancer Proteome Variation Database (CanProVar) by integrating information on protein sequence variations from various public resources, with a focus on cancer-related variations. We have also built a user-friendly interface for querying the database which can be accessed from <http://bioinfo.vanderbilt.edu/canprovar>. The current version of CanProVar comprises 41,541 nonsynonymous SNPs in 30,322 proteins from the dbSNP database and 8,570 cancer-related variations in 2,921 proteins collected from various resources. CanProVar provides quick access to published cancer-related variations in protein sequences along with related cancer samples, relevant publications, data sources, and functional information such as Gene Ontology annotations for the proteins, protein domains in which the variation occurs, and protein interaction partners with cancer-related variations. CanProVar also helps reveal functional characteristics of cancer-related variations and proteins bearing these variations. Owing to its protein-centric nature, CanProVar can serve as a bridge between genomic data and proteomics studies in human cancer.

Because shotgun proteomics data analysis usually relies on database search, adding protein mutation information into the database can help identify mutated proteins. This is especially important in cancer studies in which the detection of disease-related mutated peptides/proteins can help reveal the drivers or messengers of oncogenesis. Based on the CanProVar database, we created a novel searchable sequence database and proposed an easy to follow workflow for identifying both wide-type and cancer-related mutated peptides simultaneously from the shotgun proteomics data. The workflow has been tested for compatibility with the popular database search engines including Sequest, Mascot, X!Tandem, and MyriMatch. We also applied this workflow on three Orbi-trap LTQ data sets, two from colon cancer cell lines RKO and SW480, and one from a colon cancer specimen. As results, 76, 61 and 142 mutated peptides were detected in the RKO cell line, the SW480 cell line, and the colon cancer specimen, respectively. These mutations include 82 known cancer-related mutations on 51 peptides from 20 proteins.

# Modeling Idiopathic Pulmonary Fibrosis Disease Progression based on Gene and Protein Expression

Tien-ho Lin<sup>1</sup>, Jose D Herazo<sup>2</sup>, Kazuhisa Konishi<sup>2</sup>, Naftali Kaminski<sup>2</sup>, Ziv Bar-Joseph<sup>1</sup>

<sup>1</sup>*School of Computer Science, Carnegie Mellon University;* <sup>2</sup>*Simmons Center for Interstitial Lung Disease, University of Pittsburgh Medical School.*

Idiopathic pulmonary fibrosis (IPF) is a progressive fibrotic interstitial lung disease without a known cause and without established cure. Recently several high-throughput gene and protein expression analysis on IPF patients have been conducted to identify differentially expressed genes and proteins. However most of the analyzing tools are designed for static (snapshot) datasets.

We have proposed a method for classification of time series gene expression that takes temporal information into account, and showed improvement on classifying multiple sclerosis (MS) patients. Our method can both classify the time series expression datasets and account for individual differences in progression rates. Hidden Markov models (HMMs) is used to represent the expression profiles of the two classes. Using such a HMM we align a patient's time series gene expression to a common profile. Conceptually, a hidden state in our HMM correspond to a phase in the treatment response.

For biomarker discovery, we propose a backward stepwise feature selection method that utilizes the alignment to the HMM profiles, termed *HMM-RFE*. In the MS dataset, this feature selection method has been shown to improve classification accuracy and identify genes relevant to MS. In the IPF dataset, the selected genes can be further examined by more experiments to find out the causal factors of different disease progression outcome.

We collected time series expression of 29,807 genes and 13 proteins of 20 IPF patients for 3 to 7 visits, spanning over 2 years. The accuracy of predicting disease outcome improves with more time points and achieved 95% based on leave-one-out cross validation. Using the expression of only 13 proteins in 2 time points, our method still has 85% accuracy. For comparison we also classify the data by linear SVM that does not consider temporal ordering. For all three data sources (gene expression, protein expression, and the combined gene and protein expression), HMM outperforms SVM across all time points, indicating the importance of temporal information. The model not only predicts a patient's outcome, but also infers the disease progression of an individual and potentially shed light on the mechanism of IPF progression.

# A tri-partite clustering analysis on microRNA, gene and disease model

Chengcheng Shen, [Ying Liu](#)

*Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688*

MicroRNAs (miRNAs) have been emerged as an important regulator in post-transcriptional regulation of gene expression [1] because of their increasingly importance in development and disease. The base-pairing between miRNA and mRNA can lead to inhibition of initiation and elongation of translation and degradation of mRNAs [3]. MiRNAs target many different genes such as transcription factors [4], apoptotic genes [5] etc.; the deregulation of these genes' expression causes different kinds of diseases including cancers [6], cardiovascular diseases [7] etc. Better understanding of the relationships among miRNAs, gene targets and diseases would be of great importance to find the functions of miRNAs and potential causes of diseases. Recently, some efforts have been made to find miRNAs groups based on expression profiles [8] and miRNA regulatory modules (groups of miRNAs and genes) [9], but there still lacks a systematic analysis on relationships between miRNAs, genes and diseases

We explore the connections between miRNAs, target genes and diseases caused by miRNAs by constructing a tri-partite network. In the tri-partite network of miRNAs, their predicted target genes and the diseases caused by altered expressions of these genes, valuable knowledge about the pathogenicity of miRNAs, involved genes and related disease classes can be revealed by co-clustering miRNAs, target genes and diseases simultaneously. Tri-partite co-clustering can lead to more informative results than traditional co-clustering with only two kinds of members and pass the hidden relational information along the relation chain by considering multiple members.

Here we report a spectral co-clustering algorithm for k-partite graph to find clusters with heterogeneous members. We use the method to explore the potential relationships between miRNAs, genes and diseases. The clusters obtained from the algorithm has significantly higher density than randomly selected clusters, which means members in the same cluster are more likely to have common connections. Results also show that miRNAs in the same family based on the hairpin sequences tend to belong to the same cluster. We also validate the clustering by checking the correlation of enriched gene functions and disease classes in the same cluster. Finally widely studied miR-17-92 and its paralogs are analyzed as a case study to reveal that functions of genes and classes of diseases co-clustered with the miRNAs are in accordance with current research findings.

[1] Bandyopadhyay, S. and Bhattacharyya, M. (2009). Analyzing mirna co-expression networks to explore tf-mirna regulation. *BMC Bioinformatics*, 10, 163–178.

[2] Yoon, S. and Micheli, G. D. (2005). Prediction of regulatory modules comprising micromnas and target genes. *Bioinformatics*, 21 (Suppl 2), ii93–ii100.

## Life After Comparative Genomics; Regulatory Systems, Homeostasis, Synergy, SNPs and Disease.

Alasdair MacKenzie, Lynne Shanley, Scott Davidson, Marissa Lear, John Barrow, Gemma Halliday and Ben Wen Qing Hing.

*Gene Regulatory Systems Group, School of Medical Sciences, Institute of Medical Sciences, University of Aberdeen, Aberdeen, AB25 2ZD, Scotland, UK. E-mail- mbi167@abdn.ac.uk .*

Thanks to the efforts of different Genome Sequencing programs and the development of powerful computer algorithms, the use of comparative genomics for the identification of important gene regulatory regions is now routinely used. We believe that the next stage in increasing our understanding of the role of gene regulation in health and the development of disease must be to understand the regulatory systems (ligand-receptor, signal transduction and protein-DNA interactions) that modulate the activity of these regulatory regions<sup>1,2</sup>. Another critical step forward must be to understand how separate regulatory regions interact with each other to modulate gene expression and how these interactions may differ between individuals as a result of regulatory polymorphisms.

We used comparative genomics to understand the cell specific mechanisms that regulate neurogenic inflammation, a process associated with chronic diseases such as arthritis, inflammatory bowel disease and asthma. We found that the promoter of the TAC1 gene, that encodes substance-P (SP); an important player in neurogenic inflammation, only responds to the inflammatory stimulus; capsaicin (chilli extract), in sensory neurones in the presence of a highly conserved and remote enhancer (214kb from TAC1). Culturing of transgenic dorsal root ganglion explants and primary sensory neurones in the presence of different cell signalling agonists and antagonists demonstrated that synergy was required between this enhancer and the TAC1 promoter to allow a response to the p38MAPK pathway that modulates aspects of the inflammatory response. These novel observations have important implications for the understanding of the mechanisms underlying the development of chronic pain. We present further data demonstrating the role of a second remote TAC1 enhancer that can suppress the glucocorticoid receptor (GR) induced activity of the TAC1 promoter in the amygdala<sup>3</sup>. Intriguingly this second enhancer is also controlled by GR suggesting GR mediated homeostasis at the TAC1 locus. Because SP has anxiogenic properties when expressed in the amygdala this result has important implications for understanding the mechanisms underlying susceptibility to anxiety and chronic depression. We are currently exploring the effects of several common human SNPs on the regulatory activity of these sequences.

Finally, we have established collaborations with psychiatric geneticists at the Institute of Psychiatry in London who have carried out GWAS analysis on polymorphisms within large sample groups of patients suffering from major depressive disorder. We are currently in the process of studying the strongest "hits" from these analyses that, intriguingly, are frequently contained within highly conserved non-coding sequences

1. Miller KA, et al. (2007) *Dev Biol* 311(2):665-678.
2. Miller KA, et al. (2008) *Dev Biol* 317(2):686-694.
3. Davidson S, et al (2006) *Mol Psychiatry* 11(4): 410-421.

# Identifying Cell Lineage-Specific Gene Expression Modules

Jessica Mar<sup>1,2</sup>, Christine Wells<sup>3</sup>, John Quackenbush<sup>1,2,4</sup>

<sup>1</sup>Department of Biostatistics & Computational Biology, Dana-Farber Cancer Institute;

<sup>2</sup>Department of Biostatistics, Harvard School of Public Health; <sup>3</sup>National Centre for Adult Stem Cell Research, Eskitis Institute for Molecular and Cellular Therapies, Griffith University, <sup>4</sup>Department of Cancer Biology, Dana-Farber Cancer Institute.

A mammalian organism is made up of more than 200 highly-specialized cell types, each of which carries out a specific task within the organism. The various cell types can vary by morphology, structure, lifespan, functional ability, and much more. Despite such remarkable diversity, all cells within an organism are derived from an original precursor cell, and in most cases, share the same genome. Diversity comes about largely through differential expression programs where cells regulate the abundance of individual gene transcripts and their downstream molecules such as proteins and microRNAs. Epigenetic modification and transcription factor networks are integral to this control, but the manifestation of each cell type's unique program is represented in its transcriptional profile. One significant challenge in gene expression analysis is the identification of the functional networks and pathways that are represented and which are ultimately responsible for the cellular states we observe. To address this challenge, we developed a new method that isolates the biological processes and pathways that are differentially activated between cell types and identifies modules of interacting genes. For each of these key processes, our method constructs a distinct set of representative expression profiles which collectively describe the different transcriptional states of the cell. Finally, protein-protein interaction maps and other experimental annotation information are used to create networks that provide a mechanistic explanation for and additional hypotheses regarding the biology behind cell-specific differences. The construction of these cell lineage-specific modules also serves as a starting point for building predictive state space models that extend the work of Kauffman and others to make use of Waddington's canalization principles.

# Data Integration for High-throughput Morphological and Transcriptional Genetic Screens

Oaz Nir<sup>1,2\*</sup>, Chris Bakal<sup>3,4,5\*</sup>, Norbert Perrimon<sup>4,5</sup>, Bonnie Berger<sup>1,2,3†</sup>

<sup>1</sup>Department of Mathematics, MIT; <sup>2</sup>Harvard/MIT Division of Health Sciences and Technology; <sup>3</sup>Computer Science and Artificial Intelligence Laboratory, MIT; <sup>4</sup>Department of Genetics, Harvard Medical School; <sup>5</sup>Howard Hughes Medical Institute; \*Contributed equally; †To whom correspondence should be addressed; email: bab@mit.edu.

A recurrent theme in computational biology is the development of methods to combine multiple data sources for increased predictive power. With the emergence of high-throughput morphological screens, a key challenge is to integrate this data source with transcriptional data. Here, we apply techniques from microarray analysis to determine differential expression between group pairs defined by quantitative morphology-based class distinctions. By comparing expression data between control treatment conditions and treatment conditions displaying a particular morphological phenotype (using t-tests and SAM for differential expression, and GSEA for GO/KEGG gene set analysis), we identify genes and pathways correlated with the class distinction. We demonstrate the efficacy of these techniques by applying them to morphological data from a *Drosophila* genetic screen in tissue culture [1], microarray data from an overlapping *Drosophila* screen [2], and a variety of morphological class distinctions: control versus high single-cell morphological variability, low morphological variability, inability to form lamellipodia, inability to form adhesions, similarity to Rho1 knockout, and similarity to Rac1 knockout, respectively. We apply our framework for integrating high-throughput data from morphological and transcriptional screens to study genes/pathways involved in different morphological phenotypes and use results in the literature to validate our findings. For instance, for control versus inability to form lamellipodia, we found that the Wnt pathway, known to play a key role in neural crest migration and lamellipodia formation, is significantly down-regulated in the experimental group versus control. Further, GO categories for actin cytoskeleton and axonogenesis were down-regulated, consistent with actin-based lamellipodia structure. For control versus high single-cell morphological variability, the mTor and EGFR pathways were up-regulated in the high variability group. This finding reflects these pathways' roles in regulating cell locomotion and that cells undergoing locomotion display higher population-level morphological variability. The EGFR family has been noted for producing a diversity of signaling outputs due to multiple ligands/receptors initiating distinct pathways combinatorially. We extend this result by showing that EGFR up-regulation is correlated with high morphological variability.

Overall, we identify meaningful differential expression or pathway/functional category enrichment for multiple morphological class distinctions, thus highlighting putative mechanisms for morphological change and generating new genes of interest for future study. Based on the success of this study, we expect that these techniques will prove useful in further data integration studies.

[1] Bakal C, Aach J, Church G, Perrimon N. Quantitative Morphological Signatures Define Local Signaling Networks Regulating Cell Morphology. *Science*. 2007 Jun 22;316(5832):1753-1756.

[2] Baym M, Bakal C, Perrimon N, Berger B: High-Resolution Modeling of Cellular Signaling Networks. Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008), LNBI 4955: 257-271, 2008.

# An Association Analysis Approach to Biclustering

Gaurav Pandey<sup>1</sup>, Gowtham Atluri<sup>1</sup>, Michael Steinbach<sup>1</sup>, Chad L. Myers<sup>1</sup>, Vipin Kumar<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, University of Minnesota, Minneapolis  
(Contact: gaurav@cs.umn.edu)

One of the most important types of analyses conducted on microarray and other types of genomic data is the discovery of functional modules, i.e., groups of genes that share a common function. Although clustering is the standard approach for this task, a more effective approach is biclustering, where both genes as well as their features (experimental conditions) are clustered together. This approach is particularly useful for data sets constructed over a diverse set of experimental conditions. Several algorithms, such as ISA, SAMBA, OPSM and Cheng and Church's (CC) algorithm, have been proposed to find biclusters from microarray data, and recent evaluation studies have demonstrated that gene modules found using these algorithms are generally more functionally enriched than those found using (hierarchical) clustering.

In this work, we propose a novel algorithm for biclustering, which is based on concepts from the field of association analysis in data mining. This algorithm uses an anti-monotonic measure, named *range support*, that is designed to capture the tendency of a group of genes to co-express, preferably at higher levels of expression, across several, but not necessarily all the conditions. The higher the value of this measure for a gene group, the more likely they are to show co-expression and thus represent valid functional modules. Furthermore, since the value of this measure is guaranteed not to increase as the size of a gene group increases (anti-monotonicity property), it is possible to use a bottom-up discovery algorithm (commonly known as the Apriori algorithm) that can exhaustively and efficiently find all range support modules or patterns (RAP) in a given microarray data set that satisfy the associated constraints [1].

This algorithm has two important advantages over the existing biclustering algorithms: (i) it performs an exhaustive search for biclusters, whereas most existing algorithms employ heuristics to perform an approximate search of all possible biclusters, and (ii) due to the bottom-up nature of the discovery algorithm, a substantial number of small biclusters are found among the RAP results, while these biclusters are mostly hidden within larger biclusters found by the traditional algorithms. These advantages are reflected in the results obtained from Hughes' *et al.*'s *S. cerevisiae* microarray compendium. Here, ISA found only 6-318 biclusters at different parameter settings, while the RAP approach produced at least about 20000 biclusters at different parameter settings in almost the same time (less than 30 minutes). The RAP biclusters are also better enriched by specific GO Biological Process terms containing less than 30 yeast genes, with about 30% of the RAP biclusters being enriched with these terms at  $p\text{-values} \leq 10^{-5}$ , while only 20% of the ISA biclusters are enriched by these terms at this threshold. Furthermore, RAP biclusters capture 48 GO BP terms, such as *threonine synthesis*, that are not covered by any of the ISA biclusters. These results, valid over several sets of small classes, demonstrate the utility of our approach for efficiently discovering all valid biclusters, particularly smaller biclusters or functional modules that are often missed by existing biclustering and clustering algorithms.

[1] For details and source code, visit <http://vk.cs.umn.edu/gaurav/rap/>.

## Identification of genomic features novel to xylose-fermenting yeasts through comparative analyses of *Pichia stipitis*, *Candida tenuis*, and *Spathaspora passalidarum*

Dana J. Wohlbach<sup>1,2</sup>, Thomas W. Jeffries<sup>2,3</sup>, Alan Kuo<sup>4</sup>, Igor V. Grigoriev<sup>4</sup>, Kerrie W. Barry<sup>4</sup>, Audrey P. Gasch<sup>1,2</sup>

<sup>1</sup> Department of Genetics, University of Wisconsin, Madison; <sup>2</sup> Great Lakes Bioenergy Research Center, Madison, Wisconsin; <sup>3</sup> Department of Bacteriology, University of Wisconsin, Madison; <sup>4</sup> DOE Joint Genome Institute, Walnut Creek, California

Efficient fermentation of cellulosic feedstocks is an essential step in the production of ethanol from plant materials. The six-carbon sugar glucose and the five-carbon sugar xylose are the two most abundant monomeric carbohydrates found in hemicellulose. Although *Saccharomyces cerevisiae*, the yeast most commonly utilized for fermentation during ethanol production, is able to ferment glucose, it is unable to ferment xylose. However, several Ascomycete yeasts that both ferment and assimilate xylose have been identified including *Pichia stipitis*, whose genome has recently been sequenced.

To elucidate the genetic features that underlie the ability to ferment xylose, we performed whole-genome sequencing with the Joint Genome Institute (JGI) on two novel xylose-fermenting yeasts, *Candida tenuis* and *Spathaspora passalidarum*, and performed comparative genomic analyses between the xylose-fermenting yeasts *P. stipitis*, *C. tenuis*, *Sp. passalidarum* and other closely related non-xylose-fermenting yeasts, including *S. cerevisiae*. Here we present analysis of the genome sequences, including phylogenetic reconstruction and an examination of CUG codon usage in these yeasts. Additionally, mapping of xylose growth and fermentation phenotypes onto ortholog groups allowed us to identify several genes unique to xylose-fermenting species. We also present a comparative analysis of gene expression across species in response to glucose or xylose. We anticipate that the genomic features identified through our analysis may be candidates for engineering more efficient xylose production in *S. cerevisiae*.

# Phosphopeptide-based signatures accurately predict the response of NSCLC cell lines to tyrosine kinase inhibitors

Chang-Jiun Wu<sup>1</sup>, Martin Steffen<sup>1,2</sup>, Simon Kasif<sup>1,3,4</sup>

<sup>1</sup>Department of Biomedical Engineering; <sup>2</sup>Department of Pathology and Laboratory Medicine; <sup>3</sup>Center for Advanced Genomic Technology, Boston University; <sup>4</sup>Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology.

The primary role of tyrosine phosphorylation in cancer cell signaling transduction leads to the increasing usage of tyrosine kinase inhibitors (TKI) to suppress tumor cell proliferation or invasion. Recent advances in high-throughput technology enable the profiling of tyrosine phosphorylation status at the proteomic level. We hypothesize that signatures based on phosphopeptides will achieve strong performance at predicting cancer cell susceptibility to TKIs. By integrating the phosphoproteomic profiles of 41 lung cancer cell lines<sup>1,2</sup> and the drug sensitivity data to 4 TKIs<sup>3</sup>, we developed separate phosphopeptide signatures for each drug. The signature for Erlotinib, a reversible EGFR inhibitor, has the highest predicting performance in cross validation (accuracy=0.90). Out of the three cell lines responding to erlotinib treatment, the model correctly classified two EGFR mutated cells as responders but missed one without documented EGFR mutations. Tyrosine sites on EGFR(Y998), ERBB2(Y877), ERBB3(Y1159), and other proteins in the EGF pathway are included in the signature. Predictive models for CL387,785 (irreversible EGFR inhibitor) and HKI-272 (irreversible EGFR/HER2 inhibitor) have high cross-validation performance as well (accuracy=0.85 and 0.78 respectively). Tyrosine sites on EGFR(Y998) and other EGF pathway proteins are also included in the two signatures. The susceptibility to sorafenib, a B-Raf inhibitor, is hard to be predicted by phosphotyrosine signatures trained on the whole proteome (accuracy=0.68). However, a predictive model derived only from the EGF pathway proteins improved the accuracy to 0.81, suggesting that the activation states of EGF pathway may play a role in the cellular response to non-EGFR inhibitors. Our analysis provides effective signatures for the drug sensitivity of NSCLC cell lines, which could potentially be adapted to select lung cancer patients that will benefit from TKI treatments.

[1] Rikova et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* 131(6):1190-203 (2007).

[2] Guo et al. Signaling networks assembled by oncogenic EGFR and c-Met. *Proc Natl Acad Sci U S A*. 105(2):692-7 (2008).

[3] McDermott et al. Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proc Natl Acad Sci U S A*. 04(50):19936-41 (2007).

# Reconstruction and Validation of RefRec: a Global Model for the Yeast Molecular Interaction Network

Tommi Aho<sup>1</sup>, Henrikki Almusa<sup>2</sup>, Jukka Matilainen<sup>2</sup>, Antti Larjo<sup>1</sup>, Pekka Ruusuvuori<sup>1</sup>, Kaisa-Leena Aho<sup>1</sup>, Thomas Wilhelm<sup>3</sup>, Harri Lähdesmäki<sup>1,4</sup>, Andreas Beyer<sup>5</sup>, Manu Harju<sup>1</sup>, Sharif Chowdhury<sup>1</sup>, Kalle Leinonen<sup>1</sup>, Christophe Roos<sup>2</sup>, Olli Yli-Harja<sup>1</sup>

<sup>1</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland; <sup>2</sup>Medicel Ltd., Espoo, Finland; <sup>3</sup>Institute of Food Research, Norwich, United Kingdom; <sup>4</sup>Department of Information and Computer Science, Helsinki University of Technology, Espoo, Finland; <sup>5</sup>Biotechnology Center, Technische Universität Dresden, Dresden, Germany.

Molecular interaction networks establish all cell biological processes. The networks are under intensive research that is facilitated by new high-throughput measurement techniques for the detection, quantification, and characterization of molecules and their physical interactions. For the common model organism yeast *Saccharomyces cerevisiae*, public databases store a significant part of the accumulated information and, on the way to better understanding of the cellular processes, there is a need to integrate this information into a consistent reconstruction of the molecular interaction network.

This work presents and validates RefRec, the most comprehensive molecular interaction network reconstruction currently available for yeast. The reconstruction integrates protein synthesis pathways, a metabolic network, and a protein-protein interaction network from major biological databases. The core of the reconstruction is based on a reference object approach in which genes, transcripts, and proteins are identified using their primary sequences. This enables their unambiguous identification and non-redundant integration. The obtained total number of different molecular species and their connecting interactions is ~67,000. In order to demonstrate the capacity of RefRec for functional predictions, it was used for simulating the gene knockout damage propagation in the molecular interaction network in ~590,000 experimentally validated mutant strains. Based on the simulation results, a statistical classifier was subsequently able to correctly predict the viability of most of the strains. The results also showed that the usage of all molecular domains in the reconstruction is important for accurate phenotype prediction.

In general, the findings demonstrate the benefits of global reconstructions of molecular interaction networks. With all the molecular species and their physical interactions explicitly modeled, our reconstruction is able to serve as a valuable resource in additional analyses involving objects from multiple molecular domains. For that purpose, RefRec will be freely available in the Systems Biology Markup Language format [1] in the BioModels database [2].

[1] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, et al. (2003) The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19: 524–531.

[2] Le Novère N., Bornstein B., Broicher A., Courtot M., Donizelli M., et al. (2006) BioModels Database: A Free, Centralized Database of Curated, Published, Quantitative Kinetic Models of Biochemical and Cellular Systems *Nucleic Acids Res.*, 34: D689-D691.

## DNA motif discovery approach adapted to ChIP-chip and ChIP-Seq data.

Malika Aid<sup>1</sup>, Sylvie Mader<sup>1</sup>

<sup>1</sup>Institute for research in immunology and cancer-Montreal university-Quebec-Canada

Transcription factors (TFs) play an important role in various biological processes such as differentiation, cell cycle progression and tumorigenesis. They regulate gene transcription by binding to specific DNA sequences (cis-regulatory elements). Identifying these cis-regulatory elements is a crucial step in the understanding of gene regulatory networks. Recent developments in genomic technologies such as DNA microarrays and Chromatin immuno-precipitation followed by microarray hybridization (ChIP-chip) and DNA sequencing (ChIP-Seq) have enabled the characterization of the whole genome TFs binding sites (TFBS) and permitted the development of several computational DNA motif discovery tools. Although these various tools are widely used and have led to the discovery of novel motifs, but in practice none of them have proven to be efficient in control data sets due to a high rate of false positive and false negative predictions.

Mainly, DNA motif discovery tools use two different strategies to extract the motif patterns: Enumerative and alignment-based approaches. Each of these approaches uses a specific scoring function to evaluate motifs by comparing them to reference data sets then report those having the highest scores. Analyses conducted on simulated data using enumerative algorithms (Mdscan, Weeder) and alignment based algorithms (MEME, MotifSampler) have shown that DNA motif discovery tool scoring functions do not represent the observed characteristics of TFBS in ChIP regions. For example, they do not take into account that motifs representing real binding sites are more likely to reside near the center of the ChIP fragments. Our results showed that these scoring functions are not adapted to the discovery of TFBS in ChIP-chip/ChIP-Seq data.

We propose to implement two new scoring functions: The former scoring function measures how a given motif is distributed across the ChIP regions. It is expected that true binding sites will be enriched in specific positions (peak in the ChIP central regions) compared to what is expected by chance (reference data set). The latter scoring function is a measure of how a given motif targets the set of ChIP sequences. We expect that true TFBS will be distributed evenly throughout the ChIP sequences and are clearly more frequent compared to a reference data set.

We applied the new scoring functions on simulated data sets and on ChIP-chip and ChIP-Seq data sets. The results show that our approach enhances significantly the DNA motif discovery tools performances and significantly reduces the rate of false positive and false negative predictions.

# SSBBN: Gene Regulatory Network Construction using Spectral Subtraction Denoising, Biclustering and Bayesian Network

Fadhl M. Al-Akwa<sup>1,2</sup>, Nahed H. Solouma<sup>2</sup>, Yasser M. Kadah<sup>2</sup>

<sup>1</sup>*Biomedical Engineering Department, University of Science and Technology, Sana'a, Yemen*

<sup>2</sup>*Biomedical and System Engineering Department, Cairo University, Giza, Egypt*

The increasing development of high throughput technology like microarray, promotes researchers to study the complexity of gene regulatory network (GRN) in biological cells. GRN inference algorithms have much impact in drug development and in understanding disease ontology. The great challenges in GRN modelling are dimensionality reduction and denoising of microarray data. We propose an integrated algorithm (SSBBN) for denoising, biclustering and learning to overcome these problems. Firstly, the microarray dataset is denoised using our Spectral Subtraction novel method to decrease the false positive rate. Secondly, we divide the whole set of genes into a number of overlapped biclusters using our proposed BicAT-Plus. Thirdly, these biclusters are learned using Greedy Hill Climbing search algorithm to produce subnetworks. Finally, these subnetworks could be integrated to produce the whole Gene Regulatory Network. The proposed method was applied to time series gene expression data of *Saccharomyces Cerevisiae*. The generated network was validated via available interaction databases and the result revealed the performance of our proposed method. The approach could potentially be applied to other networks in yeast as well as higher organisms.

BicAT-Plus can be downloaded from <http://home.k-space.org/BicAT-plus.zip>

# Scalable Steady State Analysis of Boolean Biological Regulatory Networks

Ferhat Ay<sup>1</sup>, Fei Xu<sup>1</sup>, Tamer Kahveci<sup>1</sup>

<sup>1</sup>*Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611;*

**Background:** Computing the long term behavior of regulatory and signaling networks is critical in understanding how biological functions take place in organisms. Steady states of these networks determine the activity levels of individual entities in the long run. Identifying all the steady states of these networks is difficult due to the state space explosion problem.

**Methodology:** In this paper, we propose a method for identifying all the steady states of Boolean regulatory and signaling networks accurately and efficiently. We build a mathematical model that allows pruning a large portion of the state space quickly without causing any false dismissals. For the remaining state space, which is typically very small compared to the whole state space, we develop a randomized traversal method that extracts the steady states. We estimate the number of steady states, and the expected behavior of individual genes and gene pairs in steady states in an online fashion. Also, we formulate a stopping criterion that terminates the traversal as soon as user supplied percentage of the results are returned with high confidence.

**Conclusions:** This method identifies the observed steady states of Boolean biological networks computationally. Our algorithm successfully reported the G1 phases of both budding and fission yeast cell cycles. Besides, the experiments suggest that this method is useful in identifying co-expressed genes as well. By analyzing the steady state profile of Hedgehog network, we were able to find the highly co-expressed gene pair **GL1-SMO** together with other such pairs.

**Availability:** Source code of this work is available at <http://bioinformatics.cise.ufl.edu/palSteady.html>

[Original Full Length Paper]

## A scalable algorithm for discovering conserved active subnetworks across species

Raamesh Deshpande<sup>1</sup>, Shikha Sharma<sup>2</sup>, Wei-Shou Hu<sup>2</sup>, Chad L. Myers<sup>1</sup>

<sup>1</sup>*Department of Computer Science and Engineering, University of Minnesota - Twin Cities, Minneapolis, Minnesota, United States of America;* <sup>2</sup>*Department of Chemical Engineering, University of Minnesota - Twin Cities, Minneapolis, Minnesota, United States of America*

Overlaying differential changes in gene expression on protein-protein interaction (PPI) networks has proven to be useful approach to interpreting the cell's dynamic response to a changing environment. Following a landmark study by Ideker et al. (2002) that originally defined the problem of finding active subnetworks, this area has received substantial attention, including additional studies that extended this approach to larger or confidence-weighted PPI networks (Rajagopalan et al, 2005, Ulitsky et al, 2009), to incorporate expression correlation between genes, (Cabusora et al, 2005, Ulitsky et al. 2009), or that formulated the problem as an optimization problem with an integer programming solution (Dittrich et al. 2008).

Despite successes in many of these directions, the idea of overlaying lists of differentially expressed genes on networks has not yet been extended to compare specific gene expression responses across species. Comparative studies can help to identify core conserved behaviors, uncover interesting evidence for rewiring through evolution, and provide a powerful basis for general functional characterization.

We have designed a scalable, cross-species network search algorithm, that finds conserved, active subnetworks based on parallel differential expression studies in multiple species. Our approach leverages recent work from the genomic data integration community, which has developed approaches for combining heterogeneous collections of genomic data to produce weighted functional linkage networks. These networks now exist for several organisms, including higher eukaryotes such as mouse and human and typically include millions of edges (orders of magnitude more dense than PPI networks) that capture not just physical interactions but include more general functional relationships. These networks provide more comprehensive maps of the cell, but demand more efficient and scalable algorithms for active subnetwork discovery, particularly in the case of the multiple species problem.

We applied our cross-species approach to identify conserved subnetworks that are differentially active in stem cells relative to differentiated cells based on parallel gene expression studies in mouse and human. In contrast to a random expression model, we find hundreds of conserved active subnetworks enriched for functions such as cell cycle, DNA repair, and chromatin modification processes that are characteristic of stem cells. Using a variation of this approach, we also find a number of species-specific networks, which likely reflect mechanisms of stem cell function that have diverged between mouse and human. We discuss the network randomization model used for assessing the significance of subnetworks derived from our approach, and describe several case examples that illustrate the utility of comparative analysis of active subnetworks.

# Dynamic Flux Estimation – A novel framework for metabolic pathway analysis

Gautam Goel<sup>1</sup>, Luis Lopes da Fonseca<sup>2</sup>, Eberhard O. Voit<sup>3</sup>

<sup>1</sup>Center for Computational and Integrative Biology, Massachusetts General Hospital;

<sup>2</sup>Institute of Chemical and Biological Technology- ITQB, Portugal; <sup>3</sup>The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology

At the center of computational systems biology are mathematical models that capture the dynamics of biological systems and offer novel insights. The bottleneck in the construction of these models is presently the identification of model parameters that make the model consistent with observed data. In this talk I will present a novel methodological framework, titled Dynamic Flux Estimation (DFE), for estimating parameters for models of metabolic systems from time-series data.

DFE consists of two distinct phases, an entirely model-free and assumption-free data analysis phase and a model-based mathematical characterization of process representations. The model-free phase reveals inconsistencies within the data, and between data and the alleged system topology, while the model-based phase allows quantitative diagnostics of whether--or to what degree--the assumed mathematical formulations of enzyme kinetics are appropriate or in need of improvement. I will elucidate these concepts using several examples of successively increasing degree of complexity and likeness to real-life situations and demonstrate how DFE can be used a framework to integrate in-vivo time-series data with in-vitro kinetic information.

I will present our initial results suggesting that the proposed approach is more effective and robust than presently available methods for deriving metabolic models from time-series data. I will then present insights obtained from the application of this framework to address the real-world problem of analyzing glycolytic control in *Lactococcus lactis*. I will demonstrate how the temporal patterns in dynamic flux estimates, obtained from the model free phase, have raised new question about how we think the glycolytic flux is controlled. I will discuss the intricacies of selecting and fitting kinetic functions and show how simple local functional analysis can help us build a hypothesis about flux control.

# Reconstruction of a dynamic regulatory map from murine liver regeneration data

Sandro Lambeck<sup>1</sup>, Reinhard Guthke<sup>1</sup>

<sup>1</sup>Leibniz Institute for Natural Product Research and Infection Biology – Hans Knoell Institute, Beutenbergstr. 11a, D-07745 Jena, Germany

**Background:** Major efforts have been done within *systems biology* to reconstruct complex interactions in order to understand cell's mechanisms. Therefore, expression profiles obtained from perturbation experiments are increasingly used to reconstruct dynamic transcriptional regulatory networks.

**Methods:** In this work, we show how to combine a knowledge network of transcriptomic regulations taken from the cisRED database with expression profiles from different data sets describing the same process, in order to infer the key regulators and their effects. These are estimated by a system of difference equations, potentially bridging the gap from co-expression to co-regulation. We provide application for murine liver regeneration data after partial hepatectomy, because understanding the processes involved will likely assist the treatment of a variety of serious liver diseases and may have important implications for certain types of therapy.

**Results:** Systems biology approaches were developed to integrate gene expression data and *a priori* knowledge to identify gene targets for modulating liver regeneration. Analysing two datasets from different labs, we found an highly homogeneous orchestration by a large number transcription factors being also differentially expressed on the transcriptional level. Functional associated targets being highly reflected within the data, suggesting an cooperative interplay of transcription factors in core hubs of the network. We predicted an essential array of 21 transcription factors (e.g. jun, tcf4, atf3, cebpb, foxm1) being capable to rapidly trigger time dependent mitogenic cascades during the proliferation phase.

# Inferring Fault Tolerance from E-MAP Data

Diana Tatar<sup>\*1</sup>, Mark D.M. Leiserson<sup>\*1</sup>, Lenore Cowen<sup>1</sup>, Benjamin Hescott<sup>1</sup>

<sup>1</sup>Dept of Computer Science, Tufts University, Medford, MA 02155, email: hescott@cs.tufts.edu.

The well-studied between-pathway model (BPM), introduced first by Kelley and Ideker [1] is a network motif consisting of a particular pattern of genetic and physical interactions that is thought to signify a pair of redundant pathways within a protein-protein interaction network. In particular, each BPM consists of two subsets of genes (each called a *pathway*) where physical interactions tend to occur between pairs of genes within the same subset, and synthetic lethal interactions tend to occur between pairs of genes in different subsets. It has been shown by Kelley and Ideker and subsequent work [2,3,4] that BPM pathways show significant biological enrichment for functional coherence, mostly using known ontological annotation [1,2,3,4], but also recently, functional coherence of predicted BPM modules based on gene expression data has also been demonstrated [5]. All these methods are based on binary genetic interaction data, that is, a pair of proteins is in a synthetic lethality relationship, or it is not.

Epistatic miniarray profiles (E-MAPs) are a recently developed high-throughput tool capable of providing quantitative scores for synergistic or alleviating reactions between gene pairs. Complete E-MAPs have been published for a set of *S. cerevisiae* genes involved in cell cycle [6] and MAPK pathways [7]. The question thus arises as to whether the BPM paradigm can be adapted to make use of the more expressive non-binary genetic interaction data from an E-MAP, to reveal modules of fault tolerance in these new, more complete networks.

We show a natural generalization of the original BPM motif that captures an analogous notion for E-MAP genetic interaction data, and present an adaptation of the method of Brady et al. [4] to efficiently generate new motifs. As in previous studies, we measure the quality of the pathways we predict based on GO enrichment. We generate over 100 generalized BPMs based on the MAPK array, and over 200 generalized BPMs based on the Cell Cycle array, using the new method. Furthermore, over 46% of the MAPK BPMs and over 48% of the Cell Cycle BPMs exhibit GO enrichment ( $p < .01$ ) for at least one pathway.

\*Both authors contributed equally to this work.

[1] Kelley R., Ideker T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 23: 561-566.

[2] Ulitsky I., Shamir R. (2007) Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction network. *Mol Syst Bio* 3:104.

[3] Ma X., Tarone A.M., Li W. (2008) Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS ONE* 3:e1922.

[4] Brady A., Maxwell K., Daniels N., Cowen L.J. (2009) Fault Tolerance in Protein Interaction Networks: Stable Bipartite Subgraphs and Redundant Pathways. *PLoS ONE* 4(4): e5364.

[5] Hescott B., Leiserson M., Cowen L.J., Slonim D. (2009) Evaluating Between-Pathway Models with Expression Data. *Proceedings of the 13<sup>th</sup> RECOMB*, 372-385.

[6] Collins et al. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446, 806-810.

[7] Fiedler et al. (2009) Functional Organization of the *S. cerevisiae* Phosphorylation Network. *Cell* 135.

# The RNA Editing Dataflow System (REDS) for the transcriptome-wide discovery of RNA modification sites

Stefan Maas<sup>1</sup>, Christina P. Godfried Sie<sup>1</sup>, Dylan E. Dupuis<sup>1</sup>, Ivan Stoev<sup>2</sup> and Daniel Lopresti<sup>2</sup>

<sup>1</sup>*Department of Biological Sciences, Lehigh University;* <sup>2</sup>*Department of Computer Science and Engineering, Lehigh University.*

RNA editing by adenosine deamination, catalyzed by the adenosine deaminases acting on RNA (ADARs), is a posttranscriptional mechanism for the regulation of gene expression and particularly widespread in mammals. A-to-I RNA editing generates transcriptome and proteome diversity and also regulates important functional properties of neurotransmitter receptor genes in the central nervous system by changing single codons in pre-mRNA.

We have previously identified wide-spread editing of non-coding transcripts [1]. In contrast, almost all currently known cases of A-to-I RNA editing that affect protein-coding sequences have been discovered serendipitously. However, since it is expected that many more such recoding editing sites exist and to understand the overall importance of RNA editing in gene regulation, it is crucial to map RNA editing sites in a systematic way.

Here we present the RNA Editing Dataflow System (REDS), a computational pipeline that allows us to predict A-to-I RNA editing sites (as well as other types of RNA modifications) in any genome for which genomic and expression databases are available. We show that a high percentage of the predicted target sites are likely bona fide editing events and go on to experimentally validate novel recoding events in three vertebrate species.

Apart from the identification of novel editing sites, our analysis provides insights on the overall landscape of RNA editing, insights on why certain sequences and RNA folds are more prone to undergo RNA editing, and with REDS provide a computational tool that should foster progress in the discovery of RNA modification well beyond this study.

[1] Athanasiadis, A., Rich, A., and Maas, S. 2004: Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biology*, 2 (12), e391, 1-15, Epub 2004 Nov 9.

# Linking MicroRNA and mRNA Co-Expressed Clusters to Regulatory Networks in Cancer

Amin Mazloom<sup>1,2</sup>, Nabil Ahmed<sup>3</sup>, Avi Ma'ayan<sup>1,2</sup>

<sup>1</sup>Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York; <sup>2</sup>Systems Biology Center New York (SBNY); <sup>3</sup>Herricks High School, New Hyde Park, New York

miRNAs are single stranded non-coding RNA molecules consisting of approximately 21-23 nucleotides found primarily in mammalian cells. miRNAs function as down-regulators of gene expression by binding to their complimentary targeted mRNAs. Malignant cells have a significantly different miRNA expression patterns than healthy cells. As such, microRNAs represent a potential new class of drug targets. In this project we identified clusters of differentially expressed miRNAs in 60 tumor cell lines from the NCI-60 panel, and then linked these clusters with differentially co-expressed mRNAs in these tumor cells. To link clusters of co-expressed miRNAs with clusters of co-expressed mRNAs we used databases of predicted miRNA-mRNA interactions. We find clear relations between aberrantly co-expressed mRNAs with their potential group of miRNA regulators in several different cancer types, further analysis using prior biological knowledge provides mechanisms by which such relations may have come to being, potentially suggesting novels pathway to new cancer therapies.

# Exploring the Monochromatic Landscape in Yeast using Genetic Interactions and Known Pathways

Magali Michaut, Gary Bader

*Terrence Donnelly CCBR, University of Toronto, Ontario, Canada*

The eukaryotic cell is often described as a hierarchy of systems (complexes and pathways) composed of genes or proteins. These systems can be defined at multiple levels (e.g. one complex containing several subunits or different complexes interacting together). Identifying these systems and their relationships from experimental data is important to organize genes at a system level but still challenging. Hierarchical clustering or model-based network analysis has been proposed to extract modules from genetic interaction networks. These methods are useful but define a flat set of modules often describing complexes. Boundaries and system level relationships are still difficult to define.

We propose to use known pathways, complexes and their relationships as defined in Gene Ontology (GO) annotations to better define cellular system organization. We assess the coherence of systems defined by the biological process ontology using quantitative genetic interactions classified into positive and negative interactions. Previous studies have shown that complexes tend to be monochromatic (enriched either in positive or negative interactions). We use this idea and propose a score (the relative ratio of positive to negative interactions) to characterize a GO term by its monochromatic purity, based on the set of genes associated to it. We define a coverage measure to assess only the GO terms sufficiently covered by the genetic interaction data. We then generate random networks and compute z-scores to assess the likelihood of the scores obtained and extract unexpected patterns.

Using these coverage and monochromatic measures, we show that 10-20% of the known biological processes are monochromatic. The classical within-pathway model can explain GO terms with highly positive monochromatic z-scores. Nevertheless, some GO terms with highly negative monochromatic z-scores are surprising and show new connectivity patterns. Finally, we use the monochromatic purity to connect new genes to specific GO terms, based on their connectivity with the GO term. We are developing a method, based on this, for gene function prediction.

To our knowledge, this work is the first general study of the monochromatic purity landscape of known biological processes. This new approach is developed in a hierarchical way, enabling to define system boundaries at several levels and to organize a map of the yeast cell. This work will hopefully help define gene function in the context of relevant complexes and pathways and can further be used to automatically select gene sets from GO at an appropriate level of the hierarchy to best annotate a given genomic data set.

## Enrichment and aggregation of topological motifs in integrated interaction networks

Tom Michoel<sup>1,2</sup>, Bruno Nachtergaele<sup>3</sup>, Yves Van de Peer<sup>1,2</sup>

<sup>1</sup>Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium.

<sup>2</sup>Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium.

<sup>3</sup>Department of Mathematics, University of California, Davis, One Shields Avenue, Davis, CA 95616-8366, USA.

Reconstructing the organizational principles that determine the structure and function of protein interaction and regulatory networks is a key challenge of systems biology. Network motifs, small subgraphs occurring significantly more often than expected by chance, have been proposed as the basic building blocks of complex networks, including integrated networks composed of multiple types of interactions. In transcriptional regulatory networks, network motifs are known to aggregate into larger, self-contained units. This concept was extended to integrated networks and for some motifs, 'network themes' were found, frequently recurring higher-level patterns of overlapping network motifs, which characterize the structure of functional modules. These results suggest a strong connection between network motif enrichment and aggregation.

Here we introduce a novel computational method for identifying dense clusters of network motifs in integrated networks. It uses a generalized spectral approach to partition motif instances into high-scoring clusters, such that nodes may belong to multiple clusters. We used this algorithm to investigate the relation between network motif enrichment and aggregation using all possible three-node motifs in a network containing more than 50,000 curated protein-protein, transcription factor binding and phosphorylation interactions in *Saccharomyces cerevisiae*. Although our analysis confirms that well-studied motifs like the transcriptional feed-forward loop (FFL) and the coregulated interacting proteins motif are significantly enriched as well as aggregating, we find that in general network motif enrichment and aggregation are independent organizational principles. This is demonstrated at several different levels. First of all, a large-scale statistical comparison between the real networks and an ensemble of randomized networks shows that the enrichment and aggregation significance scores are uncorrelated, i.e. there exist network motifs which are significantly enriched but not aggregating, and vice versa. Secondly, we show that a single motif (e.g. the protein interaction-mediated transcriptional or post-translational feedback loops) can give rise to multiple, functionally and topologically distinct aggregated network themes. Thirdly, we show that a mixed post-translational - transcriptional FFL, which is not enriched yet significantly aggregating, most likely does not play a universal information-processing role, but is important for a specific biological process, in this case the cell cycle.

We conclude that the modular organization of biological interaction networks built from aggregating network motifs is considerably more complex than the organization of hierarchical scale-free networks or networks grown by simple duplication-divergence mechanisms, indicating a need for further refinement of these models. The network motif clustering algorithm introduced here is a powerful and versatile technique for integrating large-scale datasets and for reconstructing their modular organization, with a wide range of potential applications.

## microRNAs preferentially target dosage-sensitive genes

Martin L. Miller<sup>1</sup>, Debora S. Marks<sup>2</sup>

<sup>1</sup>*Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA;* <sup>2</sup>*Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA.*

Surprisingly, most genes are not considered “dosage sensitive” and expression values may fluctuate physiologically across cells with no known pathological effects. On the other hand, groups of genes are known to be particularly sensitive to overexpression and dysregulation may have harmful consequences, for example upregulation of oncogenes. MicroRNAs can tune the expression levels of genes, but paradoxically such regulation levels are relatively small compared to the natural fluctuation of gene expression.

It is known that dosage-sensitive genes are tightly regulated at multiple stages, including by post-transcriptional microRNA regulation. If this is the case, we may expect to see dosage-sensitive genes enriched for microRNA regulation and that this phenomena should be evolutionarily conserved.

To test this hypothesis we investigated the relationship between the number of microRNA targets in a gene and its likelihood of being a dosage-sensitive.

Across worm, fly and mammals, we find that dosage-sensitive genes are more targeted than the rest of the genome, an observation that was not found in for example essential (lethal) genes.

As a consequence, we speculate that microRNA-based perturbations will have different effects on dosage-sensitive genes compared to other genes. Preliminary observations support this as we find that oncogenes are less down-regulated than other target genes after small RNA transfection, and further, inhibiting microRNAs with antagomirs up-regulates oncogenes with microRNA targets less than expected.

# Predicting Genetic Interactions in *C. elegans* using Machine Learning

Patrycja Vasilyev Missiuro<sup>1,2</sup>, Hui Ge<sup>2</sup>, Tommi S. Jaakkola<sup>1</sup>

<sup>1</sup>Computer Science and Artificial Intelligence Lab (CSAIL), MIT, Cambridge, MA;

<sup>2</sup>Whitehead Institute, MIT, Cambridge, MA

Our main objective is the discovery of genetic interactions based on sparse and incomplete information. We develop a set of machine learning techniques to investigate and predict gene properties across a variety of *Caenorhabditis elegans* datasets.

First, we show how Bayesian sets method can be applied to gain intuition as to which datasets are the most relevant for predicting genetic interactions. In order to directly apply this method to microarray data, we extend Bayesian sets to handle continuous variables. Using Bayesian sets, we show that genetically interacting genes tend to share phenotypes but are not necessarily co-localized.

One of the major difficulties in dealing with biological data is the problem of incomplete datasets. We describe a novel application of collaborative filtering (CF) in order to predict missing values in the biological datasets. We adapt the factorization-based and the neighborhood-aware CF<sup>1</sup> to deal with a mixture of continuous and discrete entries. We use collaborative filtering to input missing values, assess how much information relevant to genetic interactions is present, and, finally, to predict genetic interactions. We also show how CF can reduce input dimensionality.

Using collaborative filtering we fill in the missing entries in the input data describing genes. Since the input matrix is no longer sparse, we are able to use Support Vector Machines to predict genetic interactions. We find that SVM with a nonlinear *rbf* kernel has greater predictive power over CF.

Overall, our approaches achieve substantially better performance than previous attempts at predicting genetic interactions. We emphasize the features of the studied datasets and explain our findings from a biological perspective. We hope that our work possesses an independent biological significance by helping one gain new insights into *C. elegans* biology: specific genes orchestrating developmental and regulatory pathways, response to stress, etc.

[1] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. Proceedings of the 13<sup>th</sup> SIGKDD international conference on Knowledge Discovery and Data Mining, pp 95-104, 2007.

# Identifying uncharacterized genes and functional networks in human autophagy

Aylwin Ng<sup>1,2</sup>, Ramnik Xavier<sup>1,2,3</sup>

<sup>1</sup>Center for Computational & Integrative Biology and <sup>2</sup>Gastrointestinal Unit, Massachusetts General Hospital and Harvard Medical School; <sup>3</sup>Broad Institute of MIT and Harvard.

Autophagy is a conserved physiological process through which defective organelles and protein aggregates are cleared. Environmental perturbation triggering starvation can induce the autophagic response. Additionally, in mammalian systems, recent evidence suggests an emerging role of autophagy in host-pathogen interactions, innate and adaptive immunity. In human, autophagy has also been implicated in chronic inflammatory bowel disease. A core autophagy gene, *ATG16L1* was identified as a Crohn's disease gene from genome-wide association studies. While the core apparatus of classical autophagy has been relatively well characterized in yeast, pathways and signaling components linking perturbation states such as infection to autophagy in human are, at present, not well understood.

To gain insights about networks and as yet unidentified ancillary effectors involved in the human autophagic process, we took a yeast-human and fly-human interolog mapping approach alongside mammalian interactomes to define a core network underlying the human autophagy process. Our analysis identified more than 30 autophagy-associated proteins, 14 of which had no previously identified roles in mammalian autophagy. We selected a candidate for complete functional validation and identified a novel interaction between ATG3, a component of the core autophagy apparatus and FNBP1L (a human ortholog of the yeast BZZ1) which we found to be important for anti-bacterial autophagy.

Further context-dependent extensions of the core autophagy network were constructed by anchoring on orthogonal data from time-series gene expression, cI-CAT-MS-based proteomic analyses, functional RNAi and genetic screens examining a variety of perturbation conditions (rapamycin, starvation, metabolic disruptions or infection) that impinge on the autophagy process. We demonstrate the utility of this multi-tiered approach towards novel molecule discovery and the construction of functionally relevant networks for autophagy in various biological contexts.

## Genome-wide identification of Transcription Factor Binding Sites using DNase-seq footprints and other annotations

Roger Pique-Regi<sup>1,\*</sup>, Jacob F. Degner<sup>1,2,\*</sup>, Athma A. Pai<sup>1</sup>, Daniel J. Gaffney<sup>1</sup>, Greg Crawford<sup>3</sup>, Yoav Gilad<sup>1</sup>, Jonathan K. Pritchard<sup>1,4</sup>

<sup>1</sup>Department of Human Genetics, The University of Chicago; <sup>2</sup>Comitee on Genetics Genomics and Systems Biology, The University of Chicago; <sup>3</sup>Institute for Genome Sciences & Policy, Duke University; <sup>4</sup>Howard Hughes Medical Institute. \*These authors contributed equally to this work.

One of the great remaining challenges in genomics is to identify and interpret the regulatory elements of the genome. Transcription factors (TF) bind to specific DNA sites, characterized by sequence motifs, to control the assembly of the transcriptional machinery and the rate of transcription. Fundamental to our understanding of the eukaryotic regulatory code, will be a complete understanding of the DNA targets of each transcription factor across tissue types and developmental stages.

Here, we propose a novel Bayesian hierarchical mixture model that uses the characteristic footprint of each TF given by genome-wide DNase-I cut-site data to predict which TF motif instances are bound. For each TF with a known sequence motif in TRANSFAC and JASPAR databases (1004 motifs) all the instances (~100 million) across the genome are retrieved. Then, an Expectation Maximization (EM) algorithm is used to learn a mixture model that captures the probability distribution of the DNase-I cut-sites around bound versus unbound motif matches. After fitting this probabilistic model, we obtain a posterior probability that a specific TF binding sites is active (i.e. bound by a TF). This framework can also include, using a logistic model, genomic annotations that can affect the prior probability, such as: motif match score, conservation score and distance to the transcription start site (TSS).

We validate our results using publicly available ChIP-seq data for six transcription factors and we achieve very good prediction performance for the majority of these TF. Compared to ChIP-seq, DNase-seq can target all TF with a known motif within the same experimental assay, there is no need to develop a specific antibody or epitope tagged protein, and using this method, there is a precise delineation of the location of TF binding. Using the proposed methods on the entire set of TFs in human B-cells, we are now developing a more-complete picture of the components of gene regulation in these cells.

# Unraveling of an ancient regulatory pathway: RNAi insensitivity in the germline of *C. elegans*

Daniel A. Pollard<sup>1,2</sup>, Maxwell J. Kramer<sup>1,2</sup>, Matthew V. Rockman<sup>1,2</sup>

<sup>1</sup>Center for Genomics & Systems Biology, New York University; <sup>2</sup>Department of Biology, New York University

Eukaryotes utilize the microRNA and RNA interference (RNAi) pathways to regulate gene expression post-transcriptionally. microRNAs regulate endogenous gene expression and play critical roles during development. The natural role for the RNAi pathway is less well understood but is believed to function in protecting the germline from mobile elements and exogenous RNA. The core molecules in these regulatory pathways are well characterized and well conserved across eukaryotes however the components and deployment of the pathways vary substantially both within and across species. The round worm *Caenorhabditis elegans* has played a central role in the elucidation of these pathways and yet the commonly used Hawaiian natural isolate is completely insensitive to RNAi in the germline. We sought to expand the characterization of the loss of RNAi sensitivity in *C. elegans* to identify new genes involved in the pathway and better understand the genetics underlying the loss of this pathway in the population.

Using quantitative trait locus mapping techniques we first examined the genetics underlying the difference in germline RNAi sensitivity between the fully sensitive N2 lab strain and the fully insensitive Hawaiian isolate. Although the insensitivity does not segregate as a single locus Mendelian trait (implying complexity), mapping revealed only one large-effect QTL, centered over the argonaut gene *ppw-1* (previously implicated in this insensitivity). We next surveyed germline RNAi sensitivity in 41 natural isolates representing all known *C. elegans* haplotypes. Germline RNAi insensitivity is common and geographically widespread in natural populations, with nearly one in four natural isolates showing insensitivity. Surprisingly, association mapping with the natural isolates implicated no significant genomic regions, suggesting a high level of genetic complexity or heterogeneity. To test if the RNAi insensitivity in the population has heterogeneous causes we performed complementation tests between the insensitive natural isolates and a *ppw-1* null strain. The *ppw-1* null both complemented and failed to complement natural isolates, suggesting germline RNAi insensitivity may have been gained multiple times through separate mechanisms. We are currently performing QTL mapping of RNAi sensitivity in the *ppw-1* complementing natural isolates.

We conclude that germline RNAi insensitivity is a widespread and complex trait in *C. elegans* with heterogeneous and potentially novel underlying molecular mechanisms. We propose that the losses of the RNAi pathway response in the germline may be the result of relaxed selection due to infrequent outcrossing and small effective population size in *C. elegans*. These results suggest that we have caught an ancient and highly conserved regulatory pathway in the process of unraveling and falling apart.

# Prediction of Chromatin Modification (CM)-related Functional Domains and Genes in Human

Shuye Pu<sup>1</sup>, Andrei Turinsky<sup>1</sup>, Shoshana J. Wodak<sup>1,2,3</sup>

<sup>1</sup>Molecular Structure and Function Program, the Hospital for Sick Children, Toronto, Ontario, Canada; <sup>2</sup>Department of Biochemistry and; <sup>3</sup>Department of Medical Genetics, University of Toronto, Ontario, Canada.

Modifications of chromatin structure and states, through ATP-dependent chromatin remodeling, histone exchange, as well as chemical modifications of DNA and histone molecules, affect positioning, composition and biochemical states of nucleosomes, which compact genomic DNA and dictate its accessibility. Chromatin modification (CM) has profound impact on all DNA-based processes such as transcription, replication, repair and recombination, and thus is key to gene regulations underlying various physiological and disease processes. CM is carried out by multi-protein complexes whose subunits are composed of distinct functional domains.

We analyzed 25 known CM-related domains (CM domain) in 5 model organisms, including the yeast *S. cerevisiae*, the worm *C. elegans*, the fly *D. melanogaster*, the mouse *M. musculus* and the human *H. sapiens*, and found that domain families involved in histone methylation, DNA methylation and histone variants are remarkably expanded in mouse and human, presumably reflecting the increased demand for cell type-specific gene regulation (mostly repression) in these organisms.

Using a domain co-occurrence network simulation procedure that emulates domain-pair duplications, we found that CM domains are not promiscuous, when corrected for domain abundance. By analyzing the pair-wise co-occurrence of CM domains with other CM and non-CM domains, we identified 47 potentially novel CM domains. Among these, 24 are DNA binding domains, whose role in CM has received little attention so far.

CM is best understood in simple eukaryotes like yeast, but very limited knowledge is available for humans. We collated lists of 312 yeast CM genes and 398 human CM genes, all supported by experimental evidence from the Gene Ontology database, protein complex databases and the literature. Using their domain composition as a feature, we predicted additional 379 putative CM genes in human using a supervised machine learning technique (Support Vector Machines). Experimental evidence for some of these genes (e.g., JARID2, ALC1, ATAC2) has been found in the recent literature (1-3), and the involvement of remainder genes in CM is currently subjected to experimental verifications. Identification of novel CM genes in human will aid our understanding of gene regulations that are important for stem cell and tumor development and treatment (This work is supported by CIHR Team Grant CTP 82940 and funding from Sickkids Foundation to SJW).

## References

1. Shirato, H., *et al.* (2009) *J Biol Chem*, **284**, 733-739.
2. Ahel, D., *et al.* (2009) *Science*, **325**, 1240-1243.
3. Guelman, S., *et al.* (2009) *Mol Cell Biol*, **29**, 1176-1188.

## Effective identification of conserved pathways in biological networks using hidden Markov models

Xiaoning Qian<sup>1</sup> and Byung-Jun Yoon<sup>2</sup>

<sup>1</sup>*Department of Computer Science & Engineering, University of South Florida;*

<sup>2</sup>*Department of Electrical & Computer Engineering, Texas A&M University.*

The advent of various high-throughput experimental techniques for measuring molecular interactions has enabled the systematic study of biological interactions on a global scale. Since biological processes are carried out by elaborate collaborations of numerous molecules that give rise to a complex network of molecular interactions, comparative analysis of these biological networks can bring important insights into the functional organization and regulatory mechanisms of biological systems.

In this paper, we present an effective framework for identifying common interaction patterns in the biological networks of different organisms based on hidden Markov models (HMMs). Given two or more networks, our method efficiently finds the top  $k$  matching paths in the respective networks, where the matching paths may contain a flexible number of consecutive insertions and deletions. Based on several protein-protein interaction (PPI) networks obtained from the Database of Interacting Proteins (DIP) and other public databases, we demonstrate that our method is able to detect biologically significant pathways that are conserved across different organisms. Our algorithm has a polynomial complexity that grows linearly with the size of the aligned paths. This enables the search for very long paths with more than 10 nodes within a few minutes on a desktop computer.

# Learning probabilistic networks of condition-specific response: Digging deep in yeast stationary phase

Sushmita Roy<sup>1</sup>, Margaret Werner-Washburne<sup>2</sup>, Terran Lane<sup>1</sup>

<sup>1</sup>UNM Computer Science; <sup>2</sup>UNM Biology

Central to the proper functioning of living systems is the ability to accurately sense environmental cues and respond to changing conditions. This ability of producing different *condition-specific responses* involves global changes of the cellular parts list (genes, proteins and metabolites). *Condition-specific networks* are networks of functional interactions describing how the parts interact as cells function under changing conditions. Condition-specific networks can provide a comparative, systems-level understanding of the role of various bio-chemical pathways in maintaining the health and well-being of a cell, as well as the failure points of these pathways causing diseases.

Recent developments in systems biology have produced numerous network-based approaches for capturing condition-specific behavior (Chuang *et al.*, 2007, Oleg *et al.*, 2007, Sanguinetti *et al.*, 2008, Bergmann *et al.*, 2004). However, these approaches often assume that the network is known, identify pair-wise co-expression relationships rather than general statistical dependencies, and often focus on differences rather than both differences and similarities across conditions. Many of these approaches infer networks for each condition separately, and do not exploit the shared information across conditions *during* network learning.

We describe a novel approach to condition-specific response analysis, Network Inference with Pooling Data (NIPD) that jointly infers networks for the different conditions in a *multiple-network* learning framework. NIPD is based on Probabilistic Graphical Models (PGMs) where edges represent pair-wise and higher-order statistical dependencies among genes. The multiple-network learning framework searches for the best set of networks using a novel score that evaluates candidate networks with respect to data from any subset of conditions, pooling data for subsets with more than one condition. This data pooling property of NIPD exploits shared information across the conditions during structure learning and enables us to learn better network structures than approaches learning networks per condition independently, especially on small training datasets.

We applied our approach to microarray data from two yeast stationary-phase cell populations, *quiescent* and *non-quiescent*. Compared to an approach inferring networks for each population independently, networks learned by NIPD were associated with many more biological processes, or were enriched in targets of known transcription factors (TFs). Many of the TFs were involved in stress response, which is consistent with the fact that the populations are under starvation stress. Comparative analysis of the inferred networks implicated respiration-related processes to be common across the two populations, whereas regulation of epigenetic expression to be specific to quiescent cells, consistent with known characteristics of these cells. We also found several cases of combinatorial interaction among single gene deletions that can be experimentally tested, and that will contribute to our understanding of differentiated cell populations in yeast stationary phase. To conclude, NIPD is a multiple-network learning framework for network-based characterization of unique and shared response patterns, thus providing a holistic picture of condition-specific response mechanisms in cells.

## **PPi module for visualization and analysis of protein-protein interfaces in Friend.**

Amit Upadhyay, Valentin A. Ilyin.

*Department of Biology, Boston College. Chestnut Hill, MA*

The interactions among proteins are largely responsible for the complexity of biological systems. Identification of interactions between proteins will help in elucidation of protein networks providing insights into biological processes which in turn finds application in drug designing, protein engineering and developing scoring functions for docking. Studying the interfaces between interacting proteins will provide greater understanding of principles governing the binding of proteins. A number of methods have been described for studying protein interfaces. But none of the methods could be effectively used in predicting protein-protein recognition sites.

We describe an application library for visualization and analysis of protein-protein interfaces based on Voronoi-Delaunay tessellation (VDT), in the Friend software (<http://ilyinlab.org/friend>, an integrated analytical application designed for simultaneous analysis and visualization of multiple structures and sequences of proteins and/or DNA/RNA). The presented PPi method using VDT is more objective as compared to those based on changes in solvent accessible surface area ( $\Delta$ SASA) and various radial cutoffs that lead to ambiguity. The library takes structural data files in PDB format as input and enables visualization of the chain-level interfaces as well as a detailed qualitative and quantitative analysis of the interfaces. There is no restriction with respect to the number of chains present in the PDB files and also interactions involving heteroatom's can be considered unlike most methods. This library can therefore be extended to study protein-ligand interactions as well.

We compared results of interface analysis by the PPi-Friend with other methods on a number protein complexes consisting of two chains. The protein structures were solvated in order to identify interactions mediated through water. The study revealed that the number of indirect interactions was very large which is often not considered in most methods. Another observation was that polarity of interface was found to be much higher as compared to other methods.

The library can be used to carry out large-scale statistical analysis of protein interfaces in order to determine the significance of these indirect interactions in prediction of protein recognition sites. The data obtained by this study can further be used in training machine-learning based approaches for protein docking.

## Redundancy and asymmetric divergence of paralogs from genome-wide analysis of genetic interactions

Benjamin VanderSluis<sup>1</sup>, Jeremy Bellay<sup>1</sup>, Gabriel Musso<sup>2</sup>, Balazs Papp<sup>3</sup>,  
Anastasia Baryshnikova<sup>2</sup>, Michael Costanzo<sup>2</sup>, Charles Boone<sup>2</sup>, Chad L.  
Myers<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, University of Minnesota - Twin Cities, Minneapolis, Minnesota, USA; <sup>2</sup>Department of Molecular Genetics, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, ON, Canada, <sup>3</sup>Institute of Biochemistry, Biological Research Center of the Hungary Academy of Sciences, Szeged, Hungary

Duplication events resulting from errors occurring during mitosis and meiosis can provide genomes with new genes which can acquire novel functionalities through subsequent selection and divergence. However, studies have shown that instead of complete functional dispersal, many duplicate pairs retain a significant amount of functional overlap over long stretches of evolutionary time. As complete redundancy between duplicates can not stably be maintained from an evolutionary standpoint, elucidation of the advantages of retained function is key to understanding how novel functions arise within the cell.

In an effort to explore the overall contribution of redundancy among extant duplicates to genetic robustness, recent studies have attempted to characterize the functional similarity of paralogs using physical and genetic interaction networks. Genetic interactions involve the perturbation of multiple genes simultaneously in search of a surprising phenotype, and hence suggest clues about the unique and/or shared functions of duplicates providing a snapshot of their divergent evolution. However, the lack of a comprehensive set of interactions has limited previous comparisons, potentially producing biased or misleading findings.

Here we present an analysis of duplicated genes based on the largest genetic interaction network collected to date, derived from Synthetic Genetic Array (SGA) screening conducted in budding yeast *Saccharomyces cerevisiae*. Using whole-genome genetic interaction profiles for over 1700 yeast genes (~5.4 million double mutants), we explored interactions involving paralogs from both whole-genome and small-scale duplication events, and describe a unified framework with which to describe duplicate retention and its consequences. Specifically, the previous observation that duplicate pairs often have poor genetic interaction profile similarity despite obviously shared functionality can be explained by phenotypic masking on behalf of a redundant partner. Further, we present direct evidence for this model and show that whole-genome genetic interactions can be used to dissect both redundant and divergent paralog functions.

In addition to identifying functional overlap among duplicates, these genome-wide genetic interaction data allowed us to confidently address the asymmetric nature of duplicate divergence. We show that genetic interactions can be used to quantify selective pressure and predict rates of sequence evolution, demonstrating that one member of a paralog pair retains the majority of the ancestral functions in most cases. The evidence presented here demonstrates that genome-wide genetic interactions studies are particularly well-suited to address central questions in evolutionary biology, and we expect this technology to be pivotal in future studies on the evolution of gene duplicates.

## Nucleosomes are positioned in exons and have histone marks suggesting co-transcriptional splicing

Claes Wadelius<sup>1</sup>, Robin Andersson<sup>2</sup>, Stefan Enroth<sup>2</sup>, Alvaro Rada-Iglesias<sup>2</sup>, Francisco de la Vega<sup>3</sup>, Kevin McKernan<sup>4</sup>, Jan Komorowski<sup>2,5</sup>

<sup>1</sup> Dept Gen & Pathology, Uppsala Univ, Uppsala, Sweden; <sup>2</sup> Linnaeus Centre for Bioinformatics, Uppsala University, Sweden; <sup>3</sup> Life Technologies, Foster City, CA.; <sup>4</sup> Life Technologies, Beverly, MA; <sup>5</sup> Interdisciplinary Centre for Mathematical and Computer Modelling, Warsaw University, Poland

It is known that nucleosomes are well positioned over the first exon in active genes, with histone modifications reflecting the transcription rate. So far positioning of nucleosomes relative to other genomic features has not been determined and which histone modifications are located along a gene has only been partly analyzed.

We reanalyzed public nucleosome position data for man and *C. elegans*, histone modification data from man and mouse as well as gene and exon expression data from man to look for distinct patterns also in internal exons. We found one well-positioned nucleosome at internal exons with a signal clearly higher than at the TSS. The peaks are centered at +94 (human) and +101 (*C. elegans*) relative to the exon start meaning that the average 5' end of a nucleosome is positioned at +20 and +27 in man and worm, respectively. We found no positioned nucleosome in exons <50 bp, which comprise <5% of human exons, but in long exons there is a nucleosome positioned at the start and end. Nucleosomes were positioned at internal exons regardless of transcription level, in contrast to the situation at the TSS.

We systematically screened 38 histone modifications to see if the nucleosomes had distinct patterns related to gene and exon expression. The H3K36me3 signal was significantly higher in exons than in the following introns in highly expressed human and mouse genes,  $p < 10e-4$  for each exon-intron comparison. This applies from the third exon and onwards. On the other hand H3K27me2 and me3 were associated with gene silencing. In highly expressed genes, high exon usage was associated with high H3K36me3 and low H3K27me2 and the opposite was found in exons with low usage. These data suggest that H3K36me3 might facilitate exon inclusion during co-transcriptional splicing and that splicing is under epigenetic control. We have generated extremely deep data sets by sequencing nucleosomes, RNA and ChIP-DNA from HepG2 cells to further evaluate the findings. As expected we find that nucleosomes are positioned in exons also in these cells. Furthermore, we have mapped >5 million 50 bp reads to splice junctions and find numerous genes with alternative splice isoforms. We are currently analyzing how these events are related to histone modifications and signals from other nuclear proteins. Our results show that exons are functional units not only defined by their coding capacity but also by the way they are packaged in nucleosomes. The factors controlling nucleosome positioning at internal exons must be under strong evolutionary constraint given the strikingly similar pattern in man and worm, with a common ancestor around 1 billion years ago.

# Microbial interaction web inference using metagenomic data

James Robert White<sup>1,2</sup>, Mihai Pop<sup>1,3</sup>

<sup>1</sup>Center for Bioinformatics and Computational Biology, University of Maryland – College Park, College Park, MD, 20742; <sup>2</sup>Applied Mathematics and Scientific Computation Program, University of Maryland – College Park; <sup>3</sup>Department of Computer Science, University of Maryland – College Park.

The central aim of metagenomics is to describe the structure and function of microbial environments using genomic data. Metagenomics projects can now generate millions of DNA sequences, providing insight into taxonomic diversity and structure, as well as gene content. 16S rRNA gene sequencing plays a key role in characterizing microbe populations throughout nature, and large-scale clinical studies are beginning to employ this technique to quantify the extent of taxonomic variation in human-associated microbial communities (e.g. skin and gut). To increase our understanding of microbial dynamics, spatial and temporal studies are required, as the current “snapshots” of microbial communities represent a static view of a potentially changing system. Longitudinal metagenomics studies will allow researchers to see in a new dimension where oscillations and community instability may reflect infections and other human diseases. Here we examine several techniques for inferring interaction networks and overall microbial dynamics from time-series metagenomic data (e.g. 16S rRNA sequences). We then present a systematic methodology for reliably predicting microbial interactions, and finally apply our methods to longitudinal 16S sequence datasets following the intestinal tract of mice on a prototypic Western (high fat/sugar) diet, revealing major interactions amongst bacteria inhabiting the distal gut. Our work illustrates the potential information mathematical modeling can contribute to the field of metagenomics.

## Global Analysis of Human Protein-DNA Interactions for Annotated and Unconventional DNA-Binding Proteins

Zhi Xie,<sup>2,#</sup> Shaohui Hu,<sup>1,4,#</sup> Akishi Onishi,<sup>3,4</sup> Xueping Yu,<sup>2</sup> Lizhi Jiang,<sup>3,4</sup> Jimmy Lin,<sup>5</sup> Hee-sool Rho,<sup>1,4</sup> Crystal Woodard,<sup>1,4</sup> Hong Wang,<sup>3,4</sup> Jun-Seop Jeong,<sup>1,4</sup> Shunyou Long,<sup>4</sup> Xiaofei He,<sup>1,4</sup> Herschel Wade<sup>6</sup>, Seth Blackshaw,<sup>3,4,\*</sup> Jiang Qian,<sup>2,\*</sup> Heng Zhu<sup>1,4,\*</sup>

<sup>1</sup>*Department of Pharmacology & Molecular Sciences,*

<sup>2</sup>*Department of Ophthalmology*

<sup>3</sup>*Department of Neuroscience,*

<sup>4</sup>*The HiT Center,*

<sup>5</sup>*Department of Cellular and Molecular Medicine,*

<sup>6</sup>*Department of Biophysics and Biophysical Chemistry,*

*Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA*

Protein-DNA interactions (PDIs) mediate a broad range of functions essential for cellular differentiation, function, and survival. However, it is still a daunting task to comprehensively identify and profile sequence-specific PDIs in complex genomes. Here, we have used a combined bioinformatics and protein microarray-based strategy to systematically characterize the human protein-DNA interactome. We identified 17,718 PDIs between 460 DNA motifs predicted to regulate transcription and 4,191 human proteins of various functional classes. Among them, we recovered many known PDIs for transcription factors (TFs). We identified a large number of unanticipated PDIs for known TFs, as well as for previously uncharacterized TFs. Analysis of PDIs for these TFs revealed a complex landscape of DNA binding specificities in TF families. Surprisingly, we also found that over three hundred unconventional DNA-binding proteins (uDBPs) -- which include RNA binding proteins, mitochondrial proteins, and protein kinases -- showed sequence-specific PDIs. A number of newly identified PDIs have also been confirmed both *in vitro* and *in vivo*. Furthermore, one *in-depth* study of uDBPs, MAPK1, using combined *in silico*, *in vitro* and *in vivo* approaches, has revealed that MAPK1 acts as a transcriptional repressor of interferon-gamma response genes in human cells, suggesting an important biological role for such proteins.

# Association of genetic features with pathways using multiple high-throughput data

Antti Ylipää<sup>1</sup>, Matti Nykter<sup>1</sup>, Olli Yli-Harja<sup>1</sup>, Wei Zhang<sup>2</sup>, Ilya Shmulevich<sup>3</sup>

<sup>1</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>2</sup>Department of Pathology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA; <sup>3</sup>Institute for Systems Biology, Seattle, WA, USA.

The advent of comprehensive databases of high-throughput genomic data is enabling a paradigm shift from concentrating on individual samples to jointly analyzing large sets of data. Instead of particular information gained by studying one sample, analyzing data from a collection of similar samples can yield deeper understanding. This, in turn, can lead to a broader and more reliable view of the disease or condition. [1]

Towards this end, we present a generic method that allows associating genetic features, like gene expression, microRNA expression, gene copy numbers, etc. to pre-defined sets of genes, such as pathways, through a series of high-throughput measurements. We demonstrate the flexibility of this framework and the advantages of jointly analyzing an ensemble of microarray data to gain new knowledge on glioblastoma multiforme (GBM), the most common brain tumor in humans.

The framework takes high-throughput data, e.g. gene expression profiles, and pre-defined sets of genes, e.g. pathways, as input. First, gene set enrichments [2] are computed for each gene set under each experimental condition (array). Thus, we obtain an  $[m \times p]$  table of  $p$ -values, where  $m$  is the number of gene sets and  $p$  is the number of conditions. We also have a  $[p \times n]$  table containing the measured values for  $n$  genetic features to be correlated with the gene sets. Subsequently, these data may further be transformed into  $-\log$  transformed  $p$ -values, e.g. by statistical testing for differential expression. Then, we compute a correlation  $p$ -value table  $[m \times n]$  for each genetic feature-gene set combination. The output may be represented as a bipartite graph with genetic features as nodes on the left side and the gene sets as nodes on the right side. Edges connect the genetic features to gene sets with the corresponding correlation  $p$ -value as an association score.

We present the method in the context of two example analyses. First, we show how microRNAs are differently associated with pathways in normal brain tissue and GBM tissue. Second, we identify several genes whose differential expression correlates with significant pathway enrichments in GBM using a set of more than 300 gene expression arrays from the Cancer Genome Atlas and 200 pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG).

From the GBM expression data, we found several potential gene-pathway associations that were previously unknown. Also, we showed that we are able to separate normal brain tissue from GBM samples by using microRNA-pathway association profiles. These examples illustrate the versatility and applicability of the developed framework.

[1] The Cancer Genome Atlas Research Network 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, vol. 455, pp. 1061-1068.

[2] Subramanian A, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A*, vol. 102, pp. 15545-15550.

# Computational Modeling of Crosstalk in Cancer Signaling Networks

Jie Zheng<sup>1</sup>, Rafal Zielinski<sup>4</sup>, Pawel F Przytycki<sup>2</sup>, David Zhang<sup>3</sup>, Jacek Capala<sup>4</sup>  
and Teresa M Przytycka<sup>1</sup>

<sup>1</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; <sup>2</sup>Columbia College Columbia University New York, NY, USA; <sup>3</sup>Department of Electrical & Computer Engineering, University of Maryland, College Park, MD, USA; <sup>4</sup>National Cancer Institute National Institutes of Health, Bethesda, MD, USA

Cellular signaling pathways interact in a number of ways forming sophisticated signaling networks. The lack of reaction coefficients necessary for detailed modeling of signal propagation raises the question whether simple parameter-free models could provide useful information about such pathways. We study the combined signaling network of three major pro-survival signaling pathways: Epidermal Growth Factor Receptor (EGFR), Insulin-like Growth Factor-1 Receptor (IGF-1R), and Insulin Receptor (IR). Our study involves static analysis and dynamic modeling of this network, as well as an experimental verification of the model by measuring the response of selected signaling molecules to differential stimulation of receptors.

To simulate the dynamics of signaling transduction, we developed an open source software tool, SimBoolNet, based on Boolean network model combined with stochastic propagation of the signal. SimBoolNet can visualize the dynamic changes of the network with color animation and time series; it also generates scatter plots with regression lines for inference of the causal relation between molecules. Most (although not all) trends suggested by SimBoolNet have been confirmed by experiments. It can be used to facilitate experimental studies of cellular signaling pathways in other studies.

---

# REGULATORY GENOMICS - SYSTEMS BIOLOGY – DREAM4

---

MIT / Broad Institute, Dec 2-6, 2009

## Booklet Addendum

Session	Presenter	Title	Page
RG1	Laiho	Genetic Control Elements of Beta Cell Autoimmunity	A2
RG1	Schultheiss	Identifying gene regulatory modules with support vector machine kernels	A3
RG1	Djebbari	Refining gene signatures	A4
RG2	Marks	Small RNA regulation	A5
RG2	Rodionov	Comparative genomic of transcriptional regulons in <i>Streptococcus</i>	A6
RG2	Imakaev	Simulating chromatin dynamics in a fractal globule	A14
DR1	Hou	Dynamic modeling for gene network inference	A7
SB1	Tai	Genome-wide expression of the electron transfer genes of <i>Shewanella oneidensis</i>	A8
SB1	Gujral	Investigation of the Dynamics of Wnt Signaling	A9
SB1	O'Callaghan	Cellular processes underpinning recombinant monoclonal antibody production by mammalian cells	A10
SB2	Wolf-Yadlin	Elucidation of Cellular Signaling Networks Downstream of Receptor Tyrosine Kinases	A11
SB2	Huttenhower	Orthology-based functional transfer in microbial communities	A12
SB2	Thiesen	Computational analysis of KRAB ZNF protein expression signatures	A13
SB2	Sevecka	Analysis of signal transduction networks using lysate microarray technology	A15

# A Genome-Wide Association Scan for Identification of Genetic Control Elements of Beta Cell Autoimmunity Using Disease Subphenotypes as Outcomes

Asta Laiho<sup>1</sup>, Attila Gyenesei<sup>1</sup>, Kati Lipponen<sup>2</sup>, Inga Pukonen<sup>2</sup>, Riitta Lahesmaa<sup>1</sup>, Mikael Knip<sup>3</sup>, Olli Simell<sup>4</sup>, Jorma Ilonen<sup>2,5</sup>, Robert Hermann<sup>2</sup>.

<sup>1</sup>Turku Centre for Biotechnology, University of Turku and ÅBO Akademi University, Finland; <sup>2</sup>Immunogenetics Laboratory, University of Turku, Finland; <sup>3</sup>Hospital for Children and Adolescent's, University of Helsinki, Finland; <sup>4</sup>Department of Paediatrics, University of Turku, Finland; <sup>5</sup>Department of Clinical Microbiology, University of Kuopio, Finland.

Type 1 (T1D) diabetes is caused by immune mediated destruction of the insulin producing beta cells. Recent whole genome scans using large case-control collections have identified more than 10 new type 1 diabetes susceptibility loci. The effect size of all these new loci is in the range of OR 1.05-1.2, therefore the significance of these single locus associations in understanding disease mechanism or in the prediction of disease risk is unclear. The major limitation of the GWAS studies is the lack of detailed characterization of the phenotypes.

We have shown previously that alternative disease pathways exist in T1D. Patients carrying the HLA DR4-DQ8 haplotype and the insulin gene (11p15.5) disease susceptibility Class I variable number of tandem repeat alleles are much more prone to develop insulin autoantibodies than those with HLA DR3-DQ2 haplotypes. Therefore, to understand the contribution of various susceptibility loci to disease pathways it is important to carry out a GWAS on subjects at different stages of disease development.

In the current study we choose a new study design to enhance statistical power of the GWAS approach. We used subphenotypes of type 1 diabetes characterized by various autoantibody marker combinations during emergence of autoimmunity. The study material comprised a Finnish birth cohort from the Diabetes Prediction and Prevention Study (DIPP) and molecular markers measured during a 10-year follow-up period were used to stratify children according disease stages and subphenotype classes. One hundred thirty subjects carrying either HLA DR4-DQ8/non-DR3-DQ2, or DR4-DQ8/DR3-DQ2 genotypes were genotyped using Affymetrix SNP arrays. Controls who did not develop signs of autoimmunity were used to fully match for HLA Class II genotype, ethnic and geographic origin from the same birth cohort.

In the DR4-DQ8/DR3-DQ2 positive cohort two genomic regions that reached genome wide significance have been detected, while in the DR4-DQ8/non-DR3-DQ2 positive cohort 5 susceptibility loci were identified. These loci were not detected in the other genotype groups, indicating pathway specific phenomena. Pathway analysis was also applied to identify the known biological interactions between the causative genes.

In conclusion, the strategy of selecting disease subphenotypes according to known susceptibility genes and associated disease biomarkers enables detection of additional disease loci in a much smaller sample size than traditional GWA studies.

# Identifying Gene Regulatory Modules with Support Vector Machine Kernels

Sebastian J. Schultheiss<sup>1,2</sup>, Wolfgang Busch<sup>1</sup>, Jan U. Lohmann<sup>1,3</sup>, Oliver Kohlbacher<sup>4</sup>, Gunnar Rättsch<sup>1</sup>

<sup>1</sup>Friedrich Miescher Laboratory of the Max Planck Society, Machine Learning in Biology Research Group, Tübingen, Germany; <sup>2</sup>Max Planck Institute for Developmental Biology, Tübingen, Germany; <sup>3</sup>University of Heidelberg, Department of Stem Cell Research, Heidelberg, Germany; <sup>4</sup>University of Tübingen, Wilhelm Schickard Institute for Computer Science, Tübingen, Germany.

We predict transcription factor (TF) target genes based on their promoter sequence. A TF binding site is a short segment (~10 bp) near a gene's regulatory region that is recognized by respective TFs. Overrepresented motifs can be identified in regulatory sequences of a set of genes that is enriched with targets for a specific TF. Gibbs-sampling methods that try to identify position weight matrices to characterize binding sites have been successful for small genomes, but are problematic in higher eukaryotes, where motifs are degenerate and form cis-regulatory modules.

Our method classifies genes as TF targets. We use *de novo* motif finding and subsequently apply a Support Vector Machine employing a kernel that captures information about the motifs, their relative location, and sequence conservation. The weighted degree kernel with shifts (WDS) computes similarity of fixed-length sequences. We extend this kernel with conservation information and information about motif co-occurrence to the Regulatory Modules (RM) kernel.

A software implementation is available, integrated into our Galaxy server [<http://galaxy.tuebingen.mpg.de>] or for download, under the name KIRMES [1]: kernel-based identification of regulatory modules in euchromatic sequences. This is a two-step process. A motif finder is applied to the user input of regulatory sequences. Around the best-matching motifs in every sequence, we excise 20 base pairs around the center. This sequence window, pairwise distances of the motifs and sequence conservation information are integrated into the RM kernel, effectively concatenating feature spaces.

Using positional oligomer importance matrices, we are able to make the output of the kernel interpretable. It returns a ranked list of sequence logos of the oligomers that contributed most to the correct classification.

We compared our method to a state-of-the-art Gibbs sampler, PRIORITY [1], on its own dataset with the published settings with respect to success classification. We achieve correct predictions on 74% of their sets vs. 63% for PRIORITY. We let KIRMES classify gene sets obtained from microarrays of Arabidopsis. Using conservation as weighting for the WDS kernel improves performance. These results illustrate the power of our approach in exploiting the relationship between motifs as well as conservation to improve the recognition of TF targets. Interpretable results and an easy-to-use web service make this a valuable tool for any researcher interested in regulation.

[1] Schultheiss SJ, Busch W, Lohmann JU, Kohlbacher O, and Rättsch G (2009) KIRMES: kernel-based identification of regulatory modules in euchromatic sequences. *Bioinformatics* 25(16):2126-2133.

## Refining gene signatures: A Bayesian approach

Amira Djebbari<sup>1</sup>, Aurélie Labbe<sup>2</sup>

<sup>1</sup>*Knowledge Discovery Group, Institute for Information Technology, National Research Council of Canada, 46 Dineen Drive, Fredericton, NB, Canada;* <sup>2</sup>*Department of Epidemiology, Biostatistics and Occupational Health, 1020 Pine Avenue West, Montréal, QC, Canada*

High density arrays evaluate DNA, RNA and protein levels at the genome and proteome scale. These high throughput experiments enable, for example, the classification of gene expression profiles with the potential to help diagnosis, prognosis, to suggest targeted treatment and to predict response to treatment. In high density arrays, the identification of relevant genes for disease classification is complicated by not only the curse of dimensionality but also the highly correlated nature of the array data. We are interested in the question of how many and which genes should be selected for a disease class prediction.

Our work consists of a Bayesian supervised statistical learning approach to refine gene signatures with a regularization which penalizes for the correlation between the variables selected. We performed an extensive simulation study where we simulated variables to predict the class and showed that we can most often recover the correct subset of genes that predict the class as compared to other methods, even when accuracy and subset size remain the same. On real microarray datasets (breast cancer, medulloblastoma and metastases datasets), we show that our approach can refine gene signatures to obtain either the same or better predictive performance than other existing methods with a smaller number of genes.

Our novel Bayesian approach includes a prior which penalizes highly correlated features in model selection and is able to extract key genes in the highly correlated context of microarray data. Our methodology is described in the context of microarray data, but can be applied to any array data (such as microRNA, for example) as a first step towards predictive modeling of cancer pathways.

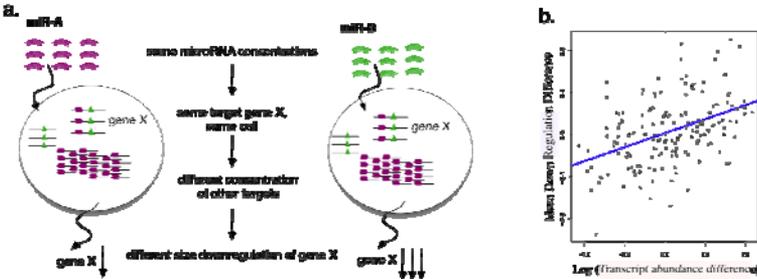
# Small RNA regulation is driven by target concentration and determined by more than binding site efficiency

Aaron Arvey<sup>1</sup>, Christina Leslie<sup>1</sup>, Debra S. Marks<sup>2</sup>

<sup>1</sup>Computational Biology Program, MSKCC, New York, NY; <sup>2</sup>Harvard Medical School, Boston, MA

According to basic chemical kinetics, regulation of mRNAs by small RNAs depends on the concentration of the partners therefore the concentration of target sites for microRNAs or siRNAs should determine both the specificity and sensitivity of such regulation. However, while small RNA ‘dosage’ has been extensively explored in many experiments, much less is known about the effects of the natural variation in the cell of concentration of target sites for each microRNA or siRNA. Our hypothesis is that the amount of competition between target sites (in mRNAs) for a limited number of active small RNAs (in RISC) should determine how much a small RNA can down-regulate each of its target mRNAs (Figure 1a)

**Figure 1.**



To test this hypothesis we considered all genes targeted by two different transfected microRNAs in the same cell type in ~50 experiments and whether the total concentration of all targets for a microRNA in a cell affects the amount of down-regulation of an individual target. Specifically, for each set of genes targeted by a pair of microRNAs, miR-A and miR-B, we compared the difference in mean log expression change in the transfections of miR-A and miR-B and the difference in total target abundance for these two microRNAs (estimated from RNA-seq data). Consistent with our hypothesis, we found a significant correlation between these two differences ( $p < 1e-3$ , empirical  $p$ -value, Figure 1b), that is, greater target abundance correlates with smaller downregulation. Similarly, we also counted the number of target sites per miRNA and the number of mRNA transcripts (using RNA-seq), and discovered a correlation between the number of mRNA targets in the cell and the mean amount of down-regulation for all the predicted targets of the microRNA (Spearman correlation 0.58)

Our results support the idea that mRNA target abundance and competition for miRNAs is a global phenomenon. Specifically, better understanding target concentration will shed insight into several critical problems to the community, including better understanding of (i) the varying levels of regulation by miRNAs, (ii) the effects of miRNA transfection studies, and (iii) designing siRNA screens and siRNA therapeutics

# Comparative genomic reconstruction of transcriptional regulons in the *Streptococcus* genus

Dmitry A. Rodionov<sup>1,2</sup>, Marat D. Kazanov<sup>1,2</sup>, Irina A. Rodionova<sup>2</sup>, Pavel S. Novichkov<sup>3</sup>, Elena S. Novichkova<sup>3</sup>, Ramy K. Aziz<sup>4</sup>, and Andrei L. Osterman<sup>2</sup>

<sup>1</sup>*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia;* <sup>2</sup>*Burnham Institute for Medical Research, La Jolla, California;* <sup>3</sup>*Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California;* <sup>4</sup>*Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt.*

Integrative comparative genomics approaches were used to infer transcriptional regulatory networks (TRNs) in *Streptococcus* species with sequenced genomes. To accomplish this goal, we combined the identification of transcription factors (TFs), TF-binding sites (TFBSs) and cross-genome comparison of regulons with the analysis of genomic and functional context inferred by metabolic reconstruction. A limited input of established regulon members was provided by publications on a particular TF in individual species. The reconstructed TRNs for the key pathways involved in central metabolism, production of energy and biomass, metal homeostasis, stress response and virulence provide a framework for the interpretation of gene expression data.

A genomic census of TFs within the group of analyzed genomes also allows estimating the scale of TRN in these species, the expected number of conserved (core) regulons and the degree of variations between individual species. The total number of putative TFs varies substantially within the group of *Streptococcus* (from 60 to 126) suggesting substantial variations in their TRNs. To a large extent these variations are associated with prophage and virulence associated regulators. First, we focused on the reconstruction of core TF regulons conserved in 8 analyzed genomes of various *Streptococcus* species, including global catabolic regulon CcpA and other regulons controlling the core metabolism, stress response and virulence. Using comparative genomics approach we identified candidate TFBSs for 38 TFs from the *Streptococcus* group. Two major diversification strategies were defined: constrained, when the regulon is either present or absent in its entirety with tightly conserved regulation of all genes, and permissive, when most genes of a regulon are conserved between genomes, whereas the conservation of respective regulatory sites is much weaker and sometimes not mandatory. Second, we analyzed the *Streptococcus* genomes to identify the RNA regulatory elements, such as known types of riboswitches and T-boxes, and found that the latter is the most widespread regulatory RNA in streptococci.

The results of this analysis for 8 *Streptococcus* genomes were captured within the RegPrecise database (<http://regprecise.lbl.gov>). We developed this novel database for capturing, visualization and analysis of predicted transcription factor regulons in prokaryotes that were reconstructed and manually curated by utilizing the comparative genomic approach. Finally we built an accurate procedure of regulon propagation and applied it to the reconstructed regulons in *Streptococcus*. This automatic propagation allows one to estimate the number of conserved regulatory interactions (TF, target genes, TFBSs) in other closely-related genomes.

# Discrete and continuous, comparative and reconstruction-based dynamic modeling for gene network inference

Ping Hou<sup>1</sup>, Zhengyu Ouyang<sup>1</sup>, Yang Zhang<sup>1</sup>, Haizhou Wang<sup>1</sup>, Joe Song<sup>1</sup>

<sup>1</sup>*Department of Computer Science, New Mexico State University*

Genome-wide expression time course and perturbation data sets allow one to learn the structure and kinetics of gene regulatory networks (GRNs) which have largely remain unknown except for a few well-studied prokaryotes. The Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges create a platform to evaluate biological network reverse engineering algorithms. We integrate discrete and continuous, comparative and reconstruction-based dynamic modeling into gene network inference and demonstrate it through DREAM4 In Silico Network Challenge 2. The Challenge provided knockout, knockdown, time course and multifactorial experimental data. The perturbed time courses enable us to apply a novel comparative modeling method that we have developed to detect gene interactions. The new method can report interaction shifts between perturbed time courses and normal time courses. Traditional reconstruction-based methods have limited statistical power especially when we have limited sample size. Using comparative modeling, one can detect subtle changes under different experimental conditions even when the data size is small. While discrete generalized logical network (GLN) is better for explaining qualitative dynamics, the continuous dynamical system model (DSM) has an advantage of capturing quantitative dynamics. Both models have the ability to produce prediction under various conditions, such as double knockouts.

First, we used comparative modeling for GLN and DSM to take advantage of the time series data given under different conditions. Discrete dynamic is suitable for GLN comparative modeling based on the provided two halves of time series. Similarly we compare different DSM models reconstructed based on the two half time series to capture continuous dynamic. We separated time series data to two categories “with perturbation” and “without perturbation”, then use both to reconstruct the models and compare with each other to find conserved interactions. We assume that the conserved ones are the inherent in the original networks, but the differential ones are caused by perturbations. We extract conserved interactions as regulatory relationship candidates for further analysis.

Next, we deal with the knockout, knockdown, and wild-type steady state data. We applied GLN model reconstruction with zero<sup>th</sup> Markov order. In data of knockout and knockdowns, when a particular gene is inhibited, expression levels of some genes would be extremely high or low in the inhibited steady state compared to the normal steady state. We assume that when the parent gene has been knocked out or down, the expression level of child gene will change accordingly. Based on this assumption, we combine the three types of steady state data for knockouts, knockdowns and wild type. Then we built GLN models from the steady state data we combined to find the candidate regulatory relationships for future modeling.

Lastly, we recovered GRNs by reconstructing a DSM made of ordinary differential equations. In the DSM reconstruction, we use the regulatory relationship candidates from the previous two steps, to reduce the search space and focus on more credible topology of GRNs. After this step, we recovered differential equations that fit reasonably well to the data in Challenge 2 of DREAM4.

# Genome-wide expression of the electron transfer genes of *Shewanella oneidensis* reveals transcriptive association with chemotaxis

Shang-Kai Tai<sup>1</sup>, Shinsheng Yuan<sup>1</sup>, Ker-Chau Li<sup>1,2</sup>

<sup>1</sup>Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan, R.O.C.; <sup>2</sup>Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA.

*Shewanella oneidensis* has the ability to produce current in microbial fuel cells. *OmcA*, *omcB* (also known as *mtrC*), *mtrA*, *mtrB*, and *gspF* are some known genes of *S. oneidensis* MR-1 that participate in electron-transfer pathways. How do the bacteria coordinate the expression of these genes? To shed light on this problem, we obtain the gene expression datasets of MR-1 that are recently public-accessible in NCBI Gene Expression Omnibus. We utilize the novel statistical method, liquid association (LA), to investigate the complex pattern of gene regulation. Through a web of information obtained by our data analysis, a network of transcriptional regulatory relationship between chemotaxis and electron transfer pathways is revealed, highlighting the important roles of the chemotaxis gene *cheA-1*, a triheme *c*-type cytochrome gene SO4572, and the magnesium transporter gene *mgtE-1*.

# A System-wide Investigation of the Dynamics of Wnt Signaling Identifies Novel Phases of Transcriptional Regulation

Taranjit S. Gujral and Gavin MacBeath

*Department of Chemistry & Chemical Biology, Harvard University, Cambridge, MA 02138,*

The American Cancer Society estimates that there will be about 150,000 new cases of colorectal cancer (CRC), causing 50,000 deaths in 2009 in the United States. The most common cause of CRCs (> 90%) is an activating mutation of the canonical Wnt signaling pathway. Several components of this signaling pathway starting from ligand binding at the cell surface to subsequent changes in gene transcription have now been identified. However, a system-wide analysis of the dynamics of Wnt signaling has yet to be performed. Here, we provide the first, quantitative system-wide analysis of over one hundred pathway-wide signaling events in response to a time course Wnt3a stimulation using quantitative western blotting and qPCR arrays. Our analyses of Wnt signaling have identified previously unrecognized signaling trends and nodes of regulation. A broad, quantitative, and dynamic study of Wnt3a stimulation revealed two phases of transcriptional regulation: an early phase in which signaling antagonists were downregulated, providing positive feedback, and a later phase in which many of these same antagonists were upregulated, attenuating signaling. The dynamic expression profiles of several response genes, including *MYC* and *CTBP1*, correlated significantly with proliferation and migration ( $P < 0.05$ ). Additionally, their levels tracked with the tumorigenicity of colon cancer cell lines (Caco-2, DLD-1, HT-29, HCT116 and SW480) and they were significantly overexpressed in colorectal adenocarcinomas ( $P < 0.05$ ). Taken together, our systematic and multi-dimensional approach has identified global signaling trends and key proteins in the Wnt signaling pathway that have functional roles in tumor progression. Selective targeting of these proteins may provide a novel strategy for developing new therapeutics for the treatment of colorectal cancer.

# Modeling Cellular Processes Underpinning Recombinant Monoclonal Antibody Production by Mammalian Cells

O'Callaghan, P.M., McLeod, J., Pybus, L., Wilkinson, S.J., James, D.C.

*Department of Chemical and Process Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK*

The increasing demand for recombinant therapeutic monoclonal antibodies (Mabs) in the clinic has placed significant pressure on the biopharmaceutical industry to develop high yielding mammalian cell-based production systems. It is anticipated that targeted engineering of recombinant mammalian cell lines to increase cell-specific Mab productivity (qMab) could help achieve the necessary increases in Mab yield from production processes to meet this demand. However, to design rational engineering strategies to increase qMab we require a significantly improved understanding of the control of flux through the complex Mab biosynthetic pathway from the recombinant heavy chain (HC) and light chain (LC) genes through to the fully-formed and secreted Mab protein (HC<sub>2</sub>LC<sub>2</sub>). Without this knowledge cell line engineering strategies continue to prove intractable.

In answer to this problem we present the first empirically-derived mathematical model of recombinant IgG<sub>4</sub> Mab synthesis using a panel of 8 Chinese Hamster Ovary cell lines varying in qMab as a model. Utilizing dynamic measurements of recombinant Mab mRNA and polypeptide intermediate pools in exponential phase cell cultures together with experimentally-determined rates of Mab intermediate degradation and processing, we have developed a comprehensive mathematical model of Mab synthesis that we have used to understand the cellular constraints on productivity and how this varies with qMab.

Our model was built and solved using Sentero, an in-house interactive modeling program which utilizes MATLAB as a simulation engine. The model combines an unstructured description of cell growth, death, nutrient uptake and metabolite production with a structured description of Mab synthesis and secretion. The structure of our hybrid model differs substantially from previously published models in which IgG<sub>4</sub> Mab synthesis proceeds via a "half Mab" (HCLC) synthesis intermediate. Based on our new experimental analysis of the *in vivo* kinetics of Mab assembly we have built our model around an alternative Mab assembly pathway in which IgG<sub>4</sub> Mab synthesis proceeds via a HC dimer intermediate (HC<sub>2</sub>) to which LC is sequentially added to form full Mab (HC<sub>2</sub>LC<sub>2</sub>). This new model predicts our empirical data with a high degree of accuracy.

The results of our mathematical modeling analysis of Mab synthesis within our panel of cell lines shows that the specific qMab constraints differ between cell lines with varying qMab, although across the panel as a whole HC transcription rate, HC mRNA stability and HC translation rate were identified as the biosynthetic parameters that exert the most control over flux through the pathway. A key advantage of our mathematical model is the ability to make *in silico* predictions of the effect of targeted cell engineering on qMab. Using single parameter sensitivity analysis to investigate the effect of increasing HC mRNA abundance and HC translation rate suggests that such engineering efforts could achieve significant increases in qMab. However, the impact of such strategies varies substantially between cell lines with different qMab. Finally, we show how our mathematical modeling approach can be used to design cell engineering strategies for generating "super-producing" cell lines with high qMab.

# High Throughput Studies for the Elucidation of Cellular Signaling Networks Downstream of Receptor Tyrosine Kinases

Alejandro Wolf Yadlin<sup>1</sup>, Mark Sevecka<sup>1,2</sup>, Taranjit Gujral<sup>1</sup>, Gavin MacBeath<sup>1</sup>

<sup>1</sup>Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA

<sup>2</sup>Currently at Whitehead Institute, MIT, Cambridge MA

Ligand binding to cell surface receptor tyrosine kinases (RTKs) initiates a cascade of signaling events regulated by dynamic phosphorylation on a multitude of intracellular proteins. It is believed that quantitative features, including intensity, timing, and duration of phosphorylation of particular residues, determine cellular response. Although all RTKs are known to activate similar downstream pathways, the way signaling networks are diversified and differentially stimulated by distinct RTKs is not well understood. Because deregulation of RTK-initiated signaling networks plays an important role in cancer, it is important to understand how different receptors interact with their downstream pathways and how they control diverse phenotypic cellular behaviors such as proliferation and transformation.

To determine at a quantitative level how different RTKs behave when placed in the same cellular background, we selected six well-studied and phylogenetically diverse RTKs: EGFR, FGFR1, IGF1R, MET, NTRK2 and PDGFR- $\beta$ . Six stable cell lines were generated by transfecting the full-length coding region into Flp-In 293 cells, resulting in average expression of ~105 receptors per cell. We used these cell lines in lysate microarray base studies, to screen a library of 400 phospho- and pan-specific antibodies.

Next, we conducted a systematic, data-rich study to elucidate quantitative features in the signaling networks downstream our six receptors of interest using the validated antibodies. We choose 20 signaling nodes to knockdown expression in each cell line using 2 different shRNAi interventions. ShRNAi-treated cells were then challenged with their cognate ligand and 11 point time courses of stimulation were collected. Lysates were then printed onto nitrocellulose slides – each containing over 27000 independent samples – for analysis with the validated antibodies. Currently we are in the process of analyzing the slides, collecting, sorting and modeling the data.

Further, we wanted to understand not only how our six receptors affected cellular signaling *in vitro*, but also how their presence might influence tumorigenicity and signaling events *in vivo*. Flp-In 293 cells lines, stably expressing each of RTKs, were injected subcutaneously into athymic mice and their ability to form tumor outgrowths was monitored. We are utilizing lysate microarray and immunohistochemistry to analyze the downstream signaling events *in vivo*.

Mathematical analysis of these *in vitro* and *in vivo* data will allow us to determine the temporal dynamics and network topology associated with each receptor and slice through the pathways' complexity to identify nodes in the network most likely to be associated with a given phenotypic response.

# Orthology-Based Functional Transfer in Microbial Communities

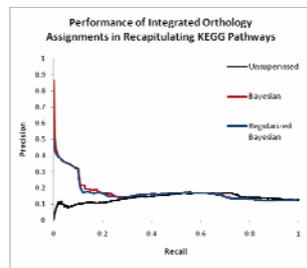
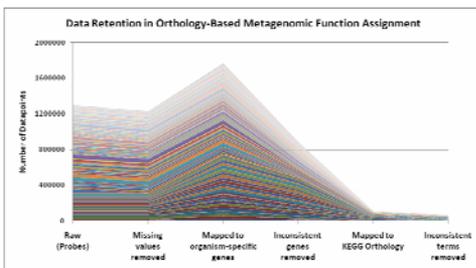
Curtis Huttenhower<sup>1</sup>

<sup>1</sup>*Department of Biostatistics, Harvard School of Public Health, Boston, MA*

Microbial communities are rapidly emerging as one of the most biologically and computationally challenging arenas in which next-generation sequencing can be applied. Human microflora have a significant impact on health and disease, and soil communities directly influence crop growth and sustainable agriculture. Very few of the microbes in these communities are culturable, and based on current metagenomic sequencing - short read sequencing directly on environmental samples - we have only begun to sample the genomic and functional diversity present in even the most common microbial communities.

The computational challenge inherent in this situation is daunting: given a collection of millions of short sequence reads, assemble an interpretable picture of the phylogenetic and functional activity present in the underlying microbial community. Current approaches to the latter portion of this problem rely mainly on orthologous sequence clusters such as COG or the KEGG Orthology to assign biological function. Here, we investigate this methodology by asking how much functional information (in the sense of proteins' biological roles, not chemical activities) is transferrable among microbial species using currently assigned orthology. We retrieved 121 microbial gene expression datasets describing ~75 species from GEO and ArrayExpress, normalized them using established methodology (Huttenhower, 2009), and mapped their genetic content to the KEGG Orthology. We then asked whether the resulting collection of KO-based datasets could recapitulate known functional interactions from KEGG pathways.

The two striking results of this study are, first, that less than 9% of the genetic content of these datasets was mappable to KO terms using very generous criteria. This number falls below 6% when the resulting KO term mappings are required to have consistent expression patterns within each dataset (analogous to multiple probes in a probeset). This is corroborated by recent metagenomic studies (Turnbaugh, 2009) in which ~5% of the sequence content has been mappable to functional terms. Second, while individual datasets are functionally consistent, the integrated organism-independent, orthology based data provides little functional information outside of extremely well-conserved areas (e.g. the ribosome). This finding is robust across multiple data integration methods: unsupervised, Bayesian, logistic regression, and several other machine learning algorithms. This calls for a closer look at current methodology for cross-species functional transfer using sequence orthology and emphasizes the need for novel, network-based methods for investigating metagenomic community function.



# Computational analysis of KRAB ZNF protein expression signatures derived from the ProteinAtlas database

Bjoern Ziem (1), Michael.Kreutzer (2), Ziliang Qian (3), Larisa L. Kiseleva (4), Cristina Al-Khalili Szogyarto (5), Mathias Uhlén(5), Yixue Li, H-J Thiesen (2),

1) *Gesellschaft für Individualisierte Medizin mbH, Rostock, Germany*, 2) *Institute of Immunology, University of Rostock, Rostock, Germany*, 3) *Shanghai Institutes for Biological Sciences (SIBS), Shanghai, China*, 4) *Larisa Kiseleva AIST Research Staff, Sequence Analysis Team, Computational Biology Research Center (CBRC) Tokyo, 135-0064, Japan*, 5) *AlbaNova University Center, Royal Institute of Technology (KTH), Stockholm, Sweden*

**Introduction:** Spatial information on cell-type specific and disease-pathway associated protein expressions related to KRAB ZNF protein functions are a prerequisite to proceed with computational analysis of KRAB ZNF gene functions. Information on protein expressions presented by the ProteinAtlas database were downloaded ([www.proteinatlas.org](http://www.proteinatlas.org)) and semi-quantitative expression values computationally analyzed, partly guided by the visualisation tool [www.toponostics.org](http://www.toponostics.org). Informative antibody repertoires were selected to determine protein expressions that are inversely correlated to the expression of KRAB ZNF genes since Krüppel-associated box (KRAB) zinc finger (ZNF) genes encode the largest mammalian family of proteins with strong transcriptional repressor activities (Margolin et al., PNAS 1994, Deuschle et al., MCB 1995; Lorenz et al, Biol Chem 2001). Semi-quantitative measures limited to four levels were used to calculate Pearson correlation distances of immunostainings of i. normal tissue, ii. cancer tissue and iii. cell types singly in respect to evaluate how expression profiles diverge from each other and to determine tumour expression profiles that are more closely or more distantly correlated with each other as well as to KRAB ZNF protein expressions. Expressions of putative candidate target genes regulated by KRAB ZNF proteins should be inversely correlated in their expression in respect to the expression of KRAB ZNF proteins.

**Methods and Results:** At an initial step, immunostainings on cell lines and cells were quantitated by making use of the Definiens software and compared to ProteinAtlas measures. To validate the quality of monovalent antibodies raised against KRAB ZNF PrEST sequences, the epitope signatures of KRAB ZNF antisera were assessed against peptides of all corresponding ZNF related PrEST as well as against unrelated protein sequences, see Lorenz et al, 2009. The phylogenetic relatedness of KRAB ZNF genes under study were determined by making use of the SysZNF database (see Ding et al., 2009: URL <http://epgd.biosino.org/SysZNF>). Putative candidate target genes inversely regulated in respect to KRAB ZNF protein expressions have been identified from these ProteinAtlas data sets. Regulatory networks directed by KRAB ZNF proteins can be extracted from the ProteinAtlas protein expression database under the assumption that immunostainings can be routinely quantitated in a standardized fashion and cross-/ specificities of monovalent antisera detecting protein families have been intensively validated.

**Summary:** I. Immunostainings should be systematically quantified by imaging software. II. Each polyclonal serum should be characterized by their epitope pattern on PrEST and related peptide sequences. III. KRAB ZNF genes have to be grouped in phylogenetic clusters before analyzing them on the toponome level. IV. Our in-silico analysis demonstrates that the ProteinAtlas database encompasses a unique reservoir of pathway information waiting to become extracted and implemented in computational analyses.

## Simulating chromatin dynamics in a fractal globule

Imakaev M.\*, Lieberman-Aiden E.\*, van Berkum N.L., Lander E.S., Dekker J., and Mirny L.A.

A recent study (Lieberman-Aiden et al., Science 2009) used the novel Hi-C technique to characterize folding of DNA inside a cell. In this study the structure of the genome on the scale of a few megabases was found to be consistent with a fractal globule conformation. The fractal globule is a dense conformation of a polymer (e.g. chromatin fiber) into which a polymer can spontaneously fold.

Here we use computer simulations to examine structural and dynamic properties of DNA folded into a fractal globule and their connection to the process of gene expression. We use Monte Carlo simulations to explore both static and dynamic properties of the fractal globule. This analysis demonstrates that the folded genome is organized into sectors, that compact chromatin domains have extensive contacts with each other, and that such folding facilitates access of freely diffusing transcription factors to proximal genes. We suggest how such organization can lead to the coordinated regulation of gene expression. Furthermore, we examine dynamic properties of the fractal globule essential for gene regulation: opening of chromatin domains, formation of loops between distant regions of the genome, and mobility of individual loci. We demonstrate that the dynamics is facilitated by the lack of knots in the folded DNA -- a unique property of the fractal globule. Biophysical modeling and theoretical analysis of chromatin folding will provide us with new insights into the nature of genomic organization and gene regulation.

## **EGFR and Beyond – Analysis of Signal Transduction Networks Using Lysate Microarray Technology**

Mark Sevecka<sup>1</sup>, Alejandro Wolf-Yadlin<sup>2</sup>, Jennifer Grenier<sup>3</sup>, David Root<sup>3</sup>, Gavin MacBeath<sup>2</sup>

<sup>1</sup>Whitehead Institute for Biomedical Research, Cambridge MA; <sup>2</sup>Department of Chemistry and Chemical Biology, Cambridge MA; <sup>3</sup>Broad Institute of Harvard and MIT, Cambridge MA

Signal transduction networks regulate and coordinate cellular behavior, thereby mediating global tissue responses to outside cues. To enable future advances in basic cell biology, as well as patient therapy, there is a strong need for technologies that can dissect cellular signaling events at a global level. The work presented here centers on the development and application of an emerging technique for quantitative systems biology: lysate microarray technology.

We present a strategy based on microarrays of cell lysates that dissects the global effects of perturbations of signaling components in the epidermal growth factor receptor (EGFR)-induced signaling network. Using carefully validated RNAi reagents, we systematically modulated the abundance of 21 central signaling components. We then determined the time-dependent response of each one of 16 central signaling nodes by lysate microarrays. Our data indicates extensive inter-pathway crosstalk, and suggests a novel network influence linking MAPK and Akt pathways. We also present evidence that phosphatases play a crucial role in mediating unexpected network influences.

Finally, to expand the utility of lysate microarray technology for systems biology applications, we present a screen for functional detection antibodies across multiple cellular contexts. Using a statistical threshold based on the observed assay variability, we identified putative 'hit' combinations of antibodies and cell lines, and further validated hits by quantitative immunoblotting. We found that functional detection antibodies can be identified for diverse protein targets in each cell type, thereby demonstrating the generality of lysate microarray technology. Our data also provide an initial map of signaling events in the different cellular contexts. We anticipate that global quantitative approaches such as this will play an important role in elucidating the complexities of signal transduction networks.

# Index of all invited, oral, poster, and paper presentations

Presenting author	Abstract title (abbrev.)	Page	Type	Conf	Time
Aerts	Regulatory network for retinal differentiation	28	Oral	RG	Thu 5:45p
Agius	Learning compact models of DNA binding specificities for	119	Poster	RG 1	Wed 8:15p
Aho	Gene expression profile of human adipose stem cells cultured in	246	Poster	SB 1	Fri 8:15p
Aho	Reconstruction and Validation of RefRec: a Global Model for the	266	Poster	SB 1	Fri 8:15p
Aid	DNA motif discovery approach adapted to ChIP-chip and ChIP-Seq	267	Poster	SB 1	Fri 8:15p
Äijö	Learning gene regulatory networks with delayed ODEs and	146	Poster	RG 1	Wed 8:15p
Akavia	A Bayesian framework to detect drivers	73	Oral	DR	Sat 3p
Akavia	Conexic: A Bayesian framework to detect drivers and their function	90	Poster	RG 1	Wed 8:15p
Al-Akwaa	SSBNN: Gene Regulatory Network Construction using Spectral	268	Poster	SB 1	Fri 8:15p
Alexopoulos	Identifying drug effects via pathway alterations	42	Paper	DR	Fri 11a
Amzallag	Comparison of gene expression time courses between light	147	Poster	RG 1	Wed 8:15p
Apri	How to analyze the robustness of biological models with oscillatory	223	Poster	SB 1	Fri 8:15p
Arunachalam	Computational discovery of Cis-regulatory elements in multiple	120	Poster	RG 1	Wed 8:15p
Ay	Analysis of Boolean regulatory networks	31	Paper	RG	Thu 6:45p
Ay	Scalable Steady State Analysis of Boolean Biological Regulatory	269	Poster	SB 1	Fri 8:15p
Barkai	Evolution of nucleosome positioning	19	Invited	RG	Thu 1p
Baryshnikova	The Genetic Landscape of a Cell	224	Poster	SB 1	Fri 8:15p
Behrens	Studying the evolution of promoters: a waiting time problem	121	Poster	RG 1	Wed 8:15p
Belcastro	CENTRO: A CoExpression NeTwoRk Omnibus for gene function and	247	Poster	SB 1	Fri 8:15p
Bellay	Decomposition of genetic interaction networks	34	Oral	RG	Thu 7:30p
Benyamini	Metabolic flux balance analysis	54	Paper	DR	Fri 6:30p
Betel	Comprehensive modeling of microRNA targets	57	Oral	DR	Fri 7:30p
		91	Poster	RG 1	Wed 8:15p
Bieler	Modeling 3D Flies	82	Oral	DR	Sat 7:15p
Biggin	Evidence for Quantitative Transcription Networks	1	Invited	RG	Wed 5:15p
		148	Poster	RG 1	Wed 8:15p
Boeke	Building Saccharomyces cerevisiae v2	59	Invited	DR	Sat 9a
Bolotin	Identification of human HNF4 target genes	22	Oral	RG	Thu 2p
Bolotin	Integrated Approach for the Identification of Human HNF4 $\alpha$ Target	122	Poster	RG 1	Wed 8:15p
Bonneau	Comparative analysis of genomics data collections: Multi-species	249	Poster	SB 1	Fri 8:15p
Borenstein	Super-metabolism microbial communities	85	Oral	DR	Sun 9:30a
Bourque	Binding site turnover in stem cells	13	Oral	RG	Thu 10a
Bristow	Exploring the CBP developmental time-course in Drosophila	92	Poster	RG 1	Wed 8:15p
Brodsky	Identification and analysis of regulatory regions	14	Oral	RG	Thu 10:45a
Brynildsen	Metabolic strategies to enhance antibiotics	63	Oral	DR	Sat 10:45a
Bugrim	Role of growth factor signaling network	64	Oral	DR	Sat 11a
Calvo	Widespread translational repression	2	Oral	RG	Wed 5:45p
Camacho	Decoding small RNA networks in bacteria	60	Oral	DR	Sat 9:30a
Candeias	Temporal Dynamics of Regulatory Networks in Drosophila	149	Poster	RG 1	Wed 8:15p
Carmel	A universal nonmonotonic relationship between gene compactness	225	Poster	SB 1	Fri 8:15p
Carson	Investigating Co-regulation Networks Using Generative Models	150	Poster	RG 1	Wed 8:15p
		226	Poster	SB 1	Fri 8:15p
Carvalho	Applications of Centroid Estimation to Regulatory Genomics	123	Poster	RG 1	Wed 8:15p
Chan	A dynamic analysis of IRS-PKR signaling	56	Paper	DR	Fri 7:15p
Chang	The intersect of mRNA, microRNA and protein dynamics upon down	93	Poster	RG 1	Wed 8:15p
Chang	In silico prediction for regulation of transcription factorson	179	Poster	DR 1	Thu 3:45p
Chang	The intersect of mRNA, microRNA and protein dynamics upon down	199	Poster	SB 1	Fri 8:15p
Chen	Empirical mode decomposition for time-series gene expression data	227	Poster	SB 1	Fri 8:15p
Chun	Reverse Engineering of Gene Regulation Network from DREAM4	151	Poster	RG 1	Wed 8:15p
		181	Poster	DR 1	Thu 3:45p
Clark	Characterizing Artificial Chemistries	228	Poster	SB 1	Fri 8:15p
Clarke	A missing ingredient in the Pho4 paradigm? Evidence for Pho4/Cbf1	124	Poster	RG 1	Wed 8:15p
Clote	RNA Structural Segmentation	94	Poster	RG 1	Wed 8:15p
Cook	Using ChIP seq to search for sequence determinants of binding	125	Poster	RG 1	Wed 8:15p
Cosgrove	Accounting for dependency within mRNA expression compendia	182	Poster	DR 1	Thu 3:45p
Culhane	Large scale analysis of stem cell expression	74	Oral	DR	Sat 3:15p
Dabrowski	Effects of motif and CNS multiplicity on gene expression subspaces	126	Poster	RG 1	Wed 8:15p
		200	Poster	SB 1	Fri 8:15p
Dalkic	Distinct topological changes of the different cancer types	250	Poster	SB 1	Fri 8:15p

Davis	Classification Trees Can Describe and Predict Conditional	152	Poster	RG 1	Wed 8:15p
Degner	Effect of read-mapping biases on detecting allele-specific expression	114	Poster	RG 1	Wed 8:15p
		229	Poster	SB 1	Fri 8:15p
Deshpande	A scalable algorithm for discovering conserved active subnetworks	270	Poster	SB 1	Fri 8:15p
Devey	Host factors involved in HCV replication	201	Poster	SB 1	Fri 8:15p
Di Bernardo	Mathematical modeling of RNA interference	230	Poster	SB 1	Fri 8:15p
Ding	Exact calculation of partition function	15	Oral	RG	Thu 11a
Dojer	Identification of cis-regulatory modules in homologous sequences	153	Poster	RG 1	Wed 8:15p
Dotu	Computing folding pathways between RNA secondary structures	231	Poster	SB 1	Fri 8:15p
Ellis	Predikin: Combining Structure and Sequence Information to Predict	183	Poster	DR 1	Thu 3:45p
Erkkilä	Incorporating spatial information of heterogeneous cell populations	251	Poster	SB 1	Fri 8:15p
Ernst	Genome-wide discovery of chromatin states	6	Oral	RG	Wed 7p
Fang	Subspace differential coexpression analysis	75	Oral	DR	Sat 3:30p
Feng	Genome-wide survey of D.melanogaster insulator proteins binding	95	Poster	RG 1	Wed 8:15p
Fontana	Combinatorial complexity in systems biology	84	Invited	DR	Sun 9a
Frogner	Learning recurrent mRNA expression patterns from	154	Poster	RG 1	Wed 8:15p
Gentles	Predicting histological transformation	30	Oral	RG	Thu 6:15p
Gentles	A pluripotency signature predicts histological transformation	202	Poster	SB 1	Fri 8:15p
Gitter	Backup in gene regulatory networks explains differences between	127	Poster	RG 1	Wed 8:15p
Goel	Dynamic Flux Estimation – A novel framework for metabolic	271	Poster	SB 1	Fri 8:15p
Goff	Genomic relationship between small RNAs and histone	96	Poster	RG 1	Wed 8:15p
Gray	Widespread RNA polymerase II recruitment	11	Oral	RG	Thu 9:30a
Greenfield	Inferring Topology and Dynamical Properties of Genome-wide	155	Poster	RG 1	Wed 8:15p
Guan	Sampling Bayesian Network with Fast Mixing MCMC	232	Poster	SB 1	Fri 8:15p
Gyenesei	Functional Inference from a Genome-Wide in situ Hybridization	156	Poster	RG 1	Wed 8:15p
		252	Poster	SB 1	Fri 8:15p
Habib	Aromatase inhibition in a transcriptional network context	157	Poster	RG 1	Wed 8:15p
		203	Poster	SB 1	Fri 8:15p
Haynes	Utilizing Global Constraints in Regulatory Network Inference	184	Poster	DR 1	Thu 3:45p
Hemberg	De novo detection of transcribed regions in mouse based on RNA	97	Poster	RG 1	Wed 8:15p
Hou	Integration of discrete and continuous, comparative and	185	Poster	DR 1	Thu 3:45p
Huggins	Design of multiple hypothesis tests for microarray data	172	Poster	RG 1	Wed 8:15p
Hurley	Combining network inference algorithms reveals insights intocancer	187	Poster	DR 1	Thu 3:45p
Ilyin	Functional annotation with TOPOFIT-DB including non-sequential	233	Poster	SB 1	Fri 8:15p
Iorio	Identifying drug mode of action from gene expression	52	Oral	DR	Fri 5:30p
Jabbari	Novel thermodynamics-based algorithm for probe-specific position	128	Poster	RG 1	Wed 8:15p
Jacobsen	Genes up-regulated after microRNA perturbation have significant	98	Poster	RG 1	Wed 8:15p
Jang	Simulation-based Perturbation Studies: Genome-Wide Cause	234	Poster	SB 1	Fri 8:15p
Ji	Genes as molecular machines: Microarray Evidence for structural	99	Poster	RG 1	Wed 8:15p
Joshi	Diverse aspects of posttranscriptional regulatory network analysis	188	Poster	DR 1	Thu 3:45p
Jothi	A link between dynamics and network architecture	29	Oral	RG	Thu 6p
Jungreis	A Computational Investigation of Widespread Stop Codon	129	Poster	RG 1	Wed 8:15p
Kadri	Evolutionary role of microRNAs in developmental gene regulatory	100	Poster	RG 1	Wed 8:15p
Karlic	Towards a Histone Code for Transcription	101	Poster	RG 1	Wed 8:15p
Kartal	Robustness as an evolutionary design principle	81	Paper	DR	Sat 7p
Kartal	Ground State Robustness as an Evolutionary Design Principlein	158	Poster	RG 1	Wed 8:15p
Kazan	Learning binding preferences	3	Paper	RG	Wed 6p
Keränen	On computational analysis of quantitative, 3D spatial expression	253	Poster	SB 1	Fri 8:15p
Kheradpour	Regulatory motifs associated with TF	17	Oral	RG	Thu 11:30a
Kim	Peptide recognition domain (PRD)	45	Oral	DR	Fri 1p
Kim	When Two Plus Two Doesn't Equal Four: Modeling Non-	235	Poster	SB 1	Fri 8:15p
King	Structure-Based Prediction of Protein-Peptide Specificity inRosetta	198	Poster	DR 1	Thu 3:45p
Kinney	Regulatory physics from DNA sequence data	35	Oral	RG	Thu 7:45p
Kivinen	Selection of an optimal set of blood biomarker proteins	254	Poster	SB 1	Fri 8:15p
Klitgord	Predicting synthetic environments	86	Oral	DR	Sun 9:45a
Komorowski	Local molecular interdependency networks underlying HIV-1	204	Poster	SB 1	Fri 8:15p
Konieczka	Evolution of the High Osmolarity Glycerol (HOG) stress response	159	Poster	RG 1	Wed 8:15p
		205	Poster	SB 1	Fri 8:15p
Krogan	Insights from interaction maps	53	Invited	DR	Fri 6p
Kumar	The Msx1 Homeoprotein Recruits Histone Methyltransferase	102	Poster	RG 1	Wed 8:15p
Kural	Identification of Noncoding Motifs Under Selection in	130	Poster	RG 1	Wed 8:15p
Lachmann	GATE: Grid Analysis for Time-Series Expression	255	Poster	SB 1	Fri 8:15p

Lahesmaa	SATB1 dictates expression of multiple genes including IL5 human	206	Poster	SB 1	Fri 8:15p
Lambeck	Network inference by considering multiple objectives: Insights	180	Poster	DR(2)	Fri 8:15p
Lambeck	Reconstruction of a dynamic regulatory map from murine	272	Poster	SB 1	Fri 8:15p
Larjo	Simulating chemotactic and metabolic response	65	Oral	DR	Sat 11:15a
Lasserre	TSS detection	25	Oral	RG	Thu 3:15p
Laurila	Protein-protein interactions improve multiple transcriptionfactor	131	Poster	RG 1	Wed 8:15p
Le	Distance functions for querying large, multi species, expression	115	Poster	RG 1	Wed 8:15p
Lee	Successful Enhancer Prediction from DNA Sequence	132	Poster	RG 1	Wed 8:15p
Lee	Evolvability of the expression pattern of the Drosophila gap	256	Poster	SB 1	Fri 8:15p
Lefebvre	A human B cell interactome	51	Oral	DR	Fri 5:15p
Leiserson	Inferring Fault Tolerance from E-MAP Data	273	Poster	SB 1	Fri 8:15p
Lemischka	Systems level approaches to stem cell fate	68	Invited	DR	Sat 1p
Li	Identifying motifs using GADEM with a starting	133	Poster	RG 1	Wed 8:15p
Li	Human Cancer Proteome Variation Database and Mutated Peptides	257	Poster	SB 1	Fri 8:15p
Lieberman-Aidan	Fractal model for chromatin dynamics	36	Oral	RG	Thu 8p
Lin	Modeling Idiopathic Pulmonary Fibrosis Disease Progression based	116	Poster	RG 1	Wed 8:15p
		258	Poster	SB 1	Fri 8:15p
Liu	Gene network analysis of diabetes	66	Oral	DR	Sat 11:30a
Liu	Prediction of Polycomb target genes in mouse embryonicstem	178	Poster	RG 1	Wed 8:15p
Liu	A tri-partite clustering analysis on microRNA, gene and disease	259	Poster	SB 1	Fri 8:15p
Ljosa	Large-scale learning of cellular phenotypes	87	Oral	DR	Sun 10a
Logsdon	Regulatory network reconstruction	61	Oral	DR	Sat 9:45a
Lorenz	Rapid Estimation of RNA Kinetics	103	Poster	RG 1	Wed 8:15p
Lu	A novel method to simulate genome-wide background noise	134	Poster	RG 1	Wed 8:15p
Maas	The RNA Editing Dataflow System (REDS) for the transcriptome	104	Poster	RG 1	Wed 8:15p
		274	Poster	SB 1	Fri 8:15p
Maclsaac	Condition specific master regulators	77	Oral	DR	Sat 5:45p
Maclsaac	Novel methods for the discovery of condition specific	160	Poster	RG(2)	Thu 3:45p
MacKenzie	Life After Comparative Genomics; Regulatory Systems, Homeostasis	173	Poster	RG(2)	Thu 3:45p
		260	Poster	SB(2)	Sat 3:45p
Mahony	Chromatin state dynamics and the acquisition of	105	Poster	RG(2)	Thu 3:45p
Majoros	Modeling the Evolution of Regulatory Elements by Simultaneous	161	Poster	RG(2)	Thu 3:45p
Mar	Identifying Cell Lineage-Specific Gene Expression Modules	261	Poster	SB(2)	Sat 3:45p
Marbach	Generating realistic benchmarks for gene	47	Oral	DR	Fri 2:45p
Marbach	Strengths and weaknesses of network inference	50	Oral	DR	Fri 5p
Marchal	De novo detection and qualification of regulatory motifs.	135	Poster	RG(2)	Thu 3:45p
Marcotte	Insights into evolution and disease	76	Invited	DR	Sat 5:15p
Martins	Regulatory Element Identification with Functional Genomic	162	Poster	RG(2)	Thu 3:45p
Marucci	Turning genetic circuit into synthetic oscillator	32	Paper	RG	Thu 7p
Mayo	Hierarchical Model of Gas Exchange within the Acinar Airwaysof	236	Poster	SB(2)	Sat 3:45p
Mazloom	Linking MicroRNA and mRNA Co-Expressed Clusters to Regulatory	275	Poster	SB(2)	Sat 3:45p
McGettigan	Identification of epigenetic changes in the brain of a rat model	106	Poster	RG(2)	Thu 3:45p
Meyer	Inferring key transcriptional regulators	5	Oral	RG	Wed 6:45p
Meyer	Meta-Analysis in Transcriptional Network Inference	190	Poster	DR(2)	Fri 8:15p
Meysman	Structural DNA for the prediction of binding	20	Oral	RG	Thu 1:30p
Michaut	Exploring the Monochromatic Landscape in Yeast using	276	Poster	SB(2)	Sat 3:45p
Michoel	Enrichment and aggregation of topological motifs in integrated	277	Poster	SB(2)	Sat 3:45p
Michor	The cell of origin of human cancers	58	Invited	DR	Fri 7:45p
Miller	microRNAs preferentially target dosage-sensitive genes	107	Poster	RG(2)	Thu 3:45p
		278	Poster	SB(2)	Sat 3:45p
Mirny	Different strategies for gene regulation	8	Oral	RG	Wed 7:30p
Missiuro	Predicting Genetic Interactions in C. elegans using MachineLearning	117	Poster	RG(2)	Thu 3:45p
		279	Poster	SB(2)	Sat 3:45p
Molina	Studying transcription bursts from modeling high temporal	237	Poster	SB(2)	Sat 3:45p
Molinelli	Models from Experiments: Combinatorial Drug Perturbations of	207	Poster	SB(2)	Sat 3:45p
Morine	Combined inter-organ transcriptomic and metabolic analysis reveals	208	Poster	SB(2)	Sat 3:45p
Morris	Using accessibility to predict RNA-binding protein targets	108	Poster	RG(2)	Thu 3:45p
Narayanan	Simultaneous clustering	39	Paper	RG	Fri 9:45a
Ng	Identifying uncharacterized genes and functional networks inhuman	280	Poster	SB(2)	Sat 3:45p
Nielsen	CATCHprofiles reveals nucleosome positioning of histone	109	Poster	RG(2)	Thu 3:45p
Nir	Data Integration for High-throughput Morphological and	262	Poster	SB(2)	Sat 3:45p
Nolan	Single cell signaling & pathology	44	Invited	DR	Fri 11:30a

Novichkov	The automatic selection of TFBS score threshold in comparative	136	Poster	RG(2)	Thu 3:45p
Pandey	An Association Analysis Approach to Biclustering	263	Poster	SB(2)	Sat 3:45p
Pando	Adaptation of a synthetic gene circuit through diverse evolutionary	238	Poster	SB(2)	Sat 3:45p
Park	Dynamic networks	55	Paper	DR	Fri 6:45p
Parnell	Network analysis of gene-diet interactions for obesity	209	Poster	SB(2)	Sat 3:45p
Peleg	Network-free Inference of knockout effects	72	Paper	DR	Sat 2:45p
Pelizzola	Human DNA methylomes at base resolution	7	Oral	RG	Wed 7:15p
Perkins	Structure of cellular networks	80	Paper	DR	Sat 6:45p
Piipari	Inference and validation of a large cis regulatory motif setusing	137	Poster	RG(2)	Thu 3:45p
Pique-Regi	Genome-wide identification of Transcription Factor Binding Sites	281	Poster	SB(2)	Sat 3:45p
Polak	Large differences in transcription associated strand asymmetries	138	Poster	RG(2)	Thu 3:45p
Pollard	Unraveling of an ancient regulatory pathway: RNAi insensitivity	163	Poster	RG(2)	Thu 3:45p
		282	Poster	SB(2)	Sat 3:45p
Pu	Prediction of Chromatin Modification (CM)-related	283	Poster	SB(2)	Sat 3:45p
Qian	Effective identification of conserved pathways in biological networks	49	Paper	DR	Fri 4:45p
		284	Poster	SB(2)	Sat 3:45p
Rach	The landscape of transcription initiation	24	Oral	RG	Thu 3p
Raj	Variability in gene expression	67	Oral	DR	Sat 11:45a
Rajewsky	Post-transcriptional gene regulation	27	Invited	RG	Thu 5:15p
Rapaport	Copy number alterations in cancer	78	Oral	DR	Sat 6p
Rautajoki	ESTOOLSDB – a comprehensive database for stem cell research	164	Poster	RG(2)	Thu 3:45p
Ray	Discriminating functionality by kernel clustering k-mers inthe	165	Poster	RG(2)	Thu 3:45p
Regan	NF-kB and Forkhead – partners and opponents	175	Poster	RG(2)	Thu 3:45p
Reimand	Generalised linear models of transcriptional regulation to predict	191	Poster	DR(2)	Fri 8:15p
Reinitz	Finding the rules by asking the right questions	83	Invited	DR	Sat 7:30p
Richardson	Design of synthetic chromosomes	88	Oral	DR	Sun 10:45a
Rieder	Spatial association of multiple coordinately expressed but	166	Poster	RG(2)	Thu 3:45p
Robine	piRNA production in a Drosophila ovary cell line	110	Poster	RG(2)	Thu 3:45p
Rossetti	Epigenetic silencing of a tumor suppressor network unmasks the	111	Poster	RG(2)	Thu 3:45p
		210	Poster	SB(2)	Sat 3:45p
Roy	Learning probabilistic networks of condition-specific response	285	Poster	SB(2)	Sat 3:45p
Russo	Global Entrainment of Transcriptional Systems to Periodic Inputs	167	Poster	RG(2)	Thu 3:45p
Saez-Rodriguez	Discrete logic modeling to link pathway	41	Oral	DR	Fri 10:45a
Saez-Rodriguez	Challenge 3: Predictive signaling	46	Oral	DR	Fri 1:15p
Sahoo	Discovery of a branchpoint between B cell and T cell development	211	Poster	SB(2)	Sat 3:45p
Sales	Bayesian Nonparametric Clustering of Temporal Gene Expression	212	Poster	SB(2)	Sat 3:45p
Schwank	Organ growth control – from classical genetics to a systemslevel	213	Poster	SB(2)	Sat 3:45p
Sealfon	Supervised learning approaches to predicting enhancer regions	139	Poster	RG(2)	Thu 3:45p
Selvarajoo	Evidence for simple governing rules in complex biological networks	239	Poster	SB(2)	Sat 3:45p
Shamir	Signaling pathways analysis tool	26	Oral	RG	Thu 3:30p
Shen	Hybrid modeling and robustness analysis of Caulobacter cellcycle	192	Poster	DR(2)	Fri 8:15p
Shlomi	Predicting metabolic engineering KO Strategies	70	Oral	DR	Sat 1:45p
Skupsky	Integration site of the HIV promoter primarily modulates	176	Poster	RG(2)	Thu 3:45p
Srinivasan	Large-scale comparative analysis of RNA structures by TOPOFIT	112	Poster	RG(2)	Thu 3:45p
Storms	The Effect of Orthology and Coregulation on Detecting Regulatory	140	Poster	RG(2)	Thu 3:45p
Sugathan	Global DNase Hypersensitivity Mapping Reveals Growth Hormone	141	Poster	RG(2)	Thu 3:45p
Taher	Function conservation in diverged noncoding elements	142	Poster	RG(2)	Thu 3:45p
Taylor	Optimizing GO-based similarity for stroke pathologies usingnetworks	193	Poster	DR(2)	Fri 8:15p
Tsai	The integrated pathway and proteomic resources for identifying	118	Poster	RG(2)	Thu 3:45p
Tuller	Reconstructing ancestral gene content	89	Oral	DR	Sun 11:00a
Turinsky	Literature curation of protein interaction	40	Oral	RG	Fri 10a
		240	Poster	SB(2)	Sat 3:45p
Ulitsky	Towards prediction of MicroRNA function	4	Paper	RG	Wed 6:15p
Upadhyay	PPI module for visualization and analysis of protein-proteininterfaces	168	Poster	RG(2)	Thu 3:45p
		286	Poster	SB(2)	Sat 3:45p
Vacic	Estimating significance of CNV-pathway associations in	214	Poster	SB(2)	Sat 3:45p
Vallania	SPLINTER: detection of rare regulatory variants using a large	177	Poster	RG(2)	Thu 3:45p
Van Mourik	Continuous-time modeling of cell fate determination in Arabidopsis	215	Poster	SB(2)	Sat 3:45p
VanderSluis	Redundancy and asymmetric divergence of paralogs from genome	287	Poster	SB(2)	Sat 3:45p
Vandn	Identification of Significantly Mutated Pathways in Cancer	216	Poster	SB(2)	Sat 3:45p
Venner	Networks of Evolutionary Template Matches for Prediction	217	Poster	SB(2)	Sat 3:45p
Vermeirssen	Composite network motifs in integrated metazoan gene regulatory	169	Poster	RG(2)	Thu 3:45p

		241	Poster	SB(2)	Sat 3:45p
Vescio	Automatic Generation of In-Silico Biological Networks: a Cellular	194	Poster	DR(2)	Fri 8:15p
Vitkup	Prediction of human disease genes	69	Oral	DR	Sat 1:30p
Wadelius	Nucleosomes are positioned in exons	21	Oral	RG	Thu 1:45p
		288	Poster	SB(2)	Sat 3:45p
Wagner	Crosstalk among receptor tyrosine kinases	43	Oral	DR	Fri 11:15a
Waldman	TP53 cancerous mutations exhibit selection for translation efficiency	174	Poster	RG(2)	Thu 3:45p
		218	Poster	SB(2)	Sat 3:45p
Wang	Reconstructing Gene Cooperation Network of Lipotoxicity by	219	Poster	SB(2)	Sat 3:45p
Wang	Bottom-up Engineering of Synthetic Gene Networks	242	Poster	SB(2)	Sat 3:45p
Wasson	An ensemble model of competitive binding	33	Oral	RG	Thu 7:15p
Waterston	Deciphering C Elegans embryonic network	9	Invited	RG	Wed 7:45p
Wetmur	Stochastic modeling for loss of imprinted mRNA expression with	243	Poster	SB(2)	Sat 3:45p
White	Transcriptional regulatory networks	37	Invited	RG	Fri 9a
White	Microbial interaction web inference using metagenomic data	289	Poster	SB(2)	Sat 3:45p
Wohlbach	Identification of genomic features novel to xylose-fermenting	170	Poster	RG(2)	Thu 3:45p
		264	Poster	SB(2)	Sat 3:45p
Wojtowicz	Genome-wide mapping and computational analysis of non-B	244	Poster	SB(2)	Sat 3:45p
Wójtowicz	Mapping of non-B DNA structures	23	Oral	RG	Thu 2:45p
Won	Genome-wide prediction of transcription factor binding sites using	143	Poster	RG(2)	Thu 3:45p
Wooten	Network Based Analysis Identifies AXIN1/PDIA2 and Endoglin	220	Poster	SB(2)	Sat 3:45p
Wu	A dynamic analysis of the integrated control in the hepatic insulin	221	Poster	SB(2)	Sat 3:45p
Wu	Phosphopeptide-based signatures accurately predict the response	265	Poster	SB(2)	Sat 3:45p
Xie	Global Analysis of human protein-DNA Interactions	38	Oral	RG	Fri 9:30a
		290	Poster	SB(2)	Sat 3:45p
Xin	Epigenetic Profiling of Human Brain Development	113	Poster	RG(2)	Thu 3:45p
Yaffe	Systems biology of DNA damage	48	Invited	DR	Fri 4:15p
Yan	Structural and Regulatory Evolution of Electrophysiological Systems	171	Poster	RG(2)	Thu 3:45p
Yeang	An integrated analysis of molecular aberrations in cancer	222	Poster	SB(2)	Sat 3:45p
Ylipää	Association of genetic features with pathways using multiple high	291	Poster	SB(2)	Sat 3:45p
Yoon	Accurate and reliable cancer classification	62	Paper	DR	Sat 10a
Young	Programming cell state	10	Invited	RG	Thu 9a
Yu	Inferring Master Regulators of Glucocorticoid-Resistance in T	195	Poster	DR(2)	Fri 8:15p
Zaslaver	Metazoan operons accelerate transcription and recovery rates	144	Poster	RG(2)	Thu 3:45p
		245	Poster	SB(2)	Sat 3:45p
Zhang	Bayesian Learning and Optimization Approaches to Learning Gene	196	Poster	DR(2)	Fri 8:15p
Zhang	A probabilistic phylogenetic model to improve regulatory network	197	Poster	DR(2)	Fri 8:15p
Zhao	Inferring binding energies	12	Paper	RG	Thu 9:45a
Zheng	Computational Modeling of Crosstalk in Cancer Signaling Networks	292	Poster	SB(2)	Sat 3:45p
Zhong	A analysis of transcription factor interactions	16	Paper	RG	Thu 11:15a
Zhong	Edgetic perturbation models of human	18	Oral	RG	Thu 11:45a
Zhou	Determinants of Transcription Factor Binding and Regulation	145	Poster	RG(2)	Thu 3:45p
Zhu	Dynamic changes in the blood transcriptional network	71	Paper	DR	Sat 2p
Zinman	New insights into cross-species conservation	79	Oral	DR	Sat 6:15p

## Author Index

(\* denotes presenting author abstracts)

Abate-Shen, Cory.....	102
Abedi, Vida.....	80
Abid, Md.....	175
Aegerter-Wilmsen, Tinri.....	213
Aerts, Stein.....	28*
Agius, Phaedra.....	57,119*
Ahlfors, Helena.....	206*
Ahmed, Nabil.....	275
Aho, Kaisa-Leena.....	246*,266*
Aho, Tommi.....	266*
Äijö, Tarmo.....	146*
Aid, Malika.....	267*
Aird, William.....	175
Aittokallio, Tero.....	206
Akavia, Uri.....	73*
Al-Akwaa, Fadhl.....	268*
Alexopoulos, Leonidas.....	41,42*,46
Alizadeh, Ash.....	30,202
Almind, Katrine.....	66
Almusa, Henrikki.....	266
Alvarez, MarianoJ.....	51
Amato, Francesco.....	194
Arnazlag, Arnaud.....	147*
Andersen, Erik.....	67
Andersson, Robin.....	21
Angenent, GC.....	215
Ankley, Gerald.....	157
Antosiewicz-Bourget, Jessica.....	7
Aow, Jonathan.....	124
Apri, Mochamad.....	223*
Aravind, L.....	29
Arkin, Adam.....	176
Arndt, Peter.....	138
Arunachalam, Manonmani.....	120*
Arvey, Aaron.....	119,A5
Aswani, A.....	1
Atherton, J.....	1
Atluri, Gowtham.....	34,263
Autio, Reija.....	164
Avraham, Karen.....	26
Ay, Ferhat.....	31*,269*
Aziz, Ramy.....	A6
Babu, MM.....	29
Baddeley, Bob.....	193
Bader, Gary.....	45,224,276
Bader, Joel.....	55,59,88
Bakal, Chris.....	262
Balaji, S.....	29
Baldock, Richard.....	156
Bar-Joseph, Ziv.....	79,115,127,172,258
Barbara-Haley, Kellie.....	11
Barkai, Naama.....	19*
Barrow, John.....	173
Barry, Kerrie.....	264
Barton, David.....	32
Baryshnikova, Anastasia.....	34,224*,287
Basler, Konrad.....	213
Basso, Katia.....	51
Bate, Ashley.....	249
Baugh, Ryan.....	144
Beagley, Nat.....	193
Beer, Michael.....	132
Behrens, Sarah.....	121*
Belcastro, Vincenzo.....	247*,52
Bellay, Jeremy.....	34*,224,287
Bencic, David.....	157
Benham, Craig.....	23,244
Benos, Panagiotis.....	100
Bentink, Stefan.....	74
Benyamini, Tomer.....	54*
Ber, Yaara.....	26
Berger, Bonnie.....	262
Bernstein, Bradley.....	6
Betel, Doron.....	57*
Bethel, Wes.....	253
Beyer, Andreas.....	266
Bezy, Olivier.....	66
Bhardwaj, Nitin.....	150
Bhattacharya, Deepta.....	211
Bianchini, Julie.....	99
Bickel, P.....	1
Bieler, Jonathan.....	82*
Biggin, Mark.....	1*,253
Birin, Hadas.....	89
Blackshaw, Seth.....	38
Blatti, Charles.....	14
Boeck, M.....	9
Boeke, Jef.....	59*,88
Boley, N.....	1
Bolotin, Eugene.....	22*,122*
Bonneau, Richard.....	155,249
Bontempi, Gianluca.....	190
Boone, Charles.....	224,287
Borecky, Ingrid.....	177
Borenstein, Elhanan.....	85*
Bosotti, Roberta.....	52
Bourque, Guillaume.....	13*
Boyle, T.....	9
Bradley, Phil.....	198
Braun, David.....	93,199
Brendel, Volker.....	134
Brent, Michael.....	184
Bristow, Christopher.....	17,92*,139,154
Brodsky, Michael.....	14*
Brown, JB.....	1
Brown, Myles.....	5
Brownstein, Zippi.....	26
Brunetti, Nicola.....	52
Brynildsen, Mark.....	63*
Bugrim, Andrej.....	64*
Burnett, John.....	176
Burute, Mithila.....	206
Busch, Wolfgang.....	A3
Califano, Andrea.....	51,195,234

Callan, Curtis	35
Calvo, Sarah	2*
Camacho, DiogoM	60*
Candeias, Rogerio	149*
Cantone, Irene	32
Capala, Jacek	292
Carmel, Liran	225*
Carpenter, Anne	87
Carson, Matthew	150*
Carvalho, Luis	123*
Casellas, Rafael	23,244
Celniker, Sue	1,253
Cha, Hye	76
Chan, Christina	56*, 219, 221, 250
Chan, Clara	129
Chan, Esther	3
Chan, Michelle	205
Chang, Betty	93*, 199*
Chang, Cheng-Wei	227
Chang, King-Jen	179*
Chang, Li-Yun	179*
Chang, William	119
Chanrion, Benjamin	113
Chen, Chaang-Ray	227*
Chen, Chieh-Chun	16
Chen, Err-Cheng	118
Chen, Guang-Wu	118
Chen, Jia	243
Chen, Kuei-Tien	118
Chen, Yangqing	71
Chinnaiyan, Arul	64
ChiTa, Tuong	22
Cho, Emerson	40
Chowdhury, Sharif	266
Chrasegaran, Srinisavan	59
Chuang, Jeffrey	130
Chun, Hyonho	181*
Chung, Ho-Ryun	101
Chuu, Chih-Pin	43
Ciaccio, Mark	43
Claeys, Annelies	28
Claeys, Marleen	135, 140
Clark, Edward	228*
Clarke, Neil	124*
Clarke, Tim	228
Clote, Peter	15, 94*, 103, 128, 231
Collins, James	63, 242
Cook, Kristen	125*
Coombes, Cice	59
Cooper, Eric	59
Corcoran, DavidL	24
Correll, Mick	74
Cosentino, Carlo	194
Cosgrove, Elissa	182*
Cosma, MariaPia	32
Costa, Allen	11
Costanzo, Michael	34, 224, 287
Cowen, Lenore	273
Cox, EdwardC	35
Crampin, Edmund	187
Crawford, Greg	281
Cuccato, Giulia	230
Culhane, AedínC	74*
Dabrowski, Michal	126*
da Fonseca, LuisLopes	271
Dalkic, Ertugrul	250*
Dalla-Favera, Riccardo	51
Dang, Thanh	20
David, Eyal	26
Davidson, Andrew	99
Davidson, Duncan	156
Davidson, Scott	173
Davidson, SM	1
Davis, Matthew	152*
DChairakaki, Aikaterini	42
De-Moor, Bart	140
de Gee, Maarten	215, 223
Degner, Jacob	114*, 281
Dekker, J	36
de la Vega, Francisco	21
Deng, Minghua	181
Deng, Youping	196
Denslow, Nancy	157
DePace, Angela	253
Deshpande, Raamesh	270*
Devay, Piroaska	201*
Dezso, Zoltan	64
di Bernardo, Diego	32, 52, 230*, 247
di Bernardo, Mario	32, 167, 230*
Dill, David	192, 211
Ding, Huiming	224
Ding, Yang	15*, 130
Diplas, Andreas	243
Djebbari, Amira	A4*
Dojer, Norbert	126, 153*
Donaldson, Ian	40
Dotu, I	94, 231*
Dougherty, Edward	62
Downen, Robert	7
Down, Thomas	137
Dramiski, Micha	204
Druley, Todd	177
Dräger, Andreas	180
Dubchak, I	136
Dupuis, Dylan	274
Dymond, Jessica	59
Ebenhöh, Oliver	81, 158
Ecker, Joseph	7
Edsall, Lee	7
Eichenberger, Patrick	249
Eisen, Michael	1, 152, 253
Elkon, Rani	26
Ellis, Jonathan	183*
Ellis, Tom	242
Elo, Laura	206
Emilsson, Valur	71

Engelen, Kristof.....	20	Greenberg, Michael.....	11,97
Enroth, Stefan.....	21	Greenfield, Alex.....	155*
Epperlein, Jonathan.....	41	Gregoretti, Francesco.....	247
Erdin, Serkan.....	217	Grenier, Jennifer.....	A15
Erkkilä, Timo.....	251*	Grigoriev, Igor.....	264
Ernst, Jason.....	6*,17	Grochow, JoshuaA.....	29
Evans, Jane.....	22,122	Gsponer, Jörg.....	29
Fang, Fang.....	16	Guan, Yongtao.....	232*
Fang, Gang.....	75*	Gujral, Taranjit.....	A9*,A11
Feldman, Marcus.....	85	Gupta, Piyush.....	87
Feng, Feng.....	212	Gusenleitner, Daniel.....	74
Feng, Xin.....	95*	Guthke, Reinhard.....	180,272
Ferrando, Adolfo.....	195	Guttman, Mitch.....	67
Fischer, Maria.....	166	Gyenesei, Attila.....	156*,A2
Fisher, Bill.....	1,253	Habib, Tanwir.....	157*
Floreano, Dario.....	47,50	Hagen, Hans.....	253
Foley, Jonathan.....	176	Haghighi, Fatemeh.....	113
Folger, Ori.....	54	Hahne, Lauri.....	164
Fontana, Walter.....	84*	Haibe-Kains, Benjamin.....	190
Fornes, Oriol.....	127	Halliday, Gemma.....	173
Fowlkes, Charless.....	1,253	Hamann, Bernd.....	253
Fraenkel, Ernest.....	77,141,160	Hammonds, Ann.....	1,253
French, Courtney.....	205	Harel, David.....	65
Frogner, Charlie.....	92,154*	Harju, Manu.....	266
Gaffney, Daniel.....	281	Harmin, David.....	11
Gaither, Alex.....	201	Hart, Ronald.....	96
Galande, Sanjeev.....	206	Hartemink, Alexer.....	33
Gao, Yuan.....	24	Hawkins, RD.....	7
Garcia-Reyero, Natália.....	157	Haynes, Brian.....	184*
Gardner, Timothy.....	182	He, HH.....	5
Garraway, Levi.....	73	He, Xiaofei.....	38
Gasch, Audrey.....	264	He, Xin.....	16
Gatta, Giusy.....	195	Hechmer, A.....	1
Gauthier, Nicholas.....	207	Hedgepetha, Alyson.....	220
Gautier, Jean.....	51	Hemberg, Martin.....	11,97*
Ge, Hui.....	117	Hendriks, CrisLuengo.....	1,253
Ge, Yongchao.....	113	Henriquez, Clara.....	253
Gennemark, Peter.....	207	Herazo, Jose.....	258
Gentles, Andrew.....	30*,202*	Hermann, Robert.....	A2
Gerstein, M.....	9	Hescott, Benjamin.....	273
Gheorghiu, Stefan.....	236	Hickinbotham, Simon.....	228
Gifford, David.....	105	Hillier, L.....	9
Gilad, Yoav.....	114,281	Hing, Ben-Wen.....	173
Gilman, Sarah.....	69	Hinman, Veronica.....	100
Gitter, Anthony.....	127*	Hoffman, Brad.....	133
Goel, Gautam.....	271*	Hon, Gary.....	7
Goh, Wee.....	124	Hong, Feng.....	16
Gokhale, Paul.....	164	Hoodless, Pamela.....	133
Gong, Ping.....	196	Horowitz, Mark.....	192
Gopalan, Banu.....	193	Hou, Ping.....	185*
Gophna, Uri.....	89	Housmanc, David.....	220
Gormley, Claire.....	208	Hsiao, Tzu-Lin.....	69
Gottimukkala, Kamal.....	206	Hsieh, Fon-Jou.....	179
Gottschling, Daniel.....	59	Hsu, Ian.....	227
Granás, David.....	12	Hu, Jianzhong.....	243
Grant, Marianne.....	175	Hu, Shaohui.....	38
Gray, Jesse.....	11*,97	Hu, Wei-Shou.....	270
Green, P.....	9	Hubbard, Tim.....	137

Huggins, Peter .....	172*	Karsunky, Holger .....	211
Hugginsa, Gordon .....	220	Kartal, Önder .....	81*,158*
Hughes, Timothy .....	3	Kasif, Simon .....	66,265
Huhtala, Heini .....	246	Kaufmann, K .....	215
Hurley, Daniel .....	187*	Kaushik, Poorvi .....	207
Huttenhower, Curtis .....	A12*	Kazan, Hilal .....	3*
Huynen, Martijn .....	109	Kazanov, Marat .....	A6
Hwa, Hsiao-Lin .....	179	Kazemian, Majid .....	14
Hwang, Catalina .....	13	Keim, Celia .....	102
Hwang-Verslues, Wendy .....	22,122	Kellis, Manolis. 6,17,92,96,129,139,149,154	
Hyman, A .....	9	Kepler, Thomas .....	212
Iakoucheva, Lilia .....	214	Keränen, Soile .....	1,253*
Ilonen, Jorma .....	A2	Khali, Ahmad .....	96
Ilyin, Valentin .....	112,233*,286	Khalil, Ahmad .....	67
Imakaev, M .....	36	Kharchenko, Peter .....	6
Immink, GH .....	215	Kheradpour, Pouya .....	6,17*,92,139,154
Ingalls, Brian .....	80	Kierczak, Marcin .....	204
Inlay, Matthew .....	211	Kim, Ah-Ram .....	83,235*
Ionides, John .....	235	Kim, Jessica .....	73
Iorio, Francesco .....	52*,247	Kim, Philip .....	45*
Isacchi, Antonella .....	52	Kim, S .....	9
Iyera, Lakshmanan .....	220	Kim, Tae-Kyung .....	11,97
Jaakkola, Tommi .....	117	Kim, Yungil .....	224
Jabbari, Hosna .....	128*	Kimchi, Adi .....	26
Jacobsen, Anders .....	98*	King, Chris .....	198*
James, DC .....	A10	Kinney, Justin .....	35*
Jang, InSock .....	234*	Kiseleva, Larisa .....	A13
Janssens, Hilde .....	83	Kivinen, Virpi .....	254*
Jayasurya, Karthik .....	120	Klamt, Steffen .....	41
JCollins, James .....	60	Klitgord, Niels .....	86*
Jeffries, Thomas .....	264	Klutstein, Michael .....	127
Jeong, Jun-Seop .....	38	Knip, Mikael .....	A2
Jessell, Thomas .....	105	Knowles, David .....	1,253
Jeyakani, Justin .....	13	Kobe, Bostjan .....	183
Ji, Sungchul .....	99*	Koh, Judice .....	224
Jiang, Bo .....	5	Kohanski, MichaelA .....	60
Jiang, Lizhi .....	38	Kohlbacher, Oliver .....	A3
Jiang, Tao .....	22,122	Kolaczyk, Eric .....	182
Jones, Richard .....	43	Koller, Daphne .....	30,202
Jones, Steven .....	133	Komorowski, Jan .....	21,204*
Joshi, Anagha .....	188*	Kong, Lingjia .....	164
Jothi, Raja .....	29*	Konieczka, Jay .....	205*
Jungreis, Irwin .....	129*	Konishi, Kazuhisa .....	258
Järvenpää, Laura .....	164	Koonin, Eugene .....	225
Kacmarczyk, Thadeous .....	249	Koppal, Anjali .....	57
Kadah, Yasser .....	268	Koronacki, Jacek .....	204
Kadri, Sabah .....	100*	Korpi, Alicia .....	130
Kaern, Mads .....	80	Kotliar, Dylan .....	73
Kahn, CR .....	66	Kouzine, Fedor .....	23,244
Kahveci, Tamer .....	31,269	Kramer, Maxwell .....	282
Kaminska, Bozena .....	126	Kreiman, Gabriel .....	11,97
Kaminski, Naftali .....	258	Kreutzer, Michael .....	A13
Kang, Jia .....	181	Krogan, Nevan .....	53*
Kaplan, T .....	1	Krogh, Anders .....	98
Karchin, Rachel .....	132	Kuang, Rui .....	75
Karlebach, Guy .....	26	Kuersten, Scott .....	11
Karlic, Rosa .....	101*	Kugler, Hillel .....	65
Karpen, Gary .....	6	Kulkarni, Tripti .....	112

Kumar, Roshan .....	102*	Lin, Jimmy.....	38
Kumar, Sudhir .....	14	Lin, Michael.....	129
Kumar, Vipin.....	75,263	Lin, Tien-ho.....	258*
Kunarso, Galih .....	13	Lindroos, Bettina .....	246
Kuo, Alan.....	264	Lindstrom, Derek .....	59
Kuo, Wen-Hung.....	179	Ling, Guoyu .....	141
Kuokkanen, Hannu.....	246	Lipponen, Kati.....	A2
Kupiec, Martin .....	89	Lipshitz, Howard .....	108
Kural, Deniz.....	130*	Lisewski, AM.....	217
Labbe, Aurélie .....	A4	Lister, Ryan.....	7
Lachmann, Alexander .....	255*	Litvin, Oren.....	73
Lahesmaa, Riitta .....	206,A2	Liu, Jun .....	5
Lai, Chao-Qiang .....	209	Liu, Li .....	157
Lai, Eric .....	110	Liu, Li-Yu.....	179
Laiho, Asta .....	A2*	Liu, Manway.....	66*
Lamb, John .....	71	Liu, Shirley.....	5
Lambeck, Sandro .....	180*,272*	Liu, Ying.....	259*
Lane, Terran.....	285	Liu, Yingchun .....	178*
Larjo, Antti .....	65*,266	Ljosa, Vebjorn .....	87*
Lasserre, Julia.....	25*,101	Loewer, Sabine .....	96
Lau, Nelson .....	110	Loftus, Brendan .....	106
Lauffenburger, Douglas.....	41,43,46	Logsdon, Benjamin .....	61*
Laukens, Kris .....	20	Lohmann, Jan .....	A3
Laurent, Jon .....	76	Long, Shunyou.....	38
Laurent, Louise .....	4	Lopresti, Daniel.....	274
Laurila, Kirsti .....	131*	Lorenz, WA .....	15,94,103*,231
Lawrence, Charles .....	123	Loyal, Goff .....	96*
Le, Hai-Son .....	115*	Lu, Hong .....	134*
Lear, Marissa .....	173	Lu, Hui.....	150
Lee, Dongwon .....	132*	Lu, Yun .....	93,199
Lee, Ho-Joon.....	256*	Lund, Riikka .....	164
Lee, Leonard .....	7	Luscombe, Nicholas .....	191
Lee, Su-In.....	30,202	Lähdesmäki, Harri.....	131,146,251,266
Lee, Yu-Chi .....	209	Ma'ayan, Avi .....	255,275
Lefebvre, Celine .....	51*	Ma, Haisu.....	181
Leinonen, Kalle .....	164,266	Ma, Zeqiang.....	257
Leiserson, Mark.....	273	Maaron, Aiveen.....	208
Lemischka, Ihor.....	68*,93,199,255	Maas, Stefan.....	274*
Lemponen, Riina .....	246	Macarthur, Ben .....	255
Leonardson, Amy.....	71	MacArthur, S.....	1
Lander, Eric .....	36,87	MacBeath, Gavin .....	A9,A11,A15
Leslie, Christina.....	57,78,119,A5	MacCoss, M.....	9
Levens, David .....	23,244	Mace, D.....	9
Levy, Ron .....	30,202	MacIsaac, Kenzie .....	77*,160*
Li, Jing.....	1,257*	MacKenzie, Alasdair .....	173*
Li, Ker-Chau .....	A8	Madar, Aviv .....	155
Li, Leping.....	133*	Mader, Sylvie .....	267
Li, Peng.....	196	Mahony, Shaun.....	105*
Li, Xiao .....	108	Majoros, WH .....	161*
Li, XY.....	1	Makarov, Vladimir .....	214
Li, Yixue .....	A13	Malhotra, Dheeraj .....	214
Liao, Hailing .....	22,122	Malik, Jitendra.....	1,253
Lichtarge, Olivier .....	217	Mar, Jessica.....	261*
Lieberman-Aiden, E .....	36*	Marbach, Daniel.....	47,50*
Liebler, Daniel .....	257	Marchal, Kathleen.....	20,135*,140
Lim, Geoffrey.....	124	Marcotte, Edward.....	76*
Lim, Wei .....	51,195	Margueron, Raphael .....	102
Limaye, Amita .....	206	Marioni, John .....	114

Markenscoff-Papadimitriou, Eirene .....	11	Morrison, Kyle .....	40
Marks, Debora .....	98,107,A5*	Moses, Eyal .....	73
Martin, Dan .....	254	Murphy, Keith .....	106
Martinez, Carlos .....	83	Murray, JI .....	9
Martins, Andre .....	162*	Murugan, An .....	35
Marucci, Lucia .....	32*	Musso, Gabriel .....	287
Matilainen, Jukka .....	266	Myers, Chad .....	75,224,263,270,287
Mavis, Swerdel .....	96	Myklebust, June .....	30,202
Mayo, Michael .....	236*	Mäkelä, Jarno .....	164
Mazloom, Amin .....	275*	Müller, Waltraud .....	166
Mazor, Tali .....	141	Nachtergaele, Bruno .....	277
Mazzoni, Esteban .....	105	Naef, Félix .....	82,147,237
McAdams, Harley .....	192	Narayanan, Manikan .....	39*
McCallion, Andrew .....	142	Nath, Aritro .....	219
McCarthy, Shane .....	214	Nelander, Sven .....	207
McCuine, Scott .....	105	Nery, Joseph .....	7
McCutchan, Michael .....	14	Nesvzhskii, Alexey .....	64
McDermott, Jason .....	193	Newberg, Lee .....	123
McGary, Kriston .....	76	Ng, Aylwin .....	280*
McGaughey, David .....	142	Ng, Chris .....	77,160
McGettigan, Paul .....	106*	Ng, Huck-Hui .....	13,16
McKernan, Kevin .....	21	Ngo, Que-Minh .....	7
McKinnon, David .....	171	Ni, Ting .....	24
McLeod, J .....	A10	Nielsen, Fiona .....	109*
McNally, James .....	166	Nir, Oaz .....	262*
Melas, IoannisN .....	42	Nkadori, Everlyne .....	114
Mendelsohna, Michael .....	220	Noble, William .....	119
Meyer, Clifford .....	5*	Nolan, Garry .....	44*
Meyer, Patrick .....	190*	Notani, Dimple .....	206
Meysman, Pieter .....	20*	Novichkov, Pavel .....	136*,A6
Mezey, Jason .....	61	Novichkova, Elena .....	A6
Michaut, Magali .....	276*	Nykter, Matti .....	254,291
Michael, Tom .....	169,188,241,277*	O'Callaghan, PM .....	A10*
Michor, Franziska .....	58*	O'Donnell, Anne .....	113
Mieczkowski, Jakub .....	126	O'Gaora, Peadar .....	208
Miettinen, Susanna .....	246	O'Shea, Erin .....	125,145,205
Mikkelsen, Tarjei .....	6	O'Sullivan, Niamh .....	106
Millar, AH .....	7	Oberholzer, Patrick .....	73
Miller, D .....	9	Ohler, Uwe .....	24,161,120
Miller, Martin .....	107*,207	Oliva, Baldo .....	127
Mirny, Leonid A .....	8*,36	Oliva, Gennaro .....	247
Mironov, A .....	136	Olson, Brian .....	88
Missiuro, Patrycja .....	117*	Omenn, Gilbert .....	64
Mitra, Robi .....	177	Ones, Thouis .....	87
Mitsos, Alexer .....	42	Onishi, Akishi .....	38
Modi, SheetalR .....	60	Ordovas, Jose .....	209
Molenaar, Jaap .....	215,223	Orlov, Yuryi .....	124
Molina, Nacho .....	237*	Osterman, Andrei .....	A6
Molinelli, Evan .....	207*	Ostler, Harry .....	155
Moloney, Aidan .....	208	Ouyang, Zhengyu .....	185
Monagle, Jolene .....	208	Ovcharenko, Ivan .....	142
Montefusco, Maria .....	220	Pagliari, David .....	2
Moore, Christopher .....	55	Pai, Athma .....	114,281
Mootha, Vamsi .....	2	Paisios, Nektarios .....	93,199
Morgan, Tom .....	92,154	Pandey, Gaurav .....	263*
Mori, MarceloA .....	66	Pando, Bernardo .....	238*
Morine, Melissa .....	208*	Papenhausen, Gerald .....	74
Morris, Quaid .....	3,108*	Papp, Balazs .....	287

Park, Peter	6	Reimand, Jüri	191*
Park, Tae	76	Reinberg, Danny	102
Park, Yongjin	55*	Reinitz, John	83*,235
Parnell, Laurence	209*	Reinke, V	9
Payneb, Douglas	220	Ren, Bing	7,143
Paz, Arnon	26	Restrepo, Simon	213
Pe'er, Dana	73	Reynolds, Clare	208
Peleg, Tal	72*	Rho, Hee-sool	38
Pelizzola, Mattia	7*	Ricci, MariaAurelia	32
Perkins, Edward	157,196	Richards, Adam	14
Perkins, Theodore	80*	Richardson, Sarah	59,88*
Perrimon, Norbert	262	Rieder, Dietmar	166*
Pey, Gaurav	34,75	Rienschke, Rick	193
Pfeifer, Peter	236	Rifkin, Scott	67
Pfiffner, Jenna	205	Rinn, John	67,96
Picard, Kermshlise	74	Robertson, Gordon	133
Picard, Shaita	74	Robine, Nicolas	110*
Pickrell, Joseph	114	Roche, Helen	208
Piipari, Matias	137*	Rockman, Matthew	282
Pique-Regi, Roger	281*	Rodionov, Dmitry	136,A6*
Plevritis, Sylvia	30,202,211	Rodionova, Irina	A6
Poggio, Tomaso	92,154	Roos, Christophe	266
Polak, Paz	138*	Root, David	A15
Pollard, Daniel	282*	Rosasco, Lorenzo	92,154
Polynikis, Athanasios	230	Rosati, Barbara	171
Pop, Mihai	289	Roseberry, Gene	193
Pozorini, Christian	82	Rossetti, Stefano	111*
Prill, Robert	46,50	Roy, Sushmita	285*
Print, Cristin	187	Run, Jin	124
Pritchard, Jonathan	114,281	Ruotti, Victor	7
Przytycka, Teresa	23,29,244,292	Ruppin, Eytan	54,72,89,174
Przytycki, Pawel	292	Russo, Giovanni	167*
Pu, Shuye	283*	Ruusuvuori, Pekka	251,266
Pukkila, Heidi	131,164	Rätsch, Gunnar	A3
Pukonen, Inga	A2	Räty, Sari	246
Pybus, L	A10	Rübel, Oliver	253
Qian, Jiang	38	Sabo, P	1
Qian, Xiaoning	49*	Sacchi, Nicoletta	111
Qian, Ziliang	A13	Sachidanam, Ravi	93
Quackenbush, John	74,261	Sachidanandam, Ravi	199
Quan, Xiao-Jiang	28	Saez-Rodriguez, Julio	41*,42,46*
Querfurth, Robert	138	Sagir, Dorit	26
Quon, Gerald	108	Sahoo, Debashis	211*
Rach, Elizabeth	24*	Sales, Ana	212*
Rada-Iglesias, Alvaro	21	Samaga, Regina	41
Raj, Arjun	67*	Samsonova, Maria	83
Rajbhari, Presha	51	Sanchez, Aminael	140
Rajewsky, Nikolaus	27*	Sander, Chris	207
Rapaport, Franck	78*	Sanfilippo, Antonio	193
Raphael, Benjamin	216	Santini, Stefania	32
Rasool, Omid	206	Sato, Mai	51
Rautajoki, Kirsi	164*	Saunders, Neil	183
Ray, Debashish	3	Schadt, Eric	39,71
Ray, Pradipta	165*	Schaffer, David	176
Razick, Sabry	40	Schaffer, Thomas	47,50
Reeder, Christopher	105	Scheideler, Marcel	166
Regan, Erzsebet	175*	Scheifele, Lisa	59
Regev, Aviv	205	Schibler, Ueli	237

Schultheiss, Sebastian	A3*	Steffen, Martin	265
Schwank, Gerald	213*	Stein, Lincoln	95
Schwarzl, Thomas	74	Steinbach, Michael	75,263
Scott, Matt	80	Stepney, Susan	228
Sealfon, Rachel	17,92,139*,154	Sternberg, Paul	144
Sebat, Jonathan	214	Stewart, Ron	7
Segrè, Daniel	86	Stocker, Gernot	166
Seita, Jun	211	Stoev, Ivan	274
Selvarajoo, Kumar	239*	Stolovitzky, Gustavo	46,47*,50
Semple, Colin	156	Stormo, Gary	12
Seppälä, Janne	164	Storms, Valerie	135,140*
Ser, Chris	57	Stunnenberg, Henk	109
Serwold, Thomas	211	Styczynski, Mark	205
Sevecka, Mark	A11,A15*	Su, Junjie	62
Shachaf, Catherine	30,202	Sugathan, Aarathi	141*
Shahbaba, Babak	30,202	Sultana, Razvan	74
Shakhnovich, Eugene	256	Sumazin, Pavel	51
Shamir, Ron	4,26*	Suter, David	237
Shanley, Lynne	173	Suuronen, Riitta	246
Shao, Zhen	178	Sze, Mei	156
Shapiro, Hagit	66	Szigarto, Cristina	A13
Shapiro, Lucy	192	Ta, Tuong	122
Sharan, Roded	72,174	Tabb, David	257
Sharma, Shikha	270	Taher, Leila	142*
Sharp, David	83	Tai, Shang-Kai	A8*
Shehu, Amarda	88	Tatar, Diana	273*
Shen, Chengcheng	259	Taylor, Ronald	193*
Shen, Xiling	192*	Tepper, Naama	70*
Shiloh, Yossi	26	Thieffry, Denis	256
Shlomi, Tomer	54,70	Thiesen, H-J	A13*
Shmulevich, Ilya	251,254,291	Thomas, S	1
Shyu, Ming-Kwang	179	Thompson, Dawn	205
Siciliano, Velia	230,247	Thompson, William	123
Sie, ChristinaGodfried	274	Thomson, James	7
Siegfried, Zehava	127	Tirosh, Itay	19
Siepel, Adam	162	Todd, Annabel	191
Simell, Olli	A2	Tomancak, Pavel	120
Siminelakis, Paraskeuas	42	Tomita, Masaru	239
Simirenko, L	1	Tomlin, C.	1
Simon, Itamar	127	Tonti-Filippini, Julian	7
Sinha, Saurabh	14,16	Toomey, Sinead	208
Skupsky, Ron	176*	Toufighi, Kiana	224
Sladek, Frances	22,122	Trajanoski, Zlatko	166
Slebos, Robbert	257	Tsai, Kun-Nan	118*
Snyder, M	9	Tsuchiya, Masa	239
Socha, Amanda	205	Tuller, Tamir	89*,174
Solouma, Nahed	268	Tuomela, Soile	206
Song, Carl	80	Tuomisto, Lauri	164
Song, Joe	185	Turinsky, Andrei	40*,283
Song, Le	165	Turner, Brian	40
Song, Shen	24	Tzur, Yossi	73
Sontag, Eduardo	167	Uhlén, Mathias	A13
Sorger, Peter	41,46	Ulitsky, Igor	4*,26
Spana, Eric	24	Upadhyay, Amit	286*
Sreekumar, Arun	64	Upfal, Eli	216
Srinivasan, Preethi	112*	Vacic, Vladimir	214*
Stamatoyannopoulos, J	1	Vallania, Francesco	177*
Stavrovskaya, E	136	van Berkum, NL	36

Vanden-Eijnden, Eric .....	155	Winkler, Jonathan .....	63
Van De Peer, Yves.....	169,188,241,277	Wodak, Shoshana .....	40,283
VanderSluis, Benjamin.....	287*	Wohlbach, Dana .....	264*
van Dijk, ADJ .....	215	Wolf-Yadlin, Alejandro .....	A11*
Vandin, Fabio .....	216*	Wolfe, Scot .....	14
van Ham, RCH .....	215	Won, Kyoung-Jae .....	143*
VanHentenryck, P .....	94,231	Woodard, Crystal .....	38
van Mourik, S .....	215*	Woods, John .....	76
van Oudenaarden, Alex .....	67,238	Wootena, Eric .....	220*
van Voorn, Gak .....	223	Wu, Chang-Jiun .....	265*
Vaquerizas, Juan .....	191	Wu, Jiantao .....	130
Vellaichamy, Adaikkalam .....	64	Wu, Joy .....	59
Venner, Eric .....	217*	Wu, Ming.....	56,221*
Vermeirssen, Vanessa .....	169*,241*	Wunderlich, Zeba.....	8
Verstuyf, Annemieke .....	140	Wuster, Arthur.....	29
Vescio, Basilio.....	194*	Wójtowicz, Damian .....	23*,244*
Vetta, Adrian .....	39	Xavier, Ramnik .....	280
Vidal, Marc .....	18	Xie, Zhi.....	38*
Villeneuve, Daniel .....	157	Xin, Yurong .....	113*
Vilo, Jaak.....	191	Xing, Eric .....	165
Vingron, Martin.....	25,101,121	Xu, Fei.....	31,269
Vitkup, Dennis .....	69*	Yadlin, Alejandro.....	A15
Vlahovick, Kristian.....	101	Yaffe, Michael .....	48*
Vogel, Christine.....	76	Yamane, Arito .....	23,244
Voit, Eberhard .....	271	Yan, Jiekun .....	28
Wade, Herschel.....	38	Yan, Qinghong.....	171*
Wadelius, Claes .....	21*	Yang, Chuhu .....	22,122
Wagner, Joel .....	43*	Yang, Xuerui .....	56,219
Waldman, Yedaël.....	174*	Ye, Zhen .....	7
Wallingford, John .....	76	Yeang, Chen-Hsiang .....	222*
Waltman, Peter .....	249*	Yli-Harja, Olli .. 65,131,164,246,254,266,291	
Wang, Haizhou .....	185	Ylipää, Antti .....	291*
Wang, Hong .....	38	Yoon, Byung-Jun .....	49,62*
Wang, Jingqiang .....	102	Yoon, Seungtai .....	214
Wang, Kai.....	51,71	Yosef, Nir .....	72
Wang, Wei.....	143	Young, Peter .....	228
Wang, Weiqing.....	207	Young, Richard .....	10*,102,105
Wang, Xiao.....	242*	Yu, Jiyang .....	195*
Wang, Xuewei .....	219*	Yu, Xueping .....	38
Wapinski, Ilan.....	205	Yuan, Guocheng .....	178
Ward, RM .....	217	Yuan, Shinsheng .....	A8
Wasson, Todd.....	33*	Zaslaver, Alon.....	144*
Waterston, RH.....	9*	Zawadzka, Malgorzata.....	126
Waxman, David.....	141	Zhang, Bing .....	257
Weber, Gunther.....	253	Zhang, Chaoyang .....	196*
Weissman, Irving.....	211	Zhang, David .....	292
Wells, Christine .....	261	Zhang, Linxia .....	219
Wen, Jean .....	98	Zhang, Wei .....	291
Werner-Washburne, Margaret .....	285	Zhang, Xianghua .....	181
Wetmur, James .....	243*	Zhang, Xiuwei .....	197*
White, James .....	289*	Zhang, Yang .....	185
White, Joe .....	74	Zhao, Hongyu .....	181
White, Kevin.....	37*	Zhao, Yue .....	12*
Wichterle, Hynek .....	105	Zhao, Z.....	9
Wiedmann, Brigitte.....	201	Zheng, Jie .....	292*
Wijnen, Herman .....	147	Zhong, Quan .....	18*
Wilhelm, Thomas .....	266	Zhong, Shan .....	79
Wilkinson, SJ.....	A10	Zhong, Sheng .....	16*

Zhou, Xu ..... 145\*

Zhu, Heng ..... 38

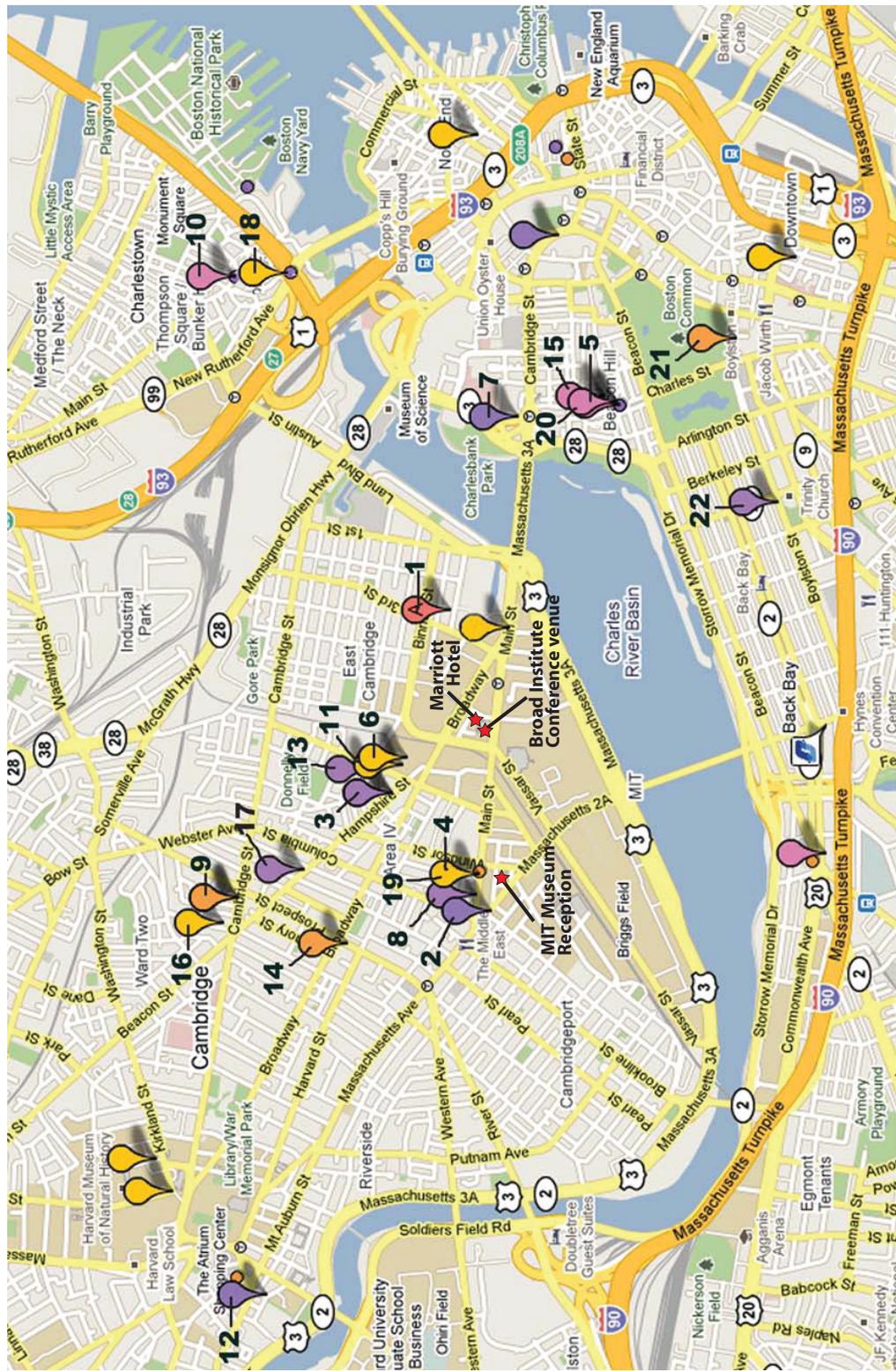
Zhu, Jun ..... 24,39,71\*

Zielinski, Rafal ..... 292

Ziems, Bjoern ..... A13

Zien, Alexer ..... 25

Zinman, Guy ..... 79\*



**Cambridge**

**MIT Museum Reception**

**Marriott Hotel**

**Broad Institute Conference venue**

12

16

9

14

3

13

11

6

1

4

19

8

2

3

20

15

5

21

22

2

9

3

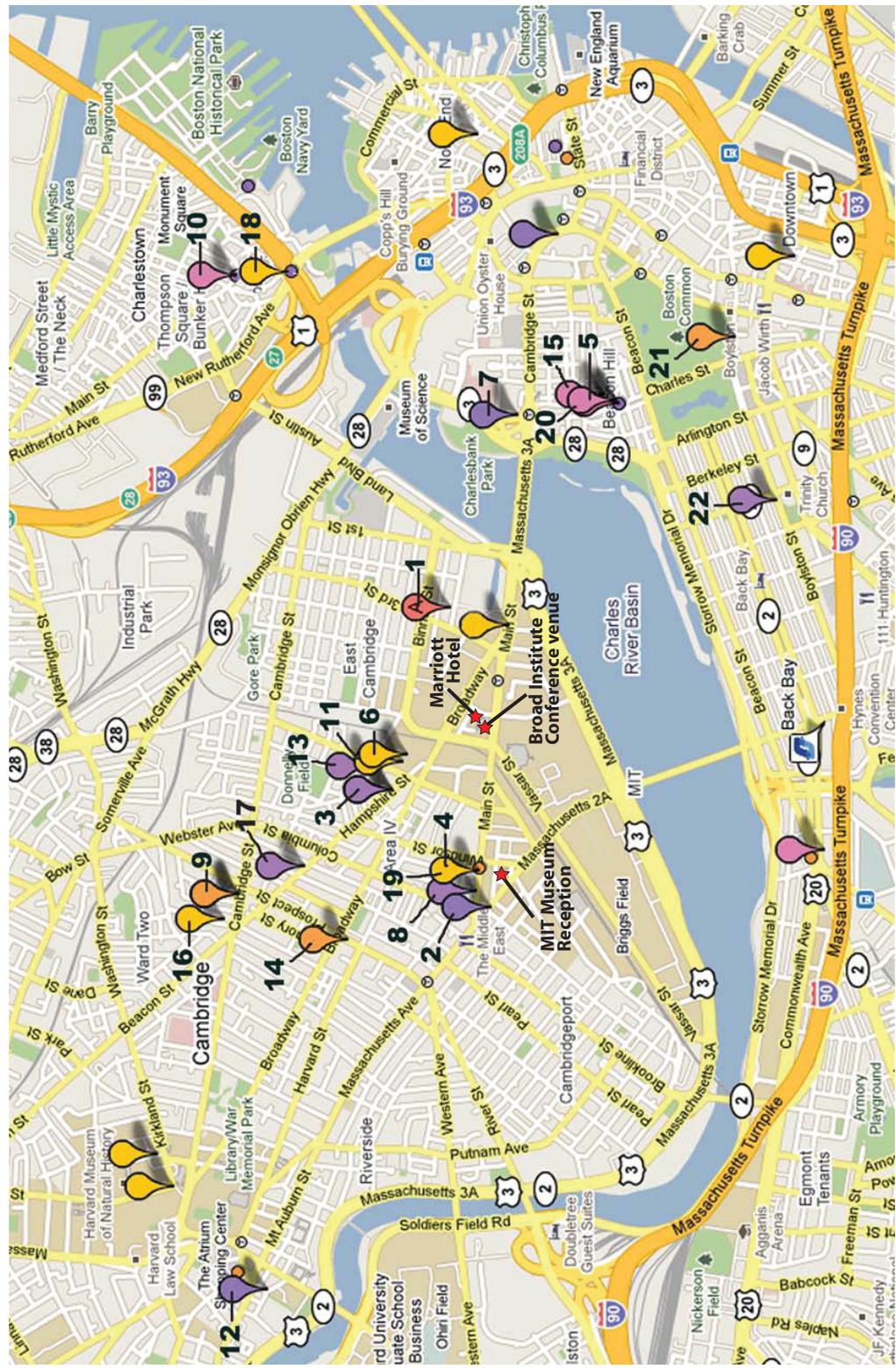
1

3

3

3

3



## Restaurant List

**1. Aceituna** - *Lebanese/Middle Eastern Cuisine*

605 W. Kendall St., Cambridge  
617-252-0707  
*Evening, Lunch*

**2. Asgard** - *Irish Cuisine*

350 Mass Ave., Cambridge  
617-577-9100  
*Evening, Lunch*

**3. Atasca** - *Portuguese Cuisine*

50 Hampshire St., Cambridge  
617-621-6992  
*Evening, Lunch*

**4. Bertucci's** - *Italian Cuisine*

799 Main St., Cambridge  
617-661-8356  
*Evening, Lunch*

**5. Bin 26 Enoteca** - *Italian Cuisine*

26 Charles St, Boston  
617-723-5939  
*Evening, Formal, Lunch*

**6. Blue Room** - *Modern American Cuisine*

One Kendall Square, Cambridge  
617-494-9034  
*Evening, Formal*

**7. Clink at Liberty Hotel** - *Modern Amer. Cuisine*

215 Charles Street, Boston  
617-224-4004  
*Evening, Formal, Lunch*

**8. Craigie on Main** - *French Cuisine*

853 Main Street, Cambridge  
617-497-5511  
*Evening, Formal, Lunch*

**9. East Coast Grill & Raw Bar** - *Seafood/Barbeque*

1271 Cambridge St., Cambridge  
617-491-6568  
*Evening, Lunch*

**10. Figs** - *Modern American Cuisine*

67 Main St., Charlestown  
617-242-2229  
*Evening, Formal, Lunch*

**11. Friendly Toast** - *American Breakfast (all day)*

One Kendall Square, Cambridge  
617-577-8668  
*Evening, Lunch*

**12. Henrietta's Table** - *Modern American Cuisine*

1 Bennett St., University Rd., Cambridge  
617-661-5005  
*Evening, Lunch*

**13. Hungry Mother** - *Modern American Cuisine*

233 Cardinal Medeiros Ave, Cambridge  
617-499-0090  
*Evening, Formal, Lunch*

**14. Koreana** - *Korean Cuisine*

154~158 Prospect at Broadway, Cambridge  
617-576-8661  
*Evening, Lunch*

**15. Lala Rokh** - *Persian Cuisine*

97 Mt Vernon St, Boston  
617-720-5511  
*Evening, Formal, Lunch*

**16. Ole Mexican Grille** - *Mexican Cuisine*

11 Springfield St, Cambridge  
617-492-4495  
*Evening, Lunch*

**17. Oleana** - *Mediterranean Cuisine*

134 Hampshire St., Cambridge  
617-661-0505  
*Evening, Lunch*

**18. Olives** - *Modern American Cuisine*

10 City Sq., Charlestown  
617-242-1999  
*Evening, Lunch*

**19. Salt's** - *Modern American Cuisine*

789 Main Street, Cambridge  
617-876-8444  
*Evening, Formal, Lunch*

**20. Ristorante Toscano** - *Italian Cuisine*

41-47 Charles St., Boston  
617-723-4090  
*Evening, Formal, Lunch*

**21. Troquet** - *French Cuisine*

140 Boylston St., Boston  
617-695-9463  
*Evening, Formal, Lunch*

**22. Viora Restaurant** - *Mediterranean Cuisine*

545 Boylston Street, Boston  
617-638-9699  
*Evening, Formal, Lunch*

## Registered Participants

### Invited Speakers

Naama Barkai  
Mark Biggin  
Jef Boeke  
Walter Fontana  
Nevan Krogan  
Ihor Lemischka  
Franziska Michor  
Garry Nolan  
Nikolaus Rajewsky  
John Reinitz  
Bob Waterston  
Kevin White  
Michael Yaffe  
Richard Young

naama.barkai@weizmann.ac.il  
mdbiggin@lbl.gov  
jboeke@jhmi.edu  
walter@hms.harvard.edu  
krogan@cmp.ucsf.edu  
ihor.lemischka@mssm.edu  
michorf@mskcc.org  
gnolan@stanford.edu  
rajewsky@mdc-berlin.de  
reinitz@odd.bio.sunysb.edu  
waterston@gs.washington.edu  
kpwhite@uchicago.edu  
myaffe@mit.edu  
young@wi.mit.edu

### Academic

Stein Aerts  
Leonidas Alexopoulos  
Dimitris Anastassiou  
Inma Barrasa  
George Bell  
Panayiotis (Takis) Benos  
Ginestra Bianconi  
Richard Bonneau  
Guillaume Bourque  
Michael Brent  
Michael Brodsky  
Andrea Califano  
Liran Carmel  
Luis Carvalho  
Christina Chan  
Jeffrey Chuang  
Neil Clarke  
Peter Clote  
Jacques Cohen  
Edmund Crampin  
Michal Dabrowski  
Vlado Dancik  
Diego Di Bernardo  
Anthony DiBiase  
Christoph Dieterich  
Amira Djebbari  
Norbert Dojer  
Ivan Dotu  
Dan Galian  
Andrew Gentles  
Ping Gong  
Suresh Gopalan  
Patricia Greninger  
Hubert Hackl  
David Harmin  
Oliver Hofmann  
Tim Hughes  
Curtis Huttenhower  
Valentin Ilyin  
Lakshmanan Iyer  
Sungchul Ji  
Raja Jothi  
Irwin Jungreis  
Tamer Kahveci  
Manolis Kellis  
Soile Keränen  
Philip Kim  
Debra Knisley  
Jeff Knisley  
Jan Komorowski  
Mark Kon  
Alexander Lachmann  
Jens Lagergren  
Harri Lähdesmäki  
Riitta Lahesmaa  
Celine Lefebvre  
Christina Leslie

Group Leader, University of Leuven  
Lecturer, National Technical University of Athens  
Professor, Columbia University  
Bioinformatics Scientist, Whitehead Institute  
Senior Bioinformatics Scientist, Whitehead Institute  
Associate Professor, University of Pittsburgh  
Assistant Professor In Physics, Northeastern University  
Ast. Prof., New York University / Courant  
Senior Group Leader, Genome Institute of Singapore  
Professor, Washington University School of Medicine  
Assistant Professor, University of Massachusetts Medical School  
Founding Director, Columbia Initiative in Systems Biology  
Senior Lecturer, The Hebrew University of Jerusalem  
Professor, Boston University  
Professor, Michigan State University  
Assistant Professor Of Biology, Boston College  
Deputy Director, Genome Institute of Singapore  
Professor, Boston College  
Professor, Brandeis University  
Senior Lecturer, Auckland Bioengineering Institute  
Adukt, Nencki Institute  
Computational Chemical Biologist, Broad Institute  
Principal Investigator, Fondazione Telethon - TIGEM  
Bioinformaticist, Children's Hospital Boston  
Group Leader, Max Delbrueck Center Berlin  
Research Officer, National Research Council Canada  
Adukt, University of Warsaw  
Visiting Researches, Boston College  
Deputy Director, NCI - Division of Cancer Biology  
Research Associate, Stanford University  
Senior Scientist, US Army Corps of Engineers  
Molecular Biology, Massachusetts General Hospital  
Data Manager, MGH Cancer Center  
Ass. Prof., Graz University of Technology  
Instructor In Neurobiology, Harvard Medical School  
Associate Director, HSPH Bioinformatics Core  
Professor, University of Toronto  
Assistant Professor, Harvard School of Public Health  
Associate Professor, Boston College  
Research Assistant Professor, Tufts University  
Professor, Rutgers University  
Principal Investigator, National Institutes of Health  
Research Scientist, MIT  
Professor, University of Florida  
Associate Professor, MIT / CSAIL / Broad Institute  
Scientist, Lawrence Berkeley National laboratory  
Assistant Professor, University of Toronto  
Professor, East Tenn State Univ  
Associate Professor, East Tennessee State University  
Prof., Linaeus Cent for Bioinformatics  
Professor, Boston University  
Systems Analyst, Mount Sinai School of Medicine  
Professor, KTH  
Professor, Tampere University of Technology  
Director, Professor, Turku Centre for Biotechnology  
Associate Research Scientist, Columbia University  
Assistant Professor, Memorial Sloan-Kettering Cancer Center

stein.aerts@med.kuleuven.be  
leo@mail.ntua.gr  
anastas@ee.columbia.edu  
ibarrasa@wi.mit.edu  
gbell@wi.mit.edu  
benos@pitt.edu  
ginestra.bianconi@gmail.com  
bonneau@nyu.edu  
bourque@gis.a-star.edu.sg  
brent@cse.wustl.edu  
michael.brodsky@umassmed.edu  
califano@c2b2.columbia.edu  
carmel@cc.huji.ac.il  
lecarval@math.bu.edu  
krischan@egr.msu.edu  
chuangj@bc.edu  
clarke@gis.a-star.edu.sg  
clote@bc.edu  
jc@cs.brandeis.edu  
e.crampin@auckland.ac.nz  
m.dabrowski@nencki.gov.pl  
vdancik@broadinstitute.org  
dibernardo@tigem.it  
adibiase@enders.tch.harvard.edu  
christoph.dieterich@mdc-berlin.de  
amira.djebbari@nrc-cnrc.gc.ca  
dojer@mimuw.edu.pl  
idotu@cs.brown.edu  
dg13w@nih.gov  
andrewg@stanford.edu  
ping.gong@us.army.mil  
gopalan@molbio.mgh.harvard.edu  
pgreninger@partners.org  
hubert.hackl@tugraz.at  
david\_harmin@hms.harvard.edu  
ohofmann@hspk.harvard.edu  
t.hughes@utoronto.ca  
chuttenh@hspk.harvard.edu  
ilyin@bc.edu  
lax.iyer@tufts.edu  
sji@rci.rutgers.edu  
jothi@mail.nih.gov  
iljung@csail.mit.edu  
tamer@cise.ufl.edu  
manoli@mit.edu  
svekeranen@lbl.gov  
pm.kim@utoronto.ca  
knisley@etsu.edu  
klejdy@etsu.edu  
jan.komorowski@lcb.uu.se  
mkon@bu.edu  
alexander.lachmann@mssm.edu  
jensl@csc.kth.se  
harri.lahdesmaki@tut.fi  
lahesmaa@idi.harvard.edu  
lefebvre@c2b2.columbia.edu  
cleslie@cbio.mskcc.org

Stuart Levine	Director, MIT BioMicro Center	levine@mit.edu
Fran Lewitter	Director, Bioinformatics & Research Computing, Whitehead Institute	lewitter@wi.mit.edu
Leping Li	Principal Investigator, National Institute of Environmental Health Sci	li3@niehs.nih.gov
Ying Liu	Assistant Professor, The University of Texas at Dallas	ying.liu@utdallas.edu
Vebjorn Ljosa	Computational Biologist, Broad Institute of MIT and Harvard	ljosa@broad.mit.edu
Xinghua Lu	Associate Professor, Medical University of South Carolina	lux@musc.edu
Stefan Maas	Assistant Professor, Lehigh University	swm31@lehigh.edu
Alasdair MacKenzie	Reader, University of Aberdeen	mbi167@abdn.ac.uk
Kathleen Marchal	Group leader, KULeuven	kathleen.marchal@biw.kuleuven.be
Edward Marcotte	Professor, University of Texas	marcotte@icmb.utexas.edu
Steven McCarroll	Assistant Professor, Harvard Medical School	mccarroll@genetics.med.harvard.edu
Jill Mesirov	Associate Director And Chief Informatics Officer, Broad Institute	mesirov@broadinstitute.org
Simon Minovitsky	Software Engineer, jgi/lbl	sminovitsky@lbl.gov
Leonid Mirny	Associate Professor, MIT	leonid@mit.edu
Alexander Mitsos	Assistant Professor, MIT	mitsos@mit.edu
Quaid Morris	Assistant Professor, University of Toronto	quaid.morris@gmail.com
Carol Munro	Senior Lecturer, University of Aberdeen	c.a.munro@abdn.ac.uk
Chad Myers	Assistant Professor, University of Minnesota	cmyers@cs.umn.edu
Pavel Novichkov	Bioinformaticist Project Se, Lawrence Berkeley National Laboratory	psnovichkov@lbl.gov
Matti Nykter	Senior Researcher, Tampere University of Technology	matti.nykter@tut.fi
Uwe Ohler	Assistant Professor, Duke University	uwe.ohler@duke.edu
David Page	Director/Professor/Investigator, Whitehead Institute/MIT/HHMI	page_admin@wi.mit.edu
Larry Parnell	Computational Biologist, JM-USDA Human Nutrition Center Aging	laurence.parnell@ars.usda.gov
Theodore Perkins	Scientist / Asst. Prof., Ottawa Hospital Research Institute	tperkins@ohri.ca
Jonathan Pritchard	Professor, university of chicago human genetics	vwebster@bsd.uchicago.edu
Xiaoning Qian	Assistant Professor, University of South Florida	xiaoning.qian@gmail.com
Ben Raphael	Assistant Professor, Brown University	braphael@brown.edu
Erzsébet Ravasz Regan	Instructor Of Medicine, Beth Israel Deaconess Medical Center	eregan@bidmc.harvard.edu
Mireille Regnier	Research Director, INRIA	mireille.regnier@inria.fr
Christopher Roos	Adjunct Professor, Tampere University of Technology	j.ch.roos@gmail.com
Thomas Sauter	Prof. Dr., University of Luxembourg	thomas.sauter@uni.lu
John Schwacke	Assistant Professor, MUSC	schwacke@musc.edu
Salvatore Sechi	Dr. Proteomics Program, NIH/NIDDK	salvatore_sechi@nih.gov
Kumar Selvarajoo	Asst Professor, Keio University	kumar@ttck.keio.ac.jp
Ron Shamir	Prof., Tel Aviv University	rshamir@tau.ac.il
Xiling Shen	Assistant Professor, Cornell University	xilingsen@ece.cornell.edu
Tomer Shlomi	Lecturer, Technion	tomersh@cs.technion.ac.il
Ilya Shlyakhter	Computational Biologist, Broad Institute	ilya@broad.mit.edu
Nahed Solouma	Associate Professor, Cairo University	nsolouma@k-space.org
Joe Song	Assistant Professor Of Computer Science, New Mexico State Univ.	joemsong@cs.nmsu.edu
Alexander Statnikov	Assistant Professor, NYU	alexander.statnikov@med.nyu.edu
Gustavo Stolovitzky	Manager, Functional Genomics And Systems Biology, IBM Research	gustavo@us.ibm.com
Alice Tay	Assistant Prof, Institute of Molecular and Cell Biology	mcbalice@imcb.a-star.edu.sg
Ronald Taylor	Research Scientist, Pacific Northwest National Laboratory	ronald.taylor@pnl.gov
Hans-Juergen Thiesen	Director, Institute of Immunology	hj.thiesen@gmx.de
Prat Thiru	Bioinformatics Analyst, Whitehead Institute	pthiru@wi.mit.edu
Dennis Vitkup	Professor, Columbia University	dv2121@columbia.edu
Claes Wadelius	Professor, Uppsala University	claes.wadelius@genpat.uu.se
Li Wang	Associate Professor, Beijing Institute of biotechnology	liwang@tsinghua.edu.cn
Katrina Waters	Senior Research Scientist, PNNL	katrina.waters@pnl.gov
David Waxman	Professor, Boston University	djw@bu.edu
James Wetmur	Professor, Mount Sinai School of Medicine	james.wetmur@mssm.edu
Chen-Hsiang Yeang	Assistant Research Fellow, Academia Sinica	chyeang@stat.sinica.edu.tw
Byung-Jun Yoon	Assistant Professor, Texas A&M, Electrical&Computer Engineering	bjoon@ece.tamu.edu
Bingbing Yuan	Bioinformatics Scientist, Whitehead Institute	byuan@wi.mit.edu
Bing Zhang	Assistant Professor, Vanderbilt University	bing.zhang@vanderbilt.edu
Chaoyang Zhang	Associate Professor, University of Southern Mississippi	chaoyang.zhang@usm.edu
Jim Zheng	Assistant Professor, Medical University of South Carolina	zhengw@musc.edu
Sheng Zhong	Assistant Professor, University of Illinois at Urbana-Champaign	szhong@illinois.edu
Jun Zhu	Director, Sage Bionetwork	junzhu_99@yahoo.com

## Industry

John Baker	Sr. Director, Pfizer Inc.	david.baker@pfizer.com
Jadwiga Bienkowska	Principal Scientist, Biogen Idec	jadwiga.bienkowska@biogenidec.com
Andrej Bugrim	Coo, GeneGo, Inc.	andrej@genego.com
Rico Caldo	Bioinformatics Scientist, Monsanto	rico.a.caldo@monsanto.com
Piroska Devay	Research Investigator li, Novartis	piroska.devay@novartis.com
Ben Gordon	Bd Manager / Scientist, Agilent	ben_gordon@agilent.com
Tanwir Habib	Research Scientist, US Army Corp of Engineers	brandy.c.stinson@usace.army.mil
Yizheng Li	Scientist, Wyeth	yizhengli88@gmail.com
Michael Mayo	Research Physicist, US Army Corp of Engineers	brandy.c.stinson@usace.army.mil
Manikandan Narayanan	Sr. Research Scientist, Merck Research Labs	manikandan_narayanan@merck.com
Ravi Pandya	Architect, Microsoft	ravip@microsoft.com
Krzysztof Potempa	Bioinformatician, Lunbeck	krpo@lunbeck.com
Kai Wang	Senior Scientist, Pfizer Inc	kai.wang4@pfizer.com

Ke Xu	Principal Scientist, Bristol-Myers Squibb	ke.xu@bms.com
Peng Yu	Research Scientist, Eli Lilly and Company	yupe@lilly.com
<b>Postdoc</b>		
Phaedra Agius	Research Fellow, MSKCC	phaedragius@gmail.com
Uri David Akavia	Post Doc, Columbia University	uda2001@columbia.edu
Arnaud Amzallag	Collaborateur Scientifique, EPFL SV ISREC UPNAE	arnaud.amzallag@epfl.ch
Mukesh Bansal	Postdoc, Columbia University	mb3113@c2b2.columbia.edu
Alex. Barabanschikov	Postdoc, Northeastern	abarabanschikov@yahoo.com
Sarah Behrens	Postdoc, Computational Molecular Biology, Max Planck Institute	sbehrens@molgen.mpg.de
Jeremy Bellay	Post-Doc Researcher, University of Minnesota	bellay@cs.umn.edu
Doron Betel	Research Fellow, Memorial Sloan-Kettering Cancer Center	betel@cbio.mskcc.org
Elhanan Borenstein	Postdoctoral Fellow, Stanford University	ebo@stanford.edu
Christopher Bristow	Post Doc, MIT	bristow@mit.edu
Mark Brynildsen	Postdoc, HHMI/Department of Biomedical Engineering, BU	mbrynilid@bu.edu
Sarah Calvo	Computational Biologist, Broad Institute	scalvo@broadinstitute.org
Diogo Camacho	Post-Doctoral Fellow, HHMI / BU	camacho@bu.edu
Li-Yun Chang	Postdoc, Obstetrics and Gynecology/National Taiwan University	panlinf@gmail.com
RUI CHANG	Postdoc, University of California San Diego	chang.ru@hotmail.com
Hyonho Chun	Post Doc, Yale University	hyonho.chun@yale.edu
Edward Clark	Research Associate, university of york	edclark@cs.york.ac.uk
Mathieu Clément-Ziza	Post-Doc, Biotec TU-Dresden	mathieu.clement-ziza@biotec.tu-dresden.de
Aedin Culhane	Research Associate, Dana Farber Cancer Institute	aedin@jimmy.harvard.edu
Minghua Deng	Visiting Scholar, Yale University	minghua.deng@yale.edu
Jason Ernst	Postdoctoral Fellow, MIT	jernst@mit.edu
Babak Faryabi	Research Fellow, NIH	faryabib@niaid.nih.gov
Jasmine Foo	Postdoctoral Fellow, Sloan Kettering	jfoo@cbio.mskcc.org
Georg Gerber	Resident, Clinical Pathology, Brigham and Women's Hospital	georg@mit.edu
Gautam Goel	Research Fellow, Massachusetts General Hospital / CCIB	gautam.goel@gmail.com
Raluca Gordon	Postdoctoral Fellow, Brigham & Women's Hospital / Harvard Medical	rgordan@rics.bwh.harvard.edu
Jesse Gray	Postdoc, Harvard Medical School	jgray@foo.net
Yongtao Guan	Postdoc, University of Chicago	ytguan@gmail.com
Attila Gyenesei	Head, Finnish DNA Microarray Centre, Turku Centre Biotechnology	attila.gyenesei@btk.fi
Martin Hemberg	Postdoc, Childrens Hospital Boston	martin.hemberg@childrens.harvard.edu
Jie Hu	Postdoctoral Research Scientist, Columbia University	jiehu@c2b2.columbia.edu
Jieun Jeong	Postdoc, Harvard School of Public Health	jjeong@hsph.harvard.edu
Georgios Kararigas	Postdoctoral Fellow, Charite Medical University	georgios.kararigas@charite.de
Adam Kiezun	Postdoctoral Fellow, Harvard Medical School	akiezun@rics.bwh.harvard.edu
Justin Kinney	Postdoc, Cold Spring Harbor Laboratory	justin.block.kinney@gmail.com
Jay Konieczka	Postdoctoral Fellow, Harvard University/Broad Institute	jkonieczka@mc.b.harvard.edu
ANIL KORKUT	Postdoctoral Research Fellow, MSKCC	akorkut@cbio.mskcc.org
Robert Kueffner	Dr. Ludwig Maximilians University	robert.kueffner@bio.fli.lmu.de
Roshan Kumar	Postdoctoral Associate, Whitehead Institute	roshan@wi.mit.edu
Erik Larsson	Postdoctoral Researcher, Memorial Sloan Kettering Cancer Center	larsson@cbio.mskcc.org
Julia Lasserre	Post-Doc, MPI for Molecular Genetics	julia.lasserre@gmail.com
Kevin Leder	Post Doc, Memorial Sloan Kettering	leder@cbio.mskcc.org
Ho-Joon Lee	Post-Doc, Harvard Medical School	hl129@hms.harvard.edu
Jing Li	Postdoc, Vanderbilt University	jing.li@vanderbilt.edu
Yingchun Liu	Research Scientist, Dana Farber Cancer Institute	yingchun_liu@dfci.harvard.edu
William Lorenz	Postdoctoral Fellow, Boston College	lorenzwi@bc.edu
Kenzie MacIsaac	Research Affiliate, MIT	macisaac@mit.edu
Shaun Mahony	Postdoctoral Associate, Massachusetts Institute of Technology	mahony@mit.edu
William Majoros	Staff Scientist, Duke University	bmajoros@duke.edu
Jessica Mar	Postdoc, Dana-Farber Cancer Institute	jmar@hsph.harvard.edu
Debora Marks	Researcher, harvard	deboramarks@gmail.com
Patrick May	Bioinformatician, Max-Planck-Institute of Molecular Plant Physiology	may@mpimp-golm.mpg.de
Amin Mazloom	Post Doctoral Fellow, Mount Sinai School of Medicine	amin.mazloom@mssm.edu
Paul McGettigan	Bioinformatician, University College Dublin	paul.mcgettigan@ucd.ie
Patricia Menendez	Postdoc, Biometris, Wageningen University	patricia.menendez@wur.nl
Cliff Meyer	Research Scientist, Dana-Farber Cancer Institute	cliff@research.dfci.harvard.edu
Patrick Meyer	Postdoc, Universite Libre de Bruxelles	pmeyer@ulb.ac.be
Magali Michaut	Postdoc, University of Toronto	magali.michaut@utoronto.ca
Tom Michael	Senior Researcher, VIB, Ghent University	tom.michael@psb.vib-ugent.be
Martin Miller	Research Fellow, cBio MSKCC	miller@cbio.mskcc.org
Antonina Mitrofanova	Postdoctoral Cifellow, Columbia University	antonina@c2b2.columbia.edu
Nacho Molina	Post-Doc, EPFL SV ISREC UPNAE	nacho.molina@epfl.ch
Thomas Mullen	Postdoc, Harvard Medical School	temullen@partners.org
Aylwin Ng	Research Fellow (Post-Doc), CCIB, MGH & Harvard	ang@ccib.mgh.harvard.edu
David Nusinow	Postdoctoral Fellow, Brigham and Women's Hospital	dnusinow@partners.org
Peter O'Callaghan	Postdoctoral Associate, Mammalian Cell Engineering, U. Sheffield	p.ocallaghan@sheffield.ac.uk
Fabio Parisi	Postdoctoral Fellow, New York University	fabio.parisi@med.nyu.edu
Mattia Pelizzola	Postdoc, Salk Institute for Biological Studies	mpelizzola@salk.edu
Roger Pique-Regi	Postdoctoral Researcher, University of Chicago	rpique@uchicago.edu
Daniel Pollard	Postdoc, New York University	dpollard@gmail.com
Robert Prill	Postdoc, IBM Research	rprill@us.ibm.com

Shuye Pu Project Manager, The Hospital for Sick Children  
 Saumyadipta Pyne Postdoc Fellow, Broad Institute of MIT & Harvard  
 Arjun Raj Postdoc, MIT  
 Franck Rapaport Post-Doctoral Position, Memorial Sloan-Kettering Cancer Center  
 Kirsi Rautajoki Senior Researcher, Tampere University of Technology  
 Dietmar Rieder Dr., Graz University of Technology  
 Nicolas Robine Postdoc, Sloan Kettering Institute  
 Dmitry Rodionov Postdoc, Inst for Inf Transm Problems, Russian Academy of Sciences  
 Stefano Rossetti Post Doc, Roswell Park Cancer Institute  
 Sushmita Roy Postdoc, UNM  
 Julio Saez-Rodriguez Research Fellow, Harvard Medical School and MIT  
 Debashis Sahoo Instructor, Stanford University  
 Gerald Schwank Scientist, University of Zurich  
 Ron Skupsky Postdoc/Scientist, UC Berkeley  
 Nicola Soranzo Ph.D., CRS4  
 Leila Taher Visiting Fellow, NCBI/NLM/NIH  
 Shang-Kai Tai Postdoctoral Fellow, Academia Sinica  
 Mikko Taipale Postdoctoral Fellow, Whitehead Institute  
 Kun-Nan Tsai Postdoctoral Fellow, Research Center for Emerging Viral Infections  
 Tamir Tuller Post-Doc, Weizmann Institute of Science  
 Andrei Turinsky Research Associate, Hospital for Sick Children  
 Igor Ulitsky Postdoctoral Associate, Whitehead Institute  
 Vladimir Vacic Post-Doctoral Fellow, Cold Spring Harbor Laboratory  
 Golnaz Vahedi Research Fellow, NIH  
 Simon Van Mourik Postdoc, Wageningen University  
 Vanessa Vermeirssen Post-Doc, VIB-Ghent University  
 Weiqing Wang Research Scholar, MSKCC  
 Xiao Wang Research Associate, Boston University  
 Ying Wang Postdoctoral Scholar, University of Chicago  
 Dana Wohlbach Research Associate, University of Wisconsin-Madison  
 Damian Wojtowicz Visiting Fellow, National Institutes of Health, NLM/NCBI  
 Kyoung Jae Won Postdoc, UCSD  
 Eric Wooten Postdoctoral Fellow, Tufts Medical Center  
 Chang-Jiun Wu Postdoctoral Fellow, Biomedical Engineering, Boston University  
 Zhi Xie Postdoctoral Research Fellow, Johns Hopkins University  
 Yurong Xin Postdoc, Columbia University  
 Wei Xu Postdoc Fellowship, Switzerland Federal Institute  
 Muhammed Yildirim Postdoctoral Associate, Whitehead Institute for Biomedical Research  
 Hossein Zare Research Fellow, NIH  
 Alon Zaslaver Postdoc, Caltech  
 Jie Zheng Postdoc, NCBI/NLM/NIH  
 Quan Zhong Research Fellow, Harvard Medical School

## Student

Anton Aboukhalil Grad Student, MIT/Bulyk Lab  
 Marit Ackermann PhD Student, TU Dresden, Biotec  
 Kaisa-Leena Aho Student, Tampere University of Technology  
 Tommi Aho Student, Tampere University of Technology  
 Malika Aid PhD Student, IRIC Montreal University  
 Tarmo Aijo Student, Tampere University of Technology  
 Robert Altschuler Graduate Student, MIT  
 MoChamad Apri PhD, Biometris, Wageningen University  
 Manonmani Arunachalam PhD Student, Max Planck Institute, Molecular Cell Biology / Genetics  
 Aaron Arvey Grad Student, Sloan Kettering  
 Gowtham Atluri Graduate Student, University of Minnesota  
 Ferhat Ay Research Assistant, University of Florida  
 Anastasia Baryshnikova PhD Student, University of Toronto  
 Vincenzo Belcastro Student, Fondazione Telethon - TIGEM  
 Tomer Benyamini Student, Tel Aviv University  
 Jonathan Bieler PhD, EPFL SV UPNAE  
 Eugene Bolotin Graduate Researcher, University of California Riverside  
 Blake Borgeson Student, Rice University  
 Rogerio Candeias Grad Student, MIT  
 Matthew Carson Graduate Student, UIC  
 Chaang-Ray Chen Graduate Student, National Tsing Hua University  
 Kristen Cook PhD Candidate, Harvard University  
 Elissa Cosgrove Graduate Student, Boston University  
 Ross Curtis PhD Student, Carnegie Mellon University  
 Ertugrul Dalkic PhD Student, MSU  
 Matthew Davis Doctoral Candidate, UC Berkeley  
 Jacob Degner Student, University of Chicago  
 Raamesh Deshpande Graduate Student, University of Minnesota - Twin Cities  
 Lei Du Dr., Harbin Medical University  
 Jonathan Ellis Post Doc, University of Queensland  
 Timo Erkkilä Researcher, Tampere University of Technology

shuyepu@sickkids.ca  
 spyne@broad.mit.edu  
 arjunrajlab@gmail.com  
 rapaport@cbio.mskcc.org  
 kirsi.rautajoki@tut.fi  
 dietmar.rieder@tugraz.at  
 robinen@mskcc.org  
 rodionov@iitp.ru  
 stefano.rossetti@roswellpark.org  
 sroy@cs.umn.edu  
 julio@hms.harvard.edu  
 saahoo@stanford.edu  
 gerald.schwank@molbio.uzh.ch  
 skupskyr@gmail.com  
 soranzo@crs4.it  
 taherl@ncbi.nlm.nih.gov  
 sktai@stat.sinica.edu.tw  
 taipale@wi.mit.edu  
 knitsai@mail.cgu.edu.tw  
 tamirtull@post.tau.ac.il  
 turinsky@sickkids.ca  
 ulitsky@wi.mit.edu  
 vacic@cshl.edu  
 vahedig@mail.nih.gov  
 simon.vanmourik@wur.nl  
 vanessa.vermeirssen@psb.vib-ugent.be  
 wqwang@cbio.mskcc.org  
 xiaow@bu.edu  
 ygwang@uchicago.edu  
 dana.wohlbach@gmail.com  
 wojtowda@ncbi.nlm.nih.gov  
 kwon@ucsd.edu  
 ewooten@tuftsmedicalcenter.org  
 terrence@bu.edu  
 xiezhi@gmail.com  
 xinyuro@pi.cpmc.columbia.edu  
 wei.xu@epfl.ch  
 yildirim@wi.mit.edu  
 hzare@mail.nih.gov  
 alonzo@caltech.edu  
 zhengj@ncbi.nlm.nih.gov  
 quan\_zhong@dfic.harvard.edu

anton1@mit.edu  
 marit.ackermann@biotec.tu-dresden.de  
 kaisa-leena.aho@tut.fi  
 tommi.aho@tut.fi  
 malika.aid@umontreal.ca  
 tarmo.ajio@tut.fi  
 raltshul@mit.edu  
 mochamad.apri@wur.nl  
 arunacha@mpi-cbg.de  
 aarvey@cbio.mskcc.org  
 gowtham@cs.umn.edu  
 fay@cise.ufl.edu  
 a.baryshnikova@utoronto.ca  
 belcastro@tigem.it  
 tomerbe1@post.tau.ac.il  
 sophie.aquilar@epfl.ch  
 ybolo001@student.ucr.edu  
 blake.borgeson@gmail.com  
 candeias@mit.edu  
 mcarso2@uic.edu  
 d948504@oz.nthu.edu.tw  
 kcook@fas.harvard.edu  
 ejburk@bu.edu  
 rcurtis@andrew.cmu.edu  
 dalkicer@msu.edu  
 matthewdavis@berkeley.edu  
 jdegner@uchicago.edu  
 rdeshpand@cs.umn.edu  
 dostone@gmail.com  
 j.ellis2@uq.edu.au  
 timo.p.erkkila@tut.fi

Gang Fang	PhD Student, University of Minnesota	gangfang@cs.umn.edu
Xin Feng	Student, Cold Spring Harbor Lab, Stony Brook University	drestion@gmail.com
Andrew Fox	PhD Student, Tufts University	andrew.fox@tufts.edu
Sarah Gilman	Student, Columbia University	srq2104@columbia.edu
Alex Greenfield	Graduate Student, New York University	agreenf1@gmail.com
Brian Haynes	Graduate Student, Washington University	bch2@cec.wustl.edu
Andrea Hodgins-Davis	Graduate Student, Yale University	andrea.hodgins-davis@yale.edu
Ping Hou	PhD Candidate, New Mexico State University	hou.zhaoping@gmail.com
Daniel Hurley	PhD Student, Auckland Bioengineering Institute	d.hurley@auckland.ac.nz
Van Anh Huynh-Thu	PhD Student, University of Liege	vahuyhn@ulg.ac.be
Francesco Iorio	Student, Fondazione Telethon - TIGEM	iorio@tigem.it
Hosna Jabbari	Ph.D. Student, University of British Columbia	hjabbari@cs.ubc.ca
Anders Jacobsen	PhD Student, Copenhagen University	andersbj@binf.ku.dk
In Sock Jang	PhD Candidate, Columbia University	ij2113@columbia.edu
Shuiwang Ji	Doctoral Student, Arizona State University	shuiwang.ji@asu.edu
Anagha Joshi	PhD Student, VIB-Gent University	anagha.joshi@psb.vib-ugent.be
Thadeous Kacmarczyk	PhD Candidate, New York University	tjk229@nyu.edu
Sabah Kadri	Graduate Student, Carnegie Mellon University	sskadri@andrew.cmu.edu
JIA KANG	Graduate Student, YALE UNIVERSITY	jia.kang@yale.edu
Rosa Karlic	PhD Student, Max Planck Institute for Molecular Genetics	karlic@molgen.mpg.de
Önder Kartal	PhD Student, Max-Planck-Institute of Molecular Plant Physiology	kartal@pimp-golm.mpg.de
Poorvi Kaushik	Graduate Student, Sloan Kettering Institute	pkaushik@cbio.mskcc.org
Hilal Kazan	PhD Student, University of Toronto	hilal@cs.toronto.edu
Aly Khan	Student, Sloan Kettering	aakhan@cbio.mskcc.org
Pouya Kheradpour	Graduate Student, MIT	pouyak@mit.edu
Ah-Ram Kim	Graduate Student, Stony Brook University	banwise@gmail.com
Chris King	Research Assistant / Graduate Student, University of Washington	chrisk1@uw.edu
Francis Kirigin	Bioinformatics Software Engineer, Genomics Systems Biology, NYU	fkirigin@nyu.edu
Virpi Kivinen	Student, Tampere University of Technology	virpi.kivinen@tut.fi
Niels Klitgord	PhD Student, BU Bioinformatics	klitgord@gmail.com
Deniz Kural	PhD Student, Boston College	kural@bc.edu
Asta Laiho	Bioinformatics Research Engineer, University of Turku	asta.laiho@btb.fi
Sandro Lambeck	PhD Student, Hans Knoell Institute	sandro.lambeck@hki-jena.de
Antti Larjo	Student, Tampere University of Technology	antti.larjo@tut.fi
Kirsti Laurila	Student, Tampere University of Technology	kirsti.laurila@tut.fi
Dongwon Lee	Graduate Student, Johns Hopkins University	dwlee@jhu.edu
Max Leiserson	Student, Tufts University	max.leiserson@gmail.com
Manway Liu	Graduate Student, Boston University, BME	manway@bu.edu
Benjamin Logsdon	Graduate Student, Cornell University	bal47@cornell.edu
Hong Lu	Research Assistant, Iowa State Univ.	luhong@iastate.edu
Eugenia Lyashenko	PhD Student, Columbia University	el2378@columbia.edu
Nikita Lytkin	Student, NYU	nikita.lytkin@med.nyu.edu
Haisu Ma	Graduate Student, Yale University	haisu.ma@yale.edu
Daniel Marbach	PhD Student, Swiss Federal Institute of Technology in Lausanne	daniel.marbach@gmail.com
Andre Martins	PhD Student, Cornell University	alm253@cornell.edu
Lucia Marucci	Student, Fondazione Telethon - TIGEM	marucci@tigem.it
Rachel McCord	Graduate Student, Harvard University	rpmccord@fas.harvard.edu
Pieter Meysman	PhD Student, K.U.Leuven	pieter.meysman@biw.kuleuven.be
Renqiang Min	Mr., Dept Computer Science, University of Toronto	minrq@cs.toronto.edu
Patrycja Misziuro	PhD Graduate, MIT Whitehead	patrycja@mit.edu
Evan Molinelli	Graduate Student, Sloan Kettering Institute	molinelli@cbio.mskcc.org
Melissa Morine	PhD Student, University College Dublin	melissa.morine@ucd.ie
Varun Narendra	Medical Student, New York University	varun.narendra@med.nyu.edu
Fiona Nielsen	PhD Student, CMBI, NCMLS	fnielsen@cmbi.ru.nl
Oaz Nir	Hst/Math, MIT	oaz@mit.edu
Brian Olson	Graduate Research Assistant, George Mason University	bolson3@gmu.edu
Zhengyu Ouyang	Student, New Mexico State University	ouyong@nmsu.edu
Gaurav Pandey	PhD Candidate, University of Minnesota	gaurav@cs.umn.edu
Bernardo Pando	Graduate Student, Massachusetts Institute of Technology	bpando@mit.edu
Georgios Papachristoudis	Student, MIT	geopapa@mit.edu
Yongjin Park	Graduate Student, Johns Hopkins University	ypark28@jhu.edu
Tal Peleg	Student, Tel-Aviv University	stalpl@gmail.com
Hedi Peterson	PhD Student, University of Tartu / Quretec Ltd	peterson@quretec.com
Matias Piipari	Graduate Student, Wellcome Trust Sanger Institute	matias.piipari@gmail.com
German Plata	Graduate Student, Columbia University	gap2118@columbia.edu
Paz Polak	PhD Student, Max Planck Institute for Molecular Genetics	polak@molgen.mpg.de
Elizabeth Rach	PhD Student, Duke University	elizabeth.rach@duke.edu
Pradipta Ray	Graduate Student, Carnegie Mellon University	pray@cs.cmu.edu
Christopher Reeder	Graduate Student, CSAIL - MIT	ccr@csail.mit.edu
Jonathan Reichel	PhD Student, Cornell/Sloan-Kettering	jbreichel@gmail.com
Jüri Reimand	PhD Fellow, University of Tartu	thcmob@ut.ee
Sarah Richardson	Graduate Student, Johns Hopkins University	notadocor@jhmi.edu
Giovanni Russo	PhD Student, University of Naples Federico II	giovanni.russo2@unina.it
Ana Paula Sales	Graduate Student, Duke University	ad44@duke.edu
Sebastian Schultheiss	Graduate Student, Friedrich Miescher Laboratory, Max Planck Society	sebi@tuebingen.mpg.de

Rachel Sealfon  
Yasin Senbabaoglu  
Preethi Srinivasan  
Valerie Storms  
Aarathi Sugathan  
Pablo Tamayo  
Diana Tatar  
Alexander Tsankov  
Amit Upadhyay  
Francesco Vallania  
Benjamin VanderSluis  
Fabio Vandin  
Eric Venner  
Basilio Vescio  
Joel Wagner  
Yedael Waldman  
Peter Waltman  
Haizhou Wang  
Xuewei Wang  
Todd Wasson  
Aaron Wenger  
James White  
Ming Wu  
Qinghong Yan  
Antti Ylipää  
Jiyang Yu  
Lu Zhang  
Xianghua Zhang  
Xiuwei Zhang  
Yue Zhao  
Xu Zhou

Graduate Student, MIT  
Student, University of Michigan  
Student, Boston College  
PhD Student, KULeuven  
Graduate Student, Boston University  
Student, Broad Institute  
Graduate Student, Tufts University  
Graduate Student, MIT/Broad Institute  
Student, Boston College  
Graduate Student, Washington University in Saint Louis  
Graduate Student, University of Minnesota  
Ph.D. Student, University of Padova  
Graduate Student, Baylor College of Medicine  
Ph.D. Student, Magna Graecia University of Catanzaro  
Graduate Student, MIT  
PhD Student, Tel Aviv University  
Student, New York University  
Student, New Mexico State University  
Ra, Michigan State University  
Graduate Student, Duke University  
Graduate Student, Stanford Univ  
PhD Candidate, University of Maryland - College Park  
Student, Michigan State University  
Graduate Student, Stony Brook University  
Student, Tampere University of Technology  
PhD Student, Columbia University  
Graduate Student, Boston College  
Student, Yale University  
PhD Student, Swiss Federal Institute of Technology  
Graduate Student, Washington University in St. Louis  
Graduate Student, Harvard University

rsealfon@mit.edu  
shenbaba@gmail.com  
preethi.srinivasan@gmail.com  
valerie.storms@esat.kuleuven.be  
aarathi@bu.edu  
tamayo@broadinstitute.org  
diana.tatar@tufts.edu  
atsankov@mit.edu  
amitupadhyay86@gmail.com  
fvallani@artsci.wustl.edu  
bvander@cs.umn.edu  
vandinfa@cs.brown.edu  
venner@bcm.edu  
b.vescio@unicz.it  
jpw@mit.edu  
yedael@gmail.com  
peter.waltman@gmail.com  
seaboat.wang@gmail.com  
xuewei@msu.edu  
tsw5@duke.edu  
awenger@stanford.edu  
whitej@umd.edu  
wuming1@msu.edu  
qyan@ic.sunysb.edu  
antti.ylipaa@tut.fi  
jy2322@c2b2.columbia.edu  
zhanglv@bc.edu  
xianghua.zhang@yale.edu  
xiuwei.zhang@epfl.ch  
yue.zhao@wustl.edu  
xzhou@fas.harvard.edu

#### Volunteer

Mary Addonizio  
Fadhl Al-Akwaa  
Adam Callahan  
Betsy Chang  
Yang Ding  
Matt Edwards  
Chuck Epstein  
Dina Faddah  
Charlie Frogner  
Anthony Gitter  
Loyal Goff  
Taran Gujral  
Christina Harview  
Peter Huggins  
Hai-Son Le  
Tien-ho Lin  
Yan Meng  
Yeison Rodriguez  
Erroll Rueckert  
Mark Sevecka  
Tal Shay  
Alejandro Wolf-Yadlin  
Shan Zhong  
Guy Zinman

Project Manager, Broad Institute  
Student, CU  
Data Analyst I, Broad Institute  
Graduate Student, Mount Sinai School of Medicine  
Graduate Student, Boston College  
Graduate Student, MIT  
Scientist, The Broad Institute  
Graduate Student, MIT/Whitehead Institute  
Graduate Student, Massachusetts Institute of Technology  
Student, Carnegie Mellon University  
Postdoctoral Fellow, MIT/Broad  
Postdoc, Harvard University  
Lab Technician, Broad Institute of Harvard/MIT  
Postdoctoral Fellow, Carnegie Mellon University  
Graduate Student, Carnegie Mellon University  
Ph.D. Student, Carnegie Mellon University  
Senior Biostatistician, Broad Institute  
Research Assistant, New York University  
Postdoc, Stanley center at the Broad  
Post-Doctoral Associate, Whitehead Institute  
Postdoc, Broad  
Post-Doc, Harvard University  
PhD Student, Carnegie Mellon University  
PhD Student, Carnegie Mellon University

mary@broadinstitute.org  
fadlmaster1@yahoo.com  
callahan@broadinstitute.org  
betsy.chang@mssm.edu  
dingyc@bc.edu  
matted@mit.edu  
epstein@broadinstitute.org  
dafaddah@mit.edu  
frogner@mit.edu  
agitter@cs.cmu.edu  
lgooff@csail.mit.edu  
gujral@chemistry.harvard.edu  
charview@broadinstitute.org  
phuggins@andrew.cmu.edu  
hple@cs.cmu.edu  
thlin@cs.cmu.edu  
ymeng@broadinstitute.org  
yeison.rodriguez@nyu.edu  
rueckert@broadinstitute.org  
sevecka@wi.mit.edu  
talshay@broadinstitute.org  
alewolf@gmail.com  
szhong@andrew.cmu.edu  
zinman@cs.cmu.edu

#### Press

Orli Bahcall  
Clare Garvey  
Joanne Kotz  
Craig Mak  
Lara Szewczak

Senior Editor, Nature Genetics  
Editor, Genome Biology  
Editor, Nature Chemical Biology  
Associate Editor, Nature Biotechnology  
Scientific Editor, Cell - Cell Press

o.bahcall@natureny.com  
clare.garvey@genomebiology.com  
j.kotz@boston.nature.com  
c.mak@us.nature.com  
lszewczak@cell.com







# RECOMB REGULATORY GENOMICS, SYSTEMS BIOLOGY, AND DREAM4

MIT / BROAD INSTITUTE  
DEC 2-6, 2009

compbio.mit.edu/recombsat

## CONFERENCE CHAIRS:

**MANOLIS KELLIS**  
**ZIV BAR-JOSEPH**  
**ANDREA CALIFANO**  
**GUSTAVO STOLOVITZKY**



Wednesday, Dec 2	
3pm	Conference check-in open, Poster session 1 set-up
5pm	Welcome Remarks
9:15	Mark Bogdan - Evidence for Quantitative Transcription Networks
9:45	Stuart E. Schulz - Evidence for a transcriptional repressor by iDRFs: implications for viral and disease
10:15	Michael Collins - Binding Preference of RNA-binding Proteins from Yeast 3' UTRs
10:45	Michael Collins - Towards Computational Prediction of MicroRNA Function and Activity
11:15	Break / Light snacks
11:45	Colin A. Meyer - Inferring Transcription Regulation using iDRFs
12:15	Colin A. Meyer - Discovery and characterization of chromatin marks from combinatorial histone marks
12:45	Michael Collins - Hetero-DNA methylation as single-base resolution marks
1:15	Michael Collins - Fundamentally different strategies of gene regulation in bacteria and eukaryotes
1:45	Bob Waterston - Deciphering the C. Elegans Embryonic Regulatory Network
8:15-9:45	Regulatory Genomics Welcome Reception: Poster Session 1 (non-abstracts, snacks, refreshments, wine, cash bar)

Thursday, Dec 3	
9am	Breakfast
9:30	Rick Young - Programming Cell State
9:45	James J. Gray - Widespread POU recruitment and transcription of enhancers during cellular differentiation
10:15	William Borner - Inferring binding energy from sequence binding sites
10:45	William Borner - Genome-wide binding site turnover in the Core Eukaryotic Genome
11:15	Colin A Meyer - FRAP
11:45	Michael Collins - Identification and analysis of de-regulatory elements based on computational genomics
12:15	Michael Collins - A novel algorithm for the Exact Calculation of Partition Functions
12:45	Michael Collins - A high-level model for analysis of transcription factor interactions
1:15	Michael Collins - Identifying and analyzing de-regulatory elements
1:45	Michael Collins - Regulatory motifs associated with feeding behavior in Drosophila
1:55	Michael Collins - Edge prediction models of human inferred diseases
12pm	Lunch Break / Networking Opportunities
1pm	Manolis Kellis - Evolution of Nucleosome Positioning
1:30	Manolis Kellis - Structural DNA motifs for the prediction of sequence motifs such as conditional random fields
1:45	Manolis Kellis - Nucleosomes are positioned in eucore and have distinct motifs suggesting a transcriptional regulatory role
2pm	Manolis Kellis - The Human HNF4A Target Class: Genes, Protein Binding, MicroRNAs
2:15	Colin A Meyer - Coffee Snacks / Fruit Breaks - Poster Set-Up for Session 1
2:45	Michael Collins - Genomes under regulatory and computational analysis of motifs: DNA Motifs in vivo
3pm	Elizabeth A. Riedl - Landscape of Transcription Initiation in Drosophila
3:15	Manolis Kellis - Transcription Initiation in Drosophila: Motifs
3:30	Manolis Kellis - TSS detection helps predict promoters
3:45	Manolis Kellis - SPICE - Signaling Pathways Knowledge-base and Analysis
3:55	Break / Light Snacks / Poster Session 1 (non-abstracts, snacks, refreshments)
4:15	Michael Collins - Systems Biology of DNA Damage and Repair
4:45	Qian Zhang - Effective Identification of Conserved Pathways in Biological Networks
5pm	Michael Collins - Evaluating strengths and weaknesses of methods for network inference
5:15	Michael Collins - Invariant Motifs in HNF and GCOM1 in Liver and Lung: Equilibria of Postulation, Gene Expression, and iDRF
5:30	Michael Collins - DRFNet: A new and powerful approach to identifying motifs from large-scale gene expression profiles
5:45	5:45 - 5:50 Welcome Remarks
6pm	Nevan J. Krogan - Genomic Insights from Protein-Protein and Genetic Interaction Maps
6:30	Tommy Burdakov - Molecular Flux Balance Analysis with Context-Specific Priors
6:45	Michael Collins - Dynamic networks from hierarchical Bayesian graph learning
7pm	Break / Light snacks
7:15	Michael Collins - How to turn a genetic circuit into a synthetic network
7:45	Michael Collins - An ensemble Model of Competitive Multi-Substrate Enzymes
7:55	Manolis Kellis - An efficient and exhaustive approach for modular decomposition of global genetic interaction networks
8:15	Justin Kim - Regulatory Proteins from DNA-sequence data
8:30	Erna Liebermann-Liebowitz - Consensus of the Tactile Genome: a model for chromatinopathies in the nucleus
8:15-9:45	Dinner out on the town

Friday, Dec 4	
9am	Breakfast
9:30	Scott Waldrop - Transcriptional regulatory networks: from development to disease
9:45	James J. Gray - Global Analysis of Human Protein-DNA Interactions by Chromatin ChIP
10:15	Manolis Kellis - Structural network clustering of multiple gene expression and physical interaction datasets
10:45	Michael Collins - Learning the structure of protein transcription factor networks
11:15	William Borner - Welcome / DREAM Registration
11:45	Michael Collins - Characterizing regulatory motifs by the path way method
12:15	Michael Collins - Functional analysis of mammalian signal transduction pathways
12:45	Michael Collins - A novel algorithm for the Exact Calculation of Partition Functions
1:15	Michael Collins - A high-level model for analysis of transcription factor interactions
1:45	Michael Collins - Identifying and analyzing de-regulatory elements
1:55	Michael Collins - Edge prediction models of human inferred diseases
12pm	Lunch Break / Networking Opportunities
1pm	Erna Liebermann-Liebowitz - Welcome Reception: Poster Session 1 (non-abstracts, snacks, refreshments)
1:30	Michael Collins - Characterizing regulatory motifs by the path way method
1:45	Michael Collins - Functional analysis of mammalian signal transduction pathways
1:55	Michael Collins - A novel algorithm for the Exact Calculation of Partition Functions
2:15	Michael Collins - A high-level model for analysis of transcription factor interactions
2:45	Michael Collins - Identifying and analyzing de-regulatory elements
2:55	Michael Collins - Edge prediction models of human inferred diseases
3pm	Robert J. Egel - Challenges I and 2: overall results
3:15	Robert J. Egel - Challenges 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100
3:55	Short Breaks - 45 min poster set-up
4:15	Michael Collins - Welcome Remarks
4:45	Michael Collins - Systems Biology of DNA Damage and Repair
5pm	Michael Collins - Evaluating strengths and weaknesses of methods for network inference
5:15	Michael Collins - Invariant Motifs in HNF and GCOM1 in Liver and Lung: Equilibria of Postulation, Gene Expression, and iDRF
5:30	Michael Collins - DRFNet: A new and powerful approach to identifying motifs from large-scale gene expression profiles
5:45	5:45 - 5:50 Welcome Remarks
6pm	Nevan J. Krogan - Genomic Insights from Protein-Protein and Genetic Interaction Maps
6:30	Tommy Burdakov - Molecular Flux Balance Analysis with Context-Specific Priors
6:45	Michael Collins - Dynamic networks from hierarchical Bayesian graph learning
7pm	Break / Light snacks
7:15	Michael Collins - How to turn a genetic circuit into a synthetic network
7:45	Michael Collins - An ensemble Model of Competitive Multi-Substrate Enzymes
7:55	Manolis Kellis - An efficient and exhaustive approach for modular decomposition of global genetic interaction networks
8:15	Justin Kim - Regulatory Proteins from DNA-sequence data
8:30	Erna Liebermann-Liebowitz - Consensus of the Tactile Genome: a model for chromatinopathies in the nucleus
8:15-9:45	Dinner out on the town

Saturday, Dec 5	
9am	Breakfast
9:30	J.M.D. Boerj - Building Saccharomyces cerevisiae v2.0: The Synthetic Yeast Genome Project
9:45	Manolis Kellis - Decoding small RNA networks in bacteria
10:15	Manolis Kellis - Predicting synthetic environments by linkage disequilibrium
10:45	Manolis Kellis - Large-scale learning of cellular phenotypes from microarray data
11:15	Colin A Meyer - Coffee Snacks / Fruit Break
11:45	Manolis Kellis - Bayesian Design of Assemblies: Modular Synthetic Circuits
12:15	Manolis Kellis - Reconstructing Ancestral Gene Content by Co-Estimation
12:45	Closing remarks and announcement of next year's venue
1:15	Closing remarks and afternoon Poster Take-down
2:15	Coffee Snacks / Fruit Breaks == Poster Set-Up for Session II
2:45	J.M.D. Boerj - Network-based Inference of Knockout Effects in Yeast
3pm	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:15	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:30	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:45	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:55	Short Breaks - 45 min poster set-up
4:15	Michael Collins - Welcome Remarks
4:45	Michael Collins - Systems Biology of DNA Damage and Repair
5pm	Michael Collins - Evaluating strengths and weaknesses of methods for network inference
5:15	Michael Collins - Invariant Motifs in HNF and GCOM1 in Liver and Lung: Equilibria of Postulation, Gene Expression, and iDRF
5:30	Michael Collins - DRFNet: A new and powerful approach to identifying motifs from large-scale gene expression profiles
5:45	5:45 - 5:50 Welcome Remarks
6pm	Nevan J. Krogan - Genomic Insights from Protein-Protein and Genetic Interaction Maps
6:30	Tommy Burdakov - Molecular Flux Balance Analysis with Context-Specific Priors
6:45	Michael Collins - Dynamic networks from hierarchical Bayesian graph learning
7pm	Break / Light snacks
7:15	Michael Collins - How to turn a genetic circuit into a synthetic network
7:45	Michael Collins - An ensemble Model of Competitive Multi-Substrate Enzymes
7:55	Manolis Kellis - An efficient and exhaustive approach for modular decomposition of global genetic interaction networks
8:15	Justin Kim - Regulatory Proteins from DNA-sequence data
8:30	Erna Liebermann-Liebowitz - Consensus of the Tactile Genome: a model for chromatinopathies in the nucleus
8:15-9:45	Regulatory Genomics Welcome Reception: Poster Session 1 (non-abstracts, snacks, refreshments, wine, cash bar)

Sunday, Dec 6	
9am	Breakfast
9:30	J.M.D. Boerj - Building Saccharomyces cerevisiae v2.0: The Synthetic Yeast Genome Project
9:45	Manolis Kellis - Decoding small RNA networks in bacteria
10:15	Manolis Kellis - Predicting synthetic environments by linkage disequilibrium
10:45	Manolis Kellis - Large-scale learning of cellular phenotypes from microarray data
11:15	Colin A Meyer - Coffee Snacks / Fruit Break
11:45	Manolis Kellis - Bayesian Design of Assemblies: Modular Synthetic Circuits
12:15	Manolis Kellis - Reconstructing Ancestral Gene Content by Co-Estimation
12:45	Closing remarks and announcement of next year's venue
1:15	Closing remarks and afternoon Poster Take-down
2:15	Coffee Snacks / Fruit Breaks == Poster Set-Up for Session II
2:45	J.M.D. Boerj - Network-based Inference of Knockout Effects in Yeast
3pm	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:15	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:30	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:45	Manolis Kellis - Consistent Framework to detect drivers of transcriptional changes in large-scale perturbation data
3:55	Short Breaks - 45 min poster set-up
4:15	Michael Collins - Welcome Remarks
4:45	Michael Collins - Systems Biology of DNA Damage and Repair
5pm	Michael Collins - Evaluating strengths and weaknesses of methods for network inference
5:15	Michael Collins - Invariant Motifs in HNF and GCOM1 in Liver and Lung: Equilibria of Postulation, Gene Expression, and iDRF
5:30	Michael Collins - DRFNet: A new and powerful approach to identifying motifs from large-scale gene expression profiles
5:45	5:45 - 5:50 Welcome Remarks
6pm	Nevan J. Krogan - Genomic Insights from Protein-Protein and Genetic Interaction Maps
6:30	Tommy Burdakov - Molecular Flux Balance Analysis with Context-Specific Priors
6:45	Michael Collins - Dynamic networks from hierarchical Bayesian graph learning
7pm	Break / Light snacks
7:15	Michael Collins - How to turn a genetic circuit into a synthetic network
7:45	Michael Collins - An ensemble Model of Competitive Multi-Substrate Enzymes
7:55	Manolis Kellis - An efficient and exhaustive approach for modular decomposition of global genetic interaction networks
8:15	Justin Kim - Regulatory Proteins from DNA-sequence data
8:30	Erna Liebermann-Liebowitz - Consensus of the Tactile Genome: a model for chromatinopathies in the nucleus
8:15-9:45	Regulatory Genomics Welcome Reception: Poster Session 1 (non-abstracts, snacks, refreshments, wine, cash bar)

