

Alzheimer's disease: early alterations in brain DNA methylation at *ANK1*, *BIN1*, *RHBDF2* and other loci

Philip L De Jager¹⁻³, Gyan Srivastava^{1,3}, Katie Lunnon^{4,5}, Jeremy Burgess^{6,7}, Leonard C Schalkwyk^{4,5}, Lei Yu⁸, Matthew L Eaton^{3,9}, Brendan T Keenan^{1,3}, Jason Ernst^{3,9}, Cristin McCabe³, Anna Tang¹, Towfique Raj¹⁻³, Joseph Replogle¹⁻³, Wendy Brodeur¹⁰, Stacey Gabriel¹⁰, High S Chai^{6,7}, Curtis Younkin⁶, Steven G Younkin⁶, Fanggeng Zou⁶, Moshe Szyf¹¹, Charles B Epstein¹², Julie A Schneider⁸, Bradley E Bernstein^{2,12,13}, Alex Meissner^{9,12,14}, Nilufer Ertekin-Taner^{6,7}, Lori B Chibnik¹⁻³, Manolis Kellis^{3,9}, Jonathan Mill^{4,5} & David A Bennett⁸

We used a collection of 708 prospectively collected autopsied brains to assess the methylation state of the brain's DNA in relation to Alzheimer's disease (AD). We found that the level of methylation at 71 of the 415,848 interrogated CpGs was significantly associated with the burden of AD pathology, including CpGs in the *ABCA7* and *BIN1* regions, which harbor known AD susceptibility variants. We validated 11 of the differentially methylated regions in an independent set of 117 subjects. Furthermore, we functionally validated these CpG associations and identified the nearby genes whose RNA expression was altered in AD: *ANK1*, *CDH23*, *DIP2A*, *RHBDF2*, *RPL13*, *SERPINF1* and *SERPINF2*. Our analyses suggest that these DNA methylation changes may have a role in the onset of AD given that we observed them in presymptomatic subjects and that six of the validated genes connect to a known AD susceptibility gene network.

Evidence is emerging that DNA methylation levels at certain CpG dinucleotides can be both highly variable across individuals and stable over time in an individual^{1,2}. This suggests that differences in DNA methylation in certain loci could be correlated with the life experiences of a given individual, such as a disease risk factor or a diagnosis³. The most compelling evidence that the epigenome may influence AD comes from the manipulation of histone deacetylases (HDAC) in model systems of AD and in off-label treatment of AD patients with HDAC inhibitors^{4,5}. To date, results of epigenomic studies of AD are sometimes conflicting and have not yet returned robust associations⁶⁻⁹. Furthermore, there is a gradual increase in methylation at many sites throughout the genome with increasing age that has to be carefully considered when studying AD^{8,10-12}. We carried out a statistically rigorous gene discovery effort to identify regions of the genome that were differentially methylated in relation to the burden of AD neuropathology. We found that 71 discrete CpG dinucleotides in the human genome exhibited altered DNA methylation levels in relation to AD, that these changes were an early feature of AD, that the transcription of genes found in these differentially methylated regions was also independently associated with AD pathology and that these differentially expressed

genes connected to a previously reported genetically defined AD susceptibility network¹³.

RESULTS

Description of subjects and data

Our data set consisted of methylation measures at 415,848 discrete CpG dinucleotides in 708 subjects. These methylation profiles were generated using the Illumina HumanMethylation450 beadset and a sample of dorsolateral prefrontal cortex obtained from each individual. Given that we dissected out the gray matter from each sample, we profiled a piece of tissue composed primarily of different neuronal populations and other parenchymal cells such as glia. These subjects were part of the Religious Order Study (ROS) or the Memory and Aging Project (MAP), two prospective cohort studies of aging that include brain donation at the time of death. Given that the subjects were cognitively non-impaired at study entry, we studied a random selection of the older population. Over the course of the study, some subjects declined cognitively and displayed a range of amyloid pathology burden at the time of death, with 60.8% of subjects meeting a pathologic diagnosis of AD¹⁴ (**Supplementary Table 1a**). To technically validate the nature of our data, we compared our

¹Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Departments of Neurology and Psychiatry, Brigham and Women's Hospital, Boston, Massachusetts, USA. ²Harvard Medical School, Boston, Massachusetts, USA. ³Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. ⁴University of Exeter Medical School, University of Exeter, Exeter, UK. ⁵Institute of Psychiatry, King's College London, London, UK. ⁶Department of Neuroscience, Mayo Clinic, Jacksonville, Florida, USA. ⁷Department of Neurology, Mayo Clinic, Jacksonville, Florida, USA. ⁸Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, USA. ⁹Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. ¹⁰Genetic Analysis Platform, Broad Institute, Cambridge, Massachusetts, USA. ¹¹Department of Pharmacology and Therapeutics, McGill University, Montreal, Québec, Canada. ¹²Epigenomics Program, Broad Institute, Cambridge, Massachusetts, USA. ¹³Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. ¹⁴Harvard Stem Cell Institute, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to P.L.D. (pdejager@partners.org) or D.A.B. (david_a_bennett@rush.edu).

Received 5 May; accepted 16 July; published online 17 August 2014; doi:10.1038/nn.3786

Illumina-derived data to genome-wide DNA methylation sequence data generated from the same brain DNA samples in four of the subjects (two non-impaired and two AD subjects): in these four sub-

jects, we found a very strong correlation (mean $r = 0.97$) between the estimated levels of methylation generated by the two technologies, consistent with prior reports¹⁵.

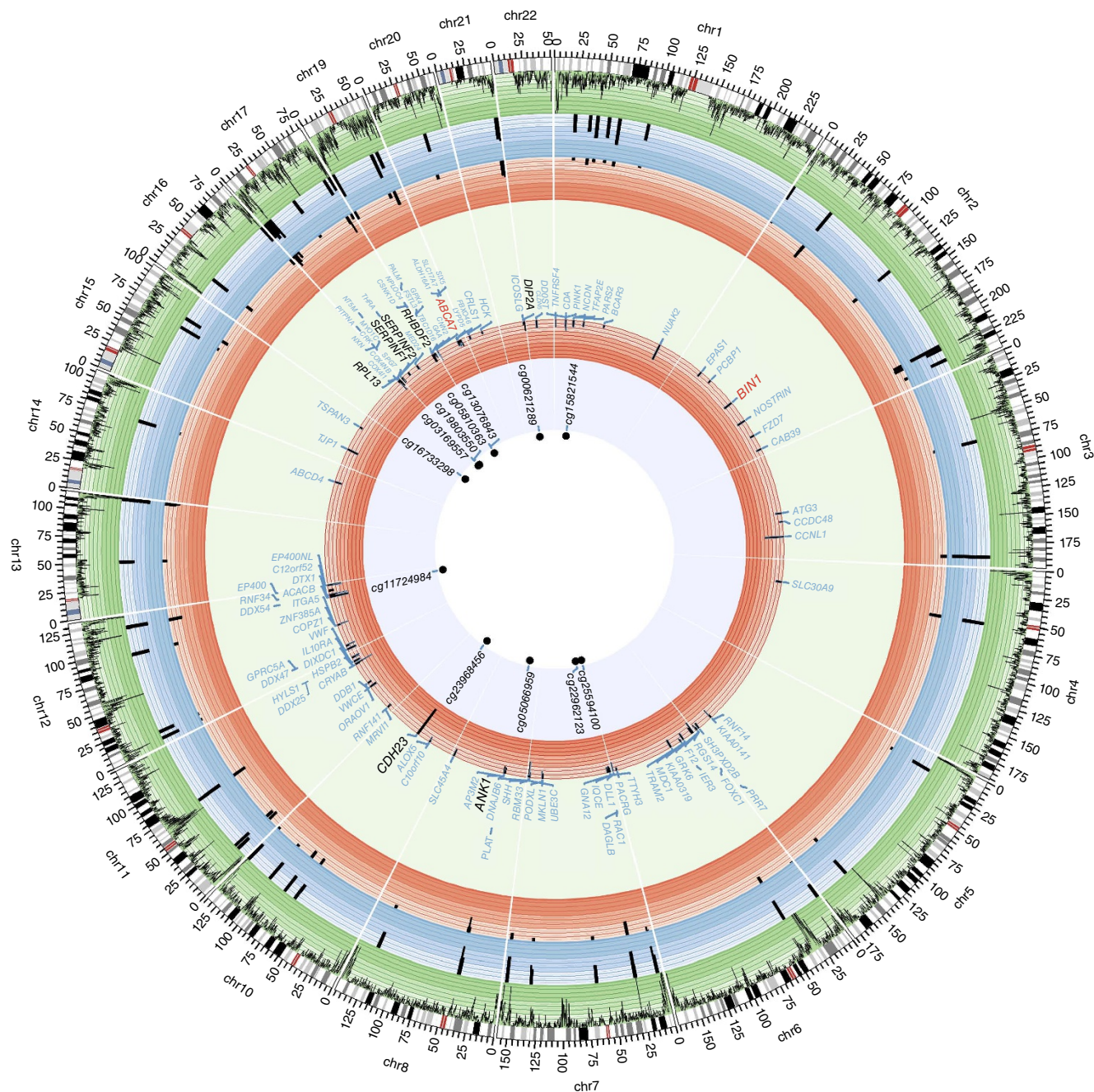


Figure 1 Summary of the genome-wide brain DNA methylation scan for NP burden and its validation using independent DNA methylation data and brain RNA data. Each sector of this diagram presents summary results of the three different analyses in a chromosome. The perimeter of this circular figure presents the physical position along each chromosome (in Mb). The cytogenetic bands of each chromosome are presented in the first circle, with the centromere highlighted in red. The next circle (green) reports the density of CpG probes successfully sampled by the Illumina beadset that are present in a given genomic segment (range, 0–200 probes per 100 kb). The blue circle reports the results of the DNA methylation scan: using a $-\log(P)$ scale (range, 0–20), we report the results for each of the 71 associated CpGs found in 60 independent differentially methylated regions (DMR) from the analysis relating DNA methylation levels to NP burden. Similarly, the first red circle reports the $-\log(P)$ (range, 0–10) for the 71 CpGs in the replication analysis. The cream-colored circle reports the names of genes found within 50 kb of each associated CpG (light blue letters). The *ABCA7* and *BIN1* regions, which harbor AD susceptibility alleles, are highlighted in red letters. The subset of the genes with differential mRNA expression in AD in the Mayo clinic data set is shown in black. The next red circle reports the results of the association of RNA expression level of these genes to a diagnosis of AD in the Mayo clinic data set ($-\log(P)$; range, 0–20). The central circle reports the set of validated CpGs. Chromosomes 9 and 18 contain no CpG that meets a threshold of genome-wide significance; thus, to enhance the clarity of the figure for the other chromosomes, these two chromosomes are not included in the figure. Not all genes found in the associated regions are listed in the figure. For clarity, only a subset of genes are selected from loci significant in the discovery analysis.

Notably, when examining the nature of human cortical methylation profiles across our subject population, we noted that the mean Pearson correlation of methylation levels for all possible subject pairs was 0.98 (Supplementary Fig. 1), suggesting that that majority of CpG sites did not show significant interindividual variation in methylation levels despite the very different life course of each of these older subjects. As expected, we found many more differences in DNA methylation profiles between our cortical samples and lymphoblastic cell lines from HapMap individuals that were profiled for assessments of data quality in our experiment (Supplementary Fig. 2).

Discovery study to identify differentially methylated regions

Our analytic strategy involves three stages (Fig. 1). In stage 1, we carried out a DNA methylation screen for chromosomal regions in which methylation levels correlate with AD pathology (Online Methods). In stage 2, we replicated the significantly associated CpGs from stage 1 in an independent set of subjects. In stage 3, we attempted to functionally validate the role of the differentially methylated regions that were replicated in stage 2 using mRNA obtained from AD and non-AD subjects. This strategy accomplished two goals: it further confirmed the role of a given differentially methylated region by showing that a meaningful biological effect (transcriptional change) related to the disease occurred in this region, and it helped to narrow down which of the genes near the differentially methylated CpGs were differentially expressed and may therefore be the target gene(s) in a given region.

In the primary analysis of our cortical methylation profiles (stage 1), we identified autosomal CpGs whose level of methylation correlates with the burden of neuritic amyloid plaques (NP), a key quantitative measure of Alzheimer's disease neuropathology. NP burden better captures the state of the brain of a deceased subject, as cognitively intact individuals display a range of NP pathology, some of which meet neuropathologic criteria for a diagnosis of AD^{16,17}. 137 CpGs were associated with the burden of NP pathology ($P < 1.20 \times 10^{-7}$; Table 1 and Supplementary Table 2). This threshold of significance accounts for the testing of all 415,848 tested CpGs by imposing a Bonferroni correction on a standard $P < 0.05$. Given that the exact number of functionally independent units of methylation in the genome is currently unknown, we chose this simple, but conservative, strategy to account for the testing of multiple hypotheses and correct for the testing of each CpG that was measured. As the proportion of neurons found in each sample was not related to AD ($P = 0.08$), we did not include this as a term in the primary analysis. However, to focus only on the most conservatively associated CpGs, we performed a secondary analysis that included the variable that captures the proportion of neurons as well as surrogate variables that capture structure in the methylation data that do not correlate with known confounders and may capture cryptic technical or other artifacts. Of the 137 CpGs discovered in the primary analysis, 71 CpGs remained significant in the more conservative secondary analysis (Table 1 and Supplementary Table 2). Some of these 71 CpGs were found in the same chromosomal segment and were highly correlated in their level of methylation.

Table 1 CpGs associated with amyloid burden: validated CpGs and CpGs in known AD loci

| CpG | Chr | Position (bp) | Discovery study | | | | Replication study | | | | Genes within 50 kb of associated CpG |
|---------------|-----|---------------|----------------------------|------------------------|----------------------------|------------------------|-------------------|------------------------|----------------------|-----------------------|---|
| | | | Estimate NP burden model 1 | <i>P</i> | Estimate NP burden model 3 | <i>P</i> | Estimate AD | <i>P</i> | Estimate Braak score | <i>P</i> | |
| cg11724984 | 12 | 121890864 | 3.02 | 4.76×10^{-9} | 4.48 | 9.75×10^{-11} | 16.7 | 1.27×10^{-7} | 3.40 | 5.44×10^{-6} | <i>RNF34, KDM2B</i> |
| cg23968456 | 10 | 73521631 | 4.97 | 3.97×10^{-10} | 4.79 | 8.97×10^{-9} | 16.8 | 1.45×10^{-5} | 1.62 | 8.27×10^{-6} | <i>CDH23, C10orf105, C10orf54</i> |
| cg15821544 | 1 | 43473840 | 3.52 | 1.17×10^{-7} | 5.11 | 8.09×10^{-11} | 18.9 | 1.46×10^{-7} | 2.47 | 2.08×10^{-4} | <i>SLC2A1, FLJ32224</i> |
| cg16733298 | 16 | 19127132 | 2.75 | 5.24×10^{-8} | 3.96 | 1.32×10^{-9} | 12.9 | 9.76×10^{-6} | 3.58 | 2.30×10^{-4} | <i>COQ7, ITPRIPL2</i> |
| cg22962123 | 7 | 27153605 | 1.7 | 1.12×10^{-7} | 3.08 | 8.79×10^{-13} | 10.8 | 5.56×10^{-8} | 5.13 | 2.79×10^{-4} | <i>HOXA1, HOTAIRM1, HOXA2, AK291164, HOXA3, AK311383, BC035889, HOXA4, LOC100133311, HOXA5, HOXA6, DQ655986, HOXA7, HOXA9, HOXA10</i> |
| cg13076843 | 17 | 74475294 | 2.35 | 1.68×10^{-9} | 2.29 | 5.99×10^{-8} | 9.32 | 5.81×10^{-7} | 3.02 | 2.99×10^{-4} | <i>UBE2O, AANAT, RHBDF2, AX747521, CYGB, PRCD</i> |
| cg25594100 | 7 | 4786943 | 3.15 | 2.54×10^{-11} | 4.22 | 3.33×10^{-13} | 14.9 | 1.83×10^{-8} | 4.09 | 4.19×10^{-4} | <i>FOXK1, AP5Z1 (KIAA0415), RADIL</i> |
| cg00621289 | 21 | 47855916 | 3.5 | 6.48×10^{-8} | 5.06 | 2.18×10^{-11} | 17.5 | 3.98×10^{-7} | 2.02 | 4.95×10^{-4} | <i>PCNT, DIP2A</i> |
| cg19803550 | 17 | 1637391 | 4.36 | 1.04×10^{-8} | 4.48 | 1.60×10^{-8} | 19 | 6.81×10^{-7} | 1.51 | 5.61×10^{-4} | <i>PRPF8, TLCD2, MIR22HG, AF070569, MIR22, WDR81, SERPINF2, SERPINF1, SMYD4</i> |
| cg03169557 | 16 | 89598950 | 4.86 | 3.99×10^{-10} | 4.88 | 3.36×10^{-9} | 18.6 | 1.67×10^{-6} | 1.52 | 6.05×10^{-4} | <i>ANKRD11, SPG7, SNORD68, RPL13, CPNE7</i> |
| cg05066959 | 8 | 41519308 | 2.69 | 7.13×10^{-14} | 2.78 | 7.56×10^{-13} | 11.2 | 4.88×10^{-10} | 3.45 | 6.48×10^{-4} | <i>AGPAT6, NKX6-3, JA429246, ANK1</i> |
| cg05810363 | 17 | 74475270 | 2.95 | 3.68×10^{-10} | 2.86 | 8.11×10^{-9} | 11.3 | 3.36×10^{-7} | 2.76 | 7.93×10^{-4} | <i>UBE2O, AANAT, RHBDF2, AX747521, CYGB, PRCD</i> |
| Known AD loci | | | | | | | | | | | |
| cg22883290 | 2 | 127800646 | 4.41 | 3.73×10^{-8} | 4.44 | 8.97×10^{-8} | 15.9 | 1.83×10^{-5} | 0.96 | 0.0067 | <i>BIN1</i> |
| cg02308560 | 19 | 10711176 | 2.19 | 3.06×10^{-8} | 3.62 | 2.45×10^{-12} | 13.7 | 1.77×10^{-8} | 3.60 | 0.011 | <i>CNN2, ABCA7, HMHA1, POLR2E, GPX4, SBNO2</i> |

The NP analysis reports the results of a linear regression analysis relating DNA methylation level to the burden of neuritic plaque. Model 1 refers to the primary analysis and Model 3 refers to the secondary analysis that included a variable for the estimated proportion of neuronal cells in the tissue and surrogate variables. The AD analysis reports the result of a logistic regression relating the level of methylation of a given CpG to a pathologic diagnosis of AD. The threshold of genome-wide significance is $P < 1.12 \times 10^{-7}$, which accounts for the testing of probes genome-wide. Listed are all genes found within a segment 50 kb upstream and downstream of the associated CpG. Those CpGs that were significant in our discovery analysis and found in AD susceptibility loci are listed in the 'Known AD loci'. Chr, chromosome; patho AD, pathologic diagnosis of AD.

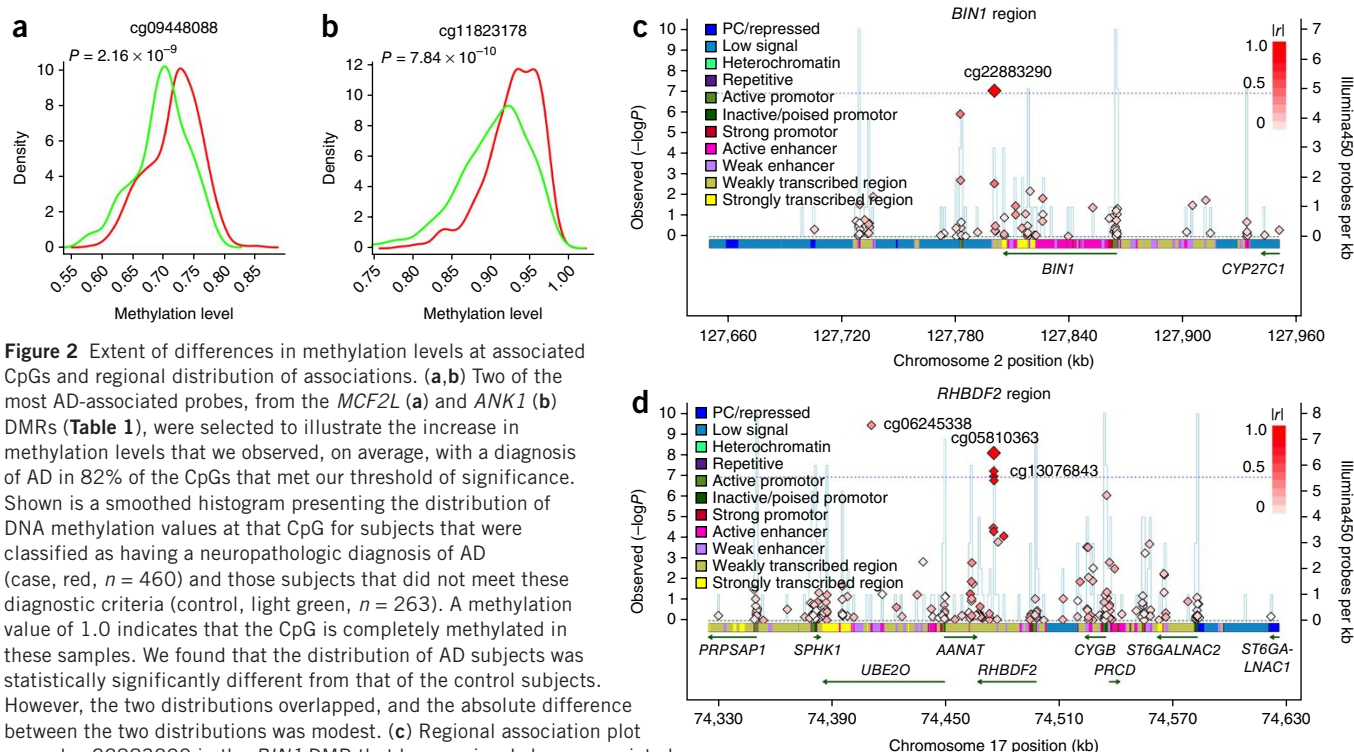


Figure 2 Extent of differences in methylation levels at associated CpGs and regional distribution of associations. (a,b) Two of the most AD-associated probes, from the *MCF2L* (a) and *ANK1* (b) DMRs (Table 1), were selected to illustrate the increase in methylation levels that we observed, on average, with a diagnosis of AD in 82% of the CpGs that met our threshold of significance. Shown is a smoothed histogram presenting the distribution of DNA methylation values at that CpG for subjects that were classified as having a neuropathologic diagnosis of AD (case, red, $n = 460$) and those subjects that did not meet these diagnostic criteria (control, light green, $n = 263$). A methylation value of 1.0 indicates that the CpG is completely methylated in these samples. We found that the distribution of AD subjects was statistically significantly different from that of the control subjects. However, the two distributions overlapped, and the absolute difference between the two distributions was modest. (c) Regional association plot around cg22883290 in the *BIN1* DMR that has previously been associated with AD susceptibility in genome-wide association studies. Each diamond represents one CpG tested in this region. The horizontal dotted blue line highlights the threshold of significance for this analysis. The vertical blue line reports the density of CpG probes at a given point. The extent to which DNA methylation level at a given CpG correlates with the level of DNA methylation of the best CpG (cg22883290) is reported using the I_{r1}^2 value. Finally, above the diagram of the genes found in this DNA segment, the chromatin state of the region is shown, as assessed in healthy, unimpaired older individuals with minimal AD-related pathology. The chromatin state was derived in 200-bp bins. Overall, the *BIN1* gene appeared to be in an open, transcribed conformation in healthy, older dorsolateral prefrontal cortex, and the associated CpG appeared to be located in a region just 3' to the gene, which was largely in a conformation found on the periphery of actively transcribed regions. (d) Regional association plot around the *RHBDF2* DMR, centered on cg13076843, which met our threshold of significance. An associated CpG was found in close proximity to two genes, and our RNA analyses suggest that it is *RHBDF2* that is the target of the DMR, as its expression was altered relative to AD (Table 2).

Altogether, the 71 CpGs were found in 60 discrete differentially methylated regions distributed throughout the genome (Fig. 1): in 7 of these 60 regions, up to three neighboring CpGs with correlated levels of methylation emerged as being significant in our analysis and probably captured the same effect.

Individually, any one of the significantly associated CpGs (Table 1 and Supplementary Table 2) had a modest effect on the brain's NP burden: on average, each the 71 CpGs explained 5.0% (range = 3.7–9.7%) of the variance in NP burden. However, this is greater than the proportion of variance explained by genetic variants associated with AD, with the exception of *APOE*. For example, in our subjects, the well-validated *CRI* susceptibility allele explains just 1% of variance in NP burden¹⁸, and all known AD variants and *APOE* $\epsilon 4$ account for 13.9% of the variance in NP burden. If we consider all 71 CpGs in one comprehensive model, they explained 28.7% of the variance in NP burden, suggesting that methylation levels of certain genomic regions are correlated and that cortical DNA methylation of a large number of discrete regions is strongly correlated with a key measure of AD neuropathology.

Notably, two of the 71 significantly associated CpGs (Supplementary Table 2) were found in loci that harbor known AD susceptibility alleles: cg22883290 in the *BIN1* locus (beta = 4.44, $P = 9.00 \times 10^{-8}$) and cg02308560 in the *ABCA7* locus (beta = 3.62, $P = 2.45 \times 10^{-12}$)^{19–22}. cg22883290 is located 5 kb from the 5' end of the *BIN1* gene and 92 kb from the index single nucleotide polymorphism (SNP), rs744373, that best captures the genetic association to AD in this region (Fig. 2c)¹⁹. The susceptibility variant rs744373 was moderately associated with

the level of methylation at cg22883290 ($P = 0.0003$). However, the CpG association with AD pathology was not driven by the variant: adjusting for rs744373 did not meaningfully change the effect size of the CpG association to NP burden (model with rs744373 as a covariate: beta = 4.37, $P = 4.91 \times 10^{-7}$). In our data set of modest size, rs744373 was not associated with AD susceptibility, and we therefore could not formally test for mediation of the SNP's association to disease by CpG methylation. In the case of *ABCA7*, the index SNP (rs3764650) was associated with NP burden²³, but had no association ($P = 0.07$) with the level of methylation at cg02308560, which is 25 kb away, so, in both of these regions, SNPs and CpGs appeared to have independent effects on AD susceptibility. Overall, risk of AD may therefore be affected by different sources of genomic variation (genetic and epigenetic) that have independent effects on the disease process.

To facilitate the interpretation of our results, we performed a secondary analysis correlating the level of methylation at these 71 CpGs with a post-mortem, neuropathologic diagnosis of AD. 22 of the NP-associated CpGs were also associated with a diagnosis of AD at a genome-wide level of significance (Table 1 and Supplementary Table 2), and all of the CpGs associated with NP burden displayed at least some evidence of association ($P < 0.001$) with AD. This is not surprising, as NP burden is one criterion for a neuropathologic AD diagnosis. We noted a polarization in the direction of these associations: 82% of the differentially methylated regions were more methylated in subjects with a diagnosis of AD. As noted above, the increased level of methylation in relation to AD at any one associated probe was modest (Fig. 2).

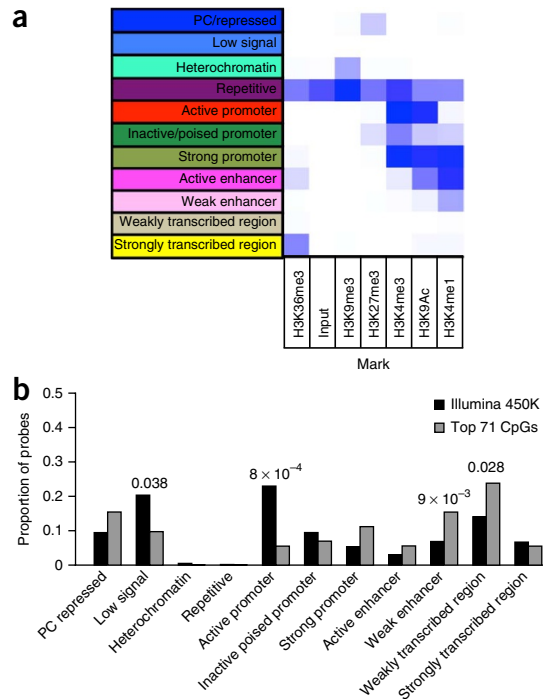


Figure 3 Distribution of CpGs associated ($P < 1.2 \times 10^{-7}$) with NP among 11 chromatin states found in mid-frontal cortex. **(a)** Chromatin map of the dorsolateral prefrontal cortex. Using data generated by the National Institute of Health's Epigenomic Roadmap effort, we assigned each chromosomal segment to 1 of 11 discrete chromatin states. MAP subjects were used that were cognitively non-impaired at the time of death and had minimal pathology on neuropathological examination. The heatmap (white, low; blue, high) graphically displays the relative abundance of sequences found in a segment of DNA after immunoprecipitation for a particular histone mark. Each chromatin state had a unique complement of histone marks. **(b)** We used the chromatin map in **a** to identify the chromatin state in which each of the interrogated CpG dinucleotides were found. The histogram compares the distribution of chromatin states found at those 71 associated CpG dinucleotides whose methylation level was associated with neuritic plaques (**Table 1**) to the overall distribution of chromatin states found in all 415,848 CpG dinucleotides that were analyzed.

significantly different from the beta values calculated from the entire sample collection (**Supplementary Table 3**). This suggests that the altered DNA methylation that we identified in our discovery study is an early feature of AD pathology and occurs in the presymptomatic stage of the disease. These DNA methylation changes are therefore not secondary to the later stages of the dementing process. The question of whether altered DNA methylation contributes to the pathologic process or is an early epiphenomenon of the neurodegenerative process remains open.

Distribution of associated CpGs among different chromatin states

To better understand the functional consequences of the associated CpGs, we interpreted our results in relation to a chromatin map of the dorsolateral prefrontal cortex, generated in collaboration with the Epigenomics Roadmap team (<http://www.roadmapepigenomics.org/>). It is derived from two MAP subjects who were cognitively non-impaired

Validation of the associated CpG in an independent sample set

To further assess the robustness of our results, we evaluated the 71 significantly associated CpGs in an independent collection of 117 subjects with a different quantitative measure of AD pathology (Braak staging)²⁴. These subjects were profiled in a sample of frontal cortex using the same Illumina HumanMethylation450 platform (**Supplementary Table 1b**). We imposed a Bonferroni correction in this analysis and found that 12 CpGs were significant in this analysis (**Table 1**); given that two of these CpGs were found in the same differentially methylated region near *RHBDF2* (**Fig. 2d**), 11 of the differentially methylated regions from the discovery study were validated (**Supplementary Fig. 3**). Thus, despite the use of a different, but related, measure of AD pathology and a much smaller sample size, we observed robust replication of our discovery screen's results. In addition to the significant CpGs, many other CpGs displayed suggestive evidence of association: when evaluating the entire set of 71 CpG, we found that the effect size of most of these CpGs was consistent across the two data sets (**Supplementary Fig. 4**), suggesting that most of these CpGs will be validated as larger sample sizes are profiled. Both the *BIN1* ($P = 0.0067$) and the *ABCA7* ($P = 0.011$) CpGs displayed suggestive evidence of replication (**Table 1**).

Cognitively non-impaired subjects display the same alterations in methylation

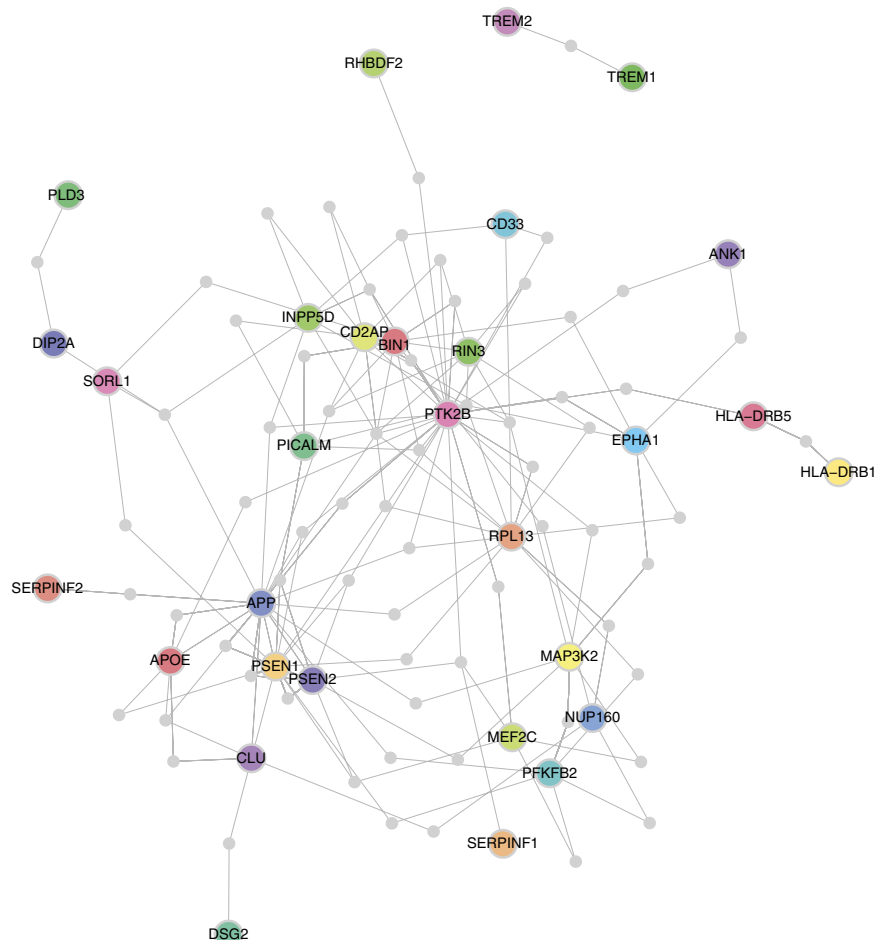
To begin to explore the question of whether the increased level of DNA methylation in the associated regions is a cause or an effect of the neurodegenerative process of AD, we limited the NP analysis to those subjects who were deemed to be cognitively non-impaired at the time of death (no AD and no mild cognitive impairment). As has been well-documented in neuropathological and imaging studies^{25,26}, a large fraction of non-impaired, older individuals demonstrated accumulation of amyloid pathology that was asymptomatic. In the subset of non-impaired subjects, the P values for the CpG associations were diminished given the reduced sample size ($n = 237$), but the beta values, which capture the magnitude of the association's effect, were not

Table 2 Analyses of transcriptional data in loci associated with amyloid burden

| Gene | Mayo Clinic subjects | |
|-----------------|----------------------|---------|
| | Beta | P |
| <i>CDH23</i> | 0.58 | <0.0001 |
| <i>DIP2A</i> | 0.07 | <0.0001 |
| <i>RPL13</i> | -0.08 | <0.0001 |
| <i>RHBDF2</i> | 0.22 | 0.0001 |
| <i>SERPINF1</i> | -0.14 | 0.0004 |
| <i>SERPINF2</i> | 0.2 | 0.0008 |
| <i>ANK1</i> | -0.17 | 0.0012 |
| <i>SMYD4</i> | 0.03 | 0.065 |
| <i>PRPF8</i> | -0.04 | 0.13 |
| <i>SLC2A1</i> | 0.02 | 0.13 |
| <i>AP5Z1</i> | 0.05 | 0.15 |
| <i>PCNT</i> | 0.03 | 0.15 |
| <i>COQ7</i> | 0.02 | 0.22 |
| <i>ANKRD11</i> | -0.02 | 0.24 |
| <i>C10orf54</i> | 0.07 | 0.25 |
| <i>FOXK1</i> | 0.02 | 0.28 |
| <i>UBE20</i> | -0.02 | 0.28 |
| <i>RNF34</i> | -0.05 | 0.30 |
| <i>WDR81</i> | -0.03 | 0.39 |
| <i>SPG7</i> | 0.04 | 0.54 |
| <i>CYGB</i> | -0.02 | 0.62 |
| <i>AGPAT6</i> | 0 | 0.9 |

Beta refers to effect size in the linear regression associating gene expression and AD diagnosis, with non-AD being the reference subject group. The significance threshold was $P < 0.0023$ given 22 tests. Results are shown for mean expression level across available probes for each gene.

Figure 4 Genes identified in our DNA methylation screen connect to a network of known AD susceptibility genes. Using protein-protein interaction data, the DAPPLE algorithm evaluated the extent of connectivity among known AD genes (susceptibility and Mendelian genes) and the seven genes found in DMRs that were also differentially expressed relative to AD. Shown are the results of an analysis allowing for one common interactor protein that is not known to be associated with AD. For example, *RHBDF2* is displayed at the top of the figure in green and connects to *PTK2B*, a protein tyrosine kinase genetically associated with AD susceptibility that has a central role in this network. Notably, *SERPINF1* and *SERPINF2* connect to different elements of the amyloid component of the network (bottom left). Furthermore, *DIP2A* connects the recently described *PLD3* gene that has a rare AD susceptibility allele and to *SORL1*, a gene with a common AD susceptibility allele, that connects to the amyloid precursor protein (APP). These interconnections are consistent with the reported effects of both *PLD3* and *SORL1* on amyloid biology and implicate *DIP2A* in the same process (see also **Supplementary Fig. 6**). The colored nodes are the proteins encoded by genes implicated in AD (genetic and epigenomic associations); the colors have no meaning. The connecting proteins not known to be associated with AD are shown in gray.



at the time of death and had minimal AD-associated pathology on post-mortem examination. Using histone-modification profiles and established methods²⁷, each 200-bp segment of the genome was annotated as being in 1 of 11 chromatin states (**Fig. 3a**) that capture the transcriptional states and putative regulatory elements found in this tissue. Using this reference map, we observed that at least some of the 71 associated CpGs were found in every chromatin state, but there was an enrichment of associated CpGs in regions predicted to be weak enhancers ($P = 0.0098$) or to be in a weakly transcribed chromatin state ($P = 0.028$) (**Fig. 3b** and **Supplementary Table 4**). Furthermore, we found a strong under-representation in regions displaying a strong promoter profile in the reference chromatin map ($P = 8 \times 10^{-4}$). These data suggest that the chromatin architecture of strong promoters that drive fundamental cellular processes of neurons and glia in the healthy brain may not be strongly altered by AD. Rather, methylation changes appear to primarily affect genomic regions that are weakly transcribed or inactive in the healthy older brain. There were no enrichments noted in different genic features or in different structures defined in relation to CpG islands (**Supplementary Fig. 5a,b**).

Functional validation of the CpG associations

Focusing on the 12 CpGs that have been validated in the replication stage, we evaluated their role in AD by assessing the level of expression of genes found in the vicinity (± 50 kb) of these differentially methylated regions (DMRs) in an independent set of 202 AD and 197 non-AD individuals assembled by the Mayo Clinic (**Supplementary Table 5**) that have RNA data from the temporal cortex. We found that the level of expression of 7 of the 21 selected genes had significant associations ($P < 0.0023$) with AD in these data (**Table 2** and **Supplementary Table 6**): *ANK1*, *CDH23*, *DIP2A*, *RHBDF2*, *RPL13*, *SERPINF1* and *SERPINF2*.

Integrating our results with known AD genes

To further evaluate the role of these seven genes in relation to well-validated AD genes, we used the DAPPLE algorithm to evaluate the connectivity of these genes with the network of known AD susceptibility genes. We previously used this method, which requires coexpression of interacting protein pairs and adjusts for gene size, and we reported the existence of an AD susceptibility network derived from protein-protein interaction data¹³. We used an updated model that includes the latest results from genome-wide association studies and the studies of rare variation. First, we found that the network of susceptibility genes from genome-wide association studies and Mendelian AD genes was significant both in terms of direct connectivity ($P = 0.0072$) and indirect connectivity (proportion of susceptibility genes sharing a common interactor, $P = 0.037$) (**Supplementary Fig. 6**). We then repeated the analysis after adding the seven genes found in the validated differentially methylated regions that also displayed altered RNA expression in AD. Several of the differentially expressed genes found in the differentially methylated regions—*ANK1*, *DIP2A*, *RHBDF2*, *RPL13*, *SERPINF1* and *SERPINF2*—were connected to the AD susceptibility network derived from genetic studies (**Fig. 4**). The direct ($P = 0.0072$) and indirect ($P = 0.042$) network connectivity remained significant in the iteration of the network analysis that included the seven genes with altered RNA expression levels.

DISCUSSION

Overall, although our study has certain limitations as a result of the Illumina platform, such as surveying only a fraction of the human

genome's CpGs and the array's inability to distinguish two closely related chromatin marks (DNA methylation and DNA hydroxymethylation), we nonetheless found several replicated, functionally validated associations between altered DNA methylation and the presymptomatic accumulation of AD pathology. These changes do not appear to be part of a generalized, genome-wide process: specific differentially methylated regions were targeted in AD and were unlikely to be found in genes that were actively transcribed in the healthy older brain. Instead, the associated regions were more likely to be in a poorly transcribed chromatin conformation in the older healthy brain (**Fig. 3b**). Altered DNA methylation (**Table 1**) and enhanced mRNA expression ($P = 1.09 \times 10^{-4}$) of the *BIN1* gene in AD in the Mayo Clinic data set links our study to a well-validated AD susceptibility locus and a recent report of enhanced *BIN1* expression in AD²⁸. The *BIN1* cg22883290 association was significant in our discovery study, and, although the results were suggestive ($P = 0.0067$) in the small replication analysis, the direction of the effect was consistent (**Table 1**), suggesting that the association is likely to validate with additional subjects. Our results therefore refine our understanding of the *BIN1* locus and suggest that different types of genomic variation (SNP and CpG) can have independent effects that integrate on the expression of *BIN1* and influence AD susceptibility. Similarly, cg02308560 in the *ABCA7* locus was associated to AD pathology, and this association was independent of the susceptibility allele found in its vicinity. These two loci illustrate the point that, although genetic variation can drive differences in DNA methylation for certain CpGs³, the associations that we observed are not driven by genetic associations with AD pathology: we also recently completed a genome-wide SNP association study with the same trait in the same subjects and found no significant genetic associations²³. Thus, our CpG associations are not driven by SNP associations.

The collocation of genetic susceptibility with CpG associations to AD pathology, along with the presence of the CpG associations in cognitively non-impaired subjects (**Supplementary Table 3**) and the connection of six of the differentially methylated genes (**Fig. 4**) to an existing AD susceptibility network (**Fig. 4**), suggest that DNA methylation changes are involved in the onset of AD. However, as with genetic studies, our epigenome-wide scan only reports associations with a trait. Thus, we cannot state that the observed changes in methylation are causal: given the plasticity of the epigenome, it is possible that these changes are an early consequence of AD pathology. Further experimental work needs to be conducted to resolve this question.

Looking at the AD network map, *DIP2A* connects directly to the known *SORL1* susceptibility gene²⁹ and indirectly to *PLD3*, a recently reported AD susceptibility gene that is otherwise not connected to the AD susceptibility network. *DIP2A* may function as a cell surface receptor protein³⁰ and, given the putative role of *PLD3* (ref. 31) and *SORL1* in amyloid processing, its relation to the burden of NP pathology may well be related to a direct effect on amyloid processing. *SERPINF1* and *SERPINF2* also connect to elements of the amyloid machinery. Notably, *SERPINF1* mRNA expression was reduced in AD (**Table 2**), and its knockdown in an *in vitro* system leads to reduced neurite outgrowth³², suggesting one potential effect of this gene.

On the other hand, *ANK1* and *RHBDF2* connect to *PTK2B*, an AD gene that is a key element of the signaling cascade involved in modulating the activation of microglia and infiltrating macrophages. Several other AD genes, such as *CD33* and *EPHA1*, connect to this molecule as well. Although little is known regarding the potential role of *ANK1*, the connection of *RHBDF2* with *PTK2B* is consistent with

the known role of this molecule in myeloid cells: it is necessary for the transport of TNF α converting enzyme (TACE, also called *ADAM17*), which releases TNF α from the cell surface³³. The absence of *RHBDF2* in mice affects the normal release of TNF α from the cell surface³⁴ and impairs systemic immune responses to pathogens³⁵. *In vitro* work also suggests that *RHBDF2* may function in regulating the substrate specificity of the TACE/*ADAM17* protease, which functions in the release of TNF α , as well as that of other proteins such as epidermal growth factor (EGF)³⁶. Its exact role in AD is not clear at this point, but our data suggest that *RHBDF2* expression is increased in the context of AD (**Table 2**). Its connection to *PTK2B* further suggests that it may be involved in the role of microglia and infiltrating macrophages in the AD pathophysiological process. Consistent with this, adjusting for an estimate of the number of microglial cells using an RNA-based model seems to account for the AD-associated differences in *RHBDF2* mRNA expression (**Supplementary Table 7a,b**).

Such cell-type adjustments using surrogate markers for different cell types^{37–39} are crude analyses, but are helpful to begin to assign some of the transcriptional alterations to certain cell types. For example, using GFAP expression as a surrogate marker for astrocytes, we found that adding a term for GFAP expression in our assessment of *CDH23* RNA expression largely abrogated the association of *CDH23* RNA expression with AD (**Supplementary Table 7b**). Given that GFAP expression is enhanced with astrocyte activation^{40,41}, we cannot distinguish whether the alteration of *CDH23* RNA expression (and presumably its altered DNA methylation) was caused by an increased number of astrocytes near neuritic plaques, the activation of these astrocytes or a combination of both effects. Regardless of the exact mechanism, in the case of *CDH23*, our DNA methylation screen uncovered a robust alteration in methylation that can now be dissected mechanistically.

The human cortex has a complex architecture that includes many different types of neurons, glia, and other cells such as microglia, peripheral immune cells and endothelial cells from cortical capillaries. The changes in DNA methylation that we observed most likely represent the altered methylation state of a subset of cells in our cortical sample, as AD pathology accumulates over several decades and only a small number of cells are affected at a given time. It is too early to confidently differentiate between three possibilities that could explain these modest, but robust, changes in methylation that occur in relation to AD pathology: a fraction of the constituent cortical cells change, such as activated astrocytes in the vicinity of neuritic plaques that overexpress *CDH23*; the relative proportion of the constituent cell populations of the cortex changes as some populations such as neurons are lost; or there is a modest influx of immune cells from the systemic circulation that alters the relative abundance of the different cortical cell populations. It is also likely that more than one of these or other, unsuspected processes may be at work.

Overall, the replication of our study's results by an independent study²⁴ and its functional validation with RNA data makes our strategy for investigating the brain's epigenome more broadly relevant to other epigenomic epidemiology studies. With clear effect sizes in hand, our results can be used to calibrate the design of future human studies in the brain or other organs. We also made the important observation that these epigenomic changes occur early in the pathologic process, while subjects displayed no cognitive impairment, but had accumulated amyloid pathology. Going forward in the aging brain, we clearly need to more precisely map the alterations of chromatin structure that contribute to AD pathophysiology and to assess, using model systems, whether remodeling the epigenome is a fruitful goal for the development of AD therapies.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the National Institute for Health (NIHR) Biomedical Research Unit in Dementia in the South London and Maudsley NHS Foundation Trust (SLaM), Brains for Dementia Research (Alzheimer Brain Bank UK) and the donors and families who made this research possible. We also would like to thank the participants of the ROS and MAP studies for their participation in these studies. Support for this research was provided by grants from the US National Institutes of Health (R01 AG036042, R01AG036836, R01 AG17917, R01AG15819, R01 AG032990, R01 AG18023, RC2 AG036547, P30 AG10161, P50 AG016574, U01 ES017155, KL2 RR024151, K25 AG041906-01). Support was also provided by the Siragusa Foundation to N.E.-T., and the Robert and Clarice Smith and Abigail Van Buren Alzheimer's Disease Research Program to N.E.-T., S.G.Y. and F.Z. This work was funded by US National Institutes of Health grant AG036039 to J.M. and an Equipment Grant from Alzheimer's Research UK.

AUTHOR CONTRIBUTIONS

C.M., A.T., W.B., S.G., C.B.E., B.E.B., A.M. and J.A.S. collected, prepared and generated data from the samples. G.S., L.B.C., J.E., B.T.K., M.K., T.R., J.R. and L.Y. performed analyses on the resulting data. K.L., L.C.S. and J.M. generated and analyzed the replication data. N.E.-T., J.B., H.S.C., C.Y., F.Z. and S.G.Y. provided and analyzed RNA data from AD and non-AD brains. P.L.D. and D.A.B. designed the study. P.L.D., D.A.B. and L.B.C. wrote the manuscript. All of the authors critically reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Wang, S.C., Oelze, B. & Schumacher, A. Age-specific epigenetic drift in late-onset Alzheimer's disease. *PLoS ONE* **3**, e2698 (2008).
- Feinberg, A.P. *et al.* Personalized epigenomic signatures that are stable over time and covary with body mass index. *Sci. Transl. Med.* **2**, 49ra67 (2010).
- Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31**, 142–147 (2013).
- Chuang, D.M., Leng, Y., Marinova, Z., Kim, H.J. & Chiu, C.T. Multiple roles of HDAC inhibition in neurodegenerative conditions. *Trends Neurosci.* **32**, 591–601 (2009).
- Gräff, J. *et al.* An epigenetic blockade of cognitive functions in the neurodegenerating brain. *Nature* **483**, 222–226 (2012).
- Chouliaras, L. *et al.* Consistent decrease in global DNA methylation and hydroxymethylation in the hippocampus of Alzheimer's disease patients. *Neurobiol. Aging* **34**, 2091–2099 (2013).
- Bakulski, K.M. *et al.* Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex. *J. Alzheimers Dis.* **29**, 571–588 (2012).
- Numata, S. *et al.* DNA methylation signatures in development and aging of the human prefrontal cortex. *Am. J. Hum. Genet.* **90**, 260–272 (2012).
- Akbarian, S., Beeri, M.S. & Haroutunian, V. Epigenetic determinants of healthy and diseased brain aging and cognition. *JAMA Neurol.* **70**, 711–718 (2013).
- Hernandez, D.G. *et al.* Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Hum. Mol. Genet.* **20**, 1164–1172 (2011).
- Bell, J.T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* **8**, e1002629 (2012).
- Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13**, R97 (2012).
- Raj, T. *et al.* Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. *Am. J. Hum. Genet.* **90**, 720–726 (2012).
- Bennett, D.A. *et al.* Selected findings from the Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* **33**, S397–S403 (2013).
- Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* **28**, 1106–1114 (2010).
- Negash, S., Bennett, D.A., Wilson, R.S., Schneider, J.A. & Arnold, S.E. Cognition and neuropathology in aging: multidimensional perspectives from the Rush Religious Orders Study and Rush Memory And Aging Project. *Curr. Alzheimer Res.* **8**, 336–340 (2011).
- Bennett, D.A., Wilson, R.S., Boyle, P.A., Buchman, A.S. & Schneider, J.A. Relation of neuropathology to cognition in persons without cognitive impairment. *Ann. Neurol.* **72**, 599–609 (2012).
- Chibnik, L.B. *et al.* CR1 is associated with amyloid plaque burden and age-related cognitive decline. *Ann. Neurol.* **69**, 560–569 (2011).
- Naj, A.C. *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.* **43**, 436–441 (2011).
- Seshadri, S. *et al.* Genome-wide analysis of genetic loci associated with Alzheimer disease. *J. Am. Med. Assoc.* **303**, 1832–1840 (2010).
- Hollingworth, P. *et al.* Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat. Genet.* **43**, 429–435 (2011).
- Lambert, J.C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094–1099 (2009).
- Shulman, J.M. *et al.* Genetic susceptibility for Alzheimer disease neuritic plaque pathology. *JAMA Neurol.* **70**, 1150–1157 (2013).
- Lunnon, K. *et al.* Cross-tissue methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease neuropathology. *Nat. Neurosci.* **17**, XXX–YYY (2014).
- Schneider, J.A., Arvanitakis, Z., Leurgans, S.E. & Bennett, D.A. The neuropathology of probable Alzheimer disease and mild cognitive impairment. *Ann. Neurol.* **66**, 200–208 (2009).
- Sperling, R.A. *et al.* Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 280–292 (2011).
- Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
- Chapuis, J. *et al.* Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. *Mol. Psychiatry* **18**, 1225–1234 (2013).
- Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
- Tanaka, M. *et al.* DIP2 disco-interacting protein 2 homolog A (*Drosophila*) is a candidate receptor for follistatin-related protein/follistatin-like 1—analysis of their binding with TGF-beta superfamily proteins. *FEBS J.* **277**, 4278–4289 (2010).
- Cruchaga, C. *et al.* Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature* **505**, 550–554 (2014).
- Roet, K.C. *et al.* A multilevel screening strategy defines a molecular fingerprint of progenerative olfactory ensheathing cells and identifies SCARB2, a protein that improves regenerative sprouting of injured sensory spinal axons. *J. Neurosci.* **33**, 11116–11135 (2013).
- Adrain, C., Zettl, M., Christova, Y., Taylor, N. & Freeman, M. Tumor necrosis factor signaling requires iRhom2 to promote trafficking and activation of TACE. *Science* **335**, 225–228 (2012).
- Siggs, O.M. *et al.* iRhom2 is required for the secretion of mouse TNFalpha. *Blood* **119**, 5769–5771 (2012).
- McIlwain, D.R. *et al.* iRhom2 regulation of TACE controls TNF-mediated protection against *Listeria* and responses to LPS. *Science* **335**, 229–232 (2012).
- Maretzky, T. *et al.* iRhom2 controls the substrate selectivity of stimulated ADAM17-dependent ectodomain shedding. *Proc. Natl. Acad. Sci. USA* **110**, 11433–11438 (2013).
- Cooper-Knock, J. *et al.* Gene expression profiling in human neurodegenerative disease. *Nat. Rev. Neurol.* **8**, 518–530 (2012).
- Simpson, J.E. *et al.* Microarray analysis of the astrocyte transcriptome in the aging brain: relationship to Alzheimer's pathology and APOE genotype. *Neurobiol. Aging* **32**, 1795–1807 (2011).
- Linnertz, C. *et al.* Genetic regulation of alpha-synuclein mRNA expression in various human brain tissues. *PLoS ONE* **4**, e7480 (2009).
- Renner, N.A. *et al.* Transient acidification and subsequent proinflammatory cytokine stimulation of astrocytes induce distinct activation phenotypes. *J. Cell. Physiol.* **228**, 1284–1294 (2013).
- Krafft, A.W. *et al.* Attenuating astrocyte activation accelerates plaque pathogenesis in APP/PS1 mice. *FASEB J.* **27**, 187–198 (2013).

ONLINE METHODS

Subjects and genotypes. The analyses in this manuscript included deceased subjects from two large, prospectively followed cohorts maintained by investigators at Rush University Medical Center: the *Religious Orders Study* (ROS) and the *Memory and Aging Project* (MAP). The ROS cohort, established in 1994, consists of more than 1,100 Catholic priests, nuns and brothers from 40 groups in 12 states who were at least 55 years of age and free of known dementia at the time of enrollment. The MAP cohort, established in 1997, consists of more than 1,600 men and women primarily from retirement facilities in the Chicago area who were at least 53 years of age and free of known dementia at the time of enrollment. All participants in ROS and MAP signed an informed consent agreeing to annual detailed clinical evaluations and cognitive tests, and the rate of follow-up exceeds 90%. Similarly, participants in both cohorts signed an Anatomical Gift Act donating their brains at the time of death. The overall autopsy rate exceeds 85%. As in previous studies¹⁷, we analyzed the ROS and MAP cohorts jointly, as they were designed to be combined, were maintained by a single investigative team and a large set of phenotypes collected were identical in both studies. All aspects of these studies were approved by the Institutional Review Boards of Rush University Medical Center and Partners Healthcare. More detailed information regarding the two cohorts can be found in previously published literature^{42,43}. Genotypes were available from prior studies and were derived from Affymetrix GeneChip 6.0 or Illumina Omni1-Quad genotypes and imputation using the HapMap reference, as previously described²³.

The replication DNA methylation analysis uses samples of prefrontal cortex (PFC) obtained from 117 individuals archived in the MRC London Neurodegenerative Disease Brain Bank (<http://www.kcl.ac.uk/iop/depts/cn/research/MRC-London-Neurodegenerative-Diseases-Brain-Bank/MRC-London-Neurodegenerative-Diseases-Brain-Bank.aspx>). All samples were dissected by trained specialists, snap-frozen and stored at -80°C . Genomic DNA was isolated from ~ 100 mg of each dissected brain region using a standard phenol-chloroform extraction method and tested for degradation and purity before analysis.

Temporal cortex expression levels for the autopsied Mayo Clinic subjects were obtained as part of a recently published brain expression GWAS (eGWAS), where the methodology is described in detail^{44,45}. Briefly, expression levels of 24,526 transcripts were measured from the temporal cortex of autopsied brains from subjects with pathologic AD ($n = 202$) and those with other brain pathologies (non-AD, $n = 197$). Total RNA extraction and QC were done using the Ambion RNAqueous kit and Agilent 2100 Bioanalyzer, respectively, according to published methods. Whole-genome DASL expression microarrays (Illumina) were used for the transcriptome measurements of RNA samples that were randomized across chips and plates using a stratified approach to ensure balance with respect to diagnosis, age, sex and RNA Integrity Numbers (RIN). Raw probe-level expression data exported from Genome Studio software (Illumina) were preprocessed with background correction, variance stabilizing transformation, quantile normalization and probe filtering using the lumi package of BioConductor^{46,47}. Preprocessed probe transcript levels were used in the downstream analysis.

Phenotypes. Our primary phenotype of interest in this manuscript was the burden of neuritic plaques, a quantitative measure of the amount of AD neuropathology in the brain at the time of death. Brain autopsies in ROS and MAP were performed across the US, as described previously^{42,43}. Bielschowsky silver stain was used to visualize neuritic plaques in tissue sections from five brain regions: the midfrontal, middle temporal, inferior parietal and entorhinal cortices, and the hippocampal CA1 sector. As in prior publications¹⁷, a quantitative composite score of neuritic plaque burden was then computed for each individual by dividing the subject's raw count in each of the five regions by the population s.d. in that same region, and then taking the average of the standardized counts across the five regions. Because the distribution of these average standardized counts is skewed, we used the square-root-transformed values in our statistical analyses.

To put the associations with neuritic plaque burden in the context of an AD diagnosis, we also assessed for associations with a neuropathologic diagnosis of AD, which is determined on post-mortem examination. Specifically, subjects were classified as having a pathologic diagnosis AD if they had intermediate or high likelihood of AD based on the National Institute on Aging (NIA)-Reagan criteria. The NIA-Reagan criteria, which integrates both the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) estimates of neuritic plaque density

and Braak staging of neurofibrillary tangle pathology, was implemented as reported⁴⁸. These diagnoses are made by board-certified neuropathologists without access to the clinical data collected during the study. All neuropathologic data was collected in a blinded fashion (relative to clinical diagnosis) by the neuropathology staff.

Experimental protocol for DNA extraction from post-mortem brain. 100-mg sections of frozen dorsolateral prefrontal cortex were obtained from each of 761 deceased subjects from the ROS and MAP studies based at the Rush Alzheimer's Disease Center. These sections were thawed on ice, and the gray matter was carefully dissected from the white matter. DNA extraction was performed using the Qiagen (cat. #51306) QIAamp DNA mini protocol. The Qubit 2.0 Fluorometer was used to quantitate the DNA. 16 μl of DNA at a concentration of 50 $\text{ng } \mu\text{l}^{-1}$ as measured by PicoGreen was used by the Broad Institute's Genomics Platform for data generation by the Illumina InfiniumHumanMethylation450 bead chip assay. The platform produces a data file by implementing the recommended procedures of the proprietary Illumina GenomeStudio software, which includes color channel normalization and background removal. All data generation was conducted by laboratory personnel who were blinded as to the clinical and neuropathological phenotypes of each subject

Subject and probe quality control. For the initial quality check of the data, we used the detection P value criteria recommended by Illumina. These P values represent the quality of the probes compared to background noise. We selected probes that had detection $P < 0.01$ for all samples to ensure the use of good quality probes. Using these criteria, we selected 470,913 out of a total of 485,577 tested probes for further analysis. Of these 470,913 probes, 460,045 are found in one of the autosomes.

However, not all probes are unique: some probes are predicted to cross-hybridize with the sex chromosomes⁴⁹ based on sequence alignment. Many of these probes showed strong association to gender in a recent report that recommends discarding probes in which 47 of 50 nucleotides match the sex chromosome sequence during sequence alignment using BLAT⁴⁹. Specifically, the authors recommend discarding 29,233 probes from the Illumina Infinium HumanMethylation450 bead chip assay that meet this sequence match criterion. We implemented this recommendation and removed the 29,233 probes from our probe list for downstream analysis. Furthermore, as noted below, we adjusted for gender in our analytic model given the well-described influence of sex on methylation levels.

In addition to cross-reactive probes, a substantial fraction of CpG probes also overlap with known SNP sites based on the 1000 Genomes database⁴⁹. Methylation level of these CpGs could be affected by a subject's genotype. However, SNPs with very low minor allele frequencies should not have major effect on the methylation data given our sample size. Thus, we removed CpGs in which a SNP with a minor allele frequency (MAF) ≥ 0.01 existed within 10 base pairs upstream or downstream of the CpG site, where single base extension occurs. Based on a previous analysis⁴⁹, we found a total of 14,964 autosomal, polymorphic CpGs with $0.01 \leq \text{MAF (EU)} \leq 0.99$ in subjects of European ancestry, such as our subjects. These CpGs were removed from consideration, leaving 415,848 autosomal CpGs for downstream analysis.

We also evaluated the 'per subject' quality of the data after the initial CpG data cleaning. The first step in this component of our quality control pipeline was to remove subjects with poor quality data from further consideration. Specifically, we used principal component analysis (PCA) to select subjects that were within ± 3 s.d. from the mean of a principal component (PC) for PC1, PC2 and PC3. PCA was performed using a random selection of 50,000 autosome probes for all of the samples. Using these criteria, we removed 13 subjects from a total of 761 subjects for which the Illumina assay was attempted. We excluded an additional 40 samples for having poor bisulfite conversion efficiency. Hence, after our 'per subject' quality check, a total of 708 samples were selected for downstream analyses.

Normalization of the data. We observed a strong batch effect in our data due to the use of two different thermocyclers during data generation. This batch effect was also confirmed by our PCA analysis. We evaluated different approaches to normalize our data, including COMBAT⁵⁰ and independent component analysis (ICA). However, when applying these approaches to our data and comparing

them with an approach in which we adjust for given confounding variables (such as batch number and age) that we find to be associated with principal components within our data, these approaches (COMBAT and ICA) are overly conservative in normalizing our data. In particular, ICA includes adjustments for many unknown variables that capture structure within a set of data and are not associated with known confounders. Thus, such ICA-derived variables could be related to elements of disease pathophysiology that is not yet understood.

ICA is a matrix factorization technique that separates a matrix into statistically independent, non-Gaussian factors using an R package *fastICA* described at <http://cran.r-project.org/web/packages/fastICA/index.html>. We applied ICA to the matrix of methylation beta values to infer the number of statistically independent components k iteratively from 1 to 40. In each instance, all samples are assigned k surrogate variables representing k entries of the mixing matrix. We then performed a refinement step similar to that used in *iSVA* (ref. 51) where for each independent component i ($1 \leq i \leq k$) we created a sub-matrix of the beta values populated only by methylation probes that were significantly associated with component i . We then performed ICA on this sub-matrix and selected the independent component most correlated with component i , and used those refined components as the surrogate variables. To set an optimal k , we explored the average variance explained in the methylation matrix by each value of k , and determined that 7 was a conservative number of components that explained an appreciable percentage of the variance. These seven surrogate variables were then used in our evaluation of normalization methods.

In our comparison of the results obtained using the two different approaches (including known technical confounders such as batch versus COMBAT and ICA normalization), we found that the top 71 CpG in the primary analysis remained significant regardless of the approach used and that, although the P value of the other 67 CpGs that were significant in our primary analysis fluctuated below our threshold of significance, they retained strongly suggestive evidence of association to the neuritic amyloid trait (95% remain at a $P < 5 \times 10^{-5}$). As a result, and given our strategy of validating the CpG analysis with two rounds of validation to minimize false positives (Fig. 1), we opted to use the batch variable approach in the primary analysis to directly address the source of technical variation and avoid over-correcting our data for variables that are not demonstrated to be associated with technical or demographic confounders. Thus, we prefer to minimize the risk of false negative results and minimize false positives with two rounds of validation.

Accounting for differences in the proportion of cell types in our tissue. Given that the proportion of neurons and other cell types found in the human cerebral cortex can change with AD, we evaluated a technique to account for the possible difference in the proportion of neuronal cells in our brain samples. As discussed in the main text, we used an R package⁵² to quantify the proportion of NeuN⁺ cells (primarily neurons) in each brain sample using DNA methylation data. We used the data from the NeuN⁺ nuclei found in this package to create convex combinations of purified profiles using nonlinear least-squares. This yielded an estimate of the proportion of NeuN⁺ nuclei in each of our samples. However, the resulting measure was not significantly associated with a pathologic diagnosis of AD ($P = 0.08$), and we therefore did not include it as a covariate in our primary analysis. However, we added it as a covariate in a secondary analysis that led us to identify the most conservatively associated subset of 71 CpG.

Data analysis and statistical modeling. For our primary and secondary analyses, we used the β values reported by the Illumina platform for each probe as the methylation level measurement for the targeted CG site in a given sample instead of M values (logistically transformed β values)⁵¹. Although M values have certain favorable statistical properties relative to β values, they are less biologically interpretable than β values⁵³. We therefore opted to use β values, which range from 0 (no methylation) to 1 (100% methylation) and show good correlation to estimates of DNA methylation derived from whole genome bisulfite sequencing approaches (see below for details). Any missing β value was imputed using a k -nearest neighbor algorithm for $k = 100$.

To perform our methylome-wide association study (MWAS) and discover differentially methylated regions associated with neuritic plaque pathology, we used a linear model, adjusting for age at death, sex, study (ROS or MAP), experimental batch and bisulfite conversion efficiency. A logistic regression was used for the AD analysis, using the same covariates. As noted above, we considered,

but elected not to include, terms for (1) the proportion of neurons in each sample or (2) surrogate variables that displayed no correlation with available technical or demographic variables. To account for the testing of multiple hypotheses, we used a Bonferroni correction that yielded a $P < 1.20 \times 10^{-7}$ as the threshold for genome-wide significance given the 415,848 autosomal CpG probes tested in our analysis. For annotation of the CpG probes, we used the hg19 human reference genome.

The bisulfite conversion efficiency term is calculated using the bisulfite conversion control probes, based on Illumina guidelines. Ten CpG sites designated by Illumina as control sites (6 CpGs targeted by type I probes and 4 CpGs targeted by type II probes), where we expect each CpG to be 100% methylated, are used to control for non-complete bisulfite conversion. The bisulfite conversion efficiency term used in the primary analysis is the median methylation estimate from the ten control sites. The bisulfite conversion term is calculated by taking the median value of the probes that Illumina provides to estimate bisulfite conversion efficiency.

To assess whether changes in DNA methylation are an early feature of AD, we compared the results of our association analysis obtained from the full set of subjects ($n = 708$) to those obtained from the subset of subjects who were cognitively non-impaired at the time of death ($n = 217$) by assessing, for a given CpG, whether the beta estimate from the non-impaired subset was different from the beta estimate derived from the entire cohort. We used the fact that the non-impaired subjects are a subset of the entire cohort and tested, using a one-sample test and the t distribution, whether the results from this subset, specifically the beta estimate taking into account the standard error, were different from those measures obtained from the entire population (that is, the total cohort of 708 subjects) from which the subset was drawn.

Bisulfite-sequencing data generation and analysis. For four individuals (two with AD and two non-impaired, each pair consists of a man and a woman), we used the same DNA sample profiled using the Illumina HumanMethylation Array to perform whole genome bisulfite sequencing. Genomic DNA was fragmented to 100–500 bp using a Covaris S2 sonicator. DNA fragments were end-repaired, A-tailed and ligated with methylated paired-end adaptors (purchased from ATDBio). Whole-genome bisulfite sequencing (WGBS) was performed as previously described⁵⁴. In short, Illumina genomic DNA adaptors were added to the fragments, and the adaptor-ligated fragments were size-selected before two rounds of 5-h bisulfite treatments using the Epitect Bisulfite kit (Qiagen). Libraries were then purified and run on the Illumina HiSeq2000 using a standard 36 base protocol⁵⁴.

The WGBS libraries were aligned using BSMAP 2.7 (ref. 55) to the hg19/GRCh37 reference assembly. Subsequently, CpG methylation calls were made using custom software⁵⁶, excluding duplicate, low-quality reads as well as reads with more than 10% mismatches. Only CpGs covered by $\geq 5\times$ reads were considered for further analysis.

Methylation profile from four samples (two with AD; two non-AD) generated from Illumina450k as well as bisulfite-sequencing, were compared by selecting a random 50,000 autosomal CpG sites that are present in both data sets for each individual. In each subject, the estimated level of methylation at each CpG measured by the two technologies was compared using a Student's t test. The results of the comparison of the technical replicates (Illumina versus WGBS) are as follows: subject #ROS20963578, $r = 0.975$; subject #MAP5797875, $r = 0.969$; subject #MAP50403446, $r = 0.971$; subject #ROS20214850, $r = 0.972$. Across the four subjects, the mean $r = 0.972$.

Chromatin state map. In collaboration with the Broad Institute's Roadmap Epigenomic Mapping Center, chromatin immunoprecipitation using antibodies targeting six different chromatin marks (H3K36me3, H3K27me3, H3K4me1, H3K4me3, H3K9ac and H3K9me3) was performed independently for two MAP subjects who were cognitively non-impaired at the time of death and had minimal evidence of AD, vascular and lewy body pathology on neuropathological examination. In both cases, samples were obtained from the dorsolateral prefrontal cortex that is also sampled in our DNA methylation scan. Chromatin was extracted from each of the cortical samples. Library construction and sequencing were performed as previously reported⁵⁷. Chromatin state maps were then learned from the sequence data (available at <http://www.roadmapepigenomics.org>). The H3K36me3 and H3K27me3 data were from donor id 149 while the

other data sets were from donor id 112. The data were dichotomized using the default settings of the *BinarizeBed* command of *ChromHMM*⁵⁸ except applying the *-center* option to the already signal-extended bed files. The models were trained by applying the *LearnModel* command with default settings from *ChromHMM*. We selected a model with 11 distinct chromatin states as the optimal model for our tissue sample.

To assess whether certain chromatin states are enriched for associations in our analysis, we compared the distribution of the 71 associated CpGs across the 11 chromatin states to that of all 415,848 probes tested using a chi-square test.

RNA data and analysis. A detailed description of the RNA data from the Mayo Clinic samples has been reported elsewhere^{44,45}. Given that the analysis here explored the relation of mRNA levels to an AD diagnosis, a multivariable linear regression model was implemented, adjusting covariates to correct for technical or biological variables, including age at death, sex, PCR plate, RIN and (RIN-RINmean)². For those analyses that corrected for the expression levels of genes that are specific for the main five cell types present in the CNS, the following probes were used as covariates: *ENO2* for neurons (ILMN_1765796), *GFAP* for astrocytes (ILMN_1697176), *CD68* for microglia (ILMN_2267914), *OLIG2* for oligodendrocytes (ILMN_1727567) and *CD34* for endothelial cells (ILMN_1732799). Some or all of these five expression levels were included to account for neuronal loss, gliosis and/or vascular tissue in the assessed brain regions, where indicated.

Pathway analysis. We constructed a protein-protein interaction (PPI) network using the online tool DAPPLE⁵⁹ to determine whether the genes identified in our DNA methylation study significantly interact with known AD associated proteins discovered in genetic studies. We compiled a list of genes associated with AD, including 25 late-onset AD GWAS genes, 3 early onset AD associated genes and *TREM1* (J. Replogle, D.A.B. & P.L.D., unpublished data). We produced a PPI network with a cutoff of 2 interacting binding degrees. DAPPLE creates direct and indirect networks of connected proteins using evidence of physical interaction from the InWeb database, which contains 169,810 high-confidence pair-wise interactions involving 12,793 proteins⁶⁰. To assess the statistical significance of PPI networks, DAPPLE applies a within-degree node-label permutation strategy to build random networks that mimic the structure of the original network and evaluates four network connectivity parameters on these random networks to generate empirical distributions for comparison to the original network. Our genetically defined AD network was significantly connected based on its Direct Edge Count ($P = 0.0072$) and Seed Common Interactors Degrees Mean ($P = 0.037$). Once the seven genes that emerge from the RNA functional validation

analysis are added to the list of genes to be considered in the analysis, the PPI network is expanded and includes six of these genes. In this iteration of the analysis, both the Direct Edge Count ($P = 0.0072$) and Common Interactors Degrees Mean ($P = 0.042$) measures remain significant.

42. Bennett, D.A. *et al.* Overview and findings from the rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–663 (2012).
43. Bennett, D.A., Schneider, J.A., Arvanitakis, Z. & Wilson, R.S. Overview and findings from the religious orders study. *Curr. Alzheimer Res.* **9**, 628–645 (2012).
44. Allen, M. *et al.* Novel late-onset Alzheimer disease loci variants associate with brain gene expression. *Neurology* **79**, 221–228 (2012).
45. Zou, F. *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet.* **8**, e1002707 (2012).
46. Du, P., Kibbe, W.A. & Lin, S.M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
47. Lin, S.M., Du, P., Huber, W. & Kibbe, W.A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.* **36**, e11 (2008).
48. Bennett, D.A. *et al.* Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* **66**, 1837–1844 (2006).
49. Chen, Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
50. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
51. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
52. Guintivano, J., Aryee, M.J. & Kaminsky, Z.A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290–302 (2013).
53. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13**, 705–719 (2012).
54. Gifford, C.A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).
55. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPPING program. *BMC Bioinformatics* **10**, 232 (2009).
56. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
57. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
58. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
59. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7**, e1001273 (2011).
60. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).