nature communications

Article

https://doi.org/10.1038/s41467-024-55698-6

regionalpcs improve discovery of DNA methylation associations with complex traits

Received: 8 April 2024

Accepted: 18 December 2024

Published online: 03 January 2025

Check for updates

Tiffany Eulalio[1] ✉, Min Woo Sun[1], Olivier Gevaert[1,2], Michael D. Greicius[3], Thomas J. Montine[4], Daniel Nachun[4,5] ✉ & Stephen B. Montgomery[1,4,5] ✉

We have developed the *regionalpcs* method, an approach for summarizing gene-level methylation. *regionalpcs* addresses the challenge of deciphering complex epigenetic mechanisms in diseases like Alzheimer's disease. In contrast to averaging, *regionalpcs* uses principal components analysis to capture complex methylation patterns across gene regions. Our method demonstrates a 54% improvement in sensitivity over averaging in simulations, providing a robust framework for identifying subtle epigenetic variations. Applying *regionalpcs* to Alzheimer's disease brain methylation data, combined with cell type deconvolution, we uncover 838 differentially methylated genes associated with neuritic plaque burden—significantly outperforming conventional methods. Integrating methylation quantitative trait loci with genome-wide association studies identified 17 genes with potential causal roles in Alzheimer's disease risk, including *MS4A4A* and *PICALM*. Available in the Bioconductor package *regionalpcs*, our approach facilitates a deeper understanding of the epigenetic landscape in Alzheimer's disease and opens avenues for research into complex diseases.

DNA methylation is a key component of the epigenome responsible for regulating gene expression, holding considerable promise as a disease biomarker and a therapeutic target. Predominantly occurring at cytosine-phosphate-guanine (CpG) sites, this epigenetic modification of DNA is influenced by both inherited genetic variation and environmental factors[1–5]. Specific methylation changes have been associated with neurological disorders, notably Alzheimer's disease (AD)[6–9], suggesting that DNA methylation may be helpful in identifying novel disease treatments. The applications of demethylating agents to treat acute myeloid leukemia and myelodysplastic syndrome are among a growing number of methylation-based therapeutics[10–13]. DNA methylation changes may also help identify genes and pathways relevant to disease biology that are more difficult to detect with gene or protein expression due to often noisier data[14,15].

Despite many identified associations between DNA methylation and disease pathogenesis, interpretation of these findings remains challenging. Complex, sometimes contradictory, relationships exist between DNA methylation and gene expression. Methylation of promoters typically induces transcriptional silencing, while methylation within gene bodies can increase gene expression[16,17]. Like other functional genomics modalities, DNA methylation is also cell type-specific[15,18]. There is considerable variability in the magnitude and direction of associations reported between methylation and disease phenotypes[19].

Compounding these challenges in interpreting DNA methylation is the granularity at which it is assessed. Existing approaches typically analyze methylation at the level of single CpG sites, ranging from methylation arrays that quantify several hundred thousand CpG sites to whole-genome methylation sequencing consisting of nearly 30 million CpG sites[20–22]. Although many studies have investigated CpG-level changes in methylation across disease states[23–27], the functional implications of methylation changes at individual CpGs often remain

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2]Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine, Stanford University, Stanford, CA, USA. [3]Department of Neurology & Neurological Sciences, Stanford University, Stanford, CA, USA. [4]Department of Pathology, Stanford University, Stanford, CA, USA. [5]These authors contributed equally: Daniel Nachun, Stephen B. Montgomery.
✉e-mail: eulalio@alumni.stanford.edu; dnachun@stanford.edu; smontgom@stanford.edu

ambiguous. Most studies map CpG sites of interest to the nearest genes to interpret their biological impact[23,28,29]. However, this approach does not account for the cumulative and regional effects of changes in multiple CpG sites on gene function and disease pathology. Some studies attempt to model regional changes in methylation by systematically segmenting the genome into regions spanning 100–1000 base pairs to identify differentially methylated regions (DMRs)[30,31]. One intrinsic limitation of DMR-based approaches is that they depend on arbitrary parameters such as region size, which can significantly affect results and their biological interpretation[32].

An alternative solution to improve interpretability is to summarize methylation at the level of specific genomic regions, such as promoters or CpG islands. Research by Cai et al. demonstrated that methylation aggregated at the gene level more effectively discriminated between disease groups than individual CpG-level measurements[33]. While averaging across CpG sites within these regions is the most common aggregation strategy[34–38], this approach may oversimplify complex correlation structures between CpG sites across a region. Kapourani and Sanguinetti underscored this oversimplification by demonstrating that large genomic regions (±7 kb around the transcription start site; TSS) exhibit complex methylation patterns that influence gene expression and that such phenomena were not adequately captured by averaging CpG-level methylation across the region[39]. Zheng et al.[40] further demonstrated that gene-region methylation profiles can delineate disease subgroups, emphasizing the nuanced complexity of methylation at the gene level. Thus, there is a need for an effective method of capturing methylation signals within gene regions to improve the detection of biologically relevant methylation changes.

We have developed the regional principal components (rPCs) method for aggregating methylation data to improve gene region methylation summaries. This method leverages principal components analysis (PCA) within genomic regions to capture methylation changes pertinent to disease with increased accuracy. By capturing multiple orthogonal axes of variance across CpG sites, rPCs offer a robust representation of methylation data. To facilitate its broad application in the research community, we have made our software available as an R package, *regionalpcs*, available on Bioconductor[41] and GitHub at https://github.com/tyeulalio/regionalpcs.

Our study used *regionalpcs* to identify cell type-specific methylation changes associated with AD and genetic variation within the ROSMAP cohort[7]. AD, a neurodegenerative disorder affecting approximately 50 million people globally, is marked by progressive memory loss and cognitive decline[42–44]. This condition severely diminishes patient quality of life, overburdens caregivers, and exhausts healthcare infrastructures. These issues highlight an urgent need for effective interventions. Presently, the absence of definitive prevention or treatment strategies underscores a gap in our comprehension of the underlying mechanisms driving AD. Research implicates DNA methylation changes as a contributor to AD pathology, with hundreds of regions identified as differentially methylated in the context of the disease[45]. The ROSMAP cohort comprising clinical, lifestyle, genetic, and methylation data from over 3200 older adult participants offers an invaluable resource for investigating the epigenetic landscape of AD, which can facilitate insights into the disease's complex etiology.

We demonstrate a comprehensive AD methylation analysis by applying regionalpcs in the ROSMAP cohort. Differential methylation (DM) analysis is used to discern complex links between DNA methylation and AD phenotypes. We then map quantitative trait loci (QTL) to identify genomic regions where genetic variants affect DNA methylation and associate those variants with the genetic risk of AD. Across these analyses, *regionalpcs* reveals cell type-specific epigenetic changes associated with AD pathology and disease risk that would not be captured by CpGs or averages alone. *regionalpcs* balance the reduced multiple testing burden and greater interpretability of regional aggregation of methylation with the granularity of CpG-level analysis. Our method is broadly applicable to any analysis associating methylation with phenotypes or mapping methylation QTLs. We anticipate that this approach will contribute to improved identification of novel treatment targets and disease risk prediction.

Results

Methylation analysis can be simplified through gene-level summaries

We developed the *regionalpcs* package to summarize methylation over genomic regions using PCA (Fig. 1a). We chose to use PCA because it is a simple, computationally efficient approach with a well-developed theoretical framework. The rPCs summarize the highly correlated CpG features into a new set of orthogonal features that capture more information about methylation in a region than averaging but still provide a low-dimensional representation of the data for downstream analyses.

PCA is commonly used to simplify high-dimensional data by mapping it to a lower-dimensional space. In our approach, PCA typically reduces the methylation data from dozens of CpG sites in a gene region to just a few rPCs that explain the most significant patterns of variation. Each principal component is a linear combination of CpG methylation values, with high-variance CpGs contributing the most to the rPCs (Supplementary Fig. 1). This ensures that rPCs retain the most informative variation for identifying methylation changes. A key additional advantage of using PCA is that there are well-established methods for identifying how many eigenvectors capture a distinguishable signal from random noise. We used the Gavish-Donoho method[46], designed to identify the optimal eigenvalue threshold for minimizing the asymptotic mean squared reconstruction error of the original data. We have also implemented the less conservative Marchenko-Pastur method as an alternative approach to choosing the number of components[47].

This work primarily summarizes methylation at the level of full gene regions to facilitate a more straightforward interpretation of biological pathways—context often lost when focusing on individual CpG sites. However, our framework can accommodate a variety of regional annotations, such as promoters, CpG islands, intergenic enhancers, or custom regions specified by the user.

rPCs detect subtle methylation differences in simulated data

We evaluated the effectiveness of rPCs for identifying DM using the *RRBS-sim*[48] framework to simulate reduced representation bisulfite sequencing (RRBS) data. RRBS data can be seen as an intermediary between the sparse coverage of methylation microarrays and the denser coverage of whole-genome bisulfite sequencing. We compared rPCs against traditional averaging for summarizing methylation changes across regions. Averages were computed as the mean methylation across all CpGs within a region, while rPCs were derived through PCA of CpG-level methylation (Fig. 1a).

In our first simulation involving 1000 regions (50 CpG sites per region, 50 individuals split evenly between cases and controls), rPCs consistently showed higher performance than averaging. When 25% of CpGs were DM, rPCs detected a median of 73.1% of DM regions, compared to just 19.1% with averages. As the proportion of DM sites increased to 75%, rPCs identified 99% of the cases, compared to a 57.4% detection rate with averages (Fig. 1b). The magnitude of methylation differences between cases and controls and sample size were critical factors in DM detection, while the number of CpGs per region had a more subtle effect. With a modest 1% methylation difference, rPCs detected DM in 18.8% of regions, more than double the performance of averaging, which detected 8.4%. When the methylation difference was increased to 9%, rPCs identified nearly all (99.7%) DM regions compared to averaging (50.1%). With a sample size of 50, rPCs detected a

https://doi.org/10.1038/s41467-024-55698-6

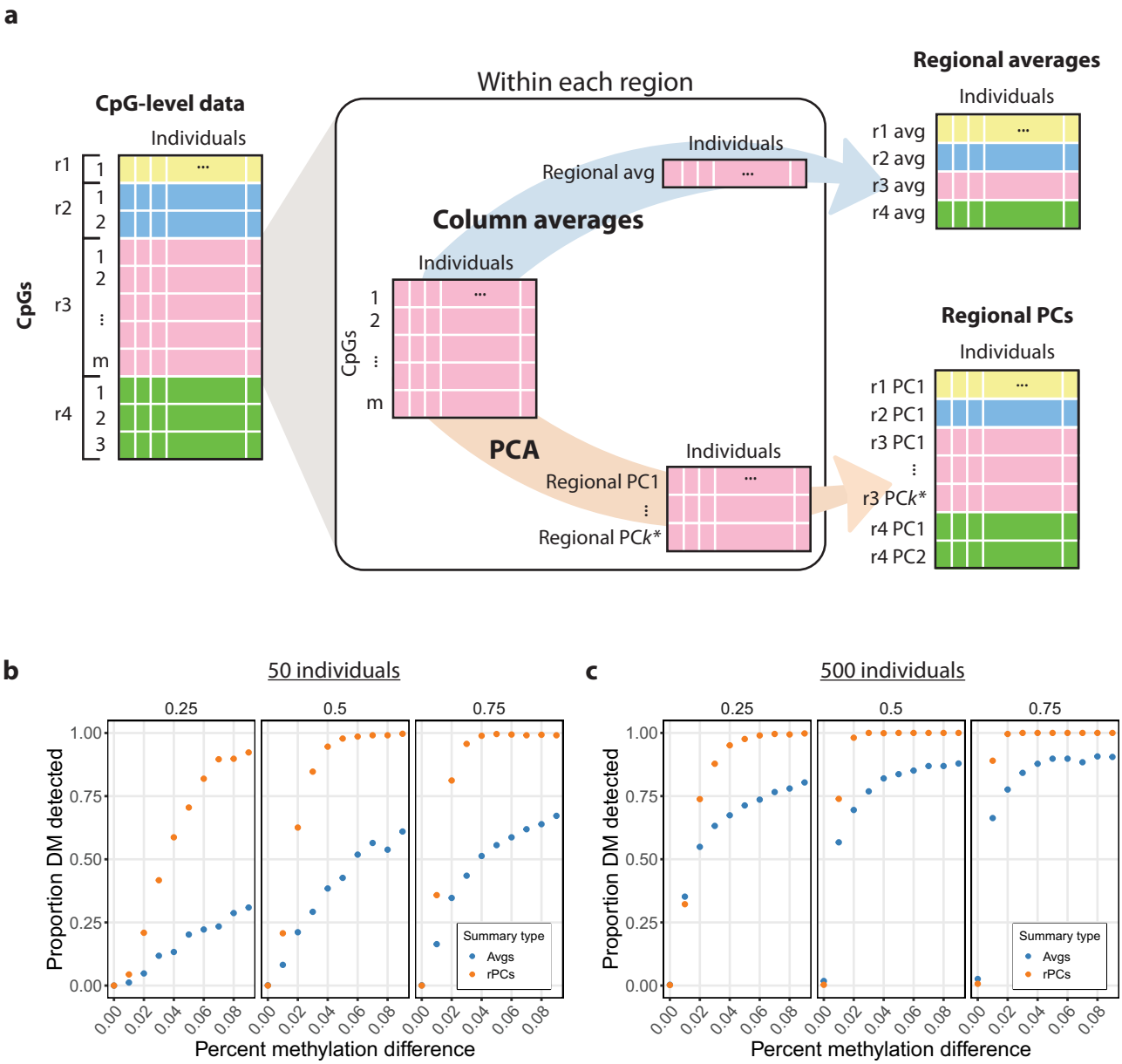

Fig. 1 | Overview of the *regionalpcs* method. a Summarizing regional methylation using averages and regional principal components (rPCs). The CpG-level data contains normalized methylation for CpGs (rows) and individuals (columns) grouped into regions based on genomic annotations (represented by colors). For regional averages, mean methylation across CpGs in each region is calculated for each individual. rPCs use principal components analysis (PCA) with the number of PCs (k^*) determined by the Gavish-Donoho method. **b**, **c** show comparative performance of summarizing regional methylation using rPCs (orange) and averages (blue) in simulations for **b** 50 individuals and **c** 500 individuals, highlighting the impact of larger sample sizes on methylation detection. Facets show the proportion of differentially methylated CpG sites. Each point indicates the proportion of regions (out of 1000 simulated) identified with significant differential methylation (y axis) at varying percent methylation differences between cases and controls (*x* axis). Simulations involved 50 CpG sites per region and samples evenly split between cases and controls. Source data are provided as a Source Data file. (PCA principal components analysis, PC principal component, Avgs averages, rPCs regionalpcs, DM differential methylation).

median of 94.4% of DM regions, compared to 32.6% with averaging. Increasing the sample size to 500 improved the detection capability of both methods, with rPCs detecting a median of 99.9% of DM regions, compared to 80.4% with averages (Fig. 1c). This finding demonstrates the improved sensitivity of rPCs, especially in studies with smaller sample sizes. In regions with 20 CpG sites, representative of RRBS data in promoter regions, rPCs detected a median of 78.2% of DM regions compared to only 45.4% with averaging. In regions with 50 CpG sites, representing full gene regions in RRBS data, the detection rate was 99.0% for rPCs compared to 59.1% for averages.

We also performed a second, larger simulation of 16,000 genes with 5% true DM genes, which is representative of the number of protein-coding genes in a typical tissue or cell type. We compared rPCs with established methods for detecting DMRs. rPCs achieved the highest precision (median 0.56) and F1 score (median 0.64), outperforming CpGs (precision 0.15, F1 score 0.26), DMRcate[30] (precision 0.54, F1 score 0.59) and methyKit[49] (precision 0.14, F1 score 0.22) (Supplementary Fig. 2). Both rPCs and DMRcate had the lowest type I error rate (0.03), while CpGs (0.27) and methylKit (0.23) had higher false positive rates. Summarizing methylation using averages failed to identify significant DM genes in the larger dataset, underscoring their limitations in complex analyses. Reducing the total number of genes to 1000 lowered the multiple testing burden, resulting in some significant associations with averages (Supplementary Fig. 3). In this

analysis, methylKit tended to have a higher false positive rate, however, adjusting the site coverage and increasing window size from 1000 (default) to 2000 reduced this false positive rate. These findings suggest that while rPCs are robust in detecting DM regions without any parameter tuning across many simulation contexts, the performance of other DMR detection methods like methylKit is more dependent on the choice of parameters and simulation design.

Recommendations for rPCs usage

For optimal use of rPCs, we recommend adhering to standard preprocessing practices, including filtering based on sequencing depth or array detection p values as well as CpG variability, to ensure high-quality methylation scores[49,50]. For normalization, users can apply the M value conversion[51] or consider the more robust inverse normal transformation (INT) to make the methylation values more normally distributed. However, while INT is more robust to outliers, it is a non-parametric method that cannot be converted back to the original methylation scale, as can be done with M values.

Our simulations demonstrate that rPCs robustly detect DM across a wide range of conditions, including the number of CpG sites within a region, the proportion of DM sites, and the magnitude of methylation differences between cases and controls (percent DM difference) (Supplementary Fig. 4). The most significant factor affecting performance was the combination of the percentage DM difference and sample size. Specifically, with a 0.09 percent methylation difference, accuracy differed by 3% between sample sizes 500 and 50. This difference in accuracy further increased to 11% at 0.04 percent methylation difference, and 27% at 0.00001 percent (Supplementary Fig. 4). Similar patterns were observed for F1 score and recall, while precision remained relatively stable across different sample sizes. At a 0.04% methylation difference, a sample size of 50 can achieve over 80% accuracy, precision, recall, and F1 score. These results offer practical guidelines for rPCs users, emphasizing the importance of balancing sample size and expected methylation differences for robust detection of DM regions.

Application of *regionalpcs* to AD

We used the *regionalpcs* package to summarize methylation in an AD cohort using real data. We analyzed paired DNA methylation and whole-genome sequencing data from a subset of 563 older individuals from the ROSMAP cohort[7]. Data were obtained post-mortem from the dorsolateral prefrontal cortex (DLPFC). Individuals were classified through annual and post-mortem clinical assessments into Alzheimer's Disease Dementia (ADD, 42.8%), Mild Cognitive Impairment (MCI, 25.4%), No Cognitive Impairment (NCI, 30.2%), or Other Dementia (1.6%) categories (Supplementary Table 1). In addition, neuropathological indices for neurofibrillary tangles[52] and neuropathological diagnoses based on CERAD scores for neuritic plaques were also recorded[53].

We applied *regionalpcs* to identify methylation changes relevant to AD at multiple scales of analysis (Fig. 2). Starting with DNA methylation data from bulk tissue of the DLPFC, we utilized cell type deconvolution to isolate cell type-specific signals (Fig. 2a, b). Subsequently, we summarized this methylation data across various gene regions (Fig. 2c, d). We assessed how methylation summarization at the gene level using *regionalpcs* compared against traditional averaging and single CpG analysis in detecting changes linked to neuritic plaque burden (Fig. 2e). Moreover, we evaluated the effectiveness of *regionalpcs* versus averages for mapping meQTLs and integrating these findings with AD GWAS data (Fig. 2f, g).

Imputing cell type-specific DNA methylation from bulk tissue

Before applying the *regionalpcs* method on the AD cohort, we accounted for mixed cell type signals in the methylation data. Previous investigations into DM and methylation QTLs have primarily used bulk tissue samples consisting of a heterogeneous mixture of cell types[7,54–56]. Although these studies have yielded valuable insights, attributing the changes in methylation to cell type proportions or cell type specificity remained challenging[57]. Cell type heterogeneity often acts as a confounding factor in functional genomics data collected from bulk tissue samples[58,59].

We used EPISCORE[60] to estimate the proportions of each cell type. Aligning with existing literature[61–63], neurons were the most abundant cell type, followed by astrocytes, oligodendrocyte/OPCs, and endothelial cells. Microglia were not detected, likely due to their relatively low abundance in the DLPFC brain region[62]. Our subsequent analyses focused on the four cell types with reliably estimated proportions (Fig. 3a). Validation with an alternative method[64] for estimating neuron proportions demonstrated high consistency (Pearson $r = 0.8$, p value $< 1.2 \times 10^{-8}$; Supplementary Fig. 5).

We applied tensor composition analysis (TCA)[57], to decompose the bulk methylation data into cell type-specific profiles. Uniform Manifold approximation and projection (UMAP) analysis[65] showed distinct clustering of bulk and cell type-specific methylation profiles (Fig. 3b). K-means clustering analysis[66] showed higher mean Silhouette scores for cell type-specific clusters compared to the bulk, indicating more homogenous cell type-specific clusters (astrocytes = 0.85, endothelial cells = 0.85, neurons = 0.79, oligo/OPC = 0.78, bulk cells = 0.67).

Clustering analysis validated that our deconvolved methylation profiles clustered closely with published nuclei-sorted methylation profiles from matching cell types[26]. 45 of 53 nuclei-sorted neuron samples clustered with our deconvolved neurons and 40 of 42 nuclei-sorted oligodendrocyte samples clustered with our deconvolved oligodendrocyte/OPCs (Supplementary Table 2).

Summarizing brain methylation using *regionalpcs*

We used the *regionalpcs* package to summarize methylation signals across 16,417 protein-coding genes, focusing on the full gene region (5 kb upstream of the TSS to the end of the 3' untranslated region) (Fig. 2d). Additionally, we examined three sub-regions—promoters, 5 kb upstream (of TSS), and gene bodies. The count of CpGs per gene varied widely, with a median of 15 CpGs and a range from one to 872 CpGs (Supplementary Table 3). Each gene was summarized by one average value and one and thirteen rPCs, with an average of 1.04 rPCs per gene, explaining a median of 15% variance (Supplementary Table 3).

We confirmed that rPCs primarily reflected regional methylation signals rather than global factors like batch, sex, or age. We observed consistent correlation distributions among rPCs of the same rank using a correlation analysis of rPCs spanning multiple genes (see Methods). Briefly, we calculated the correlation distribution among rPCs of the same rank. If specific rPCs mainly reflect global signals, we would expect different distribution patterns across rPC ranks 1 through 3 (an illustrative example is provided in Supplementary Fig. 6a). We initially found that the rPCs were capturing methylome-wide signals as reflected by the top four methylome-wide PCs (Supplementary Fig. 6b, c). Following the removal of global methylation PCs, we observed minimal correlation among all rPC ranks and no notable differences between them (Supplementary Fig. 6d). These findings indicate that our rPCs represent region-specific signals.

The number of rPCs positively correlated with gene region complexity measures such as the number of CpGs (Pearson $r = 0.74$; Fig. 3d; Supplementary Fig. 7), CpG density (Pearson $r = 0.19$), gene length (Pearson $r = 0.12$), and signal variance measured by median absolute deviation (Pearson $r = 0.099$) (all p values < 0.0001). The bulk data yielded the fewest rPCs, suggesting that cell-type deconvolution aids in distinguishing methylation signals from noise.

This gene level summarization substantially reduced the number of features from 271,223 CpGs to 17,341 rPCs and 16,417 averages for

https://doi.org/10.1038/s41467-024-55698-6

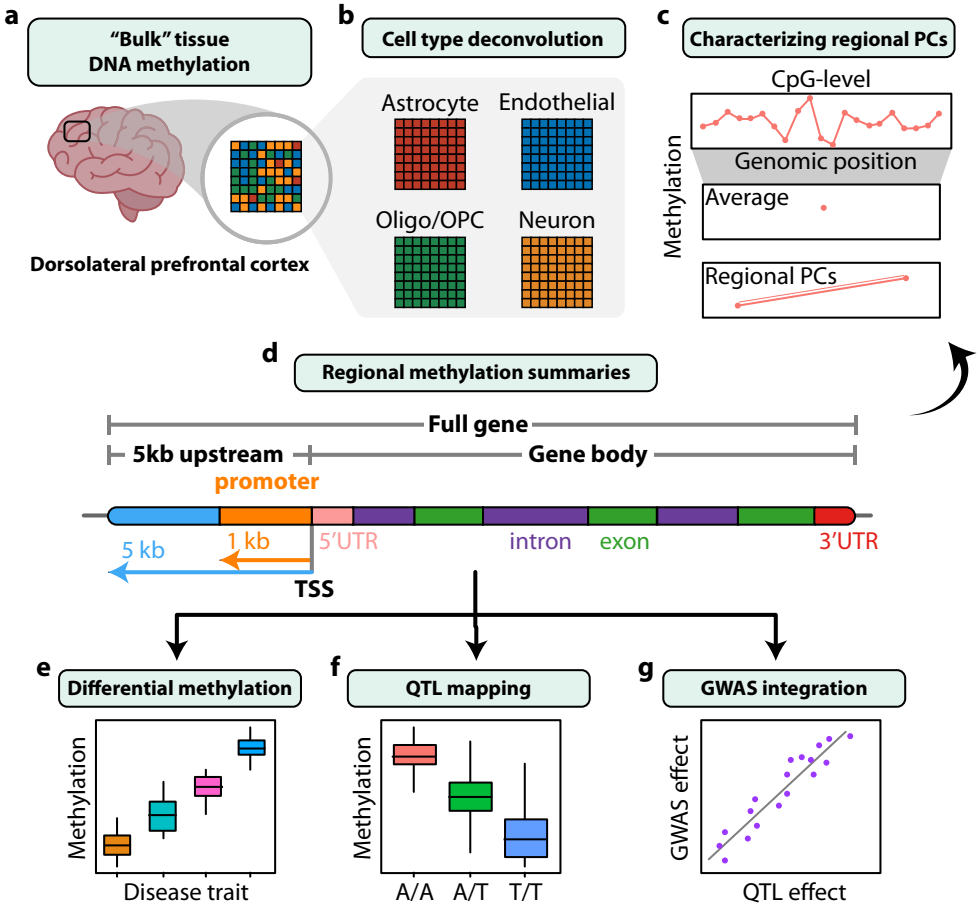

Fig. 2 | DNA methylation analysis in AD using the ROSMAP cohort. a Bulk methylation data were quantified from brain tissue using microarrays. **b** Cell type deconvolution isolated four cell type-specific signals from the bulk methylation data. **c** Regional CpG-level methylation data was summarized using averaging and *regionalpcs*. **d** Methylation was summarized across four types of gene regions: promoters, gene body, 5 kb upstream (of the transcription start site - TSS), and the full gene. **e**–**g** Toy examples illustrating the workflow for: **e** Differential methylation analysis was used to find changes in gene-level methylation across disease states and phenotypic traits within our cohort. **f** Quantitative trait loci (QTL) mapping was used with regional methylation summaries to identify genes with methylation levels associated with genomic variants. **g** A genome-wide association study (GWAS) for Alzheimer's disease (AD) risk was integrated with methylation QTLs to identify genetic variants impacting both methylation risk and AD risk. (GWAS genome-wide association study, kb kilobase pair, Oligo/OPC oligodendrocyte/oligodendrocyte progenitor cells, PCs principal components, TSS transcription start site, UTR untranslated region, QTL quantitative trait loci).

endothelial cells, with similar reductions for other cell types (Supplementary Table 4). This approach efficiently reduced multiple testing challenges and computational demands.

rPCs identify more AD-relevant DM

Our comparative DM analysis utilized *regionalpcs* to identify more disease-relevant methylation changes than traditional averaging or CpG-level assessments. We found notable DM associated with AD, mainly linked to neocortical neuritic plaques (CERAD score) and neurofibrillary degeneration (Braak stage). Interestingly, fewer associations were observed with *APOE* genotype and ADD diagnosis, suggesting that methylation endophenotypes might offer clearer biological distinctions than these clinical phenotypes alone (Supplementary Fig. 8).

For our comparison of methylation summary methods, we focused on the proportion of significantly differentially methylated features identified by each method. This approach accounts for the variability in the number of CpGs and rPCs representing each gene rather than using a single average value per gene (Fig. 2c). Across all cell types, rPCs consistently identified a significantly higher proportion of differentially methylated features associated with neuritic plaque density than both averages and individual CpGs (Supplementary Fig. 9). Specifically, rPCs identified a median of four times greater proportion of DM features than averages and 24 times greater than CpGs (Supplementary Table 5), highlighting increased power to capture methylation changes associated with neuritic plaque load.

At the gene level, rPCs detected more differentially methylated genes than averages. In astrocytes, rPCs identified 638 DM genes versus the 192 detected by averages (Supplementary Fig. 8, Supplementary Table 6). Astrocytes were the cell type with the largest number of associated genes for both neuritic plaques and tau, supporting previous studies underlining their critical role in AD[67–71]. Differentially methylated genes associated with neuritic plaques were also found in endothelial cells, neurons, and oligodendrocytes/OPCs (Fig. 4a). In contrast, analysis of bulk data identified few DM genes–10 using rPCs and none using averages–highlighting the inherent challenge in discerning methylation associations within mixed cell populations. Most genes (62%) associated with neuritic plaque burden were exclusive to a single-cell type, with astrocytes having the greatest number of unique genes (125 with averages, 324 with rPCs).

To assess gene-level results, CpG-level DM analysis requires a subsequent step of mapping DM CpGs to corresponding genes. When considering a gene region as differentially methylated if any CpG within the region is DM, CpG-level analysis identified multiple genes in astrocytes ($N = 1013$), exceeding rPCs ($N = 638$). However, CpGs seemed to capture localized changes–identifying a median of one DM

https://doi.org/10.1038/s41467-024-55698-6

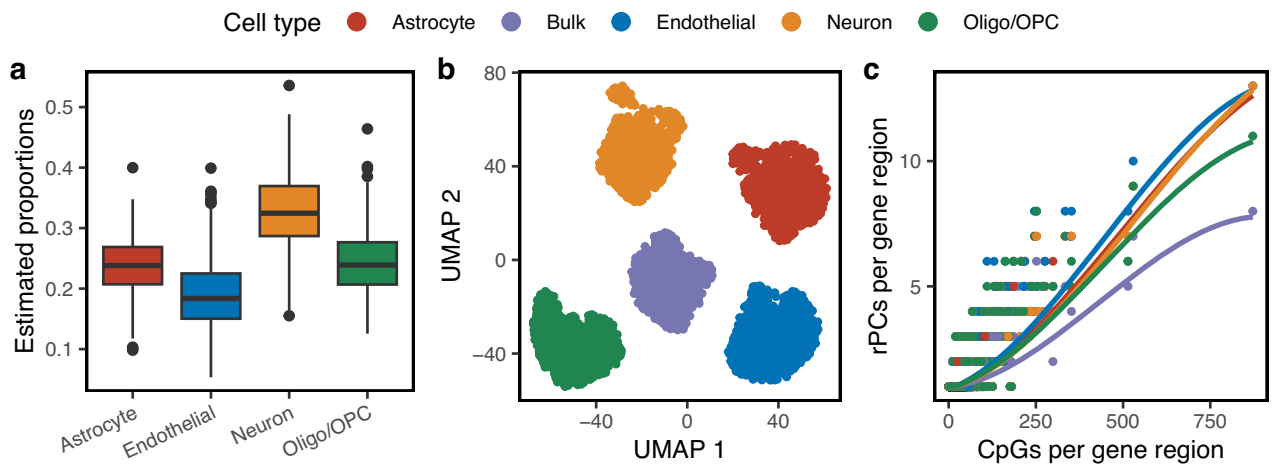

Fig. 3 | Cell type deconvolution imputes cell type-specific signals from bulk tissue DNA methylation. a Boxplots show the estimated cell type proportion distributions across four primary brain cell groups: astrocytes, endothelial cells, neurons, and oligodendrocytes/oligodendrocyte progenitor cells (Oligo/OPC). Each cell type group contains estimated proportions from 563 individuals. Horizontal lines within the interquartile range (IQR) boxes indicate the median values for each group, and whiskers represent 1.5*IQR from the IQR values. **b** UMAP visualization of deconvolved cell type-specific data across all samples. **c** Scatter plot of the association between the number of CpGs and the number of *regionalpcs* (rPCs) representing the full gene regions, colored by cell type. LOESS lines depict data trends for each cell type. All plots are derived from biological replicates ($N = 563$ individuals) and no technical replicates. Source data are provided as a Source Data file. (Oligo/OPC oligodendrocyte/oligodendrocyte progenitor cells, rPCs regionalpcs).

CpG within a full gene region consisting of a median of 25 total CpGs. While CpGs identified a similar amount of DM genes in oligo/OPCs (rPCs 354, CpGs 355), rPCs identified more genes in endothelial cells (rPCs 259, CpGs 180), neurons (rPCs 117, CpGs 46), and bulk (rPCs 10, CpGs 5).

We used the AD association score from the Open Targets database[72] to assess the relevance of these differentially methylated genes to AD. This AD association score reflects the level of evidence for each gene's involvement in AD. We correlated these scores with p values from DM analysis using rPCs or averages. A positive correlation was expected between higher AD association scores and more significant methylation changes. Indeed, positive correlations were observed across nearly all cell types for both rPCs, averages (Fig. 4b), and for CpGs. However, in bulk tissue, p values from averaging were negatively correlated with AD association scores. A linear interaction model revealed that rPCs showed significantly stronger correlations with AD-related genes than averages (p value $= 4.2 \times 10^{-6}$) and CpGs (p value $= 9.06 \times 10^{-7}$). Of the top 5 percent of genes ranked by AD association score (348 genes), 41 genes were identified exclusively with rPCs, compared to just one by averages alone; seven genes were identified by both methods (Fig. 4c). Among these, eight genes recognized by the National Institute on Aging Genetics of AD (NIAGADS) (https://adsp.niagads.org/gvc-top-hits-list/), such as *SORL1*, *PILRA*, and *PSENEN*, were identified by rPCs. Notably, the *PSENEN* gene[73] showed a significant association with neuritic plaque load only when analyzed with rPCs (Benjamini-Hochberg (BH) p value = 0.007), whereas no significance was found with averages (BH p value = 0.521) (Supplementary Fig. 10).

We also compared the results identified by the DMR detection method, DMRcate. DMRcate is built on limma-based analysis[74], which closely aligns with the methods used for our gene-level summaries and CpG-level analysis. At the gene-level, the CpG analysis mapped to the most unique genes ($N = 669$), followed by rPCs ($N = 261$), DMRcate ($N = 99$) and averages ($N = 37$) (Supplementary Fig. 11). A substantial number of genes were identified as DM by both CpGs and rPCs ($N = 228$), and a moderate set of genes were identified by CpGs, rPCs, and averages, but not DMRcate ($N = 88$). There were very few genes identified in common between DMRcate and any of the other summary methods. As there were no ground truth labels for these genes, further interpretation of these results is challenging. Nonetheless, the differences observed between the CpG-, gene-, and DMR-level approaches suggest that each method provides valuable information depending on the specific context of the investigation.

rPCs identify more significant associations between methylation and gene expression

We conducted an expression quantitative trait methylation (eQTM) analysis to assess the biological significance of rPCs in identifying genes with significant methylation-expression associations. Methylation was summarized using rPCs, averages, and individual CpGs. While CpGs identified slightly more eQTM genes (2–5% more, Supplementary Table 7), rPCs showed significantly larger median effect sizes for eQTMs in the full gene (11% increase) and gene body regions (19% increase; two-sided t-test, BH p value = 0.012 for both, Supplementary Fig. 12). Averages also displayed a 16% increase in effect size over CpGs in the full gene region (two-sided t-test, BH p value = 0.056), supporting the efficacy of aggregating methylation data[33].

Effect sizes were significantly greater in the full gene and gene body regions for rPCs, with median increases of 6–7% and 20% compared to preTSS and promoter regions when using rPCs, respectively (two-sided t test, BH p value < 0.0001). Averages also showed significant increase effect sizes in these regions, while CpG-level eQTMs did not show significant differences between region types. These findings suggest that summarizing methylation across entire gene regions provides stronger gene expression correlations than focusing on individual CpGs or smaller sub-regions.

rPCs identify more gene-level methylation QTLs

Next, we mapped methylation quantitative trait loci (meQTL) using rPCs, averages, and CpG-level analysis to assess their ability to capture genetic effects on gene-level DNA methylation across 16,417 genes. Focusing on common germline variants within one megabase (Mb) of the TSS, rPCs detected ~15% more genes with significant associations with at least one genetic variant (meGenes) than averages (Fig. 5). This represented a median increase of 1608 meGenes identified by rPCs across different cell types (Supplementary Data 1). A substantial proportion of the genes evaluated were meGenes—75% for rPCs and 66% for averages. There was minimal evidence of cell type specificity of

https://doi.org/10.1038/s41467-024-55698-6

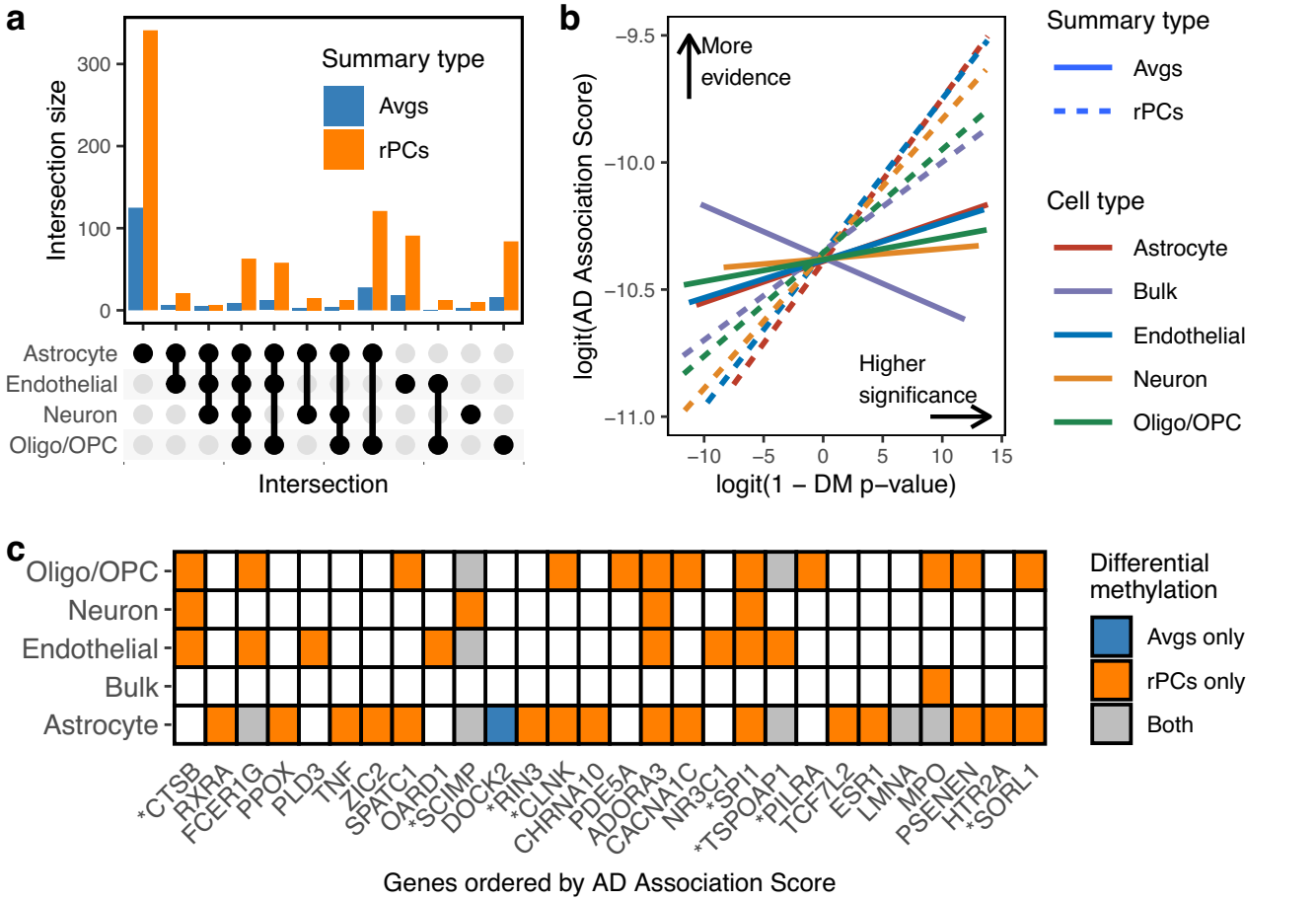

Fig. 4 | Regional PCs identify more associations between gene-level methylation and neuritic plaque burden. a UpSet plot showing the intersection of associations between gene-level methylation and neuritic plaque burden across cell types. **b** Line plot of the correlation between differential methylation significance levels and AD association scores, using dashed lines for rPCs and solid lines for averages, color-coded by cell types. **c** Tile plot presents genes with high evidence for AD association, based on the top 5% of AD association scores derived from the Open Targets AD gene list. Genes marked with * are found in the National Institute on Aging Genetics of Alzheimer's Disease (NIAGADS) list of high-priority AD genes. Color-coded tiles indicate significant associations between gene methylation and neuritic plaque burden (linear regression, Benjamini-Hochberg corrected p value < 0.05) by each summary type: blue tiles for rPCs only, orange tiles for averages, and gray tiles for genes identified by both methods. Data are based on $n = 563$ biological replicates, and no technical replicates were used. Source data are proved as a Source Data file. (avgs averages, AD Alzheimer's disease, DM differential methylation, Oligo/OPC oligodendrocytes/oligodendrocyte progenitor cells, rPCs regionalpcs).

meGenes using averages or rPCs across all region types because most genes were meGenes across all cell types (Supplementary Fig. 13).

We refined these associations by performing statistical fine-mapping to identify credible sets of causal variants which likely influence methylation. After fine-mapping, a small fraction (4%) of meGenes harbored strong credible sets with a cluster posterior inclusion probability (CPIP) >0.5 (Fig. 5). rPCs outperformed averages, identifying a median of 463 meGenes, compared to 386 with averages across cell types. The overlap of meGenes (median of 95) between rPCs and averages was limited, suggesting that each approach captures distinct signals across cell types (Supplementary Fig. 14a–e). Despite differences in meGene detection, the median credible set size remained similar across methods–69 for rPCs and 64 for averages. Similarities across set sizes indicate that the choice of summary method for methylation does not alter the size of credible sets.

CpG-level meQTLs showed a considerable overlap of fine-mapped credible sets with those identified by rPCs. At a CPIP > 0.8, a median of 20 credible sets overlapped between CpGs and rPCs across cell types in the full gene region. Averages shared 14 credible sets with rPCs and 12 with CpGs. The median percentage of overlapping variants within credible sets was 41% between CpGs and rPCs, 38% between CpGs and averages, and 91% between rPCs and averages. CpGs captured more localized signals, typically identifying one meCpG per full gene region, with a median of 24 total CpGs in those regions. rPCs uniquely identified a median of four credible sets not detected by CpG-level or averages. rPCs identified the highest number of unique credible sets in neurons ($N = 8$), and the fewest in bulk tissue ($N = 2$).

The set of meGenes with strong credible sets was more cell type-specific than the full set of meGenes identified with QTL mapping. Endothelial cells had the greatest number of unique meGenes with strong credible sets (rPCs 259, averages 189), while oligodendrocyte/OPCs had the fewest (rPCs 94, averages 43) (Supplementary Fig. 13). rPCs identified 7% more cell type-specific meGenes with strong credible sets compared to averages, suggesting that rPCs can identify more cell type-specific meQTLs than averages.

rPCs identify more causal associations between methylation and AD risk through GWAS integration

We integrated our meQTLs with an AD GWAS[75] to identify potential causal associations between gene-level DNA methylation and AD risk. Our pipeline consists of three parts: colocalization with fastENLOC[76], instrumental variable (IV) analysis with probabilistic transcriptome-wide association study (PTWAS)[77], and the integration of these two analyses with INTACT[78].

https://doi.org/10.1038/s41467-024-55698-6

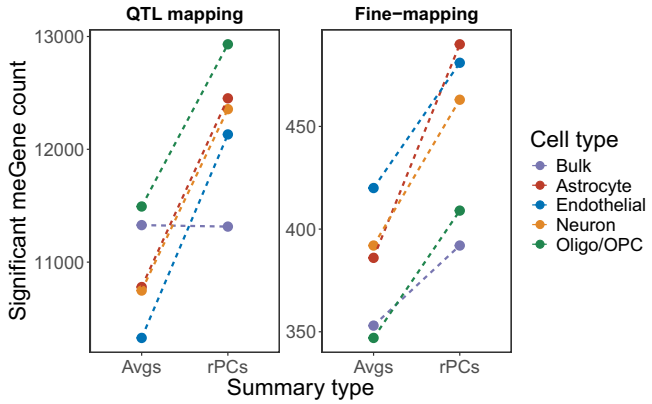

Fig. 5 | rPCs identify more meGenes than averages in meQTL. rPCs identified more genes with significant associations between methylation levels and genetic variants (meGenes) than averages. QTL mapping was performed using linear models in tensorQTL, and fine-mapping was conducted using a Bayesian framework with DAP-G. The y axis shows the number of genes identified as having significant associations with FDR-adjusted *p* values below 0.05 (QTL mapping) and genes further refined through fine-mapping with cluster posterior inclusion probabilities (CPIPs) exceeding 0.5 (Fine-mapping). Color variations denote different cell types. Using rPCs consistently results in a higher identification rate for genes of interest across both QTL mapping and fine-mapping stages. Source data are proved as a Source Data file. (avgs averages, oligo/OPC oligodendrocyte/oligodendrocyte progenitor cells, QTL quantitative trait loci, rPCs regionalpcs).

Using colocalization, we identified individual variants and loci likely causal for both a methylation QTL and AD. Approximately 1% of genes with strong meQTL credible sets colocalized with AD at a gene-level colocalization probability (GLCP) threshold of >0.5 (Supplementary Data 1), including four genes in the AD Open Targets database with an AD Association Score above the 88th percentile. One of these genes, *MS4A6A*, showed strong colocalization in neurons and is listed as a high-priority AD gene by NIAGADS. *MS4A6A* is a four-transmembrane domain protein reported to be differentially expressed in AD brains[79,80]. Colocalized genes were mostly cell type-specific, with astrocytes and neurons having the highest number of identified genes for rPCs and averages, respectively (Supplementary Fig. 15). These genes also had smaller credible sets than genes that were not colocalized (36% for rPCs, 44% for averages) (Supplementary Data 2), suggesting that colocalized genes tend to have smaller meQTL credible sets. None of these colocalized genes were within one centimorgan (cM) of the *APOE* locus, indicating independence from *APOE*-related risk variants.

We identified causal relationships between meQTLs and AD risk using IV analysis with PTWAS. Consistent with prior findings[81], PTWAS identified more genes associated with AD (43 genes) compared to colocalization (6 genes) at a genome-wide significance threshold (p value $< 5 \times 10^{-8}$, Supplementary Data 1). However, a substantial portion of PTWAS-identified genes (44% for rPCs, 32% for averages) were within 1 cM of *APOE*, likely due partly to "LD hitchhiking"[82]. LD hitchhiking can occur when strong meQTLs near *APOE* may not be risk variants for AD but are strongly correlated with AD risk variants in *APOE*. These discrepancies motivated the use of INTACT[78] to reconcile differences between colocalization and IV analysis. [81]Using INTACT, we identified 20 unique genes compared to 17 with colocalization and 95 with PTWAS (Supplementary Fig. 16).

Across four cell types and bulk tissue, we identified 17 high-confidence genes (posterior probabilities >0.6) with putative causal relationships between methylation and AD using rPCs, 12 of which were identified exclusively by rPCs (Fig. 6). Among these, *MS4A4A* and *PICALM* were high-priority AD genes identified by NIAGADS. *RELB*, identified as causal in oligodendrocyte/OPCs, was recently found to be an eGene in a cell type-specific expression QTL (eQTL) analysis[83], suggesting a potential multi-omic network affecting AD risk.

Among INTACT-identified genes, a substantial portion (33% for rPCs, 25% for averages) were within 1 cM of the *APOE* variant. Incorporating colocalization ensures genes are less likely to be false positives due to LD hitchhiking as was observed in our IV analysis. These results support previous studies that implicated DNA methylation in the vicinity of *APOE* in AD pathogenesis[84,85].

Methylation associations with neuritic plaques and genetic variation across different region types are complementary

We applied the rPCs method to summarize methylation in three specific sub-regions—promoters, 5 kb upstream (of TSS), and gene bodies—to explore methylation effects across different gene regions.

Across all cell types and traits, a greater number of differentially methylated genes were identified in the full gene and gene body regions compared to the promoter and preTSS regions (Supplementary Fig. 17). In astrocytes, where the strongest neuritic plaque-associated DM signal was detected, rPCs identified over three times more DM genes in the full gene ($N = 638$) and gene body ($N = 559$) regions than in the promoter ($N = 189$) and preTSS ($N = 93$) regions. Many DM genes were unique to specific regions (Supplementary Fig. 17).

Both eQTM and meQTL analyses reinforced these observations. More significant eQTMs were observed in the full gene and gene body regions, indicating that methylation across the entire gene contributes to gene expression variation (Supplementary Fig. 12). Gene-level methylation summaries in these larger regions also showed stronger associations with gene expression than in preTSS and promoter regions. Integration of meQTLs with GWAS revealed region-specific differences. For example, the AD gene *TOMM40*[86–88] showed high INTACT probability of 0.831 when summarized over the promoter compared to zero when summarized over the full gene region in oligodendrocytes/OPCs (Supplementary Fig. 18). These results indicate that while summarizing over full genes yields the largest number of associations between methylation and both disease phenotypes and genetic variation, there may still be value in analyzing smaller regions.

Discussion

In this study, we introduced rPCs as a novel approach for summarizing and interpreting gene-level methylation. rPCs provide an improved alternative to traditional CpG-level analysis and regional averaging. By focusing on methylation changes across various gene regions, we achieved a more complete understanding of methylation dynamics across the genome. The functional regions analyzed included promoters, 5 kb upstream (of TSS) areas, gene bodies, and full genes. Our method's effectiveness was further enhanced by employing cell type deconvolution, allowing us to summarize imputed cell type-specific methylation profiles derived from bulk tissue data. The accuracy of our deconvolution was supported by clustering our imputed cell-type-specific profiles with nuclei-sorted data.

Leveraging the *RRBS-sim* simulation framework, our comparative analysis highlighted the improvement of rPCs over traditional averaging in identifying differentially methylated regions. These simulations were designed to mimic RRBS data. Simulations varied in CpG site counts, sample sizes, proportions of differentially methylated sites, and mean methylation differences between cases and controls. The rPCs consistently outperformed averages in detecting differentially methylated regions, especially in scenarios with smaller sample sizes and fewer differentially methylated sites. This increased sensitivity of rPCs to subtle methylation differences was observed without yielding false positive results. These findings underscore the robustness of rPCs across a diverse range of study designs and types of methylation data.

https://doi.org/10.1038/s41467-024-55698-6

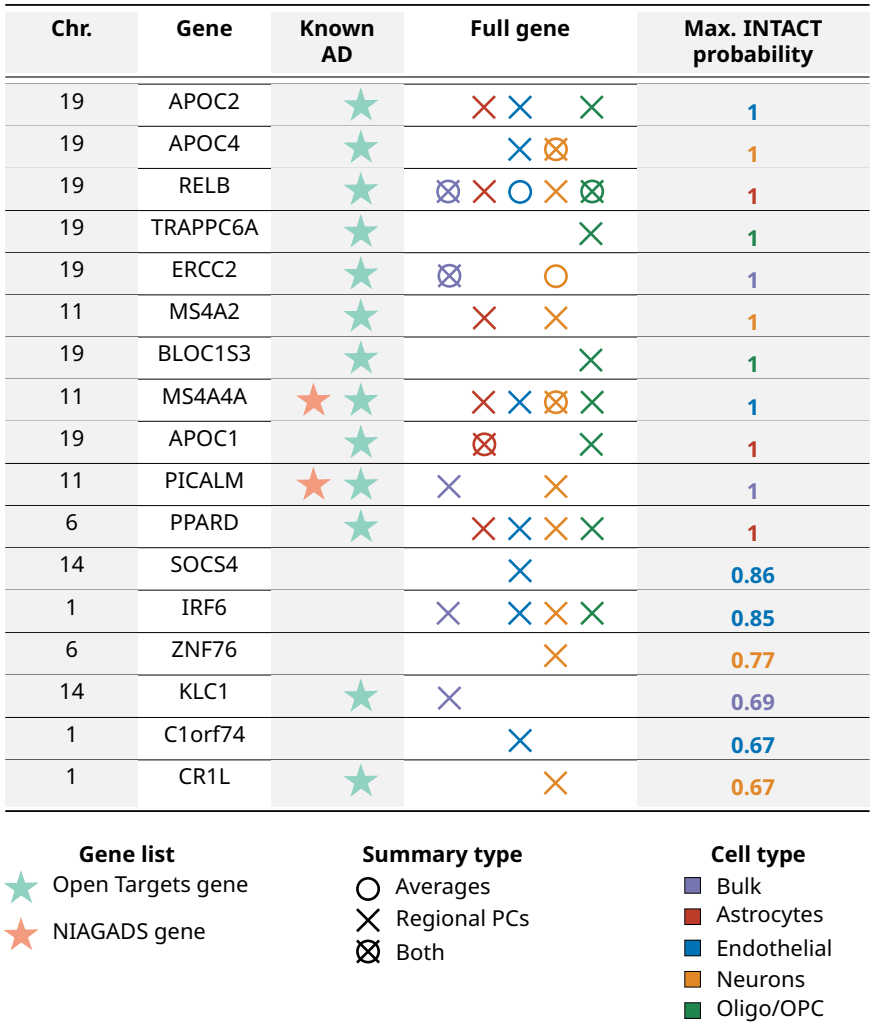

Fig. 6 | Regional PCs identify more candidate causal links between gene-level methylation and Alzheimer's disease risk across gene region types. Only genes with INTACT posterior probabilities >0.6 using rPCs are shown. Genes identified using rPCs (denoted as 'X') and those identified by averages (marked with 'O') are color-coded according to the cell type in which they were identified. Maximum INTACT probabilities for each gene are presented and color-coded to the respective cell type. The chromosome for each gene is in the 'Chr.' column. Source data are proved as a Source Data file. (AD Alzheimer's disease, Chr. Chromosome, Max maximum, NIAGADS National Institute on Aging Genetics of Alzheimer's Disease, Oligo/OPC oligodendrocyte/oligodendrocyte progenitor cells).

We applied rPCs to the cell type-specific methylation data from the ROSMAP cohort to demonstrate its effectiveness in identifying AD-relevant methylation changes. We observed a stronger correlation between the degree of DM detected with established gene relevance to AD pathology using rPCs compared to averages. The gene-level associations we discovered were predominantly cell type-specific, with astrocytes showing the most substantial methylation signal associated with CERAD score. Although some disease-relevant genes emerged from bulk methylation data, cell type-specific analyses yielded far richer sets of genes. We found relatively few associations of methylation changes with ADD diagnosis and *APOE* genotype compared to neuritic plaque burden and Braak stage. These results support previous findings[63] that endophenotypes derived from proteinopathy measures are more consistently associated with molecular traits than clinical diagnoses and individual genetic markers. Mapping meQTLs with rPCs further identified more genes with significant associations between DNA methylation and genetic variation than traditional averaging methods. Integration with AD GWAS through colocalization, IV analysis, and INTACT revealed seventeen unique genes potentially mediating AD via DNA methylation, 12 of which were exclusively identified by rPCs. The unique genes included *MS4A4A* and *PICALM*, both identified by NIAGADS as high-priority AD-related genes.

The DMR results from both the simulation and AD analyses demonstrated that rPCs consistently outperformed other methods in terms of precision and F1 score. While it is possible that parameter tuning or alternative simulation procedures could influence these outcomes, one of the key strengths of rPCs lies in their minimal need for parameter tuning. This simplicity not only contributes to robust performance across different datasets but also makes the method easier to use.

The rPCs framework was designed to be broadly applicable across methylation datasets with minimal adaptation. rPCs leverage the correlation structure of CpG methylation within well-defined gene boundaries. This local correlation of methylation sites can be influenced by a number of cis- factors including genetics, TF binding and chromatin state[89]. While the focus of our application of rPCs has been on brain regions and AD, our method does not make any assumptions about the biological context or study design of the data. The defined gene boundaries used in rPCs are consistent (or can be easily adapted) across human tissues, which indicates that this method could be applied to other contexts with minimal adaptation. However, we recognize that the success of rPCs in different settings will depend on specific tissue characteristics, such as the methylation landscape and the biological questions at hand. As with any analytical method, we

recommend that future researchers perform the necessary validation steps that ensure that rPCs are appropriate for their specific study design and data. We believe that, with these considerations in mind, rPCs can serve as a powerful tool for uncovering methylation patterns in various biological contexts.

While the rPCs method improves upon existing approaches for aggregating methylation by several metrics, interpreting the contribution of individual features remains challenging. Future extensions of this method could leverage approaches such as non-negative matrix factorization, factor analysis, or independent components analysis to generate more interpretable summaries. However, unlike PCA, they do not guarantee orthogonality between factors and there are no analytic approaches to choosing the number of components. We chose to use singular value decomposition (SVD) to estimate our eigenvectors, which can be viewed as a frequentist approach to PCA. This approach does not give us a full generative model that would allow us to quantify uncertainty about the latent space representation of regional methylation, but SVD is more computationally tractable for the larger number of regions we studied. Nevertheless, broadening our framework to include frequentist and probabilistic implementations of dimensionality reduction methods is a promising future direction for our method.

Our initial focus on methylation microarrays sets the stage for expanding rPCs to whole-genome methylation sequencing (WGMS) and other epigenomic datasets. The present scarcity of large-scale WGMS datasets limits comprehensive analyses, but as more extensive datasets emerge, we expect our methodology to significantly enhance our understanding of the role of methylation in complex diseases. Beyond WGMS, we foresee the application of rPCs to a diverse array of epigenomic data, including ATAC-seq, ChIP-seq, and Hi-C techniques. This expansion will enable a holistic view of the epigenetic landscape and foster the development of models capable of integrating insights across various epigenomic modalities.

The application of our *regionalpcs* methodology to brain data highlights a notable challenge: the absence of detectable microglial signals. This challenge is likely due to the relatively low abundance of microglia in the DLPFC. Despite this limitation, computational deconvolution has demonstrated promise in mapping expression QTLs in microglial populations[90], suggesting future research paths. Employing techniques such as nuclei sorting or intact cell isolation is essential for acquiring sufficient samples for methylation and other functional genomic studies, particularly for cell types like microglia. Subsequent validation of our findings using these refined cellular populations will be critical in corroborating the genes and pathways we have implicated in AD. Such validation is the first step in the development of novel therapeutic strategies and enhancing our capacity to predict AD risk, thereby advancing the field's move from broad observations to targeted interventions.

In conclusion, the rPCs method is a significant advancement in the analysis of methylation data, particularly for complex diseases like AD. Our findings underscore the value of rPCs for uncovering the complex mechanisms underlying disease, providing novel research directions that could lead to targeted therapeutic interventions and refined disease risk assessments. Further, the ability to use rPCs at a single gene locus may aid in increasing power to identify targeted methylation changes using dCas9. The versatility of rPCs extends its potential impact beyond AD to encompass a broad spectrum of conditions where methylation plays a pivotal role. With the *regionalpcs* Bioconductor package, we offer the scientific community an accessible tool to advance epigenetic research, fostering discoveries that could transform our approach to disease understanding and treatment.

Methods

Ethics & inclusion statement

This study complies with all relevant ethical regulations. Research involving human data was approved by the Institutional Review Board (IRB) at Rush University Medical Center. The ROSMAP study protocol was reviewed and approved by the IRB at Rush University. All participants provided informed consent for longitudinal clinical evaluation, brain tissue donation, and repository consent for data and biospecimen sharing. Data used in this research were accessed through Synapse AD Knowledge Portal under an approved data use agreement that ensures compliance with privacy and confidentiality requirements. This research did not involve any stigmatizing, incriminating, or discriminatory risk to participants.

Statistics & reproducibility

Sample sizes were determined by the availability of data within the ROSMAP cohort, a well-established AD resource. No data were excluded from the analyses except for those failing quality control checks. All statistical tests, including linear models for DM and QTL mapping, were performed in R (version 4.2.2) and Python (version 3.10.9), with a Benjamini–Hochberg correction applied where applicable. Other statistical tests include Student's T test, Fisher's exact, and the Kruskal–Wallis test. All tests were two-sided, and p values of 0.05 or lower were considered significant. Randomization and blinding were not applicable as the study involved observational data.

Regional principal components

Let $\mathbf{X}_r \in \mathbb{R}^{n \times m_r}$ be a matrix of n samples and m_r CpG sites for distinct regions $r \in \{1, 2, \ldots, R\}$, where each entry of the matrix corresponds to a methylation quantification value such as a beta value or an M value. Note that in contrast to Fig. 1, in this section we will work with $\mathbf{X}_r$ instead of $\mathbf{X}_r^T$. Without loss of generality, we assume that $\mathbf{X}_r$ has been transformed by the INT

$$\Phi^{-1}\left\{\frac{\mathrm{rank}(u_i) - k}{n - 2k + 1}\right\} \tag{1}$$

where Φ^{-1} is the probit function, and $k \in (0, 1/2)$ is an adjustable offset. The INT guarantees that each column is normally distributed when there are no ties in the data. To compute the rPCs of region r, we first perform SVD on $\mathbf{X}_r$ to decompose the matrix into,

$$\mathbf{X}_r = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^T \tag{2}$$

where $\mathbf{U}_r$ is an $n \times n$ orthogonal matrix whose columns are the left singular vectors of $\mathbf{X}_r$, $\boldsymbol{\Sigma}_r$ is an $n \times m_r$ diagonal matrix containing singular values of $\mathbf{X}_r$ in descending order, and $\mathbf{V}_r$ is an $m_r \times m_r$ orthogonal matrix whose columns are the right singular vectors of $\mathbf{X}_r$. The eigenvalues of the covariance matrix $\mathbf{X}_r^T\mathbf{X}_r$ are the squares of the singular values along the diagonal of $\boldsymbol{\Sigma}_r$.

We use the Gavish-Donoho method[46] to estimate the optimal number k^* of principal components needed to capture the essential variability in the data. The Gavish-Donoho method estimates the optimal number of principal components purely as a function of the ratio of number of columns to the number of rows of the matrix as well as estimated noise in the matrix. We select the first k^* right singular vectors (columns of $\mathbf{V}_r$) and denote the resulting matrix as

$$\mathbf{V}_r^* = \left[\mathbf{v}_r^1, \mathbf{v}_r^2, \cdots, \mathbf{v}_r^{k^*}\right] \tag{3}$$

where $\mathbf{v}_r^j$ is the jth column of $\mathbf{V}_r$. The columns of $\mathbf{V}_r^*$ are the first k^* principal components of $\mathbf{X}_r$. Finally, we project $\mathbf{X}_r$ onto the principal components $\mathbf{V}_r^*$ to compute the principal component scores,

$$\mathbf{Z}_r = \mathbf{X}_r \mathbf{V}_r^* \tag{4}$$

where $\mathbf{Z}_r$ is an $n \times k^*$ matrix, representing the transformed coordinates of the original data $\mathbf{X}_r$ in the reduced-dimensional space spanned by the first k^* principal components contained in $\mathbf{V}_r^*$. Each row

corresponds to a sample in X_r. We refer to each column of $\mathbf{Z}_r$ as the rPCs, which are used to summarize regional methylation for all downstream comparative analyses.

Simulation analysis

We employed the *RRBS-sim* tool, developed by ENCODE[48], to simulate RRBS data. This tool models CpG site locations using a hidden Markov model, sequencing coverage through a two-component gamma distribution mixture, and spatial correlations between sites via a Gaussian variogram.

Our simulations focused on methylation within gene regions. We explored four key parameters: the count of CpG sites, the sample size, the proportion of differentially methylated (DM) CpG sites, and the methylation percentage difference between case and control groups at DM sites. For CpG site counts, we referenced an ENCODE RRBS dataset from a K562 cell line, determining the typical number of CpGs within full gene (mean = 50, median = 29) and promoter regions (mean = 19, median = 12) using the TxDb.Hsapiens.UCSC.hg19.knownGene Bioconductor package[91]. Based on these findings, we conducted simulations with 20 and 50 CpG sites. Sample sizes of 50, 500, and 5000 were chosen to represent both small and large datasets. We tested DM site proportions at 0.25, 0.50, and 0.75, and DM percentage differences from 1 to 9%, in 1% increments. For differentially methylated regions, we set the length at half the number of CpG sites, ensuring a minimum of 2% CpG density.

In total, 162 unique parameter combinations were used to simulate methylation beta values for 1000 regions per set. Samples were equally divided into cases and controls. Methylation data for each CpG site underwent normalization using the *RankNorm* function from the RNOmni R package[92]. Methylation was summarized for each region using both averages and rPCs.

For DM analysis, we utilized the *lmFit* function from the limma package[74], fitting linear models to the summarized methylation data of each gene region. Stability in test statistics was achieved using the empirical Bayes approach via limma's *ebayes* function, and p values were Bonferroni-adjusted for multiple testing. Regions with adjusted p values below 0.05 were classified as significantly differentially methylated.

To assess the accuracy in identifying true negative results, we generated a matrix comprising regions with varied DM levels. This included 700 regions with no methylation difference (true negatives) and 300 regions with methylation differences varying from 1×10^{-3} to 9% at intervals of 1% (30 each). Methylation was summarized using both regional averages and rPCs, followed by DM analysis.

Simulation-based DMR analysis

We simulated RRBS methylation dataset to compare DMR methods with rPCs. The simulations were conducted for varying numbers of genes (N = 100, 1000, and 16,000) with fixed sample sizes of 50, 500, or 5000. In each simulation, 5% of the genes were designated as differentially methylated, resulting in 5, 50, or 800 DM genes, respectively.

Simulation parameters were randomly assigned to genes with replacement. For DM genes, the methylation difference between cases and controls ranged from 0.01 to 0.1 in increments of 0.01, while non-DM genes had no methylation differences. DMRs were defined with lengths of either 20 or 50 base pairs, and the proportion of DM sites within these regions varied, with values set at 0.25, 0.5, or 0.75. Genes were randomly assigned to chromosomes 1-22, with a random order of genes on each chromosome.

DM analyses were performed using both gene-level summaries and DMR analysis through methylKit and DMRcate. For methylKit, we simulated 30x coverage to mimic a high-quality dataset, as the tool requires coverage values at each site. The coverage value greatly affected the number of detected DMRs, with higher coverage (100×) producing many false positives. We assessed various window and step sizes in methylKit, including 50, 1000 (default), 2000, 4000, and 6000, and selected a window size of 2000, as it provided the best balance across performance metrics. While increasing the window size reduced false positives, it also decreased the number of true positives. To assess the robustness of the results, the simulation process was repeated 100 times, allowing for the calculation of error bars for the performance metrics.

Results were evaluated at a gene level, with any gene showing a significant DM hit being considered significant. This comparison was important for assessing performance on both individual CpGs and DMRs. We applied an FDR threshold of p value < 0.05, and no coverage minimum was used in methylKit to maintain consistency with other methods being compared.

Study cohort

To explore the relationship between DNA methylation and AD, we analyzed a subset of the ROSMAP cohort with DNA methylation and whole-genome sequencing (WGS) data. DNA methylation and WGS data were obtained from the Synapse AD Knowledge Portal following the submission of a Data Use Agreement (https://adknowledgeportal.org). We applied quality control filtering to the data (detailed below). Specifically, we executed a series of filtering steps on both the DNA methylation and WGS data, opting for individuals who possessed data for both and adhered to our filtering standards. This filtering procedure resulted in a final cohort of 563 individuals, all of whom self-identified as white. The average age of the cohort was 86.4 years, with females comprising 63% of the cohort (N = 356, Supplementary Table 1). Each individual in this cohort provided a single sample for methylation and genotype analysis, forming the basis for all subsequent analyses in this study.

Preprocessing genotype data

Subject genotypes were obtained from ROSMAP WGS VCF files (Synapse ID: syn11707420) generated by The Whole-Genome Sequence Harmonization Study, a multi-institutional effort to harmonize genotype data from several AD cohorts[93]. WGS libraries were constructed using the KAPA Hyper Library Preparation Kit and sequenced on an Illumina HiSeq X sequencer. Raw sequence reads were aligned to the GRCh37 human reference genome using the Burrows-Wheeler Aligner (BWA-MEM v0.7.08). The aligned data were processed following the GATK best practices workflow, which includes haplotype calling, joint genotyping with GATK v3.5, and variant filtering with Variant Quality Score Recalibration (VQSR).

Variant filtering. Variants were normalized with BCFtools[94] and lifted over from GRCh37 to GRCh38 using LiftoverVcf from GATK v4.0.10.0[95]. We used PLINK2[96] to remove multiallelic variants, variants with more than 5% missing genotype calls, variants with minor allele frequency less than 0.01, variants that failed the haplotype-based test for non-random missing genotype data with a p value threshold of 1×10^{-9}, variants that deviated from Hardy-Weinberg equilibrium at p value < 0.001, variants on the sex chromosomes, and variants that are not at CpG sites. Our filtering pipeline yielded 9,470,439 high-quality autosomal variants for subsequent analyses.

Sample filtering. Utilizing PLINK[97], we excluded samples exhibiting missing call rates exceeding 5%. We also used the *SmartPCA* algorithm from EIGENSTRAT[98] to detect population outliers using the default parameters of ten principal components and six standard deviations; none were discovered in our cohort.

Selection and interpretation of ROSMAP phenotypic traits

We chose four phenotypic traits from the ROSMAP cohort to test for associations with DNA methylation. These traits were a CERAD score-

based diagnosis of ADD, Braak staging for neurofibrillary degeneration, clinical diagnosis of ADD, and *APOE* genotype (Supplementary Table 1).

The CERAD score[53], a tool from the Consortium to Establish a Registry for AD, offers a semiquantitative assessment of neuritic plaque density in neocortical regions from autopsy brain samples. Based on the CERAD score, a score between one through four was assigned to each individual. This score, treated as a continuous value in our study, has integer values ranging from one (indicating definite AD) to four (indicating no AD). The distribution of subjects across CERAD score categories are Definite AD ($N = 170$), Probable AD ($N = 195$), Possible AD ($N = 59$), and No AD ($N = 139$) (Supplementary Table 1).

Braak staging[52] evaluates the anatomical extent of neurofibrillary degeneration, with integer scores from zero to six reflecting the stereotypical distribution of neurofibrillary tangles across limbic and cerebral cortical regions. Since our focus was on methylation data from the DLPFC, we discretized the Braak stage into a categorical variable indicating the presence (≥5, $N = 132$) or absence (<5, $N = 431$) of neurofibrillary tangles involving a neocortical region, which includes the DLPFC. Approximately 80% of the cohort had tangles extending into limbic regions (Braak stage III to VI), while only 24% had tangles in neocortical regions, which include the DLPFC (Braak stages V and VI).

To create a comprehensive diagnosis variable, we combined clinical and pathology scores. Our clinical measure was the final consensus cognitive diagnosis variable (cogdx), determined by a dementia-specialist neurologist based on all clinical data available, excluding post-mortem information. Our pathology scores were the previously described CERAD score and Braak stage. We labeled subjects as "AD" if they were diagnosed with cognitive impairment or dementia (cogdx 2–5), exhibited a high Braak stage (V or VI), implying the spread of neurofibrillary tangles to neocortical regions, and had a low CERAD score (≤2) indicating probable or definite AD based on neuritic plaques ($N = 217$). Subjects were designated as "Control" if they had a clinical diagnosis of NCI (cogdx = 1), a low Braak stage (≤3), and a high CERAD score (≥3) ($N = 80$). In analyses of our diagnosis variable, we omitted individuals who did not fit into either the "AD" or "Control" categories as we found this indicative of a discrepancy between the clinical and pathological variables ($N = 266$).

In our analyses of the *APOE* genotype, we focused on individuals with *ϵ3/ϵ3* ($N = 340$) and *ϵ3/ϵ4* ($N = 129$) genotypes so that we could focus on exploring the impact of a single *ϵ4* allele, which is associated with an increased risk of AD. The *ϵ3* allele is the most common and is used as the reference in estimating relative risk, while the rare *ϵ2* allele is reported to be protective against AD[86,99]. We excluded individuals carrying the *ϵ2* allele from our *APOE* analyses because we did not have a sufficient sample to robustly analyze these subjects as a distinct group.

Preprocessing methylation data

IDAT files of DNA methylation microarray from ROSMAP (Synapse ID: syn7357283) were generated as described in De Jager et al.[7]. DNA was extracted from frozen DLPFC sections of deceased ROSMAP participants. The Illumina Infinium HumanMethylation450 bead chip assay was used to quantify DNA methylation.

We used the minfi package[100] for normal-exponential convolution using out-of-band probes dye bias correction, mapping of probes to the genome, and removal of probes in sex chromosomes or probes overlapping any genetic variant in Illumina and dbSNP references[101]. The ewastools package[102] was used to compute detection p values, and we used the wateRmelon package[103] to remove probes with detection p values > 0.01 or bead count <3. We used the Beta Mixture Quantile dilation (BMIQ) normalization method[104] to correct for biases resulting from differences between type 1 and type 2 probes in the Infinium HumanMethylation450k platform. Lastly, we removed probes overlapping any genetic variant called in any subject from the WGS genotype data from the ROSMAP cohort. After all filtering steps, 365,899 probes were used for downstream analysis.

Samples were excluded if their bisulfite conversion efficiency was below 80% or if they failed the ewastools quality control metrics estimated following Illumina's BeadArray Controls Reporter Software Guide. In total, 87 samples were removed because they did not pass these quality control measures.

Following these initial normalization and quality control steps, the methylation was further processed using the following steps: cell type deconvolution to estimate cell type-specific methylation profiles, normalization and cleaning applied within each cell type-specific dataset, and then summarization of regional methylation using *regionalpcs* and averages. Effects of confounding variables were addressed during cell type deconvolution, data cleaning, and in the DM and mQTL models. Extensive details of these steps can be found in the methods sections below.

Cell type deconvolution

We used cell type deconvolution to estimate cell type-specific methylation from the processed and filtered bulk tissue methylation data. We estimated cell type-specific methylation from 365,899 probes and 563 samples for which matched whole-genome sequencing data was available. Our deconvolution pipeline has two steps: estimation of cell type proportions and imputation of cell type-specific methylation profiles. We used the EpiSCORE R package[60] to estimate the cell type proportions for six major brain cell types: astrocytes, endothelial cells, neurons, oligodendrocytes, oligodendrocyte progenitor cells, and microglia, using the brain reference panel provided by the EpiSCORE R package. As detailed in ref. 60, this reference imputed cell type-specific methylation at specific CpGs from a single-cell RNA-seq dataset.

We estimated cell-type proportions for each sample using the *wRPC* function from EpiSCORE. Our estimated cell type proportion for microglia across all samples was zero, likely due to their low proportions in the DLPFC, so we had to exclude microglia from further analysis. Additionally, the proportions of oligodendrocytes were extremely low (mean = 0.003%, median = 0%), with only four individuals having a non-zero estimated proportion. This is likely because of the high similarity to oligodendrocyte progenitor cells, for which we estimated proportions consistent with the expected abundance of oligodendrocytes. Under the assumption that it was difficult for EpiSCORE to distinguish between these similar cell types, we combined oligodendrocytes and oligodendrocyte progenitors into a single proportion. This resulted in cell type proportion estimates for four cell types: astrocytes, endothelial cells, neurons, and oligodendrocyte/oligodendrocyte progenitor cells.

To validate the estimated neuron proportions, we used an independent brain reference panel and estimation method published by Houseman et al.[64] for estimating cell type proportions from the *estimateCellCounts* function in the minfi package. This reference panel only distinguishes neurons from non-neurons but was generated from methylation microarray data from sorted nuclei from brain samples. We found that the neuron proportions estimated by the EpiSCORE and Houseman methods were highly correlated (Pearson $r = 0.8$, p value < 0.001). We used the cell type proportion estimated by EpiSCORE for downstream analysis because these proportions were more granular for non-neuronal cell types.

We used the TCA[61] package for the imputation of cell type-specific methylation profiles. TCA uses a multivariate regression model that decomposes a matrix of bulk tissue methylation beta values into a tensor of cell type-specific beta values given a vector of cell type proportions for each sample. Our inputs to TCA were the filtered bulk tissue methylation data and the cell type proportions estimated by EpiSCORE.

TCA accommodates the inclusion of several types of covariates in the deconvolution model. TCA accounts for variables that may affect

methylation at the cell-type level (C1) and global level (C2). We evaluated models with age and sex as C1 covariates, batch, post-mortem interval (PMI), and study as C2 covariates, or a combination of C1 and C2 covariates. The model containing only C2 covariates emerged as the best fit since the C1 covariates seemed to exaggerate the effects of age and sex in downstream DM analysis. We used other options for TCA to constrain the model's mean parameter during optimization and refit the estimated cell-type proportions iteratively. The resulting refitted cell type proportions were used as covariates to model cell type proportions in downstream analyses. The final output of TCA was imputed CpG-level methylation beta values in all subjects for the four cell types whose proportions we estimated with EpiSCORE.

Normalization and cleaning of deconvolved methylation data

We removed probes with zero variance for each cell type-specific dataset and the original bulk tissue dataset. We applied the rank-based INT using the *RankNorm* function from the RNOmni package[92]. Transforming methylation beta values to M values or inverse normal transformed values is a standard approach to reduce heteroscedasticity and give the methylation values an approximately Gaussian distribution for downstream statistical analyses[51]. We compared the non-parametric INT to the logit transformation for M values and found comparable results, consistent with previous reports comparing these approaches[105]. We chose to use INT because it does not produce large, transformed values for beta values close to 0 or 1, as seen with M values.

We identified technical variables strongly correlated with our methylation data by estimating the top methylome-wide principal components (PCs) from the full dataset. The number of PCs was chosen using the Gavish-Donoho method[46], described in more detail in the "*Summarizing regional DNA methylation*" section. We correlated traits with PCs using Spearman's rank correlation for numeric variables and the Kruskal–Wallis test for categorical variables. We used a Bonferroni-adjusted p value < 0.05 as our significance cutoff. We found that batch, PMI, study (ROS or MAP), sex, age at death, and cell type proportions were highly correlated with one or more PCs in datasets. We used the *removeBatchEffect* function from the limma package[74] to remove the effects of the top four methylome-wide PCs and all of the previously described covariates except for sex, which we instead included as a covariate in downstream analyses. After removing these covariates, some correlations persisted between the technical variables and methylome-wide PCs, and we incorporated them as covariates in all downstream analyses.

Visualizing deconvolved cell type methylation profiles

The cell type-specific and bulk tissue methylation profiles were visualized with the UMAP method for dimensionality reduction[65] as implemented in the uwot package[106]. We transformed the deconvolved methylation data into M values and used principal components analysis on the union of all cell type-specific datasets and the bulk tissue dataset using the default parameters of the *pca* function from the PCAtools package[107]. To determine the number of principal components to retain, we used the Gavish-Donoho method[46], described in the "*Summarizing regional DNA methylation*" section. The UMAP projection was estimated with the *umap* function from the uwot package, setting the number of neighbors to 500, the spread to 10, and the minimum distance to 5.

Annotating gene region types

Our investigation focused on DNA methylation patterns across four gene regions: promoters (spanning from 1 kb upstream of the TSS to the TSS itself), gene bodies (ranging from the TSS to the downstream end of the 3′ untranslated region (UTR)), 5 kb upstream region (extending from 5 kb upstream of the TSS to the TSS), and the full gene (a union of the 5 kb upstream and gene body regions). We delineated these regions using data from the Annotatr package[108], Gencode[109], and the UCSC Table Browser[110], all of which used the GRCh38 reference genome build.

We used the Gencode v32 gene annotations for *Homo sapiens* to select genes and transcript isoforms. We filtered this annotation for protein-coding transcripts with a transcript support level less than four. For genes whose canonical isoform passed these filters, we used this isoform, and for genes with no canonical isoform after filtering, we used the longest available isoform instead. We chose the four gene region types described above because these regions are generally consistent across transcript isoforms, in contrast to more variable features such as exons and introns.

Summarizing regional DNA methylation

We designed an approach to summarize gene-level DNA methylation within the four gene regions with two different methods: averaging and *regionalpcs* (rPCs). We extracted the genomic positions of each probe from the GRCh37 annotations of the HumanMethylation450K array provided by RnBeads package[34]. We remapped the genomic positions to GRCh38 using the GenomicRanges and liftOver packages[111,112], and CpGs were assigned to region types using the *findOverlaps* function from the GenomicRanges package.

We summarized methylation with averaging by calculating the mean across all CpGs falling within a region type. We summarized methylation with rPCs by estimating the principal components of CpGs within a region using the *pca* function from the PCAtools package. We used the Gavish-Donoho method implemented by the *chooseGavishDonoho* function in PCAtools to select the optimal number of principal components to represent methylation in a gene region. The Gavish-Donoho method[46] estimates the optimal number of principal components purely as a function of the ratio of number of columns to the number of rows of the matrix as well as estimated noise in the matrix. Because our matrices are always scaled and centered, the noise parameter always had a value very close to 1. We assessed both the Gavish-Donoho and Marchenko-Pasteur methods[47] for selecting the number of principal components and chose the Gavish-Donoho method because it was more conservative.

We applied the averaging and rPC methods for each gene and region type which included at least one CpG. In cases where a region type for a gene contained only one CpG, we used the original CpG methylation value to represent that gene in summarized datasets for averages and rPCs. In total, we summarized 20 datasets using averages and rPCs, encompassing all four region types and four cell types as well as bulk tissue methylation.

Assessment of rPCs for potential global methylation factors

We performed a correlation analysis between rPCs across multiple genes to test for the possibility that the rPCs may capture global methylation effects rather than regional variation. We estimated the correlation between the first three rPCs (ordered by the proportion of variance explained) for each gene with those of genes located on different chromosomes. Our underlying hypothesis posited that if rPC1 primarily captured global methylation effects, it would exhibit higher inter-gene correlations compared to other rPCs.

We estimated partial correlations that adjusted for several confounding variables: sex, age, PMI, study, batch, and cell type proportions. This analysis revealed marked disparities in the distribution of correlations, with rPC1 displaying the highest inter-gene correlations (Supplementary Fig. 7A).

We next tested if incorporating the top global methylation PCs into our partial correlation model would reduce the inter-gene correlations we observed. The number of global PCs was determined using the Gavish-Donoho method. The inclusion of these global PCs attenuated the previously observed correlation discrepancies among the

rPCs, demonstrating that these rPCs were capturing global signals that were also reflected in the top global PCs (Supplementary Fig. 7B).

We sequentially excluded each global PC from the model, re-estimated partial correlations for all other variables, and determined that the first four global PCs drove the shared signal. When any of these global PCs were omitted, inter-gene correlation distributions were discordant between rPCs (Supplementary Fig. 7C).

We refined our methylation summarization model by removing the first four global with the *removeBatchEffect* function from the limma package prior to estimating the final set of rPCs. The adjustment was validated with a partial correlation analysis, which found negligible disparities in inter-gene correlations among the newly derived rPCs (Supplementary Fig. 7D). These results suggest that the rPCs estimated after removing the top global PCs were no longer influenced by global methylation patterns and instead captured regional methylation signals.

DM analysis

We performed a comprehensive DM analysis of the processed methylation data for each region type and cell type, using averages and rPCs to summarize methylation. Our disease phenotypes described in the "*Selection and interpretation of ROSMAP phenotypic traits*" section were CERAD score (a measure of neuritic plaques), Braak stage (a measure of Tau protein), AD diagnosis, and *APOE* genotype.

In addition to the previously described disease phenotypes, our model included age of death, PMI, study (ROS or MAP), batch, sex, and cell type proportions we estimated for all four cell types. We also included the top global methylation principal components (PCs) computed from the full CpG-level methylation data. These global PCs were calculated separately for each of the four cell type datasets across all CpGs and samples, and the number of PCs was determined using the Gavish-Donoho method described in the "*Summarizing Gene-level DNA Methylation*" section. To prevent the removal of disease-associated signals from the methylation data, we omitted global PCs if they were significantly correlated with the outcome trait and not with batch, PMI, and study (Bonferroni-adjusted p value < 0.1). We used Spearman rank correlation for numeric variables and the Kruskal-Wallis test for categorical variables.

We identified differentially methylated genes with the *lmFit* function from the limma package by fitting linear models to the methylation data for each summarized gene region. The methylation was normalized using rank-based INT from the *RankNorm* function in the RNOmni package. We used the empirical Bayes method implemented by the *ebayes* function in the limma package to obtain more stable estimates of the test statistics, and we used the Benjamini-Hochberg method for multiple test adjustments.

DMR analysis on the ROSMAP data was performed using the R package DMRcate. CpG-level methylation data were preprocessed the same as CpG-level analyses, including the use of INT. The same covariates were included as in the DM analysis for CpGs, rPCs, and averages, including the global methylation PCs. DMR analysis was restricted to astrocytes with the CERAD score outcome, as this analysis showed the most consistent signal in the summary-type DM analyses. DMRcate was run using default parameter settings. Significant DMRs were overlapped with genes based on full gene region annotations used for regional methylation summaries. A gene was considered a significant hit if any DMR overlapped with it.

Identifying high-confidence AD genes

We used the Open Targets database[72] to identify a high-confidence set of genes associated with AD. This comprehensive resource compiles information on the relationship between genes and diseases from a range of sources. We used a curated list of 6,595 genes linked to AD based on seven distinct evidence categories: genetic associations, somatic mutations, drugs, pathways systems biology, text mining, RNA expression, and animal models. Open Targets assigns an overall association score to each gene on the list according to the available evidence.

Mapping cis-QTLs

We used tensorQTL[113] to map cis-methylation QTLs (meQTLs) for each gene in every region type and cell type. tensorQTL is a reimplementation of FastQTL[114] that uses linear models to identify correlations between alternate allele count and molecular traits such as methylation. We considered common germline variants with a threshold of 1% minor allele frequency (MAF) and restricted analysis to biallelic single nucleotide polymorphisms. For every gene, we tested for associations between alternate allele count and gene-level methylation summaries within a cis-window of 1 million base pairs upstream and downstream of the TSS for both averages and rPCs.

We incorporated several covariates into the QTL mapping model to control for potential confounding factors, including age of death, PMI, batch, study (ROS or MAP), sex, and cell type proportions for our four cell types of interest. We also included the top 10 genotype principal components (PCs) and the top methylation PCs, determined using the Gavish-Donoho method[46] as previously described. We used the *get_significant_pairs* function from tensorQTL with default parameters to extract significant meQTLs with a false discovery rate (FDR) threshold of 0.05 that accounted for the number of genes and variants tested. We used the beta-approximated p value distribution estimated by tensorQTL as a faster approximation of permutation testing.

Fine-mapping QTLs

We used fine-mapping to identify credible sets of causal variants in our meQTLs for gene region-level methylation summaries estimated with averages and rPCs. First, we used TORUS[115] to estimate priors for each variant based on their distance to theTSS. We applied the default TORUS settings for binning and enrichment score calculations without additional parameter tuning. Next, we used DAP-G[116] to identify credible sets from the priors generated by TORUS and the individual-level genotype and methylation data we also used for meQTL mapping. DAP-G uses a Bayesian framework to estimate the posterior probabilities of each genetic variant associated with a molecular trait such as methylation. We selected an R^2 cutoff of 0.25 as the minimal LD correlation needed for two variants to be included within the same credible set. This threshold is commonly used in fine-mapping studies and serves as the default value in popular tools like Susie, FINEMAP, and DAP-G[117,118]. It balances the inclusion of meaningful correlations with reducing noise from low LD. Credible sets represent a group of genetic variants that contain a causal association with a trait independent of any other credible set, and a single gene can contain multiple credible sets. Because the variants within the credible set are in LD with each other, the exact causal variant cannot be determined solely from QTL mapping unless the credible set contains a single variant. Our threshold for defining a strong credible set in downstream analyses was a posterior inclusion probability (PIP) of 0.5 or greater.

Fine-mapping GWAS

To integrate our meQTLs with genome-wide association studies (GWAS) of AD risk, we first needed to fine-map the GWAS summary statistics. We used DAP-G[116] for this task similar to how we used it for fine-mapping the QTL data.

We used a published meta-analysis of AD risk GWAS[75]. We converted the genomic positions in the GWAS summary statistics from GRCh37 to GRCh38 using the GATK v4.0.10.0 LiftoverVcf tool to match genomic coordinates in our meQTL analysis. We matched 8,114,981 variants in the GWAS summary statistics to the variants tested in our meQTL analysis and removed 6912 variants with inconsistent reference/alternate alleles.

Since we only had summary statistics rather than individual-level data, we used linkage disequilibrium (LD) blocks for fine-mapping. We

used a published set of 1361 LD block annotations in GRCh38 coordinates[119] and intersected them with the GWAS variants in the summary statistics using the GenomicRanges package.

We used PLINK 1.90b6.21 to create LD matrices within each LD block and ran DAP-G with default parameters. Similar to our QTL fine mapping, we identified credible sets across several LD blocks representing independent AD risk signals.

GWAS integration

We integrated our meQTLs with an AD GWAS[75] to identify putative causal associations between gene-level DNA methylation and AD risk. In this framework, we can view the perturbation of DNA methylation by germline genetic variants as quantified by meQTLs as a "natural experiment" that tests how changes in DNA methylation affect the risk of developing AD. Our GWAS integration pipeline is comprised of three parts: colocalization with fastENLOC[76], IV analysis with PTWAS[77], and the probabilistic integration of these two analyses with INTACT[78].

Colocalization. We used colocalization with fastENLOC[76] to identify genes with evidence of shared causal variants between meQTLs and AD risk. fastENLOC uses a Bayesian framework to estimate the posterior probability of colocalization between a QTL and a GWAS trait while also estimating the overall GWAS heritability enrichment of QTL variants across all genes tested for QTLs.

The inputs for this analysis were the fine-mapped QTL and AD GWAS summary statistics generated in the previous steps. We limited our colocalization analysis to fine-mapped SNPs assigned to credible sets in the meQTL or GWAS summary statistics with PIPs $< 1 \times 10^{-4}$. We used the default parameters for fastENLOC and set the "total variants" flag to 12,688,339 variants.

We used a GLCP threshold of 0.50 to decide which genes were colocalized with AD risk. The GLCP is a locus-level measure of colocalization which addresses some of the limitations found in variant-level colocalization analysis, such as the potential for excessive type 2 errors. It quantifies the probability that a particular genomic locus harbors both a molecular QTL with a causal effect and a causal GWAS variant. This approach has more power to detect colocalization between methylation QTLs and AD, improving our understanding of the molecular basis of AD.

IV analysis with PTWAS. We used PTWAS[77] to explore potential causal relationships between DNA methylation and AD risk. The premise of IV analysis is that genetic variants identified as risk variants for AD in a GWAS mediate their effects through changes in DNA methylation. PTWAS constructs a composite IV by integrating weights from multiple independent methylation QTLs for each gene. The weights are derived from their respective posterior effect sizes from our fine-mapping meQTL analysis. A causal association Z score is computed as a weighted sum of GWAS trait Z scores for each independent instrument in the tested gene.

We followed the published PTWAS pipeline which first calculates weights for each gene with the *ptwas_builder* function from DAP-G using the individual-level data previously used to map meQTLs and results generated via DAP-G. Effect sizes and standard errors for the Wightman et al. GWAS were computed based on the reported z scores for individual genetic variants. We used GAMBIT[120] to perform the PTWAS scan step with the 1000 Genomes project, phase 3 version 5 LD reference panel for Europeans[121], chosen for its phase and ancestry match to the ROSMAP cohort. The LD panel was converted from GRCh37 to GRCh38 genome build using the liftOver tool, and any mismatches between reference and alternative alleles were rectified using PLINK v1.90b6.21. The significance of putative causal genes was defined by the z scores generated in the PTWAS scan. We set a genome-wide significance threshold considering genes with an absolute z score greater than 5.45 (corresponding to a p value < 5e−8) as having a likely causal association with AD.

Integration of colocalization and TWAS with INTACT. We used the Integration of TWAS and Colocalization (INTACT) approach to integrate the results from our colocalization and PTWAS analyses[78]. Colocalization is designed to identify QTLs that share causal variants with a GWAS trait. In contrast, TWAS is designed to uncover molecular features such as methylation mediating causal variants' effects on AD risk. INTACT is designed to reconcile the discrepancies observed when colocalization and TWAS are used separately by multiplying the posterior probabilities from colocalization and gene-trait associations Z scores from TWAS. This integration provides a single metric to quantify the probability of a causal association between genetic variations, methylation changes and AD traits. As inputs to INTACT, we used the GLCP estimated by fastENLOC and the gene-level Z score statistics estimated by PTWAS. We used the default settings for INTACT with a linear prior function with a truncation threshold of 0.05 and we filtered the posterior probabilities at a threshold of 0.6. This filtered set of genes was used as our final list with strong evidence of a causal relationship between DNA methylation and AD risk.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The original whole-genome sequencing data from ROSMAP (Synapse ID: syn11707420) and the DNA methylation microarray IDAT files (Synapse ID: syn7357283) are accessible through the Synapse AD Knowledge Portal (https://adknowledgeportal.org) under controlled access. Access requires a valid AD Knowledge Portal Data Use Certificate to protect the privacy and confidentiality of study participants. For more information on data access, please refer to the AD Knowledge Portal's (https://adknowledgeportal.synapse.org/Data%20Access). The AD Knowledge Portal provides a platform for accessing data, analyses, and tools generated by the Accelerating Medicines Partnership (AMP-AD) Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-science practices and accelerate translational learning. The data, analyses, and tools are shared early in the research cycle without a publication embargo on secondary use. Data is available for general research use according to the following requirements for data access and data attribution (https://adknowledgeportal.synapse.org/Data%20Access). For access to content described in this manuscript see: https://doi.org/10.7303/syn3219045. To increase accessibility, we have deposited summarized results that do not contain individual-level data on Zenodo, where they are available without a data use agreement. These summarized datasets include DM results[122], DAP-G fine-mapped QTL results[123], and variant-level colocalization and PTWAS results[124]. QTL results are provided by cell type for regionalpcs and averages[125–129] and CpGs[130–134]. Gene-level scores from GWAS integration steps (colocalization, PTWAS, and INTACT) can be found in Supplementary Data 3. Source data are provided with this paper.

Code availability

The *regionalpcs* R package, is available through Bioconductor at https://bioconductor.org/packages/release/bioc/html/regionalpcs.html[41] or on GitHub (https://github.com/tyeulalio/regionalpcs). The code used to generate the analyses in this paper are available online at Zenodo[135] (v.1.0.0, https://doi.org/10.5281/zenodo.14004154).

References

1. Cedernaes, J. et al. Acute sleep loss results in tissue-specific alterations in genome-wide DNA methylation state and metabolic fuel utilization in humans. *Sci. Adv.* **4**, eaar8590 (2018).
2. Lahtinen, A. et al. A distinctive DNA methylation pattern in insufficient sleep. *Sci. Rep.* **9**, 1193 (2019).

3. Lee, K. & Pausova, Z. Cigarette smoking and DNA methylation. *Front. Genet.* **4**, 55185 (2013).
4. Zeilinger, S. et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *Plos One* **8**, e63812 (2013).
5. Gaine, M. E., Chatterjee, S. & Abel, T. Sleep deprivation and the epigenome. *Front. Neural Circuits* **12**, 14 (2018).
6. Coppieters, N. et al. Global changes in DNA methylation and hydroxymethylation in Alzheimer's disease human brain. *Neurobiol. Aging* **35**, 1334–1344 (2014).
7. De Jager, P. L. et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat. Neurosci.* **17**, 1156–1163 (2014).
8. Mastroeni, D. et al. Epigenetic changes in Alzheimer's disease: decrements in DNA methylation. *Neurobiol. Aging* **31**, 2025–2037 (2010).
9. Irier, H. A. & Jin, P. Dynamics of DNA methylation in aging and Alzheimer's disease. *DNA Cell Biol.* **31**, S–42 (2012).
10. Saba, H. I. & Wijermans, P. W. Decitabine in myelodysplastic syndromes. *Semin Hematol.* **42**, S23–S31 (2005).
11. Oki, Y., Aoki, E. & Issa, J.-P. J. Decitabine–bedside to bench. *Crit. Rev. Oncol. Hematol.* **61**, 140–152 (2007).
12. Keating, G. M. Azacitidine: a review of its use in higher-risk myelodysplastic syndromes/acute myeloid leukaemia. *Drugs* **69**, 2501–2518 (2009).
13. Leone, G., Voso, M. T., Teofili, L. & Lübbert, M. Inhibitors of DNA methylation in the treatment of hematological malignancies and MDS. *Clin. Immunol.* **109**, 89–102 (2003).
14. Baron, U. et al. DNA methylation analysis as a tool for cell typing. *Epigenetics* **1**, 55–60 (2006).
15. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **19**, 129–147 (2018).
16. Kass, S. U., Landsberger, N. & Wolffe, A. P. DNA methylation directs a time-dependent repression of transcription initiation. *Curr. Biol.* **7**, 157–165 (1997).
17. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacol* **38**, 23–38 (2013).
18. Loyfer, N. et al. A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).
19. Nikolac Perkovic, M. et al. Epigenetics of Alzheimer's disease. *Biomolecules* **11**, 195 (2021).
20. Illumina. Infinium MethylationEPIC v2.0 BeadChip data sheet. (2022).
21. Meissner, A. et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
22. Adusumalli, S., Mohd Omar, M. F., Soong, R. & Benoukraf, T. Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief. Bioinformatics* **16**, 369–379 (2015).
23. Altuna, M. et al. DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis. *Clin. Epigenetics* **11**, 91 (2019).
24. Li, P. et al. Epigenetic dysregulation of enhancers in neurons is associated with Alzheimer's disease pathology and cognitive symptoms. *Nat. Commun.* **10**, 2246 (2019).
25. Zhang, L. et al. Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex implicates the immune processes in Alzheimer's disease. *Nat. Commun.* **11**, 6114 (2020).
26. Mendizabal, I. et al. Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biol.* **20**, 135 (2019).
27. Gasparoni, G. et al. DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. *Epigenetics Chromatin* **11**, 41 (2018).
28. Wilkinson, G. S. et al. DNA methylation predicts age and provides insight into exceptional longevity of bats. *Nat. Commun.* **12**, 1615 (2021).
29. Konki, M. et al. Peripheral blood DNA methylation differences in twin pairs discordant for Alzheimer's disease. *Clin. Epigenetics* **11**, 130 (2019).
30. Peters, T. J. et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* **8**, 6 (2015).
31. Jaffe, A. E. et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J. Epidemiol.* **41**, 200–209 (2012).
32. Lun, A. T. L. & Smyth, G. K. De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res.* **42**, e95 (2014).
33. Cai, J. et al. A comprehensive comparison of residue-level methylation levels with the regression-based gene-level methylation estimations by ReGear. *Brief. Bioinformatics* **22**, bbaa253 (2021).
34. Müller, F. et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* **20**, 55 (2019).
35. Wang, T. et al. A systematic study of normalization methods for Infinium 450 K methylation data using whole-genome bisulfite sequencing data. *Epigenetics* **10**, 662–669 (2015).
36. Gull, N. et al. DNA methylation and transcriptomic features are preserved throughout disease recurrence and chemoresistance in high grade serous ovarian cancers. *J. Exp. Clin. Cancer Res.* **41**, 232 (2022).
37. Bhalla, S., Kaur, H., Dhall, A. & Raghava, G. P. S. Prediction and analysis of skin cancer progression using genomics profiles of patients. *Sci. Rep.* **9**, 15790 (2019).
38. Schultz, M. D., Schmitz, R. J. & Ecker, J. R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28**, 583–585 (2012).
39. Kapourani, C.-A. & Sanguinetti, G. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics* **32**, i405–i412 (2016).
40. Zheng, Y., Jun, J., Brennan, K. & Gevaert, O. EpiMix is an integrative tool for epigenomic subtyping using DNA methylation. *Cell Rep. Methods* **3**, 100515 (2023).
41. Eulalio, T. regionalpcs: summarizing regional methylation with regional principal components analysis. *Bioconductor version: Release (3.18)* https://doi.org/10.18129/B9.bioc.regionalpcs (2023).
42. Qazi, T. J., Quan, Z., Mir, A. & Qing, H. Epigenetics in Alzheimer's disease: perspective of DNA methylation. *Mol. Neurobiol.* **55**, 1026–1044 (2018).
43. Scheltens, P. et al. Alzheimer's disease. *Lancet* **397**, 1577–1590 (2021).
44. DeTure, M. A. & Dickson, D. W. The neuropathological diagnosis of Alzheimer's disease. *Mol. Neurodegener.* **14**, 32 (2019).
45. Smith, R. G., Pishva, E. & Shireby, G. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. *Nat. Commun.* **12**, 3517 (2021).
46. Gavish, M. & Donoho, D. L. The Optimal Hard Threshold for Singular Values is 4/sqrt(3), *IEEE Transactions on Information Theory*, **60**, 5040–5053 (2014).
47. Marcenko, V. A. & Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Math. USSR Sb.* **1**, 457 (1967).
48. Lacey, M. R., Baribault, C. & Ehrlich, M. Modeling, simulation and analysis of methylation profiles from reduced representation bisulfite sequencing experiments. *Stat. Appl. Genet. Mol. Biol.* **12**, 723–742 (2013).

49. Akalin, A. et al. E. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
50. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
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. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).
53. Mirra, S. S. et al. The consortium to establish a registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41**, 479–486 (1991).
54. Oliva, M. et al. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* **55**, 112–122 (2023).
55. Bild, D. E. et al. Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156**, 871–881 (2002).
56. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
57. Rahmani, E. et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat. Commun.* **10**, 3417 (2019).
58. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
59. 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).
60. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol.* **21**, 221 (2020).
61. Jew, B. et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).
62. Patrick, E. et al. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput. Biol.* **16**, e1008120 (2020).
63. Lang, A.-L. et al. Methylation differences in Alzheimer's disease neuropathologic change in the aged human brain. *Acta Neuropathol. Commun.* **10**, 174 (2022).
64. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
65. McInnes, L. et al. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, **3**, 861 (2018).
66. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
67. Orre, M. et al. Isolation of glia from Alzheimer's mice reveals inflammation and dysfunction. *Neurobiol. Aging* **35**, 2746–2760 (2014).
68. Zhao, J., O'Connor, T. & Vassar, R. The contribution of activated astrocytes to Aβ production: implications for Alzheimer's disease pathogenesis. *J. Neuroinflamm.* **8**, 150 (2011).
69. Jo, S. et al. GABA from reactive astrocytes impairs memory in mouse models of Alzheimer's disease. *Nat. Med.* **20**, 886–896 (2014).
70. Ceyzériat, K. et al. Modulation of astrocyte reactivity improves functional deficits in mouse models of Alzheimer's disease. *Acta Neuropathol. Commun.* **6**, 104 (2018).
71. Monterey, M. D., Wei, H., Wu, X. & Wu, J. Q. The many faces of astrocytes in Alzheimer's disease. *Front Neurol.* **12**, 619626 (2021).
72. Koscielny, G. et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
73. Tábuas-Pereira, M., Santana, I., Guerreiro, R. & Brás, J. Alzheimer's disease genetics: review of novel loci associated with disease. *Curr. Genet. Med. Rep.* **8**, 1–16 (2020).
74. Ritchie, M. E. et al. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
75. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
76. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLOS Genet.* **13**, e1006646 (2017).
77. Zhang, Y., Quick, C., Yu, K. & Barbeira, A. GTEx Consortium, Luca, F., Pique-Regi, R., Kyung Im, H. & Wen, X. PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. *Genome Biol.* **21**, 232 (2020).
78. Okamoto, J. et al. Probabilistic integration of transcriptome-wide association studies and colocalization analysis identifies key molecular pathways of complex traits. *Am. J. Hum. Genet.* **110**, 44–57 (2023).
79. Lacher, S. E. et al. A hypermorphic antioxidant response element is associated with increased MS4A6A expression and Alzheimer's disease. *Redox Biol.* **14**, 686–693 (2017).
80. 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).
81. Hukku, A., Sampson, M. G., Luca, F., Pique-Regi, R. & Wen, X. Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility. *Am. J. Hum. Genet.* **109**, 825–837 (2022).
82. Zhao, S. et al. Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. *Nat Genet*, **56**, 336–347 (2024).
83. Patel, D. et al. Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl. Psychiatry* **11**, 250 (2021).
84. Marioni, R. E. et al. GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 1–7 (2018).
85. Tulloch, J. et al. Glia-specific APOE epigenetic changes in the Alzheimer's disease brain. *Brain Res.* **1698**, 179–186 (2018).
86. Strittmatter, W. J. & Roses, A. D. Apolipoprotein E and Alzheimer's disease. *Annu. Rev. Neurosci.* **19**, 53–77 (1996).
87. Chiba-Falek, O., Gottschalk, W. K. & Lutz, M. W. The effects of the TOMM40 poly-T alleles on Alzheimer's disease phenotypes. *Alzheimer's Dement.* **14**, 692–698 (2018).
88. Shao, Y. et al. DNA methylation of TOMM40-APOE-APOC2 in Alzheimer's disease. *J. Hum. Genet* **63**, 459–471 (2018).
89. Watkins, S. H. et al. DNA co-methylation has a stable structure and is related to specific aspects of genome regulation. Preprint at https://doi.org/10.1101/2022.03.16.484648 (2022).
90. Young, A. M. H. et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet* **53**, 861–868 (2021).
91. Bioconductor core team and bioconductor package maintainer. TxDb.Hsapiens.UCSC.hg38.knownGene: annotation package for TxDb object(s). *R package version 3.10.0*. https://doi.org/10.18129/B9.bioc.TxDb.Hsapiens.UCSC.hg38.knownGene (2019).
92. McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation

for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**, 1262–1272 (2020).
93. The Whole Genome Sequence Harmonization Study. AD Knowledge Portal. https://adknowledgeportal.synapse.org/Explore/Studies/DetailsPage/StudyDetails?Study=syn22264775. Accessed 1 Sept. 2023.
94. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
95. *Picard Tools - By Broad Institute*. http://broadinstitute.github.io/picard/. (2022).
96. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
97. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet* **81**, 559–575 (2007).
98. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
99. Safieh, M., Korczyn, A. D. & Michaelson, D. M. ApoE4: an emerging therapeutic target for Alzheimer's disease. *BMC Med.* **17**, 64 (2019).
100. Aryee, M. J. et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
101. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
102. Murat, K. et al. Ewastools: infinium human methylation BeadChip pipeline for population epigenetics integrated into galaxy. *Gigascience* **9**, giaa049 (2020).
103. Pidsley, R. et al. A data-driven approach to preprocessing Illumina 450 K methylation array data. *BMC Genomics* **14**, 293 (2013).
104. Teschendorff, A. E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).
105. van Rooij, J. et al. Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. *Genome Biol.* **20**, 235 (2019).
106. Tang, J., Liu, J., Zhang, M. & Mei, Q. Visualizing large-scale and high-dimensional data. In: *Proceedings of the 25th International Conference on World Wide Web* 287–297 https://doi.org/10.1145/2872427.2883041. (2016).
107. Blighe, K. PCAtools: everything principal component analysis. https://www.bioconductor.org/packages/devel/bioc/vignettes/PCAtools/inst/doc/PCAtools.html (2023).
108. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
109. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
110. Karolchik, D. et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
111. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
112. liftOver: Changing genomic coordinate systems with rtracklayer::liftOver. *Bioconductor* http://bioconductor.org/packages/liftOver/. (2018).
113. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
114. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
115. Molecular, Q. T. L. discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).
116. Lee, Y., Luca, F., Pique-Regi, R. & Wen, X. Bayesian multi-SNP genetic association analysis: control of FDR and use of summary statistics. Preprint at https://doi.org/10.1101/316471 (2018).
117. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet.* **18**, e1010299 (2022).
118. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
119. MacDonald, J. W., Harrison, T., Bammler, T. K., Mancuso, N. & Lindström, S. An updated map of GRCh38 linkage disequilibrium blocks based on European ancestry data. Preprint at https://doi.org/10.1101/2022.03.04.483057 (2022).
120. Quick, C., Wen, X., Abecasis, G., Boehnke, M. & Kang, H. M. Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis. *PLOS Genet.* **16**, e1009060 (2020).
121. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
122. Eulalio, T. et al. Regionalpcs ROSMAP differential methylation results. *Zenodo* https://doi.org/10.5281/zenodo.14004291 (2024).
123. Eulalio, T. et al. Regionalpcs ROSMAP Fine-mapped QTLs. *Zenodo* https://doi.org/10.5281/zenodo.14020225 (2024).
124. Eulalio, T. et al. Regionalpcs ROSMAP GWAS integration. *Zenodo* https://doi.org/10.5281/zenodo.14016038 (2024).
125. Eulalio, T. et al. ROSMAP meQTL results for endothelial cells with regionalpcs and averages. *Zenodo* https://doi.org/10.5281/zenodo.14027678 (2024).
126. Eulalio, T. et al. ROSMAP meQTL results for bulk cells with regionalpcs and averages. *Zenodo* https://doi.org/10.5281/zenodo.14027807 (2024).
127. Eulalio, T. et al. ROSMAP meQTL results for neurons with regionalpcs and averages. *Zenodo* https://doi.org/10.5281/zenodo.14029321 (2024).
128. Eulalio, T. et al. ROSMAP meQTL results for oligodendrocytes with regionalpcs and averages. *Zenodo* https://doi.org/10.5281/zenodo.14029154 (2024).
129. Eulalio, T. et al. ROSMAP meQTL results for astrocytes with regionalpcs and averages. *Zenodo* https://doi.org/10.5281/zenodo.14028091 (2024).
130. Eulalio, T. et al. ROSMAP meQTL results for neurons with CpGs. *Zenodo* https://doi.org/10.5281/zenodo.14029286 (2024).
131. Eulalio, T. et al. ROSMAP meQTL results for oligodendrocytes with CpGs. *Zenodo* https://doi.org/10.5281/zenodo.14029094 (2024).
132. Eulalio, T. et al. ROSMAP meQTL results for astrocytes with CpGs. *Zenodo* https://doi.org/10.5281/zenodo.14027962 (2024).
133. Eulalio, T. et al. ROSMAP meQTL results for bulk tissue cells with CpGs. *Zenodo* https://doi.org/10.5281/zenodo.14027753 (2024).
134. Eulalio, T. et al. ROSMAP meQTL results for endothelial cells with CpGs. *Zenodo* https://doi.org/10.5281/zenodo.14027718 (2024).
135. Eulalio, T. Y. & Sun, M. W. regionalpcs improve discovery of DNA methylation associations with complex traits. *Zenodo* https://doi.org/10.5281/ZENODO.14004153 (2024).

Acknowledgements

We extend our gratitude to the participants of the ROSMAP study and their families, whose generosity made this research possible. We acknowledge the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, for providing access to data collected from the Religious Orders Study and the Memory and Aging Project. Special thanks to the National Institute on Aging (NIA) for funding the creation and sharing of the ROSMAP dataset (grants P30AG10161, R01AG15819, R01AG17917, R01AG36836). The results published here are in whole or in part based on data obtained from the AD Knowledge Portal. We extend our sincere appreciation to Jake Chang (Department of Biomedical Data

Science, Stanford University) and Chansen Hesia for their significant contributions to this study. Jake Chang played a crucial role in the early development of the IV analysis. Chansen Hesia provided steadfast support throughout the project's duration and contributed to the review of the final manuscript. We appreciate the technical support provided by Stanford University and the SCG Informatics Cluster for computational resources. Lastly, we're grateful to the Bioconductor community for their indispensable tools and forums that supported this work. This research was supported by funding from the National Institute of Aging grants R01AG066490 (T.E., D.N., S.B.M.) and U01AG072573 (T.E., D.N., S.B.M.); National Institute of Health (NIH) T15 NLM007033 (T.E., MWS) and T32AG047126 (D.N.) training grants, R01AG066490 (T.E., D.N., S.B.M.), U01AG072573 (T.E., D.N., S.B.M.), R01MH125244 (T.E., D.N., S.B.M.), NIH NHGRI IGVF Program grant U01HG012069-03 (T.E., D.N., S.B.M.); and project grant P30AG066515 (M.D.G.).

Author contributions

T.E., D.N., and S.B.M. were responsible for the study's conception. The design was developed by T.E., M.W.S., O.G., M.D.G., D.N., and S.B.M. T.J.M. and S.B.M. obtained the necessary funding. Data collection, preprocessing, methodology development, and statistical analysis were conducted by T.E., D.N., and S.B.M., who also developed and implemented simulation analyses with M.W.S. The manuscript's first draft was written by T.E., M.W.S., D.N., and S.B.M., and all authors participated in the review and editing process. Each author has approved the final article.

Competing interests

S.B.M. is an advisor to Character Bio, MyOme, PhiTech and Tenaya Therapeutics. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-55698-6.

Correspondence and requests for materials should be addressed to Tiffany Eulalio, Daniel Nachun or Stephen B. Montgomery.

Peer review information *Nature Communications* thanks Linn Gillberg, Guoqiang Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.