scMD: cell type deconvolution using single-cell DNA methylation references

Manqi Cai[1], Jingtian Zhou[2,3], Chris McKennan[4], and Jiebiao Wang[1,5,*]

[1]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA
[2]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, USA
[3]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA
[4]Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA
[5]Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, USA
[*]Correspondence: jbwang@pitt.edu

Abstract

The proliferation of single-cell RNA sequencing data has led to the widespread use of cellular deconvolution, aiding the extraction of cell type-specific information from extensive bulk data. However, those advances have been mostly limited to transcriptomic data. With recent development in single-cell DNA methylation (scDNAm), new avenues have been opened for deconvolving bulk DNAm data, particularly for solid tissues like the brain that lack cell-type references. Due to technical limitations, current scDNAm sequences represent a small proportion of the whole genome for each single cell, and those detected regions differ across cells. This makes scDNAm data ultra-high dimensional and ultra-sparse. To deal with these challenges, we introduce scMD (single cell Methylation Deconvolution), a cellular deconvolution framework to reliably estimate cell type fractions from tissue-level DNAm data. To analyze large-scale complex scDNAm data, scMD employs a statistical approach to aggregate scDNAm data at the cell cluster level, identify cell-type marker DNAm sites, and create a precise cell-type signature matrix that surpasses state-of-the-art sorted-cell or RNA-derived references. Through thorough benchmarking in several datasets, we demonstrate scMD's superior performance in estimating cellular fractions from bulk DNAm data. With scMD-estimated cellular fractions, we identify cell type fractions and cell type-specific differentially methylated cytosines associated with Alzheimer's disease.

Background

Tissue-level quantification of omics has gained popularity in the last decades because of its mature technology and affordable cost. Numerous studies on tissue-level omics, such as gene expression and DNA methylation (DNAm), provide rich resources to help answer interesting biological questions. However, bulk omics data are generated from a mixture of cells, meaning tissue-level analyses are often confounded by cellular heterogeneity, and cell type-specific (CTS) signals are obscured. While labor-intensive technologies such as flow cytometry and immunohistochemistry (IHC) can help measure cell type compositions, they are costly and more challenging for solid tissues[1]. As a cost-efficient alternative, *in silico* cellular deconvolution methods have been developed to recover the cell type composition of bulk omics data, allowing us to adjust for confounding cellular heterogeneity and infer CTS associations from bulk data[2,3,4].

Recent advances in single-cell technology have fueled numerous studies, leveraging high throughput single-cell RNA sequencing (scRNA-seq) as a reference to estimate cellular fractions in bulk RNA-seq data[5,6]. However, this progress in scRNA-seq stands in stark contrast to single-cell DNA methylation (scDNAm), which remains less studied. As a consequence, DNAm-based cell proportion estimates are often imprecise and can only be obtained for coarse cell types compared to RNA-based deconvolution. For example, deconvolving brain DNAm has been predominantly restricted to references derived from two cell types: neurons and non-neurons[7].

Recently, EpiSCORE[8] was proposed to deconvolve brain DNAm into six cell types. It employs scRNA-seq data to create a proxy signature for DNAm at the gene level. Specifically, EpiSCORE uses a scRNA-seq-derived reference to impute the DNAm at the promoter regions of marker genes and runs deconvolution based on these imputed signatures. However, not all CpGs in the promoter region are CTS, and EpiSCORE's imputation function mapping marker gene counts to promoter DNAm is not CTS. These compromise the cell type-specificity and accuracy of their DNAm signature, which are critical for the fidelity of deconvolution[9].

Fortunately, scDNAm has been emerging in the last few years, especially for the brain[10,11,12,13]. The data exhibits strong cell type specificity, offering the potential to deconvolve tissue-level DNAm data. However, due to technical limitations, these methods usually detect only a small fraction of the genome in each single cell ($\sim$5% of all CpG sites), and the regions being detected could be highly variable between cells. Consequentially, the data is ultrahigh-dimensional and sparse, presenting considerable computational challenges.

To address these issues, we developed scMD (single cell Methylation Deconvolution), which uses scDNAm data to generate a high-quality DNAm reference and deconvolve bulk DNAm data. scMD leverages the strong cell type-specificity exhibited by scDNAm markers to perform high-resolution and accurate cellular deconvolution. Critically, scMD addresses the statistical and methodological hurdles that accompany scDNAm data, including its ultrahigh-dimensionality and sparsity, to identify cell-type marker CpGs and construct a signature that is amenable to bulk DNAm data. We use six real bulk DNAm datasets to illustrate scMD's superior performance over existing methods, where we show its ability to better estimate cellular fractions and infer Alzheimer's disease-related cell types. With scMD, we can complement bulk DNAm analyses with estimated cellular fractions to deconfound tissue-level analyses and enable CTS analyses.

Results

Overview of scMD

Here we provide an overview of scMD, which uses scDNAm data to construct a DNAm signature amenable to bulk data and perform deconvolution (Fig. 1). The most challenging aspect of scDNAm is its high dimensionality and sparsity, which arises because only a small fraction ($\sim$ 5%) of the roughly 53 million DNAm sites are measured in each cell (Supplementary Table S1). The set of measured sites is cell-specific, meaning cell-type marker selection and signature matrix generation tools that require fully observed data, like those traditionally employed in scRNA-seq data[14,15], are not applicable in scDNAm. To address this, we subset sites observed in bulk data, e.g., CpGs on Illumina's 450k and EPIC arrays or in whole genome bisulfite sequencing (WGBS), and aggregate them across cells of the same type to obtain a much smaller and more computationally tractable cell cluster-level dataset. With methylated and unmethylated read counts, we then use Fisher's exact test to identify cell-type marker CpGs from cluster-level scDNAm data (Methods). This results in CTS p-values that compare one cell type with all other cell types. We finally define our signature matrix to be the beta values of marker sites in each cell type.

In contrast to existing DNAm-based deconvolution approaches that segregate brain tissue into coarse cell types (neurons and non-neurons)[7] or use RNA-derived signatures[16], our method takes advantage of recent advancements in brain scDNAm resources[11,13] to construct the first brain scDNAm signature matrices encompassing seven distinct cell types: astrocytes, endothelial cells, excitatory neurons, inhibitory neurons, microglia, oligodendrocytes, and oligodendrocyte progenitor cells (OPC) (Fig. 1). After constructing the DNAm signature matched with the target bulk DNAm data, we utilize our previously developed robust and accurate ensemble cellular deconvolution method, EnsDeconv[9], to integrate different scDNAm references, data transformations, and deconvolution algorithms. We explore all suitable combinations of these factors and utilize CTS robust regression to obtain the optimal ensemble of cellular fractions.

Validating scMD using sorted-cell data

We assessed the accuracy of scMD in deconvolution using three different sorted-cell datasets derived from various DNAm platforms. This evaluation allowed us to understand scMD's performance across multiple technologies and gauge its proficiency in accurately deconvolving different purified-cell samples. We first tested scMD with the dataset from Mendizabal *et al.*[17], which quantified WGBS DNAm from sorted neurons (NeuN+) samples and OLIG2+ samples that indicate oligodendrocytes and OPC. We then utilized the datasets from Guintivano *et al.*[18] and Gasparoni *et al.*[19], both containing sorted-cell DNAm samples from NeuN+ and non-neurons (NeuN-). All samples from the three datasets have definitive fractions of non-neurons, neurons, or the sum of oligodendrocytes and OPC. These datasets provided an opportunity to accurately measure scMD's performance in identifying and distinguishing between various major brain cell types. Further details about the validation datasets and the approaches employed for evaluating the performance of scMD are outlined in the Methods section and Supplementary Table S2.

We carried out a comparative analysis of scMD with EpiSCORE[16]. Tested on the Mendizabal dataset[17], scMD almost perfectly fits the data, accurately deconvolving the neuron and oligodendrocyte samples (Fig. 2a). Compared to EpiSCORE, which has difficulty differentiating OLIG2+ and other cell types, our proposed method excelled in effectively identifying sorted-cell samples as their corresponding cell types. This suggests that scMD can effectively harness signals from the originally sparse scDNAm data. We also evaluated its accuracy in deconvolving 450k array-based samples available from Guintivano *et al.*[18] (Fig. 2b) and Gasparoni *et al.*[19] (Fig. 2c), which underscore scMD's ability to accurately deconvolve both NeuN+ and NeuN- samples, thereby demonstrating its versatility and efficiency in brain cell deconvolution.

scMD accurately estimates cellular fractions in cerebral cortex

To gain deeper insights into the performance of scMD, we conducted a comprehensive comparison of scMD with various other deconvolution methods using real bulk data with IHC-measured cell counts of four cell types from cerebral cortex samples that were part of the Religious Orders Study (ROS)[20]. We also used our signatures as input into existing deconvolution methods to demonstrate the importance of our novel signature matrices and illustrate the fidelity of EnsDeconv when applied to DNAm.

On average, scMD significantly outperforms EpiSCORE (Fig. 3a). Especially EpiSCORE exhibits a low correlation with the measured fractions of microglia and astrocytes. This is because EpiSCORE consistently estimates microglia fractions to be zero and tends to overestimate astrocyte fractions (Fig. 3b). In contrast, the fractions estimated by scMD and those measured through IHC are consistent, especially for astrocytes and microglia (Fig. 3c). This alignment underscores the importance of accurately estimating microglia fractions, as microglia is a crucial brain cell type implicated in multiple diseases, such as Alzheimer's disease[21]. Results also show that provided they utilize our scDNAm-based signature, existing deconvolution methods also outperform EpiSCORE (Fig. 3a), thereby further illustrating the accuracy of our signatures. We do note, however, that scMD, which utilizes EnsDeconv to perform deconvolution, outperforms all methods.

Consistent cellular fractions estimated from DNAm and mRNA

While it is ideal to validate scMD with measured cell counts, the resources are limited to major cell types and small sample sizes given the challenges of counting cell types in solid tissues like the brain. Instead, the deconvolution of RNA-seq data has been well benchmarked and thus can be used as "gold standard"[9]. In addition to the ROS data, we further validated scMD in more cell types and a different platform with the dataset from Markunas *et al.*[22], which sequenced paired DNAm of Illumina EPIC arrays and RNA-seq bulk data from the nucleus accumbens (NAc) of 211 individuals. Even though we do not have measured cell counts, the cellular fractions are available from deconvolving paired mRNA data. The intuition is that if we possess paired DNAm and RNA bulk data from the same tissue samples, we should observe high concordance between the estimated cellular fractions from these two omics types, given that there is a single true cellular composition for a tissue sample.

With the above rationale, we first estimated cellular fractions using RNA data[9] as the gold standard of cellular fractions for benchmarking. Equipped with our newly constructed signature matrices, we deconvolved the NAc bulk DNAm and RNA data and examined the deconvolution results between these two omics types. We obtained a strong correlation between estimated RNA- and DNAm-based fractions when we employed scMD to deconvolve DNAm samples and EnsDeconv on RNA samples (Fig. 4a). Except for OPC, all correlations exhibited were above 0.5. The correlation was especially noticeable among major cell types such as neurons and oligodendrocytes, where correlations of 0.82 and 0.89 were observed. Furthermore, the correlation remained high (0.74) even for the less common endothelial cells. In contrast, when using EpiSCORE to infer cellular fractions from DNAm, the correlations are lower than those of scMD across all cell types (Fig. 4b). Notably, EpiSCORE consistently estimates microglia fractions to be approximately zero, and its correlations for astrocytes and OPC are both negative.

scMD identifies cell types associated with Alzheimer's disease

To demonstrate the utility of scMD-estimated cellular fractions, we tested their associations with clinical phenotypes related to Alzheimer's disease (AD). We utilized the brain DNAm data from Mount Sinai Brain Bank (MSBB), which also collected variables such as age, Clinical Dementia Rating (CDR), the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) score, and Braak stage—a widely used classification system indicating the progression of AD. The CDR was employed as an assessment tool to evaluate dementia and cognitive status, assigning ratings on a scale of 0 to 5, which correspond to escalating levels of severity in pathology[23]. CERAD score is a four-level semi-quantitative measure of neuritic plaques. Braak stage categorizes the advancement of neurofibrillary tangles and amyloid plaques in the brain, with stages ranging from 0 to 6, representing increasing levels of pathology severity[24,25].

Given the neurodegeneration that accompanies AD, comparing cell-type fractions across age and various AD phenotypes is therefore of scientific interest. We conducted a comprehensive study examining the correlation of various phenotypes in MSBB with estimated cellular fractions using scMD and EpiSCORE (Fig. 5a and Supplementary Table S3). As expected, scMD detected a significant decrease in OPC and inhibitory neurons with aging, but EpiSCORE did not identify any cell types associated with age. Among the three AD-related phenotypes, we found the most differential fraction signals in clinical dementia rating. With scMD, we observed significantly increased fractions of microglia and oligodendrocyte and decreased fractions of OPC, excitatory, and inhibitory neurons, while EpiSCORE only identified a significant increase in astrocytes and oligodendrocytes and a decrease in neurons.

Interestingly, as two aspects of AD, neuritic plaques and neurofibrillary tangles show strikingly different differential fraction results. Both scMD and EpiSCORE did not identify any cell types associated with neuritic plaques (as indicated by CERAD score), but there are some cell types associated with neurofibrillary tangles (as measured by Braak score). For instance, scMD-estimated microglia and excitatory neuron proportions increase and decrease as the Braak stage increases, respectively (Fig. 5b), and inhibitory neuron proportions exhibit little change. The observed increase in microglia proportions suggests an enhanced immune response and neuroinflammation, which are known to be critical in neurodegenerative disorders

like AD[21]. Additionally, the substantial decline in excitatory neurons is a compelling finding. Excitatory neurons play a crucial role in signal transmission and neural communication within the brain. The reduction in their cell count implies potential disruptions in synaptic activity and impaired neuronal function in affected brain regions. These findings align with previous research emphasizing neuronal loss as a sign of neurodegenerative disease. Similarly, we also used EpiSCORE to estimate cell fractions from MSBB data and identified cell types associated with AD. EpiSCORE identified oligodendrocytes and OPC associated with the Braak score. Consistent with Figure 4b, EpiSCORE estimates microglia proportions to be almost all zero and therefore not able to infer a significant correlation between their proportions and Braak stage. While the decrease in neurons among AD patients is confirmed with EpiSCORE (Fig. 5c), it lacks the resolution to show the decrease is primarily driven by excitatory neurons since it does not estimate neuronal subtypes.

Furthermore, scMD-estimated cellular fractions enable CTS differential methylation analyses. We used CellDMC[26] to identify cell type-specific differentially methylated cytosines (CTS-DMCs) (Supplementary Table S4 and Supplementary File). With scMD-estimated cellular fractions, we identified 38 CTS-DMCs in microglia associated with age (Fig. 5d) and 13 DMCs in OPC with FDR $< .05$. Notably, among the most significant CpGs in microglia, cg18574144 is within the gene body of *THOP1*, which is currently under investigation as a potential biomarker for Alzheimer's disease[27]. For CDR, we detected 221 DMCs in astrocytes and 20 DMCs in OPC. We also identified dozens of DMCs in excitatory and inhibitory neurons and oligodendrocytes associated with CERAD (neuritic plaques) and in inhibitory neurons for Braak score (neurofibrillary tangles).

Discussion

The scMD method that we have developed presents a significant step forward in the ability to analyze and understand the cellular heterogeneity of the brain at the molecular level using DNAm data. By constructing signature matrices for seven distinct brain cell types, scMD offers a much finer level of detail than previous deconvolution methods. Our method goes beyond existing approaches by effectively leveraging recent advancements in scDNAm data resources, bridging the gap between single-cell and bulk DNAm data. This utilization of single-cell data in the generation of our signature matrices captures the intrinsic cellular heterogeneity of the brain, which is an important consideration in the study of various brain-related diseases and conditions.

The accuracy of scMD is reflected in its high performance in various validation studies across different DNAm platforms. First, scMD consistently outperformed other approaches in the deconvolution of purified-cell and bulk datasets, highlighting its robustness and potential for widespread application. Furthermore, scMD demonstrated high concordance between the RNA-estimated fractions and DNAm-estimated fractions, suggesting that scMD is successful in capturing useful signals from the original sparse scDNAm data. Lastly, we showed that scMD can precisely identify microglia and excitatory neurons associated with AD.

Despite the evident promise shown by scMD, it is essential to recognize certain challenges and limitations that may warrant future work. First, there may be a computational burden presented by processing scDNAm data due to its high dimensionality. The massive volumes

of raw data require computational resources to process and reduce storage and memory. However, given the availability of scDNAm data on public online platforms, we have mitigated this issue through our parallel computation approach. This method allows for the rapid and efficient processing of many cells at once during the construction of a cell-type signature matrix. Another potential issue lies in the performance of the model being contingent on the quality and scope of the reference single-cell data used to build the signature matrices. Despite incorporating diverse sources of data to create our matrices, representation of certain cell types, particularly rare ones, may be limited. Furthermore, the method's efficacy in tissues and conditions not represented in the training data awaits further evaluation. As scDNAm is increasingly applied to various tissue types, we anticipate a broadened use of scMD beyond the brain.

Conclusion

We present a robust and versatile tool for researchers to deconvolve bulk DNAm data with scDNAm references. By offering more accurate, detailed, and efficient analyses of brain cell composition from DNAm data, our method opens up new avenues for exploring the molecular underpinnings of brain function and pathology. In our future work, we plan to refine and expand the capabilities of scMD. We aim to incorporate additional cell types and explore various tissue types with the expansion of scDNAm. Additionally, we aim to integrate other omics data to gain novel insights into cellular heterogeneity in the brain and other tissues.

Methods

Details of the proposed scMD framework

Processing scDNAm data

Our goal is to build high-quality DNAm references using scDNAm technology. The initial step towards achieving this objective involves harmonizing the differences in DNAm technologies between traditional bulk DNAm data and the noisier scDNAm data. Compared to bulk DNAm data, scDNAm is significantly sparser and higher in dimensionality. The challenge is to bridge the gap between the dimensionality of the traditional bulk DNAm data and that of the scDNAm data. Traditional bulk DNAm data typically use arrays of 450k or 850k CpG sites, while scDNAm is characterized by its sparse quantification across billions of genomic locations. To address this issue, we employ a strategy that involves subsetting the DNAm sites sequenced with 450k or 850k arrays or WGBS. This approach serves two-fold. Firstly, it accelerates the overall process. Secondly, it simplifies the process of identifying marker CpGs, which are crucial for various analyses in the field of epigenetics. By employing this technique, we can successfully reduce the dimensionality of scDNAm data to make it comparable to its bulk counterpart. After achieving a reduced-dimension scDNAm dataset, the subsequent step is to derive CTS p-values to identify marker CpGs. Given the inherent sparsity of scDNAm data, characterized by missing values, it is not feasible to identify specific markers and construct signatures in the same manner as with scRNA-seq data.

To generate a DNAm signature matrix akin to a reference, we first aggregated the methylated counts and coverage of DNAm of each cell type:

$$mc_{ik} = \sum_{j=1}^{N_k} mc_{ijk}, \quad cov_{ik} = \sum_{j=1}^{N_k} cov_{ijk},$$

where mc_{ik} represents the cumulative count of all methylcytosine for the i^{th} CpG site and k^{th} cell type, N_k denotes the number of cells belonging to the cell type k, and cov_{ik} is the total cytosine basecalls, incorporating both methylcytosine and unmethylcytosine for the i^{th} DNAm site and k^{th} cell type. Subsequently, we carry out two-sided Fisher's exact tests for each cell type across all DNAm sites. To differentiate cell type k from all other cell types for a given DNAm site i, we formulated the following table:

	Cell type k	Other cell types	Row total
Methylcytosine	$M_{ik} = mc_{ik}$	$\sum_{k' \neq k} M_{ik'} = \sum_{k' \neq k} mc_{ik'}$	$\sum_{k} M_{ik}$
Unmethylcytosine	$U_{ik} = cov_{ik} - mc_{ik}$	$\sum_{k' \neq k} U_{ik'} = \sum_{k' \neq k} cov_{ik'} - \sum_{k' \neq k} mc_{ik'}$	$\sum_{k} U_{ik}$
Column total	cov_{ik}	$\sum_{k' \neq k} cov_{ik'}$	$\sum_{k} cov_{ik}$

The p-value of Fisher's exact test corresponding to cell type k at the DNAm site i is derived as

$$p_{ik} = \frac{\left(\sum_{k} M_{ik}\right)! \left(\sum_{k} U_{ik}\right)! \left(\sum_{k' \neq k} cov_{ik}\right)! (cov_{ik})!}{(M_{ik})! \left(\sum_{k' \neq k} M_{ik'}\right)! (U_{ik})! \left(\sum_{k' \neq k} U_{ik'}\right)! \left(\sum_{k} cov_{ik}\right)!}.$$

Once we calculated the CTS p-values for each cell type across all DNAm sites, we arranged these values in ascending order. Based on a detailed evaluation of two sorted-cell datasets, we selected the top 100 marker DNAm sites for each cell type based on their p-values (Supplementary Fig. S1). This aligns with existing methods, such as minfi[2], which opted to select the top 100 differentially methylated marker DNAm sites per cell type. We believe this strategic selection offers dual benefits. It not only accelerates the deconvolution process, making the computational burden manageable for extensive bulk DNAm datasets, especially for WGBS bulk DNAm data, but also enhances accessibility for the broader scientific community. By reducing computational costs, our approach alleviates the challenges for researchers, particularly those in resource-constrained settings, when handling large datasets.

Here we use an example to illustrate how scMD handles the large-scale raw scDNAm data. The original compressed raw scDNAm files for more than 4,200 nuclei from Lee *et al.*[11] totaled a substantial 717 GB. After filtering for only CpG sites, the data was condensed to 183.2 GB. Further refinement at the cluster level reduced the data size to a more manageable 17.4 GB before loading into the R environment. Through the application of parallel computation across 20 nodes, we were able to generate an 850k-based signature within approximately 10 minutes. Supplementary Table S1 provides detailed information on the number of DNAm sites before and after processing.

Ensemble deconvolution (EnsDeconv)

After obtaining the signature matrix from scDNAm data as described earlier, the subsequent crucial step involves performing deconvolution on DNAm datasets. To accomplish this, we employed our previously developed method EnsDeconv[9]. In essence, EnsDeconv represents a deconvolution technique that draws inspiration from ensemble learning, wherein the outputs of multiple deconvolution algorithms are combined to achieve enhanced estimation accuracy. EnsDeconv focuses on important factors such as the choice of reference datasets, data transformations, and deconvolution methods. EnsDeconv implements CTS robust regression to synthesize results from different deconvolution settings, resulting in more robust and accurate results than randomly choosing one setting. Taking into account all possible combinations of the aforementioned factors in deconvolution, we leveraged ensemble learning to generate $\hat{\boldsymbol{P}}_1, \ldots, \hat{\boldsymbol{P}}_D$, representing the estimated cellular proportions from each of the D scenarios. In this context, we define a scenario as a specific setting with a particular reference dataset, transformations approach, and deconvolution method. We treat the ensemble learning problem as a robust regression problem:

$$\underset{\substack{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_K\in[0,1]^S \\ (\boldsymbol{W}_1,\cdots,\boldsymbol{W}_K)\mathbf{1}_K=\mathbf{1}_S}}{\operatorname{argmin}} \sum_d \sum_{k=1}^{K} \|\hat{\boldsymbol{W}}_{dk} - \boldsymbol{W}_k\|_2,$$

where $\boldsymbol{W}_k$ denotes the k-th cell type's ensemble fraction for S samples, $\hat{\boldsymbol{W}}_{dk}$ represents the estimate for the k-th cell type fraction in the d-th deconvolution scenario, and $\|\boldsymbol{v}\|_2 = (\sum_i \boldsymbol{v}_i^2)^{1/2}$ is the vector equivalent of absolute deviation.

In this study, we utilized two scDNAm references, Lee *et al.*[11] and Tian *et al.*[13], to implement the EnsDeconv approach. In terms of data transformations, scDNAm adopts both beta-value and M-value transformations. In addition, our implementation of EnsDeconv incorporated nine diverse deconvolution methods, each founded on unique theoretical bases and specifically designed for various purposes. A portion of these techniques was originally developed for deconvolving bulk DNAm data, e.g., quadratic programming[28] and robust partial correlations (RPC)[29]. In parallel, we also integrated several deconvolution methods primarily designed for RNA-seq experiments. These included the robust regression technique from FARDEEP[30], support-vector regression from CIBERSORT[31], the penalized regression method with elastic net regularization featured in DCQ[32], a log-normal model from ICeDT[33], and non-negative least squares (NNLS).

DNA methylation datasets

Brain scDNAm datasets

In this study, we began by generating a reference for scDNAm using snmC-seq data obtained from Lee *et al.*[11]. The dataset utilized in this study is comprised of 4,238 single human brain prefrontal cortex cells, enabling the simultaneous capture of chromatin organization and DNA methylation information. The scDNAm data was downloaded from the GEO database (GSE130711). To ensure data quality, we utilized the cell-type annotation provided by the authors and excluded any cells marked as outliers, resulting in a total of 4,234 cells for further analysis. The distribution of cell-type annotations in the remaining dataset consisted of

670 inhibitory neurons (InN), 945 excitatory neurons (ExN), 1,250 oligodendrocytes (Oligo), 449 astrocytes (Astro), 416 microglial cells (Micro), 315 endothelial cells (Endo), and 189 oligodendrocyte progenitor cells (OPC). To map the scDNAm data to the DNAm sites in the bulk DNAm dataset, we specifically considered cytosines in the CG context while excluding those in the CH context. Additionally, we incorporated data from a newly collected dataset Tian *et al.*[13], which employed snmC-seq3 technology to profile whole-genome DNAm data. We obtained the cluster-level data from the frontal cortex for the same seven cell types as Lee *et al.*[11].

Sorted-cell brain DNAm datasets for validation

Descriptions of DNAm validation datasets used in this part are summarized in Supplementary Table S2. In order to assess the accuracy of our signature matrices, we used three sorted-cell datasets that contained either sorted neuron samples and non-neuron samples or oligodendrocyte samples as validation datasets. The first dataset Mendizabal[17] is a whole-genome bisulfite sequencing (WGBS) postmortem human brain dataset. It is composed of two cell populations: NeuN+ and OLIG2+. We focus on healthy controls, and the sample size is 25 and 20 respectively for the two cell types. The data is downloaded from GEO (GSE108066). The second dataset Guintivano[18] is an Illumina Human 450k Methylation dataset and profiled in the postmortem frontal cortex of two different cellular populations (NeuN+ vs. NeuN-) generated from 29 individuals using flow sorting. We downloaded the Guintivano data through the Bioconductor package FlowSorted.DLPFC.450k. The third dataset Gasparoni[19] is an Illumina Human 450k Methylation dataset that contains 62 sorted-cell frontal cortex brain samples, including 31 NeuN+ samples and 31 NeuN- samples. Gasparoni data is available at GEO (GSE66351). We processed the Mendizabal data by extracting a subset of DNAm sites that corresponds to the specific locations matched with scMD and EpiSCORE references respectively. We prepossessed the Gasparoni, and Guintivano data using the minfi package[7]. We evaluated the performance of scMD and EpiSCORE using total mean absolute error (MAE) comparing estimated and measured fractions.

Bulk brain DNAm datasets for validation

The bulk DNA methylation (DNAm) data for the Mount Sinai Brain Bank (MSBB)[23] were obtained from Synapse (ID: syn21347197). This data encompasses 201 tissue samples derived from the parahippocampal gyrus region of the brain (Brodmann area 36), and processed using the Illumina 850k platform. We first deconvolved the MSBB DNAm data, subsequently examining the relationship between cellular fractions and the Braak stage of Alzheimer's disease (AD).

We also used brain DNAm data from the Religious Orders Study (ROS), specifically from the dorsolateral prefrontal cortex (DLPFC) tissue of 49 senior donors. This dataset incorporates both bulk DNAm data, captured through a 450k array as described by De Jager *et al.*[20], and measured cell-type fractions as reported by Patrick *et al.*[34]. The study measured the proportions of four distinct cell types, namely astrocyte, microglia, neuron, and oligodendrocyte. Note that we excluded endothelial cells since prior studies confirmed their poor quality of measured cell counts[9].

In addition, paired bulk DNAm and RNA data from the nucleus accumbens (NAc) were obtained from public repositories, GSE147040 and GSE171936, respectively. The DNAm data for NAc was profiled using the Infinium MethylationEPIC/850k microarray following the guidelines provided by the manufacturer. The raw idat files were subsequently processed and normalized using the minfi R package.

Single-cell RNA-sequencing (scRNA-seq) references to deconvolve bulk RNA-seq data

In our prior research, we compiled a selection of scRNA-seq reference datasets[9]. For the present study, we utilized the brain scRNA-seq data curated by STAB[35] from three studies: Darmanis *et al.*[36], Hodge *et al.*[37], and Habib *et al.*[38]. Deconvolution results were then obtained via EnsDeconv. The single deconvolution methods implemented in EnsDeconv to derive the RNA estimated fraction extend beyond those used in DNAm EnsDeconv. Notably, we excluded deconvolution methods initially intended for DNAm deconvolution, including Houseman *et al.*[28] and RPC. This included an additional hybrid scale method—dtangle[39]—and two deconvolution approaches specifically designed for scRNA-seq references: MuSiC[5] and Bisque[6]. These methods are not used in the DNAm deconvolution due to their methodological incompatibility with DNAm data.

EpiSCORE

For a comprehensive comparison, we incorporated the DNAm reference matrix from Zhu *et al.*[16], which employs single-cell RNA sequencing (scRNA-seq) information and obtained using EpiSCORE[8]. The resources needed to generate this reference matrix, including the code and data, were directly sourced from the Code Ocean repository: https://codeocean.com/capsule/2549317/tree/v3.

Availability of data and materials

- scMD is publicly hosted on GitHub (https://github.com/randel/scMD).
- The processed signatures of different platforms can be downloaded from (https://github.com/randel/scMD/tree/main/Processed_data_450k850k).
- EnsDeconv (https://github.com/randel/EnsDeconv).
- The Lee scDNAm data can be obtained through (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130711).
- The Tian scDNAm data is available at http://neomorph.salk.edu/wtian/hba-data/.
- The Guintivano DNAm data can be downloaded from a Bioconductor package named FlowSorted.DLPFC.450k.

- The Gasparoni DNAm data is publicly available on (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66351).
- The Mendizabal DNAm data is publicly available on (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108066).
- The NAc bulk DNAm and RNA data are publicly available on (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147040) and (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171936).
- The ROS and MSBB bulk data and related clinical data are publicly available on AD Knowledge Portal (https://adknowledgeportal.synapse.org/).
- EpiSCORE (https://github.com/aet21/EpiSCORE).

References

1. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology* **25,** 571–578 (2013).
2. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology* **15,** R31 (2014).
3. Zheng, X., Zhang, N., Wu, H.-J. & Wu, H. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biology* **18** (2017).
4. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications* **11,** 1–14 (2020).
5. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications* **10,** 1–9 (2019).
6. Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., Sul, J. H., Pietiläinen, K. H., Pajukanta, P. & Halperin, E. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications* **11,** 1–11 (2020).
7. Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. & Irizarry, R. A. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30,** 1363–1369 (2014).
8. 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 biology* **21,** 1–33 (2020).

9. Cai, M. *et al.* Robust and accurate estimation of cellular fraction from tissue omics data via ensemble deconvolution. *Bioinformatics* **38,** 3004–3010 (2022).

10. Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P., *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357,** 600–604 (2017).

11. Lee, D.-S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J. R., Fitzpatrick, C., O'Connor, C., Dixon, J. R., *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature methods* **16,** 999–1006 (2019).

12. Luo, C., Liu, H., Xie, F., Armand, E. J., Siletti, K., Bakken, T. E., Fang, R., Doyle, W. I., Stuart, T., Hodge, R. D., *et al.* Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell genomics* **2,** 100107 (2022).

13. Tian, W., Zhou, J., Bartlett, A., Zeng, Q., Liu, H., Castanon, R. G., Kenworthy, M., Altshul, J., Valadon, C., Aldridge, A., *et al.* Epigenomic complexity of the human brain revealed by single-cell DNA methylomes and 3D genome structures. *bioRxiv,* 2022–11 (2022).

14. Delaney, C., Schnell, A., Cammarata, L. V., Yao-Smith, A., Regev, A., Kuchroo, V. K. & Singer, M. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Molecular Systems Biology* **15** (2019).

15. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).

16. Zhu, T., Liu, J., Beck, S., Pan, S., Capper, D., Lechner, M., Thirlwell, C., Breeze, C. E. & Teschendorff, A. E. A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. *Nature methods* **19,** 296–306 (2022).

17. Mendizabal, I., Berto, S., Usui, N., Toriumi, K., Chatterjee, P., Douglas, C., Huh, I., Jeong, H., Layman, T., Tamminga, C. A., *et al.* Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome biology* **20,** 1–21 (2019).

18. Guintivano, J., Aryee, M. J. & Kaminsky, Z. A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8,** 290–302 (2013).

19. Gasparoni, G., Bultmann, S., Lutsik, P., Kraus, T. F., Sordon, S., Vlcek, J., Dietinger, V., Steinmaurer, M., Haider, M., Mulholland, C. B., *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,** 1–19 (2018).

20. De Jager, P. L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., Eaton, M. L., Keenan, B. T., Ernst, J., McCabe, C., *et al.* Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature neuroscience* **17,** 1156–1163 (2014).

21. Navarro, V., Sanchez-Mejias, E., Jimenez, S., Muñoz-Castro, C., Sanchez-Varo, R., Davila, J. C., Vizuete, M., Gutierrez, A. & Vitorica, J. Microglia in Alzheimer's disease: activated, dysfunctional or degenerative. *Frontiers in aging neuroscience* **10,** 140 (2018).

22. Markunas, C. A., Semick, S. A., Quach, B. C., Tao, R., Deep-Soboslay, A., Carnes, M. U., Bierut, L. J., Hyde, T. M., Kleinman, J. E., Johnson, E. O., *et al.* Genome-wide DNA methylation differences in nucleus accumbens of smokers vs. nonsmokers. *Neuropsychopharmacology* **46,** 554–560 (2021).

23. Wang, M., Beckmann, N. D., Roussos, P., Wang, E., Zhou, X., Wang, Q., Ming, C., Neff, R., Ma, W., Fullard, J. F., *et al.* The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Scientific data* **5,** 1–16 (2018).

24. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta neuropathologica* **82,** 239–259 (1991).

25. Bennett, D., Schneider, J., Arvanitakis, Z., Kelly, J., Aggarwal, N., Shah, R. & Wilson, R. Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* **66,** 1837–1844 (2006).

26. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nature methods* **15,** 1059–1066 (2018).

27. Hok-A-Hin, Y. S., Bolsewig, K., Ruiters, D. N., Lleó, A., Alcolea, D., Lemstra, A. W., van der Flier, W. M., Teunissen, C. E. & Del Campo, M. Thimet oligopeptidase as a potential CSF biomarker for Alzheimer's disease: A cross-platform validation study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **15,** e12456 (2023).

28. Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. & Kelsey, K. T. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13,** 1–16 (2012).

29. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC bioinformatics* **18,** 1–14 (2017).

30. Hao, Y., Yan, M., Heath, B. R., Lei, Y. L. & Xie, Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS computational biology* **15,** e1006976 (2019).

31. Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M. & Alizadeh, A. A. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12,** 453–457 (2015).

32. Altboum, Z. *et al.* Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular Systems Biology* **10,** 720 (2014).

33. Wilson, D. R., Jin, C., Ibrahim, J. G. & Sun, W. ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *Journal of the American Statistical Association* **115,** 1055–1065 (2020).

34. Patrick, E., Taga, M., Ergun, A., Ng, B., Casazza, W., Cimpean, M., Yung, C., Schneider, J. A., Bennett, D. A., Gaiteri, C., *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Computational Biology* **16,** e1008120 (2020).

35. Song, L., Pan, S., Zhang, Z., Jia, L., Chen, W.-H. & Zhao, X.-M. STAB: a spatio-temporal cell atlas of the human brain. *Nucleic Acids Research* **49,** D1029–D1037 (2021).

36. Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Gephart, M. G. H., Barres, B. A. & Quake, S. R. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* **112,** 7285–7290 (2015).

37. Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573,** 61–68 (2019).

38. Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., Ardlie, K., *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* **14,** 955–958 (2017).

39. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. Dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35,** 2093–2099 (2019).

Acknowledgements

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. The results published here are in part based on data obtained from the AD Knowledge Portal. Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG61356 (whole-genome sequencing, targeted proteomics, ROSMAP AMP-AD) and the Illinois Department of Public Health (ROSMAP). The MSBB data were generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by Dr. Eric Schadt from Mount Sinai School of Medicine.

Funding

This research was funded in part through NIH's R01AG080590, R03OD034501, R01MH123184, and UL1TR001857.

Author contributions

This study was conceived of and led by J.W. Jointly with J.W. and C.M., M.C. designed the algorithm, implemented the scMD software, and led the data analyses. J.Z. helped provide

data and scientific insights on scDNAm. M.C., J.W., C.M., and J.Z. wrote the paper.

Competing interests

The authors declare no competing interests.

Figures

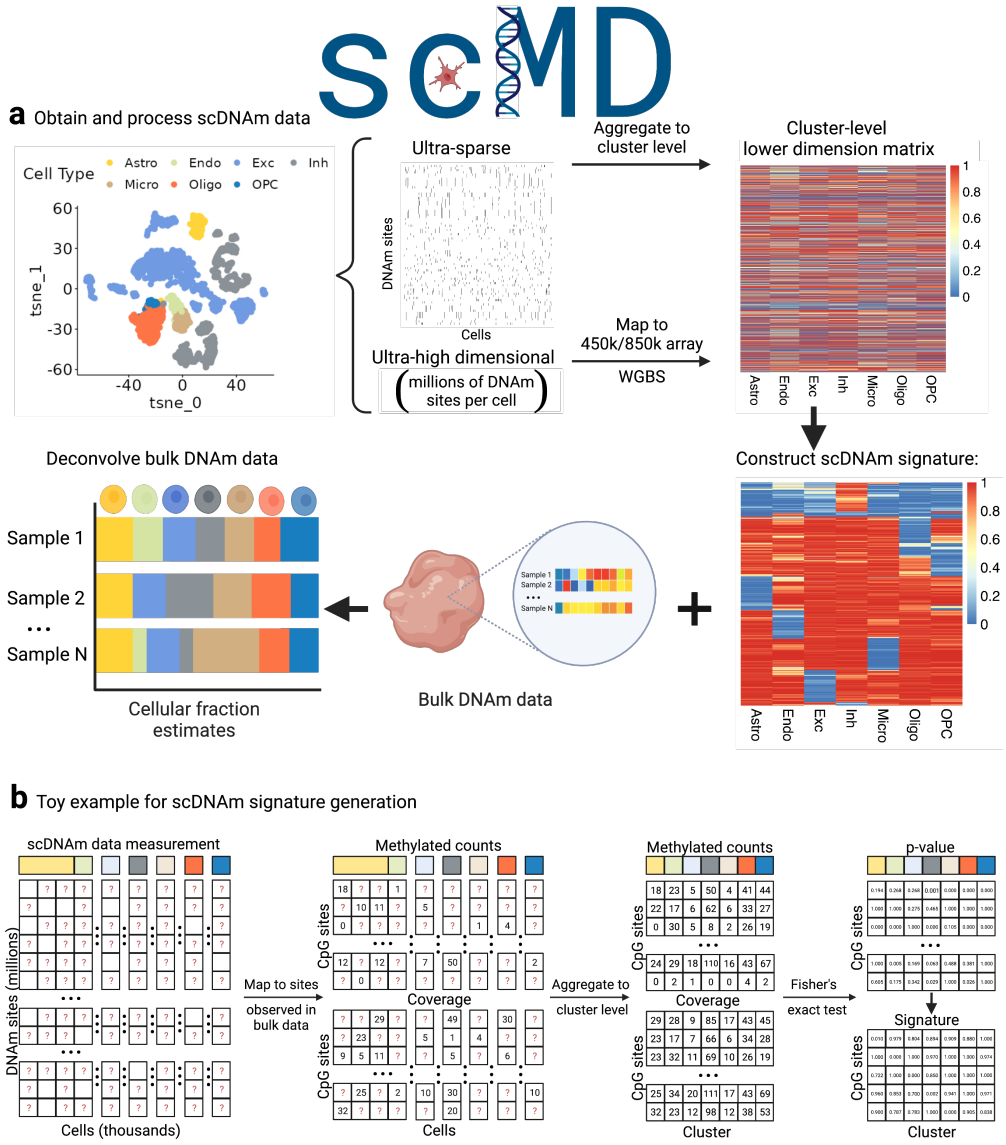

Figure 1: **a**, Schematic representation of the proposed scMD framework. The scMD process is comprised of several steps. First, with single-cell DNA methylation (scDNAm) data and cell cluster labels, the data is filtered to include only DNAm sites present in the 450k or 850k array or WGBS sequencing, addressing the challenge of high dimensionality. Second, the data is aggregated at the cluster level to mitigate the issue of sparsity. Third, Fisher's exact test is utilized to identify marker CpGs by comparing each cell type against all other cell types. Finally, based on the resulting p-values, a distinctive scDNAm signature is constructed. Here we show seven cell types: astrocytes (Astro), endothelial cells (Endo), excitatory neurons (ExN), inhibitory neurons (InN), microglia (Micro), oligodendrocytes (Oligo), and oligodendrocyte precursor cells (OPC). **b**, Detailed demonstration of building DNAm signature matrix from high-dimensional and sparse scDNAm data. Question marks in matrices denote missing data. Column annotations denote cell types.

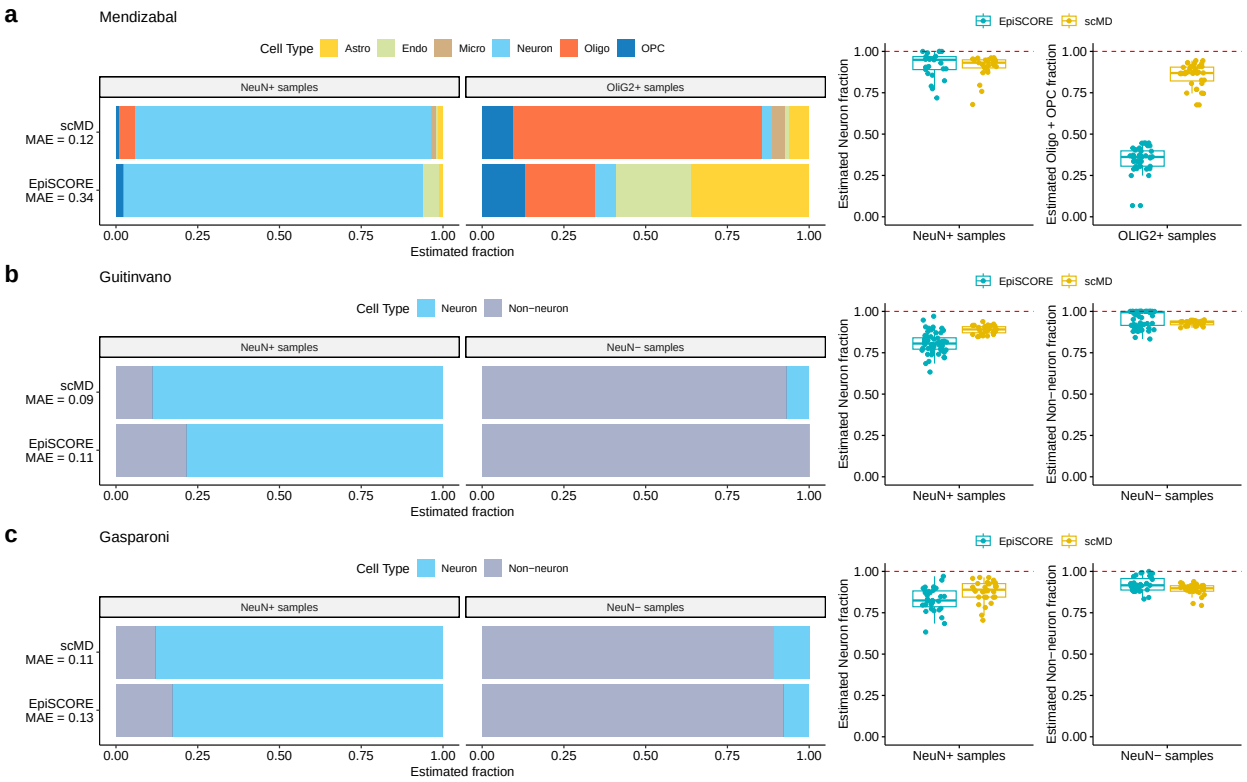

Figure 2: Validating cell-type DNAm signature from scDNAm data on sorted-cell data. **a**, Validation on Mendizabal *et al.*[17]. Bar plots of mean estimated cellular fractions across NeuN+ and OLIG2+ samples using scMD and EpiSCORE. Different cell types are annotated with different colors. Box plots of cellular fractions in sorted NeuN+ and OLIG2+ samples are shown on the right. Different colors represent different methods. **b**, Validation on Guintivano *et al.*[18]. **c**, Validation on Gasparoni *et al.*[19]. Note that for benchmarking, we aggregated the fraction estimates of cell subtypes to generate the fractions of broader cell types.

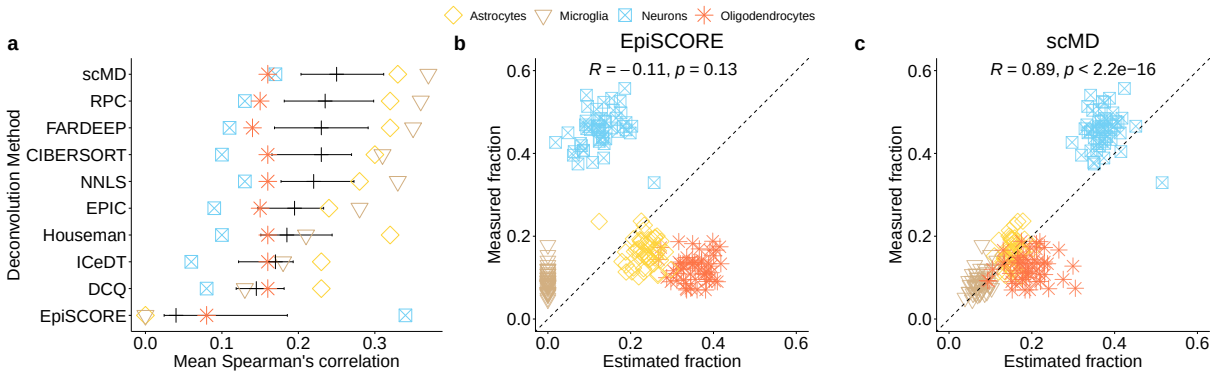

Figure 3: **a**, Benchmarking of scMD and other deconvolution methods on ROS data. Except for EpiSCORE which uses its RNA-derived reference, all other methods use our derived scDNAm references. For scMD, each dot denotes one correlation for each cell type. For other methods, each dot represents the average of Spearman's correlation across scenarios in each cell type. A scenario is defined as a particular setting with a specific deconvolution method and reference dataset. The black vertical line shows the mean of the average Spearman's correlation across scenarios, and the horizontal lines present means ± standard error of the mean. Scatterplots (b) and (c) illustrate the relationship between the estimated fractions of EpiSCORE and scMD (x-axis) against the corresponding fractions measured using immunohistochemistry (IHC) in ROS data (y-axis). Note that for benchmarking, we aggregated scMD's fraction estimates of neuronal subtypes to generate the neuronal fractions.

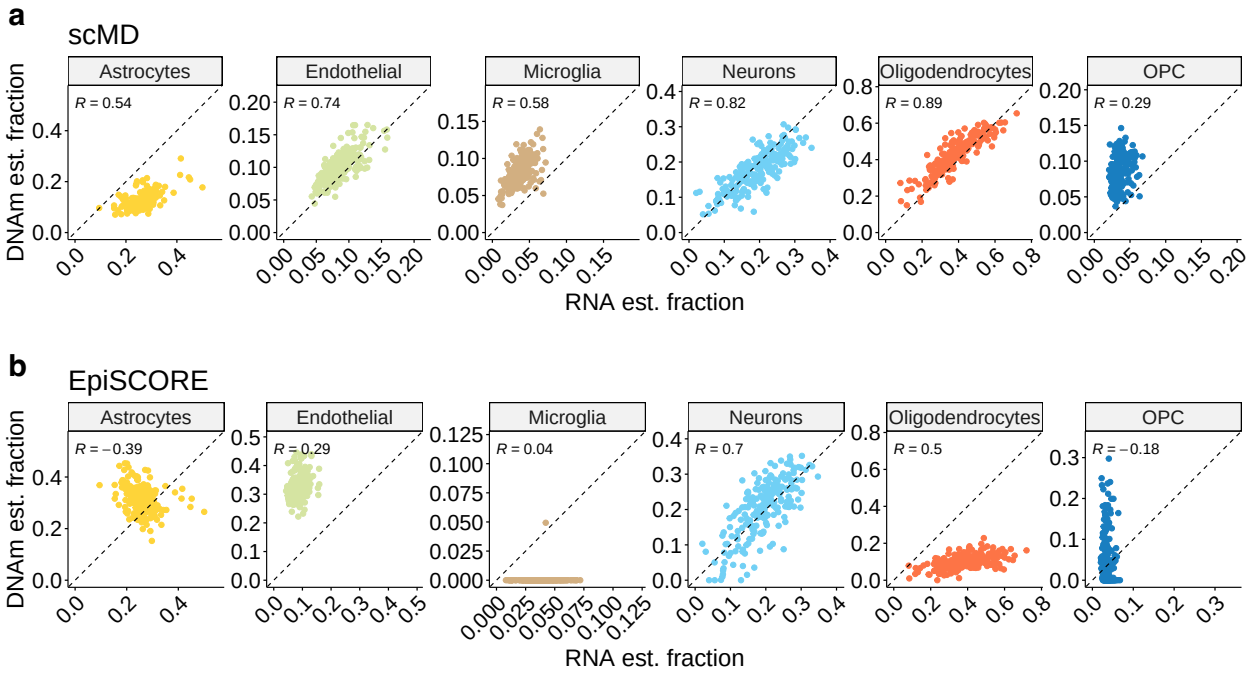

Figure 4: Deconvolution of bulk NAc data using scMD **(a)** and EpiSCORE **(b)** into six cell types. Scatter plots comparing the estimated fraction (est. fraction) obtained for each sample with RNA data (x-axis) using EnsDeconv vs. DNAm data (y-axis).

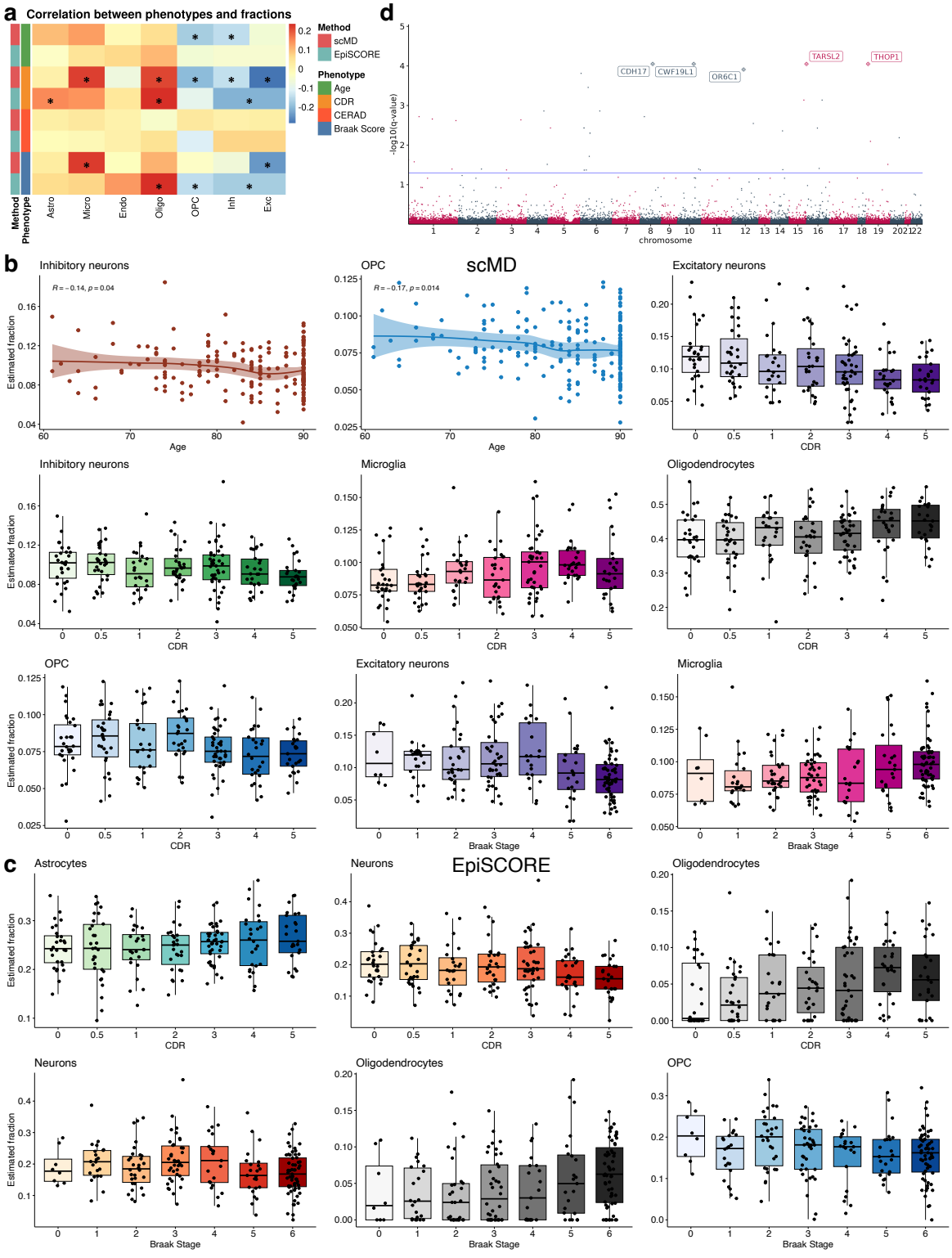

Figure 5: Identifying differential cellular fractions and methylated cytosines with the Mount Sinai Brain Bank (MSBB) data. **a**, Correlation between cellular fractions and age and AD phenotypes. * p-value < 0.05. **b**, scMD identified pairs of phenotypes and cellular fractions. **c**, EpiSCORE identified pairs of phenotypes and cellular fractions. **d**, Differentially methylated cytosines in microglia associated with aging using scMD estimated cellular fractions.