MSA: scalable DNA methylation screening BeadChip for high-throughput trait association studies

David C Goldberg[1], Cameron Cloud[1], Sol Moe Lee[1], Bret Barnes[2], Steven Gruber[2], Elliot Kim[1], Anita Pottekat[2], Max Westphal[2], Luana McAuliffe[2], Elisa Majournie[2], Manesh KalayilManian[2], Qingdi Zhu[2], Christine Tran[2], Mark Hansen[2], Jared B Parker[3], Rahul M Kohli[3], Rishi Porecha[2], Nicole Renke[2], Wanding Zhou[1,4,*]

[1]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, PA, 19104, USA, [2]Illumina, Inc. San Diego, CA 92122, USA, [3]Department of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA, [4]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

*To whom correspondence should be addressed. Tel: +1 215 590 0473; Email: wanding.zhou@pennmedicine.upenn.edu

SUMMARY

The Infinium DNA Methylation BeadChips have significantly contributed to population-scale epigenetics research by enabling epigenome-wide trait association discoveries. Here, we design, describe, and experimentally verify a new iteration of this technology, the Methylation Screening Array (MSA), to focus on human trait screening and discovery. This array utilizes extensive data from previous Infinium platform-based epigenome-wide association studies (EWAS). It incorporates knowledge from the latest single-cell and cell type-resolution whole genome methylome profiles. The MSA is engineered to achieve scalable screening of epigenetics-trait association in an ultra-high sample throughput. Our design encompassed diverse human trait associations, including those with genetic, cellular, environmental, and demographical variables and human diseases such as genetic, neurodegenerative, cardiovascular, infectious, and immune diseases. We comprehensively evaluated this array's reproducibility, accuracy, and capacity for cell-type deconvolution and supporting 5-hydroxymethylation profiling in diverse human tissues. Our first atlas data using this platform uncovered the complex chromatin and tissue contexts of DNA modification variations and genetic variants linked to human phenotypes.

Availability: Informatics for MSA data preprocessing and functional analysis is available in the R/Bioconductor package *SeSAMe* (version 3.22+): https://bioconductor.org/packages/release/bioc/html/sesame.html

The complete MSA manifest, design criteria, technical, human trait, and functional annotations are available at https://zwdzwd.github.io/InfiniumAnnotation

The generated human cell line, primary tissue methylome profiles (N=420), and EM-seq data are available in the Gene Expression Omnibus with accession GSE264438 and GSE267407.

INTRODUCTION

The dynamic patterns of genome-wide cytosine modifications, including 5-methylcytosine and 5-hydroxymethylcytosine in genomic DNA, play a critical role in gene expression regulation[1], genome stability maintenance[2], and organismal development[3]. Through these roles, DNA methylation has been extensively associated with cellular and physiological human traits[4] and is increasingly utilized as a biomarker in translational research and clinical applications[5,6]. Notable examples include applying DNA methylation to classify cancer and rare diseases[7–10], liquid biopsy-based disease diagnosis[11], and assessing disease hazard through methylation risk scores[12] and forensic analysis[13]. Analysis of DNA methylation profiles is also crucial for elucidating gene transcription mechanisms[14], understanding cell identity maintenance[15], studying variations in cell composition[16], and investigating gene-environment interactions within populations[4].

Selecting the optimal DNA methylation assay technology hinges on the application context[17], challenged by the large size of the human genome, the complex methylation biology across genomic regions, and prevalent inter-cellular heterogeneity in tissues[18]. The most comprehensive DNA methylation profiling assay is single-cell whole genome methylation sequencing (scWGMS), which offers unparalleled detail by providing base-resolution data for individual cells[19]. However, the high costs and technical complexity of scWGMS limit its use primarily to the high-resolution reference epigenome mapping in a limited set of samples[20]. It is currently not economically feasible to implement scWGMS for population-wide studies. Consequently, alternative methodologies are more frequently used, trading off genome coverage, base resolution, or cell-type resolution to reduce costs and technical demands. These include methods for profiling bulk tissues[21] or FACS-purified cells (e.g., bulk deep WGBS)[22], targeted genome capture (e.g., RRBS[23]), and the use of data science techniques to interpret sparse data sets (e.g., low-pass sequencing[24]).

Among the various approaches, the Infinium DNA methylation BeadChip, which offers base resolution data with high quantitative precision, has become particularly prominent for large-scale methylation discovery and screening efforts due to its ease of experiment and data analysis[25]. This platform has been central to consortia such as The Cancer Genome Atlas (TCGA) and has amassed over 80,000 HM450 methylomes[26] and a comparable number of EPIC array methylation profiles in the Gene Expression Omnibus (GEO). While sequencing-based methods are preferred for case-specific and mechanistic studies, Infinium arrays are the platform of choice for population-level research, including meQTL studies[27,28], epigenetic risk scoring[29,30], and other epigenome-wide association studies (EWAS)[31,32]. This is partly due to the necessity in population studies to cover many samples with nuanced variations in methylation levels and, in others, to dissect multiple cohort covariates such as sex, age, genetic background, and tissue type. The technology's high-throughput sample processing platform has facilitated rapid clinical applications[33], proving successful in cancer diagnosis[10], cell-free liquid biopsies[34], and forensic investigations[13]. It has also been instrumental in creating the most extensive mammalian DNA methylome atlas to date[35–37].

While array technologies can rapidly generate simple-to-analyze and inexpensive data sets, they rely on static probe designs that fix the target CpG space to those selected during the array's development. This permits cross-study comparisons, but designs should be periodically updated to incorporate newly discovered, trait-linked CpGs from various sequencing modalities that cannot yet scale to the population level. Whole genome methylation sequencing of human cells and tissues has recently elucidated cell type methylation at unprecedented resolution[22]. This necessitates a new Infinium BeadChip that incorporates these recent discoveries and facilitates their analysis in large-scale human population studies and clinical settings[20,22,38,39].

Furthermore, most predictive models hinge on a small number of trait associates. For example, most epigenetic clock models used hundreds of CpGs and reached high prediction accuracy[40]. Minimalistic approaches were taken in epigenetic clock construction[41], cell type deconvolution, and cancer classification[42]. These observations suggest the feasibility of designing condensed arrays that can profile orders of magnitude more samples while retaining the predictive value of generated data.

To respond to this need, we present the rational, systematic design and first implementation of the Methylation Screening Array (MSA), the latest Infinium BeadChip iteration. Compared to previous Infinium BeadChips, MSA has concentrated its coverage on trait-associated methylation (5.6 trait associations per site vs. 2.2 in EPICv2, Methods) and cell-identity-associated methylation variations (3.7 cell signatures per site vs. 2.3 in EPICv2, with an additional 48 novel cell type contrasts that can be made). Half of the design targeted what was reported in prior EWAS studies. The other half leverages the latest single-cell and bulk whole genome methylation profiling efforts that deeply characterize diverse human cell types. This dual approach enables high-resolution cell-type deconvolution, supported by reference methylation panels and predictive models we have rigorously benchmarked in this study. Compared to the 8-sample plate design used in previous methylation arrays, MSA is built on a 48-sample EX methylation platform to achieve ultra-high sample throughput at a lower cost per sample while screening for more traits per probe site. Evaluation of the array's accuracy and reproducibility confirms its robustness for population-scale applications. We showed that the clock models originally developed on previous Infinium platforms are compatible with MSA data. Applying MSA to various human tissues has demonstrated its capacity to dissect cell composition and shed insight into mitotic aging and 5-hydroxymethylcytosine dynamics in heterogeneous tissue types.

RESULTS

Rational, systematic design of MSA

We designed the MSA array by compactly consolidating human trait-associated loci identified in previous EWAS studies and novel probe designs targeting diverse methylation biology (Figure 1A). After post-manufacture quality control, the MSA array contains 284,317 unique probe sets targeting 269,094 genomic loci. 145,318 loci overlap what is targeted by the EPICv2 platform (Figure S1A). More SNP-targeting probe sets and nearly as many CpH probes were incorporated relative to EPICv2 (Figure S1B). Human trait-associated methylations were identified by mining EWAS databases and literature, prioritizing the diversity of trait coverage and statistical significance (Methods). We broadly classified all EWAS hits into 16 trait groups (Figure 1B, S1C). As expected from the design, MSA is highly enriched by EWAS associations across human traits (Figure 1C), reflecting the platform's targeted design and compact size.

For new CpGs that previous Infinium platforms have not targeted, we leveraged existing WGBS data sets to identify CpGs associated with cell type, cis-regulatory elements, correlation with chromatin accessibility and gene expression, 5-hydroxymethylation and additional methylation features (Figure S1D, Methods). We emphasized high-confidence cell-type-specific methylation discriminants to facilitate the deconvolution of complex heterogeneous tissue types and the study of cell-specific processes. Using pseudo bulk and sorted methylomes from brain[43–45], pan tissue[22], and blood cells[46], we performed hierarchical, non-parametric analyses to identify CpG discriminants for the different cell types (Methods). These analyses identified thousands of hyper and hypomethylated signatures across hundreds of cell types (Figure S1E). Compared to EPICv2, MSA contains more markers per cell type comparison group despite the smaller size (Figure 1D). These differences are especially pronounced for rarer cell types or comparison

groups with relatively few designable genome-wide markers. For example, our analysis of WGBS data identified 34 high-quality markers of the SRGAP1 subtype of VIP interneurons derived from the caudal ganglionic eminence. We incorporated 31 markers onto MSA, whereas EPICv2 contains three (Figure 1D).

Like the EPICv2 array, the MSA design is highly enriched in the promoter, enhancer, and transcriptionally active regions. It is strongly depleted from quiescent, heterochromatic, and ZNF regions (as annotated by the full stack ChromHMM[47]) (Figure S1F, Table S1). The two platforms are less represented by open-sea CGI sites but have a higher proportion of cis-regulatory element coverage (as annotated by ENCODE[48]) (Figure S1G, Table S1). MSA has a slightly increased proportion of proximal (5.6% vs. 3.45%) and distal (16.2% vs. 10.1%) enhancer elements and marginally less coverage of CpG island (12.4% vs. 16.2%) sites compared to EPICv2. Compared to EPICv2, MSA CpH probes were designed by analyzing brain cell type-specific methylomes with more prevalent CpH methylation. The queried cytosines are more linked to brain and neuron functions, implicating genes critical for neuron development and synaptic signaling (Figure 1E).

Lastly, MSA contains at least one probe linked to 14,964 genes (overlapping or within 1500bp of the transcription start site), nearly as many as the larger EPICv2 array (Figure S1H). The 772 genes on EPICv2 but not MSA were enriched in olfactory receptors and highly polymorphic genes whose readings are often confounded by genetic polymorphism[49] (Figure S1I). In summary, the MSA assay targets human trait-associated methylations and novel sites where methylation is predicted to be dynamic, cell type-specific, and biologically relevant.

MSA is highly reproducible and accurate.

To assess the technical performance of the MSA BeadChip, we generated 146 methylomes for eight cell lines (GM12878, HCT116, HeLa, Jurkat, K562, LNCaP, MCF7, and Raji) with replicates. Probe success rates for most of these 146 methylomes surpassed 90% (Methods, Figure S1J). Probe detection rates were robust to 50 ng of input DNA but declined to <60% for three samples with ~30 ng of input DNA (Figure S1J).

For all cell lines, we observed high correlation coefficients between samples of the same line regardless of the laboratory of cell culture (Figure 1F). The correlations between different cell lines were significantly lower, reflecting the different cellular origin, ploidy, and epigenomic properties of the different lines. For the GM12878 and HCT116 cell lines, we generated technical replicate methylomes using the same DNA sample and computed F1 scores based on binarized methylation levels (Methods). The technical replicates had highly similar methylation profiles, with F1 scores of 0.99 and 0.95 for GM12878 and HCT116, respectively (Figure S1K).

Like the EPICv2 BeadChip, the MSA array includes replicate probe designs that target the same 122-mer genomic loci but may vary in the other design details[50]. The replicate designs have the same prefix but alternative suffixes that describe the chemistry and target strand specifications[49]. For each of the 8,523 replicate probe groups, we calculated the standard deviation (SD) of replicate probes within cell line samples and compared the means of these SDs to the SDs of non-replicate probes (Figure S1L). Replicate probes had a low mean standard deviation of 0.02 compared to non-replicate probes, suggesting that the replicate probes can produce consistent methylation measurements.

Next, we evaluated the accuracy of MSA by comparing MSA beta values with methylation titration standards. For each titration, the beta value distributions centered on the target titration level (Figure 1G). We additionally compared our cell line methylomes from MSA to methylomes of the same DNA samples generated using an EM sequencing protocol[51,52] (Figure 1H). Beta values were again highly correlated within but not across cell lines. Lastly, we compared the

MSA cell line methylomes to publicly available WGBS data of the same cell lines and observed higher intra-cell line correlations than between cell lines (Figure S1M). These experiments confirm that MSA measurements are accurate and consistent with ground truth titrations and WGBS data.

MSA uncovers tissue-specific methylation biology.

We generated methylomes for five different sorted immune cell types (CD4 T, CD8 T, Total T cells, NK cells, Monocytes), peripheral blood mononuclear cells (PBMCs), and six different human tissue types (skin, pancreas, heart, colon, cerebellum, motor cortex). We performed unsupervised clustering using principal component analysis (PCA) to explore their global methylome similarities. Related cell and tissue types were highly colocalized (Figure 2A).

Cell type proportions are often the main drivers of bulk tissue EWAS results[53]. Using reference-based deconvolution, we tested whether our MSA tissue methylomes could be resolved into their constituent cell types (Methods). We first constructed a reference by overlapping the MSA probes with cell-specific differentially methylated regions (DMRs) previously identified[22]. The sorted immune cells were predicted to contain >90% of the matching cell type, consistent with typical yields from the standard purification protocol (Methods, Figure S2A). The cell proportion estimates of bulk tissues aligned with the reported tissue types (Figure 2B). Skin samples were largely composed of epidermal keratinocytes, while the heart sample was predicted to contain cardiomyocytes, heart fibroblasts, and endothelial cells. Colon samples were predicted to contain a high proportion of colon epithelial cells. However, two samples were predicted as small intestine epithelial cells and treated as small intestine samples thereafter. While most pancreatic tissues were estimated as acinar and ductal cells, the most populous cell types of the organ[54], one sample had a higher fraction of endothelial cells (denoted as PS1 in Figure 2B), and another had a higher fraction of immune cells (denoted as PS2 in Figure 2B). PS2 is clustered closer to the sorted immune cells from the PCA analysis (Figure 3A), suggesting excessive blood contamination, and was excluded from downstream analysis.

Next, we performed one-vs-all non-parametric supervised analyses of the tissues (Methods) and identified thousands of CpG discriminants uniquely methylated in the target tissue type (Figure 2C). Most CpG signatures were hypomethylated compared to the remaining tissues (Figure S2B). These tissue-specific probe sets were highly enriched in the cell-specific CpG signature lists curated by analyzing the publicly available single and sorted cell data sets during array design (Figure S2C, Methods), validating the design process and the performance of the selected probes in discriminating the target cell types.

To explore the role of tissue-specific methylation markers in the corresponding tissue biology, we analyzed the chromatin state distributions and gene linkages of the CpG sets. We first compared them with the full stack ChromHMM states, a universal genome annotation learned from over 1,000 data sets comprising diverse cell types[47] (Figure 2D). Hypermethylated tissue signatures were generally absent from enhancers and were enriched in promoter and bivalent promoter states, while hypomethylated markers were enriched in enhancers and gene bodies. The signatures are strongly enriched in the chromatin state associated with the matching cell type. For example, cerebellum and motor cortex signatures are enriched in EnhA6, representing brain enhancers. In contrast, colon and small intestine signatures were strongly enriched in EnhA14/A15, liver/digestive/intestine enhancers (liver not profiled). The EnhA7 state is annotated to blood enhancers, and the monocyte, NK cell, CD4+, and CD8+ T cell signatures were specifically enriched in this state. In addition to tissue-specific chromatin states, the signatures colocalized with the corresponding tissue-specific transcription factor binding sites (Figure S2D, Methods). Lastly, many signature CpGs were significantly enriched in marker genes for the corresponding cell types (Figure 2E, Methods). Collectively, our MSA data

uncovered the epigenome signatures at tissue-specific transcription factor binding sites and genes that regulate the corresponding tissue biology.

MSA reveals dynamic 5-hydroxymethylation biology in human tissues.

The standard array preparation uses bisulfite conversion, which does not discriminate 5-methylcytosine (5mC) from 5-hydroxymethylation (5hmC)[55]. To test if MSA is compatible with 5hmC profiling, we employed a modified ACE seq protocol in a subset of the human tissues[56] (Methods). The derived 5hmC levels were globally anti-correlated with the proliferation rate of the tissue, being most abundant in neuron-enriched brain tissues, followed by the heart and pancreas, and lowest in the skin, colon, and small intestine (Figure S3A). Across chromatin states, 5hmC levels peaked in H3K36me3/H3K79me2 marked gene body enhancers and actively transcribed states (Figure 3A). Meta gene analysis showed a rapid depletion of 5hmC levels near the TSS, which rebounded and peaked in gene bodies (Figure S3B). To validate 5hmC measurements, we compared probe sets selected for tissue-specific 5hmC levels using 5hmC-Seal[39] and hmC-CATCH[38]. While brain tissues had high 5hmC levels across most design groups, the non-brain tissues had the highest 5hmC in the designed tissue groups (Figure 3B).

Next, we expanded the tissue-specific 5hmC analysis to all probes. tSNE analysis of 5hmC profiles showed a separation according to tissue type (Figure 3C). Supervised analysis (Methods) identified tissue-specific 5hmC sites in the brain (N=56,354 cerebellum, N=75,241 motor cortex), heart (N=1,168), and pancreas (N=235) (Figure 3D). There were relatively few markers for colon (N=2) and small intestine (N=5), consistent with the low global levels of 5hmC in these tissues. Intriguingly, tissue-specific 5hmCs are also enriched in tissue-specific gain identified from bisulfite conversion-based data (Figure 3E), suggesting that a subset of tissue-specific hypermethylated CpG sites in bisulfite-conversion-based analysis are likely sites of 5hmC accumulation. Future studies may elucidate the unique role of 5hmC in establishing and maintaining cellular identity compared to 5mC and tissue-specific unmodified cytosine.

We identified the genes over which the tissue-specific 5hmC markers were localized. Interestingly, some genes contained markers for multiple tissue types, and these markers had unique distributions along the gene according to the tissue. For example, our 5hmC analyses found markers for the cerebellum, motor cortex, heart, and pancreas in *NCOR2*, a broadly expressed transcriptional repressor. While much of the gene had 5hmC in the cerebellum and motor cortex, distinct areas had specific accumulation or depletion of 5hmC in the pancreas and heart, reflecting an enhancer origin (Figure S3C).

MSA predicts age and mitotic history using epigenetic clocks.

DNA methylation can be used to infer chronological age and age-related abnormalities[57], and whether diseases of aging or environmental exposures are associated with accelerated epigenetic aging is an active area of DNA methylation research[40]. We sought to include common epigenetic clock probes in the MSA design. While a small percentage was lost due to manufacturing dropout, the twelve targeted epigenetic clocks have a minimum of 95% coverage (Figure 4A). We computed epigenetic age estimates on 730 normal tissues (Methods) using the full set of clock probes and the subset of probes on the MSA array to test whether the small fraction of lost probes would impact predictions. For the clocks tested, we observed high correlations and negligible differences in absolute prediction error between the full clock probe sets and the MSA representations (Figure 4B, S4A).

We also computed epigenetic clock estimates on the tissue and sorted immune cell methylomes that we generated. The multi-tissue Horvath clock predictions were highly correlated with the chronological age of the samples (Pearson's r=0.87, Figure 4C). The predictions were as good or better than the full Horvath clock applied to normal tissues HM450 data above (Pearson's

r=0.78, Figure S4B). When computed on all samples, the Horvath skin clock[58] predictions were slightly less correlated with chronological age (r=0.57). However, when analyzing only skin and blood, the tissues on which the clock was trained, this correlation increased to 0.93. Colon tissues with a high proportion of epithelial and blood cell types[59] are also predicted with high correlation. The Hannum and Levine clocks measure the chronological and phenotypic age, respectively[60,61]. We did not have phenotypic measurements for our tissues to infer these metrics, but the predicted ages correlated with chronological age, as previously reported. The four clocks tested on our samples contained replicate probe designs. We found that predictions did not change whether a single design was randomly selected for prediction or whether the methylation was averaged over all replicates (Figure S4C). This robustness is due to the low standard deviation in beta values between replicate probe designs within samples (Figure S4D).

Lastly, we analyzed the mitotic histories of the different tissue methylomes using a subset of PRC2 target CpGs[62] and partially methylated domains (PMDs) to track the cumulative cell divisions of the tissue. Applying the models to our tissue and immune cell methylomes yielded division rates consistent with the relative proliferative activity of these tissues reported in the literature based on radioisotope labeling[63]. For example, the colon, small intestine, and T cells had the highest division rate score, consistent with the high cellular turnover of these tissues (Figure 4D). In contrast, tissues with higher fractions of post-mitotic cell types, such as the motor cortex, cerebellum, and heart, had the lowest division rates. The estimates of mitotic activity using the PMD methylations largely correlated with those obtained from the PRC2 model. Interestingly, pancreatic tissues showed relatively low PMD methylation compared to other tissues and predictions based on average PRC2 target methylations. These effects were not fully explained by global methylation differences, which were minor for tissues of similar mitotic activity based on the EpiTOC2 model (Figure S4E). Acinar cell PMDs in sorted WGBS data[22] were similarly hypomethylated relative to other cell types (Figure S4F). The physiological cause or consequence of this PMD hypomethylation in acinar cell biology warrants further investigation.

MSA methylomes reveal strong tissue contexts of human trait associations.

Leveraging the trait association focus of MSA, we evaluated the capacity of MSA data to perform functional annotation of EWAS hits. In this analysis, we focused on the tissue context using the primary tissue profiles produced in this study. We first note that for the traits investigated in the curated studies, trait-associated probes are more often significantly enriched in enhancers and promoters[47] but underrepresented in heterochromatic and repressive genomes (Figure S5A), consistent with their roles in transcriptional regulation. Traits characterized by genomic alterations (e.g., Down's syndrome), cell proliferation (e.g., malignancy), and frequent toxin exposure (e.g., smoking) had distinct and recurring chromatin feature enrichment (Figure 5A). In contrast, complex disease traits, e.g., diabetes and Alzheimer's disease, are varied in chromatin state enrichment across studies.

As expected, the enhancer and promoter-associated probes are more variably methylated across primary human tissue types (Figure S5B). To test whether such variation reveals the tissue context of each trait, we grouped CpGs by their associated traits and compared the methylation levels across tissue types (Figure 5B). An intriguing correspondence between the perceived tissue context and the methylation rank emerged. For example, CpGs associated with Alzheimer's disease showed the most extreme methylation in brain tissues compared to other tissue types (Figure 5B). Sites with a putative positive disease effect size have the highest methylation readings in the brain, whereas sites with reduced methylation in diseases were least methylated in brain tissues. Similarly, probes associated with irritable bowel syndrome (IBS) were most methylated in the colon and small intestinal tissues. These results suggest a propensity of trait-associated CpGs to colocalize with differential methylations specific to the

tissue that manifest the trait phenotype, underscoring the importance of tissue context when conducting EWASs.

We also investigated the extent to which GWAS variants colocalize with tissue-specific methylation. We tested the enrichment of trait-associated SNPs in the one-vs-all cell-specific methylation signatures on MSA (Methods). These analyses identified multiple genetic variants associated with a tissue-specific trait co-localizing with the methylation signature of the corresponding tissue type. For example, SNPs associated with blood glucose and diabetes were colocalized with methylation markers for pancreatic cell types, while cholesterol variants were localized to hepatocyte-specific methylations (Figure 5C). Diverse autoimmune disorders were enriched in CpG markers for regulatory T cells, which are involved in immune system homeostasis and autoimmune suppression[64]. Whether the genetic variants implicated in these diseases directly impact nearby tissue-specific methylation to perturb gene expression and function requires follow-up studies.

DISCUSSION

The Infinium DNA methylation BeadChip is a broadly used and accessible assay in human population studies. It has enabled trait association discoveries and predictive models such as epigenetic clocks, risk scores, and disease classifiers. Previous Infinium BeadChips have been designed to target genomic features, such as gene promoters, gene bodies, and cis-regulatory elements. While methylation variation at these genomic features is indeed associated with human traits, evenly covering genomic elements is not as economical for trait screening applications as in discovery and hypothesis generation settings.

The existing methylation-based screening of most human traits requires relatively few loci. For instance, the Horvath clock for chronological age used 353 CpGs[57]. Other epigenetic clocks use feature numbers ranging from a few CpGs to ten thousand CpGs[60], much smaller in number than existing Infinium array capacities[65]. The feasibility of such minimalistic approaches has also been established in cancer classification[42] and cell type deconvolutions[66] and demonstrates high inference precision. The development of MSA can be seen as a balanced approach to DNA methylome-based trait screening for the benefit of profiling larger human populations.

Besides offering a balanced approach in trait screening, MSA also represents an upgrade of Infinium array content to bridge deep high cell-type resolution profiling and cost-effective population screening. While offering greater cell type variation and genome-wide details, single-cell methylome profiling cannot be scaled to population settings. MSA is designed to translate the cell type-specific knowledge from single-cell and bulk whole-genome methylome profiles for use in the population setting.

Computational cell-type deconvolutions are powerful methods for interrogating tissue composition variation in development and disease. The expanded cell-specific CpG markers and refined annotation in MSA enhance deconvolution granularity compared to EWAS studies based on previous Infinium platforms. For example, the commonly used CETS algorithm for estimating brain cell proportions estimates NeuN+:NeuN- proportions without predicting trait-relevant subtypes[67]. We designed cell-specific probes discriminating 174 unique cell types (82 brain cell types, 51 pan tissue, 41 blood) and anticipate that these markers will enable high-resolution deconvolution, augmenting the study of selectively vulnerable or rare cell populations in complex diseases and tissue types. Our results and other recent work have identified an enrichment of genetic variants associated with complex traits within cell-specific DMRs[20]. It is not clear the extent to which methylation changes in these cell-specific DMRs may perturb the

functioning of the disease-relevant cell types. We anticipate that MSA will permit such investigations.

Previous efforts have established the compatibility of Infinium arrays with other base conversion protocols, such as Tet-assisted bisulfite conversion, to profile 5hmC modifications[68,69]. Our analysis suggested that the new MSA array is compatible with the tandem bisulfite-A3A conversion for 5hmC profiling. We applied the 5hmC profiling to neuronal and peripheral human tissues. The tissue-specificity mirrors previous sequencing-based 5hmC profiles, suggesting the feasibility of using methylation arrays to implement 5hmC profiling in large sample sets. Our data also underscores the high cell type specificity of 5hmC signals, which could be used to trace cell identity and tissue composition changes.

As a first application, our analysis was limited in validating trait-associated probes selected due to limited metadata availability. However, we found that probes associated with some traits in the literature were variably methylated in the corresponding tissue types we profiled or had a strong tissue context according to the beta value rank by tissue type (Figure 5). Attempting to design a consolidated array, we were also limited in the number of the CpG sites we could include and thus genomic feature and trait coverage. As more WGBS and array-based methylomes are generated, future designs may refine the most relevant trait and cell type-implicated CpG sites to maximize screening and discovery power most economically.

CONCLUSION

We systematically developed, benchmarked, and applied MSA, a novel Infinium BeadChip assay consolidating trait-associated probes from the extensive EWAS literature, single-cell and bulk whole genome methylome profiles. Our benchmark revealed MSA as an accurate, reproducible, scalable, next-generation Infinium human methylation BeadChips targeting trait discovery in population settings. Our first application uncovered the cell type context of human EWAS and GWAS discoveries and dynamic 5hmC association in peripheral tissues. We anticipate MSA to be a valuable tool for methylation screening in large human populations for trait associations and broadly dissecting the cell-type-specific mechanisms of human diseases.

METHODS

CpG Probe Selection

Probe designability We aligned unmethylated and methylated probe sequences to the hg38 genome using the BISCUIT tool suite[70]. To identify uniquely mapping sequences, subsequences of 30,35,40 and the entire 50nt probe sequence were aligned, and only probe designs where all subsequences had mapping quality >20 for both the methylated and unmethylated allele were considered. For these 19,253,974 uniquely mapping CpGs, design scores reflecting hybridization efficiency and melting temperature were computed, and 13,891,035 CpGs with design scores > .3 were retained. Any probe sequence that contained common SNPs (dbSNP Build 151)[71] within 5nt of the 3' end was removed. Sequences with more than six additional CpGs were also removed to prevent hybridization interference due to variable methylation of neighboring CpGs. 9,993,793 CpGs remained from this preprocessing ("Designable Probes"), from which all array content was subsequently selected. When possible, high-quality probes (design score >= .6) were prioritized.

Cis-regulatory elements: Human GRCh38 candidate cis-regulatory element (CRE) annotations were downloaded from the ENCODE Project Consortium[72] and intersected with designable CpG sites. The methylation range for each CpG was computed across sorted

immune[46] and pan tissue[22] cell types. CpGs that did not show a range > .4 were filtered out. The remaining CpGs were grouped by CRE type and sorted by methylation range. 30,000 CpGs total were sampled with a bias toward enhancer elements (dELS: 64%; pELS: 21%; CTCF Only, CTCF-bound:11%; PLS:2%; DNAse-H3K4me3:2%).

Monoallelic/intermediate methylation: 180 bulk adult normal WGBS samples (Table S2) were analyzed to identify candidate monoallelically methylated CpG sites. Autosomal CpGs with minimum coverage of 20 reads and mean methylation >.3 and <.7 across 140 of the 180 samples were considered intermediate methylation and intersected with the designable probe list. 207 pan-tissue sorted cell WGBS methylomes from Loyfer et al [22] were also analyzed for intermediate methylation, and designable CpGs with mean methylation >.3 and <.7 across 180 of the 207 samples were selected.

XCI-linked CpGs: 76 high coverage (>20 million CpGs) normal female WGBS samples (Table S2) were analyzed to identify X-chromosome CpG sites with intermediate methylation across samples (0.3 < methylation < .7). An additional 95 normal male WGBS samples were analyzed to identify X chromosome CpG sites fully unmethylated (< .3 methylation across 50 samples) or fully methylated (>.7). The CpG sites intermediately methylated in female samples but unmethylated or fully methylated in male samples were intersected with the high-quality probe list.

Cell type-specific methylation: BED/bigWig files for single cell brain[43–45], sorted pan tissue[22], and sorted immune cell WGBS data[46] were downloaded and used for marker identification. To reduce the sparsity of single-cell brain data, pseudo bulk methylomes were generated by averaging methylation over the cell type labels obtained by unsupervised clustering analysis previously reported. One vs. all comparisons were performed across major cell type groups and hierarchically within major groups to identify subtype markers. Wilcoxon rank sum testing was performed between the target and out groups at each CpG site to identify cell-specific markers. Designable CpG sites with an AUC = 1 and a delta beta >= .3 between the in and out groups were selected, and markers were capped at 80 CpGs per cell type contrast. Hyper and hypomethylated signatures were balanced when possible.

5hmC: 5hmC-Seal[39] and hmC-CATCH[38] 5hmC peaks were downloaded. Genomic intervals were intersected with the designable CpG list. For 5hmC-Seal data, the 5hmC CpG signal was treated as a binary value (1 if within a significant peak, 0 if not). For hmc-CATCH data, the peak coverage was applied to CpGs within the peak, and samples were scaled according to the total coverage. Tissue-specific 5hmC sites were identified as previously described for the WGBS data. To identify 5hmC sites along a continuum of tissue specificity, the top 10K most highly covered CpGs in each sample from the hmC-CATCH data[38] were collected and binned according to the frequencies the CpG was in the top 10K across the 60 samples. 11 bins of 5 tissue count intervals (e.g., 1-5, 6-10, …, 55-60 tissues) were sampled equally, with sampling capped at 200 CpGs per bin.

Cell-specific CpH methylation: Genes with cell-specific mCH methylation were downloaded[43], and the top ten genes with the highest AUROC were selected for each cell type. Gene coordinates were intersected with CAC cytosines, the most prevalent mCH context found in neurons. 20 cytosines were sampled from each gene for each cell type.

DNA methylation-gene expression correlations: Matched WGBS / Gene expression data from the Roadmap Epigenomics Mapping Consortium were used to compute the Spearman correlation between CpGs in the high-quality designability list and genes within 10KB of the CpG. CpGs were then ranked by the *P*-value of the correlation, standard deviation and expression levels of the gene, and absolute value of the correlation. The top 2,500 CpGs negatively correlated with the expression of the linked gene, and the top 2,500 positively

correlated CpGs were selected. TCGA normal tissues[73] were also analyzed to identify correlated linked CpG-Gene pairs. CpGs with a correlation >= 0.6 or <= -0.7 and a *P*-value < .05 were additionally included (901 positively correlated, 1,620 negatively correlated).

DNA methylation-chromatin accessibility correlations: Matched DNA-chromatin accessibility data were downloaded from Luo *et al.* 2022[43], and Spearman correlations were computed between the accessibility peaks and CpG methylation sites. Correlations with *P*-values < .05 and |Spearman's rho| > .5 were selected, and the CpGs intersected with the high-quality designability list.

CoRSIVs: Genomic coordinates for CoRSIVs were downloaded[74,75] and intersected with high-quality designable probes.

Solo-WCGW in partially methylated domains: CpGs in the WCGW context (flanked by A or T) in common PMDs were downloaded from Zhou *et al.* 2018[21] and intersected with high-quality designable probes. This subset was further intersected with CpG islands, and 6,000 probes were randomly sampled.

meQTLs: meQTL data was downloaded from the GoDMC database[28], and CpGs were ranked according to the number of times a CpG was associated with a meQTL. The top 10K CpGs were selected. An additional 20K meQTLs were randomly sampled from Hawe *et al.* 2021[27].

Imprinting-associated DMRs: Differentially methylated regions associated with monoallelically expressed genes were downloaded from Skaar *et al.* 2012[76] and lifted to GRCh38 coordinates. The DMRs were intersected with the designable probes list.

Y-linked genes: 180 high coverage (>20 million CpGs) Human WGBS samples (Table S2) were analyzed to identify variably methylated Y-linked genes. The Y chromosome CpGs were intersected with designable probes and subsequently intersected with all Y chromosome genes (GENCODE V39). The variance across the 180 samples was computed at every remaining CpG site. For each gene, the top 20 most variable probes were selected.

Human trait associations: 1,067 EWAS studies were curated from the literature and EWAS databases (EWAS catalog[31], EWAS atlas[32]). A subset of high-priority studies was identified according to sample number and statistical significance, diversity of trait coverage, citation number, and the journal impact factor. All probes, or the top 2500 most significant probes, were selected from high-priority studies. The top 100 most significant probes were selected from all remaining curated studies. Study titles and trait annotations were queried for regular expressions to consolidate all selected studies/traits into 16 major trait groups.

Sample Preparation

Tissue dissection: Fresh frozen tissue samples were obtained from the Cooperative Human Tissue Network (CHTN), and 30-50mg of tissue were dissected on dry ice.

Cell line culture: GM12878, K562 (CCL-243), LNCaP (CRL-1740), and HCT116 (CCL-247) cells (Source 1) were obtained from American Type Culture Collection (ATCC, Manassas, VA, USA). 1-4 x 10^6 cells were plated and cultured for 6 days with fresh media added 2-3 days. K562 cells were cultured in Iscove's Modified Dulbecco's Medium (30-2005, ATCC), 10% Fetal Bovine Serum (FBS) (45000-736, Gibco), and 1% penicillin/streptomycin (15140122, Gibco). LNCaP cells were cultured in Roswell Park Memorial Institute Medium (RPMI-1640) (30-2001, ATCC), 10% FBS, and 1% penicillin/streptomycin (15140122, Gibco). GM12878 cells were cultured with RPI-1640 (72400047, Invitrogen), and 15% Fetal Bovine Serum (Gibco, 45000-736), 1% GlutaMAX™ (Gibco, 35050061), and 1% penicillin/streptomycin (15140122, Gibco). HCT116 cells were cultured in McCoy's 5a medium modified (ATCC,30-2007), 10% Fetal Bovine Serum (FBS) (45000-736, Gibco), and 1% penicillin/streptomycin (15140122, Gibco). All

cells were maintained in a 37°C incubator with 5% CO2 and cultured at a 75 cm2 culture flask (Fisher, BD353136)

DNA extraction: Genomic DNA was extracted from 30-70 mg of tissue or 5.0 x 10^6 cells for Source 1 cell lines using commercially available QIAGEN QIAamp Mini Kit (QIAGEN, 51304), following the manufacturer's protocol. DNA was quantified using a Qubit 4 Fluorometer (Invitrogen). For Source 2 and Source 3 cell lines, genomic DNA was purchased from BioChain Institute (HeLa - #D1255811, Raji - #D1255840, Jurkat - #D1255815, MCF7 - #D1255830, K562 - #D1255820)

Immune cell purification: Sorted immune cells were purified by the Human Immunology Core at the University of Pennsylvania following STEMCELL Technologies RosetteSep Enrichment Cocktail protocols (https://cdn.stemcell.com/media/files/pis/10000000545-PIS_02.pdf). PBMCs were isolated using a Lymphoprep ficol layer.

Methylation titration controls: 10 ng of fully methylated human blood (Thermo Scientific, SD1131) and Jurkat (Thermo Scientific, SD1121) genomic DNA were amplified using the Repli-g Mini Kit (QIAGEN, 150023) according to the manufacturer's protocol. Following quantification with a Qubit 4 Fluorometer, 500ng of unamplified and amplified DNA were combined for the 50% control. Human pre-mixed calibration standards (0,5,10,25,50,75,100%) were purchased from EpigenDx (EpigenDx 80-8060H_PreMix), and 200ng / titration was used for testing.

EM sequencing of cell line DNA: Genomic DNA from the GM12878, K562, and HCT116 cell lines were extracted according to the QIAGEN QIAmp Mini Kit Protocol. The three samples were then mechanically sheared to 300 base pairs using the M220 Focused-ultrasonicator (Covaris, 500295) and methylated lambda control DNA. 200ng of each sample was enzymatically converted using the NEBNext® Enzymatic Methyl-seq Kit (NEB, E7120) with the manufacturer's protocol. The samples were then indexed during PCR amplification during PCR amplification using EM-Seq™ index primers (NEB 7140). The indexed libraries (200 ng each) were pooled and used as input for the Twist NGS Methylation Detection System for target enrichment. A pre-hybridization solution of blockers and enhancers was created to prepare the pool for hybridization (Twist Bioscience, 104180). The DNA was hybridized with the Twist Human Methylome Panel (Twist Bioscience, 105520**)**, and the targets were bound with streptavidin beads (Twist Bioscience, 100983), followed by a post-capture amplification. The enriched libraries were sequenced to 20X on the Illumina Novaseq 6000 PE150 platform.

5hmC profiling: Using the EZ DNA Methylation Kit (Zymo Research, D5001), 500 ng of each sample was bisulfite converted and purified following the manufacturer's protocol. The samples were then denatured with DMSO at 95°C for 5 minutes and snap-cooled on dry ice. The samples were deaminated using APOBEC3A (A3A) purified following previously published protocol[77] over 2 hours at 37°C. After incubation, the samples were purified using the Oligo Clean and Concentrator Kit (Zymo Research, D4060), following the manufacturer's protocol. Two cycles of whole genome amplification were performed using 50 U of Klenow Fragment (3'→5' exo-) (NEB, M0212M), dNTP solution mix (Bio-Rad, #1708874), and Random Primer 6 (NEB, S1230S). The samples were finally purified using AMPure XP Beads (Beckman Coulter Life Sciences, A63881).

MSA Data Analysis

Data preprocessing: All data preprocessing was done using the *SeSAMe* R package (version 1.22.0)[78]. A manifest address file was generated using the MSA manifest available at https://github.com/zhou-lab/InfiniumAnnotationV1/raw/main/Anno/MSA/MSA.hg38.manifest.tsv.gz and the *sesameAnno_buildAddressFile* function. Beta values were extracted from raw IDAT files using

the *openSesame* function with the built address file and default parameters. Probe detection rates were obtained using the *probeSuccessRate* argument with the *openSesame* function. One sample with probe detection rates < 0.7 was excluded from analyses.

Trait enrichment testing: 2,398,372 EWAS hits were curated from the literature and EWAS databases[31,32] and used as a background for enrichment testing. Traits were annotated to 16 major trait groups by searching for regular expression terms relevant to the trait group within the study or trait descriptions. The odds ratio enrichment in these trait groups was computed for 3 query sets: 1) EPICv2 probes, retained MSA probes from prior Infinium platforms, and a random set of probes equal in size to the retained MSA probes. log2 odds ratio was plotted for each platform across trait groups. For testing the enrichment of MSA and EPICv2 probes in total trait associated probes, all EWAS probes were rank-ordered according to how many traits the probes associated with. The MSA and EPICv2 probes were each tested as a query against the ranked probe list using a modified gene set enrichment approach[79] using the *knowYourCG* R package (version 1.0.0).

Gene linkage and ontology analysis: The MSA and EPICv2 manifests were downloaded (https://github.com/zhou-lab/InfiniumAnnotationV1/raw/main/Anno/), and probe coordinates expanded 1500bp upstream of the probe start site. The manifests were then intersected with GENCODE.v41 GTF files to identify linked genes. Gene ontology testing was performed for protein-coding genes using Enrichr[80]. For CpH probe-linked genes, only genes with a minimum of 2 probes per gene were analyzed.

Sample reproducibility and accuracy: Pearson correlation coefficients were computed across cell line samples. Correlation matrices were plotted in heatmaps. For pairwise replicate comparisons, beta values were first binarized as 1 if beta > 0.5 and 0 if beta < 0.5. F1 scores for the binarized vectors were computed using the MLmetrics package (1.1.3).

Cell deconvolution: Reference-based cellular deconvolution was performed using the EpiDISH R package[81] with the robust partial correlations (RPC) method. The centDHSbloodDMC.m matrix provided within the package was used as a reference for sorted immune cell deconvolution. Cell-specific unmethylated DMRs from Loyfer et al. 2023[22] were lifted over to the hg38 genome and intersected with the MSA manifest. The reference beta values used the table's target vs. background methylation values. Cellular proportion estimates were plotted in heat maps.

Identification of tissue-specific markers: One-vs-all tissue type comparisons were performed for sorted immune cells and bulk tissues. Wilcoxon rank sum testing between the target and out-group was performed at each CpG site. CpGs with >.1 missingness in the target group were excluded. The AUC for discriminating between the target and the out-groups was computed. Only CpGs with a delta beta >.3 were selected as cell markers. For visualization, the top 200 hypo and hypermethylated CpGs sorted by AUC and delta beta were selected for each tissue type. For 5hmC samples, the same analysis was performed, but only CpGs with zero missingness in the target group were considered, and a delta beta > .15 was used as a threshold to be considered a marker.

Tissue-specific CpG - transcription factor binding site analysis: BED files containing TFBS peaks were downloaded from ReMap 2022 (https://remap.univ-amu.fr[82]). The peaks for each transcription factor were intersected with all MSA CpGs to create CpG-TFBS links. Tissue signatures were tested for enrichment in the TFBS CpG sets using Fisher's exact test with all MSA probes as the background.

Tissue-specific CpG – marker gene enrichment testing: CpG sets for genes linked to hypo and hypermethylated tissue signatures were constructed using the *buildGeneDBs* function from

the knowYourCG R package (version 1.0.0) and the MSA manifest (https://zwdzwd.github.io/InfiniumAnnotation) with a max distance of 10KB. Fisher's exact testing was performed to calculate the enrichment of tissue-specific CpGs in the CpGs associated with linked gene sets. For each tissue signature set, the top 20 most significant gene enrichments were overlapped with all cell type marker genes from the CellMarker 2.0 database (https://academic.oup.com/nar/article/51/D1/D870/6775381[83]) and the top 200 highest overlap pairs were selected for visualization.

Epigenetic clocks: Using default parameters, the *DNAmAge* function from the *methylclock* package (version 1.8.0)[84] was used to estimate the epigenetic age of 730 TCGA normal tissues. The probe IDs from this beta value matrix were then intersected with the list of unique MSA probe IDs (after removing probe suffixes) to get only the probes found on the MSA platform. The epigenetic clock predictions were computed again. For tissues profiled on MSA, probe ID suffixes were removed, a single duplicate probe ID was selected, methylation was averaged over all duplicate probe IDs, and the betas were tested with the same *DNAmAge*() function with default parameters. HypoClock and EpiTOC2 estimates were computed by tissue type group using the data and code provided by the authors at https://zenodo.org/records/2632938.

Analysis of EWAS hit chromatin state contexts: Each set of EWAS trait probes in the curated studies was tested for enrichment in 100 full-stack ChromHMM chromatin states [47] using Fisher's exact test. The total pool of curated EWAS hits was used as a background set. The number of traits-chromatin state associations with FDR < .05 was computed for each chromatin state and plotted. 17 studies representing 6 trait groups were selected, and the enrichment across chromatin states was plotted in heat maps.

Chromatin context analysis of EWAS methylations: The standard deviation of all probes was computed using the tissue methylomes generated on MSA and sorted to create a ranked probe list. Selected full-stack ChromHMM states were intersected with the list of total EWAS hits and tested as queries against the ranked probe list using a modified gene set enrichment approach[79] using the *knowYourCG* R package (version 1.0.0).

Tissue context analysis of EWAS methylations: For each set of EWAS trait probes in the curated studies, we computed the standard deviation of the probes using the tissue methylomes we generated using MSA. Trait sets were sorted according to the average standard deviations, and the most variable traits were selected for further analysis. In these traits, the rank for each sample was computed according to beta values. The mean rank of each tissue type group was computed for every CpG in the trait, and the distributions of ranks for each tissue type were plotted.

GWAS co-localization with tissue-specific methylations: GWAS summary statistics were downloaded from the NHGRI-EBI GWAS catalog[85] (version 1.0.2.1). The top 3000 unique disease/trait categories with the most SNPs were grouped and tested as independent queries against each one-vs-all tissue/cell-specific CpG set from the curated lists incorporated into the final MSA design. SNPs and CpG sites were expanded by 5kbps in upstream and downstream directions, and genomic interval overlaps were computed using the *IRanges* package (version 2.36.0). The total number of CpG intervals for all tissue signatures was used as a background set, and Fisher's Exact test was performed for enrichment testing.

DECLARATION OF INTERESTS

W.Z. received MSA BeadChips from Illumina Inc. for research. B.B., S.G., A.P., A.W., E.M., M.K., M.P., Q.Z., M.H., R.P., and N.R. are Illumina employees. United States Patent Application

No. 63/596,091 has been submitted and covers the methods and findings discussed in this research.

FUNDING

National Institute of Health [R35-GM146978 to W.Z.] [R01-HG010646 to R.M.K]

ACKNOWLEDGEMENTS

The authors thank Lynn Chen, Max Eldabbas, and Emileigh Maddox of the Human Immunology Core at the Perelman School of Medicine at the University of Pennsylvania for assistance with immune cell purification. The HIC is supported in part by NIH P30 AI045008 and P30 CA016520. HIC RRID: SCR 022380. The authors thank the NCI Cooperative Human Tissue Network (CHTN) for providing human tissue samples. Other investigators may have received specimens from the same tissue specimens. RRID: SCR_004446.

AUTHOR CONTRIBUTIONS

Conceptualization, W.Z., N.R., R.P., and D.C.G.; Methodology, D.C.G., W.Z.; Formal Analysis, D.C.G., and W.Z.; Investigation, D.C.G., C.C, E.K., S.L., M.H., M.K. S.G., A.P., B.B., M.W, and L.M.; Resources, R.K., J.B.P., M.H., M.K., Q.Z., S.G. and A.P.; Writing – Original Draft, D.C.G. and W.Z.; Writing – Review & Editing, D.C.G., W.Z., NR, RP and RMK, Funding Acquisition, N.R., R.P. and W.Z.; Supervision, W.Z.

FIGURES LEGENDS

Figure 1: MSA design workflow and major trait groups. (A) The screening process is to identify designable probes (left), targeted trait groups, and methylation features (right). (B) Major trait categories are incorporated into array content with a subset of represented sub-trait groups. Two trait categories (Sex and Other) are omitted. (C) Set enrichment analysis showing the running sum statistics of MSA and EPICv2 probes down a ranked list of EWAS hits probes, ranked according to the number of trait associations (top) and the enrichment of retained sites on MSA in all annotated major trait groups compared with EPICv2 and a random selection of Infinium probes (bottom). (D) Number of CpGs per cell type contrast on MSA vs. EPICv2 for contrasts with few (<500) high-quality whole genome markers. (E) Gene ontology for biological process results for genes linked to CpH probes (minimum 2 probes per gene) on MSA and EPICv2. (F) Heatmap of beta value correlations between cell line samples profiled on MSA. (G) Density plots of beta values for methylation titration standards. (H) Heatmap of beta value correlations between cell line samples profiled on MSA and with an EM sequencing protocol.

Figure 2: MSA reveals tissue-specific methylation biology and tissue compositions. (A) PCA plot showing unsupervised clustering of sorted immune cells and bulk tissues profiled on MSA. PS1 and PS2 refer to pancreas tissue samples 207854510042_R07C03 and 207854510042_R09C03, respectively. (B) Heatmap showing cell type proportion estimates obtained by methylation-based deconvolution on the sorted immune cell and bulk tissues (columns) profiled on MSA (C) Heatmaps showing beta values of tissue-specific CpGs (rows) over bulk and sorted immune cells (columns). (D) Enrichment of hyper and hypomethylated tissue-specific CpGs in different full-stack ChromHMM chromatin states (FDR < .05) (E)

Heatmap showing the overlap of genes linked to tissue-specific CpGs identified from MSA profiled tissues (rows) with marker genes for similar cell types based on publicly available scRNA databases (columns).

Figure 3: 5hmC analysis of human tissues with MSA. (A) Dot plot showing mean 5hmC levels across full stack ChromHMM states for each sample, colored by tissue type. (B) Heatmap showing mean 5hmC levels for each sample (rows) across the 5hmC tissue-specific design groups created during array development (columns) (C) tSNE plot showing unsupervised clustering of bulk tissues profiled for 5hmC. PS1 and PS2 refer to pancreas tissue samples 207854510042_R07C03 and 207854510042_R09C03, respectively. (D) Heatmap showing representative one vs. all 5hmC signatures (E) Heatmap showing the enrichment of 5hmC signatures in tissue-specific hypermodified (5mC + 5hmC) signatures curated during array development from WGBS data

Figure 4: Epigenetic aging and mitotic history analysis of MSA profiled tissues. (A) The **p**roportion of epigenetic clock probes retained on the MSA platform (B) Correlation of age estimates using full clocks and MSA probe-only clocks on TCGA normal tissues (C) Correlation of predicted age computed with four different epigenetic clocks with actual age for tissues profiled on MSA. (D) Correlation of cell division estimates computed with PRC2 methylation and solo-WCGW methylation for MSA profiled tissues

Figure 5: Tissue context of human trait associations. (A) Heatmap showing the enrichment of publicly available trait-associated probes across different genomic features (B) Distributions of mean beta value rank for each tissue type over trait-associated CpGs (C) Enrichment of a subset of designed tissue-specific methylation sets on MSA in the SNP sets associated with different GWAS traits

REFERENCES

1. Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21. 10.1101/gad.947102.

2. Chen, R.Z., Pettersson, U., Beard, C., Jackson-Grusby, L., and Jaenisch, R. (1998). DNA hypomethylation leads to elevated mutation rates. Nature *395*, 89–93. 10.1038/25779.

3. Greenberg, M.V.C., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol. *20*, 590–607. 10.1038/s41580-019-0159-6.

4. Flanagan, J.M. (2015). Epigenome-wide association studies (EWAS): past, present, and future. Methods Mol. Biol. *1238*, 51–63. 10.1007/978-1-4939-1804-1_3.

5. Locke, W.J., Guanzon, D., Ma, C., Liew, Y.J., Duesing, K.R., Fung, K.Y.C., and Ross, J.P. (2019). DNA methylation cancer biomarkers: translation to the clinic. Front. Genet. *10*, 1150. 10.3389/fgene.2019.01150.

6. Laird, P.W. (2003). The power and the promise of DNA methylation markers. Nat. Rev. Cancer *3*, 253–266. 10.1038/nrc1045.

7. Levy, M.A., McConkey, H., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Bralo, M.P., Cappuccio, G., Ciolfi, A., Clarke, A., et al. (2022). Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. HGG Adv. *3*, 100075. 10.1016/j.xhgg.2021.100075.

8. Rots, D., Chater-Diehl, E., Dingemans, A.J.M., Goodman, S.J., Siu, M.T., Cytrynbaum, C., Choufani, S., Hoang, N., Walker, S., Awamleh, Z., et al. (2021). Truncating SRCAP variants outside the Floating-Harbor syndrome locus cause a distinct neurodevelopmental disorder with a specific DNA methylation signature. Am. J. Hum. Genet. *108*, 1053–1068. 10.1016/j.ajhg.2021.04.008.

9. Sadikovic, B., Levy, M.A., Kerkhof, J., Aref-Eshghi, E., Schenkel, L., Stuart, A., McConkey, H., Henneman, P., Venema, A., Schwartz, C.E., et al. (2021). Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. Genet. Med. *23*, 1065–1074. 10.1038/s41436-020-01096-4.

10. Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature *555*, 469–474. 10.1038/nature26000.

11. Kerachian, M.A., Azghandi, M., Mozaffari-Jovin, S., and Thierry, A.R. (2021). Guidelines for pre-analytical conditions for assessing the methylation of circulating cell-free DNA. Clin. Epigenetics *13*, 193. 10.1186/s13148-021-01182-7.

12. Chen, J., Gatev, E., Everson, T., Conneely, K.N., Koen, N., Epstein, M.P., Kobor, M.S., Zar, H.J., Stein, D.J., and Hüls, A. (2023). Pruning and thresholding approach for methylation risk scores in multi-ancestry populations. Epigenetics *18*, 2187172. 10.1080/15592294.2023.2187172.

13. Mannens, M.M.A.M., Lombardi, M.P., Alders, M., Henneman, P., and Bliek, J. (2022). Further introduction of DNA methylation (dnam) arrays in regular diagnostics. Front. Genet. *13*, 831452. 10.3389/fgene.2022.831452.

14. Zeng, Y., Jain, R., Lam, M., Ahmed, M., Guo, H., Xu, W., Zhong, Y., Wei, G.-H., Xu, W., and He, H.H. (2023). DNA methylation modulated genetic variant effect on gene transcriptional regulation. Genome Biol. *24*, 285. 10.1186/s13059-023-03130-5.

15. Kaluscha, S., Domcke, S., Wirbelauer, C., Stadler, M.B., Durdu, S., Burger, L., and Schübeler, D. (2022). Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. Nat. Genet. *54*, 1895–1906. 10.1038/s41588-022-01241-6.

16. Luo, Q., Dwaraka, V.B., Chen, Q., Tong, H., Zhu, T., Seale, K., Raffaele, J.M., Zheng, S.C., Mendez, T.L., Chen, Y., et al. (2023). A meta-analysis of immune-cell fractions at high resolution reveals novel associations with common phenotypes and health outcomes. Genome Med. *15*, 59. 10.1186/s13073-023-01211-5.

17. Laird, P.W. (2010). Principles and challenges of genomewide DNA methylation analysis. Nat. Rev. Genet. *11*, 191–203. 10.1038/nrg2732.

18. Iqbal, W., and Zhou, W. (2023). Computational Methods for Single-cell DNA Methylome Analysis. Genomics Proteomics Bioinformatics *21*, 48–66. 10.1016/j.gpb.2022.05.007.

19. Karemaker, I.D., and Vermeulen, M. (2018). Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. Trends Biotechnol. *36*, 952–965. 10.1016/j.tibtech.2018.04.002.

20. Tian, W., Zhou, J., Bartlett, A., Zeng, Q., Liu, H., Castanon, R.G., Kenworthy, M., Altshul, J., Valadon, C., Aldridge, A., et al. (2023). Single-cell DNA methylation and 3D genome architecture in the human brain. Science *382*, eadf5357. 10.1126/science.adf5357.

21. Zhou, W., Dinh, H.Q., Ramjan, Z., Weisenberger, D.J., Nicolet, C.M., Shen, H., Laird, P.W., and Berman, B.P. (2018). DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat. Genet. *50*, 591–602. 10.1038/s41588-018-0073-4.

22. Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., et al. (2023). A DNA methylation atlas of normal human cell types. Nature *613*, 355–364. 10.1038/s41586-022-05580-6.

23. Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. *33*, 5868–5877. 10.1093/nar/gki901.

24. Vermeulen, C., Pagès-Gallego, M., Kester, L., Kranendonk, M.E.G., Wesseling, P., Verburg, N., de Witt Hamer, P., Kooi, E.J., Dankmeijer, L., van der Lugt, J., et al. (2023). Ultra-fast deep-learned CNS tumour classification during surgery. Nature *622*, 842–849. 10.1038/s41586-023-06615-2.

25. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288–295. 10.1016/j.ygeno.2011.07.007.

26. Maden, S.K., Thompson, R.F., Hansen, K.D., and Nellore, A. (2021). Human methylome variation across Infinium 450K data on the Gene Expression Omnibus. NAR Genom. Bioinform. *3*, lqab025. 10.1093/nargab/lqab025.

27. Hawe, J.S., Wilson, R., Schmid, K.T., Zhou, L., Lakshmanan, L.N., Lehne, B.C., Kühnel, B., Scott, W.R., Wielscher, M., Yew, Y.W., et al. (2022). Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nat. Genet. *54*, 18–29. 10.1038/s41588-021-00969-x.

28. Min, J.L., Hemani, G., Hannon, E., Dekkers, K.F., Castillo-Fernandez, J., Luijk, R., Carnero-Montoro, E., Lawson, D.J., Burrows, K., Suderman, M., et al. (2021). Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat. Genet. *53*, 1311–1321. 10.1038/s41588-021-00923-x.

29. Thompson, M., Hill, B.L., Rakocz, N., Chiang, J.N., Geschwind, D., Sankararaman, S., Hofer, I., Cannesson, M., Zaitlen, N., and Halperin, E. (2022). Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. NPJ Genom. Med. *7*, 50. 10.1038/s41525-022-00320-1.

30. Aref-Eshghi, E., Kerkhof, J., Pedro, V.P., Groupe DI France, Barat-Houari, M., Ruiz-Pallares, N., Andrau, J.-C., Lacombe, D., Van-Gils, J., Fergelot, P., et al. (2020). Evaluation of DNA methylation episignatures for diagnosis and phenotype correlations in 42 mendelian neurodevelopmental disorders. Am. J. Hum. Genet. *106*, 356–370. 10.1016/j.ajhg.2020.01.019.

31. Battram, T., Yousefi, P., Crawford, G., Prince, C., Sheikhali Babaei, M., Sharp, G., Hatcher, C., Vega-Salas, M.J., Khodabakhsh, S., Whitehurst, O., et al. (2022). The EWAS Catalog: a database of epigenome-wide association studies. Wellcome Open Res. *7*, 41. 10.12688/wellcomeopenres.17598.2.

32. Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., Li, R., Xia, L., Zhang, T., Niu, G., et al. (2019). EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res. *47*, D983–D988. 10.1093/nar/gky1027.

33. Berdasco, M., and Esteller, M. (2019). Clinical epigenetics: seizing opportunities for translation. Nat. Rev. Genet. *20*, 109–127. 10.1038/s41576-018-0074-2.

34. Li, H.-T., Xu, L., Weisenberger, D.J., Li, M., Zhou, W., Peng, C.-C., Stachelek, K., Cobrinik, D., Liang, G., and Berry, J.L. (2022). Characterizing DNA methylation signatures of retinoblastoma using aqueous humor liquid biopsy. Nat. Commun. *13*, 5523. 10.1038/s41467-022-33248-2.

35. Haghani, A., Li, C.Z., Robeck, T.R., Zhang, J., Lu, A.T., Ablaeva, J., Acosta-Rodríguez, V.A., Adams, D.M., Alagaili, A.N., Almunia, J., et al. (2023). DNA methylation networks underlying mammalian traits. Science *381*, eabq5693. 10.1126/science.abq5693.

36. Ding, W., Kaur, D., Horvath, S., and Zhou, W. (2023). Comparative epigenome analysis using Infinium DNA methylation BeadChips. Brief. Bioinformatics *24*. 10.1093/bib/bbac617.

37. Arneson, A., Haghani, A., Thompson, M.J., Pellegrini, M., Kwon, S.B., Vu, H., Maciejewski, E., Yao, M., Li, C.Z., Lu, A.T., et al. (2022). A mammalian methylation array for profiling methylation levels at conserved sequences. Nat. Commun. *13*, 783. 10.1038/s41467-022-28355-z.

38. He, B., Zhang, C., Zhang, X., Fan, Y., Zeng, H., Liu, J., Meng, H., Bai, D., Peng, J., Zhang, Q., et al. (2021). Tissue-specific 5-hydroxymethylcytosine landscape of the human genome. Nat. Commun. *12*, 4249. 10.1038/s41467-021-24425-w.

39. Cui, X.-L., Nie, J., Ku, J., Dougherty, U., West-Szymanski, D.C., Collin, F., Ellison, C.K., Sieh, L., Ning, Y., Deng, Z., et al. (2020). A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. Nat. Commun. *11*, 6161. 10.1038/s41467-020-20001-w.

40. Horvath, S., and Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat. Rev. Genet. *19*, 371–384. 10.1038/s41576-018-0004-3.

41. Weidner, C.I., Lin, Q., Koch, C.M., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D.O., Jöckel, K.-H., Erbel, R., Mühleisen, T.W., et al. (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. *15*, R24. 10.1186/gb-2014-15-2-r24.

42. Xia, D., Leon, A.J., Cabanero, M., Pugh, T.J., Tsao, M.S., Rath, P., Siu, L.L.-Y., Yu, C., Bedard, P.L., Shepherd, F.A., et al. (2020). Minimalist approaches to cancer tissue-of-origin classification by DNA methylation. Mod. Pathol. *33*, 1874–1888. 10.1038/s41379-020-0547-7.

43. 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. (2022). Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. Cell Genomics *2*. 10.1016/j.xgen.2022.100107.

44. 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. (2019). Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat. Methods *16*, 999–1006. 10.1038/s41592-019-0547-z.

45. 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. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604. 10.1126/science.aan3351.

46. Martens, J.H.A., and Stunnenberg, H.G. (2013). BLUEPRINT: mapping human blood cell epigenomes. Haematologica *98*, 1487–1489. 10.3324/haematol.2013.094243.

47. Vu, H., and Ernst, J. (2022). Universal annotation of the human genome through integration of over a thousand epigenomic datasets. Genome Biol. *23*, 9. 10.1186/s13059-021-02572-z.

48. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2022). Author Correction: Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *605*, E3. 10.1038/s41586-021-04226-3.

49. Zhou, W., Hinoue, T., Barnes, B., Mitchell, O., Iqbal, W., Lee, S.M., Foy, K.K., Lee, K.-H., Moyer, E.J., VanderArk, A., et al. (2022). DNA methylation dynamics and dysregulation delineated by high-throughput profiling in the mouse. Cell Genomics *2*. 10.1016/j.xgen.2022.100144.

50. Kaur, D., Lee, S.M., Goldberg, D., Spix, N.J., Hinoue, T., Li, H.-T., Dwaraka, V.B., Smith, R., Shen, H., Liang, G., et al. (2023). Comprehensive evaluation of the infinium human methylationepic v2 beadchip. Epigenetics Commun. *3*. 10.1186/s43682-023-00021-5.

51. Lee, S.M., Loo, C.E., Prasasya, R.D., Bartolomei, M.S., Kohli, R.M., and Zhou, W. (2024). Low-input and single-cell methods for Infinium DNA methylation BeadChips. Nucleic Acids Res. *52*, e38. 10.1093/nar/gkae127.

52. Vaisvila, R., Ponnaluri, V.K.C., Sun, Z., Langhorst, B.W., Saleh, L., Guan, S., Dai, N., Campbell, M.A., Sexton, B.S., Marks, K., et al. (2021). Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res. *31*, 1280–1289. 10.1101/gr.266551.120.

53. Groarke, E.M., and Young, N.S. (2019). Aging and Hematopoiesis. Clin. Geriatr. Med. *35*, 285–293. 10.1016/j.cger.2019.03.001.

54. Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., et al. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst. *3*, 385-394.e3. 10.1016/j.cels.2016.09.002.

55. Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R., and Rao, A. (2010). The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS ONE *5*, e8888. 10.1371/journal.pone.0008888.

56. Schutsky, E.K., DeNizio, J.E., Hu, P., Liu, M.Y., Nabel, C.S., Fabyanic, E.B., Hwang, Y., Bushman, F.D., Wu, H., and Kohli, R.M. (2018). Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. Nat. Biotechnol. *36*, 1083–1090. 10.1038/nbt.4204.

57. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115. 10.1186/gb-2013-14-10-r115.

58. Horvath, S., Oshima, J., Martin, G.M., Lu, A.T., Quach, A., Cohen, H., Felton, S., Matsuyama, M., Lowe, D., Kabacik, S., et al. (2018). Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY) *10*, 1758–1775. 10.18632/aging.101508.

59. Elmentaite, R., Kumasaka, N., Roberts, K., Fleming, A., Dann, E., King, H.W., Kleshchevnikov, V., Dabrowska, M., Pritchard, S., Bolt, L., et al. (2021). Cells of the human intestinal tract mapped across space and time. Nature *597*, 250–255. 10.1038/s41586-021-03852-1.

60. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell *49*, 359–367. 10.1016/j.molcel.2012.10.016.

61. Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY) *10*, 573–591. 10.18632/aging.101414.

62. Teschendorff, A.E. (2020). A comparison of epigenetic mitotic-like clocks for cancer risk prediction. Genome Med. *12*, 56. 10.1186/s13073-020-00752-3.

63. Sender, R., and Milo, R. (2021). The distribution of cellular turnover in the human body. Nat. Med. *27*, 45–48. 10.1038/s41591-020-01182-9.

64. Rocamora-Reverte, L., Melzer, F.L., Würzner, R., and Weinberger, B. (2020). The complex role of regulatory T cells in immunity and aging. Front. Immunol. *11*, 616949. 10.3389/fimmu.2020.616949.

65. Zhou, W., Laird, P.W., and Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. *45*, e22. 10.1093/nar/gkw967.

66. Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat. Commun. *9*, 5068. 10.1038/s41467-018-07466-6.

67. Guintivano, J., Arad, M., Gould, T.D., Payne, J.L., and Kaminsky, Z.A. (2014). Antenatal prediction of postpartum depression with blood DNA methylation biomarkers. Mol. Psychiatry *19*, 560–567. 10.1038/mp.2013.62.

68. Zhang, Z., Lee, M.K., Perreard, L., Kelsey, K.T., Christensen, B.C., and Salas, L.A. (2022). Navigating the hydroxymethylome: experimental biases and quality control tools for the tandem bisulfite and oxidative bisulfite Illumina microarrays. Epigenomics *14*, 139–152. 10.2217/epi-2021-0490.

69. Tiedemann, R.L., Eden, H.E., Huang, Z., Robertson, K.D., and Rothbart, S.B. (2021). Distinguishing active versus passive DNA demethylation using illumina methylationepic beadchip microarrays. Methods Mol. Biol. *2272*, 97–140. 10.1007/978-1-0716-1294-1_7.

70. Zhou, W., Johnson, B.K., Morrison, J., Beddows, I., Eapen, J., Katsman, E., Semwal, A., Habib, W.A., Heo, L., Laird, P.W., et al. (2024). BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. Nucleic Acids Res. *52*, e32. 10.1093/nar/gkae097.

71. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. 10.1093/nar/29.1.308.

72. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. 10.1038/s41586-020-2493-4.

73. Gao, G.F., Parker, J.S., Reynolds, S.M., Silva, T.C., Wang, L.-B., Zhou, W., Akbani, R., Bailey, M., Balu, S., Berman, B.P., et al. (2019). Before and after: comparison of legacy and harmonized TCGA genomic data commons' data. Cell Syst. *9*, 24-34.e10. 10.1016/j.cels.2019.06.006.

74. Gunasekara, C.J., MacKay, H., Scott, C.A., Li, S., Laritsky, E., Baker, M.S., Grimm, S.L., Jun, G., Li, Y., Chen, R., et al. (2023). Systemic interindividual epigenetic variation in humans is associated with transposable elements and under strong genetic control. Genome Biol. *24*, 2. 10.1186/s13059-022-02827-3.

75. Gunasekara, C.J., Scott, C.A., Laritsky, E., Baker, M.S., MacKay, H., Duryea, J.D., Kessler, N.J., Hellenthal, G., Wood, A.C., Hodges, K.R., et al. (2019). A genomic atlas of systemic interindividual epigenetic variation in humans. Genome Biol. *20*, 105. 10.1186/s13059-019-1708-1.

76. Skaar, D.A., Li, Y., Bernal, A.J., Hoyo, C., Murphy, S.K., and Jirtle, R.L. (2012). The human imprintome: regulatory mechanisms, methods of ascertainment, and roles in disease susceptibility. ILAR J. *53*, 341–358. 10.1093/ilar.53.3-4.341.

77. Wang, T., Luo, M., Berrios, K.N., Schutsky, E.K., Wu, H., and Kohli, R.M. (2021). Bisulfite-Free Sequencing of 5-Hydroxymethylcytosine with APOBEC-Coupled Epigenetic Sequencing (ACE-Seq). Methods Mol. Biol. *2198*, 349–367. 10.1007/978-1-0716-0876-0_27.

78. Zhou, W., Triche, T.J., Laird, P.W., and Shen, H. (2018). SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. *46*, e123. 10.1093/nar/gky691.

79. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set

enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA *102*, 15545–15550. 10.1073/pnas.0506580102.

80. Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M., et al. (2021). Gene Set Knowledge Discovery with Enrichr. Curr. Protoc. *1*, e90. 10.1002/cpz1.90.

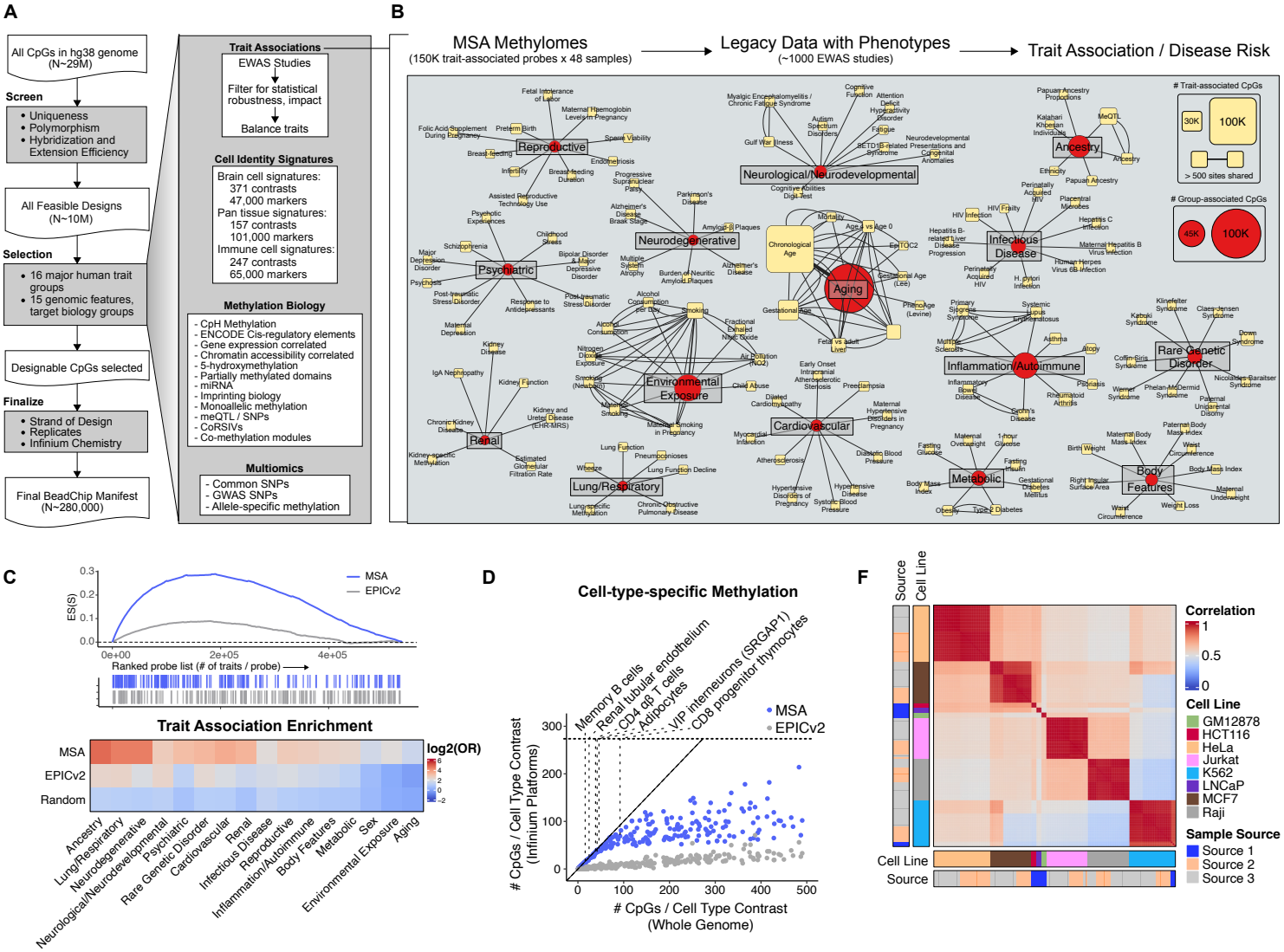
81. Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics *18*, 105. 10.1186/s12859-017-1511-5.

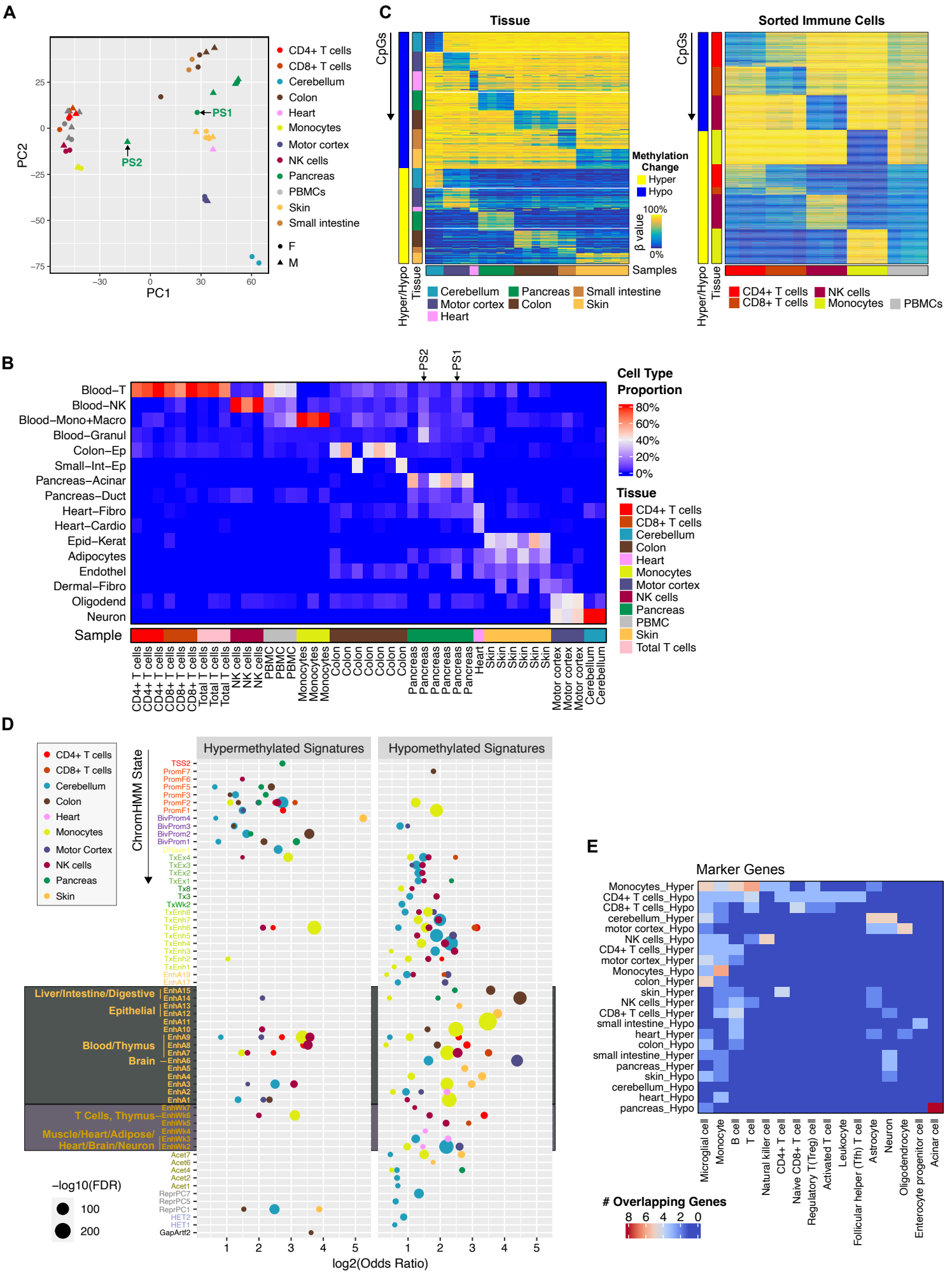
82. Hammal, F., de Langen, P., Bergon, A., Lopez, F., and Ballester, B. (2022). ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Res. *50*, D316–D325. 10.1093/nar/gkab996.

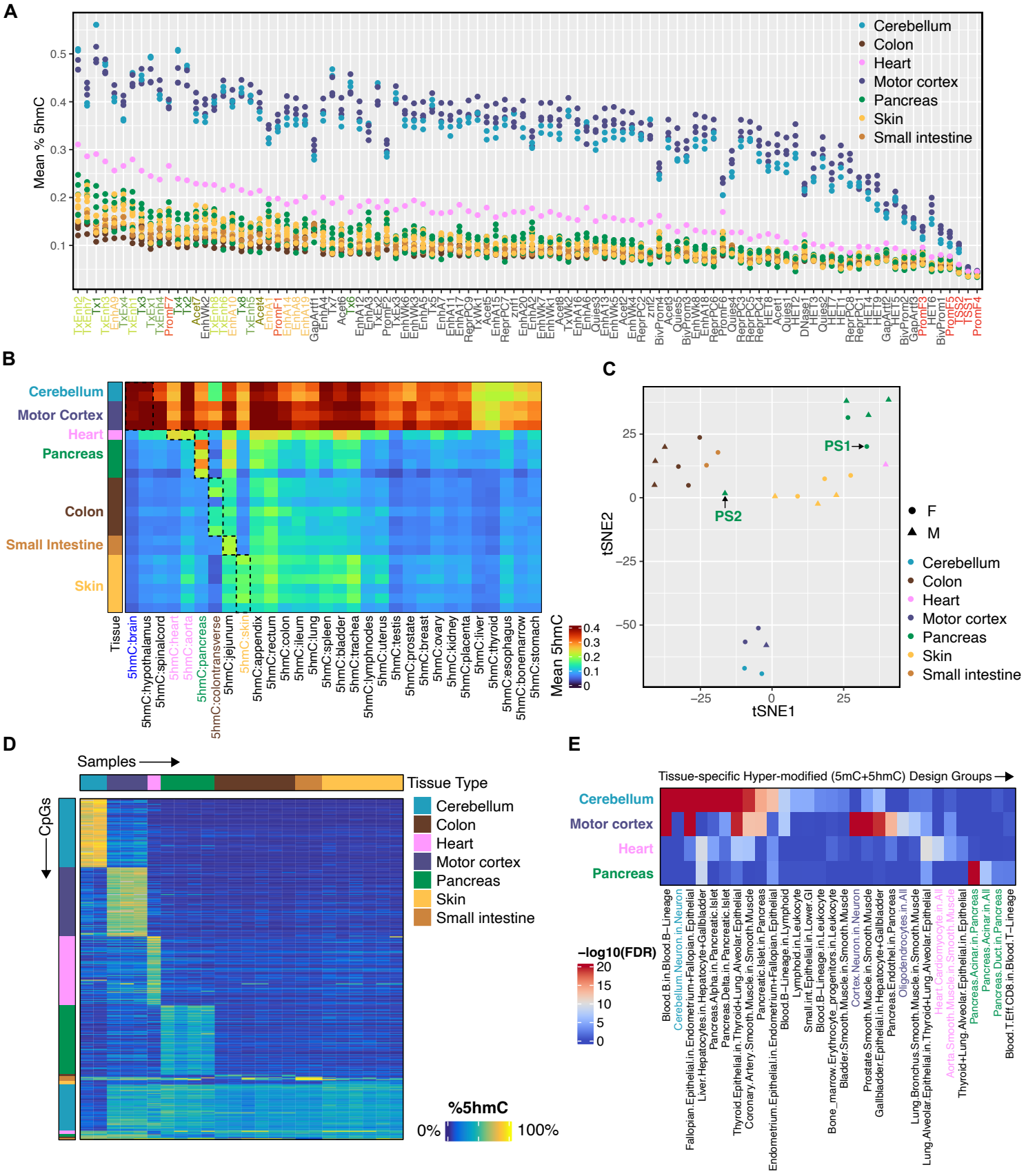
83. Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., et al. (2023). CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. *51*, D870–D876. 10.1093/nar/gkac947.

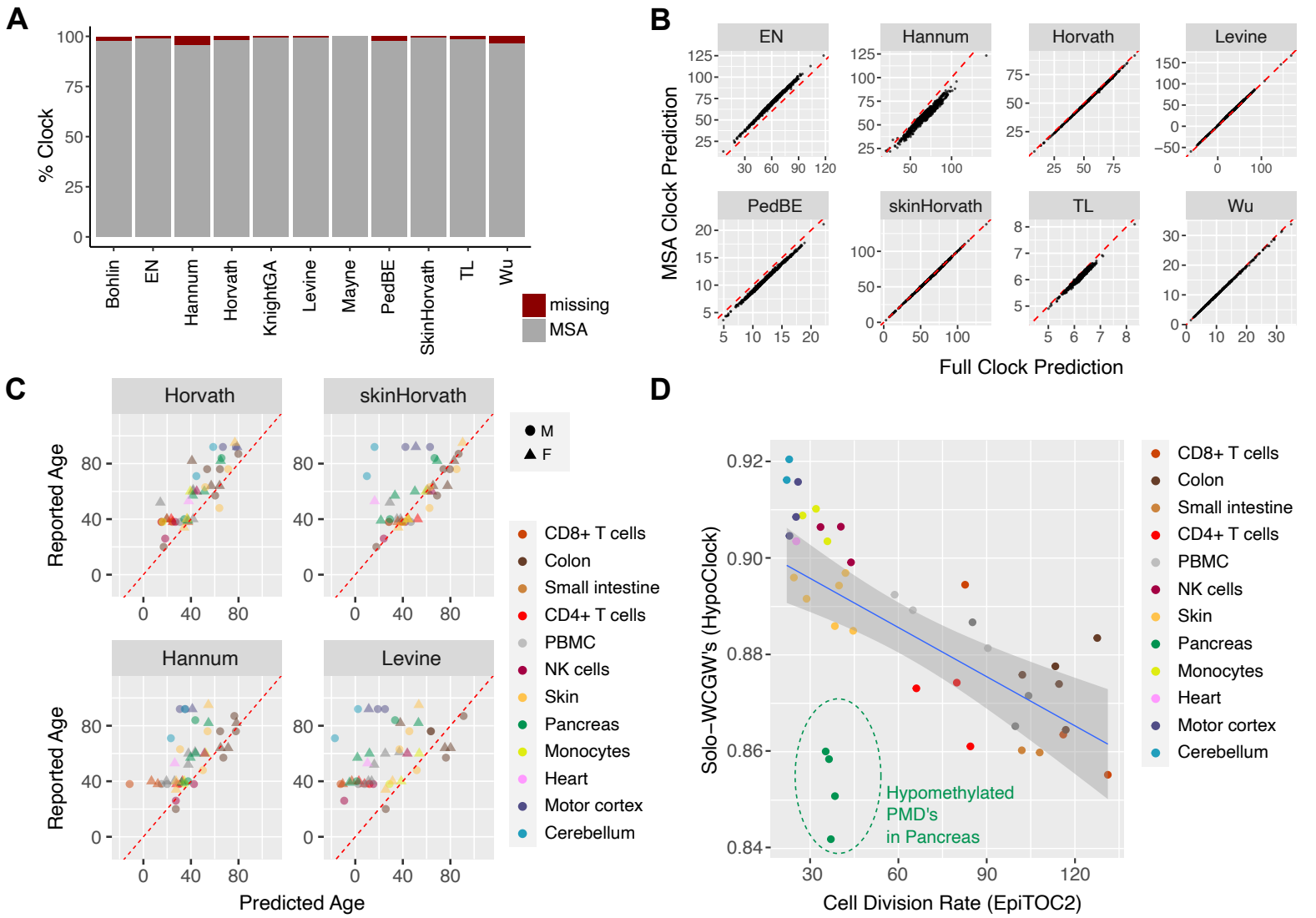
84. Pelegí-Sisó, D., de Prado, P., Ronkainen, J., Bustamante, M., and González, J.R. (2021). methylclock: a Bioconductor package to estimate DNA methylation age. Bioinformatics *37*, 1759–1760. 10.1093/bioinformatics/btaa825.

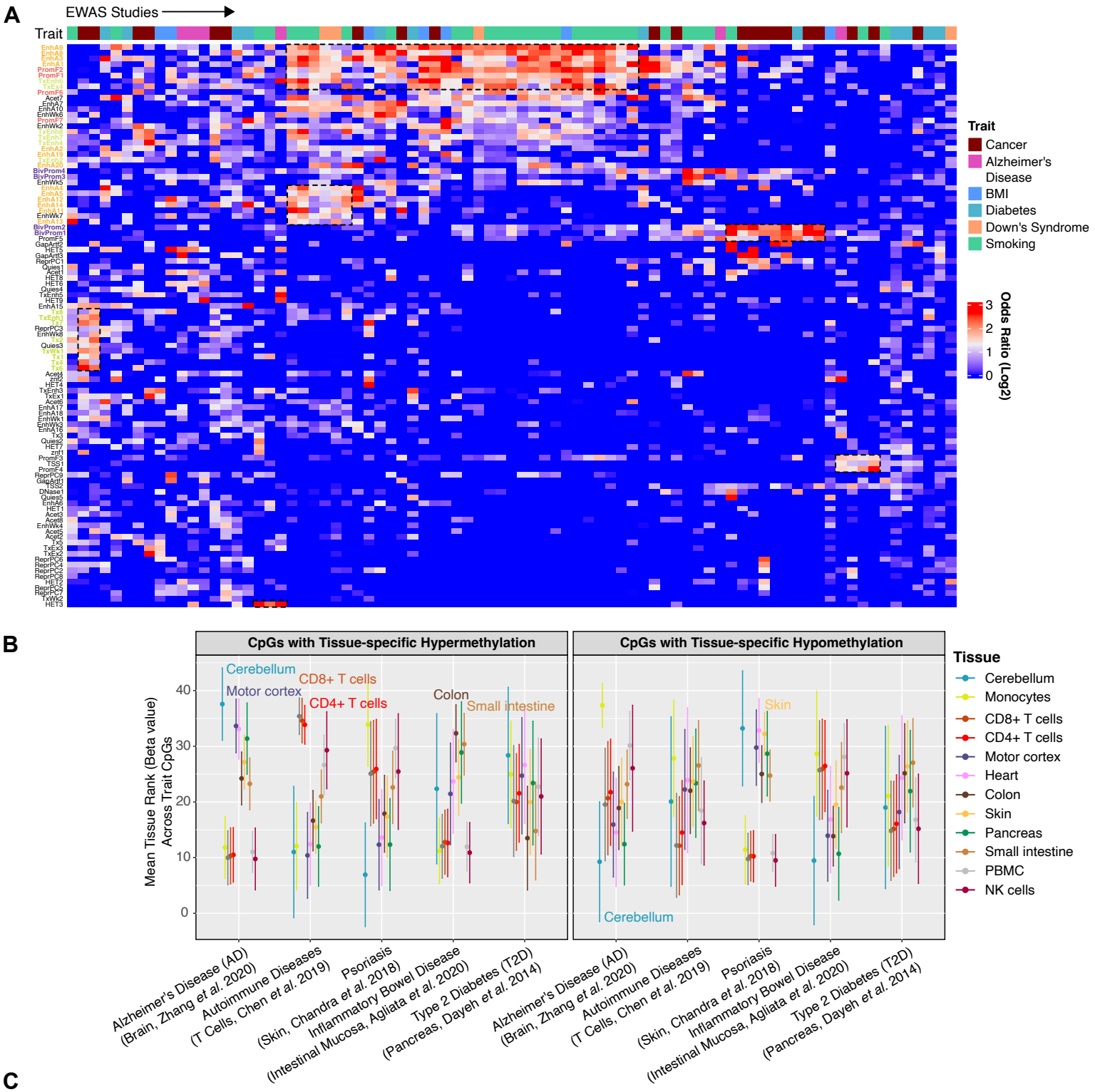
85. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. *51*, D977–D985. 10.1093/nar/gkac1010.

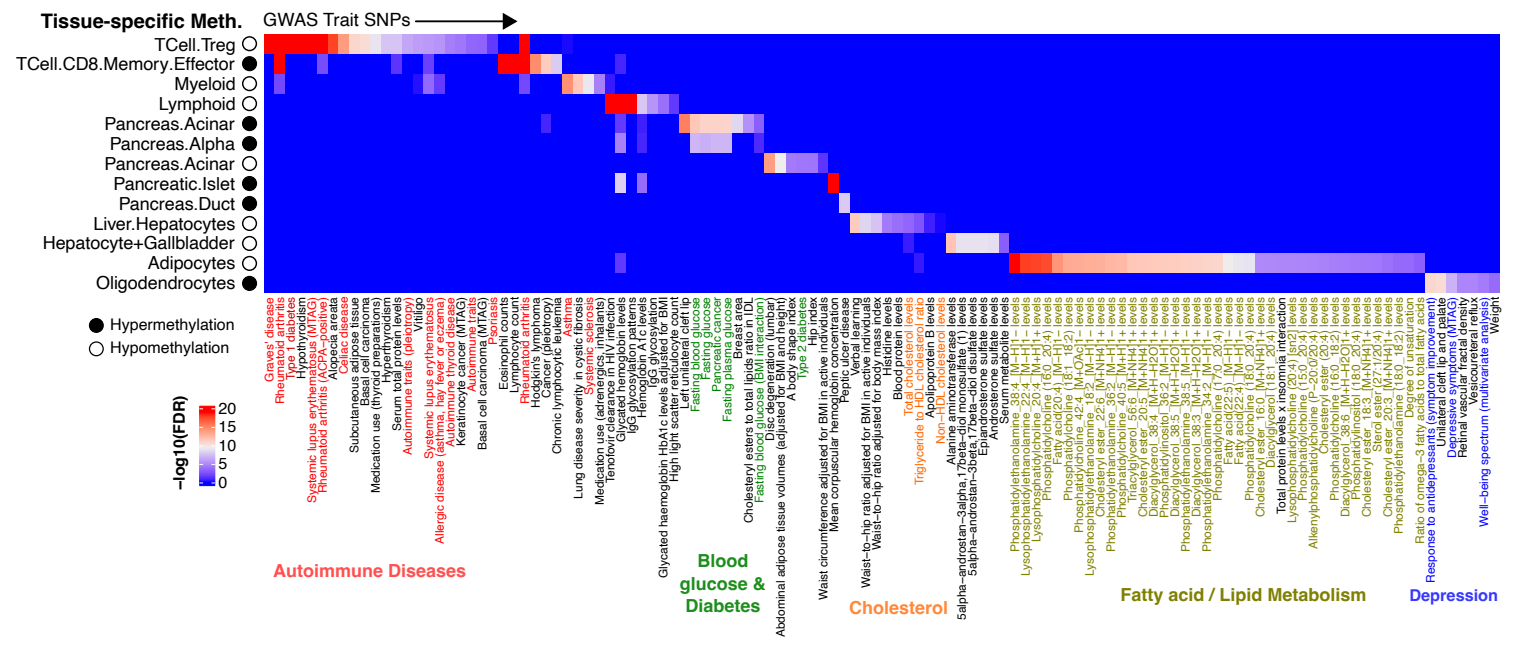

A
All CpGs in hg38 genome (N~29M)
Screen
• Uniqueness
• Polymorphism
• Hybridization and Extension Efficiency
All Feasible Designs (N~10M)
Selection
• 16 major human trait groups
• 15 genomic features, target biology groups
Designable CpGs selected
Finalize
• Strand of Design
• Replicates
• Infinium Chemistry
Final BeadChip Manifest (N~280,000)
Trait Associations
EWAS Studies
Filter for statistical robustness, impact
Balance traits
Cell Identity Signatures
Brain cell signatures:
371 contrasts
47,000 markers
Pan tissue signatures:
157 contrasts
101,000 markers
Immune cell signatures:
247 contrasts
65,000 markers
Methylation Biology
- CpH Methylation
- ENCODE Cis-regulatory elements
- Gene expression correlated
- Chromatin accessibility correlated
- 5-hydroxymethylation
- Partially methylated domains
- miRNA
- Imprinting biology
- Monoallelic methylation
- meQTL / SNPs
- CoRSIVs
- Co-methylation modules
Multiomics
- Common SNPs
- GWAS SNPs
- Allele-specific methylation
B
MSA Methylomes
(150K trait-associated probes x 48 samples)
Legacy Data with Phenotypes
(~1000 EWAS studies)
Trait Association / Disease Risk
Trait-associated CpGs
30K
100K
> 500 sites shared
Group-associated CpGs
45K
100K
Reproductive
Neurological/Neurodevelopmental
Ancestry
Neurodegenerative
Psychiatric
Chronological Age
Aging
Infectious Disease
Environmental Exposure
Inflammation/Autoimmune
Rare Genetic Disorder
Cardiovascular
Renal
Metabolic
Body Features
Lung/Respiratory
C
MSA
EPICv2
ES(S)
Ranked probe list (# of traits / probe)
Trait Association Enrichment
MSA
EPICv2
Random
log2(OR)
Ancestry
Lung/Respiratory
Neurodegenerative
Neurodevelopmental
Psychiatric
Rare Genetic Disorder
Cardiovascular
Renal
Infectious Disease
Reproductive
Inflammation/Autoimmune
Body Features
Metabolic
Sex
Environmental Exposure
Aging
D
Cell-type-specific Methylation
Memory B cells
Renal tubular endothelium
CD4 αβ T cells
Adipocytes
VIP interneurons (SRGAP1)
CD8 progenitor thymocytes
MSA
EPICv2
CpGs / Cell Type Contrast (Infinium Platforms)
CpGs / Cell Type Contrast (Whole Genome)
F
Source
Cell Line
Correlation
Cell Line
GM12878
HCT116
HeLa
Jurkat
K562
LNCaP
MCF7
Raji
Sample Source
Source 1
Source 2
Source 3
E
GO Terms (neuron-related)
Regulation Of Small GTPase Mediated Signal Transduction
Axonogenesis
Modulation Of Chemical Synaptic Transmission
Cellular Response To Calcium Ion
Nervous System Development
Response To Transforming Growth Factor Beta
Neuron Migration
Chondroitin Sulfate Biosynthetic Process
Proteoglycan Metabolic Process
Neuron Projection Guidance
Axon Guidance
Cellular Response To Growth Factor Stimulus
Calcium Ion Transmembrane Import Into Cytosol
Positive Regulation Of Osteoblast Differentiation
Response To Calcium Ion
Positive Regulation Of Nucleic Acid−Templated Transcription
Positive Regulation Of DNA−templated Transcription
Regulation Of Neuron Projection Development
Positive Regulation Of Myoblast Differentiation
Regulation Of DNA Repair
Positive Regulation Of Double−Strand Break Repair
Chromatin Remodeling
Regulation Of Double−Strand Break Repair
Regulation Of Cell Cycle Process
DNA Repair
CpH probes
MSA
EPICv2
Odds Ratio
Overlap
−log10 (P-value)
G
EpigenDx
Human Blood
Jurkat
β value
% Titrated Methylation Level
Methylation Level
H
EM-seq
GM12878
HCT116
K562
Correlation
Cell Line
GM12878
HCT116
K562
Source
Source 1
Source 2
Source 3


A
PC2
PC1
PS1
PS2
CD4+ T cells
CD8+ T cells
Cerebellum
Colon
Heart
Monocytes
Motor cortex
NK cells
Pancreas
PBMCs
Skin
Small intestine
F
M
C
Tissue
Sorted Immune Cells
CpGs
Methylation Change
Hyper
Hypo
β value
100%
0%
Samples
Hyper/Hypo
Tissue
Cerebellum
Pancreas
Small intestine
Motor cortex
Colon
Skin
Heart
CD4+ T cells
NK cells
CD8+ T cells
Monocytes
PBMCs
B
Blood−T
Blood−NK
Blood−Mono+Macro
Blood−Granul
Colon−Ep
Small−Int−Ep
Pancreas−Acinar
Pancreas−Duct
Heart−Fibro
Heart−Cardio
Epid−Kerat
Adipocytes
Endothel
Dermal−Fibro
Oligodend
Neuron
Sample
Cell Type Proportion
80%
60%
40%
20%
0%
Tissue
D
Hypermethylated Signatures
Hypomethylated Signatures
ChromHMM State
Liver/Intestine/Digestive
Epithelial
Blood/Thymus
Brain
T Cells, Thymus
Muscle/Heart/Adipose/
Heart/Brain/Neuron
−log10(FDR)
100
200
log2(Odds Ratio)
E
Marker Genes
Monocytes_Hyper
CD4+ T cells_Hypo
CD8+ T cells_Hypo
cerebellum_Hyper
motor cortex_Hypo
NK cells_Hypo
CD4+ T cells_Hyper
motor cortex_Hyper
Monocytes_Hypo
colon_Hyper
skin_Hyper
NK cells_Hyper
CD8+ T cells_Hyper
small intestine_Hypo
heart_Hyper
colon_Hypo
small intestine_Hyper
pancreas_Hyper
skin_Hypo
cerebellum_Hypo
heart_Hypo
pancreas_Hypo
Microglial cell
Monocyte
B cell
T cell
Natural killer cell
CD4+ T cell
Naive CD8+ T cell
Regulatory T(Treg) cell
Activated T cell
Leukocyte
Follicular helper (Tfh) T cell
Astrocyte
Neuron
Oligodendrocyte
Enterocyte progenitor cell
Acinar cell
Overlapping Genes


A
Mean % 5hmC
Cerebellum
Colon
Heart
Motor cortex
Pancreas
Skin
Small intestine
B
Cerebellum
Motor Cortex
Heart
Pancreas
Colon
Small Intestine
Skin
Tissue
Mean 5hmC
C
PS1
PS2
tSNE1
tSNE2
F
M
Cerebellum
Colon
Heart
Motor cortex
Pancreas
Skin
Small intestine
D
Samples
CpGs
Tissue Type
Cerebellum
Colon
Heart
Motor cortex
Pancreas
Skin
Small intestine
%5hmC
0%
100%
E
Tissue-specific Hyper-modified (5mC+5hmC) Design Groups
Cerebellum
Motor cortex
Heart
Pancreas
−log10(FDR)


A
% Clock
Bohlin
EN
Hannum
Horvath
KnightGA
Levine
Mayne
PedBE
SkinHorvath
TL
Wu
missing
MSA
B
EN
Hannum
Horvath
Levine
PedBE
skinHorvath
TL
Wu
MSA Clock Prediction
Full Clock Prediction
C
Horvath
skinHorvath
Hannum
Levine
Reported Age
Predicted Age
M
F
CD8+ T cells
Colon
Small intestine
CD4+ T cells
PBMC
NK cells
Skin
Pancreas
Monocytes
Heart
Motor cortex
Cerebellum
D
Solo−WCGW's (HypoClock)
Cell Division Rate (EpiTOC2)
Hypomethylated
PMD's
in Pancreas
CD8+ T cells
Colon
Small intestine
CD4+ T cells
PBMC
NK cells
Skin
Pancreas
Monocytes
Heart
Motor cortex
Cerebellum


A
EWAS Studies
Trait
Cancer
Alzheimer's Disease
BMI
Diabetes
Down's Syndrome
Smoking
Odds Ratio (Log2)
B
CpGs with Tissue-specific Hypermethylation
CpGs with Tissue-specific Hypomethylation
Mean Tissue Rank (Beta value) Across Trait CpGs
Cerebellum
Motor cortex
CD8+ T cells
CD4+ T cells
Colon
Small intestine
Skin
Tissue
Cerebellum
Monocytes
CD8+ T cells
CD4+ T cells
Motor cortex
Heart
Colon
Skin
Pancreas
Small intestine
PBMC
NK cells
Alzheimer's Disease (AD) (Brain, Zhang et al. 2020)
Autoimmune Diseases (T Cells, Chen et al. 2019)
Psoriasis (Skin, Chandra et al. 2018)
Inflammatory Bowel Disease (Intestinal Mucosa, Agliata et al. 2020)
Type 2 Diabetes (T2D) (Pancreas, Dayeh et al. 2014)


C
Tissue-specific Meth.
GWAS Trait SNPs
TCell.Treg
TCell.CD8.Memory.Effector
Myeloid
Lymphoid
Pancreas.Acinar
Pancreas.Alpha
Pancreas.Acinar
Pancreatic.Islet
Pancreas.Duct
Liver.Hepatocytes
Hepatocyte+Gallbladder
Adipocytes
Oligodendrocytes
Hypermethylation
Hypomethylation
−log10(FDR)
Autoimmune Diseases
Blood glucose & Diabetes
Cholesterol
Fatty acid / Lipid Metabolism
Depression