An atlas of genetic effects on the monocyte methylome across European and African populations

Running title: Multi-ancestry monocyte methylome study

Wanheng Zhang[1,2,#], Xiao Zhang[3,#], Chuan Qiu[3], Zichen Zhang[1], Kuan-Jui Su[3], Zhe Luo[3], Minghui Liu[4], Bingxin Zhao[5], Lang Wu[6], Qing Tian[3], Hui Shen[3,*], Chong Wu[1,7,*] and Hong-Wen Deng[3,*]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

[2]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[3]Division of Biomedical Informatics and Genomics, Tulane Center of Biomedical Informatics and Genomics, Deming Department of Medicine, Tulane University, New Orleans, LA 70112, USA

[4]Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[5]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA

[6]Cancer Epidemiology Division, Population Sciences in the Pacific Program, University of Hawaii Cancer Center, University of Hawaii at Manoa, Honolulu, HI, 96813, USA

[7]Institute for Data Science in Oncology, The UT MD Anderson Cancer Center, Houston, TX, 77030, USA.

[#]*These authors contribute equally*

**Corresponding authors:*

Hui Shen (hshen3@tulane.edu), Chong Wu (cwu18@mdanderson.org) and Hong-Wen Deng (hdeng2@tulane.edu)

Hui Shen, Ph.D.
Professor
Associate Director, Center for Biomedical Informatics & Genomics
Associate Director, Tulane Integrated Institute of Data and Health Sciences (TIIDHS)
School of Medicine, Tulane University
1440 Canal Street, Suite 1621, New Orleans, LA 70112, USA
Phone: 504-988-6987
Email: hshen3@tulane.edu

Chong Wu, Ph.D.
Assistant Professor
Department of Biostatistics, University of Texas MD Anderson Cancer Center
7007 Bertner Avenue, Unit 1689, Houston, TX 77030
Phone: 713-409-5160
Email: cwu18@mdanderson.org

Hong-Wen Deng, Ph.D.
Professor
Aron Family Endowed Chair
Director, Center for Biomedical Informatics & Genomics
Director, Tulane Integrated Institute of Data and Health Sciences (TIIDHS)
School of Medicine, Tulane University
1440 Canal Street, Suite 1621, New Orleans, LA 70112, USA
Phone: 504-988-1310
Email: hdeng2@tulane.edu

Abstract

Elucidating the genetic architecture of DNA methylation is crucial for decoding complex disease etiology. However, current epigenomic studies are often limited by incomplete methylation coverage and heterogeneous tissue samples. Here, we present the first comprehensive, multi-ancestry human methylome atlas of purified human monocytes, generated through integrated whole-genome bisulfite sequencing and whole-genome sequencing from 298 European Americans (EA) and 160 African Americans (AA). By analyzing over 25 million methylation sites, we identified 1,383,250 and 1,721,167 methylation quantitative trait loci (meQTLs) in *cis*-regions for EA and AA populations, respectively, revealing both shared (880,108 sites) and population-specific regulatory patterns. Furthermore, we developed population-specific DNAm imputation models, enabling methylome-wide association studies (MWAS) for 1,976,046 and 2,657,581 methylation sites in EA and AA, respectively. These models were validated through multi-ancestry analysis of 41 complex traits from the Million Veteran Program. The identified meQTLs, MWAS models, and data resources are freely available at www.gcbhub.org and https://osf.io/gct57/.

Keywords: meQTL; heritability; WGBS; MWAS; MVP.

Introduction

Unraveling the functional consequences of genetic variation on complex human diseases is a fundamental challenge in human genetics. While genome-wide association studies (GWAS) have identified many single nucleotide polymorphism (SNP)-trait associations across a range of complex diseases and traits,[1] the molecular pathways through which these variations contribute to disease risk remain largely elusive[2,3]. This knowledge gap hinders the advancement of therapeutic targets[4] and underscores the need for studying multi-omics biomarkers[5]. DNA methylation (DNAm), a relatively stable epigenetic mark, has emerged as a critical molecular biomarker in this endeavor. DNAm not only provides insights into the current state of human health[6] but also facilitates the estimation of epigenetic age[7–9] and partially captures the effects of lifestyle and environment exposures[10] on disease pathogenesis. Recognizing its potential, extensive efforts have been made to collect comprehensive, population-scale datasets of the human genome and methylome.[11–19]

However, these pursuits have been constrained by several key limitations. Firstly, the reliance on Illumina BeadChip platforms[20] restricts the analysis to only 3% of the approximately 30 million CpG sites in the human methylome[21–24], due to its focus on only 450,000 or 935,000 predefined CpG methylation sites. This constraint significantly limits our understanding of the methylome's extensive landscape. Secondly, the majority of methylation studies have examined whole blood samples[25–28] or normal bulk tissue[29–31], adjusting the estimated cell type proportions as covariates[32,33]. This approach inherently limits the examination of methylation patterns in specific cell types, which could be crucial for understanding disease mechanisms.[34,35] Monocytes, in particular, play central roles in various physiological and pathological processes, including bone remodeling[36,37], neurodegenerative disorders such as Alzheimer's disease[38,39], inflammatory conditions like rheumatoid arthritis[40] and inflammatory bowel disease, cardiovascular diseases[41], cancer[42], and infectious diseases, making them a highly relevant cell type for investigating the link between genetic variation, DNAm, and complex diseases. Thirdly, DNAm and genetic variations could be ancestry-specific[43–45], highlighting the critical need to study DNAm across multi-ancestry populations to fully characterize their relationships.

To overcome these limitations and accurately characterize the human methylome in purified monocytes across European Americans (EA) and African Americans (AA) ancestry, we performed whole genome bisulfite sequencing (WGBS) and whole genome sequencing (WGS) on 298 independent individuals of EA and 160 independent individuals of AA in the Louisiana Osteoporosis Study[46]. Our comprehensive approach encompassed several key analyses. For each ancestry group, we conducted methylation quantitative trait loci (meQTL) analyses to identify loci in both *cis*- and *trans*-region where genetic variations are associated with DNAm changes, and compared results across ancestries to explore shared and unique genetic architecture. Further, we calculated *cis*-heritability (*cis*-h^2) of DNAm to quantify the proportion of variance in methylation levels that can be explained by *cis*-acting genetic variants. To link our findings to complex traits, we conducted methylome-wide association studies (MWASs) using an instrumental variable regression framework with two steps.[47,48] First, we built DNAm prediction, or imputation models using penalized regression. Second, we tested the association between the predicted DNAm and 41 phenotypes using GWAS summary statistics from the Million Veteran

Program (MVP) dataset[49] in both AA and EA. The results were further validated by colocalization and Mendelian randomization (MR) analyses. This multi-faceted approach provides new insights into the genetic regulation of monocyte-specific DNAm and its relationship to complex traits across diverse ancestries. The overview of our study design is presented in Fig.1, and the developed models and data resources are freely accessible at www.gcbhub.org.

Results

Data collection and processing

We recruited 495 male subjects aged 20-51 years (185 AA and 310 EA) from the ongoing Tulane Louisiana Osteoporosis Study cohort[46] with strict inclusion/exclusion criteria; see Method section and Supplementary Table 1 for details. We restricted our analyses to subjects with purified peripheral blood monocytes (PBMs). Sequencing reads were processed to remove adaptor sequences, contamination, and low-quality bases, then aligned to the human reference genome GRCh38/hg38. In total, methylation levels for over 25 million CpG sites across the genome were quantified (see Methods for details). DNAm β-values underwent rank-based inverse normal transformation[50] to normalize the data distribution. We also implemented comprehensive quality control (QC) for the genetic data. Single nucleotide polymorphisms (SNPs) were excluded based on several criteria: deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$), minor allele frequency (MAF) less than 0.01, and presence of mismatched alleles. Additionally, to mitigate confounding effects from population structure, we removed individuals with a degree of relatedness, defined as an identity by state (IBS) greater than 0.125. This resulted in a final cohort of 160 AA and 298 EA. For the genetic data, we extracted the top ten principal components (PCs) to capture the major axes of genetic variation. Similarly, principal component analysis (PCA) was conducted on the DNAm data, focusing on the 20,000 most variable CpG sites to obtain the top ten nongenetic PCs.[11] Adjusting for nongenetic PCs derived from the DNAm data can significantly improve the power to detect true associations[51], and the risk of introducing collider bias that could impact the results is minimal[11]. As a result, a total of 13,238,663 SNPs were utilized in the AA cohort and 8,513,381 SNPs in the EA cohort.

Identification of *cis*-meQTL reveals methylation landscapes across populations through whole genome bisulfite sequencing (WGBS)

Leveraging the comprehensive coverage of WGBS, we conducted an extensive analysis of methylation quantitative trait loci (*cis*-meQTLs) across 25,721,231 CpG sites in African American (AA, n=160) and European American (EA, n=298) populations. *Cis*-regions were defined as regions extending ±1 Mb from the CpG sites. Using MatrixEQTL[52], we identified *cis*-acting methylation quantitative trait loci (*cis*-meQTL) by testing the association between DNAm and genotype via linear regression, adjusting for confounders including age, BMI, smoking, alcohol consumption, (proportions of B cells, monocytes, neutrophils), and population structure (ten genetic and nongenetic principal components). Furthermore, we controlled for ten genetic and ten nongenetic principal components.

We identified 58,472,084 and 127,178,161 *cis*-meQTL associations ($P < 1 \times 10^{-8}$)[11] in the AA and EA populations, respectively. On average, each CpG site had 2.27 *cis*-meQTL mappings in the AA population and 4.94 in the EA population. In total, 1,721,167 CpG sites in the AA and 1,383,250 in the EA population exhibited at least one *cis*-meQTL, representing 6.7% and 5.4% of all sites tested, respectively, with 880,108 (51.1% of AA and 63.6% of EA) CpG sites shared between populations, suggesting both common and population-specific genetic regulation of DNA methylation.

Our analysis uncovered several key patterns in the genetic architecture of methylation regulation. First, we observed a negative correlation between MAF and effect size (Fig. 2a), with AA showing stronger correlation (correlation $= -0.45,\ P < 2.2 \times 10^{-16}$) compared to the EA population (correlation $= -0.43,\ P < 2.2 \times 10^{-16}$) and larger median effect sizes (AA: median = 0.90, EA: median = 0.65, $P < 1 \times 10^{-16}$ by Wilcoxon sum rank test). These findings extend the principles observed in GWAS studies[53,54] to methylation regulation, while highlighting important population differences. Second, we found that genetic variants closer to CpG sites typically showed stronger associations, with effect sizes decreasing as distance increased (Fig. 2b). This spatial relationship was consistent across populations (Supplementary Fig. 1). Third, investigation of genomic context using Annotatr[55] revealed that majority of CpG sites with *cis*-meQTLs were located in open sea regions (AA: 94.7%, EA: 94.1%), with smaller proportions in CpG islands (AA: 0.9%, EA: 1.1%) and shores/shelves (AA: 4.4%, EA: 4.8%). The enrichment of meQTLs in open sea regions aligns with previous findings[18], suggesting these regions may be more susceptible to genetic regulation.

Furthermore, we found substantial population-specific genetic regulation of methylation. While 30.0% of *cis*-meQTLs in the AA population were rare or nonexistent in Europeans, only 8.1% of EA *cis*-meQTLs were rare or nonexistent in Africans. We defined rare as two or fewer individuals carrying the same variant in the Phase-3 1000 Genome Project[56] (1000G) of the corresponding ancestry. This disparity was even more pronounced for sentinel *cis*-meQTLs, defined as the most significant SNP per CpG site, with 49% of AA variants being rare or absent in Europeans compared to 19% of EA variants in Africans (Fig. 2d). Despite these differences, for shared variants, effect sizes showed high concordance between populations (correlation $= 0.96, P < 2.2 \times 10^{-16}$, Fig. 2e).

To identify likely causal variants underlying significant *cis*-meQTL, we conducted population-specific fine-mapping for 1,655,943 and 1,334,508 CpG sites with at least one *cis*-meQTL in AA and EA populations using SuSiE[57]. We found that the number of variants within the 95% credible sets was notably smaller for the AA population, with a median of 13 and an interquartile range (IQR) of 8,045, as opposed to the EA population, which exhibited a median of 27 and an IQR of 5,316 (Fig. 2f). This observation aligns with previous pQTL study[54] for plasma proteins and likely reflects lower linkage disequilibrium in African populations, though the smaller AA sample size may contribute to this difference.

Cis-h^2 of CpG sites and imputation models

To quantify genetic contributions to DNA methylation variation, we estimated *cis*-heritability (*cis*-h^2, the proportion of methylation variation explained by *cis*-acting SNPs within 1 Mbp) using

restricted maximum likelihood (REML) in GCTA[58]. Among all the 25,721,231 CpG sites in WGBS, we found that 50.5% and 57.0% CpG sites exhibited *cis*-h^2 greater than 0.01, 21.1% and 33.8% CpG sites with a *cis*-h^2 between 0.01 and 0.1, and 28.3% and 9.2% CpG sites with a *cis*-h^2 greater than 0.1 for the AA and EA populations, respectively (Fig. 3a). The mean values for *cis*-h^2 of AA population (0.09) is significantly higher than that of EA population (0.04) ($P < 1 \times 10^{-16}$ by a Wilcoxon sum rank test). Besides, we investigated the p-values for the *cis*-h^2 estimated from REML and found that among CpG sites with *cis*-h^2 greater than 0.01, 16.4% and 18.2% of them having p-values less than 0.05, while among CpG sites with *cis*-h^2 greater than 0.5, 36.8% and 55.1% of them having p-values less than 0.05, for AA and EA respectively (Fig. 3b).

Integration of heritability estimates with meQTL findings revealed strong concordance. For highly heritable sites (*cis*-h^2 > 0.5), the majority showed significant *cis*-meQTL associations (86% in AA, 99% in EA; Fig. 3c). CpG sites with identified *cis*-meQTLs exhibited substantially higher mean heritability (AA: 0.48, EA: 0.33) compared to those without (AA: 0.06, EA: 0.02; Fig. 3d). While AA showed higher overall heritability, we identified fewer *cis*-meQTL associations compared to EA. To investigate this apparent discrepancy, we performed a controlled analysis by randomly sampling 160 individuals from the EA population to match the AA sample size. This reduced analysis identified 48,086,725 *cis*-meQTLs (1.87 per CpG site) in EA, fewer than the 2.27 associations per site in AA, indicating that the original disparity resulted from sample size difference. This finding highlights the importance of increasing sample sizes in diverse populations to reduce disparity.

To evaluate methylation quantification approaches, we compared heritability patterns between WGBS and two common array platforms: the MethylationEPIC Infinium v2.0[59], which contains over 935,000 CpG sites (referred to as the 900K set), and the HumanMethylation450 Infinium assay[60], encompassing around 450,000 sites (referred to as the 450K set). Recent update from EPIC to EPIC v2 have complicated integrating new data with previous Infinium array platforms, such as the EPIC and the HM450K, and has been used in recent studies.[61–63] The two datasets are provided at https://zwdzwd.github.io/InfiniumAnnotation. The distribution of *cis*-h^2 values for CpG sites in the 450K and 900K sets closely mirrored that of the overall WGBS dataset, albeit with slightly lower values (mean *cis*-h^2 of 900K/450K: AA = 0.08/0.08, EA = 0.03/0.03; Fig. 3e). This comparison demonstrates that while array-based platforms effectively sample representative CpG sites, WGBS provides comprehensive genome-wide methylation profiling, enabling identification of additional heritable sites and potentially novel regulatory regions that may be missed by targeted approaches.

Multi-ancestry MWAS of complex traits in MVP

We first built population-specific methylation prediction models for CpG sites with *cis*-h^2 > 0.01 using Elastic Net[64] with nested cross-validation[65]. We selected SNPs within 500 kb of target CpG sites as predictors for computational efficiency. We built DNAm prediction models (with R^2 value of at least 0.01) for 2,657,581 and 1,976,046 CpG sites in AA and EA populations, with 1,067,816 shared CpG sites. The mean accuracy for models for DNAm prediction, measured by R^2, was 0.19 and 0.17 for AA and EA populations, respectively, which is consistently higher compared to

array-covered sites (450K/900K mean R^2: AA=0.11/0.10; EA = 0.10/0.09). This finding aligns with the observed heritability patterns and demonstrating the enhanced predictive power gained through comprehensive methylome coverage.

Using these prediction models, we conducted methylome-wide association studies (MWAS) for 41 phenotypes using GWAS summary statistics from the Million Veteran Program (MVP) dataset[49]. Similar to transcriptome-wide association studies (TWASs)[66–68] which assesses gene expression levels and their associations with traits, and proteome-wide association studies (PWASs)[3,54] that focus on the proteomic landscape to understand protein variations and their disease linkages, MWASs focus on the association between DNAm and complex traits, and operates through a two-step instrumental variable regression framework[47,48]. After Bonferroni correction across all models and phenotypes (P cutoff is 4.9×10^{-10} for AA and 6.6×10^{-10} for EA), we identified 34,334 significant methylation-phenotype associations (Supplementary Tables 2.1 and 2.2), with EA contributing 31,336 and AA contributing 2,998 associations (Fig. 4a), primarily because the sample size of EA GWAS data (mean = 261,202) is much larger than that of AA (mean = 54,271). Notably, WGBS enabled identification of numerous associations beyond array-based platforms: for AA, 2,944 (98.1%) and 2,926 (97.6%) significant associations were not captured by the 450K and 900K methylation panels, respectively. Similarly, for EA, 30,835 (98.4%) and 30,394 (97.0%) associations were beyond the coverage of these panels. This highlights our study's extensive coverage, significantly encompassing a broader scope of the methylome, indicating a comprehensive mapping of methylation associations across ancestries.

Among the identified associations, 559 unique methylation-phenotype association pairs were significant in both ancestries; the majority (537 out of 559) of these associations displayed consistent directions of effects (Fig. 4b). This consistency underscores a possible shared epigenetic mechanism across ancestries. These consistent associations encompassed 433 distinct DNAm sites and were observed across 12 phenotypes categorized into six broad groups: lipids, cardiovascular diseases (CVD), anthropometric measurements, metabolic conditions, renal traits, and addiction. Examination of genomic context revealed that most shared associations (529 of 559) involved CpG sites in open sea regions, with fewer sites in shores (30), shelves (24), and islands (15). This distribution suggests potential regulatory elements in traditionally less-studied genomic regions and highlights the importance of comprehensive methylome coverage for discovering functional variations.

To aggregate CpG-level associations into gene-level insights, We mapped CpG sites from all associations to the genes and used Aggregated Cauchy Association Test (ACAT)[69] to detect significant gene-phenotype associations.[70] After Bonferroni correction (P cutoff is 2.7×10^{-6} for AA and EA), we identified 8,670 significant associations for EA population encompassing 30 traits and 7 categories (Supplementary Table 3.2). We also identified 1,109 significant associations for AA population encompassing 17 traits and 7 categories (Supplementary Table 3.1). Among these, 747 associations are significant in both ancestries, and they encompassed 14 traits in 6 categories (Supplementary Table 4). We have made all meQTL results, MWAS models, association results, and optimized codes for replicating all results publicly available to facilitate

future research endeavors (www.gcbhub.org and https://osf.io/gct57/). To illustrate the biological relevance of our findings, we examined Type 2 Diabetes (T2D) in detail.

For Type 2 Diabetes (T2D), we identified 13 genes significantly associated with T2D in both ancestries (Supplementary Table 4). The most significant associations included several established diabetes-related genes with distinct molecular functions: *TCF7L2*, whose association with T2D has been consistently replicated across populations with diverse genetic origins[71] and experimentally validated[72]; *ANK1,* a known T2D susceptibility locus[3] implicated in insulin resistance, a key feature in the development of T2D;[74] and *FTO, which influences T2D risk primarily through BMI and obesity regulation*.[75] We also identified novel associations with potential mechanistic roles: *NKX6-3*, which, along with other specific pancreatic islet β-cell transcription factors, is sensitive to oxidative stress, a condition associated with β-cell dysfunction in both Type 1 and Type 2 diabetes.[76] Previous GWAS studies suggest that *TH* may contribute to T2D susceptibility and insulin resistance. [77,78] Notably, our analysis revealed seven T2D-associated genes unique to the AA population (Supplementary Table 3.1). Among these, *BCKDHA*, encoding the branched-chain alpha-keto acid (BCAA) dehydrogenase (BCKD) subunit, is particularly interesting. This finding connects with previous observations of elevated BCAA levels in both human and animal models of obesity[79,80], and the documented suppression of BCKD complex activity in liver and adipose tissue during obesity[81]. Given BCKD's expression across multiple tissues including monocytes, this ancestry-specific association suggests potential immune system involvement in T2D pathogenesis that may be particularly relevant in AA populations.

Comparison with array-based approaches demonstrated the superior detection power of WGBS. In the EA population, analysis restricted to array-covered sites identified only 1,354 associations using the 450K array (1,105 overlapping with WGBS) and 1,929 associations using the 900K array (1,644 overlapping). The disparity was even more pronounced in the AA population, where arrays detected only 136 (450K) and 200 (900K) associations, with 102 and 153 overlapping WGBS findings, respectively. This limitation of array-based approaches was particularly evident in T2D analysis, where arrays identified only three genes (*TCF7L2*, *FTO*, and *KCNC2*), missing numerous biologically relevant associations detected through WGBS, as discussed earlier.

Concordance with colocalization and MR analysis

To validate the causal nature of our MWAS associations, we performed complementary analyses using Mendelian Randomization (MR)[92] and Bayesian colocalization analyses[93] for AA and EA separately. After Bonferroni correction across all CpG sites and phenotypes (P cutoff is 1.2×10^{-9} for AA and 2.3×10^{-9} for EA), we identified 36,292 likely causal CpG-phenotype pairs in the EA and 1,804 in the AA using MR. The substantial overlap between MR-supported relationships and our MWAS findings—68.6% (21,483/31,334) in EA and 42.3% (1,270/2,998) in AA populations—provides strong evidence for the biological relevance of our identified associations. Notably, among the 537 MWAS associations consistently detected across both ancestries, 89.9% (483) demonstrated causal relationships through MR, suggesting particularly robust cross-ancestry effects.

Complementary Bayesian colocalization analysis revealed shared genetic architecture between MWAS signals and trait associations, with 36.8% (11,517 out of 31,336) of EA and 9.4% (281 out of 2,998) of AA associations showing strong evidence for colocalization (posterior probability (PPH4) exceeding 0.7). The higher colocalization rate in EA likely reflects greater statistical power due to larger sample sizes in both meQTL and GWAS data. Of particular interest, 42.8% (230 out of 537) of cross-ancestry consistent MWAS associations demonstrated colocalization, indicating that a substantial proportion of these associations are likely driven by the same causal variants in both populations. (Supplementary Table 2.1 and 2.2)

Identification of *trans*-meQTL for whole genome DNAm sites across two populations

Given our modest sample sizes (AA: n=160; EA: n=298), we conducted an exploratory analysis of *trans*-meQTL associations to provide preliminary insights into long-range methylation regulation. Using a conservative threshold of 1×10^{-14}. our analysis revealed 1,664,615 trans-meQTL associations involving 159,053 CpG sites (0.6% of all sites tested) in the AA and 2,484,481 *trans*-meQTL associations involving 103,768 CpG sites (0.4% of all sites tested) in the EA. The average density of associations differed slightly between populations, with 0.06 and 0.10 trans-meQTL mappings per CpG site in AA and EA populations, respectively. Notably, we observed minimal overlap between populations, with only 1,952 shared CpG sites showing *trans*-meQTL associations. The majority of associations were population-specific, with 98.8% (157,101) unique to AA and 98.1% (101,816) unique to EA populations, though these patterns should be interpreted cautiously given the sample size limitations.

Initial characterization of these putative *trans*-meQTL associations revealed several patterns that warrant further investigation in larger cohorts. We observed a negative correlation between MAF and effect size when MAF > 0.05 (AA: correlation $= -0.47,\ P < 2.2 \times 10^{-16}$, EA: correlation $= -0.55,\ P < 2.2 \times 10^{-16}$). Interestingly, variants with low MAFs (~0.01) showed unexpectedly small effect sizes (Fig. 5a). The relationship between effect size and genomic distance showed stronger distance-dependent attenuation compared to *cis*-meQTLs (Fig. 5b), while effect sizes of shared trans-meQTLs demonstrated high concordance between populations (correlation $= 0.97, P < 2 \times 10^{-16}$, Fig. 5c).

Discussion

We present the first comprehensive analysis of genetic influences on monocyte DNA methylation across African American (AA, n=160) and European American (EA, n=298) populations using whole genome bisulfite sequencing (WGBS). This study makes several key contributions to the field of epigenetics. First, by analyzing over 25 million CpG sites, we identified approximately ten times more *cis*-meQTL associations than previous studies[11], providing unprecedented resolution of methylation regulation. Second, our population-specific DNAm imputation models, validated using the MVP dataset of 41 complex traits, demonstrate how integrative MWAS can complement GWAS in identifying associated CpG sites and genes and potential therapeutic targets. Third, our use of purified monocytes enables precise characterization of cell-type-specific methylation patterns, critical for understanding immune-mediated diseases.

Our cross-ancestry analysis reveals important insights into population-specific methylation regulation. The AA population showed higher *cis*-h^2 of DNAm levels compared to EA, and about 30% of *cis*-meQTLs detected in the AA population were rare or absent in Europeans, while only 8% of EA-specific *cis*-meQTLs were rare in Africans. These findings underscore the critical importance of diverse ancestry representation in epigenetic studies and suggest potential mechanisms for population-specific disease susceptibility.

The comprehensive coverage (over 25 million CpG sites) of WGBS enabled discovery of numerous methylation markers beyond conventional array-based approaches, yielding novel biological insights. Integration of these CpG sites to the genes revealed gene-trait associations. For example, while we validated established Type 2 diabetes-associated genes like TCF7L2 and FTO, the comprehensive coverage also led to identification of novel candidate genes. Notably, we identified AA-specific associations such as *BCKDHA*, which may help explain ancestry-specific differences in disease susceptibility.

Our findings complement recent work showing that the DNAm landscape is primarily determined by cell-type-specific regulatory programs[94]. While cell-type identity provides the broad framework for methylation patterns, we demonstrate that genetic variation plays a crucial role in fine-tuning these patterns, even within a homogeneous cell population like monocytes. This suggests a hierarchical model of methylation regulation where cell-type programs establish the baseline pattern, which is then modulated by genetic factors in a population-specific manner. Our analysis of purified monocytes provides direct evidence for this model, revealing how genetic variation can influence methylation patterns within a defined cellular context.

Several limitations of our study warrant discussion and suggest directions for future research. First, the relatively modest sample sizes, particularly in the AA cohort, may have led to underestimation of *cis*-meQTL associations and limited our ability to detect *trans*-effects, which typically have smaller effect sizes and require larger samples to detect[95]. However, it is important to note that our study is still one of the largest studies to date, especially for the comparison of EA and AA populations. Second, the male-predominant composition of both our cohort (all males) and the MVP (approximately 90% male) dataset may mask sex-specific methylation patterns. While DNAm patterns have been shown to be relatively consistent between genders[96], subtle differences may still exist.[97] Third, while our focus on purified monocytes provides valuable insights into immune cell regulation, future studies should expand to other cell types and integrate transcriptomic data to build a more comprehensive picture of epigenetic regulation. This foundation in monocyte methylation patterns sets the stage for broader investigation of tissue-specific methylation landscapes and their roles in health and disease.

Figures

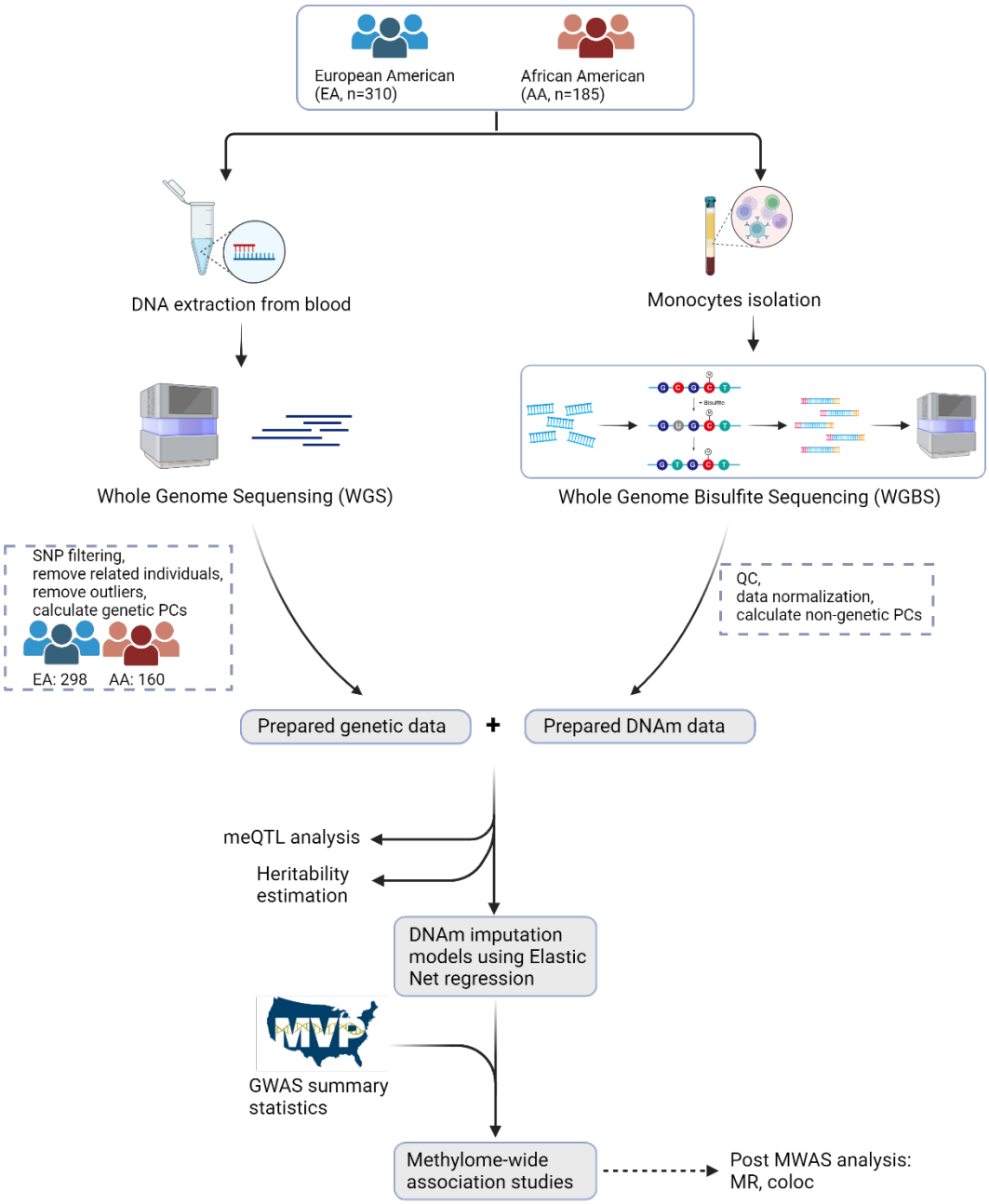

Fig.1 Overview of the data collection, preparation, and analysis pipeline.

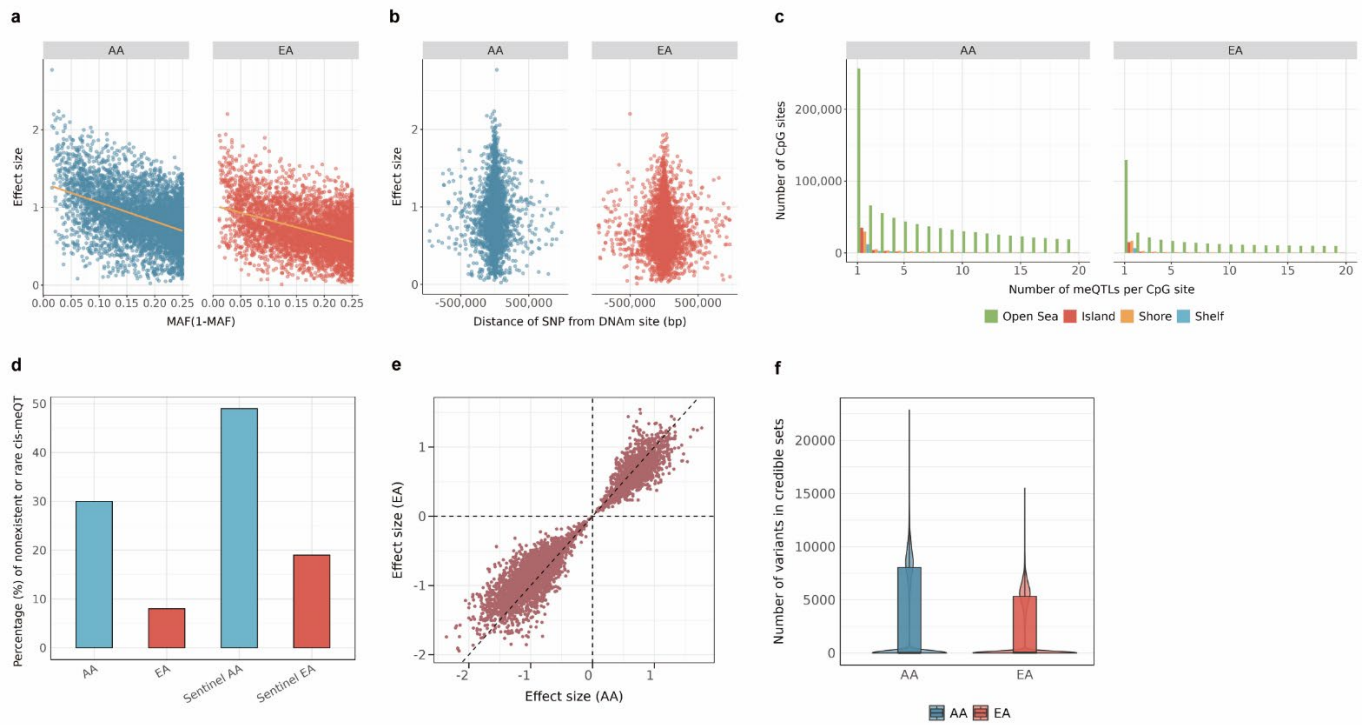

Fig.2 *cis*-meQTL and fine-mapping analysis across AA and EA populations. (a) Relationship between *cis*-meQTL effect sizes and minor allele frequencies (MAF(1-MAF)), with regression trends shown in orange. **(b)**Effect size distribution as a function of distance between SNP and DNAm site in AA (left) and EA (right) populations, measured in base pairs. **(c)** Genomic distribution of CpG sites with *cis*-meQTLs. The bar chart categorizes CpG sites based on the number of *cis*-meQTLs identified, with annotations for four genomic regions: CpG islands, which are regions with a high CpG density; CpG shores, located within 2 kb of islands; CpG shelves, extending an additional 2 kb from shores; and Open Sea, which represents areas more distal to the islands, shores, and shelves. Each region is color-coded. **(d)** Percentage of nonexistent or rare *cis*-meQTLs across ancestries. It displays the proportion of identified *cis*-meQTLs that are absent or rare within the contrasting population from the 1000 Genomes Project. The four bars represent *cis*-meQTL associations for AA, EA, sentinel *cis*-meQTLs for AA and EA, respectively. Sentinel *cis*-meQTL is defined as the most significant SNPs for each CpG site. **(e)** Correlation pf *cis*-meQTL effect sizes between AA and EA population. **(f)** Fine-mapping results showing the distribution of credible set sizes from SuSiE analysis for CpG sites with significant *cis*-meQTLs in both populations. Box plots show median, quartiles, and 1.5×IQR whiskers; violin width indicates data density.

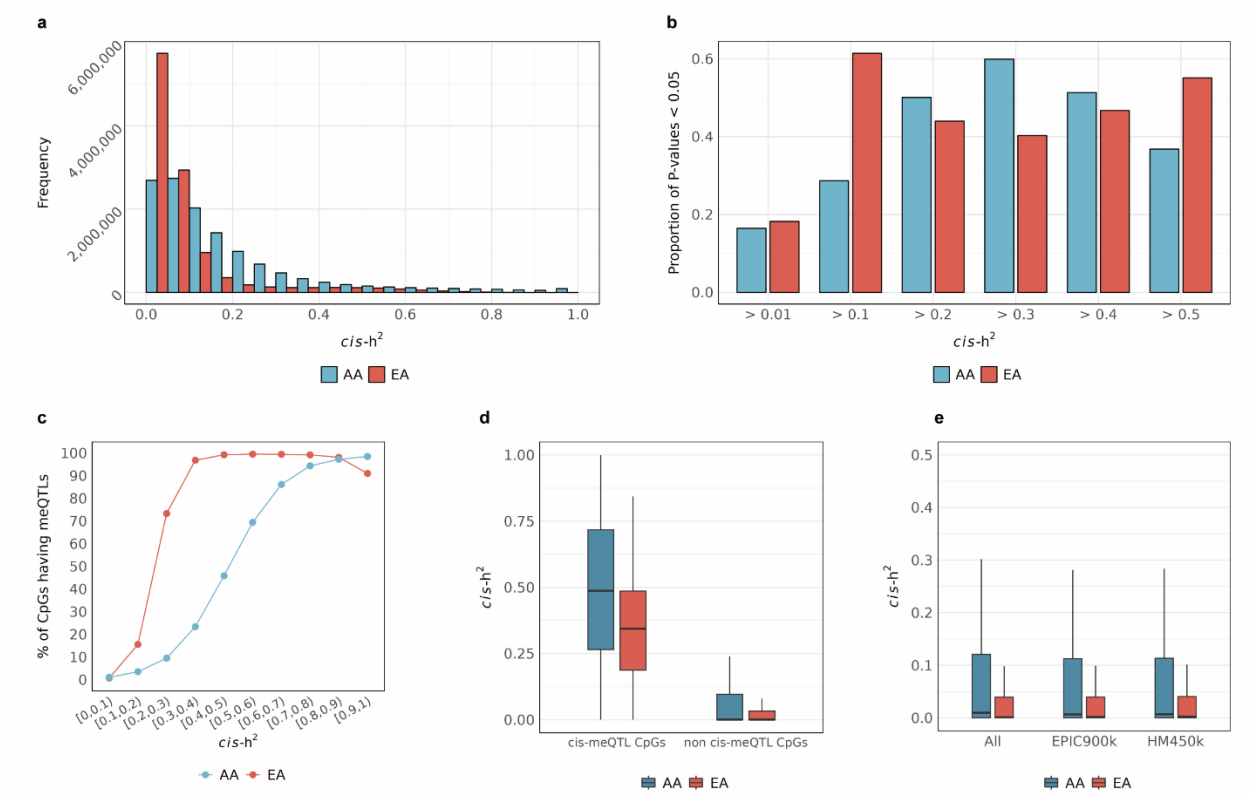

Fig.3 *cis*-h^2. (a) Distribution of *cis*-heritability estimates across CpG sites with *cis*-h^2 > 0.01in AA and EA populations. **(b)** Proportions of CpG sites achieving statistical significance ($P < 0.05$) across varying *cis*-h^2 thresholds, demonstrating the relationship between heritability magnitude and detection power. **(c)** Relationship between CpG site *cis*-h^2 and the presence of significant *cis*-meQTL associations, illustrating the concordance between heritability estimates and genetic regulation. **(d)** Comparative analysis of *cis*-h^2 distribution among CpG sites with and without significant *cis*-meQTL in AA and EA ancestries. **(e)** Comparison of *cis*-h^2 distributions between WGBS data and conventional array-based platforms (EPIC900K and HM450K), highlighting the comprehensive coverage of WGBS. Bos plots show median, quartiles, and 1.5×IQR whiskers.

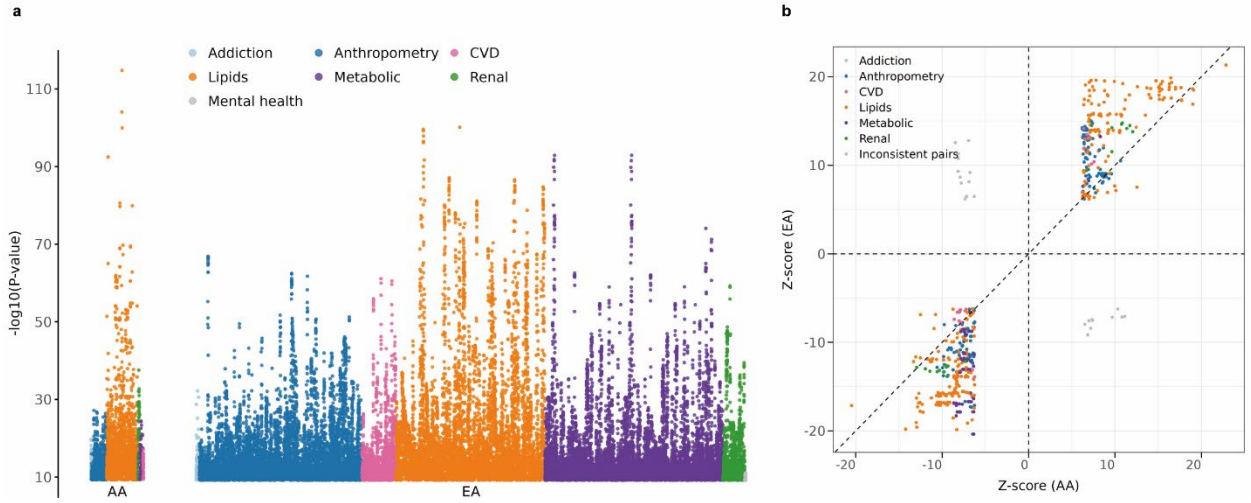

Fig.4 MWAS results for 41 phenotypes in the MVP dataset. **(a)** Distribution of significant MWAS associations after Bonferroni correction (P cutoff is 4.9×10^{-10} for AA and 6.6×10^{-10} for EA) across different trait categories. **(b)** Cross-ancestry comparison of effect sizes (Z-scores) for CpG-phenotype associations significant in both populations. Colored points represent consistent directional effects across ancestries, while grey points indicate discordant directions. Different colors denote distinct phenotype categories.

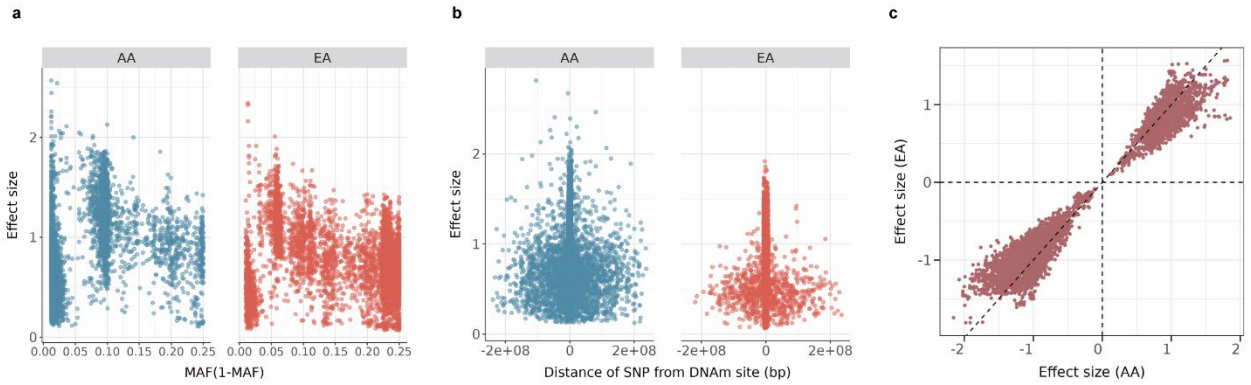

Fig.5 Characterization of *trans*-meQTL effects across African American (AA) and European American (EA) populations. (a) Relationship between *trans*-meQTL effect sizes and MAF(1-MAF), demonstrating the impact of allele frequency on long-range methylation. Orange regression lines indicate overall trends. **(b)** Effect sizes distribution as a function of genomic distance between SNP and DNAm site in AA (left) and EA (right). **(c)** Correlation of *trans*-meQTL effect sizes between AA and EA populations for shared associations, demonstrating the consistency of trans effect across ancestries.

Methods

Study subjects

We recruited a total of 495 male subjects, aged 37.2 ± 8.8 years (mean ± SD), from our ongoing Tulane Louisiana Osteoporosis Study cohort. The study population consisted of 310 (62.6%) self-identified EA and 185 (37.4%) AA. The mean height and weight of the subjects were 175.8 ± 7.1 cm and 83.4 ± 16.2 kg, respectively. Within the cohort, 355 (71.7%) subjects reported smoking, 342 (69.1%) reported alcohol consumption, and 368 (74.3%) reported regular exercise. The racial/ethnic composition and lifestyle factors were representative of the general population in the study area (Supplementary Table 1).

We excluded individuals with preexisting conditions relevant to bone mass development and immune system, including: (1) cerebral vascular disease, (2) diabetes mellitus, except for easily controlled, noninsulin dependent cases, (3) chronic renal or liver failure, (4) chronic lung disease, (5) chronic obstructive pulmonary disease, (6) any metabolic or inherited bone diseases (e.g., hyper/hypoparathyroidism, Paget's disease, osteomalacia, osteogenesis imperfecta, and hypochondrogenesis), (7) collagen disorder (e.g., rheumatoid arthritis, except for minor cases that involve only hand joint and wrist), (8) chronic gastrointestinal disease, (9) alcohol abuses, (10) treatment with corticosteroid or anticonvulsant therapy for more than 6 months duration, (11) antibiotic usage, (12) gastroenteritis, (13) major surgeries, (14) intercontinental travel in the past 3 months, (15) autoimmune or autoimmune-related diseases (e.g., multiple sclerosis), (16) immune-deficiency conditions (e.g., HIV infection), (17) haematopoietic and lymphoreticular malignancies (e.g., leukaemias), (18) active periods of asthma, or (19) influenza, infected within one week of recruitment. All qualified individuals signed an informed consent document before any data and biosample collection. The study was approved by the Tulane University Institutional Review Board (IRB #: 10-184088).

Whole Genome Sequencing (WGS)

DNA for WGS was extracted from the blood using the Gentra Puregene Blood Kit (Qiagen, USA). Concentration and quality of all the extracted DNA were assessed using Nanodrop 1000 and the samples were kept at -80 °C until further use. Briefly, 300 ng genomic DNA was used as input. The workflow for library preparation consists of DNA Nanoballs (DNBs) generation through ligation-mediated polymerase chain reaction (LM-PCR), single-strand separation, cyclization, and rolling circle amplification procedures. Two sequencing depths were utilized for this study: 130 samples were sequenced to an average depth of 22x, while another subset was sequenced to an average depth of 15x. The WGS was conducted using DNBSEQ-500™ sequencing technology platform (BGI Americans Corporation, Cambridge, MA, USA), with 350 bp paired-end reads in length. Each sample's cleaned and aligned data was mapped to the human reference genome (GRCh38/hg38) using Burrows-Wheeler Aligner (BWA, v0.7.12) software[98]. For accurate variant calling, we adhered to the recommended Best Practices ,for variant analysis using the Genome Analysis Toolkit (GATK, v4.0.3)[99,100]. HaplotypeCaller of GATK was employed to identify genomic variations, and the variant quality score recalibration (VQSR) was applied to obtain high-confident variant calls.[99,100]

Isolation of Monocytes, their Genomic DNA, and Total RNA

In the present study, we focused specifically on peripheral blood monocytes (PBMs), which can act as osteoclast precursors and play important roles in regulating bone metabolism.[101] Briefly, peripheral blood mononuclear cells (PBMCs) were firstly separated from ~60 ml freshly collected peripheral blood, by a density gradient centrifugation method using Histopaque-1077 (Sigma-Aldrich, USA). The PBMCs were washed repeatedly with 2 mM EDTA in PBS, before being dissolved in 0.5% BSA and 2 mM EDTA in PBS. PBMs were then isolated from the PBMCs with a Monocyte Isolation Kit II (Miltenyi Biotec Gmbh, Bergisch Glagbach, Germany) according to the manufacturer's protocol. The kit depleted unwanted cells (such as T and B cells) from PBMCs, leaving PBMs free of the surface-bound antibody and beads with minimum disturbance. The isolated PBMs were visually checked for purity and counted under microscope. The genomic DNA used for WGBS and total RNA used for RNA-seq were extracted from the freshly isolated PBMs with the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen, USA) following the manufacturer's protocol and kept at -80 °C until further use.

Whole Genome Bisulfite Sequencing (WGBS)

DNAm profiles were determined by WGBS according to previously published protocols.[102] Briefly, 100 ng genomic DNA isolated from PBMs was fragmented by sonication using a Bioruptor (Diagenode, Belgium) to a mean size of approximately 250 bp, followed by the blunt-ending, dA addition to 3'-end, and adaptor ligation. The ligated libraries were bisulfite converted using the EZ DNA Methylation-Gold kit (Zymo Research Corp, USA). The WGBS was conducted on DNAs extracted from isolated PBMs at an average read depth of 20x using HighSeq 4000 Illumina platform (BGI Americans Corporation, Cambridge, MA, USA).

Data filtering includes removing adaptor sequences, contamination, and low-quality reads from raw reads. Raw reads were also excluded if the number of unknown bases exceeded 10% or if the ratio of bases with a quality score less than 20 exceeded 10%, in order to obtain high-quality and cleaned data. The cleaned and aligned data was mapped to the human reference genome (GRCh37/hg19) using BSMAP[103], which was later converted to GRCh38/hg38 using UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). The methylation level of a CpG dinucleotide was determined by the ratio of the number of methylated reads to the total number of reads covering a particular cytosine site.

Bulk transcriptomics (RNA-seq) and cell deconvolution

For the RNA sequencing experiment, RNA integrity was assessed using the Agilent Technologies 2100 Bioanalyzer. Libraries for RNA-seq were prepared following Illumina's TruSeq-stranded-total-RNA-sample preparation. Briefly, 500 ng RNA was used as input. The workflow consists of rRNA removal, cDNA generation, and end repair to generate blunt ends, A-tailing, adaptor ligation and PCR amplification. The libraries were pooled and diluted to 2 nM in EB buffer and then denatured using the Illumina protocol. The denatured libraries were diluted to 10 pM by pre-chilled hybridization buffer and loaded onto Illumina NovaSeq 6000 sequencing system (LC Sciences, Hangzhou, China) using a paired-read recipe according to the manufacturer's instructions.

Quality control analysis and quantification of the sequencing library were performed using Agilent Technologies 2100 Bioanalyzer High Sensitivity DNA Chip. Paired-ended sequencing was performed. Cutadapt[104] was used to remove the reads that contained adaptor contamination, low quality bases, and undetermined bases. The sequence quality was verified using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Bowtie2[105] and HISAT2[106] were used to map read to the genome of *Homo sapiens* (v96; GRCh38/hg38). The mapped reads of each sample were assembled using StringTie[107]. Then, all transcriptomes from the samples were merged to reconstruct a comprehensive transcriptome using perl scripts and gffcompare (https://github.com/gpertea/gffcompare/). After the final transcriptome was generated, StringTie was used to estimate the expression levels of all transcripts in transcripts per million (TPM).[107]

Cell deconvolution analysis was conducted using the Estimated the Proportion of Immune and Cancer (EPIC v1.1.7)[108], which utilized its embedded circulating immune cells (BRef) as the reference profile to estimate the fraction of mRNA contribution to the bulk by each cell type. The reference profile, purified from PBMs or whole blood, includes B cells, CD4+ T cells, CD8+ T cells, monocytes, neutrophils, and natural killer cells. In addition to the default list of signature genes, 10 additional user-selected cell marker genes were incorporated to enhance the deconvolution of cell proportions by EPIC. These added cell marker genes are *CD3D/E/G*, *CD4*, *CD8A*, *CD14*, *CD19*, *CD86*, and *FCGR3A/B*.

Data preparation

Genotype data

We followed the GoDMC pipeline[109] for the genotype data processing. Each study performed quality control on genotype data for chromosomes 1-22. SNPs that failed the Hardy-Weinburg equilibrium ($P < 10^{-6}$), and had an MAF < 0.01 were removed. Then we removed the duplicated SNPs and SNPs with mismatched alleles and recoded indel alleles to I and D. Next, we calculated the first 10 genetic PCs on the SNPs extracted from HapMap3 SNPs without long-range linkage disequilibrium (LD) (MAF > 0.2). Ancestry outliers that deviated by 7 s.d. from the mean were removed. After outlier removal, we recalculated genetic PCs for use in subsequent analyses. To identify relatedness in unrelated datasets, we calculated genome-wide average identity by state using PLINK1.90. Participants with identity by state > 0.125 were removed. As a result, we had 160 and 298 samples left in AA and EA, respectively.

DNAm data normalization

To transform our DNAm data into a distribution that more closely approximates a Gaussian distribution, we applied a rank-based inverse normal transformation (INT)[50]. This method comprised two steps. In the first, the observations are transformed onto the probability scale using the empirical cumulative distribution function (ECDF). In the second, the observations are transformed onto the real line, as Z-scores, using the probit function.

Co-variates

To ensure robust analysis and account for potential confounding factors, we implemented a comprehensive covariate adjustment strategy. First, we adjusted for basic demographic and lifestyle factors including age, BMI, smoking status, and alcohol consumption. Second, although

we used PBMs with high overall purity (mean=98.8%, range=90.1-99.9%), we recognized the importance of accounting for potential cellular heterogeneity. Therefore, we included the proportions of three cell types showing the most variation (B cells, monocytes, and neutrophils) as covariates in our analysis. Finally, to control for population structure and technical variation, we incorporated ten genetic principal components and ten nongenetic principal components. The nongenetic principal components were derived from a principal component analysis of the 20,000 most variable DNAm sites.

meQTL analyses

We performed a comprehensive analysis of all *cis*- and *trans*-meQTL mappings using the MatrixEQTL[52] package in R on 25 million normalized DNAm sites in African (n = 160) and European (n = 298) ancestries, separately. We defined the *cis*-regions to be within ±1 Mb of the CpG sites. The DNAm data was firstly regressed out the covariates defined in the previous section. Then, for each DNAm CpG sites j, the residual value, y_{ji}, was regressed against each SNP k:

$$y_{ji} = \alpha_{jk} + \beta_{jk} x_{ki} + \epsilon_{jki}$$

where genotype values x_{ki} were standardized to have a mean of zero and a standard deviation of one, α_{jk} was the intercept term, and β_{jk} was the effect estimate of each SNP k on each residualized CpG site j. We used 1×10^{-8} as the p-value cutoff for identification of significant *cis*-meQTL mappings and used 1×10^{-14} for *trans*-meQTL mappings.

Fine-mapping

To ascertain the potentially causal variants influencing DNAm, we conducted fine-mapping on *cis*-SNPs associated with each CpG site. This was specifically focused on 12,706,905 and 11,041,146 CpG sites linked with at least one *cis*-meQTL in AA and EA populations, respectively. Utilizing SuSiE[57] from susieR package in R for this analysis, we were able to deduce single effect components or credible sets for each CpG site and its corresponding variants. These credible sets carry a 95% likelihood of encompassing at least one variant exerting a nonzero causal impact. We limited the number of credible sets to a maximum of ten (L = 10), which is based on the hypothesis that up to ten variants could potentially regulate a single CpG site.

***Cis*-h^2**

We employed Genome-wide Complex Trait Analysis[58] (GCTA, version 15) to estimate the h^2 – the proportion of phenotypic variance attributable to genetic factors – for each CpG site among the total of 25 million analyzed. Again, *cis*-regions are defined to be within ±1 Mb of the CpG sites. This estimation was performed using SNPs exclusively from *cis*-regions. GCTA operates by constructing a genetic relationship matrix (GRM) from SNP data, which encapsulates the degree of genetic similarity between pairs of individuals in the study. We then utilized this GRM within a restricted maximum likelihood (REML) analysis framework to estimate h^2. This approach allows for the separation of the phenotypic variance into components attributed to genetic variance (captured by the SNPs in the *cis*-regions) and residual variance. By focusing on *cis*-regional SNPs, our analysis specifically targets the genetic contribution to DNAm variance at each CpG site, providing insights into the genetic underpinnings of epigenetic modifications. To validate and enhance the robustness of our results, we also conducted a comparative analysis of the *cis*-h^2 of CpG sites within our WGBS dataset against those listed from two platforms: MethylationEPIC

Infinium v2.0 that covers over 935,000 CpG methylation sites[59] (900K) and Infinium HumanMethylation450 Beadchip[60] (450K).

DNAm imputation model

In our study, we focused on CpG sites exhibiting a *cis*-h^2 threshold above 0.01 and associated with a minimum of 10 *cis*-SNPs. A 500-kp window was adopted around each target CpG site to enhance computational efficiency during model training. For the genotype data, following others[68,110], we omitted SNPs with a MAF under 1%, SNPs with ambiguous strand orientation, those containing insertions or deletions, and SNPs absent in the LD reference panel derived from the 1000 Genomes Project[56]. Our DNAm imputation model was constructed using a penalized regression approach, integrating methylation and genotype data within the defined *cis*-regions. Let Y be the n-dimensional vector representing the methylation data for a particular CpG site, and let $\boldsymbol{X}$ be the $n \times k$ matrix representing the genotype data for k *cis*-SNPs associated with this CpG site:

$$Y = \boldsymbol{X}w + \varepsilon$$

where w is a $k \times 1$ vector of effect size to be estimated, and ε is the random noise with a mean of zero. The objective function $f(w)$ for a penalized regression to estimate w is:

$$f(w) = \frac{(Y - \boldsymbol{X}w)'(Y - \boldsymbol{X}w)}{N} + J_\lambda(w)$$

where $J_\lambda(w)$ represents a penalty term that regularizes the coefficients to prevent overfitting. Here, we used Elastic-net[64] penalty, which combined both L1 and L2 regularization terms:

$$J_\lambda(w) = \lambda(\alpha \sum |w_j| + \frac{1-\alpha}{2} \sum w_j^2)$$

where λ is the tuning parameter that controls the overall strength of the penalty, α is the elastic-net mixing parameter, which is 0.5 in this study, that determines the trade-off between L1 penalty (lasso) and L2 penalty (ridge). λ is chosen via cross-validation.

The performance of the imputation models was evaluated by predictive R^2, the squared Pearson correlation coefficient between genetically predicted and directly measured DNAm data in nested cross validation[111]. We only considered models with R^2 greater than 0.01 and had a corresponding set of more than 10 *cis*-acting SNPs in the subsequent analysis[47].

MWAS association testing in MVP database

In this section we would like to test the association between predicted methylation levels and trait of interest. In our study, when only summary-level GWAS data of phenotypes were available, we adopted the methodology outlined in existing studies[112,113] to mitigate the discrepancies that often arise from employing LD matrices from reference panels, which may not accurately reflect the LD structure inherent in the GWAS data. In our MWAS association test, we estimated the effect size $\hat{\gamma}$ and its variance between the DNAm and phenotype by:

$$\hat{\gamma} = \frac{\hat{w}'Z/\sqrt{n_s}}{\sigma_r} \text{ and } \widehat{Var}(\hat{\gamma}) = \left(\frac{1}{n_s} + \frac{1}{n_r}\right)\hat{\gamma}^2 + \frac{\zeta^2}{n_s\sigma_r}$$

Where Z is the vectors of z-scores from GWAS, $\hat{w}$ is the estimated weights from the DNAm imputation model, n_s is the sample size of the GWAS data, $\sigma_r = \frac{(G_r\hat{w})'(G_r\hat{w})}{n_r}$, $\zeta^2 = 1 - \frac{2\hat{w}'Z\hat{\gamma}}{\sqrt{n_s}} + \sigma_r\hat{\gamma}^2$, G_r is the standardized genotype matrix of the population reference panel, and n_r is the sample size.

We performed this MWAS analysis to 41 phenotypes in the MVP[49] database where GWAS summary statistics were available for both African and European ancestries. Supplementary Table 5 summarized the information of these GWAS summary statistics. In summary, the MVP GWAS summary statistics were obtained from dbGaP (study accession: phs001672.v3.p1). Prior to conducting association tests, we implemented quality control on GWAS summary statistics. These included the exclusion of duplicate records, verification and correction of strand orientation for alleles, and the removal of ambiguous alleles, ensuring the integrity and reliability of our association analysis. We also excluded the CpG sites with the number of non-zero weights from DNAm imputation models smaller than 10 to ensure the reliability of the study.

Post-MWAS analyses

We compared our MWAS results with MR analysis in line with the methodology outlined by Zhao, et al.[114]. We selected a set of *cis*-meQTLs as the MR instrumental variables. We first restricted our analysis to the common set of variants that were shared by *cis*-meQTLs and GWAS summary statistics. To avoid the potential issue of collinearity, we removed *cis*-meQTLs strongly correlated with index genetic variants ($r^2 > 0.001$) by applying LD clumping. Then we excluded those who are not significant ($P > 10^{-8}$) *cis*-meQTLs. Second, we applied the Steiger filter[115] to exclude instrumental variables with potential reverse causality. After selecting the instrumental variables, we applied either the Wald ratio (when only one instrumental variable was available) or inverse variance weighting[116] to test the causal link between DNAm and phenotype.

We also conducted Bayesian colocalization analyses[93] on the significant CpG sites identified in our MWAS and MR analysis to estimate the posterior probability that the protein and phenotype shared the same causal variant, using summary-level *cis*-pQTLs and GWAS data. Specifically, we used the coloc R package (v5.2.2), with its default setups in the coloc function, to estimate the posterior probability of both protein and phenotype being influenced by the same causal variant (i.e., the PPH4). We chose PPH4 > 0.7 as the threshold. DNAm satisfying this threshold would suggest a shared causal variant for the *cis*-meQTLs and GWAS associations.

ACKNOWLEDGEMENTS

This work is partially supported by grants from the NIH (U19AG055373, R01AG061917, R01AR069055, P20GM109036, R01CA263494). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Buniello, A. *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. **47**, D1005–D1012 (2019).
2. The GTEx Consortium *et al*. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
3. Sun, B. B. *et al*. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
4. Liggett, S. B. *et al*. A polymorphism within a conserved β_1-adrenergic receptor motif alters cardiac function and β-blocker response in human heart failure. *Proc. Natl. Acad. Sci.* **103**, 11288–11293 (2006).
5. Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D. & Cox, L. A. The need for multi-omics biomarker signatures in precision medicine. *Int. J. Mol. Sci.* **20**, 4781 (2019).
6. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
7. Unnikrishnan, A. *et al*. The role of DNA methylation in epigenetics of aging. *Pharmacol. Ther.* **195**, 172–185 (2019).
8. Lu, A. T. *et al*. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303 (2019).
9. Marioni, R. E. *et al*. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).

10. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).

11. Min, J. L. *et al*. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet*. **53**, 1311–1321 (2021).

12. Feingold, E. A. *et al*. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**, 636–640 (2004).

13. Bernstein, B. E. *et al*. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol*. **28**, 1045–1048 (2010).

14. Stunnenberg, H. G. *et al*. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).

15. Moss, J. *et al*. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun*. **9**, 5068 (2018).

16. Investigators, A. The atherosclerosis risk in communit (ARIC) study: design and objectives. *Am. J. Epidemiol*. **129**, 687–702 (1989).

17. Dawber, T. R., Meadors, G. F. & Moore, F. E. Epidemiological Approaches to Heart Disease: The Framingham Study. *Am. J. Public Health Nations Health* **41**, 279–286 (1951).

18. Shang, L. *et al*. meQTL mapping in the GENOA study reveals genetic determinants of DNA methylation in African Americans. *Nat. Commun*. **14**, 2711 (2023).

19. Darrell 1, C. G. A. R. N. T. source sites: D. U. M. S. M. R. 1 F. A. 2 B., J. 3 4 5, E. U. V. M. E. G. 3 4 5 B. D. J. 5 6 M. M. G. 3 O. J., Norman 8, H. F. H. M. T. 7 L., Oliver 11, M. A. C. C. A. K. 9 A. Y. W. 10 B. & Michael 13, U. of C. S. F. V. S. 12 B. M. 13 P. Comprehensive genomic

characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

20. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
21. de Moura, M. C. *et al.* Epigenome-wide association study of COVID-19 severity with respiratory failure. *EBioMedicine* **66**, (2021).
22. Chu, A. Y. *et al.* Epigenome-wide association studies identify DNA methylation associated with kidney function. *Nat. Commun.* **8**, 1286 (2017).
23. Zillich, L. *et al.* Epigenome-wide association study of alcohol use disorder in five brain regions. *Neuropsychopharmacology* **47**, 832–839 (2022).
24. Barbu, M. C. *et al.* Methylome-wide association study of antidepressant use in Generation Scotland and the Netherlands Twin Register implicates the innate immune system. *Mol. Psychiatry* **27**, 1647–1657 (2022).
25. Huan, T. *et al.* Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.* **10**, 4267 (2019).
26. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
27. Lin, H. *et al.* Methylome-wide association study of atrial fibrillation in Framingham Heart Study. *Sci. Rep.* **7**, 40377 (2017).
28. Swedish Schizophrenia Consortium *et al.* High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biol.* **16**, 291 (2015).

29. Shi, J. *et al.* Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat. Commun.* **5**, 3365 (2014).

30. Dai, J. Y. *et al.* DNA methylation and cis-regulation of gene expression by prostate cancer risk SNPs. *PLoS Genet.* **16**, e1008667 (2020).

31. Schulz, H. *et al.* Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. *Nat. Commun.* **8**, 1511 (2017).

32. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

33. Fortin, J.-P., Triche Jr, T. J. & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).

34. Jakubzick, C. V., Randolph, G. J. & Henson, P. M. Monocyte differentiation and antigen-presenting functions. *Nat. Rev. Immunol.* **17**, 349–362 (2017).

35. Gordon, S. & Taylor, P. R. Monocyte and macrophage heterogeneity. *Nat. Rev. Immunol.* **5**, 953–964 (2005).

36. Tao, L. *et al.* Single-cell RNA sequencing reveals that an imbalance in monocyte subsets rather than changes in gene expression patterns is a feature of postmenopausal osteoporosis. *J. Bone Miner. Res.* zjae065 (2024).

37. Zhou, Y., Deng, H.-W. & Shen, H. Circulating monocytes: an appropriate model for bone-related study. *Osteoporos. Int.* **26**, 2561–2572 (2015).

38. Marventano, I. *et al.* A complex proinflammatory role for peripheral monocytes in Alzheimer's disease. *J. Alzheimers Dis.* **38**, 403–413 (2014).

39. Yan, P. *et al.* Peripheral monocyte–derived cells counter amyloid plaque pathogenesis in a mouse model of Alzheimer's disease. *J. Clin. Invest.* **132**, (2022).

40. Kawanaka, N. *et al.* CD14+,CD16+ blood monocytes and joint inflammation in rheumatoid arthritis. *Arthritis Rheum.* **46**, 2578–2586 (2002).

41. Pamukcu, B., Lip, G. Y. H., Devitt, A., Griffiths, H. & Shantsila, E. The role of monocytes in atherosclerotic coronary artery disease. *Ann. Med.* **42**, 394–403 (2010).

42. Chittezhath, M. *et al.* Molecular profiling reveals a tumor-promoting phenotype of monocytes and macrophages in human cancer progression. *Immunity* **41**, 815–829 (2014).

43. Fraser, H. B., Lam, L. L., Neumann, S. M. & Kobor, M. S. Population-specificity of human DNA methylation. *Genome Biol.* **13**, R8 (2012).

44. Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* **47**, 1282–1293 (2015).

45. Mozhui, K., Smith, A. K. & Tylavsky, F. A. Ancestry dependent DNA methylation and influence of maternal nutrition. *PloS One* **10**, e0118466 (2015).

46. Greenbaum, J. *et al.* A multiethnic whole genome sequencing study to identify novel loci for bone mineral density. *Hum. Mol. Genet.* **31**, 1067–1081 (2022).

47. Wu, L. *et al.* An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk. *Nat. Commun.* **11**, 3905 (2020).

48. Baselmans, B. M. *et al*. Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet*. **51**, 445–451 (2019).

49. Gaziano, J. M. *et al*. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol*. **70**, 214–223 (2016).

50. Pain, O., Dudbridge, F. & Ronald, A. Are your covariates under control? How normalization can re-introduce covariate effects. *bioRxiv* 137232 (2017).

51. Dahl, A., Guillemot, V., Mefford, J., Aschard, H. & Zaitlen, N. Adjusting for principal components of molecular phenotypes induces replicating false positives. *Genetics* **211**, 1179–1189 (2019).

52. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

53. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol*. **18**, 77 (2017).

54. Zhang, J. *et al*. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet*. **54**, 593–602 (2022).

55. Cavalcante, R. G. & Sartor, M. A. Annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).

56. Consortium, 1000 Genomes Project. *A Global Reference for Human Genetic Variation*. *Nature* vol. 526 68 (Nature Publishing Group, 2015).

57. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

58. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

59. Noguera-Castells, A., García-Prieto, C. A., Álvarez-Errico, D. & Esteller, M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* **18**, 2185742 (2023).

60. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *New Genomic Technol. Appl.* **98**, 288–295 (2011).

61. Chen, B. H. & Zhou, W. mLiftOver: Harmonizing Data Across Infinium DNA Methylation Platforms. *bioRxiv* 2024–03 (2024).

62. Satomi, K., Ichimura, K. & Shibahara, J. Decoding the DNA methylome of central nervous system tumors: An emerging modality for integrated diagnosis. *Pathol. Int.* **74**, 51–67 (2024).

63. Vavourakis, C. D., Herzog, C. M. & Widschwendter, M. Devising reliable and accurate epigenetic clocks: choosing the optimal computational solution. (2023).

64. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).

65. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

66. Mai, J., Lu, M., Gao, Q., Zeng, J. & Xiao, J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. *Commun. Biol.* **6**, 899 (2023).

67. Yao, S. *et al.* A transcriptome-wide association study identifies susceptibility genes for Parkinson's disease. *Npj Park. Dis.* **7**, 79 (2021).

68. Wu, L. *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* **50**, 968–978 (2018).

69. Liu, Y. *et al.* ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).

70. Wu, C., Bradley, J., Li, Y., Wu, L. & Deng, H.-W. A gene-level methylome-wide association analysis identifies novel Alzheimer's disease genes. *Bioinformatics* **37**, 1933–1940 (2021).

71. del Bosque-Plata, L., Martínez-Martínez, E., Espinoza-Camacho, M. Á. & Gragnoli, C. The role of TCF7L2 in type 2 diabetes. *Diabetes* **70**, 1220–1228 (2021).

72. Savic, D. *et al.* Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Res.* **21**, 1417–1425 (2011).

73. Yan, R. *et al.* A novel type 2 diabetes risk allele increases the promoter activity of the muscle-specific small ankyrin 1 gene. *Sci. Rep.* **6**, 25105 (2016).

74. Harder, M. N. *et al.* Type 2 diabetes risk alleles near BCAR1 and in ANK1 associate with decreased β-cell function whereas risk alleles near ANKRD55 and GRB14 associate with decreased insulin sensitivity in the Danish Inter99 cohort. *J. Clin. Endocrinol. Metab.* **98**, E801–E806 (2013).

75. Freathy, R. M. *et al.* Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes* **57**, 1419–1426 (2008).

76. Guo, S. *et al.* Inactivation of specific β cell transcription factors in type 2 diabetes. *J. Clin. Invest.* **123**, 3305–3316 (2013).

77. Saxena, R. *et al.* Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. *Am. J. Hum. Genet.* **90**, 410–425 (2012).

78. Hanson, R. L. *et al.* 1642-P: Colocalization Analyses of Genetic Associations of Type 2 Diabetes with DNA Methylation on Chromosome 11p in American Indians. *Diabetes* **69**, (2020).

79. Caballero, B., Finer, N. & Wurtman, R. J. Plasma amino acids and insulin levels in obesity: response to carbohydrate intake and tryptophan supplements. *Metabolism* **37**, 672–676 (1988).

80. Wijekoon, E. P., Skinner, C., Brosnan, M. E. & Brosnan, J. T. Amino acid metabolism in the Zucker diabetic fatty rat: effects of insulin resistance and of type 2 diabetes. *Can. J. Physiol. Pharmacol.* **82**, 506–514 (2004).

81. She, P. *et al.* Obesity-related elevations in plasma leucine are associated with alterations in enzymes involved in branched-chain amino acid metabolism. *Am. J. Physiol.-Endocrinol. Metab.* **293**, E1552–E1563 (2007).

82. Libby, P. & Theroux, P. Pathophysiology of Coronary Artery Disease. *Circulation* **111**, 3481–3488 (2005).

83. Ghattas, A., Griffiths, H. R., Devitt, A., Lip, G. Y. H. & Shantsila, E. Monocytes in Coronary Artery Disease and Atherosclerosis. *J. Am. Coll. Cardiol.* **62**, 1541–1551 (2013).

84. Schlitt, A. *et al.* CD14+CD16+ monocytes in coronary artery disease and their relationship to serum TNF-α levels. *Thromb. Haemost.* **92**, 419–424 (2004).

85. Gil, J. & Peters, G. Regulation of the INK4b–ARF–INK4a tumour suppressor locus: all for one or one for all. *Nat. Rev. Mol. Cell Biol.* **7**, 667–677 (2006).

86. Huang, K. *et al.* Effects of *CDKN2B-AS1* polymorphisms on the susceptibility to coronary heart disease. *Mol. Genet. Genomic Med.* **7**, e955 (2019).

87. Yin, Y. *et al.* SDF-1α involved in mobilization and recruitment of endothelial progenitor cells after arterial injury in mice. *Cardiovasc. Pathol.* **19**, 218–227 (2010).

88. Döring, Y., Pawig, L., Weber, C. & Noels, H. The CXCL12/CXCR4 chemokine ligand/receptor axis in cardiovascular disease. *Front. Physiol.* **5**, 88349 (2014).

89. Runmin, G. *et al.* Genetic variation of CXCR4 and risk of coronary artery disease: epidemiological study and functional validation of CRISPR/Cas9 system. *Oncotarget* **9**, 14077 (2018).

90. Collaboration, I. G. C. E. R. F. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *The Lancet* **379**, 1205–1213 (2012).

91. Chen, C. *et al.* Serum TGF-β1 and SMAD3 levels are closely associated with coronary artery disease. *BMC Cardiovasc. Disord.* **14**, 18 (2014).

92. Zhao, H. *et al.* Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genomics* **2**, (2022).

93. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

94. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).

95. Tsai, P.-C. & Bell, J. T. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int. J. Epidemiol.* **44**, 1429–1441 (2015).

96. Singmann, P. *et al.* Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics Chromatin* **8**, 43 (2015).

97. Grant, O. A., Wang, Y., Kumari, M., Zabet, N. R. & Schalkwyk, L. Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array. *Clin. Epigenetics* **14**, 62 (2022).

98. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).

99. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

100. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

101. Gordon, S. & Taylor, P. R. Monocyte and macrophage heterogeneity. *Nat. Rev. Immunol.* **5**, 953–964 (2005).

102. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).

103. Xi, Y. & Li, W. B. whole genome bisulfite sequence MAPping program. BMC 1027. *Bioinformatics* **10**, 1028 (2009).

104. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).

105. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

106. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

107. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

108. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *elife* **6**, e26476 (2017).

109. Min, J. L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.* **53**, 1311–1321 (2021).

110. Aberg, K. A. *et al.* Convergence of evidence from a methylome-wide CpG-SNP association study and GWAS of major depressive disorder. *Transl. Psychiatry* **8**, 162 (2018).

111. Krstajic, D., Buturovic, L. J., Leahy, D. E. & Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminformatics* **6**, 10 (2014).

112. Xue, H., Shen, X. & Pan, W. Causal Inference in Transcriptome-Wide Association Studies with Invalid Instruments and GWAS Summary Data. *J. Am. Stat. Assoc.* **118**, 1525–1537 (2023).

113. Wu, C., Zhang, Z., Yang, X. & Zhao, B. Large-scale imputation models for multi-ancestry proteome-wide association analysis. *bioRxiv* 2023–10 (2023).

114. Zhao, H. *et al.* Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genomics* **2**, (2022).

115. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).

116. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).