Genome Biology

METHOD

Open Access

HBI: a hierarchical Bayesian interaction model to estimate cell-type-specific methylation quantitative trait loci incorporating priors from cell-sorted bisulfite sequencing data

Youshu Cheng[1,2], Biao Cai[1], Hongyu Li[1], Xinyu Zhang[2,3], Gypsyamber D'Souza[4], Sadeep Shrestha[5], Andrew Edmonds[6], Jacquelyn Meyers[7], Margaret Fischl[8], Seble Kassaye[9], Kathryn Anastos[10], Mardge Cohen[11], Bradley E. Aouizerat[12,13], Ke Xu[2,3*†] and Hongyu Zhao[1,2*†]

†Ke Xu and Hongyu Zhao contributed equally to this work.

*Correspondence: ke.xu@yale.edu; hongyu.zhao@yale.edu

[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT 06511, USA
[2] VA Connecticut Healthcare System, West Haven, CT 06516, USA
Full list of author information is available at the end of the article

Abstract

Methylation quantitative trait loci (meQTLs) quantify the effects of genetic variants on DNA methylation levels. However, most published studies utilize bulk methylation datasets composed of different cell types and limit our understanding of cell-type-specific methylation regulation. We propose a hierarchical Bayesian interaction (HBI) model to infer cell-type-specific meQTLs, which integrates a large-scale bulk methylation data and a small-scale cell-type-specific methylation data. Through simulations, we show that HBI enhances the estimation of cell-type-specific meQTLs. In real data analyses, we demonstrate that HBI can further improve the functional annotation of genetic variants and identify biologically relevant cell types for complex traits.

Keywords: Methylation quantitative trait loci, Cell-type-specific DNA methylation, hierarchical Bayesian interaction model, Cell-sorted methylation sequencing data, Colocalization

Background

DNA methylation (DNAm) is one of the most widely studied epigenetic modifications that capture the cumulative effects of environmental and genetic factors. DNAm regulates cellular differentiation and gene expression and plays a key role in human development and disease etiology [1, 2]. Single-nucleotide polymorphisms (SNPs) associated with DNAm levels are known as methylation quantitative trait loci (meQTLs) [3–6], which capture and represent the complex interplay between the genome and methylome.

To reveal cellular mechanisms for DNAm patterns and their link to complex traits, it is important to study cell-type-specific (CTS) genetic effects on DNAm (CTS-meQTL). For example, SNP rs174548, which is mapped on *FADS1*, a key enzyme in the metabolism of fatty acids, is associated with asthma [1]. At the same time, its effect on methylation at cg21709803 is the strongest in CD8+ T-cells. These results suggest a possible effect of rs174548 on asthma via immune dysregulation and fatty acid metabolism through methylation in CD8+ T-cells [1]. However, most meQTL studies to date have used bulk samples composed of distinct cell types [7–9]. MeQTLs identified from bulk DNAm samples reflect the aggregated genetic effects across all cell types, which provide no insights for genetic regulations in individual cell types. This approach is especially problematic for rare or less abundant cell types. The high cost and technical limitations for both cell sorting and single-cell DNAm approaches hinder the collection of large-scale, CTS methylation profiles, and limit our ability to move meQTL studies from the "bulk level" to the "cell type level."

Given the difficulty in generating large-scale CTS methylome data to directly estimate CTS effects and the broad availability of many bulk methylation datasets, several statistical methods have been developed to infer CTS meQTLs from bulk data. These methods can be classified into two categories. Methods in the first category estimate sample-level CTS DNAm profiles from bulk data in the first step, and then test the associations with outcomes of interest using the deconvoluted data for each cell type. Tensor Composition Analysis (TCA), a frequentist approach in this category, was originally designed to identify CTS differentially methylated CpG sites in epigenome-wide association studies of phenotypes (CTS-EWAS) [10]. There is also a similar algorithm designed for gene expression data [11], named Bayesian MIND (bMIND), which further incorporates information from single-cell RNA sequencing (scRNA-seq) data as a prior to refine the estimation of CTS expression for each bulk sample. bMIND innovatively integrates large-scale bulk data and small-scale CTS expression data from scRNA-seq to estimate CTS expression for large-scale bulk samples. In contrast, methods in the second category are based on an interaction model to test the interaction between cell type fractions and variables of interest without deconvolution. Examples include CellDMC [12], which focuses on the interaction between cell type fractions and phenotypes (CTS-EWAS). Westra et al. also proposed an interaction model to estimate CTS expression quantitative trait loci (CTS-eQTL) [13].

Here we introduce a hierarchical Bayesian interaction model (HBI) to infer CTS meQTLs from bulk methylation data. Our model allows the incorporation of cell-type-specific DNAm data from a relatively small number of samples to improve the performance of HBI. Compared with bMIND, which utilizes Bayesian techniques to infer the posterior mean of sample-level CTS expression (or as easily for methylation), the goal of HBI is instead to infer the posterior mean of CTS genetic effects by placing sparse hierarchical priors on regression coefficients for the interaction terms. In our model, we employ hierarchical double-exponential priors to induce different shrinkage for different variables, which corresponds to the Bayesian adaptive lasso [14]. If cell-type-specific data are available for a small number of samples (e.g., 5–10% of the sample size in bulk data), the algorithm can incorporate this information to further refine the estimates for CTS genetic effects in the larger-scale bulk samples. In our case, cell-sorted Methylation Capture sequencing data (MC-seq) is used to derive CTS DNAm and since it offers the unique advantage of directly measuring CTS methylomes, incorporating strong and robust signals from the MC-seq data will improve the estimation of CTS-meQTLs.

We show in simulations that HBI improves the estimation of CTS genetic effects when compared to other state-of-the-art methods [10–12]. We apply our method to identify cis CTS-meQTL using data from samples in the Women's Interagency HIV Study (WIHS) ($n_{\text{bulk}}=431$, $n_{\text{CTS}}=47$), now the MACS/WIHS Combined Cohort Study (MWCCS) [15]. To demonstrate the utility of our method, we use an independent meQTL dataset derived from CTS methylation data [1] to evaluate the replication of HBI-identified signals. Finally, we perform downstream analyses to improve the annotation of functional genetic variants and to reveal the cellular specificity of complex traits.

Results

Estimation of CTS-meQTLs using HBI

A linear regression framework including interaction terms between genotype/phenotype and estimated cell type fractions has been applied to identify CTS-QTL or CTS- differentially methylated CpGs [12, 13]. Here, based on this idea, we propose HBI to incorporate prior information from CTS DNAm data and to improve the estimation of CTS-meQTLs (Fig. 1). We place hierarchical double-exponential priors on regression coefficients for the interaction terms:

$$\beta_k|\tau_k^2 \sim N(\mu_k, \tau_k^2),$$

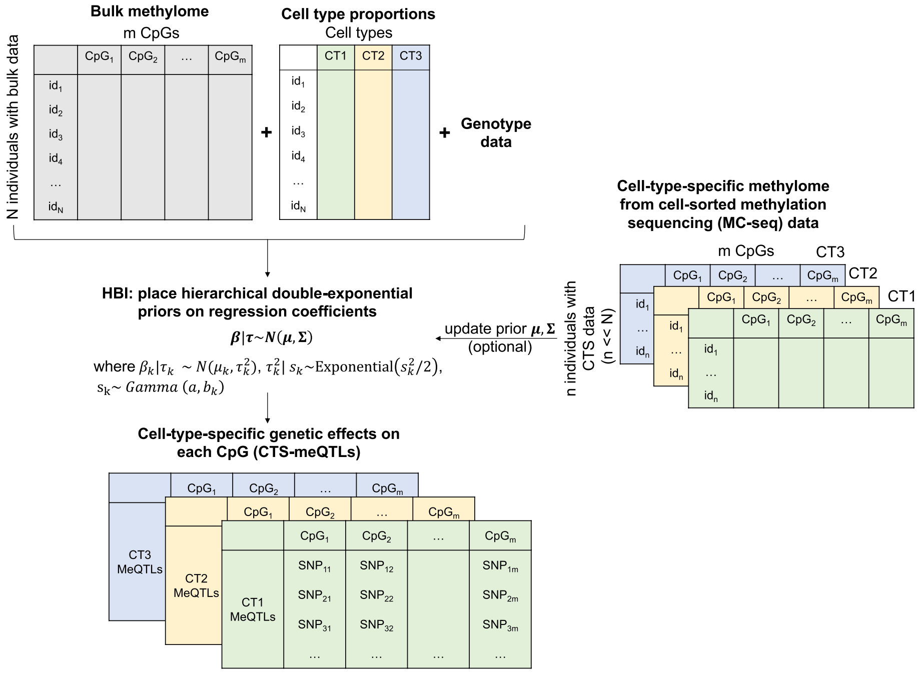

Bulk methylome
m CpGs
N individuals with bulk data
CpG_1 CpG_2 … CpG_m
id_1 id_2 id_3 id_4 … id_N
+
Cell type proportions
Cell types
CT1 CT2 CT3
+
Genotype data
HBI: place hierarchical double-exponential priors on regression coefficients
$\boldsymbol{\beta}|\boldsymbol{\tau}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$
where $\beta_k|\tau_k \sim N(\mu_k, \tau_k^2)$, $\tau_k^2|\, s_k \sim \text{Exponential}(s_k^2/2)$, $s_k \sim Gamma\ (a, b_k)$
update prior $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ (optional)
Cell-type-specific methylome from cell-sorted methylation sequencing (MC-seq) data
m CpGs
CT3 CT2 CT1
n individuals with CTS data (n << N)
id_1 … id_n
Cell-type-specific genetic effects on each CpG (CTS-meQTLs)
CT3 MeQTLs
CT2 MeQTLs
CT1 MeQTLs
SNP_{11} SNP_{12} SNP_{1m}
SNP_{21} SNP_{22} SNP_{2m}
SNP_{31} SNP_{32} SNP_{3m}

Fig. 1 Overview of the hierarchical Bayesian interaction model (HBI) to infer cell type-specific (CTS) meQTLs. With bulk methylation data and cell type proportions (we present an example of three cell types: CT1, CT2, CT3), HBI employs an interaction model with sparse hierarchical priors placed on the regression coefficients for the interaction terms. If the CTS DNA methylation data (in our case, generated by methylation capture-sequencing using cells sorted from PBMC using flow cytometry) are available for a small group of samples, HBI will further incorporate the information into priors to refine the estimates for CTS genetic effects in the larger-scale bulk samples

$$\tau_k^2 | s_k \sim \mathrm{Exp}\left(\frac{s_k^2}{2}\right),$$

where β_k is the regression coefficient on the interaction term between genotype and cell type proportion for the k th cell type. Marginalizing over τ_k^2, β_k conditional on s_k follows a double-exponential distribution:

$$\beta_k | s_k \sim DE(\mu_k, 1/s_k),$$

where parameter s_k controls the degree of shrinkage. If s_k is a fixed value for $k = 1,2,\ldots,K$, each variable will be shrunk to the same degree. Here we model s_k as a hyperparameter to allow for variable-specific penalty:

$$s_k \sim \mathrm{Gamma}(\mathrm{a}, \mathrm{b_k}),$$

where a and b_k are chosen based on empirical experiments (Methods).

In the case where only bulk data are available, we set $\mu_k = 0$ for $k = 1,2,\ldots,K$ and the model would be similar to the adaptive Lasso approach [14, 16]: the regression coefficients for interaction terms are shrunk to 0 and the degree of shrinkage differs for different variables. Such shrinkage helps to take the sparsity of genetic effects into consideration. When the CTS methylomes are available for a small number of samples, we can first get a rough estimate of the genetic effect in the k th cell type $\widehat{\beta}_{k,seq}$ using the small set. Then instead of setting $\mu_k = 0$ and shrinking the coefficient to zero, we can shrink it to a more meaningful value by updating the prior mean μ_k :

$$\mu_\kappa = weight \cdot \widehat{\beta}_{k,seq} + \left(1 - weight\right) \cdot 0,$$

where μ_k is a weighted sum between $\widehat{\beta}_{k,seq}$ (observed results from CTS methylomes) and zero (prior beliefs), while $weight = 1 - p_{adjust}$ and p_{adjust} is the p-value adjusted using the Bonferroni correction (Methods). Similar to other studies that propose weights based on posterior probabilities [17], the weights in our model are assigned based on p-values as p-value is a probability measuring the evidence against the null hypothesis ($\beta_{k,seq} = 0$) [18] and can reflect the stability of the estimator $\widehat{\beta}_{k,seq}$. Intuitively, a small p-value closer to zero indicates $\widehat{\beta}_{k,seq}$ estimated using the CTS DNAm data is relatively strong. In this case, we shrink the coefficient more towards $\mu_k = \widehat{\beta}_{k,seq}$. In contrast, a large p-value closer to one indicates $\widehat{\beta}_{k,seq}$ is not significantly different from zero, and thus we shrink it more towards $\mu_k = 0$.

Along with updating prior means, we can also update prior variances:

$$\boldsymbol{\beta}|\boldsymbol{\tau} \sim \boldsymbol{N}\left(\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \cdots & \rho_{1K}\tau_1\tau_K \\ \rho_{12}\tau_2\tau_1 & \tau_2^2 & \cdots & \rho_{2K}\tau_2\tau_K \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}\tau_K\tau_1 & \rho_{2K}\tau_K\tau_2 & \cdots & \tau_K^2 \end{bmatrix}\right),$$

where ρ_{jk} can be updated as the genetic correlation between cell type k methylation and cell type j methylation, which can also be estimated from the CTS methylomes provided. The prior variance without CTS data can be seen as a special case with all $\rho_{jk} = 0$. Of note, as the detection of genetic effects is always much harder in less abundant cell types, the incorporation of the estimated genetic correlation aims to improve the power in the less abundant cell types by borrowing information from the more abundant cell types.

More details of our model are summarized in "Methods". We note the following key features for HBI: (1) because only a few SNPs among hundreds of SNPs surrounding a CpG may have detectable effects, placing a sharp prior centered at 0 helps to take the sparsity of genetic effects into consideration; (2) local shrinkage parameters s_k and τ_k^2 make the algorithm more flexible: the degree of shrinkage could differ among variables; and (3) priors could be updated to incorporate information from CTS DNAm data, if they are available for a small group of samples.

HBI improves performance in simulations

We evaluated the performance of HBI in estimating CTS-meQTLs through extensive simulations. We considered three scenarios: (1) there are genetic effects only in the major/most abundant cell type; (2) there are genetic effects only in the minor/least abundant cell type; and (3) there are correlated genetic effects in all cell types. We compared HBI to other state-of-the-art methods: bMIND, TCA, and the basic interaction model fitted by ordinary least squares (OLS) (Methods). In each scenario, we assessed the correlation between the estimated and true effect sizes, mean squared error (MSE) between the estimated and true effect sizes, power, and false discovery rate (FDR) as a function of the proportion of causal SNPs.

HBI improved the point estimation of CTS-meQTLs by achieving higher correlation and lower MSE (Fig. 2). For example, in scenario 1 when the proportion of causal SNPs fixed at 10%, the median of correlation across 10 simulations was 0.72 for bMIND with only bulk data (denoted as "bMIND"), 0.71 for bMIND with CTS data incorporated (denoted as "bMIND_CTS-prior"), 0.68 for TCA, 0.77 for basic interaction model, 0.94 for HBI with only bulk data (denoted as "HBI"), and 0.94 for HBI with CTS data incorporated (denoted as "HBI_CTS-prior"). Across all scenarios, HBI generally achieved higher power compared with other methods. We note that in scenario 1, when genetic effects only occur in the most abundant cell type, further incorporating CTS DNAm data to update priors had comparable power to the case without incorporating CTS DNAm data. For example, in scenario 1 when the proportion of causal SNPs fixed at 10%, the median of power across 10 simulations was both 0.65 for HBI with only bulk data (denoted as "HBI") and HBI with CTS data incorporated (denoted as "HBI_CTS-prior"). In contrast, in scenarios 2 and 3, when there were genetic effects in the minor/least abundant cell type, incorporating information from CTS DNAm data helped to improve the power. For example, in scenario 2 when the proportion of causal SNPs fixed at 10%, the median of power across 10 simulations was 0.15 for HBI with only bulk data (denoted as "HBI"), and 0.24 for HBI with CTS data incorporated (denoted as "HBI_CTS-prior"). In each scenario, we varied

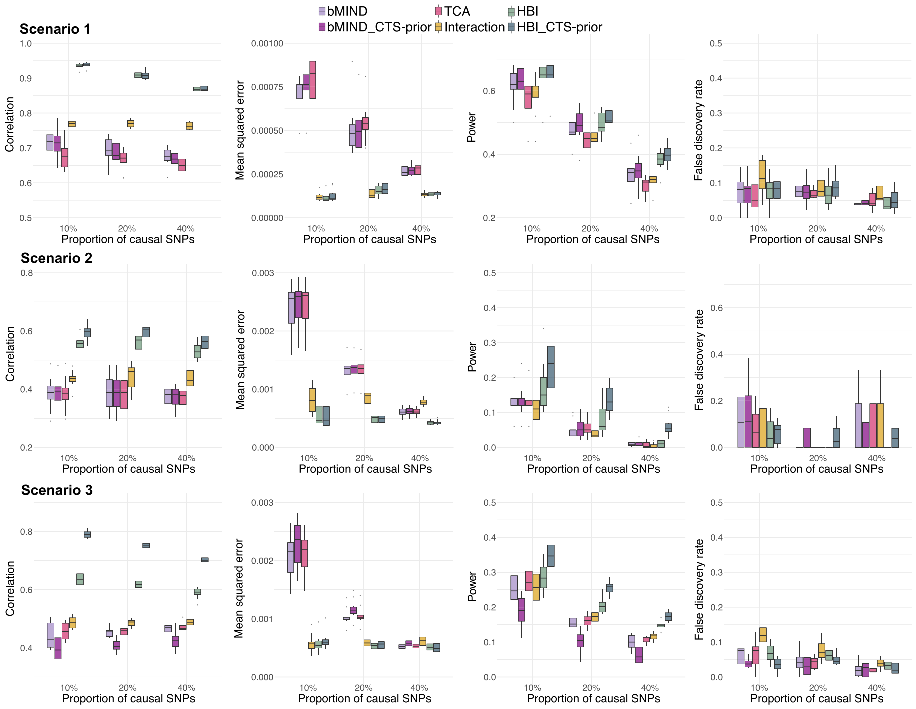


Fig. 2 Comparisons of performance in estimating cell type-specific (CTS)-meQTLs. From top to bottom: scenarios with genetic effects only in the most abundant cell type (Scenario 1), only in the least abundant cell type (Scenario 2), and with correlated genetic effects in all cell types (Scenario 3) are shown. From left to right: correlation between estimated and true effect sizes, mean squared error (MSE) between estimated and true effect sizes, power, and false discovery rate (FDR) as a function of the proportion of causal SNPs. HBI_CTS-prior, bMIND_CTS-prior represent the version of the corresponding methods with CTS DNA methylation data incorporated

the proportion of causal SNPs from 10 to 20% to 40%, to compare the performance of these methods when the genetic effects became more polygenic. As expected, the power for all methods decreased as the proportion of causal SNPs increased. When the overall genetic effect (heritability) was fixed and diluted on a larger number of SNPs, it generally became harder to detect signals. We also note that the performance for bMIND and TCA was a result of fitting conditional models (Methods). In the case of fitting marginal models for bMIND and TCA, we observed inflated FDR, especially in scenarios 1 and 2 when there were genetic effects only in one single cell type (Additional file 1: Fig S1).

Of note, all methods included here relied on cell type proportions, but in reality the biological “ground truth” of the cell type proportions is rarely available and the computationally estimated proportions [19, 20] are used directly, which introduces additional noise. Therefore, to further evaluate the robustness of all methods when “noisy” cell type proportions (random error was added to the true proportions) were given, we repeated the simulation scenario 3 but with noisy proportions inputted for all methods (Methods). With the increase in the noise added to cell type proportions, the correlation and power decreased while the MSE increased (Additional file 1: Fig S2), as expected. HBI was robust in this “noisy” setting by achieving the highest correlation and power and lowest MSE among all the methods considered. Additional results comparing the performance across different allele frequencies and different numbers of SNP-CpG pairs are summarized in Additional file 1: Fig S3-4.

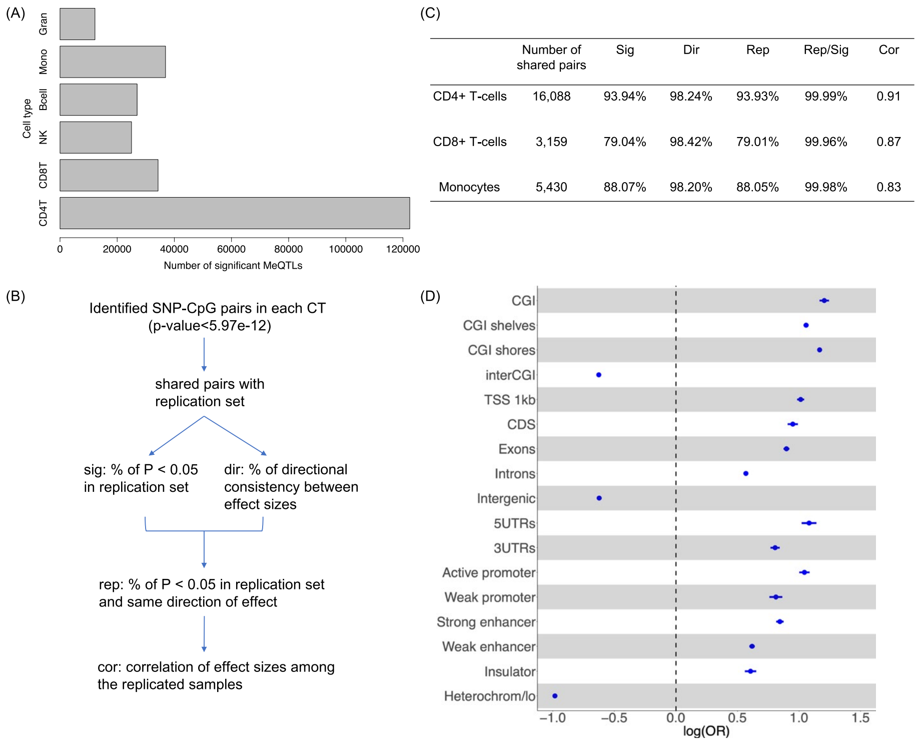
To further investigate the performance of HBI in real data for which we also have estimates of the "ground truth", we applied those methods to the Religious Orders Study and Memory and Aging Project (ROSMAP) gene expression and genotype data and the "ground truth" for QTLs was estimated with its single-cell RNA seq data [21]. We noted that HBI can also be used to identify CTS expression quantitative trait loci (eQTL). Similar patterns were observed, where HBI achieved higher power across different scenarios (Additional file 1: Fig S5).

Genome-wide CTS-meQTLs identification using HBI

To identify genome-wide CTS-meQTLs, we applied HBI to the WIHS cohort with matched genotype data and bulk DNAm data measured in peripheral blood mononuclear cells (PBMC) using the Illumina HumanMethylation EPIC beadchip ($n = 431$) (Additional file 2: Table S1). Furthermore, for a separate group of WIHS participants ($n = 47$), one aliquot of PBMC underwent CTS separation to obtain CD4+ T-cells ($n = 28$), CD8+ T-cells ($n = 28$), or monocytes ($n = 27$). The demographic characteristics of the WIHS participants are described in Additional file 2: Table S1. DNAm from each sorted cell type was profiled using Agilent SureSelectXT Methyl-seq, and the derived priors from these CTS DNAm data were incorporated in HBI (Methods). Significant *cis*-meQTLs were selected as those reaching genome-wide significance level ($p < 6E-12$; Bonferroni correction). The computational time for applying HBI is summarized in Additional file 1: Fig S6, and the median computational time was about 4 min.

HBI identified a total of 122,387 significant meQTLs in CD4+ T-cells, 34,310 in CD8+ T-cells, 25,020 in natural killer cells, 26,972 in B cells, 36,919 in monocytes, and 12,231 in granulocytes (Fig. 3A) (Additional file 3: Table S2). To replicate our identified CTS-meQTLs, we used publicly available data for meQTLs in isolated white blood cell subsets (CD4+ T-cells, CD8+ T-cells, monocytes) ($n = 60$ individuals) [1], and we defined replicated meQTLs as those with $p < 0.05$ and consistent direction of effect in this replication sample (Fig. 3B). We showed that among the shared SNP-CpG pairs in the replication sample, 98.2–98.4% had a consistent direction of effect and 79.0–93.9% were replicated (Fig. 3C). Of note, in all cell types, more than 99% of significant pairs ($p < 0.05$) had a consistent direction of effect (Rep/Sig), indicating a high level of consistency between our results in the WIHS sample and those in the replication sample. We also investigated the replication rates using the version of HBI without priors incorporated from the WIHS participants with CTS data ($n = 47$), and did parallel analyses using SNPs in high LD to increase the number of shared pairs in the replication sample (Additional file 4: Table S3). An additional data with larger sample sizes for meQTLs in isolated blood cells (CD4+ T-cells, monocytes) ($n = 197$ individuals) was also used for replication [22]. The similar pattern was observed for replication rates of HBI (97.31% in CD4+ T-cells, 92.40% in monocytes) (Additional file 4: Table S3).

Integrating annotations of CpG islands (CGI), genomic functional regions, and open chromatin states with our derived CTS-meQTL, we observed that compared with SNPs that are not meQTLs (non-meQTLs), our identified meQTLs across all cell



	Number of shared pairs	Sig	Dir	Rep	Rep/Sig	Cor
CD4+ T-cells	16,088	93.94%	98.24%	93.93%	99.99%	0.91
CD8+ T-cells	3,159	79.04%	98.42%	79.01%	99.96%	0.87
Monocytes	5,430	88.07%	98.20%	88.05%	99.98%	0.83

Fig. 3 Overview of cell type-specific (CTS)-meQTLs identification using the hierarchical Bayesian interaction model (HBI). **A** Bar chart shows the number of HBI-identified meQTLs in each cell type (p<6E−12).**B** Flow chart indicates the replication of HBI identified CTS-meQTLs in an independent dataset for meQTLs in isolated white blood cell subsets (CD4+ T-cells, CD8+ T-cells, monocytes). **C** Table summarizes the replication results. **D** Functional enrichment for meQTLs across all cell types in CpG island (CGI) regions, gene body regions, and gene regulatory regions. The logarithm of odds ratio (OR) with 95% confidence interval is presented. TSS 1 kb:<1 kb upstream of the transcription start site (TSS); CDS: coding sequence; UTR: untranslated exon region; Heterochrom/lo: regions that exhibit heterochromatic or heterochromatin-like characteristics; CD4T: CD4+ T-cells; CD8T: CD8+ T-cells; NK: natural killer cells; Mono: monocytes; Gran: granulocytes

types were enriched in important regulatory regions, such as active promotors and strong enhancers (Fig. 3D). Conversely, our meQTLs were depleted in regions with few active genes, including intergenic regions and regions with heterochromatic characteristics, as previously reported [7]. Of note, meQTLs identified in each cell type exhibited similar functional enrichment patterns and are summarized in Additional file 1: Fig S7.

Using QIAGEN Ingenuity Pathway Analysis (IPA) to perform pathway enrichment analyses of genes mapped by the significant meQTLs in each cell type [23], we found that the antigen presentation pathway was significant in multiple cell types: CD4+ T-cells (p=2.95E−05), CD8+ T-cells (p=1.12E−09), B cells (p=9.55E−06), natural killer cells (p=7.94E−07) and monocytes (p=1.41E−10) (Additional file 5: Table S4). Other identified pathways included the pulmonary fibrosis idiopathic signaling pathway in CD4+ T-cells (p=9.33E−06), the multiple sclerosis signaling pathway and the IL-15 production pathway in CD8+ T-cells (p=9.55E−07 and p=3.63E−05, respectively). These significant pathways indicated that the identified CTS-meQTLs by HBI might play a role in regulating immunity-related functions and activities.

CTS meQTLs colocalize with risk variants for complex traits

While for most meQTLs the direct impact on complex traits has not been widely reported [24], there have been studies showing that some meQTLs are associated with complex traits and may identify underlying pathways and mechanisms related to diseases [7, 8, 25]. To systematically identify potential associations between meQTLs and complex traits, we applied HyPrColoc (Hypothesis Prioritization for multi-trait Colocalization) [26] to perform an meQTL-GWAS colocalization analysis in each cell type. We integrated the HBI-identified CTS-meQTLs with 57 GWAS datasets in four categories of blood cell counts, cardiometabolic, immune, and allergy [27].

A total of 2972 significant meQTL-GWAS colocalizations (posterior probability for colocalization (PPFC) > 0.50) were identified across all GWASs and cell types (Additional file 6: Table S5A-F). Taking a further look into the number of meQTL-GWAS colocalizations per trait across all cell types, we found that GWAS traits in the category of blood cell counts had a larger number of colocalizations compared with traits in other categories (Additional file 1: Fig S8). This abundance of colocalizations was expected as the *cis*-meQTLs were identified in cell types from whole blood. To further illustrate how the meQTL-GWAS colocalizations could differ across cell types, we summarize one example in Fig. 4. The variant rs2395178 in the *HLA-DRA* gene was identified as a CD8+ T-cell-specific meQTL for cg00886432 ($p = 5.46\text{E}-12$). As expected, we observed that rs2395178 showed a stronger correlation with DNAm in participants with a high abundance of CD8+ T-cells (Fig. 4A). Meanwhile, our colocalization analyses revealed that rs2395178 was colocalized between methylation at cg00886432 in CD8+ T-cells and type I diabetes (T1D) (PPFC = 0.9802) (Fig. 4B), while no significant colocalization was observed in other cell types. Of note, polymorphisms at the *HLA-DQ* and *HLA-DR* regions have been recognized as the major genetic determinants of T1D [28]. Taken together, these results suggest that integrating DNA methylome and genome data may help link *HLA-DR* gene function in CD8+ T-cells to T1D.

MeQTL-GWAS colocalizations exhibit enrichment in trait-relevant cell types

To further investigate meQTL associations with traits in multiple cell types, we performed enrichment analyses to study if the meQTL-GWAS colocalizations for each trait were enriched in certain cell types. Specifically, for each trait we defined the enrichment score in one cell type as the ratio between the percentage of colocalized GWAS-meQTLs in that cell type and the percentage of meQTLs in that cell type (see Methods). As the absolute number of colocalizations in each cell type was largely driven by the number of identified meQTLs in that cell type and cannot be compared directly, here the enrichment score was defined as the ratio between two percentages, which allowed us to compare this value across different cell types. We further excluded granulocytes due to the low number of colocalizations identified across traits (Additional file 1: Fig S8B), which indicated the less stable signals identified in this least abundant cell-type. We also evaluated the enrichment score for meQTLs derived at the bulk PBMC level (Additional file 6: Table S5G) to further evaluate whether CTS-meQTLs can reveal more cell-specific information.

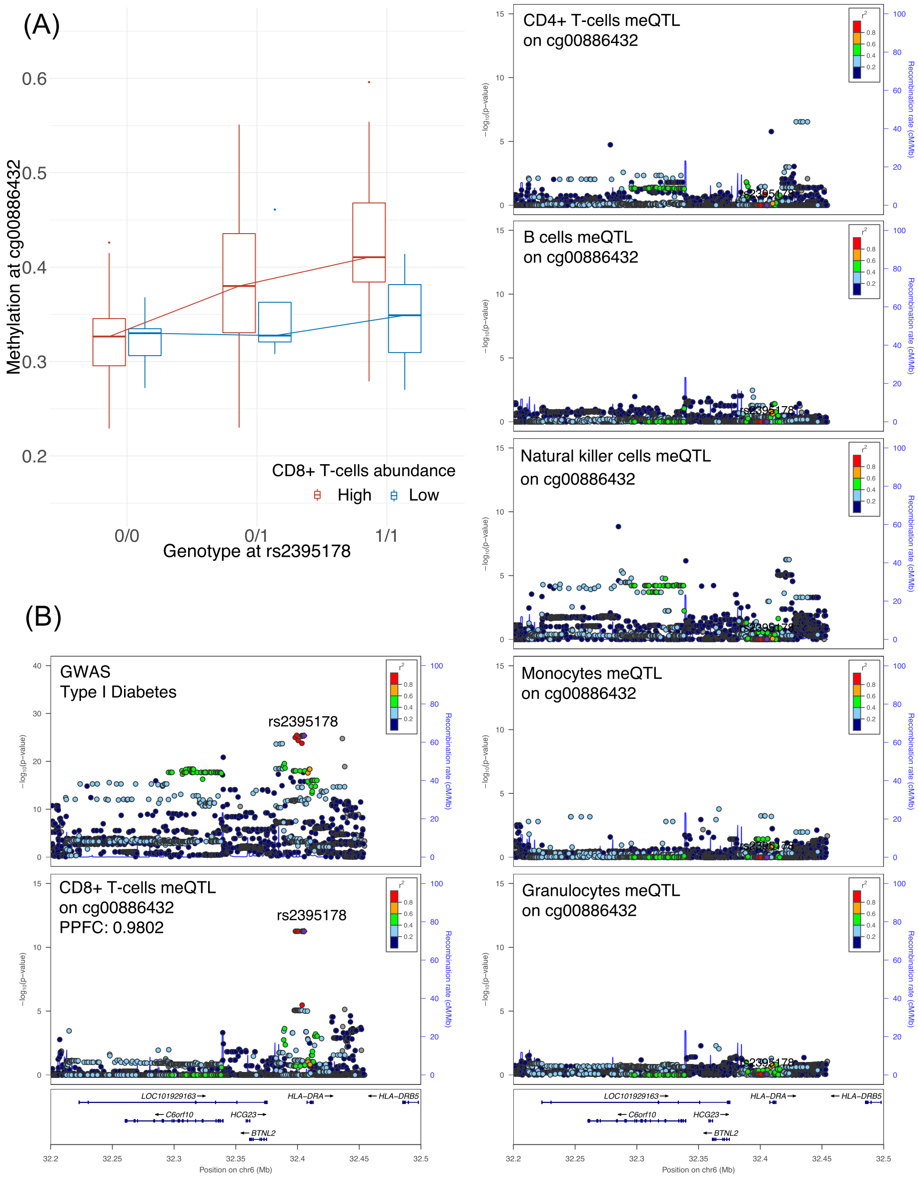


Fig. 4 Example of the rs2395178-cg00886432 locus and colocalization results with type I diabetes (T1D). **A** An association plot for the rs2395178-cg00886432 locus, separated into individuals with high and low abundance of CD8+ T-cells (above and below the median, respectively). The *y* axis shows methylation beta-values, while the *x* axis shows genotypes. **B** LocusZoom plots for the association of rs2395178 (mapped to *HLA-DRA*) with phenotypes/molecular traits. Panels illustrate the association of the SNP with GWAS T1D, cg00886432 meQTL signal in CD8+ T-cells, CD4+ T-cells, B cells, natural killer cells, monocytes, and granulocytes. The genetic variant rs2395178 was identified as a colocalized SNP between T1D and cg00886432 meQTL signal in CD8+ T-cells (posterior probability for colocalization (PPFC) is shown)

We summarize the traits with colocalizations enriched in at least one cell type in Fig. 5A, which we listed out in Additional file 7: Table S6A. To evaluate whether the enrichment results matched existing biological knowledge, we performed heritability enrichment analyses across the same GWAS traits using GenoSkyline-Plus [29], which could be viewed as an independent tool to identify biologically relevant cell types for complex traits. We found that our significant findings generally agreed with the results

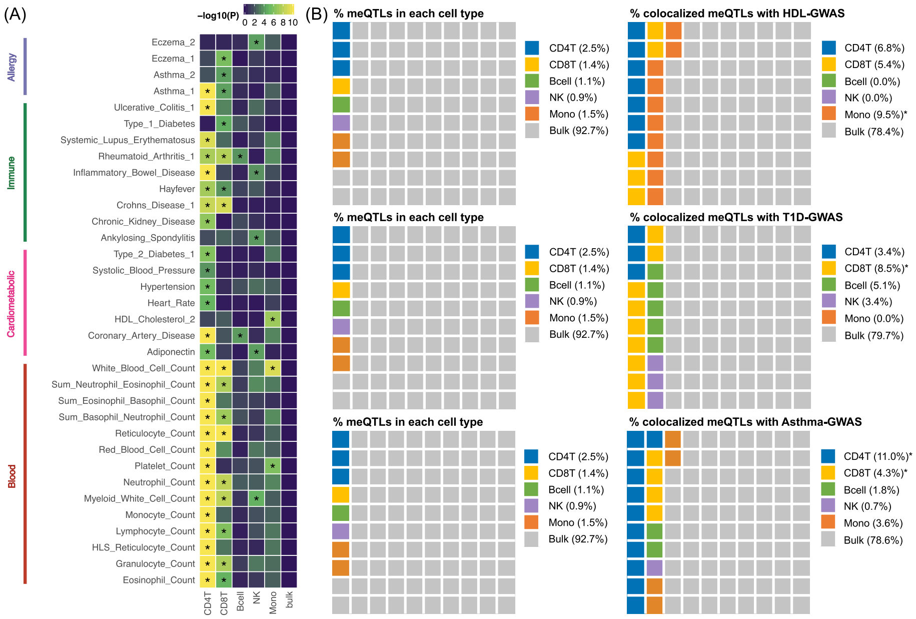


Fig. 5 Enrichment analyses for MeQTL-GWAS colocalizations. **A** Colocalization enrichment results across six cell types for traits with colocalizations enriched in at least one cell type. Asterisks highlight significance after Bonferroni correction. **B** Examples of colocalization enrichments in three traits. From left to right: the percentage of meQTLs covered by each cell type, and the percentage of colocalized meQTLs in that cell type. GWAS: genome-wide association studies; CD4T: CD4+ T-cells; CD8T: CD8+ T-cells; NK: natural killer cells; Mono: monocytes; bulk: a mixture of cell types from peripheral blood mononuclear cells (PBMC)

of heritability enrichment analyses: 85.2% of our identified cell types with enriched colocalizations were replicated by GenoSkyline-Plus (Additional file 7: Table S6B). This indicates that the colocalizations between HBI-identified CTS-meQTLs and GWASs do help to reveal biologically relevant cell types for complex traits. For example, high-density lipoprotein cholesterol (HDL) had colocalization enrichment in monocytes: monocytes only covered 1.5% of the total meQTLs but accounted for 9.5% of the colocalized meQTLs (enrichment = 6.32; $p = 1.23\text{E}-07$) (Fig. 5B). Of note, GenoSkyline-Plus also identified heritability enrichment in monocytes for HDL ($p = 1.31\text{E}-06$) (Additional file 7: Table S6B). In T1D, we identified colocalization enrichment in CD8+ T-cells (enrichment = 6.15; $p = 1.95\text{E}-05$) (Fig. 5B), while the heritability for this trait was also enriched in CD8+ T-cells ($p = 4.77\text{E}-02$) (Additional file 7: Table S6B). This finding is further supported by the evidence that autoreactive CD8+ T-cells play a fundamental role in the progression of T1D by the destruction of pancreatic beta cells [30]. In addition, asthma was enriched in colocalizations derived from CD4+ T-cells (enrichment = 4.34; $p = 5.14\text{E}-19$) and CD8+ T-cells (enrichment = 3.10; $p = 4.91\text{E}-05$) meQTLs, which was also replicated by GenoSkyline-Plus ($p = 3.43\text{E}-03$ and $p = 3.92\text{E}-05$, respectively) (Additional file 7: Table S6B). Interestingly, the association between meQTLs and asthma has been investigated by Hawe et al., who also employed colocalization and reported a shared causal variant rs174548 for methylation at cg21709803 in CD8+ T-cells and asthma [1]. Here our colocalization results in CD8+ T-cells replicated and extended their findings by identifying a nearby risk variant rs174587 (PPFC = 0.858) (Additional file 6: Table S5B), which impacts both DNAm at cg21709803 and asthma.

From Fig. 5B, we show that more bulk meQTLs were identified than CTS meQTLs, which was consistent with simulations (Additional file 1: Fig S9), but we observed no colocalization enrichments (Fig. 5A). This suggests that the meQTLs identified in bulk tissue are a mixture of signals from different cell types, thus masking the CTS information. Altogether, those results suggest that our identified CTS meQTLs can provide more insight into the cellular specificity of complex traits and aid the characterization of trait etiology.

Discussion

We have developed HBI to infer CTS meQTLs from bulk methylation data, with priors derived from CTS methylation data in a small group of samples. As far as we are aware, our model is the first one to integrate large-scale bulk DNAm data and small-scale CTS DNAm data to estimate CTS-meQTLs. We show through simulations that HBI improves the estimation of CTS genetic effects. Applying our method to samples contributed by participants from the WIHS cohort [15], we systematically characterized the genome-wide SNP-CpG associations in multiple cell types of PBMCs. Through colocalization and enrichment analyses, we demonstrate the utility of HBI to improve the annotation of functional genetic variants and enhance the understanding of the cellular specificity of complex traits.

We considered extensive simulation scenarios to compare the performance of different methods in detecting CTS QTLs. As TCA and bMIND were initially developed to detect differentially expressed or differentially methylated signals between comparison groups (e.g., cases versus controls) [10, 11], the differential effect was on a single phenotype of interest in their simulations. In contrast, in our simulations the genetic effects on a CpG were distributed across a number of SNPs and each SNP carried a small effect. Therefore, our simulation aims to detect all causal SNPs, which is more challenging than detecting the association with one single phenotype, and the simulation results may offer a more comprehensive evaluation of the performance of different methods to detect CTS-QTLs than those considered in other studies [10, 11, 31].

The simulation results show that all methods had decreased performance in scenario 2 (the least abundant cell type harbored genetic effects). This is not surprising as the information from rare cell types is in general more limited in a bulk sample, and thus the statistical instability for estimating signals in rare cell types is much larger than that in abundant cell types. In this case, incorporating CTS DNAm information did help to alleviate this problem; we show that HBI with CTS information incorporated into priors (i.e., HBI_CTS-prior) was more powerful than other methods. Specifically, to improve the power to infer meQTL in rare cell types by borrowing information from more abundant cell types, we used the small group with CTS methylation data to estimate ρ_{jk}, the genetic correlation between cell type k methylation and cell type j methylation, and incorporated the estimated genetic correlation into the prior variance. As cell-sorted MC-seq data offer the unique advantage of directly measuring CTS methylomes, incorporating strong and robust signals from such data improves the estimation of CTS-meQTLs, especially in rare cell types.

We observed inflated FDR when fitting the marginal model for bMIND and TCA. As discussed by the authors of TCA, deconvoluted CTS methylation profiles are expected to be correlated among different cell types [32], which results in false discoveries in the non-causal cell type when using the marginal test model. To mitigate this problem, the TCA authors advised applying a marginal conditional test to account for the other cell types [32], which was also used in our simulations. The developers of bMIND also proposed an alternative testing procedure, in which they only detected the top cell type with the minimal differential expressed (DE) *p*-value within a gene. This testing procedure was supported by some single-nucleus RNA-sequencing (snRNA-Seq) studies [33, 34] which reported that most CTS DE genes are only differentially expressed in one single cell type. In contrast, our meQTL analyses aimed to capture not only meQTLs that are specific in one single cell type but also meQTLs that are shared across multiple or all cell types. Previous studies have reported the existence of a substantial proportion of meQTLs that exhibit this shared pattern across diverse cell types [1, 35]. Therefore, in our simulations, we did not adopt the alternative testing procedure proposed by bMIND. Instead, the marginal conditional test model was fitted for TCA and bMIND to control FDR and to provide a fair comparison with HBI. Although HBI generally outperformed other methods in our QTL-based simulations, we note that the deconvolution step in TCA and bMIND can output sample-level CTS profiles which enable other sample-level analyses (e.g., CTS co-expression networks), while methods based on the interaction model, like HBI, do not have this additional output.

In real data applications, the use of stringent statistical thresholds and independent replication datasets [1] enables the identification of CTS-meQTLs with high confidence and generalizability. Specifically, our identified meQTLs were supported by high replication rates in isolated CD4+ T-cells, CD8+ T-cells, and monocytes. We highlighted one example of the potential of our approach to identify biologically relevant cell types for complex traits. The colocalization analyses between meQTLs and GWASs for T1D identified several SNPs in the HLA region. For example, rs2395178 in *HLA-DRA* was identified as a CD8+ T-cell specific meQTL for cg00886432. *HLA-DRA* belongs to the human leukocyte antigen (HLA) complex family, which plays an important role in antigen presentation and immune defense [36], and is well known as the major genetic determinant of T1D [28]. Our colocalization analyses revealed that rs2395178 was shared between methylation at cg00886432 in CD8+ T-cells and T1D (PPFC = 0.9802). There has also been evidence for the contribution of CD8+ T-cells to the progression of T1D by the destruction of pancreatic beta cells [30]. Altogether, our downstream analyses helped to explain the relationship between this SNP-CpG locus and T1D, especially in CD8+ T-cells. We also noted that the colocalized signals might be driven by haplotype structures or LD, as multiple studies identified strong signals for T1D at nearby SNPs in the HLA region (e.g., rs9271365 mapped to *HLA-DQA1*) [37–39], which are close to but not identical to the colocalized SNPs that we identified. Similarly, for other complex traits, we also identified biologically relevant cell types through meQTL-GWAS colocalization, and our findings strongly agreed with heritability enrichment analyses [29]. We also investigated the computational time of HBI in this real data application. The computational time increased linearly with the number of SNP-CpG pairs in one CpG. The median of pairs in one CpG was 1624 and the median of computational time was 4.05 min.

We acknowledge several limitations of our study. First, similar to other methods [10–12], our model depends on cell type proportions W_k as an input. Currently, the method described by Houseman et al. [19, 40] was widely applied to estimate this W_k for blood samples. There are also some efforts to quantify W_k for non-blood samples (i.e., brain samples) [41]. However, those estimated proportions are used directly to approximate the biological ground truth, which introduces additional error. For example, we identified significant meQTLs for granulocytes, which theoretically should have been filtered out in PBMC. This may result from technical issues including granulocyte contamination during PBMC processing [42], and the inaccuracy in the estimated granulocyte proportions. To alleviate this issue, we plan to extend the statistical model to estimate the cell type proportions and incorporate the uncertainty in the estimated proportions at the same time. This approach will broaden the applications as we will not rely on other algorithms to estimate cell type proportions, and help to obtain more accurate results as the uncertainty in the estimated proportions is considered and adjusted. Second, we used an meQTL dataset obtained from experimentally isolated white blood cells [1] as the "gold standard" to replicate our findings. However, their CpG data were generated using the Illumina Human Methylation 450 K BeadChip while our results were based on the Illumina Infinium Methylation EPIC BeadChip. Additionally, only a total of 11.2 million SNP-CpG pairs that were preselected in their bulk meQTL analysis were available. As a result, not all our significant results were represented in their database, even though we utilized SNPs in high LD to increase the number of shared pairs. Third, in colocalization enrichment analysis, we did not observe significant results for some traits (i.e., no significant enrichments for heart attack or stroke). The potential reasons might be that our CTS-meQTLs were only derived from white blood cells and may not cover the causal cell types, and the small number of identified colocalizations in some traits may impact our results as well. Therefore, re-applying HBI on a dataset with a larger sample size and a wider range of causal cell types will help to obtain a more powered and complete CTS-meQTL catalog [35].

Conclusions

HBI provides a statistical strategy to leverage bulk data and CTS MC-seq data to improve the estimation of CTS meQTLs. Through in-depth real data analyses, we linked the methylome and genome data and illustrated the power of HBI to identify biologically relevant cell types for complex traits. We believe that HBI can have wide applications in identifying CTS meQTLs and annotating functional genetic variants.

Methods

Statistical model

We model the relationship between methylation level at one CpG and one SNP as:

$$M = \sum_{c=1}^{C} \gamma_c X_c + \sum_{k=1}^{K} \alpha_k W_k + \sum_{k=1}^{K} \beta_k (W_k \cdot G) + \epsilon, \tag{1}$$

where M is the bulk methylation level, W_k is the proportion of the k th cell type, G is the genotype of the SNP (the number of alternative alleles) of interest, X_c represents the c th covariate, such as age, sex, or ancestry, ϵ is a normally distributed error, and $\alpha_k, \beta_k, \gamma_c$ are regression coefficients. The coefficients of the interaction terms β_k are of primary interest: intuitively, if there exist genetic effects of DNAm in cell type k, the observed association between methylation and genotype should be stronger in samples with a higher fraction of cell type k compared to samples with lower fractions [12]. Of note, this model without intercept is equivalent to the following one, due to the fact that cell type proportions add up to 1:

$$\mathrm{M} = \sum_{\mathrm{c}=1}^{\mathrm{C}} \widetilde{\gamma}_{\mathrm{c}} \mathrm{X}_{\mathrm{c}} + \widetilde{\alpha}_0 + \sum_{\mathrm{k}=1}^{\mathrm{K}-1} \widetilde{\alpha}_{\mathrm{k}} \mathrm{W}_{\mathrm{k}} + \widetilde{\beta}_0 \mathrm{G} + \sum_{\mathrm{k}=1}^{\mathrm{K}-1} \widetilde{\beta}_{\mathrm{k}} (\mathrm{W}_{\mathrm{k}} \cdot \mathrm{G}) + \epsilon. \tag{2}$$

The difference between the two models lies in the interpretation of coefficients. In model (1), β_k represents the genetic effects on DNAm in cell type k ($k = 1,2,\ldots,K$). In model (2), $\widetilde{\beta}_0$ represents the genetic effects on DNAm in cell type K and $\widetilde{\beta}_k$ represents the changes in genetic effects in cell type k ($k = 1,2,\ldots,K-1$) compared to the effect $\widetilde{\beta}_0$ in cell type K. Therefore, $\widetilde{\beta}_0 + \widetilde{\beta}_k$ corresponds to the genetic effects in cell type k ($k = 1,2,\ldots,K-1$). For simplicity, we use model (1) in the following derivations.

In order to take the sparsity of genetic effects into consideration and to update information derived from CTS methylation data from a small group of samples, a hierarchical framework is used to construct priors for regression coefficients [43]. To achieve optimal performance, we recommend that the small group of samples with CTS methylation data can be a subset of the overall samples with bulk data, or be similar samples drawn from the same study or cohort as the bulk samples. We first assume a multivariate normal distribution for coefficients for interaction terms $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \middle| \boldsymbol{\tau} = \left[\begin{matrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{matrix}\right] \middle| \boldsymbol{\tau} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- with general prior (no CTS methylation data)

$$\mu = 0, \Sigma = \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_K^2 \end{bmatrix} \tag{3}$$

- with prior derived from CTS methylation data of a small group of samples

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{bmatrix}, \Sigma = \begin{bmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \cdots & \rho_{1K}\tau_1\tau_K \\ \rho_{12}\tau_2\tau_1 & \tau_2^2 & \cdots & \rho_{2K}\tau_2\tau_K \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}\tau_K\tau_1 & \rho_{2K}\tau_K\tau_2 & \cdots & \tau_K^2 \end{bmatrix} \tag{4}$$

where μ_k and ρ_{jk} are updated from the CTS methylation data in the following ways:

(1) $\mu_k = w_\mu \cdot \widehat{\beta}_{k,seq} + (1 - w_\mu) \cdot 0$, where $\widehat{\beta}_{k,seq}$ is the estimated effect when regressing the cell type k methylome on the SNP, and $w_\mu = 1 - p_{adjust}$ where p_{adjust} is the

p-value adjusted by Benjamini & Hochberg (BH) or Bonferroni [17], as defined by users.

(2) $\rho_{jk} = w_\rho \cdot \widehat{\rho}_{jk,seq} + (1 - w_\rho) \cdot 0$, where $\widehat{\rho}_{jk,seq}$ is the estimated genetic correlation between cell type k methylation and cell type j methylation, and $w_\rho = 1 - p_{adjust}$ where p_{adjust} is the corresponding p-value for $\widehat{\rho}_{jk,seq}$, adjusted by Benjamini & Hochberg (BH) or Bonferroni [17], as defined by users.

In both settings with and without CTS methylomes derived from CTS DNAm data, the variable-specific parameter τ_k^2 controls the degree of shrinkage: as τ_k^2 gets close to 0, β_k is shrunk to μ_k, while as τ_k^2 gets larger, the amount of shrinkage will be smaller. We further model τ_k^2 using the exponential distribution with variable-specific hyperparameters s_k:

$$\tau_k^2 | s_k \sim Exp\left(\frac{s_k^2}{2}\right), \tag{5}$$

where s_k was modelled using a gamma distribution as a hyper-prior:

$$s_k \sim Gamma(\mathrm{a}, \mathrm{b_k}). \tag{6}$$

In this way, we allow different degrees of shrinkage for different variables by introducing the variable-specific parameters s_k and τ_k^2. We also derive the conditional posterior distributions of s_k and τ_k^2 as follows:

$$s_k | \beta_k \sim Gamma(\mathrm{a} + 1, \mathrm{b_k} + |\beta_k - \mu_\mathrm{k}|), \tag{7}$$

$$1/\tau_k^2 | s_k, \beta_k \sim Inverse\ Gaussian\left(\frac{s_k}{|\beta_\mathrm{k} - \mu_\mathrm{k}|}, s_k^2\right). \tag{8}$$

These will be used in the model fitting algorithm.

EM-IWLS algorithm for model fitting and inference

We fit the hierarchical Bayesian interaction model by a modified iterative weighted least squares (IWLS) algorithm, proposed by Yi et al. [43]. Compared with usual IWLS, the new method incorporates an expectation–maximization (EM) algorithm that treats the unknown variances τ_k^2 and the hyperparameter s_k as missing data and estimates the $\boldsymbol{\beta}$ by averaging over these missing values; hence, it is also referred to as the EM-IWLS algorithm.

In each iteration of the E-step, we update the missing values (s_k, τ_k^2) by their conditional expectations derived from (7) and (8). In the M-step, we update $\boldsymbol{\beta}$ by maximizing the expected log-likelihood. We need to incorporate the prior $\boldsymbol{\beta}|\boldsymbol{\tau}$ into the normal likelihood as additional data points [44]. Let J denote the total number of variables: $(J - K)$ covariates (e.g., α_k, γ_c) included to address potential confounding, and K covariates ($\boldsymbol{\beta}$) of our interest, and let $\boldsymbol{\theta} = \left[\boldsymbol{\gamma}^{\boldsymbol{T}}, \boldsymbol{\alpha}^{\boldsymbol{T}}, \boldsymbol{\beta}^{\boldsymbol{T}}\right]^{\boldsymbol{T}} \in \boldsymbol{R}^J$. Model (1) could be expressed as:

$$M = X\theta + \epsilon, \tag{9}$$

where $\boldsymbol{X} \in \boldsymbol{R}^{n \times J}$ is the original design matrix in model (1).

Then we update the regression coefficients by running the augmented linear regression:

$$\boldsymbol{y}_{*} \sim N(\boldsymbol{X}_{*}\boldsymbol{\theta}, \phi\boldsymbol{\Sigma}_{*}), \tag{10}$$

where $\boldsymbol{y}_{*} = \left[\boldsymbol{M}^{T}, 0^{T}, \boldsymbol{\mu}^{T}\right]^{\boldsymbol{T}}$ is an $((n+J) \times 1)$ vector of methylation levels for n samples and prior means for J covariates, $X_{*} = \begin{bmatrix} X \\ I_J \end{bmatrix}$ is an $((n+J) \times J)$ matrix constructed by the design matrix $\boldsymbol{X}$ in (9) and the identity matrix, and $\boldsymbol{\Sigma}_{*} = \begin{bmatrix} \boldsymbol{I_n} & 0 & 0 \\ 0 & \frac{1}{\boldsymbol{\phi}}\boldsymbol{\Sigma}_{22} & 0 \\ 0 & 0 & \frac{1}{\boldsymbol{\phi}}\boldsymbol{\Sigma}_{33} \end{bmatrix}$ is an $((n+J) \times (n+J))$ matrix with $\boldsymbol{\Sigma}_{22} = \mathrm{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_{(J-K)}^2)$ and $\boldsymbol{\Sigma}_{33} = \boldsymbol{\Sigma}$ is the $(K \times K)$ prior variance matrix for $\boldsymbol{\beta}|\boldsymbol{\tau}$. Then in each iteration, we can update the estimates:

$$\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{X}_{*}^{T}\boldsymbol{\Sigma}_{*}^{-1}\boldsymbol{X}_{*}\right)^{-1}\boldsymbol{X}_{*}^{T}\boldsymbol{\Sigma}_{*}^{-1}\boldsymbol{y}_{*}, \tag{11}$$

$$\widehat{\phi} = \frac{1}{\boldsymbol{n}}\left(\boldsymbol{y}_{*} - \boldsymbol{X}_{*}\widehat{\boldsymbol{\theta}}\right)^{\boldsymbol{T}}\boldsymbol{\Sigma}_{*}^{-1}\left(\boldsymbol{y}_{*} - \boldsymbol{X}_{*}\widehat{\boldsymbol{\theta}}\right), \tag{12}$$

until convergence. We can also get the variance of regression coefficients:

$$var\left(\widehat{\boldsymbol{\theta}}\right) = \left(\boldsymbol{X}_{*}^{T}\boldsymbol{\Sigma}_{*}^{-1}\boldsymbol{X}_{*}\right)^{-1}\widehat{\phi}. \tag{13}$$

The EM-IWLS algorithm is summarized as follows.

```
EM-IWLS Algorithm

Input: $X$, M.
Initialization: calculate values $\boldsymbol{\theta}^{\mathbf{0}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{M}$ and set $\phi^0 = 1$.
Repeat the following steps:
    1. E Step: update the missing values $(s_k, \tau_k^2)$ by their conditional expectations,
        1.1 For $\boldsymbol{\beta}$: $s_k^{(t)} = E(s_k|\beta_k^{(t-1)}) = \frac{a+1}{\mathrm{b_k}+|\beta_\mathrm{k}^{(\mathrm{t}-1)}-\mu_\mathrm{k}|}$ ,
                $1/\tau_k^{2(t)} = E(1/\tau_k^2|s_k^{(t)}, \beta_k^{(t-1)}) = \frac{s_k^{(t)}}{|\beta_\mathrm{k}^{(\mathrm{t}-1)}-\mu_\mathrm{k}|}$
        1.2 For $\boldsymbol{\alpha}, \boldsymbol{\gamma}$: $s_k^{(t)} = s_0$ are prefixed at small values (e.g., 0.001) instead of
            modelled as hyperparameters to ensure the smallest amount of shrinkage
            for these covariates.
                $1/\tau_k^{2(t)} = E(1/\tau_k^2|s_k^{(t)}, \alpha_k^{(t-1)}) = \frac{s_0}{|\alpha_\mathrm{k}^{(\mathrm{t}-1)}|}$
    2. M Step: based on $1/\tau_k^{2(t)}$, update $\boldsymbol{\Sigma}_{*}^{(t)}$, then update $\boldsymbol{\theta}^{(\boldsymbol{t})}$and $\phi^{(t)}$ according to
       (11) and (12).
Stop if the algorithm has converged.
Output: $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.
```

We define convergence as each element of $|\boldsymbol{\theta}^{(\boldsymbol{t})} - \boldsymbol{\theta}^{(\boldsymbol{t}-1)}|$ smaller than δ, with δ to be a small value (e.g., $1\mathrm{E}-05$). $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{var}\left(\widehat{\boldsymbol{\theta}}\right)$ can then be obtained from the last updates.

For the choice of $(\mathrm{a}, \mathrm{b_k})$, we fix $\mathrm{a} = 0.5$ as the default since the overall degree of shrinkage can be determined by $\mathrm{b_k}$ [43]. For the user-defined $\mathrm{b_k}$, we suggest taking the sample size and the abundance of the corresponding cell type into consideration. For moderate sample size (e.g., $\mathrm{n} = O\left(10^2\right)$), we suggest $\mathrm{b_k} = 0.2$ for most cell types (average of cell type proportions > 10%), and $\mathrm{b_k} = 5$ to induce a less informative prior for the least abundant cell type (average of cell type proportions < 5%). Otherwise, the estimation of the coefficient for the least abundant cell type might be overwhelmingly driven by the prior. For larger sample sizes (e.g., $n = O(10^4)$) or much more abundant cell type, $\mathrm{b_k}$ could be decreased accordingly.

Simulation settings

In this section, we introduce the simulation procedure to evaluate the performance of HBI. CTS DNAm in our simulations were generated based on genotype data from the Wellcome Trust Case Control Consortium (WTCCC) ($n = 15{,}918$) [45]. In the cell type with genetic effects, the heritability of the DNAm was fixed as 0.3, the effect sizes of the causal SNPs were generated by a multivariate normal distribution [46], and GCTA [47] was applied to simulate the DNAm in this cell type. We also generated cell type proportions using a Dirichlet distribution for three cell types with parameters 5.30, 1.27, and 1.62. These parameters were chosen based on the suggestion from Li et al. that the mean cell type composition standard deviation is around 0.13, which was estimated from the Cibersort blood true proportions [48]. Then, for each sample, the bulk DNAm levels were computed as a weighted sum of the simulated CTS DNAm levels, weighted by the corresponding cell type proportions, plus an independent and identically distributed (iid) noise term $\epsilon \sim N(0, 0.01)$.

We considered three main scenarios, and in each of them, we assumed that the total number of SNPs near the simulated CpG site to be 500 and varied the proportion of causal SNPs from 10% to 20% to 40%. All the SNPs were randomly selected from chromosome 12 and all have minor allele frequency (MAF) > 0.01. To investigate whether variants with lower frequency have low power and high false positives, we further divided the SNPs into variants with low frequency ($0.01 \leq \mathrm{MAF} < 0.05$) and common variants ($\mathrm{MAF} \geq 0.05$), and assessed their performance separately as supplementary results. Each simulation setting was repeated 10 times.

(1) Scenario 1: there were genetic effects only in the major/most abundant cell type.
(2) Scenario 2: there were genetic effects only in the minor/least abundant cell type.
(3) Scenario 3: there were correlated genetic effects in all three cell types, and the genetic correlation among the cell types was set to 0.5.

We compared our method HBI with TCA [10], bMIND [11], and the basic interaction model, which fits model (1) directly using OLS and is similar to the CellDMC algorithm [12]. For HBI and the basic interaction model, we inputted the simulated bulk DNAm and cell type proportions and directly obtained the genetic effects for each cell type as the estimated coefficients for the interaction terms ($W_k \cdot G$). The choices of the HBI parameters in the hyper prior $Gamma(\mathrm{a}, \mathrm{b_k})$ were as follows: $\mathrm{a} = 0.5$ for all cell types, $\mathrm{b_k} = 0.005$ for the major cell type, $\mathrm{b_k} = 0.1$ for the other two cell types. For TCA and bMIND, we first inputted the bulk DNAm and cell type proportions to get deconvoluted CTS DNAm, and then tested the association between CTS DNAm and genotype using PLINK [49] to fit the following two models:

1. Marginal model which regresses the deconvoluted DNAm for cell j, $\widehat{Z_j}$, on the genotype G (equivalent to *marginal test* in TCA) [32]:

$$\widehat{Z}_j \sim G.$$

2. Conditional model which regresses the deconvoluted DNAm for cell j on the genotype with DNAm for all other cell types controlled (equivalent to *marginal conditional test* in TCA):

$$\widehat{Z}_j \sim G + \sum_{l \neq j} \widehat{Z}_l.$$

The coefficients for genotype G would then be the estimated genetic effects in cell j. CTS-meQTLs were identified with FDR controlled at 0.05 for each cell type. In each setting, the performance of different methods was compared in terms of correlation between the estimated and true effect sizes, the MSE between the estimated and true effect sizes, power, and FDR calculated as follows:

$$power = \frac{\#\ identified\ true\ signals\ in\ all\ cell\ types}{\#\ true\ signals\ in\ all\ cell\ types},$$

$$false\ discovery\ rate = \frac{\#\ identified\ false\ signals\ in\ all\ cell\ types}{\#\ identified\ signals\ in\ all\ cell\ types}$$

Both HBI and bMIND had the optional step to incorporate CTS information to update priors (derived from cell-sorted MC-seq data in our case and from scRNA-seq data in bMIND's original case). Here we also assumed that for a small proportion (5%) of all samples, their CTS methylation data were available. In each simulation setting, we further compared HBI and bMIND both without this information incorporated and with this information incorporated (denoted as HBI_CTS-prior, bMIND_CTS-prior).

Since all methods included here relied on cell type proportions, we further evaluated the robustness of all methods when noisy cell type proportions were given. With the proportion of causal SNPs fixed as 20% in scenario 3, we randomly simulated noise from a left-truncated normal distribution (truncation point is zero), added noise to the true cell type proportions, and then normalized the sum of proportions to be 1. Two additional simulation settings were performed as we adjusted the standard deviation of the added noise so that the generated noisy cell type proportions would have mean absolute error (MAE) of 0.05 and 0.1, respectively. In addition, to assess the effect of the number of SNPs near the simulated CpG site, we performed additional simulations and varied the number of total SNPs from 500, 1000, to 2000, with the proportion of causal SNPs fixed as 10%.

To further investigate the simulation performance of HBI using real data, we utilized the samples in ROSMAP data with matched gene expression and genotype [21] ($n = 290$). We first estimated the "ground truth" using its single-cell RNA seq data. We included three cell types: excitatory neurons, inhibitory neurons, and oligodendrocytes, and estimated their eQTLs separately. We then extracted the significant eQTLs (Bonferroni-adjusted $p < 0.05$) fitting into 3 scenarios: (1) eQTLs only in excitatory neurons (simulated as the major cell type in pseudo-bulk), (2) eQTLs only in oligodendrocytes (simulated as the minor cell type in pseudo-bulk), and (3) eQTLs in all three cell types. The effect sizes of those eQTLs estimated by single-cell RNA seq data were treated as "ground truth". Pseudo-bulk data consisting of the 3 cell types were then created as the input for TCA, bMIND, the basic interaction model, and HBI. In each repeat, we randomly sampled 500 eQTLs in each scenario and applied all the methods to evaluate their power to correctly identify those eQTLs. Similarly, the performance was compared in terms of correlation between the estimated and true effect sizes, the MSE between the estimated and true effect sizes, power, and FDR.

Study cohort for real data applications

The Women's Interagency HIV Study (WIHS), now a part of MWCCS, is a multi-center, prospective, observational cohort study [15]. All participants are women with HIV or at risk for HIV acquisition. Informed consent was provided by all WIHS participants via protocols approved by institutional review committees at each affiliated institution. In our analysis, participants with matched genetic data and bulk DNA methylation measured in PBMC ($n=431$) and a separate group of participants with CTS DNA methylation data ($n=47$) were included. Demographic and clinical characteristics are summarized in Additional file 2: Table S1.

Genotyping, imputation, and quality control

The WIHS sample were genotyped using the Infinium Omni2.5 Bead-Chip that targeted approximately 2.4 million SNPs. Minimac4 was used for imputation with the 1000 Genomes Project 3 as the reference panel [50, 51]. We removed SNPs with minor allele frequency <0.05, missing rate $>5\%$, imputation quality $r^2<0.8$, and those that deviated significantly from Hardy–Weinberg equilibrium ($p<1e-6$). As a result, approximately 4.6 million SNPs passed QC and were used for CTS-meQTL estimation.

DNA methylation

DNA methylation measured using DNA isolated from PBMC was profiled using the Illumina Infinium MethylationEPIC BeadChip. We followed methods described in Lehne et al. [52] to perform methylation normalization and adjust for potential batch effects. A total of 852,073 CpGs for the 431 individuals passed quality control steps and were used as bulk DNAm data. We applied the method described by Houseman et al. to estimate the cell type proportions for CD4+ T-cells, CD8+ T-cells, natural killer cells, B cells, monocytes, and granulocytes [19, 40]. Another separate group of the WIHS cohort ($n=47$) were isolated for CD4+ T-cells, CD8+ T-cells, and monocytes. DNAm for each sorted cell type was profiled by the Agilent SureSelectXT Methyl-seq. After quality control and extracting CpGs that overlapped on both platforms, we had 390,851 CpGs measured in CD4+ T-cells ($n=28$), 385,679 CpGs measured in CD8+ T-cells ($n=28$), and 407,646 CpGs measured in monocytes ($n=27$), which were used as CTS DNAm data to update priors.

CTS-meQTL estimation and replication

We applied HBI to identify CTS meQTLs in the WIHS cohort for six cell types: CD4+ T-cells, CD8+ T-cells, natural killer cells, B cells, monocytes, and granulocytes.

For each CpG, we considered the following model for SNPs from 500 kb upstream to 500 kb downstream [53–56]:

$$M = \sum_{C=1}^{C} \gamma_c X_c + \sum_{k=1}^{6} \alpha_k W_k + \sum_{k=1}^{6} \beta_k (W_k \cdot G),$$

where M is the bulk methylation M-value, W_k is the cell type proportion of the k th cell type, G is the genotype of the SNP, X_c is a collection of previously identified relevant covariates: age, estimated global ancestry, local ancestry [55], tobacco use, alcohol consumption, HIV infection status, $\log_{10}$ of HIV RNA viral load, the top 5 genotype principal components (PCs), and the top 10 PCs on DNA methylation levels of control probe. HBI was applied to estimate the regression coefficients in the above model, and for CD4+ T-cells, CD8+ T-cells, and monocytes, we further incorporated the priors derived from the CTS methylation data available in a small group of subjects. The choices of the parameters in the hyper prior $Gamma(\mathrm{a}, \mathrm{b_k})$ were as follows: $\mathrm{a} = 0.5$ for all cell types, $\mathrm{b_k} = 5$ for granulocytes, $\mathrm{b_k} = 0.2$ for natural killer cells, B cells, monocytes, and $\mathrm{b_k} = 0.05$ for CD4+ T-cells, CD8+ T-cells. Among the 852,073 CpGs, a total of 1.4 billion SNP-CpG pairs were tested, and significant meQTLs were selected using Bonferroni correction ($p < 0.05/1{,}384{,}706{,}562/6 = 6.02\mathrm{E} - 12$). Due to the low proportion of granulocytes, we also conducted a sensitivity analysis with five-cell-type decomposition (proportion of granulocytes removed). CpGs on chromosome 22 were used in the sensitivity analysis and results are summarized in Additional file 1: Fig S10.

Independent data were used to replicate our identified CTS-meQTLs. We downloaded datasets for meQTLs in isolated white blood cell subsets (i.e., CD4+ T-cells, CD8+ T-cells, monocytes, neutrophils) ($n = 60$ individuals) [1]. For our identified SNP-CpG pairs in the respective cell types, we calculated the percentage of pairs that were significant in the replication set ($p < 0.05$), the percentage of pairs with directional consistency in effect sizes, and the percentage of replicated pairs ($p < 0.05$ and same effect direction). Among the replicated pairs, we also calculated the correlations of the effect sizes. Considering the limited sample size for this dataset ($n = 60$) [1], we also included another data with larger sample sizes for meQTLs in isolated blood cells (CD4+ T-cells, monocytes) ($n = 197$) [22] as supplementary results for replication.

To investigate the replication rates of the version of HBI with only bulk data, we conducted parallel analyses using HBI without priors incorporated from the WIHS participants with CTS data ($n = 47$). In addition, to increase the number of shared SNP-CpG pairs between our results and the replication data, we further utilized SNPs in LD to match pairs. Specifically, if one of our significant pairs SNP1-CpG1 could not be directly matched to the replication data, we would search for SNPs in LD (r2 > 0.6) with this SNP1. If we found one SNP in high LD (i.e., SNP2) and SNP2-CpG1 was present in the replication data, then this original pair SNP1-CpG1 could be matched to replication data. In this way, we increased the number of shared pairs among our identified CTS-meQTLs. We also included other methods for comparison (conditional models for bMIND and TCA).

MeQTL enrichment in genomic functional annotations

For all the variants tested for SNP-CpG associations, we used annotatr [57] and its built-in annotation databases to make CpG annotations (CGI, CGI shelves, CGI shores, inter CGI regions), gene body annotations (regions < 1 kb upstream of the transcription start site, coding sequence, exons, introns, intergenic regions, 5′UTRs, 3′UTRs), gene regulatory and open chromatin annotations (active promotor, weak promotor, strong enhancer, weak enhancer, insulator, regions with heterochromatic or heterochromatin-like characteristics). For gene regulatory and open chromatin annotations [58], we used the database for the K562 cell line, which is commonly used to study hematopoiesis [59]. To test whether the identified meQTLs were enriched in some functional regions, we performed functional enrichment analysis using Fisher's exact test [60, 61]. A 2 × 2 contingency table was built as follows:

	MeQTLs	**Non-meQTLs**	**Row sum**
In functional region R	MR	R-MR	R
Not in functional region R	M-MR	T-R-(M-MR)	T-R
Column sum	M	T-M	T

The total sum of the contingency table (T) was the number of variants that were tested for SNP-CpG associations. The number of identified meQTLs that were mapped in one specific functional region corresponds to the upper-left cell of the table (MR). The remaining three cells of the table can be calculated based on MR and the row/column sums. Based on the 2 × 2 contingency table, we tested whether the meQTLs were enriched in the functional region more often than by chance expected by the genome background (non-meQTLs) [62, 63]. For each cell type, we performed this analysis separately and derived the enrichment estimates as log of odds ratios and its 95% confidence intervals. Enrichment across all cell types was conducted by combining CTS-meQTLs into a union set comprising meQTLs identified in at least one cell type.

Pathway analyses based on identified CTS-meQTLs

For the identified meQTLs in each cell type, we used ANNOVAR to map variants to their nearest gene, and for variants in intergenic regions, the closest gene was kept [64]. Pathway enrichment analyses were conducted with QIAGEN Ingenuity Pathway Analysis (IPA) (QIAGEN Inc., https://digitalinsights.qiagen.com/IPA) [23]. In each cell type, we reported significant pathways at Bonferroni-adjusted $p < 0.05$.

Colocalization of meQTL with GWAS loci

To identify potential associations between meQTLs and complex traits, we applied HyPrColoc (Hypothesis Prioritization for multi-trait Colocalization) [26] in multiple genomic regions. We downloaded GWAS summary statistics published by Barbeira et al. [27], and used the 57 traits in the categories of blood cell counts, cardiometabolic, immune, and allergy. Since colocalization reports the posterior probability that two traits are colocalized in a specific linkage disequilibrium (LD) region [26, 65], we first performed clumping on the meQTLs identified in each cell type. For each CpG, highly correlated genetic variants were clustered into one clump with an LD $r^2 > 0.1$ [66], resulting in 7766 meQTL clumps for CD4+ T-cells, 4211 clumps for CD8+ T-cells, 4568 clumps for monocytes, 3219 clumps for B cells, 2649 clumps for natural killer cells, and 1821 clumps for granulocytes. For each cell type, the genetic variants in each meQTL clump were matched with GWAS summary statistics. Then for each meQTL-GWAS region pair, HyPrColoc was applied on the effect size and the corresponding standard errors. The PPFC was used to identify significant (PPFC > 0.50) colocalizations [35, 67].

Cell type-specific enrichment in meQTL-GWAS colocalizations

To investigate the cellular specificity of complex traits, we performed enrichment analyses to study if the meQTL-GWAS colocalizations for each trait were enriched in certain cell types. Here, meQTLs in granulocytes were excluded due to low numbers of colocalizations identified across traits, and meQTLs in bulk level (282,965 clumps) were included to assess if CTS-meQTLs can reveal more cellular-specific information. We also excluded three traits with a very small number of meQTL-GWAS colocalizations (< 10) across all cell types. As a result, 54 of the 57 GWAS traits remained in the enrichment analyses. For each trait, in each cell type the enrichment score was defined as the ratio between the percentage of meQTL-GWAS colocalizations (colocalized meQTL clumps) in that cell type and the percentage of meQTL clumps covered by that cell type:

$$Enrichment_k = \frac{\%\ colocalized\ meQTL\ clumps\ in\ cell\ type\ k}{\%\ meQTL\ clumps\ in\ cell\ type\ k}$$

To determine significant colocalization enrichments in certain cell types, the test for equality of proportions with continuity correction was performed to test if $Enrichment_k > 1$ ($p < 0.05/54/6 = 1.5E-04$). To evaluate the identified enrichment results, for the same GWAS traits we also performed heritability enrichment analyses using 66 functional annotations from GenoSkyline-Plus (v1.0.0) [29]. For our identified traits with colocalizations enriched in certain cell types, we determined if the heritability of this trait also enriched in this cell type at $p < 0.05$.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03411-7.

Additional file 1: Supplementary Figures. Supplementary figures 1-10

Additional file 2: Table S1. Demographic information for the WIHS participants

Additional file 3: Table S2. Results for the significant cell-type-specific meQTLs

Additional file 4: Table S3. Replication results for the identified cell-type-specific meQTLs

Additional file 5: Table S4. Canonical Pathways identified using genes mapped by the meQTLs

Additional file 6: Table S5. GWAS-meQTL colocalizations in each cell type

Additional file 7: Table S6. Colocalization enrichment results

Additional file 8: Review history

Acknowledgements

The authors appreciate the support of the WIHS cohort and Yale Center of Genomic Analysis. The simulation results are in part based on ROSMAP gene expression and genomic variants data obtained from the AD Knowledge Portal (https://adknowledgeportal.org), (https://www.synapse.org/#!Synapse:syn23446022). ROSMAP study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through

funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic).
Data in the application part of this manuscript were collected by the Women's Interagency HIV Study (WIHS), now the MWCCS. The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH). MWCCS (Principal Investigators): Atlanta CRS (Ighovwerha Ofotokun, Anandi Sheth, and Gina Wingood), U01-HL146241; Baltimore CRS (Todd Brown and Joseph Margolick), U01-HL146201; Bronx CRS (Kathryn Anastos, David Hanna, and Anjali Sharma), U01-HL146204; Brooklyn CRS (Deborah Gustafson and Tracey Wilson), U01-HL146202; Data Analysis and Coordination Center (Gypsyamber D'Souza, Stephen Gange and Elizabeth Topper), U01-HL146193; Chicago-Cook County CRS (Mardge Cohen and Audrey French), U01-HL146245; Chicago-Northwestern CRS (Steven Wolinsky, Frank Palella, and Valentina Stosor), U01-HL146240; Northern California CRS (Bradley Aouizerat, Jennifer Price, and Phyllis Tien), U01-HL146242; Los Angeles CRS (Roger Detels and Matthew Mimiaga), U01-HL146333; Metropolitan Washington CRS (Seble Kassaye and Daniel Merenstein), U01-HL146205; Miami CRS (Maria Alcaide, Margaret Fischl, and Deborah Jones), U01-HL146203; Pittsburgh CRS (Jeremy Martinson and Charles Rinaldo), U01-HL146208; UAB-MS CRS (Mirjam-Colette Kempf, Jodie Dionne-Odom, Deborah Konkle-Parker, and James B. Brock), U01-HL146192; UNC CRS (Adaora Adimora and Michelle Floris-Moore), U01-HL146194. The MWCCS is funded primarily by the National Heart, Lung, and Blood Institute (NHLBI), with additional co-funding from the Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), National Institute On Aging (NIA), National Institute Of Dental & Craniofacial Research (NIDCR), National Institute Of Allergy And Infectious Diseases (NIAID), National Institute Of Neurological Disorders And Stroke (NINDS), National Institute Of Mental Health (NIMH), National Institute On Drug Abuse (NIDA), National Institute Of Nursing Research (NINR), National Cancer Institute (NCI), National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Institute on Deafness and Other Communication Disorders (NIDCD), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute on Minority Health and Health Disparities (NIMHD), and in coordination and alignment with the research priorities of the National Institutes of Health, Office of AIDS Research (OAR). MWCCS data collection is also supported by UL1-TR000004 (UCSF CTSA), UL1-TR003098 (JHU ICTR), UL1-TR001881 (UCLA CTSI), P30-AI-050409 (Atlanta CFAR), P30-AI-073961 (Miami CFAR), P30-AI-050410 (UNC CFAR), P30-AI-027767 (UAB CFAR), P30-MH-116867 (Miami CHARM), UL1-TR001409 (DC CTSA), KL2-TR001432 (DC CTSA), and TL1-TR001431 (DC CTSA).
The authors gratefully acknowledge the contributions of the study participants and dedication of the staff at the MWCCS sites.

Review history

The review history is available as Additional file 8.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

YC, BC, and HL developed the statistical framework. YC implemented the algorithm and performed statistical analysis. HL and XZ assisted in data analysis. BEA, GD, SS, AE, JM, MF, SK, KA, and MC provided DNA samples and contributed to manuscript preparation. KX and HZ were responsible for the study design. KX advised on sample preparation and the biological interpretation of findings. HZ advised on statistical and genetics issues. YC drafted the manuscript. All authors contributed to manuscript editing and approved the manuscript.

Funding

The project was primarily supported by the National Institute on Drug Abuse (R03DA039745, R01DA038632, R01DA047063, R01DA047820). Supported was also provided by NIH grants U01HG013840 and R01 GM134005, and NSF grant DMS1902903.

Availability of data and materials

Genotype data used in the simulation were downloaded from the Wellcome Trust Case–Control Consortium (https://www.wtccc.org.uk) [68]. Real data used in the simulation were from the ROSMAP gene expression and genomic variants data (https://www.synapse.org/#!Synapse:syn23446022) [21, 69]. Those are third party data. Genotype and DNA methylation data in the application part were from the WIHS cohort, which has been identified as one with multiple vulnerabilities (e.g., racial/ethnic minority women, coinfected). The data was generated by the MWCCS sites (not belong to third party). Whereas participants from the cohort who contributed to the findings summarized in this manuscript provided written consent for genetic studies, said consent was collected prior to the most recent guidelines and requirements for data sharing. The WIHS cohort operates under an alternative data sharing plan registered with the National Institutes of Health and access to data can be requested by submitting a Concept Sheet. The instructions for the Concept Sheet submission could be found at https://www.statepi.jhsph.edu/mwccs/wp-content/uploads/2023/10/MWCCS-Concept-Sheet-and-Publication-Policies_10423.pdf. Investigator(s) should work with the Principal Investigator (PI) of a MWCCS site or MWCCS liaison to draft the concept sheet. External investigators may first request a MWCCS liaison from the Data Analysis and Coordination Center (DACC) at MWCCS@jhu.edu). The accession number for the WIHS in dbGaP genomic data is now provided (phs001503) [70]. The cohort is currently being re-approached to obtain informed consent for sharing of their data. This has been consistent with other genomic studies in the WIHS cohort. The GWAS summary data used in the meQTL-GWAS colocalizations can be downloaded at Zenodo (https://doi.org/10.5281/zenodo.3629742) [71]. HBI algorithm is publicly available at https://github.com/YoushuCheng/HBI. The code has also been deposited at Zenodo with https://doi.org/10.5281/zenodo.13131440 [72]. The repository is released under the MIT license.

Declarations

Ethics approval and consent to participate

The study was determined as non-Human Subject by Yale Human Investigation Committee. All data in this study are de-identified. Informed consent was provided by all WIHS participants via protocols approved by institutional review committees at each affiliated institution.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06511, USA. [2]VA Connecticut Healthcare System, West Haven, CT 06516, USA. [3]Department of Psychiatry, Yale School of Medicine, New Haven, CT 06511, USA. [4]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [5]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA. [6]The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [7]Department of Psychiatry, SUNY Downstate Health Sciences University School of Medicine, Brooklyn, NY, USA. [8]Department of Medicine, University of Miami School of Medicine, Miami, FL, USA. [9]Division of Infectious Diseases and Tropical Medicine, Georgetown University, Washington, DC, USA. [10]Department of Medicine, Albert Einstein College of Medicine, New York, NY, USA. [11]Hektoen Institute for Medical Research, Chicago, IL, USA. [12]Bluestone Center for Clinical Research, College of Dentistry, New York University, New York, NY, USA. [13]Department of Oral and Maxillofacial Surgery, College of Dentistry, New York University, New York, NY, USA.

Received: 19 December 2023 Accepted: 30 September 2024
Published online: 15 October 2024

References

1. Hawe JS, Wilson R, Schmid KT, Zhou L, Lakshmanan LN, Lehne BC, et al. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nat Genet. 2022;54(1):18–29.
2. Hongyu L, Jiawei W, Dianne AC, Jennifer LM, David LC, José Jaime M-M, et al. Functional annotation of the human PTSD methylome identifies tissue-specific epigenetic variation across subcortical brain regions. medRxiv. 2023:2023.04.18.23288704. https://doi.org/10.1101/2023.04.18.23288704.
3. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131–8.
4. McClay JL, Shabalin AA, Dozmorov MG, Adkins DE, Kumar G, Nerella S, et al. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. Genome Biol. 2015;16:291.
5. Lemire M, Zaidi SH, Ban M, Ge B, Aïssi D, Germain M, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. Nat Commun. 2015;6:6326.
6. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61.
7. Huan T, Joehanes R, Song C, Peng F, Guo Y, Mendelson M, et al. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat Commun. 2019;10(1):4267.
8. Gao X, Thomsen H, Zhang Y, Breitling LP, Brenner H. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. Clin Epigenetics. 2017;9:87.
9. Perzel Mandell KA, Eagles NJ, Wilton R, Price AJ, Semick SA, Collado-Torres L, et al. Genome-wide sequencing-based identification of methylation quantitative trait loci and their role in schizophrenia risk. Nat Commun. 2021;12(1):5251.
10. Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nat Commun. 2019;10(1):3417.
11. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. Genome Res. 2021;31(10):1807–18.
12. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. Nat Methods. 2018;15(12):1059–66.
13. Westra HJ, Arends D, Esko T, Peters MJ, Schurmann C, Schramm K, et al. Cell specific eQTL analysis without sorting cells. PLoS Genet. 2015;11(5):e1005223.
14. Leng C, Tran M-N, Nott D. Bayesian adaptive Lasso. Ann Inst Stat Math. 2014;66(2):221–44.
15. Barkan SE, Melnick SL, Preston-Martin S, Weber K, Kalish LA, Miotti P, et al. The women's interagency HIV study. WIHS Collab Study Group Epidemiol. 1998;9(2):117–25.
16. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–29.
17. Iain MJ, Bernard WS. Needles and straw in haystacks: empirical bayes estimates of possibly sparse sequences. Ann Stat. 2004;32(4):1594–649.
18. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12(3):179–85.
19. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1):86.

20. Rahmani E, Schweiger R, Shenhav L, Wingert T, Hofer I, Gabel E, et al. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. Genome Biol. 2018;19(1):141.
21. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. J Alzheimers Dis. 2018;64(s1):S161–89.
22. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016;167(5):1398-414.e24.
23. Krämer A, Green J, Pollard J Jr, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics. 2014;30(4):523–30.
24. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet. 2021;53(9):1311–21.
25. Morrow JD, Glass K, Cho MH, Hersh CP, Pinto-Plata V, Celli B, et al. Human lung DNA methylation quantitative trait loci colocalize with chronic obstructive pulmonary disease genome-wide association loci. Am J Respir Crit Care Med. 2018;197(10):1275–84.
26. Foley CN, Staley JR, Breen PG, Sun BB, Kirk PDW, Burgess S, et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. Nat Commun. 2021;12(1):764.
27. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Kim-Hellmuth S, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol. 2021;22(1):49.
28. Noble JA, Valdes AM. Genetics of the HLA region in the prediction of type 1 diabetes. Curr Diab Rep. 2011;11(6):533–42.
29. Lu Q, Powles RL, Abdallah S, Ou D, Wang Q, Hu Y, et al. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLoS Genet. 2017;13(7):e1006933.
30. Tsai S, Shameli A, Santamaria P. CD8+ T cells in type 1 diabetes. Adv Immunol. 2008;100:79–124.
31. Chen L, Li Z, Wu H. CeDAR: incorporating cell type hierarchy improves cell type-specific differential analyses in bulk omics data. Genome Biol. 2023;24(1):37.
32. Elior R, Brandon J, Regev S, Brooke R, Lindsey AC, Lisa FB, et al. Calling differential DNA methylation at cell-type resolution: addressing misconceptions and best practices. bioRxiv. 2021:2021.02.14.431168. https://doi.org/10.1101/2021.02.14.431168.
33. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. Science. 2019;364(6441):685–9.
34. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570(7761):332–7.
35. Oliva M, Demanelis K, Lu Y, Chernoff M, Jasmine F, Ahsan H, et al. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. Nat Genet. 2023;55(1):112–22.
36. van Lith M, McEwen-Smith RM, Benham AM. HLA-DP, HLA-DQ, and HLA-DR have different requirements for invariant chain and HLA-DM. J Biol Chem. 2010;285(52):40800–8.
37. Qu H-Q, Qu J, Bradfield J, Marchand L, Glessner J, Chang X, et al. Genetic architecture of type 1 diabetes with low genetic risk score informed by 41 unreported loci. Commun Biol. 2021;4(1):908.
38. Pociot F. Type 1 diabetes genome-wide association studies: not to be lost in translation. Clin Transl Immunol. 2017;6(12):e162.
39. Michalek DA, Tern C, Zhou W, Robertson CC, Farber E, Campolieto P, et al. A multi-ancestry genome-wide association study in type 1 diabetes. Hum Mol Genet. 2024;33(11):958–68.
40. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15(2): R31.
41. Hannon E, Dempster EL, Davies JP, Chioza B, Blake GET, Burrage J, et al. Quantifying the proportion of different cell types in the human cortex using DNA methylation profiles. BMC Biol. 2024;22(1):17.
42. Agashe C, Chiang D, Grishin A, Masilamani M, Jones SM, Wood RA, et al. Impact of granulocyte contamination on PBMC integrity of shipped blood samples: Implications for multi-center studies monitoring regulatory T cells. J Immunol Methods. 2017;449:23–7.
43. Yi N, Ma S. Hierarchical shrinkage priors and model fitting for high-dimensional generalized linear models. Stat Appl Genet Mol Biol. 2012;11(6). https://doi.org/10.1515/1544-6115.1803.
44. Andrew G, Aleks J, Maria Grazia P, Yu-Sung S. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Statist. 2008;2(4):1360–83.
45. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
46. Yiliang Z, Youshu C, Yixuan Y, Wei J, Qiongshi L, Hongyu Z. Estimating genetic correlation jointly using individual-level and summary-level GWAS data. bioRxiv. 2021:2021.08.18.456908. https://doi.org/10.1101/2021.08.18.456908.
47. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
48. Li Z, Guo Z, Cheng Y, Jin P, Wu H. Robust partial reference-free cell composition estimation from tissue expression. Bioinformatics. 2020;36(11):3431–8.
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
50. Siva N. 1000 Genomes project. Nat Biotechnol. 2008;26(3):256.
51. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.
52. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan ST, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 2015;16:37.

53. Schulz H, Ruppert A-K, Herms S, Wolf C, Mirza-Schreiber N, Stegle O, et al. Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. Nat Commun. 2017;8(1):1511.
54. Pierce BL, Tong L, Argos M, Demanelis K, Jasmine F, Rakibuz-Zaman M, et al. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. Nat Commun. 2018;9(1):804.
55. Li B, Aouizerat BE, Cheng Y, Anastos K, Justice AC, Zhao H, et al. Incorporating local ancestry improves identification of ancestry-associated methylation signatures and meQTLs in African Americans. Commun Biol. 2022;5(1):401.
56. Drong AW, Nicholson G, Hedman AK, Meduri E, Grundberg E, Small KS, et al. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS ONE. 2013;8(2):e55923.
57. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. Bioinformatics. 2017;33(15):2381–3.
58. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215–6.
59. Klein E, Ben-Bassat H, Neumann H, Ralph P, Zeuthen J, Polliack A, et al. Properties of the K562 cell line, derived from a patient with chronic myeloid leukemia. Int J Cancer. 1976;18(4):421–31.
60. Fisher RA. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. J Roy Stat Soc. 1922;85(1):87–94.
61. Bedrick EJ, Hill JR. [A Survey of Exact Inference for Contingency Tables]: Comment. Stat Sci. 1992;7(1):153–7.
62. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
63. da Huang W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
64. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
65. Cheng Y, Dao C, Zhou H, Li B, Kember RL, Toikumo S, et al. Multi-trait genome-wide association analyses leveraging alcohol use disorder findings identify novel loci for smoking behaviors in the million veteran program. Transl Psychiatry. 2023;13(1):148.
66. Cheng Y, Justice A, Wang Z, Li B, Hancock DB, Johnson EO, et al. Cis-meQTL for cocaine use-associated DNA methylation in an HIV-positive cohort show pleiotropic effects on multiple traits. BMC Genomics. 2023;24(1):556.
67. Thom CS, Voight BF. Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes. BMC Med Genomics. 2020;13(1):89.
68. Wellcome trust case control consortium. 2009. https://www.wtccc.org.uk.
69. AMP-AD knowledge portal. 2014. https://adknowledgeportal.synapse.org.
70. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. 2007. https://www.ncbi.nlm.nih.gov/gap.
71. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Ardlie K, et al. Publicly available GWAS summary statistics, harmonized and imputed to GTEx v8' variant reference. 2020. Zenodo. https://doi.org/10.5281/zenodo.3629742.
72. Cheng Y. YoushuCheng/HBI: HBI (v1.0.0). Zenodo. 2024. https://doi.org/10.5281/zenodo.13131440.

Publisher's Note