# MIMOSA: A resource consisting of improved methylome imputation models increases power to identify DNA methylation-phenotype associations

Hunter J. Melton[1], Zichen Zhang[1], Hong-Wen Deng[2], Lang Wu[3*], Chong Wu[4*]

[1] Department of Statistics, Florida State University

[2] Center of Bioinformatics and Genomics, Tulane University

[3] Cancer Epidemiology Division, University of Hawaii Cancer Center

[4] Department of Biostatistics, University of Texas MD Anderson Cancer Center

**Abstract**

DNA methylation has been shown to be involved in the etiology of many complex diseases, yet the specific key underlying methylation sites remain largely unknown. One strategy to identify putative causal CpG sites and enhance disease etiology understanding is to conduct methylome-wide association studies (MWASs), in which predicted or measured DNA methylation that is associated with complex diseases can be identified. However, current MWAS models are trained with relatively small reference datasets, limiting the ability to adequately handle CpG sites with low genetic heritability. Here, we introduce a new resource, MWAS Imputing Methylome Obliging Summary-level mQTLs and Associated LD matrices (MIMOSA), a set of models that substantially improve the prediction accuracy of DNA methylation and subsequent MWAS power through the use of a large, summary-level mQTL dataset provided by the Genetics of DNA Methylation Consortium (GoDMC). With the analyses of GWAS summary statistics for 28 complex traits and diseases, we demonstrate that MIMOSA considerably increases the accuracy of DNA methylation prediction in blood, crafts fruitful prediction models for low heritability CpG sites, and determines markedly more CpG site-phenotype associations than preceding methods.

*Corresponding author: CW (cwu18@mdanderson.org) and LW (lwu@cc.hawaii.edu)

# 1 Introduction

DNA methylation (DNAm), the epigenetic mechanism of adding methyl groups ($CH_3$) to the genome, serves an essential role in the regulation of gene expression [1]. Inter-individual variation in DNAm has been shown to be linked to the etiology of many complex traits through conventional epidemiological studies. However, its impact on disease development is not yet well-characterized, primarily due to several common biases encountered in observational studies, including selection bias, unmeasured confounding factors, and reverse causation. One strategy to decrease such limitations is employing methylome-wide association studies (MWASs) [2–4]. Similar to transcriptome-wide association studies (TWASs) [5–8], MWAS employs an instrument variable regression setup involving two steps. First, MWAS builds DNA methylation prediction models for a set of CpG sites using a methylation reference panel where both genetic and DNA methylation data are available. Second, MWAS tests the association between predicted DNA methylation levels and traits of interest.

While appealing, current MWASs, to our knowledge, rely on individual-level reference panels to build DNAm prediction models, thus limiting statistical power due to the limited sample sizes of these reference panels. To amerliorate this limitation, we introduce the MWAS Imputing Methylome Obliging Summary-Level mQTLs and Associated LD matrices (MIMOSA), a novel and comprehensive MWAS resource that improves DNAm imputation accuracy beyond existing models and can be used to boost MWAS power. Inspired by a new TWAS method SUMMIT [8] that builds expression prediction models using summary-level expression quantitative trait loci (eQTLs) data, we build DNAm imputation models using summary-level methylation quantitative trait loci (mQTLs) data in blood tissue ($n = 27,750$ individuals) from the Genetics of DNA Methylation Consortium (GoDMC) [9]. Likely due to limitations of available sample sizes of individual-level methylation reference panels, only imputation models with $R^2 > 0.01$ are included in downstream association analyses [2, 4], as methylation at CpG sites with lower heritability was not previously able to be accurately modeled. In contrast, MIMOSA includes imputation models with $0.005 < R^2 \leq 0.01$ (which is recommended by SUMMIT), in an effort to test CpG sites with low heritability that have been largely ignored by existing MWAS models.

Through analyses of GWAS summary statistics for 28 complex traits and diseases, we demonstrate that MIMOSA successfully and substantially increases the accuracy of DNAm imputation in blood and hence subsequently identifies markedly more CpG site-trait associations than existing methods. A database of our newly built DNA methylation imputation models (MIMOSA) and our association results is available as a resource to the broad community.

# 2 Results

## 2.1 MIMOSA summary

By leveraging summary-level mQTL data from GoDMC, we built MIMOSA, a set of comprehensive DNAm prediction models. Specifically, following SUMMIT [8], we trained DNAm prediction models for each CpG site using five penalized regression methods: the elastic net [10], MNet [11], the smoothly clipped absolute deviation (SCAD) [12], the minimax concave penalty (MCP) [13], and LASSO [14]. We draw summary-level mQTL data in blood tissue ($n = 27,750$ individuals) from the GoDMC [9], which is based on Illumina HumanMethylation BeadChips and Epic arrays. Then, the models are tuned and validated using independent datasets from the Framingham Heart Study (FHS) [15]. Specifically, we use FHS subcohort 1 ($n = 797$) for tuning and FHS subcohort 2 ($n = 798$) for validation. The FHS datasets encompass 370,785 CpG sites, 123,375 were included in the summary-level training data from GoDMC. Lastly, models trained by MIMOSA could be used in a standard MWAS: we demonstrated this by testing the association between imputed methylation and phenotype for each methylation imputation model with satisfactory performance ($R^2 > 0.005$), resulting in a set of CpG site-trait associations. The model framework is shown in Figure 1.

## 2.2 MIMOSA increases accuracy in methylome imputation models

We compared the imputation accuracy between MIMOSA and the models created by Baselmans et al. [2] (hereafter referred to as the Baselmans models) for whole blood using independent samples from the Framingham Heart Study (FHS). The MIMOSA models were trained on summary-level mQTL data from GoDMC [9], which is itself based on a sample size of 27,750 subjects, and tuned using individual-level mQTL data (subcohort 1, $n = 797$) from the FHS [15]. In total, MIMOSA

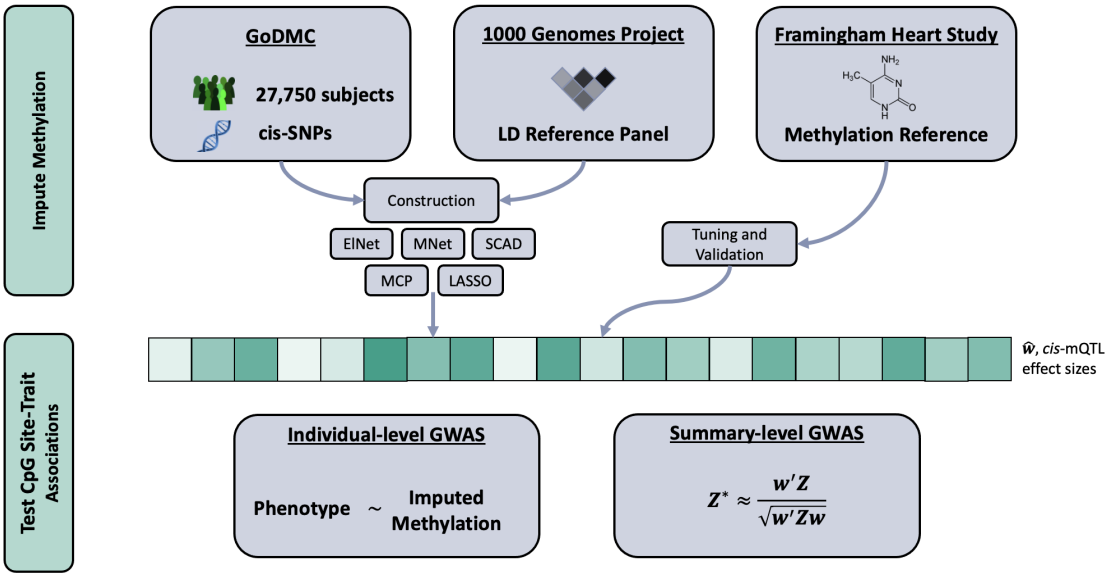



Figure 1: **MIMOSA Workflow.** MIMOSA evaluates CpG site-trait associations by: 1st) training methylome imputation models; 2nd) testing associations between imputed methylation levels and phenotype

created models for 123,375 CpG sites. The Baselmans models [2] were trained on individual-level mQTL data from the BIOS Consortium, based on samples from 4,008 subjects. We evaluated the performance of both methods using individual-level data (subcohort 2, $n = 798$) from the FHS [15]. To make a fair comparison, we examined the number of CpG sites with $R^2 > 0.01$ in the imputation models for the overlapping CpG site list (i.e., CpG sites with models created by both methods). Of the 109,119 candidate CpG sites in this list, MIMOSA built satisfactory ($R^2 > 0.01$) imputation models for 74,551, and Baselmans for 61,202. Of the 61,202 CpG sites for which the Baselmans method built a satisfactory model, MIMOSA successfully created models for 59,121 (96.6%). We conducted a paired, one-sided Wilcoxon rank-sum test to compare $R^2$ for the full set of 109,119 CpG sites; MIMOSA significantly improved prediction accuracy overall over the Baselmans method ($p < 2.2 \times 10^{-16}$). Furthermore, MIMOSA yielded better imputation models for 57,637 out of 76,632 (75.2%) CpG sites with $R^2 > 0.01$ by either model. Figure 2 demonstrates the improved imputation accuracy of MIMOSA models. It is common in TWAS or MWAS to analyze genes or CpG sites with $R^2 > 0.01$ to meet the relevance assumption. In contrast, we recommend including CpG sites with $R^2 > 0.005$ in analyses (see [**Zhang2022SUMMIT**] for justifications). By using this more liberal threshold, we increased the number of testable imputation models by an

additional 11,589 CpG sites.

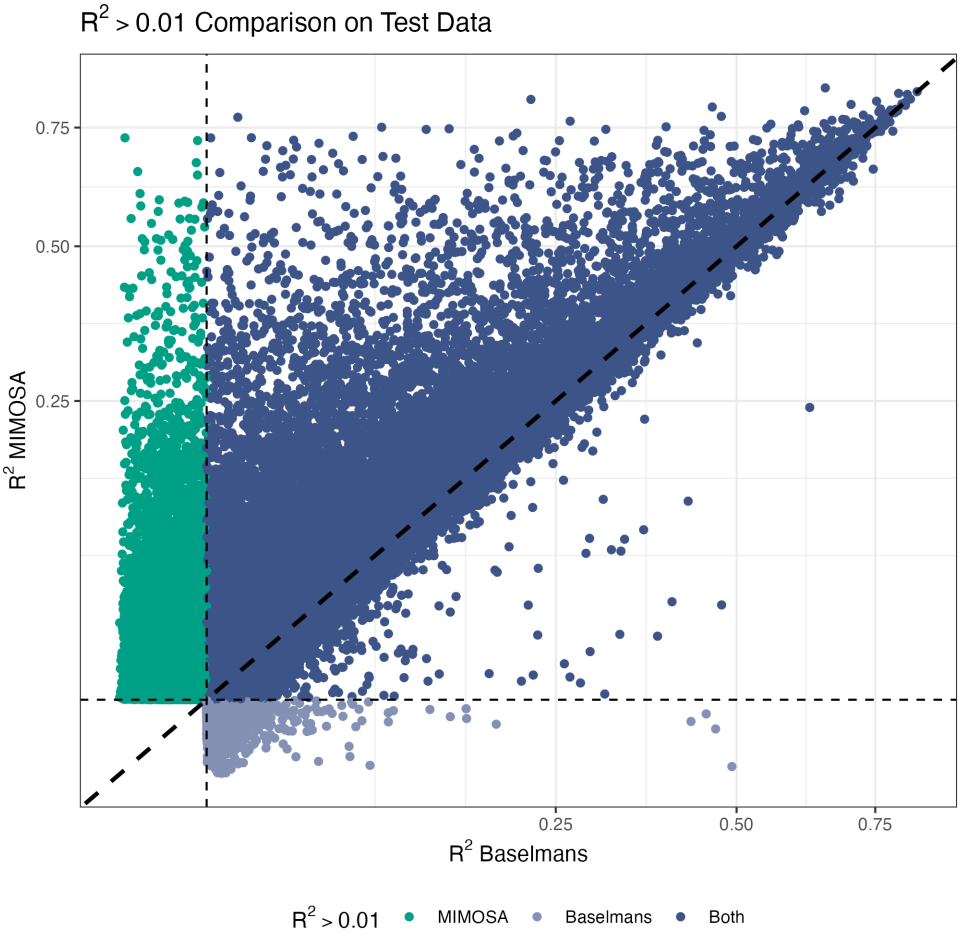



Figure 2: $R^2$ **Comparison.** A comparison of best imputation model $R^2$ on test data for all CpG sites for which either MIMOSA or Baselmans hit $R^2 > 0.01$ threshold.

## 2.3 MIMOSA identifies more CpG site-trait associations

To assess the capability of MIMOSA to identify significant associations, we applied MIMOSA DNAm prediction models to the GWAS summary statistics of 28 complex traits and compared the results with those using the Baselmans models. The significant results are summarized in Supplementary Table 1. As discussed previously, MIMOSA assesses CpG sites with imputation $R^2 > 0.005$; however, we focused only on CpG sites with $R^2 > 0.01$ here for a fair comparison. This slightly favored the competing method, as we used a more conservative threshold on MIMOSA. With this restriction, MIMOSA identified a total of 15,241 significant associations, a significant

increase over the Baselmans method, which identified 8,421 ($p = 1.3 \times 10^{-4}$ per the Wilcoxon signed-rank test). MIMOSA further pinpointed 2,812 associations derived from CpG sites with $0.005 < R^2 \leq 0.01$, bringing the overall association total to 18,053.

Since the two models have distinct sets of candidate CpG sites, we also compared MWAS results on the shared set of CpG sites for which both methods had imputation $R^2 > 0.01$. For these 59,121 candidate sites, MIMOSA again identified significantly more associations (14,290 vs. 6,912, $p = 3.2 \times 10^{-5}$). For the 6,912 associations identified by the Baselmans method, MIMOSA successfully identified 5,271 (76.3 %) under the Bonferroni correction and 6,213 (89.9 %) under a nominal p value threshold $p < 0.05$.

Furthermore, MIMOSA effectively analyzed CpG sites with lower heritability ($0.005 \leq R^2 < 0.01$). A total of 86,140 CpG sites had $R^2 \geq 0.005$, of which 11,589 had $0.005 \leq R^2 < 0.01$. For these CpG sites with lower heritability ($0.005 \leq R^2 < 0.01$), we found 2,812 significant CpG site-trait associations. For the remaining 74,551 CpG sites with $R^2 \geq 0.01$, we identified 15,241 significant associations. This finding indicates that CpG sites with lower $R^2$ (low heritability) are just as integral as those with higher $R^2$, as the proportion of significant CpG site-trait associations were very similar, 24.3% vs. 20.4%. Notably, a similar finding has been reported for gene expression [8, 16]; researchers report that genes with low expression heritability have substantially larger causal effect sizes on complex traits [16].

## 3 Discussion

We built a comprehensive source of DNAm prediction models MIMOSA, which is created following the SUMMIT method [8] with large, summary-level mQTL datasets. MIMOSA notably improved imputation accuracy for the methylome and, through this, increased power to identify associations in MWAS. We demonstrated through the analysis of GWAS results for complex traits that MIMOSA provided an appreciable step forward from previous MWAS models. MIMOSA increased accuracy in methylation imputation, identified more CpG site-trait associations, and was able to analyze CpG sites with lower heritability (i.e., imputation $0.005 \leq R^2 \leq 0.01$).

MWAS can be viewed as one type of instrumental variable (IV) regression, which seeks causal

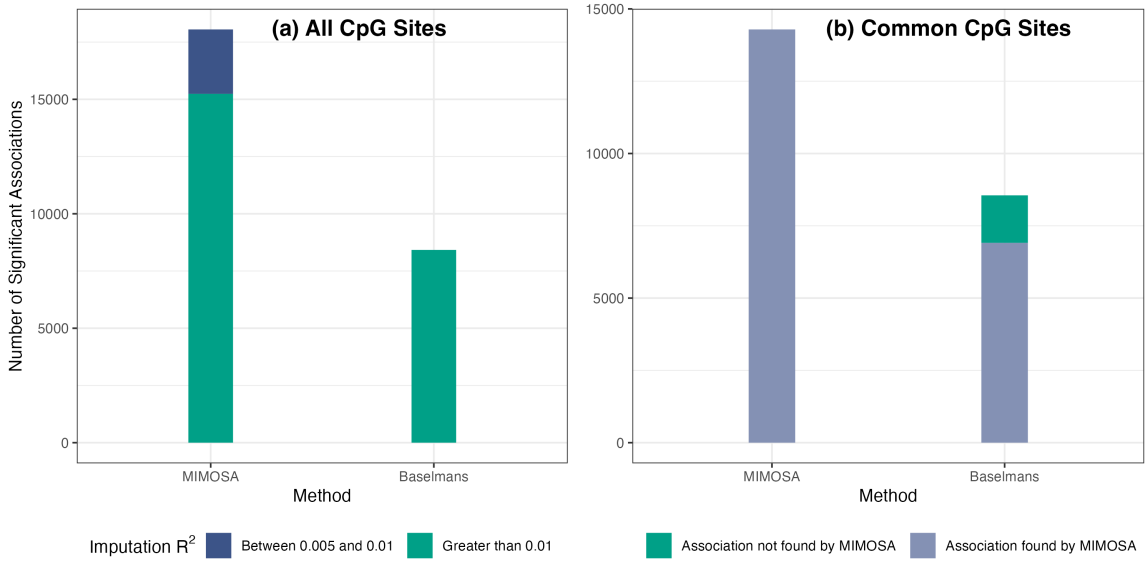



Figure 3: **Evaluation of MIMOSA on real data.** a) Number of significant associations determined by MIMOSA and Baselmans for 28 GWAS using all CpG sites. b) Number of significant associations determined by MIMOSA and Baselmans for 28 GWAS using shared CpG sites.

relationships between exposures and outcomes using instrumental variables [17, 18]. From an instrumental variable perspective, the optimal instrument is the genetic prediction score combining information from multiple SNPs and providing the best prediction accuracy [19]. This explains why the MIMOSA models, which feature better prediction accuracy, increase the statistical power to identify (causal) associations in MWAS. On the other hand, it is important to note that causal interpretations are valid only when all genetic variants used to predict DNAm are valid IVs [5]. Horizontal pleiotropy is widespread [20] and violates instrument variable assumptions. It is important to model the potential violation of valid IV assumptions and we leave such interesting and exciting topics to future research.

The new MIMOSA resource is largely motivated by our previous work SUMMIT [8], where we build gene expression prediction models for TWAS using SUMMIT with summary-level eQTL data. As such, to avoid redundancy, we eschewed a simulation study for this project. Given that the MIMOSA models are created according to the SUMMIT method and the Baselmans models are created by following the TWAS-FUSION [5] method, the performance of these two sets of models is well-explored through the simulation study accompanying SUMMIT. While MIMOSA follows SUMMIT, the resource is equally important and widely needed by the community, as powerful

DNA methylation imputation models such as those featured here will help us gain substantial insights into the etiology of complex diseases [21, 22].

We conclude by discussing the limitations of the current study. First, MIMOSA is built on mQTL data from whole blood tissue from subjects of European ancestry. As a result, the provided methylation imputation weights are appropriate for identifying CpG sites in whole blood for European ancestry. Blood cells play a key role in the immune system and are relevant to many phenotypes, so this is a natural place to start, but the MIMOSA framework could be easily applied to other tissues or subjects with different ancestry when such models have been built. Second, the summary-level mQTL data used in MIMOSA are based on microarray technology (Illumina HumanMethylation BeadChips and Epic arrays), limiting our analysis to only 1.5% of the ∼28 million DNAm sites across the genome [23]. With the recent advancement in whole-genome bisulfite sequencing, we expect the percentage of analyzable CpG sites to grow substantially in the future. The framework can be equally applied once such summary-level datasets become available to the community. Additionally, we focused on cis-SNPs to build imputation models and did not take into consideration the effect of trans-mQTLs. Possibilities for future work include incorporating such trans information or leveraging methylation data from other tissues, similar to TWAS methods UTMOST [7] and MR-JTI [24]. We anticipate either direction will further improve the performance, though substantial additional work may be needed. For example, the effect size of trans-mQTLs is often weaker and is more likely to be impacted by horizontal pleiotropy [9]. Leveraging information from other tissues may also violate the valid IV assumptions because the used IV may affect several tissues simultaneously. Finally, as discussed above, the causal argument of MWAS results (including using MIMOSA models) carry weight only when instrument variable assumptions hold and caution is warranted when interpreting MWAS results.

# 4 Methods

We apply SUMMIT to build the MIMOSA models, a set of DNA methylation prediction models using summary-level mQTL data. For completeness, we briefly describe the methods here. A more detailed discussion may be found in Zhang et. al. [8].

### 4.1 Penalized regression model for methylation imputation

Consider the linear regression model for estimating DNA methylation levels shown below:

$$Y = \sum_{i=1}^{p} w_i X_i + \epsilon,$$

where $Y$ is the $N$-dimensional vector of DNAm levels for a given CpG site, corrected for covariates such as age, sex, and genetic principal components. $X = (X_1', ..., X_p')'$ is the standardized genotype matrix of p cis-SNPs around the CpG site. $w = (w_1, ..., w_p)'$ is the mQTL effect size, and $\epsilon$ is mean-zero random noise.

We estimate the effect size $w$ via five distinct penalized regressions, specifically LASSO [14], elastic net [10], the minimax concave penalty (MCP) [13], the smoothly clipped absolute deviation (SCAD) [12], and MNet [11] . Thus, we seek to minimize the objective function

$$f(w) = \frac{1}{N}(Y - Xw)'(Y - Xw) + J_\lambda(w) = \frac{1}{N}Y'Y + \frac{1}{N}w'(X'X)w - \frac{2}{N}w'X'Y + J_\lambda(w), \qquad (1)$$

where $J_\lambda(w)$ is the primary penalty term. Then (1) contains $\frac{1}{N}X'X$, the linkage disequilibrium (LD) matrix for the SNPs, and $\frac{1}{N}X'Y$, the standardized marginal effect size for SNPs (correlation between SNPs and methylation levels). Denote the LD matrix with $R$ and the standardized marginal effects with $r = (r_1, ..., r_p)'$. We may also ignore $\frac{1}{N}Y'Y$, as it is not a function of $w$. Thus, reworking the objective function, we have

$$f(w) = w'Rw - 2w'r + J_\lambda(w).$$

We estimate $r$ with $\hat{r}$, which comes from z-scores calculated from the summary-level mQTL database. We estimate $R$ with $\hat{R}$, which comes from an LD reference panel (e.g., the 1000 Genomes Project [25]) used with a shrinkage estimator. Then the final objective function is

$$\tilde{f}(w) = w'\hat{R}w - 2w'\hat{r} + \theta w'w + J_\lambda(w),$$

where $\theta w'w,\ \theta \geq 0$ is an additional $L_2$ penalty designed to ensure a unique solution for $\hat{w}$. We then

optimize $\tilde{f}$.

The solution to the optimization, $\hat{w}$, is found according to the coordinate descent algorithm [26], which is known to converge to a local minimum. The choice of $\theta$ is restricted to the set $\{0.1, 0.2, ..., 0.9\}$, and a warm start is applied to generate a solution path for $\lambda$. These tuning parameters are chosen to maximize $R^2$ in a training dataset.

## 4.2 Methylation model training, tuning, and testing

The methylation imputation models were trained on summary-level mQTL data from GoDMC [9]. The mQTL associations, simply the effect sizes of SNPs on DNAm, are based on a sample size of 27,750 subjects of European ancestry from 36 independent studies, and 102,965,711 candidate *cis*-mQTLs were identified ($p < 1 \times 10^{-5}$, $\pm 1$ Mb from DNAm site) [9]. DNAm was profiled in blood with Illumina HumanMethylation450 BeadChips or EPIC arrays in at minimum 100 European ancestry subjects. The DNAm data were normalized and quality-controlled (described further in [9]) and adjusted for age, sex, predicted cell counts, predicted smoking, genetic PCs, and nongenetic PCs. In the case of related subjects, each of the same steps was performed in addition to fitting the genetic kinship matrix. Finally, residuals from the previous analyses were centered and standardized to have mean 0 and variance 1. Additionally, SNPs located in the MHC region (26 MB - 34 MB on chromosome 6) were removed due to its known complex LD structure [27].

To select tuning parameters for the model, such as the previously mentioned $\lambda$, and $\theta$, we employ methylation and genotype data from the Framingham Heart Study subcohort 1 ($n = 797$) [15]. In brief, DNAm was profiled using the Illumina HumanMethylation450 BeadChip. We focused on unrelated individuals of European ancestry. The DNAm data were quality controlled and normalized (as described in [4]) and adjusted for age, sex, six cell-type composition variables, and the top ten principal components (PCs) from the genotype data. The best tuning parameters were selected by examining $R^2$, the squared correlation between predicted and observed methylation.

To validate the model choices, we employed an independent subset of data from the Framingham Heart Study subcohort 2 [15], with no overlap from the previous tuning data ($n = 798$). The validation data was processed in accordance with the tuning data, and we selected models for methylation imputation with $R^2 \geq 0.005$.

### 4.3 Applying MIMOSA to GWAS

The MIMOSA models can be applied to both individual- and summary-level GWAS data. For individual-level GWAS data, use a generalized linear regression model:

$$f(E[P|X, C]) = \alpha C + \beta X\hat{w},$$

where $P$ is the phenotype for the GWAS, $X$ is the genotype data (and thus $X\hat{w}$ is the imputed methylation), $C$ is the covariance matrix, and $f$ is a chosen link function. This expression is used to test null hypothesis $H_0 : \beta = 0$, i.e., does methylation affect the phenotype.

For summary-level GWAS data, apply a burden-style test:

$$\hat{Z} = \frac{Z\hat{w}}{\sqrt{\hat{w}' L \hat{w}}},$$

where $Z$ is a vector of z-scores for all analyzed SNPs, $\hat{w}$ is the estimated effect size, and $L$ is the LD matrix for those SNPs (following SUMMIT, $L$ is estimated by a population LD reference panel such as from the 1000 Genomes project [25]).

While the above discussion on applying the MIMOSA models assumes only a single DNAm prediction model, MIMOSA can include up to five. When multiple methods build satisfactory prediction models for a single CpG site (DNAm prediction $R^2 > 0.005$), following SUMMIT, we conduct an association test as in the above procedure. The results are consolidated for the CpG site with the Cauchy combination test [28]. In detail, assuming $K$ satisfactory models, the test statistic

$$T = \sum_{i=1}^{K} \tilde{R_i}^2 \tan\left[(0.5 - p_i)\pi\right],$$

approximately follows the standard Cauchy distribution, and $\tilde{R_i}^2 = R_i^2 / \sum_{j=1}^{K} R_j^2$ and $p_i$ is the p-value for the $i$th model. The consolidated p-value is found with $p = 0.5 - \arctan(T)/\pi$. Notably, the test statistic $T$ is weighted by the DNAm prediction accuracy, so better fitting models are more impactful in the consolidated p-value.

### 4.4 Comparison with existing methods

To demonstrate efficacy and utility, we compared MIMOSA with the methylation imputation model weights generated by Baselmans et al. [2]. These weights were built from individual-level, whole-blood methylation data from the BIOS Consortium using 4,008 samples measured via Illumina 450k arrays. For each CpG site with an mQTL, a DNAm imputation model was created using LASSO penalized regression [14] with all nearby (within $2.5 \times 10^4$ bp) SNPs. Then, with the imputation models, summary-level GWAS for a given phenotype, and LD reference from the 1000 Genomes Project [25], a burden-style test based on test-statistic $\hat{Z} = \frac{Z\hat{w}}{\sqrt{\hat{w}' L \hat{w}}}$ is performed, exactly as with MIMOSA.

For the comparison, we first examined methylation prediction accuracy on the aforementioned validation set of data from the FHS, conducted by examining $R^2$ at each CpG site for which both methods possess weights. We tested the difference in the accuracies between the two methods with the Wilcoxon signed-rank test, which compares the locations of matched samples to determine if their population means differ. We further compared the methods by analyzing GWAS summary-level data for 28 complex traits. For each method and trait, we used the Bonferroni correction to determine significance level thresholds. In an effort to make an unbiased comparison, we further compared the methods for their shared CpG site set and used identical Bonferroni corrected significance cutoffs. We then tested the difference in the number of significant CpG sites for the two methods with the Wilcoxon signed-rank test

### Code availability statement

Code to use the MIMOSA models in MWAS is available on GitHub(`https://github.com/ChongWuLab/MIMOSA/tree/main`. Code and data to reproduce the results given in this study are available on OSF.IO `https://osf.io/4swdq/?view_only=1e824972813a40e78d5af25f4b2c7154`.

### Data availability

The mQTL summary data are available at `http://mqtldb.godmc.org.uk/downloads`. The datasets of FHS Offspring Cohort are publicly available via dbGaP (www.ncbi.nlm.nih.gov/gap):

dbGaP Study Accession: phs000342 and phs000724. The UK Biobank is an open-access resource but requires registration, available at `https://www.ukbiobank.ac.uk/researchers/`. The Baselmans models can be downloaded from `http://bbmri.researchlumc.nl/atlas/#data`. The 1000 Genomes Project data can be downloaded from `https://www.internationalgenome.org/data`. The genetic distance data for 1000 Genomes Project can be downloaded from `https://github.com/joepickrell/1000-genomes-genetic-maps`. The MIMOSA models are available at OSF.IO `https://osf.io/4swdq/?view_only=1e824972813a40e78d5af25f4b2c7154`. Real data results are available at `https://chongwulab.shinyapps.io/SUMMIT-app/`, where they can be easily downloaded.

### Acknowledgement

This study was partially supported by National Institutes of Health (R03 AG070669). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This study was conducted using the UK Biobank resource under Application Number 48240 (`https://www.ukbiobank.ac.uk/researchers/`). We thank Dr. Jihwan Ha at University of Hawaii Cancer Center for his help for this study. The authors would like to thank all of the individuals for their participation in the GWASs and UK Biobank and all the researchers, clinicians, technicians and administrative staff for their contribution to the studies and for making their GWAS summary results publicly available.

## References

[1] Peter A. Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature reviews. Genetics* 13.7 (July 2012), pp. 484–492. ISSN: 1471-0064.

[2] Bart M.L. Baselmans et al. "Multivariate genome-wide analyses of the well-being spectrum". In: *Nature Genetics* 51.3 (2019), pp. 445–451. ISSN: 15461718.

[3] Miruna C. Barbu et al. "Methylome-wide association study of antidepressant use in Generation Scotland and the Netherlands Twin Register implicates the innate immune system". In: *Molecular Psychiatry* October (2021). ISSN: 14765578.

[4] Lang Wu et al. "An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk". In: *Nature Communications 2020 11:1* 11.1 (Aug. 2020), pp. 1–11. ISSN: 2041-1723.

[5] Alexander Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature Genetics* (2016). ISSN: 15461718.

[6] Eric R. Gamazon et al. "A gene-based association method for mapping traits using reference transcriptome data". In: *Nature Genetics* 47.9 (2015), pp. 1091–1098. ISSN: 15461718.

[7] Yiming Hu et al. "A statistical framework for cross-tissue transcriptome-wide association analysis". In: *Nature Genetics 2019 51:3* 51.3 (Feb. 2019), pp. 568–576. ISSN: 1546-1718.

[8] Zichen Zhang et al. "SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification". In: *Nature Communications 2022 13:1* 13 (1 Oct. 2022), pp. 1–12. ISSN: 2041-1723.

[9] Josine L. Min et al. "Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation". In: *Nature Genetics* 53.9 (2021), pp. 1311–1321. ISSN: 15461718.

[10] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (Apr. 2005), pp. 301–320. ISSN: 1467-9868.

[11] Jian Huang et al. "THE Mnet METHOD FOR VARIABLE SELECTION". In: *Statistica Sinica* 26 (2016), pp. 903–923.

[12] Jianqing Fan and Runze Li. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". In: *Journal of the American Statistical Association* 96 (456 Dec. 2001), pp. 1348–1360. ISSN: 0162-1459.

[13] Cun-Hui Zhang. "Nearly unbiased variable selection under minimax concave penalty". In: *The Annals of Statistics* 38 (2 2010), pp. 894–942.

[14] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Jan. 1996), pp. 267–288. ISSN: 2517-6161.

[15] William B. Kannel et al. "An Investigation of Coronary Heart Disease in Families: The Framingham Offspring Study". In: *American Journal of Epidemiology* 110.3 (Sept. 1979), pp. 281–290. ISSN: 0002-9262.

[16] Douglas W. Yao et al. "Quantifying genetic effects on disease mediated by assayed gene expression levels". In: *Nature Genetics 2020 52:6* 52.6 (May 2020), pp. 626–633. ISSN: 1546-1718.

[17] Brandon L. Pierce and Stephen Burgess. "Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators". In: *American journal of epidemiology* 178.7 (Oct. 2013), pp. 1177–1184. ISSN: 1476-6256.

[18] George D. Smith and Shah Ebrahim. "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" In: *International Journal of Epidemiology* 32.1 (Feb. 2003), pp. 1–22. ISSN: 0300-5771.

[19] A. Belloni et al. "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain". In: *Econometrica* 80 (6 Nov. 2012), pp. 2369–2429. ISSN: 1468-0262.

[20] Marie Verbanck et al. "Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases". In: *Nature Genetics 2018 50:5* 50.5 (Apr. 2018), pp. 693–698. ISSN: 1546-1718.

[21] Simone Wahl et al. "Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity". In: *Nature 2016 541:7635* 541.7635 (Dec. 2016), pp. 81–86. ISSN: 1476-4687.

[22] John C. Chambers et al. "Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study". In: *The lancet. Diabetes & endocrinology* 3.7 (July 2015), p. 526. ISSN: 22138595.

[23] Ruth Pidsley et al. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". In: *Genome Biology* 17 (1 Oct. 2016), pp. 1–17. ISSN: 1474760X.

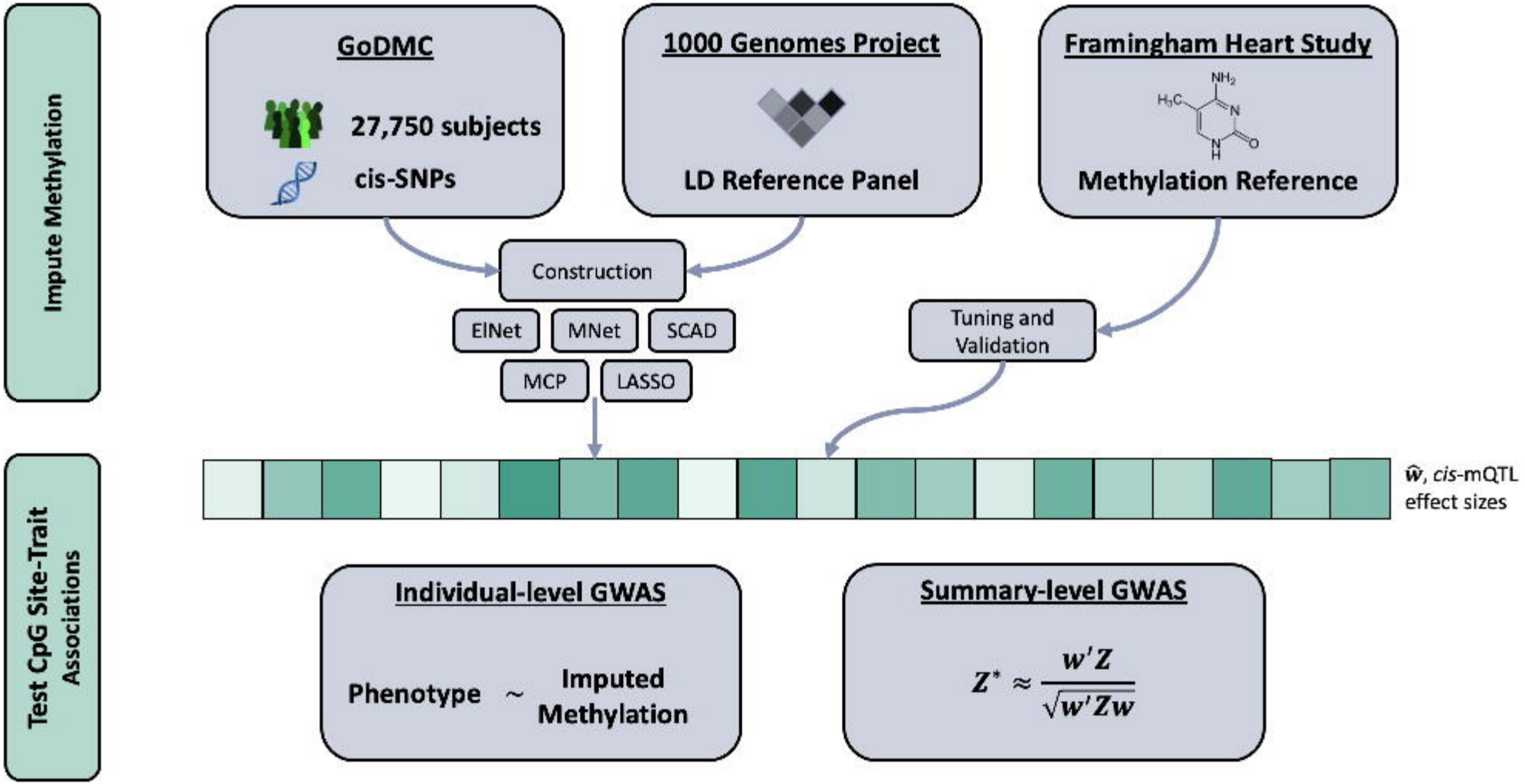
[24] Dan Zhou et al. "A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis". In: *Nature Genetics 2020 52:11* 52.11 (Oct. 2020), pp. 1239–1246. ISSN: 1546-1718.

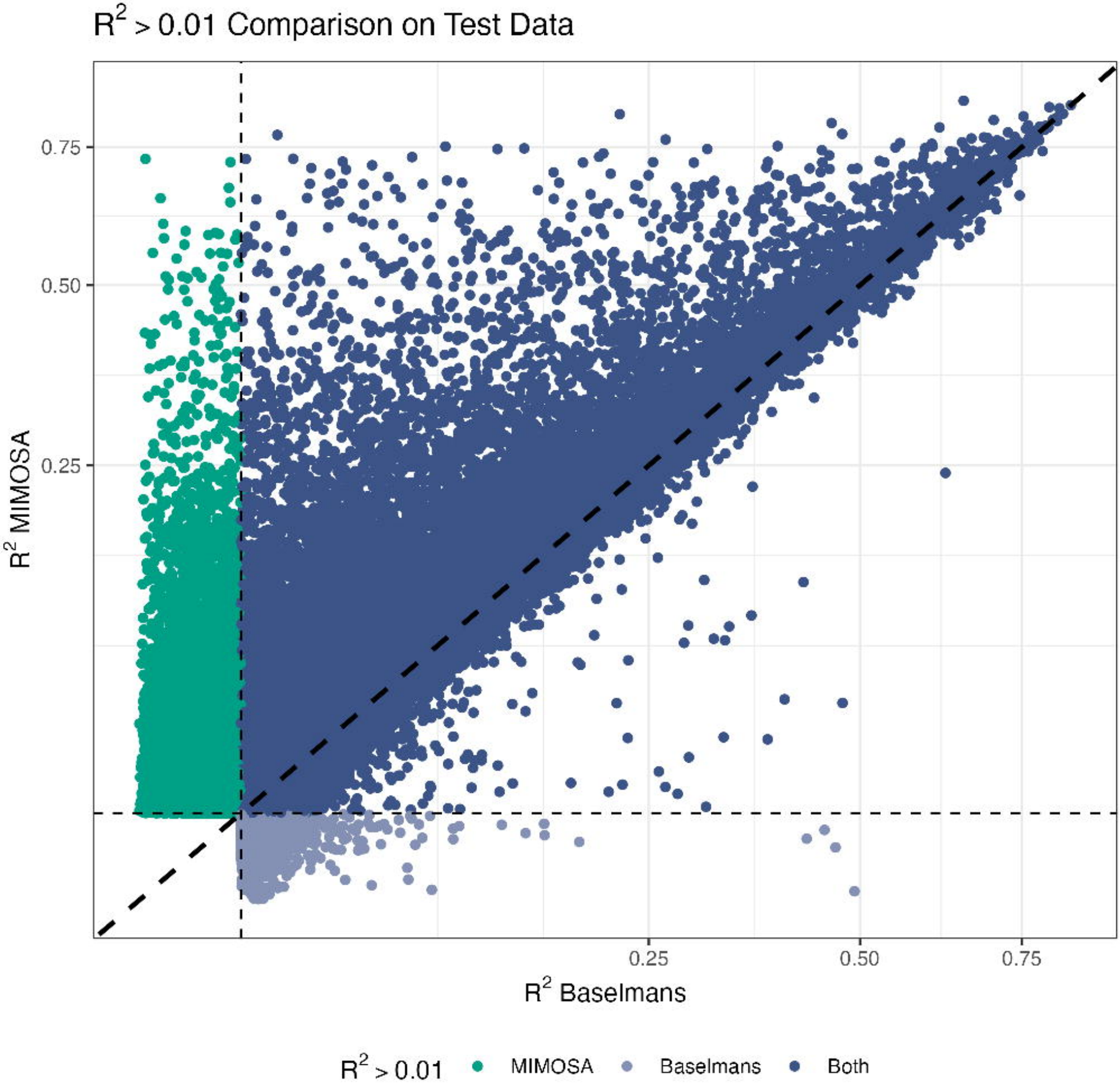
[25] 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature 2015 526:7571* 526.7571 (Sept. 2015), pp. 68–74. ISSN: 1476-4687.

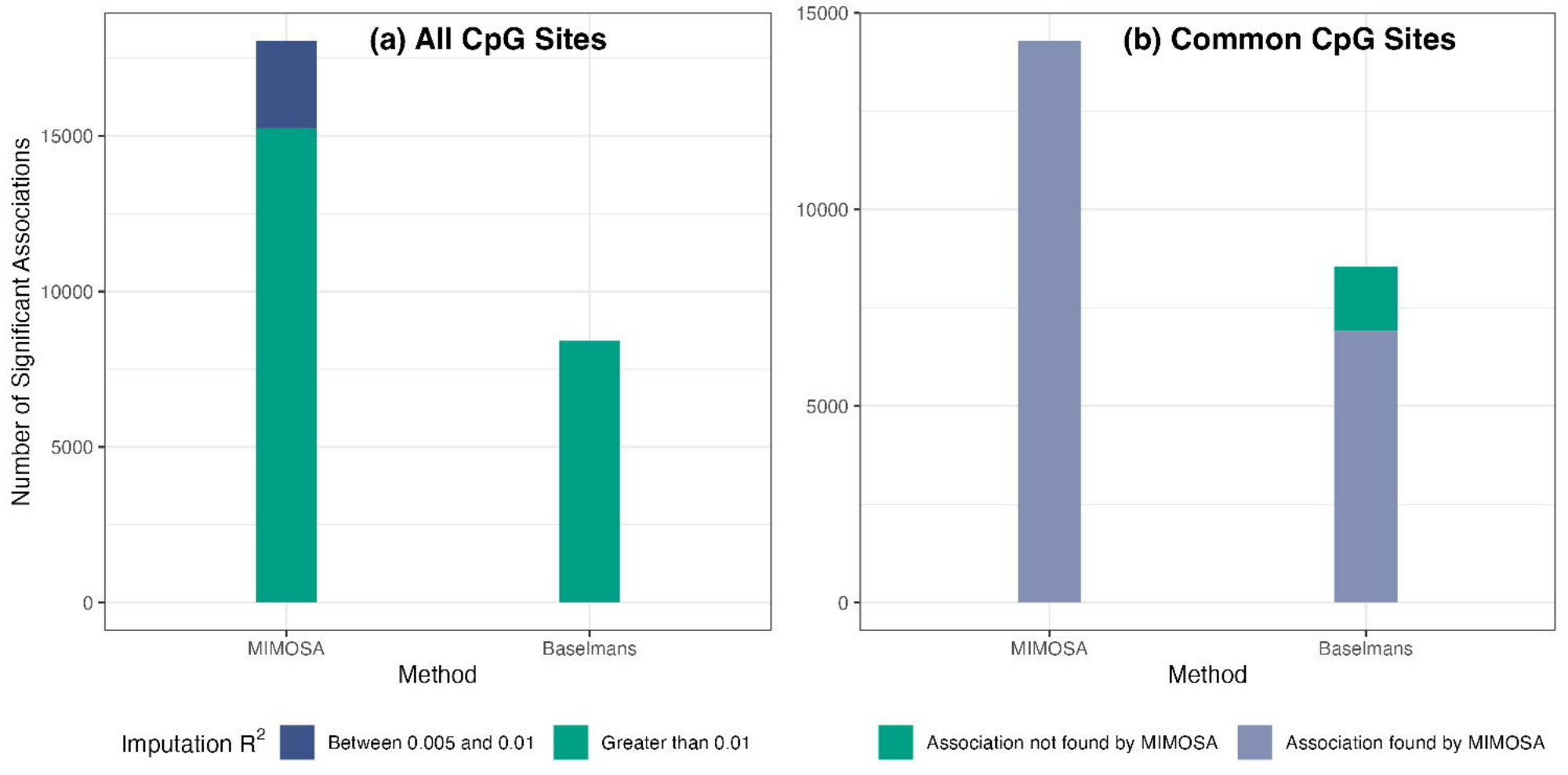
[26] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of statistical software* 33.1 (2010), p. 1. ISSN: 15487660.

[27] Jie Zheng et al. "Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases". In: *Nature Genetics* 52.10 (2020), pp. 1122–1131. ISSN: 15461718.

[28] Xuanyao Liu, Yang I. Li, and Jonathan K. Pritchard. "Trans Effects on Gene Expression Can Drive Omnigenic Inheritance". In: *Cell* 177.4 (May 2019), 1022–1034.e6. ISSN: 1097-4172.


Impute Methylation
GoDMC
27,750 subjects
cis-SNPs
1000 Genomes Project
LD Reference Panel
Framingham Heart Study
Methylation Reference
Construction
ElNet
MNet
SCAD
MCP
LASSO
Tuning and Validation
$\hat{w}$, cis-mQTL effect sizes
Test CpG Site-Trait Associations
Individual-level GWAS
Phenotype ~ Imputed Methylation
Summary-level GWAS
$Z^* \approx \frac{w'Z}{\sqrt{w'Zw}}$


$R^2 > 0.01$ Comparison on Test Data
0.75
0.50
0.25
$R^2$ MIMOSA
0.25
0.50
0.75
$R^2$ Baselmans
$R^2 > 0.01$
MIMOSA
Baselmans
Both


(a) All CpG Sites
(b) Common CpG Sites
Number of Significant Associations
15000
10000
5000
0
MIMOSA
Baselmans
Method
Imputation $R^2$
Between 0.005 and 0.01
Greater than 0.01
Association not found by MIMOSA
Association found by MIMOSA