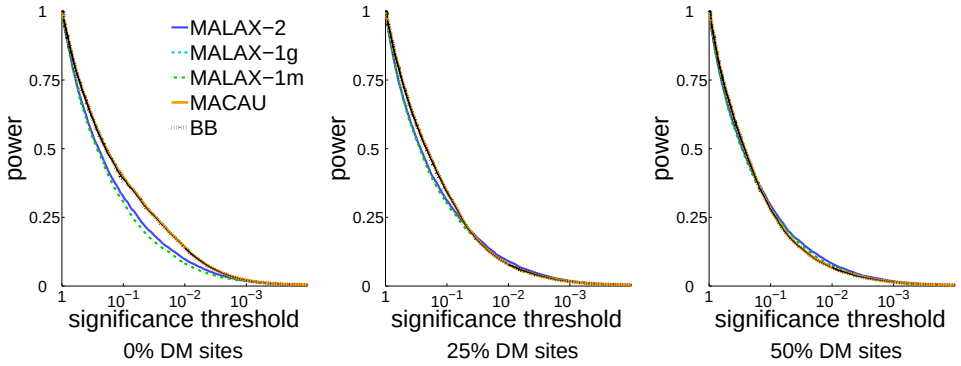
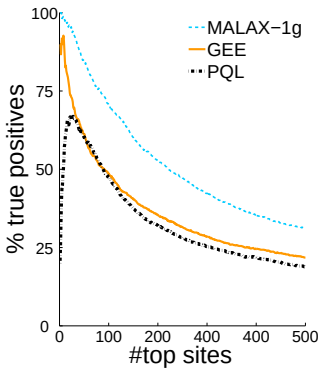
Association testing in bisulfite sequencing data via a Laplace approximation: Supplementary Information

1 Supplementary Figures



Supplementary Figure S1: Power measures for the evaluated methods under various proportions of differentially methylated sites. All results are averaged over 10 simulated data sets.



Supplementary Figure S2: The detection power of MALAX-1g and two additional evaluated methods under simulated data sets with no DM sites. The additional methods are GEE (based on CARAT) and PQL (based on GMMAT), as described in the supplementary note. All results are averaged over 10 simulated data sets.

2 Experiments with Additional Methods

We evaluated two additional methods for generalized linear mixed model (GLMM) approximation. Both methods were recently proposed for case-control association studies, but can be readily adapted to work with count-bases responses instead of binary responses.

The first method is based on the penalized quasi likelihood algorithm, as implemented in the GMMAT package [1]. This method is similar to the Laplace approximation but applies additional approximations, which results in faster computations at the cost of reduced accuracy.

The second method uses a generalized estimating equations (GEE) approach, as implemented in the CARAT package [2]. This method only models the first two moments of the model likelihood. We used a modified version of CARAT that was adapted to use a binomial instead of a binary response. We evaluated both a standard version and a version with an additional dispersion parameter, as described in [2]; both versions yielded effectively the same results.

GMMAT and CARAT were evaluated in a setting with no differentially methylated (DM) sites, which was the easiest setting in our experiments. Consequently, both methods used one variance component associated with a genetic similarity matrix, along with one variance component associated with the identity matrix, which can account for independent over-dispersion. The results demonstrate that MALAX-1g substantially outperformed both methods (Supplementary Figure S2). We therefore did not consider these methods in the remainder of the experiments. We note that we also evaluated versions of these approaches that use beta-binomial instead of binomial responses, but these versions were less accurate than the binomial ones.

3 Gradient Computation

The gradient of the approximate log likelihood described in the main text is required both for approximating the Hessian and for the maximum likelihood estimation procedure. Here we derive the gradient computation in detail. We first explicitly write the approximate log likelihood as follows:

$$\log P\left(\boldsymbol{y}^j \mid \boldsymbol{x}, \boldsymbol{W}, \boldsymbol{r}^j\right) \approx \underbrace{-\frac{1}{2}\left(\hat{\boldsymbol{l}}-\boldsymbol{m}\right)^T \boldsymbol{G}^{-1}\left(\hat{\boldsymbol{l}}-\boldsymbol{m}\right)}_{L_1} + \underbrace{\sum_{i=1}^{n} \log P\left(y_i^j \mid \hat{l}_i\right)}_{L_2} \underbrace{-\frac{1}{2}\log|\boldsymbol{B}|}_{L_3}, \tag{1}$$

where $\boldsymbol{B} \triangleq \boldsymbol{G}\boldsymbol{A}$. We denote the three terms on the right hand side of the above equation as L_1, L_2 and L_3, respectively. Note that the quantities $\boldsymbol{G}$ and $\boldsymbol{m}$ in L_1 depend explicitly on the model parameters, but $\hat{\boldsymbol{l}}$ and $\boldsymbol{A}$ also implicitly depend on these parameters, where the dependence of $\boldsymbol{A}$ is mediated entirely through its dependence on $\hat{\boldsymbol{l}}$. We therefore divide the partial derivative according to each parameter θ (which can represent variance components or fixed effects) into its explicit and implicit components, by using the chain rule as follows:

$$\frac{\partial \log P\left(\boldsymbol{y}^j \mid \boldsymbol{x}, \boldsymbol{W}, \boldsymbol{r}^j\right)}{\partial \theta} = \frac{\partial L_1}{\partial \theta}\Big|_{\text{explicit}} + \sum_{i=1}^{n} \frac{\partial \log P\left(\boldsymbol{y}^j \mid \boldsymbol{x}, \boldsymbol{W}, \boldsymbol{r}^j\right)}{\partial \hat{l}_i} \frac{\partial \hat{l}_i}{\partial \theta}. \tag{2}$$

We first derive the explicit components:

$$\begin{aligned}
\frac{\partial L_1}{\partial \sigma_v^2}|_{\text{explicit}} &= -\frac{1}{2}\left(\hat{\boldsymbol{l}}-\boldsymbol{m}\right)^T \boldsymbol{G}^{-1}\boldsymbol{K}_v\boldsymbol{G}^{-1}\left(\hat{\boldsymbol{l}}-\boldsymbol{m}\right) - \frac{1}{2}\text{tr}\left[\boldsymbol{G}^{-1}\boldsymbol{K}_v\right] \\
\frac{\partial L_1}{\partial \sigma_e^2}|_{\text{explicit}} &= -\frac{1}{2}\left(\hat{\boldsymbol{l}}-\boldsymbol{m}\right)^T \boldsymbol{G}^{-2}\left(\hat{\boldsymbol{l}}-\boldsymbol{m}\right) - \frac{1}{2}\text{tr}\left[\boldsymbol{G}^{-1}\right] \\
\frac{\partial L_1}{\partial \boldsymbol{\gamma}}|_{\text{explicit}} &= -\left(\hat{\boldsymbol{l}}-\boldsymbol{m}\right)^T \boldsymbol{G}^{-1}\boldsymbol{C}.
\end{aligned} \tag{3}$$

Here, $\boldsymbol{C}$ is the matrix of covariates for the entire sample (the matrix $\boldsymbol{W}$ with an appended column $\boldsymbol{x}$), and $\boldsymbol{\gamma} = \left[\boldsymbol{\alpha}^T \ \beta\right]^T$ is the vector of all fixed effects.

To derive the implicit components, we first note that $\frac{\partial (L_1+L_2)}{\partial \boldsymbol{l}}|_{\boldsymbol{l}=\hat{\boldsymbol{l}}} = \boldsymbol{0}$ by definition. We therefore only need to compute $\frac{\partial \log|\boldsymbol{B}|}{\partial \hat{l}_i}\frac{\partial \hat{l}_i}{\partial \theta}$. To derive $\log|\boldsymbol{B}|$, we first explicitly write the negative Hessian $\boldsymbol{A}$ as follows:

$$\boldsymbol{A} \triangleq -\nabla\nabla \log \text{P}\left(\boldsymbol{l} \mid \boldsymbol{x}, \boldsymbol{W}, \boldsymbol{y}^j, \boldsymbol{r}^j\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}} = -\nabla\nabla \log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}} + \boldsymbol{G}^{-1}. \tag{4}$$

Using this explicit notation, $\frac{\partial \log|\boldsymbol{B}|}{\partial \hat{l}_i}$ is given by:

$$\frac{\partial \log|\boldsymbol{B}|}{\partial \hat{l}_i} = -\frac{1}{2}\text{tr}\left[\boldsymbol{B}^{-1}\frac{\partial B}{\partial \hat{l}_i}|_{\boldsymbol{l}=\hat{\boldsymbol{l}}}\right] = \frac{1}{2}\text{tr}\left[\boldsymbol{B}^{-1}\boldsymbol{G}\text{diag}\left(\frac{\partial^3}{\partial l_i^3}\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}}\right)\right], \tag{5}$$

where $\frac{\partial B}{\partial \hat{l}_i}$ is a matrix of element-wise partial derivatives.

Finally, $\frac{\partial \hat{\boldsymbol{l}}_i}{\partial \theta}$ can be derived by using the fact that $\hat{\boldsymbol{l}} = \mathbf{G}\nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}} + \boldsymbol{m}$, as explained in the main text. Therefore, for every parameter θ we have:

$$\frac{\partial \hat{\boldsymbol{l}}}{\partial \theta} = \frac{\partial \boldsymbol{G}}{\partial \theta}\nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}} + \boldsymbol{G}\frac{\partial\left(\nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}}\right)}{\partial \hat{\boldsymbol{l}}}\frac{\partial \hat{\boldsymbol{l}}}{\partial \theta} + \frac{\partial \boldsymbol{m}}{\partial \theta}, \tag{6}$$

where the partial derivatives over matrices and vectors are computed element-wise. Writing these gradients explicitly for the model parameters, we have:

$$\begin{aligned}
\frac{\partial \hat{\boldsymbol{l}}}{\partial \sigma_v^2} &= \boldsymbol{K}_v\nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}} + \boldsymbol{G}\nabla\nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}}\frac{\partial \hat{\boldsymbol{l}}}{\partial \sigma_v^2} \\
\frac{\partial \hat{\boldsymbol{l}}}{\partial \sigma_e^2} &= \nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}} + \boldsymbol{G}\nabla\nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}}\frac{\partial \hat{\boldsymbol{l}}}{\partial \sigma_e^2} \\
\frac{\partial \hat{\boldsymbol{l}}}{\partial \alpha_k} &= \boldsymbol{G}\nabla\nabla\left(\log P\left(\boldsymbol{y}^j \mid \boldsymbol{l}\right)\right)|_{\boldsymbol{l}=\hat{\boldsymbol{l}}}\frac{\partial \hat{\boldsymbol{l}}}{\partial \alpha_k} + \boldsymbol{C}_k,
\end{aligned} \tag{7}$$

where $\boldsymbol{C}_k$ is the k^{th} column of $\boldsymbol{C}$.

A careful analysis reveals that the only matrix that needs to be inverted is $\boldsymbol{B}$ [3]. This is also the matrix whose determinant needs to be evaluated in the likelihood evaluation. Both quantities can be readily evaluated given the Cholesky decomposition of $\boldsymbol{B}$, whose computation scales cubically with the sample size. Hence, gradient computation incurs only a minor additional computational overhead over the likelihood approximation.

4 Data Simulations

The simulations procedure consisted of several stages: Genotype simulations, covariates simulations, cell-type simulations, phenotype simulation, methylation simulations and observed reads simulations. We describe each stage in turn.

4.1 Genome simulations

We simulated population structure via the Balding-Nichols model [4], wherein allele frequencies in the range [0.05, 0.5] were randomly drawn for an ancestral population, and frequencies for several subpopulations were drawn from a Beta distribution with parameters $f\left(1-F_{ST}\right)/F_{ST}$ and $\left(1-f\right)\left(1-F_{ST}\right)/F_{ST}$, where f is the minor allele frequency in the ancestral population.

Next, we generated a mixture vector $\boldsymbol{M}^i$ for every individual i from a symmetric Dirichlet distribution with a concentration parameter 0.25, such that $\sum_u \boldsymbol{M}^i_u = 1$, and u iterates over subpopulations. Finally, 60,000 SNPs were generated for every individual assuming Hardy-Weinberg equilibrium by sampling every SNP j of individual i from $\mathrm{Bin}(2, p^i_j)$, where $p^i_j = \sum_u \boldsymbol{M}^i_u f^j_u$, and f^j_u is the MAF of SNP j in population u.

4.2 Covariate simulations

A vector of covariates for every individual i, $\boldsymbol{C}^i$, was generated by sampling each entry from a standard normal distribution, and adding an additional entry with the value 1.0 as an intercept.

4.3 Cell-type simulations

A vector of cell-types for every individual i, $\boldsymbol{T}^i$, was generated from a non-symmetric Dirichlet distribution with concentration parameters 1.33 , 0.93 , 4.299, 2.696, 16.344. These values were fitted using estimated cell type proportions for the GSE42861 dataset [5], a 450K array data set. Specifically, we obtained cell proportion estimates of five cell types (granulocytes, monocytes, B cells, NK cells, T cells) using the default implementation available in the minfi package [6], defined and assembled for the 450K array [7] based on the approach suggested by [8] and a 450K reference data set [9].

4.4 Phenotype simulation

To generate a phenotype, we first sampled effect sizes for the factors affecting the phenotype. Specifically, we sampled an effect size $\gamma^{\mathrm{snp}}_l \sim \mathcal{N}(0, \sigma^2_{\mathrm{snp}})$ for every causal SNP l, an effect size $\gamma^{\mathrm{pop}}_u \sim \mathcal{N}(0, \sigma^2_{\mathrm{pop}})$ for every subpopulation u (intended to model shared environmental factors within the subpopulations), and an effect size $\gamma^{\mathrm{cell}}_h \sim \mathcal{N}(0, \sigma^2_{\mathrm{cell}})$ for every cell-type h.

Afterwards, the phenotype y^i for every individual i was generated as follows:

$$y^i = + \sum_{l \in \text{Causal}} s_l^i \gamma_l^{\text{snp}} + \sum_j m_u^i \gamma_u^{\text{pop}} + \sum_h t_h^i \gamma_h^{\text{cell}} + \epsilon^i, \tag{8}$$

where s_l^i, m_u^i and t_h^i are the l^{th} causal SNP, u^{th} subpopulation ancestry fraction and h^{th} cell-type fraction of individual i, respectively, $\epsilon^i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is an environmental effect for individual i where σ_ϵ^2 guarantees that the phenotype variance is 1.0, and l iterates over indices of causal SNPs only. All effects were normalized to have a zero mean and unit variance.

4.5 Methylation simulation

Methylation levels π_j^i for every site j of every individual i were generated as follows: First, either 0%, 25% or 50% of the sites were randomly selected to be differentially methylated. Next, for every site j we generated effect sizes for the factors affecting the methylation. Specifically, we sampled an effect size $\delta_k^{j,\text{covar}} \sim \mathcal{N}(0, \nu_{\text{covar}}^{2,j})$ for every covariate k and site j, an effect size $\delta_l^{j,\text{snp}} \sim \mathcal{N}(0, \nu_{\text{snp}}^{2,j})$ for every causal SNP l and site j, an effect size $\delta_u^{j,\text{pop}} \sim \mathcal{N}(0, \nu_{\text{pop}}^{2,j})$ for every subpopulation u and site j (intended to model shared environmental factors within the subpopulations), an effect size $\delta_h^{j,\text{cell}} \sim \mathcal{N}(0, \nu_{\text{cell}}^{2,j})$ for every cell-type h and site j, and an effect size $\delta^{j,\text{pheno}} \sim \mathcal{N}(0, \nu_{\text{pheno}}^{2,j})$ for the phenotype. We also generated an environmental effect $e_j^i \sim \mathcal{N}(0, \nu_{\text{env}}^{2,j})$ for every site j of individual i.

Next, we generated $\text{logit}\left(\pi_j^{i,h}\right)$ for every cell type h of site j of individual i as follows:

$$\text{logit}\left(\pi_j^{i,h}\right) = \sum_k c_k^i \delta_k^{k,\text{covar}} + \sum_{l \in \text{Causal}} s_l^i \delta_l^{j,\text{snp}} + \sum_u m_u^i \delta_u^{j,\text{pop}} + \delta_h^{j,\text{cell}} + y^i \delta^{j,\text{pheno}} + e_j^i. \tag{9}$$

In the next step, the methylation level of every cell type h of site j of individual i, $\pi_j^{i,h}$, was computed as $\pi_j^{i,h} = 1/\left(1 + \exp\left(-\text{logit}\left(\pi_j^{i,h}\right)\right)\right)$. Finally, a combined methylation level π_j^i for every individual i of every site j was computed via:

$$\pi_j^i = \sum_h t_h^i \pi_j^{i,h}. \tag{10}$$

4.6 Observed reads simulation

To simulate observed reads, we first sampled a total number of reads r_j^i for every site j of individual i from a negative Binomial distribution with parameters $n = 1.135515$, $p = 0.047623$. These values were fitted from the distribution of total number of reads of the Baboons data studied in the original MACAU paper [10].

Afterwards, an observed number of reads y_j^i for every site j of individual i was sampled from $\text{Bin}(r_j^i, \pi_j^i)$.

4.7 Default parameters

Unless otherwise stated, the default parameters used in the simulations are the following: Each data set consisted of 200 individuals, 3 covariates (including an intercept), 60,000 SNPs, 500 causal SNPs, 10,000 non-causal sites and 500 causal sites. The ancestry of every individual was a mixture of four different populations, sampled from a symmetric Dirichlet distribution with concentration parameter of 0.25. Additionally, every individual contained a mixture of 5 cell types, sampled from a Dirichlet distribution with parameters 1.33 , 0.93 , 4.299, 2.696, 16.344.

The default effect variances were the following: $\sigma^2_{\text{snp}} = 0.25$ / #SNPs, $\sigma^2_{\text{pop}} = 0.025$ / #populations, $\sigma^2_{\text{cell}} = 0.25$ / #cell-types, $\nu^{2,j}_{\text{covar}} = 0.01$ / #covariates, $\nu^{2,j}_{\text{snp}} = 1$ / #SNPs, $\nu^{2,j}_{\text{pop}} = 0.1$ / #populations, $\nu^{2,j}_{\text{pheno}} = 0.025$, $\nu^{2,j}_{\text{env}} = 0.1$, $\nu^{2,j}_{\text{cell}} = 5.0$.

References

[1] H. Chen et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am. J. Hum. Genet.* 98(4) (2016), 653–66.

[2] D. Jiang, S. Zhong, and M.S. McPeek. Retrospective binary-trait association test elucidates genetic architecture of Crohn disease. *Am. J. Hum. Genet.* 98(2) (2016), 243–255.

[3] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[4] D.J. Balding and R.A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96(1-2) (1995), 3–12.

[5] Y. Liu et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* 31(2) (2013), 142–147.

[6] M.J. Aryee et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10) (2014), 1363–1369.

[7] A.E. Jaffe and R.A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* 15(2) (2014), 1.

[8] E.A. Houseman et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform.* 13(1) (2012), 1.

[9] L.E. Reinius et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLOS ONE* 7(7) (2012), e41361.

[10] A.J. Lea, J. Tung, and X. Zhou. A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data. *PLoS Genet.* 11(11) (2015), e1005650.