

A Bayesian Exploration Into More Flexible *trans*-Methylation Quantitative Trait Locus (trans-mQTL) Mapping

DNA Methylation

DNA methylation (DNAm) occurring at genomic CpG sites is a crucial epigenetic process implicated in aging, complex disease, and more. Certain single nucleotide polymorphisms (SNPs) called **methylation quantitative trait loci (mQTLs)** help explain variation in DNAm; in particular, *trans* genetic variation accounts for the **vast majority** of that which explains DNAm variability (Gaunt et al., 2016).

mQTL Mapping Limitations

Current statistical methods struggle to adequately map the *trans* type of mQTLs due to their relatively **weak distal effects**. Limitations have been addressed in adjacent fields of *trans*-expression and protein QTL (eQTL, pQTL) mapping but have yet to be applied to *trans*-mQTLs:

Univariate framework

Independent modeling of all candidate mQTL associations **neglects** known pleiotropic and polygenic mechanisms.

Assumption of linearity

Such an assumption may be **too stringent** for real-life manifestations of weak effects.

Research Question

How can alternative methodologies from previous work in the related fields of *trans*-eQTL and pQTL mapping be **combined** into a **robust, scalable statistical workflow** to more effectively map *trans*-mQTLs?

bayesmqtl Novelty Combines Approaches From Previous Work

Modeling Stage 1: Beta Mixture Models to Represent DNA Methylation

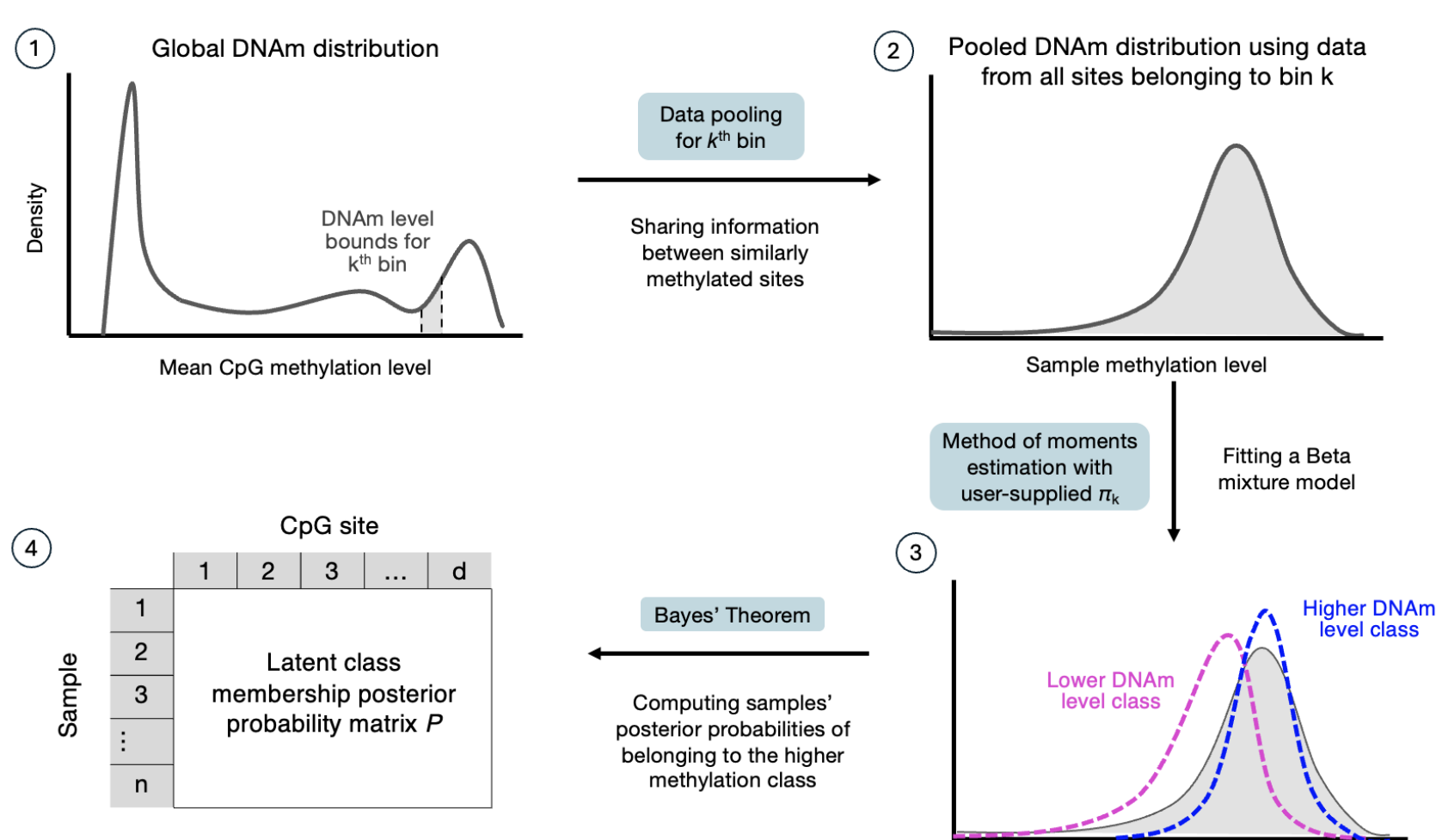


Figure created by finalist using PowerPoint, 2024.

Stage 2: Joint Bayesian Latent Class Membership Modeling

Posterior probabilities are probit-transformed into response variables z_t , whose values are determined by predictor SNPs via **Bayesian multiple linear regression**.

$$z_t | \theta_t, \tau_t \sim \mathcal{N}(X\theta_t, I_n \tau_t^{-1}), \quad \tau_t \stackrel{\text{ind}}{\sim} \text{Gamma}(\eta_t, \kappa_t),$$

$$\theta_{st} | \sigma_s^2 \sim \mathcal{N}(0, \sigma_s^2), \quad \sigma_s^2 \sim \text{Exp}(\lambda_s).$$

Model parameters are estimated using a **mean-field variational inference (MFVI) algorithm**, specifically modified by a **temperature** parameter T to yield an **annealed** variant.

$$q(\Theta) = \left\{ \prod_{t=1}^d \prod_{s=1}^p q(\theta_{st}) \right\} \left\{ \prod_{s=1}^p q(\sigma_s^2) \right\} \left\{ \prod_{t=1}^d q(\tau_t) \right\},$$

$$\log q_T(\Theta_j) = T^{-1} E_{-j} \{ \log p(z, \Theta) \} + \text{cst}, \quad j = 1, \dots, J.$$

The **posterior means** of the latent class membership regression coefficients θ are used to infer mQTL associations. A SNP behaves as an mQTL to a CpG if it alters the **probability** of a sample at that CpG **belonging to the higher methylation class**.

Design choice motivations

Beta Mixture Models

Beta distributions can capture the **heteroskedastic** and **bounded** nature of real-life DNAm data (beta values). Measuring the probability of each sample belonging to the higher methylation class at each CpG yields new data that permits **more flexible** treatment of mQTL effects.

Hierarchical Structure

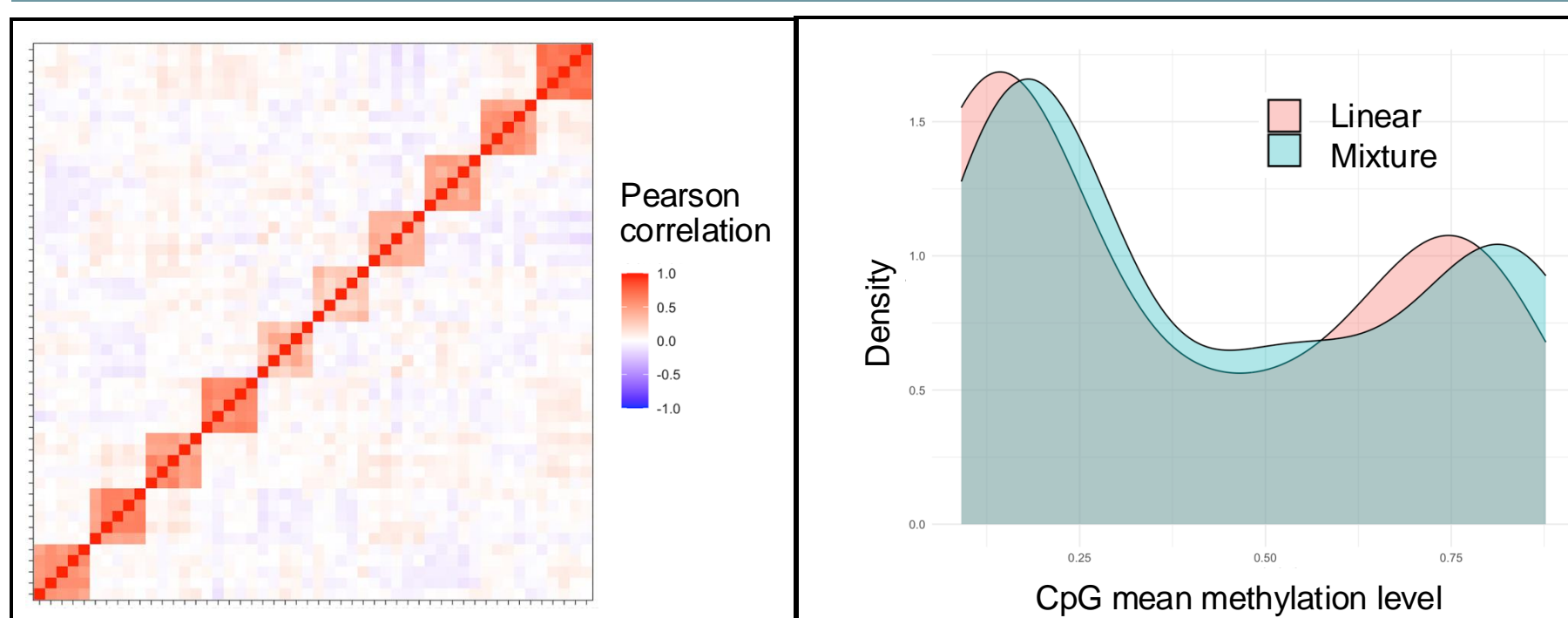
SNP-shared residual variances and CpG-shared effect size variances permit the representation of **polygenic** and **pleiotropic** effects, respectively. The **Laplace prior** on regression coefficients also enforces **sparsity** whilst averting the computational burden of a two-state prior.

Annealed MFVI

MFVI with conjugate priors ensures a **fast, deterministic** parameter estimation algorithm that scales **linearly** with respect to the number of SNPs and CpGs. Annealing **improves exploration** of highly multimodal parameter spaces, as is common in large genomic data settings.

Model Evaluation and Project Impact

Principled Data Simulation



Left: $p = 50$ SNPs in Hardy-Weinberg equilibrium, equicorrelated by 5-SNP blocks for $n = 400$ individuals, generated using the **echoseq** R package (Ruffieux 2020). Random choice of 10 labels for the "active" predictors (SNPs associated with ≥ 1 response). **Right:** $d = 50$ CpGs with DNAm levels generated under either a mixture model (SNPs determine likelihood of higher methylation class) or additive dose-effect (SNPs directly determine DNAm level) assumption. Random choice of 20 labels for the "active" responses. Both figures created by finalist using R, 2024.

Results

	Mixture model scenario		Additive dose-effect model scenario	
	AUROC	Runtime	AUROC	Runtime
bayesmqtl	0.685	0.231 seconds	0.577	0.215 seconds
Matrix eQTL	0.776	0.011 seconds	0.608	0.012 seconds
LOCUS	0.821	0.028 seconds	0.608	0.030 seconds

Table created by finalist, 2024.

Areas of Improvement

Although **bayesmqtl** exhibits the weakest statistical performance and slowest runtime in all simulation schemes, the differences in performance **vastly shrink** in the additive dose-effect scenario. Thus, **bayesmqtl** may significantly outperform existing methods given testing on real data and necessary modifications in the following:

Statistical Power

More informed, generalizable binning procedure and method of choosing the mixture proportion π_k

Computational Efficiency

Implement computationally intensive subroutines in **C++**

Other one-state sparse prior distribution on the regression coefficients (e.g., **horseshoe prior**)

Parallelize parameter estimation across multiple CPU cores

Key Takeaways

- Presented the **first** study addressing methodological limitations in the specific context of *trans*-mQTL mapping.
- bayesmqtl** encapsulates a novel suite of mathematical techniques to address this gap within a **highly flexible, practical** software implementation.
- With necessary modification and testing on real-life datasets, **bayesmqtl** holds **significant promise** for improved *trans*-mQTL identification, ultimately facilitating **deeper understanding** of complex disease pathophysiology, aging, and more.