A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data

AJ Lea[1], SC Alberts[1,2], J Tung[1,2,3,4]*, X Zhou[5,6]*†

*These authors contributed equally to this work.

1. Department of Biology, Duke University, Box 90338, Durham, NC 27708, USA

2. Institute of Primate Research, National Museums of Kenya, P. O. Box 24481, Karen 00502, Nairobi, Kenya

3. Department of Evolutionary Anthropology, Duke University, Box 90383, Durham, NC 27708, USA

4. Duke University Population Research Institute, Duke University, Box 90420, Durham, NC 27708, USA

5. Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109

6. Center for Statistical Genetics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109

†Corresponding author email: xzhousph@umich.edu

Other author emails: amanda.lea@duke.edu, alberts@duke.edu, jt5@duke.edu

Abstract

Identifying sources of variation in DNA methylation levels is important for understanding gene regulation. Recently, bisulfite sequencing has become a popular tool for estimating DNA methylation levels at base-pair resolution, and for investigating the major drivers of epigenetic variation. However, modeling bisulfite sequencing data presents several challenges. Methylation levels are estimated from proportional read counts, yet coverage can vary dramatically across sites and samples. Further, methylation levels are influenced by genetic variation, and controlling for genetic covariance (e.g., kinship or population structure) is crucial for avoiding potential false positives. To address these challenges, we combine a binomial mixed model with an efficient sampling-based algorithm (MACAU) for approximate parameter estimation and *p*-value computation. This framework allows us to account for both the over-dispersed, count-based nature of bisulfite sequencing data, as well as genetic relatedness among individuals. Furthermore, by leveraging the advantages of an auxiliary variable-based sampling algorithm and recent mixed model innovations, MACAU substantially reduces computational complexity and can thus be applied to large, genome-wide data sets. Using simulations and two real data sets (whole genome bisulfite sequencing (WGBS) data from *Arabidopsis thaliana* and reduced representation bisulfite sequencing (RRBS) data from baboons), we show that, compared to existing approaches, our method provides better calibrated test statistics in the presence of population structure. Further, it improves power to detect differentially methylated sites: in the RRBS data set, MACAU detected 1.6-fold more age-associated CpG sites than a beta-binomial model (the next best approach). Changes in these sites are consistent with known age-related shifts in DNA methylation levels, and are enriched near genes that are differentially expressed with age in the same population. Taken together, our results indicate that MACAU is an effective tool for analyzing bisulfite sequencing data, with particular salience to analyses of structured populations. MACAU is freely available at www.xzlab.org/software.html.

Introduction

DNA methylation — the covalent addition of methyl groups to cytosine bases — is a major epigenetic gene regulatory mechanism utilized by a wide variety of species. DNA methylation levels predict gene expression patterns, are involved in genomic imprinting and X-inactivation, and function to suppress the activity of transposable elements [1–3]. In addition, DNA methylation is essential for normal development [4–7]. For example, mutant *Arabidopsis* plants with reduced levels of DNA methylation display a range of abnormalities including reduced overall size, altered leaf size and shape, and reduced fertility [4–6]. In humans, DNA methylation levels are strongly linked to disease, including major public health burdens such as diabetes [8,9], Alzheimer's disease [10,11], and many forms of cancer [8,12–16]. These observations point to a central role for DNA methylation in shaping genome architecture, influencing development, and driving trait variation. Consequently, there is substantial interest in characterizing the genome-wide distribution of DNA methylation marks, and particularly, in identifying the genetic [17–20] and environmental [21–24] factors that explain variation in DNA methylation levels.

Recently, high-throughput sequencing based approaches have increased the feasibility, and consequently the popularity, of measuring DNA methylation levels. These methods, which include whole genome bisulfite sequencing (WGBS or BS-seq) [25], reduced representation bisulfite sequencing (RRBS) [26,27], and sequence capture followed by bisulfite conversion [28,29], produce base-pair resolution estimates of DNA methylation levels at genome-wide scales. All such methods rely on the differential sensitivity of methylated versus unmethylated cytosines to the chemical sodium bisulfite. Specifically, sodium bisulfite converts unmethylated cytosines to uracil (and ultimately thymine following PCR), while methylated cytosines are protected from conversion. Estimates of DNA methylation levels for each cytosine base can thus be obtained directly through high-throughput sequencing. Specifically, DNA methylation levels are estimated as the ratio of mapped cytosine reads (reflecting an originally methylated version of the base) to the total number of mapped reads at the same target (reflecting both methylated and unmethylated versions of the base).

The raw data produced by bisulfite sequencing methods are therefore count data, in which both the number of methylated reads and the total coverage at a site contain useful information. Higher total coverage corresponds to a more reliable estimate of the true DNA methylation level; however, in a typical experiment, total coverage can vary dramatically (e.g., by several orders of magnitude) across individuals and sites (Fig. S1). Many commonly used analysis methods, including all tools initially designed for array-based data [30,31], ignore this variability by converting counts to percentages or proportions (e.g., t-tests, Mann-Whitney U tests, or linear models, Table 1). Thus, a site at which 5 of 10 reads are designated as methylated (i.e., read as a cytosine) is treated identically to a site at which 50 of 100 reads are designated as methylated. This assumption reduces the power to uncover true predictors of variation in DNA methylation levels, because it ignores substantial sources of error in DNA methylation level estimates.

To address this problem, several recently introduced methods for differential DNA methylation analysis implement a beta-binomial model (e.g., ‘DSS: Dispersion

Shrinkage for Sequencing data' [32], 'RADMeth: Regression Analysis of Differential Methylation' [33], and 'MOABS: Model Based Analysis of Bisulfite Sequencing data' [34]). These methods model the binomial nature of bisulfite sequencing data, while taking into account the well-known problem of over-dispersion in sequencing reads. Because they work directly on count data, they can reliably account for variation in read coverage across sites and individuals. Consequently, beta-binomial methods consistently provide increased power to detect true associations between genetic or environmental sources of variance and DNA methylation levels [32–34].

However, beta-binomial-based methods only model over-dispersion due to independent variation, making them unsuited to studying DNA methylation variation in data sets affected by population structure or kinship. Taking these sources of structure into account is important because genetic variation is well known to exert strong and pervasive effects on DNA methylation levels [18,20,35,36]. In humans, methylation levels at more than ten thousand CpG sites are influenced by local genetic variation [19], and DNA methylation levels in whole blood are 18%-20% heritable on average, with the heritability estimates for the most heritable loci (top 10%) averaging around 68% [35,36]. As a result, DNA methylation levels will frequently covary with genetic relatedness (either kinship or population structure), and failure to account for this covariance could lead to spurious associations or reduced power to detect true effects. This phenomenon has been extensively documented for genotype-phenotype association studies [37–41], and controlling for genetic covariance between samples is now a basic requirement for these types of analyses. Similar logic applies to analyses of gene regulatory phenotypes, and studies of gene expression variation often do take genetic structure into account by using mixed model approaches [42–44]. However, despite growing interest in environmental epigenetics and epigenome-wide association studies (EWAS), no methods exist that appropriately control for genetic effects on DNA methylation levels in bisulfite sequencing data sets (Table 1).

To address this gap, we present a binomial mixed model (BMM) that accounts for both covariance between samples and extra over-dispersion caused by independent noise. We also present an efficient, sampling-based inference algorithm to accompany this model, called MACAU (Mixed model association for count data via data augmentation). MACAU works directly on binomially distributed count data and uses random effects to model relatedness/population structure and over-dispersion. Hence, MACAU enables parameter estimation and hypothesis testing in a wide variety of settings. To illustrate the advantages of our approach, we compared MACAU's performance with currently available methods using both simulated data and two real data sets (publicly available *Arabidopsis thaliana* WGBS data [45] and newly generated RRBS data from wild baboons, *Papio cynocephalus*). We found that MACAU appropriately controls for type I error and provides increased power compared to alternative methods, which either fail to account for the count nature of bisulfite sequencing data (e.g., linear mixed models [38,39,46,47]) or fail to account for genetic relatedness (e.g., beta-binomial models).

Results

The binomial mixed model and the MACAU algorithm

Here, we briefly describe the model and the algorithm. Additional details are provided in Text S1.

To detect differentially methylated sites, we model each potential target of DNA methylation individually (i.e., we model each CpG site one at a time). For each site, we consider the following binomial mixed model (BMM):

$$y_i = Bin(r_i, \pi_i),$$

where r_i is the total read count for *i*th individual; y_i is the methylated read count for that individual, constrained to be an integer value less than or equal to r_i; and π_i is an unknown parameter that represents the true proportion of methylated reads for the individual at the site. We use a logit link to model π_i as a linear function of several parameters:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{w}_i^T\boldsymbol{\alpha} + x_i\beta + g_i + e_i,$$

$$\boldsymbol{g} = (g_1, \cdots, g_n)^T \sim MVN(0,\ \sigma^2 h^2 \boldsymbol{K}),$$

$$\boldsymbol{e} = (e_1, \cdots, e_n)^T \sim MVN(0, \sigma^2(1-h^2)\boldsymbol{I}),$$

where $\boldsymbol{w}_i$ is a *c*-vector of covariates including an intercept and $\boldsymbol{\alpha}$ is a *c*-vector of corresponding coefficients; x_i is the predictor of interest and β is its coefficient; $\boldsymbol{g}$ is an *n*-vector of genetic random effects that model correlation due to population structure or kinship; $\boldsymbol{e}$ is an *n*-vector of environmental residual errors that model independent variation; $\boldsymbol{K}$ is a known *n* by *n* relatedness matrix that can be calculated based on pedigree or genotype data and that has been standardized to ensure $tr(\boldsymbol{K})/n = 1$ (this ensures that h^2 lies between 0 and 1, and can be interpreted as heritability, see [48]; *tr* denotes the trace norm); $\boldsymbol{I}$ is an *n* by *n* identity matrix; $\sigma^2 h^2$ is the genetic variance component; $\sigma^2(1-h^2)$ is the environmental variance component; h^2 is the heritability of the logit transformed methylation proportion (i.e. $logit(\boldsymbol{\pi})$); and MVN denotes the multivariate normal distribution.

Both $\boldsymbol{g}$ and $\boldsymbol{e}$ model over-dispersion (i.e., the increased variance in the data that is not explained by the binomial model). However, they model different aspects of over-dispersion: $\boldsymbol{e}$ models the variation that is due to independent environmental noise (a known problem in data sets based on sequencing reads: [49–52]), while $\boldsymbol{g}$ models the variation that is explained by kinship or population structure. Effectively, our model improves and generalizes the beta-binomial model by introducing this extra $\boldsymbol{g}$ term to model individual relatedness due to population structure or stratification.

We are interested in testing the null hypothesis that the predictor of interest has no effect on DNA methylation levels: $\boldsymbol{H_0}: \beta = 0$. This test requires obtaining the maximum likelihood estimate $\hat{\beta}$ from the model. Unlike its linear counterpart, estimating $\hat{\beta}$ from the binomial mixed model is notoriously difficult, as the joint likelihood consists of an *n*-dimensional integral that cannot be solved analytically [53]. Standard frequentist approaches rely on numerical integration [54] or Laplace approximation [55,56], but neither strategy scales well with the increasing dimension of the integral, which in our case is equal to the sample size. Because of this problem, frequentist approaches often produce biased estimates and overly narrow (i.e., anti-conservative) confidence intervals [57–61]. To overcome this problem, we instead use a Markov chain Monte Carlo (MCMC) algorithm-based approach for inference. After drawing accurate posterior samples, we rely on the asymptotic normality of both the likelihood and the posterior distributions [62] to further obtain the approximate maximum likelihood estimate $\hat{\beta}$ and its standard error se($\hat{\beta}$). This procedure allows us to construct approximate Wald test statistics and *p*-values for hypothesis testing. Despite the stochastic nature of the procedure, the MCMC errors are small enough to ensure stable p-value computation across multiple MCMC runs (Fig. S2).

For efficient, approximate *p*-value computation, we developed a novel MCMC algorithm based on an auxiliary variable representation of the binomial distribution [63–65] (Text S1). Our main contribution here is a framework that approximates the distribution of these latent variables (Fig. S3, Table S1-S2) and allows us to take advantage of recent innovations for fitting mixed effects models [38,46,47,66] (Text S1). These modifications substantially reduce the computational burden of fitting the BMM. Our algorithm reduces per-MCMC iteration computational complexity from cubic to quadratic with respect to the sample size. This results in an over 50-fold speed up compared with the popular software MCMCglmm [67] (Table S3) and makes our implementation of the BMM efficient for data sets ranging up to hundreds of individuals and millions of sites.

Because our model effectively includes the beta-binomial model as a special case, we expect it to perform similarly to the beta-binomial model in settings in which population structure is absent (we say “effectively” because strictly speaking, the beta-binomial model uses a beta distribution to model independent noise while we use a normal distribution). However, we expect our model to outperform the beta binomial in settings in which population structure is present. In addition, in the presence of population stratification, we expect the beta-binomial model to produce inflated test statistics (thus increasing the false positive rate) while our model should provide calibrated ones. Below, we test these predictions using both simulations and real data applications.

Count-based models perform well in the absence of genetic effects on DNA methylation levels

We first compared the performance of the BMM implemented in MACAU with the performance of other currently available methods for analyzing bisulfite sequencing data in the absence of genetic effects. Intuitively, since the BMM models count data and effectively includes the beta-binomial model as a special case, we expected it to perform similarly to the beta-binomial model; further, we expected both models to outperform methods that do not model counts. To test our prediction, we simulated the effect of a predictor variable on DNA methylation levels across 5000 CpG sites (4500 true negatives and 500 true positives). To approximate the distribution of a predictor variable in a real population, and because we analyze age-associated variation in DNA methylation levels in a baboon RRBS data set in detail below, we conducted our simulation using known age values sampled from the same baboon population. For all simulations, we set the effect of genetic variation on DNA methylation levels equal to zero, which is equivalent to setting either (i) the heritability of DNA methylation levels to zero (unlikely based on prior findings [35,36]), or (ii) studying completely unrelated individuals in the absence of population structure. To explore MACAU's performance across a range of conditions, we simulated age effects on DNA methylation levels across three different effect sizes (percent of variance in DNA methylation explained (PVE) = 5%, 10%, or 15%) and three different sample sizes (n = 20, 50, and 80).

Because age is naturally modeled as a continuous variable, we focused our comparisons only on approaches that could accommodate continuous predictor variables (comparisons in which we artificially binarized age, which allowed us to include a larger set of approaches, produced qualitatively similar results: Fig. S4). Specifically, in addition to the BMM implemented in MACAU, we considered the performance of a beta-binomial model, a linear model, a binomial model, and a linear mixed model (implemented in the software GEMMA [46]). As expected, we found that MACAU performed similarly to the beta-binomial model, and that these two approaches consistently detected more true positive age effects on DNA methylation levels (at a 10% empirical FDR) than all other methods (Figs. S5-S6). For example, in the “easiest” case we simulated (PVE = 15%, n = 80), we found that the beta-binomial model detected 30% of simulated true positives, while the BMM implemented in MACAU detected 27.8%. The slight loss of power in the BMM is a consequence of the smaller degrees of freedom caused by the additional genetic variance component. In comparison, the linear model detected 21.2% of true positives; the linear mixed effects model, 14%; and the binomial model, 8.4% (Fig. S5). The binomial model exhibits low power when FDR is used to control for multiple hypothesis testing due to poor type I error calibration, as has been previously reported [33]. Area under a receiver operating characteristic curve (AUC) was also consistently very similar between the beta-binomial and MACAU (Fig. S6), although the advantage of the count-based methods was less clear by this measure. This reduced contrast is because AUC is based on true positive-false positive trade-offs across a range of p-value thresholds; methods can consequently yield high AUCs even when they harbor little power to detect true positives at FDR thresholds that are frequently used in practice. Taken together, our simulations suggest a general advantage to count-based models for samples that contain no genetic structure. Further, the differences in performance between the beta-binomial model and the BMM implemented in MACAU were consistently small in this setting (Figs. S5-S6).

Binomial mixed models control for false positive associations that arise from population structure

Next, we investigated the performance of each method in the presence of population structure. When DNA methylation levels are heritable and the predictor variable of interest is confounded with population structure, false positive associations should arise if genetic covariance between samples is not modeled. Because the BMM accounts for population structure while the beta-binomial model does not, we therefore expected MACAU to produce well-calibrated test statistics and the beta-binomial model to produce inflated test statistics. To test this prediction, we drew on publicly available phenotype data and SNP genotype data for 24 *Arabidopsis thaliana* accessions [68,69] in which leaf tissue samples were recently subjected to whole genome bisulfite sequencing [45]. Among these accessions, a secondary dormancy phenotype (measured as the slope between the germination percentages of non-dormant seeds after one and six weeks of cold treatment) is correlated with population structure ($R^2 = 0.38$ against the first principal component of the genotype matrix for these accessions; $p = 7.84 \times 10^{-4}$; Fig. S7). Because secondary dormancy is associated with environmental conditions that are experienced after the seed has already dispersed, we have no expectation that secondary dormancy should be associated with DNA methylation levels in leaf tissue. Consequently, we used the true distribution of secondary dormancy characteristics and the true genetic structure among these 24 accessions to simulate a dataset that consisted entirely of true negatives. Specifically, we simulated data sets (containing 4000 sites each) in which the association between secondary dormancy and DNA methylation levels in leaf tissue was always equal to 0, but the effect of genetic variation on DNA methylation levels was either moderate ($h^2 = 0.3$) or large ($h^2 = 0.6$). Thus, in these data sets, population structure could confound the relationship between the predictor variable (the capacity for secondary dormancy) and DNA methylation levels if not taken into account.

As predicted, we found that the BMM implemented in MACAU appropriately controlled for genetic effects on DNA methylation levels: whether DNA methylation levels were moderately ($h^2 = 0.3$) or strongly ($h^2 = 0.6$) heritable, MACAU did not detect any sites associated with secondary dormancy at a relatively liberal false discovery rate threshold of 20% (whether calculated against empirical permutations or calculated using the R package *qvalue* [32]). In addition, the *p*-value distributions for secondary dormancy effects on DNA methylation levels, in both simulations, did not differ from the expected uniform distribution (Fig. 1; Kolmogorov-Smirnov (KS) test when $h^2 = 0.3$: D = 0.015, p = 0.909; when $h^2 = 0.6$: D = 0.016, p = 0.874; genomic control factors: 0.90 when $h^2 = 0.3$, 0.93 when $h^2 = 0.6$). In contrast, when we analyzed the same simulated data sets with a beta-binomial model, we erroneously detected 2 CpG sites associated with secondary dormancy when heritability was set to 0.3, and 4 CpG sites when heritability was set to 0.6 (at a 20% FDR in both cases). More concerningly, the distributions of *p*-values produced by the beta-binomial model were significantly different from the expected uniform distribution and skewed towards low (significant) values (KS test when $h^2 = 0.3$: D = 0.084, p = 1.75 x 10^{-8}; when $h^2 = 0.6$: D = 0.096, p = 2.80 x 10^{-11}; genomic control factors: 1.18 when $h^2 = 0.3$, 1.32 when $h^2 = 0.6$). These numbers suggest an increasing problem with false positives as the heritability of DNA methylation levels increases.

To investigate the calibration of test statistics in a real data set, we next analyzed the relationship between the secondary dormancy phenotype and publicly available WGBS data for the same 24 *Arabidopsis* accessions (n = 830,676 CpG sites tested [32,33,34]). We again compared the performance of a simple linear model, a binomial model, a beta-binomial model, the BMM implemented in MACAU, and an LMM implemented in GEMMA. Again illustrating its poor handling of Type I error, the binomial model detected more than 100,000 secondary dormancy-associated sites at a 10% empirical FDR threshold, respectively, with a genomic control factor of 3.81. A beta-binomial model substantially improved over the binomial model, but still detected 39 secondary dormancy-associated sites at a 20% empirical FDR threshold, and 150 sites and 690 sites at a 10% or 20% FDR *qvalue* threshold, respectively (genomic control factor = 1.16). Given the clear confounding of population structure and secondary dormancy in this sample, as well as the results of our simulations, these associations are probably spurious. In contrast, MACAU, the linear mixed model (GEMMA), and the simple linear model did not identify any CpG sites associated with secondary dormancy, either at a 10% or a 20% false discovery rate threshold (Fig. 1; genomic control factors: MACAU – 0.89, GEMMA – 0.97, Linear model – 0.99). Based on our earlier simulations, the similarity of performance among the three models likely stems from different reasons: both the linear model and the linear mixed model are more lowly powered to detect positive hits (either true positives or false positives), whereas MACAU combines both the increased power conferred by modeling the raw count data with appropriate controls for population structure (see Fig. 1 and results below).

MACAU provides increased power to detect true positives in the presence of kinship

We next investigated the power of different approaches to detect truly differentially methylated sites in the presence of relatedness. Because it appropriately models genetic similarity between relatives, we expected the BMM implemented in MACAU to exhibit improved power over the other methods. To test this prediction, we returned to the baboon data set that was the focus of our initial simulations. Instead of assuming no genetic contribution to variation in DNA methylation levels, here we instead simulated moderate to large genetic effects (h^2 = 0.3 and 0.6 respectively, as in the *Arabidopsis* simulation above). We simulated relatedness values based on the distribution of relatedness values within a single mixed-sex baboon social group. Female baboons remain in their natal groups throughout their lives, producing relatedness values that are primarily due to matrilineal descent. The resulting genetic structure is one in which females tend to be more closely related to each other, on average, than males or male-female dyads [70], but in which not all females are related (because multiple matrilines co-reside in a single group). Thus, baboon social groups contain a large set of unrelated dyads, some pairs of close relatives, and some distant relatives (Fig. S8). We simulated an effect of age on DNA methylation levels in a data set consisting of 80 baboons with known ages and dyadic relatedness levels. We simulated a range of non-zero effect sizes (percent variance explained by age = 5%, 10%, or 15%) for 5000 CpG sites, containing 500 true positives and 4500 true negatives. We chose these parameters to mimic the distribution of effect sizes observed in real data sets, which can range from small to substantial but which are generally limited to a minority of sites [9,17,36,71].

In simulations in which age had a moderate effect on DNA methylation levels (PVE = 10%), MACAU detected 11.4% (when $h^2 = 0.3$) and 20.6% (when $h^2 = 0.6$) of simulated true positives at a 10% empirical FDR. In comparison, the beta-binomial model (the next best model) detected 8.2% and 10.4% of true positives, respectively (Fig. 2). As in the simulations, we again observed that a simple binomial model was prone to type I error, which resulted in failure to detect true age-associated sites when empirical FDRs were calculated against permuted data. Our additional simulations at PVE = 5% or PVE = 15% confirmed MACAU's advantage over other methods across a range of effect sizes (Fig. S9). As expected, the magnitude of this advantage was positively correlated with the heritability of DNA methylation levels.

Age-associated DNA methylation levels in wild baboons

Finally, we applied MACAU to a real RRBS data set that we generated from 50 wild baboons, drawn from the same population used to parameterize the simulations above. This data set included 433,871 CpG sites, enriched (as expected in RRBS data sets [26,27]) for putatively functional regions of the genome (e.g., genes, gene promoters, CpG islands: Fig S11). We used these data to investigate the epigenetic signature of age at sampling (range = 1.76 – 18.01 years in our sample, Table S4); we focused on age because it is a known predictor of DNA methylation levels in humans and other animals [35,72,73] and because DNA methylation changes with age are well characterized [35,36,74–76]. Consequently, we were able to not only assess MACAU's power to detect statistically age-associated sites, but also test its ability to identify known age-related signatures in DNA methylation data.

As in our simulations, we found that MACAU provided increased power to detect age effects in the presence of familial relatedness. We detected 1.6-fold more age-associated CpG sites at a 10% empirical FDR using MACAU compared to the results of a beta-binomial model, the next best approach (1.4-fold more sites at a 20% empirical FDR; Fig. 3 and Fig. S10). This advantage was consistently observed across all FDR thresholds we considered, except for relatively low (<7.5%) empirical FDR thresholds, when all of the methods were very low powered as a result of the modest sample size.

We performed several analyses to investigate the likely validity and functional importance of the age-associated CpG sites we identified. Based on the results of previous studies, we expected that age-associated sites in CpG islands would tend to gain methylation with age [75,76], while sites in other regions of the genome (e.g., CpG island shores, gene bodies) would tend to lose methylation with age [75,76]. In addition, we expected that, in whole blood, bivalent/poised promoters should gain DNA methylation with age, while enhancers should lose methylation with age (as discussed in [74,75,77]). Our results conformed to these patterns: sites in CpG islands tended to gain methylation with age (71.4% of sites were positively correlated with age); and sites in promoters, CpG island shores, and gene bodies tended to lose methylation with age (72.7%, 75.4%, and 75.2% of sites were negatively correlated with age, respectively; Fig. 3). In addition, we found that positively correlated, age-associated sites were highly enriched in chromatin states associated with bivalent/poised promoters (as defined by the Roadmap Epigenomics Project [78]). Specifically, age-associated CpG sites in bivalent/poised promoters were 3.4 times more likely to show increases in DNA methylation with age, compared to age-associated CpG sites in other regions ($p < 10^{-10}$, Fisher's exact test). Furthermore, negatively correlated age-associated sites (i.e., sites where DNA methylation levels decreased with age) were strongly enriched in enhancers (defined as sites either marked by H3K4me1 in human PBMCs [79] or sites within chromatin states annotated as 'enhancers' by the Roadmap Epigenomics Project [78], $p = 2 \times 10^{-4}$, Fisher's exact test).

Finally, we reasoned that true positive age-associated CpG sites should also contain information about age-associated gene expression levels. To test this hypothesis, we turned to previously generated whole blood RNA-seq data [42] from the same baboon population (n = 63; only four baboons in the RNA-seq data set were also included in the DNA methylation data set). Overall, we observed a strong enrichment of differentially methylated CpG sites in or near (within 10 kb) blood-expressed genes (n = 12,018 genes), compared to the background set of all CpG sites near genes (Fisher's exact test, $p < 10^{-10}$). Further, CpG sites near age-associated genes (n = 1396 genes, 10% FDR) were 30.5% more likely to be differentially methylated with age compared to the background set of all CpG sites near genes (Fisher's exact test, p = 0.032).

Discussion

DNA methylation levels can have potent effects on downstream gene regulation, and, in doing so, can shape key behavioral, physiological, and disease-related phenotypes [8,21,80–82]. These observations have motivated an increasing number of DNA methylation studies in humans and other organisms, highlighting the need for sophisticated statistical methods that can accommodate the complexities of a broad array of data sets. Here, we demonstrate that the binomial mixed model implemented in our software MACAU can (i) effectively control for confounding relationships between genetic background and a predictor variable of interest and (ii) provide increased power to detect true sources of variance in DNA methylation levels in data sets that contain kinship or population structure. In addition, MACAU provides increased flexibility over current count-based methods that cannot accommodate biological replicates (e.g., Fisher's exact test), continuous predictor variables (e.g., DSS, MOABS, RadMeth), or biological or technical covariates (e.g., MOABS, DSS; see also Table 1). Given the increasing interest in both the environmental [22,71,83] and genetic [17,18,20,84] architecture of DNA methylation levels, we believe MACAU will be a useful tool for generalizing epigenomic studies to a larger range of populations. MACAU is particularly well suited to data sets that contain related individuals or population structure; notably, several major population genomic resources contain structure of these kinds (e.g., the HapMap population samples [85], the Human Genome Diversity Panel [86], and the 1000 Genomes Project in humans [87]; the Hybrid Mouse Diversity Panel [88]; and the 1001 Genomes Project in *Arabidopsis* [89]).

Indeed, our results suggest MACAU is a useful tool even in data sets that are less affected by genetic structure, or when the heritability of DNA methylation levels is unclear. Because the beta-binomial model is incorporated as a special case, MACAU exhibits only a slight loss of power relative to a beta-binomial model without random effects when $h^2 = 0$, while conferring better power and better test statistic calibration when $h^2 > 0$ (Fig. S5-S6; Fig. 1). Previous studies in humans have shown that, while the heritability of DNA methylation levels varies across loci, an appreciable proportion of loci are either modestly (h^2 >= 0.3: 21.06% of all CpG sites) or highly (h^2 >= 0.6: 6.95% of all CpG sites) heritable [36,90]; further, DNA methylation QTLs are widespread across the genome [19,35,84]. Thus, because investigators will rarely have *a priori* knowledge of the heritability of DNA methylation levels at a given locus, and because the advantage of a beta-binomial model is small even when heritability is zero, we recommend applying MACAU in cases where genetic effects on DNA methylation levels are poorly understood. In addition, our model provides a natural framework for incorporating the spatial dependency of DNA methylation levels across neighboring sites [91,92], which we expect to increase power even further [91,92]. However, we do note that, even with the efficient algorithm implemented here, fitting the binomial mixed model (or its extensions) remains more computationally expensive than other approaches for moderately sized datasets (Table S3). While it remains appropriate for the sample sizes used in current studies (e.g., dozens to hundreds of individuals), rapid increases in sample size—especially in the context of EWAS—strongly motivate additional algorithm development to scale up the binomial mixed model for data sets that include thousands or tens of thousands of individuals.

Although we developed MACAU with the analysis of bisulfite sequencing data in mind, we note that a count-based binomial mixed model may be an appropriate tool in other settings as well. For example, allele-specific gene expression (ASE) is often measured in RNA-seq data by comparing the number of reads originating from a given variant to the total number of mapped reads for that site [66,93–95]. The structure of these data are highly similar to the structure of bisulfite sequencing data, which focus on counts of methylated versus total reads. Unsurprisingly, beta-binomial models have also emerged as one of the most popular methods for estimating ASE values [95–97]. Researchers interested in the predictors of variation in ASE levels—which could include *trans*-acting genetic effects, environmental conditions, or properties of the individual (e.g., sex or disease status)—might also benefit from using MACAU. Recent work from the TwinsUK study motivates the need for such a model: Grundberg et al. demonstrated a strong heritable component to ASE levels [98], which could be effectively taken into account using the random effects approach implemented here.

Finally, linear mixed models have also been recently proposed to account for cell type heterogeneity in epigenome-wide association studies focused on array data [99]. In this framework, the random effect covariance structure is based on overall covariance in DNA methylation levels between samples, which is assumed to be largely attributable to variation in tissue composition. MACAU provides a potential avenue for extending these ideas to sequencing-based data sets.

Materials and Methods

Arabidopsis thaliana whole genome bisulfite sequencing (WGBS) data set

We downloaded publicly available WGBS data generated by Schmitz et al. [45], as well as previously published SNP genotype data [69] and secondary dormancy data [68] for 24 *Arabidopsis* accessions. We used the SNP genotype data (specifically, 188,093 sites with minor allele frequency >5%) to construct a pairwise genetic relatedness matrix, K, as the product of a standardized genotype matrix [48] (implemented with a built-in function in MACAU). We used this estimate of K for both the simulations and our analyses of the real WGBS data.

In these analyses, we focused on CpG sites measured in ≥50% of accessions, and excluded sites that were constitutively hypermethylated (average DNA methylation level >0.90) or hypomethylated (average DNA methylation level <0.10, following [71,99]). We also excluded highly invariable sites (i.e., sites where the standard deviation of DNA methylation levels fell in the lowest 5% of the overall data set) and sites with very low coverage (i.e., sites where the mean coverage fell in the lowest quartile for the overall data set, below a mean of 3.34 reads). After filtering, the final data set consisted of 830,676 sites.

Baboon reduced representation bisulfite sequencing (RRBS) data set

Study subjects and sample collection. To investigate age effects on DNA methylation levels, in both real and simulated data sets, we drew on data and samples from a wild population of yellow baboons in the Amboseli ecosystem of southern Kenya. This population has been monitored for over four decades by the Amboseli Baboon Research Project (ABRP) [100], and the ages of animals born in the study population (n = 37, 74% of the data set) are therefore known to within a few days' error. For animals that immigrate into the study population, ages are estimated from morphological features by trained observers (n = 13, 26% of the data set) [101]. Pairwise relatedness values were available from previously collected microsatellite data (14 highly variable loci) [102,103] analyzed with the program COANCESTRY [104]. Using the age and relatedness data sets, we simulated age effects on DNA methylation levels for either n = 50 or n = 80 baboons from a single social group. In addition, we used previously collected blood samples from the

Amboseli population, paired with age and microsatellite genotype records, to investigate age effects on DNA methylation levels in a newly generated RRBS data set.

To generate the new RRBS data, we used whole blood samples collected from 50 animals (35 males and 15 females) by the ABRP between 1989 and 2011 following well-established procedures [42,105,106]. Briefly, animals were immobilized by an anesthetic-bearing dart delivered through a hand-held blow gun. They were then quickly transferred to a processing site for blood sample collection. Following sample collection, study subjects were allowed to regain consciousness in a covered holding cage until they were fully recovered from the effects of the anesthetic. Upon recovery, study subjects were released near their social group and closely monitored. Blood samples were stored at the field site or at an ABRP-affiliated lab at the University of Nairobi until they were transported to the United States.

Importantly, given the large range in sample collection dates, we observed no correlation between the age of our study subjects at sample collection and sample age (i.e., time since the collection date; Spearman rank correlation, p = 0.779). Further, to ensure that variation in sample collection dates did not influence our results, we also controlled for sample age as a covariate in our final analyses of the RRBS dataset (see *Analysis of age-related changes in DNA methylation levels*).

RRBS data generation and low-level processing. Genomic DNA was extracted from whole blood samples using the DNeasy Blood and Tissue Kit

(QIAGEN) according to the manufacturer's instructions. RRBS libraries were created from 180 ng of genomic DNA per individual, following the protocol by Boyle et al. [26]. In addition, 1 ng of unmethylated lambda phage DNA (Sigma Aldrich) was incorporated into each library to assess the efficiency of the bisulfite conversion. All RRBS libraries were sequenced using 100 bp single end sequencing on an Illumina HiSeq 2000 platform, yielding a mean of 28.97 ±8.97 million reads per analyzed sample (range: 9.59 – 79.78 million reads; Table S4).

We removed adaptor contamination and low-quality bases from all reads using the program TRIMMOMATIC [107]. We then mapped the trimmed reads to the olive baboon genome (*Panu* 2.0) using BSMAP, a tool designed for high-throughput DNA methylation data [108]. We used a Python script packaged with BSMAP to extract the number of reads as cytosine (reflecting an originally methylated base) and the total read count for each individual and CpG site. We performed the same set of filtering steps described for the *Arabidopsis* WGBS data set to produce our final data set for the baboons. Specifically, we excluded sites that were constitutively hypermethylated or hypomethylated, sites that were highly invariable, and sites that had low average coverage across individuals (in this case, the lowest quartile for mean coverage levels was 4.74 reads). The final filtered data set consisted of 433,871 CpG sites.

Simulations

To simulate the methylated read counts and total read counts that result from WGBS and RRBS, we performed the following procedure:

First, we simulated the proportion of methylated reads for each site. To do so, we drew secondary dormancy values or age values, x, as the predictor of interest, from the actual values for the *Arabidopsis* accessions or from the baboon population, respectively. For each CpG site, we simulated the DNA methylation level, π, as a linear function of x and its effect size (β), as well as the effects of genetic variation (g) and random environmental variation (e), passed through a logit link (based on the model described in the Results section).

For the baboon RRBS simulations, we determined *K* from 14 highly variable microsatellite loci [102,103], focusing on the true values for either n = 50 or n = 80 baboons drawn from a single social group in the Amboseli population (i.e., the same population we sampled in the real RRBS dataset). For the *Arabidopsis* WGBS simulations, *K* was determined from publicly available SNP genotype data [69]. For each simulation, we set h^2 to 0, 0.3, or 0.6 to simulate no, modest, or highly heritable DNA methylation levels. We also estimated the variance term σ^2 from the real data sets. Specifically, we took the mean estimate of σ^2 across all sites (as calculated in MACAU) for each real data set, and used this value as the fixed value of σ^2 in the corresponding simulations.

Next, for each site, we simulated total read counts r_i for each individual *i* from a negative binomial distribution that models the extra variation observed in the real data:

$$r_i \sim NB(t, p),$$

where *t* and *p* are site specific parameters estimated from the real data. Specifically, we generated 10,000 sets of *t* and *p* parameters by fitting a negative binomial distribution to the total read count data from 10,000 randomly selected CpG sites in the real baboon RRBS data set or the real *Arabidopsis* data set, using the function 'fitdistr' in the R package *MASS* [109]. To simulate counts for a given CpG site, we randomly selected one of these parameter sets to produce the total number of reads. Finally, we simulated the number of methylated reads for each individual at that locus (y) by drawing from a binomial distribution parameterized by the number of total reads (r) and the DNA methylation level (π).

Comparison of MACAU to existing methods

For all simulated and real data sets, we used raw methylated and total read counts to compare the results of a beta-binomial model (using a custom R script), a binomial model (implemented via 'glm' in R), and the binomial mixed model implemented in MACAU. For computation time comparison, we also used the MCMCglmm software that implements the binomial mixed model [67]. In addition, we used the same count data to run a Fisher's exact test (implemented in R), DSS [32], and RadMeth [33] in the subset of analyses that utilized these programs. Finally, to analyze simulated and real data sets using a linear model (implemented using '*lm'* in R) or the linear mixed model implemented in GEMMA [46], we estimated DNA methylation levels by dividing the number of methylated reads by the total read count for each individual and CpG site. We then quantile normalized the resulting proportions for each CpG site to a standard normal distribution, and imputed any missing data using the K-nearest neighbors algorithm in the R package *impute* [110].

To compute empirical false discovery rates in simulated data, we divided the number of false positives detected at a given *p*-value threshold by the total number of sites called by the model as significant at that threshold (i.e., the sum of false positives and true positives). To compute empirical false discovery rates in the real data, in which the false positives and true positives were unknown, we used permutations. Specifically, we permuted the predictor variable for each data set four times, reran our analyses, and then calculated the false discovery rate as the average number of sites detected at a given *p*-value threshold in the permuted data divided by the total number of sites detected at that threshold in the real data. For simulated data sets only, we also calculated the area under the receiver operating characteristic curve (AUC) to produce a measure of the overall tradeoff between detecting true positives and calling false positives.

Analysis of age-related changes in DNA methylation levels

Our initial analyses of the baboon RRBS dataset focused only on the relative ability of each method to detect age-associated sites. For these analyses, we therefore did not control for other biological covariates that may contribute to variance in DNA methylation levels (note that biological covariates cannot be incorporated into several implementations of the beta-binomial model [32,34]: see Table 1). However, to investigate patterns of age-related changes in DNA methylation levels, and to compare them to previously described patterns in the literature, we wished to control for such covariates. To do so, we reran the differential methylation analysis in MACAU, this time controlling for sex, sample age, and efficiency of the bisulfite conversion rate estimated from the lambda phage spike-in.

First, we investigated whether age-associated sites were enriched in functionally coherent regions of the genome, many of which have previously been identified as age-related [35,75,76]. To do so, we defined gene bodies as the regions between the 5'-most transcription start site (TSS) and 3'-most transcription end site (TES) of each gene using *Panu* 2.0 annotations from Ensembl [111]. We defined promoter regions as the 2 kb upstream of the TSS. CpG were annotated based on the UCSC Genome Browser track for baboon [112], with CpG island shores defined as the 2 kb regions flanking either side of the CpG island boundary (following [27,113,114]). Finally, because no enhancer annotations are available that are specific to baboons, we used H3K4me1 ChIP-seq data generated by ENCODE (from human peripheral blood mononuclear cells) to define enhancer regions [79]. In addition, we used chromatin state annotations from the Roadmap Epigenomics Project (also generated from human peripheral blood mononuclear cells) to further investigate biases in the locations of age-associated sites [78]. Using these annotation sets, we performed Fisher's Exact Tests to ask whether age-associated sites were enriched or underrepresented in specific genomic regions.

Second, we asked whether differentially methylated sites were more likely to fall close to blood-expressed genes. For this comparison, we drew on previously published RNA-seq data, generated from whole blood samples collected in the Amboseli baboon population [42]. We defined blood-expressed genes as those genes that had non-zero counts in more than 10% of individuals in the RNA-seq data sets, and that had mean read counts greater than or equal to 10. We then compared the number of differentially methylated CpG sites near blood-expressed genes (i.e., within the gene body or within 10 kb of the gene TSS or TES) to the number of differentially methylated CpG sites near genes that were not expressed in blood, using a Fisher's Exact Test.

Finally, we investigated whether CpG sites that occur near genes that are differentially expressed with age were also more likely to be differentially methylated with age. For this comparison, we defined 'age-associated genes' as genes differentially expressed with age (at a 10% FDR) in the RNA-seq data set [42]. We compared the number of differentially methylated CpG sites near blood-expressed, age-associated genes to the number of differentially methylated CpG sites near genes that were not within this set of genes, again using a Fisher's Exact Test.

Software and data availability

The MACAU software and a custom script for implementing a beta-binomial model in R is available at: www.xzlab.org/software.html. Previously published data sets are available at http://bergelson.uchicago.edu/regmap-data/regmap.html/ (*Arabidopsis* SNP genotype data), http://www.ncbi.nlm.nih.gov/geo/ (*Arabidopsis* WGBS data: GSE43857, Baboon RNA-seq data: GSE63788); and http://www.nature.com/nature/journal/v465/n7298/full/nature08800.html#supplementary-information (*Arabidopsis* phenotype data). The baboon RRBS data set will be made publicly available at the NCBI Gene Expression Omnibus upon manuscript acceptance.

Acknowledgments

We thank the Kenya Wildlife Services, Institute of Primate Research, National Museums of Kenya, National Council for Science and Technology, members of the Amboseli-Longido pastoralist communities, Tortilis Camp, and Ker & Downey Safaris for their assistance in Kenya. We also thank Jeanne Altmann for general support and access to the Amboseli data set and samples, Raphael Mututua, Serah Sayialel, Kinyua Warutere, Mercy Akinyi, Tim Wango, and Vivian Oudu for invaluable assistance with sample collection; Matthew Stephens and Sayan Mukherjee for insight and support on previous versions of MACAU; and Dan Runcie for useful suggestions on data applications. Finally, we thank the Baylor College of Medicine Human Genome Sequencing Center for access to the current version of the baboon genome assembly (*Panu 2.0*).

Funding

This work was supported by a seed grant from the Duke Population Research Institute to JT (a component of 5R24-HD065563-03 to S. Sanders); NIH grants R21-AG049936 and R03-AG045459-01 to JT and SCA; NSF grant BCS-1455808 to JT and AJL; and a start-up fund from the University of Michigan to XZ.

References

1. Mohandas T, Sparkes R, Shapiro L. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. Science. 1981;211: 393–396.

2. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. Nature. 1993;366: 362–365. doi:10.1038/366362a0

3. Jones P. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13: 484–92. doi:10.1038/nrg3230

4. Kakutani T, Jeddeloh J, Richards EJ. Characterization of an Arabidopsis thaliana DNA hypomethylation mutant. Nucleic Acids Res. 1995;23: 130–137.

5. Ronemus MJ, Galbiati M, Ticknor C, Chen J, Dellaporta SL. Demethylation-induced developmental pleiotropy in Arabidopsis. Science. 1996;273: 654–657. doi:10.1126/science.273.5275.654

6. Finnegan EJ, Peacock WJ, Dennis ES. Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development. Proc Natl Acad Sci. 1996;93: 8449–8454. doi:10.1073/pnas.93.16.8449

7. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;69: 915–926. doi:10.1016/0092-8674(92)90611-F

8. Rakyan VK, Beyan H, Down T, Hawa MI, Maslau S, Aden D, et al. Identification of type 1 Diabetes-associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet. 2011;7: 1–9. doi:10.1371/journal.pgen.1002300

9. Dayeh T, Volkov P, Salö S, Hall E, Nilsson E, Olsson AH, et al. Genome-wide Dna methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. PLoS Genet. 2014;10. doi:10.1371/journal.pgen.1004160

10. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci. 2014;17: 1156–1163. doi:10.1038/nn.3786

11. Bakulskia K, Dolinoya D, Sartorb M, Paulsond H, Konend J, Liebermane A, et al. Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex. J Alzheimers Dis. 2012;29: 1–28. doi:10.3233/JAD-2012-111223.Genome-Wide

12. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31: 142–147. doi:10.1038/nbt.2487

13. Irizarry R, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41: 178–86. doi:10.1038/ng.298

14. Gluckman PD, Hanson M, Buklijas T, Low FM, Beedle AS. Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. Nat Rev Endocrinol. 2009;5: 401–8. doi:10.1038/nrendo.2009.102

15. Suarez-Alvarez B, Rodriguez RM, Fraga MF, López-Larrea C. DNA methylation: a promising landscape for immune system-related diseases. Trends Genet. 2012;28: 506–14. doi:10.1016/j.tig.2012.06.005

16. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14: R21. doi:10.1186/gb-2013-14-3-r21

17. Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. Genome Res. 2014; doi:10.1101/gr.176933.114

18. Bell JT, Pai A, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12: R10. doi:10.1186/gb-2011-12-1-r10

19. Banovich NE, Lan X, Mcvicker G, Degner JF, Blischak JD, Roux J, et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. 2014;10: 1–12. doi:10.1371/journal.pgen.1004663

20. Dubin MJ, Zhang P, Meng D, Remigereau M, Osborne EJ, Casale FP, et al. DNA methylation variation in Arabidopsis has a genetic basis and appears to be involved in local adaptation. eLife. 2015;4: e05255. doi:10.7554/eLife.05255

21. Weaver ICG, Cervoni N, Champagne F a, D'Alessio AC, Sharma S, Seckl JR, et al. Epigenetic programming by maternal behavior. Nat Neurosci. 2004;7: 847–54. doi:10.1038/nn1276

22. Waterland R a, Kellermayer R, Laritsky E, Rayco-Solon P, Harris RA, Travisano M, et al. Season of conception in rural gambia affects DNA

methylation at putative human metastable epialleles. PLoS Genet. 2010;6: e1001252. doi:10.1371/journal.pgen.1001252

23. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci. 2008;105: 17046–9. doi:10.1073/pnas.0806560105

24. Wolff GL, Kodell RL, Moore SR, Cooney C. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. Am Soc Exp Biol. 1998;12: 949–57.

25. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008;452: 215–219. doi:10.1038/nature06745

26. Boyle P, Clement K, Gu H, Smith Z. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol. 2012;13: R92. doi:10.1186/gb-2012-13-10-R92

27. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011;6: 468–81. doi:10.1038/nprot.2010.190

28. Ivanov M, Kals M, Kacevska M, Metspalu A, Ingelman-Sundberg M, Milani L. In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. Nucleic Acids Res. 2013;41. doi:10.1093/nar/gks1467

29. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. na. 2009;27: 353–60. doi:10.1038/nbt.1530

30. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30: 1363–1369. doi:10.1093/bioinformatics/btu049

31. Du P, Zhang X, Huang C-C, Jafari N, Kibbe W, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11: 587. doi:10.1186/1471-2105-11-587

32. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014;42: 1–11. doi:10.1093/nar/gku154

33. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics. 2014;15: 215. doi:10.1186/1471-2105-15-215

34. Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, et al. MOABS: model based analysis of bisulfite sequencing data. Genome Biol. 2014;15: R38. doi:10.1186/gb-2014-15-2-r38

35. Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 2012;8. doi:10.1371/journal.pgen.1002629

36. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 2014;15: R73. doi:10.1186/gb-2014-15-5-r73

37. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456: 98–101. doi:10.1038/nature07566

38. Kang HM, Zaitlen N, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178: 1709–23. doi:10.1534/genetics.107.080101

39. Kang H, Sul J, Zaitlen N, Kong S, Freimer NB, Sabatti C, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42: 348–354. doi:10.1038/ng.548.Variance

40. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847

41. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38: 203–8. doi:10.1038/ng1702

42. Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y. The genetic architecture of gene expression levels in wild baboons. eLife. 2015;4: 1–22. doi:10.7554/eLife.04729

43. Turner L, Harr B. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. eLife. 2014;3: e02504. doi:10.7554/eLife.02504

44. Tung J, Barreiro LB, Johnson ZP, Hansen KD, Michopoulos V, Toufexis D, et al. Social environment is associated with gene regulatory variation in the rhesus macaque immune system. Proc Natl Acad Sci. 2012;109: 6490–5. doi:10.1073/pnas.1202734109

45. Schmitz RJ, Schultz MD, Urich M, Nery JR, Pelizzola M, Libiger O, et al. Patterns of population epigenomic diversity. Nature. 2013; doi:10.1038/nature11968

46. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44: 821–4. doi:10.1038/ng.2310

47. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8. doi:10.1038/nmeth.1681

48. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003264

49. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–40. doi:10.1093/bioinformatics/btp616

50. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. BioMed Central Ltd; 2010;11: R106. doi:10.1186/gb-2010-11-10-r106

51. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14: R95. doi:10.1186/gb-2013-14-9-r95

52. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007;23: 2881–2887. doi:10.1093/bioinformatics/btm453

53. McCulloch C. Joint modelling of mixed outcome types using latent variables. Stat Methods Med Res. 2008;17: 53–73. doi:10.1177/0962280207081240

54. Pinheiro JC, Chao EC. Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. J Comput Graph Stat. 2006;15: 58–81. doi:10.1198/106186006X96962

55. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88: 9–25.

56. Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika. 1991;78: 45–51.

57. Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. J R Stat Soc Ser A. 1996;159: 505–513.

58. Rodriguez G, Goldman N. Improved estimation procedures for multilevel models with binary response: {A} case-study. J R Stat Soc Ser A. 2001;164: 339–355.

59. Browne WJ, Draper D. A comparison of {B}ayesian and likelihood-based methods for fitting multilevel models. Bayesian Anal. 2006;3: 473–514.

60. Jang W, Lim J. A numerical study of {PQL} estimation biases in generalized linear mixed models under heterogeneity of random effects. Commun Stat - Simul Comput. 2009;38: 692–702.

61. Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. Biostatistics. 2010;11: 397–412.

62. Schwartz L. On Bayes procedures. Z Wahrscheinlichkeitstheorie. 1965;4: 10–26. doi:10.1007/BF00535479

63. Frühwirth-Schnatter S, Frühwirth R, Held L, Rue H. Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Stat Comput. 2009;19: 479–492. doi:10.1007/s11222-008-9109-4

64. Scott SL. Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. Stat Pap. 2011;52: 87–109. doi:10.1007/s00362-009-0205-0

65. Fruhwirth-Schnatter S, Fruhwirth R. Data augmentation and MCMC for binary and multinomial logit models. In: Kneib T, Tutz G, editors. Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir. New York: Springer; 2010. pp. 111–132. doi:10.1007/978-3-7908-2413-1

66. Pirinen M, Donnelly P, Spencer CC. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann Appl Stat. 2013;7: 369–390. doi:10.1214/12-AOAS586

67. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Softw. 2010;33: 1–22.

68. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465: 627–631. doi:10.1038/nature08800

69. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet. 2012;44: 212–216. doi:10.1038/ng.1042

70. Altmann J, Alberts S, Haines S, Dubach J, Muruthi PM, Coote T, et al. Behavior predicts genetic structure in a wild primate group. Proc Natl Acad Sci. 1996;93: 5797–5801.

71. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, et al. Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci. 2012;109: 17253–60. doi:10.1073/pnas.1121249109

72. Yu JM, Wu X, Gimble JM, Guan X, Freitas M, Bunnell B. Age-related changes in mesenchymal stem cells derived from rhesus macaque bone marrow. Aging Cell. 2011;10: 66–79. doi:10.1111/j.1474-9726.2010.00646.x

73. Maegawa S, Hinkal G, Kim HS, Shen L, Zhang L, Zhang J, et al. Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. 2010;20: 332–340. doi:10.1101/gr.096826.109

74. Winnefeld M, Lyko F. The aging epigenome: DNA methylation from the cradle to the grave. Genome Biol. 2012;13: 165. doi:10.1186/gb4033

75. Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, et al. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biol. 2013;14: R102. doi:10.1186/gb-2013-14-9-r102

76. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CPG island context. PLoS Genet. 2009;5. doi:10.1371/journal.pgen.1000602

77. Rakyan VK, Down TA, Maslau S, Andrew T, Yang T, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;4: 434–439.

78. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317–330. doi:10.1038/nature14248

79. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

80. Murgatroyd C, Patchev A V, Wu Y, Micale V, Bockmühl Y, Fischer D, et al. Dynamic DNA methylation programs persistent adverse effects of early-life stress. Nat Neurosci. 2009;12: 1559–66. doi:10.1038/nn.2436

81. Ikegame T, Bundo M, Murata Y, Kasai K, Kato T, Iwamoto K. DNA methylation of the BDNF gene and its relevance to psychiatric disorders. J Hum Genet. 2013;58: 434–8. doi:10.1038/jhg.2013.65

82. Elliott E, Ezra-Nevo G, Regev L, Neufeld-Cohen A, Chen A. Resilience to social stress coincides with functional DNA methylation of the CRF gene in adult mice. Nat Neurosci. 2010;13: 1351–3. doi:10.1038/nn.2642

83. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet. 2011;13: 97–109. doi:10.1038/nrg3142

84. Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun. 2014;5: 3365. doi:10.1038/ncomms4365

85. The International HapMap Consortium. The International HapMap Project. Nature. 2003;426: 789–796. doi:10.1038/nature02168

86. Cann H, Toma D, Cazes L, Legrand M, Morel V, Piouffre L, et al. A human genome diversity cell line panel. Science. 2002;296: 261–2. doi:http://dx.doi.org/10.1108/17506200710779521

87. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;135: 0–9. doi:10.1038/nature11632

88. Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, et al. A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res. 2010; 281–290. doi:10.1101/gr.099234.109

89. Weigel D, Mott R. The 1001 genomes project for Arabidopsis thaliana. Genome Biol. 2009;10: 107. doi:10.1186/gb-2009-10-5-107

90. Quon G, Lippert C, Heckerman D, Listgarten J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. Nucleic Acids Res. 2013;41: 2095–2104. doi:10.1093/nar/gks1449

91. Akalin A, Kormaksson M. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. BioMed Central Ltd; 2012;13: R87. doi:10.1186/gb-2012-13-10-R87

92. Hansen K, Langmead B, Irizarry R. BSmooth : from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. BioMed Central Ltd; 2012;13: R83. doi:10.1186/gb-2012-13-10-R83

93. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24: 14–24. doi:10.1101/gr.155192.113

94. Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, et al. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat Genet. 2015;47: 353–360. doi:10.1038/ng.3222

95. Pickrell JJK, Marioni JJC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464: 768–772. doi:10.1038/nature08872.Understanding

96. Skelly D, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res. 2011;21: 1728–1737. doi:10.1101/gr.119784.110

97. Harvey C, Moyebrailean G, Davis O, Wen X, Luca F, Pique-Regi R. QuASAR: Quantitative allele specific analysis of reads. Bioinformatics. 2014; 1–7. doi:10.1101/007492

98. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44: 1084–1089. doi:10.1038/ng.2394

99. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nat Methods. 2014;11: 309–11. doi:10.1038/nmeth.2815

100. Alberts SC, Altmann J. The Amboseli Baboon Research Project: 40 years of continuity and change. In: Kappeler P, Watts DP, editors. Long-Term Field Studies of Primates. New York: Springer; 2012. pp. 261–288.

101. Altmann J, Altmann S, Hausfater G. Physical maturation and age estimates of yellow baboons, Papio cynocephalus, in Amboseli National Park, Kenya. Am J Primatol. 1981;1: 389–399. doi:10.1002/ajp.1350010404

102. Buchan JC, Alberts SC, Silk JB, Altmann J. True paternal care in a multi-male primate society. Nature. 2003;425: 179–81. doi:10.1038/nature01866

103. Alberts SC, Buchan JC, Altmann J. Sexual selection in wild baboons: from mating opportunities to paternity success. Anim Behav. 2006;72: 1177–1196. doi:10.1016/j.anbehav.2006.05.001

104. Wang J. COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. Mol Ecol Resour. 2011;11: 141–5. doi:10.1111/j.1755-0998.2010.02885.x

105. Tung J, Primus A, Bouley AJ, Severson TF, Alberts SC, Wray G. Evolution of a malaria resistance gene in wild primates. Nature. 2009;460: 388–91. doi:10.1038/nature08149

106. Tung J, Akinyi MY, Mutura S, Altmann J, Wray G, Alberts SC. Allele-specific gene expression in a wild nonhuman primate population. Mol Ecol. 2011;20: 725–39. doi:10.1111/j.1365-294X.2010.04970.x

107. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

108. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10: 232. doi:10.1186/1471-2105-10-232

109. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth. New York, NY: Springer; 2002.

110. Hastie T, Tibshirani R, Narasimhan B, Chu G. Impute: imputation for microarray data. R package version 1.42.0. 2015.

111. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2014;43: D662–D669. doi:10.1093/nar/gku1010

112. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42: 764–770. doi:10.1093/nar/gkt1168

113. Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, Hvilsom C, et al. Dynamics of DNA methylation in recent human and great ape evolution. PLoS Genet. 2013;9: e1003763. doi:10.1371/journal.pgen.1003763

114. Rönn T, Volkov P, Davegårdh C, Dayeh T, Hall E, Olsson AH, et al. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. PLoS Genet. 2013;9: e1003572. doi:10.1371/journal.pgen.1003572

Table 1. Current approaches for identifying differentially methylated loci in bisulfite sequencing data sets.

Statistical method	Directly models counts?	Controls for biological covariates?	Controls for genetic covariance?	Programs that implement the method
t-test or Wilcoxon rank-sum test	No	No	No	R and many others
Fisher's exact test	Yes	No	No	R and many others
Binomial regression	Yes	Yes	No	R and many others
Linear regression	No	Yes	No	R and many others
Beta-binomial model	Yes	Some[1]	No	DSS [32], MOABS [34], RadMeth [33]
Linear mixed model	No	Yes	Yes	GEMMA [46], EMMA [38], EMMAX [39], FaST-LMM [47]
Binomial mixed model	Yes	Yes	Yes	MACAU

[1] Only RadMeth; the implementations of the beta-binomial model in MOABS and DSS do not allow the user to control for covariates.

Figure 1. MACAU appropriately controls for genetic covariance in simulated and real WGBS data and eliminates false positive identification of differentially methylated sites. (A-B, D-E) The distribution of p-values for 4000 simulated true negative sites (n = 24 accessions; effect of secondary dormancy on DNA methylation levels = 0). For each simulation, h^2 was set to 0.3 (A, D) or 0.6 (B, E). Simulated data were analyzed with a beta-binomial model (A-B) or MACAU (D-E), and compared against the expected uniform distribution. (C, F) QQ-plots comparing the p-value distributions for (i) a model testing for effects of secondary dormancy on DNA methylation levels in real WGBS data, plotted on the y-axis; and (ii) the same model when the secondary dormancy values were permuted across individuals, plotted on the x-axis. The genomic control factor, λ, is shown for each set of results.

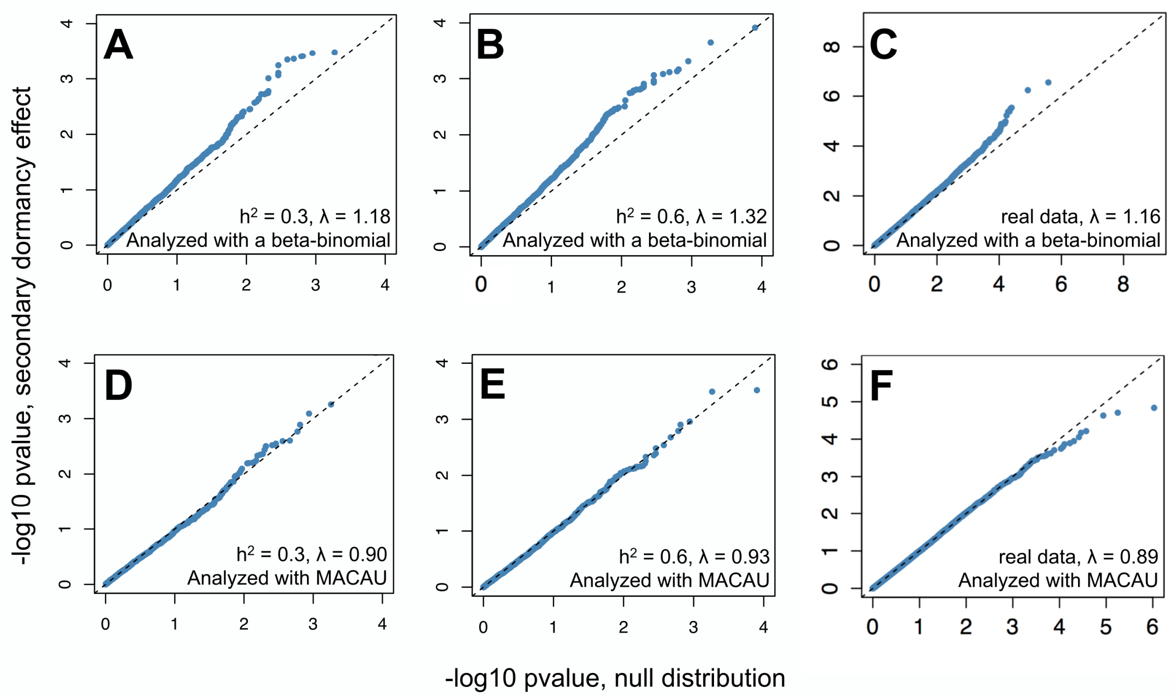


Figure 2. MACAU exhibits increased power to detect differential methylation when DNA methylation levels are heritable. Receiver operating characteristic (ROC) curves and true positive rates at a 10% false discovery rate threshold for simulated age effects on DNA methylation levels at (A-C) simulated sites with moderately heritable DNA methylation levels ($h^2 = 0.3$) and (D-F) simulated sites with highly heritable DNA methylation levels ($h^2 = 0.6$). Panels B and E are enlarged versions of panels A and D, respectively, focusing on false positive rates <0.1. Each simulated dataset contained n=80 individuals and 5000 simulated CpG sites, with 500 true positives (percent variance explained by age = 10%) and 4500 true negatives. A binomial model could not detect true positives at a false positive rate below 0.10 (when $h^2 = 0.3$) or below 0.9 (when $h^2 = 0.6$); the binomial is therefore not shown in panel B, and only shown for large x-values in panel E.

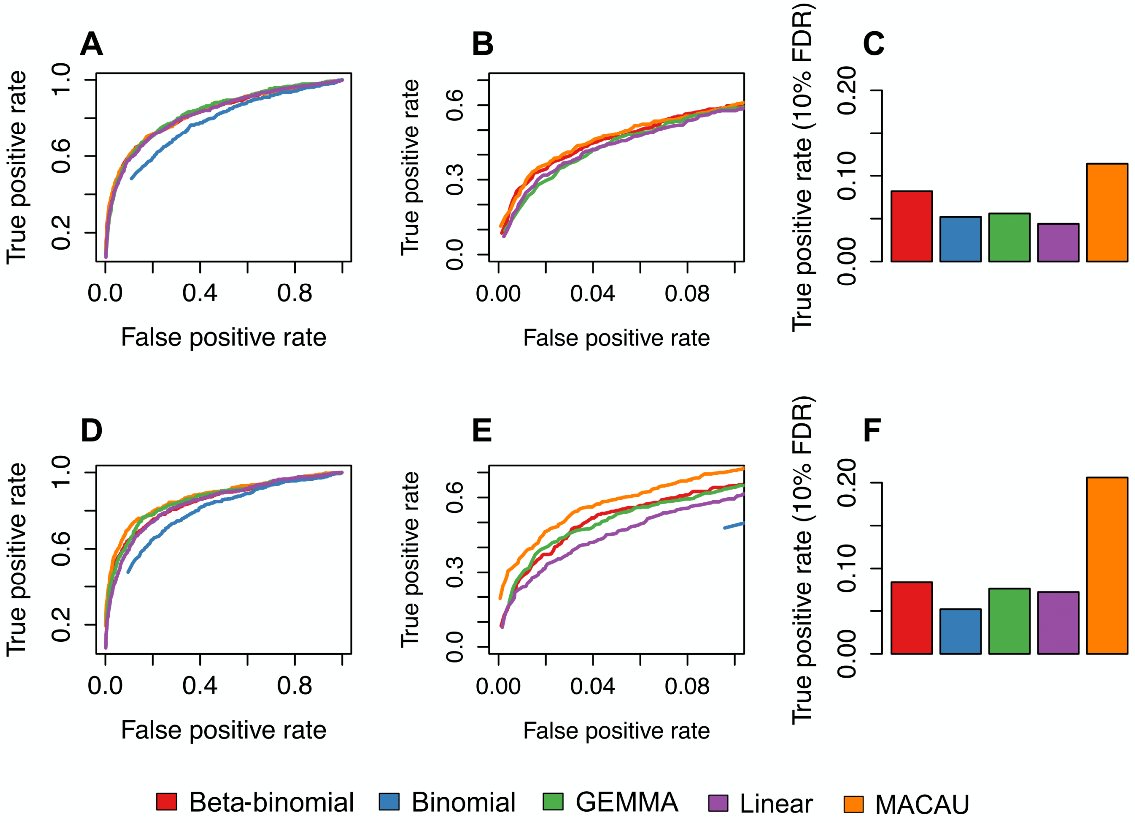
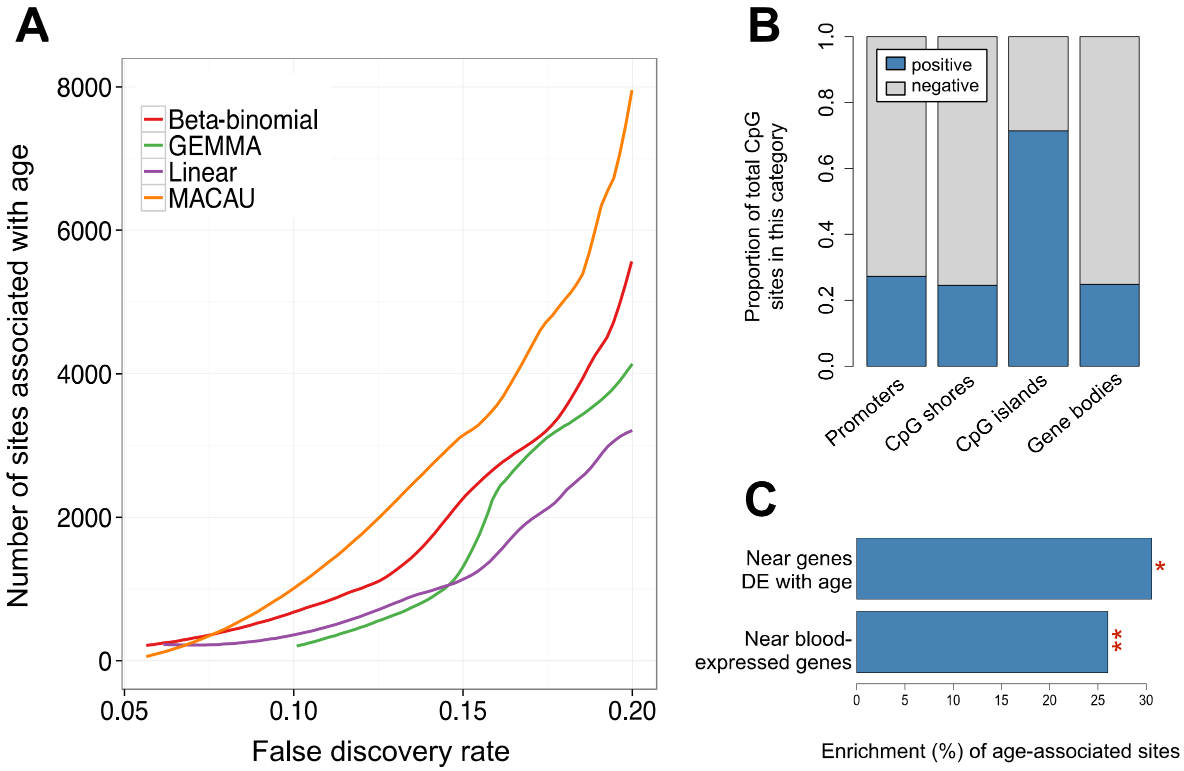
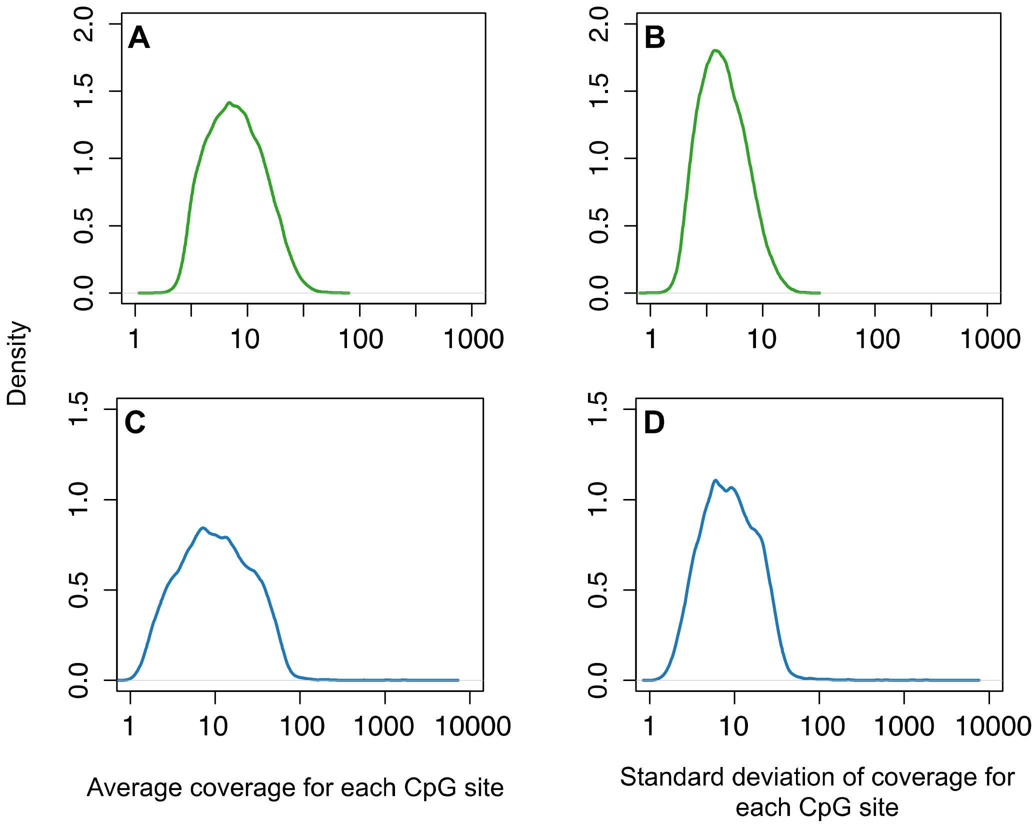


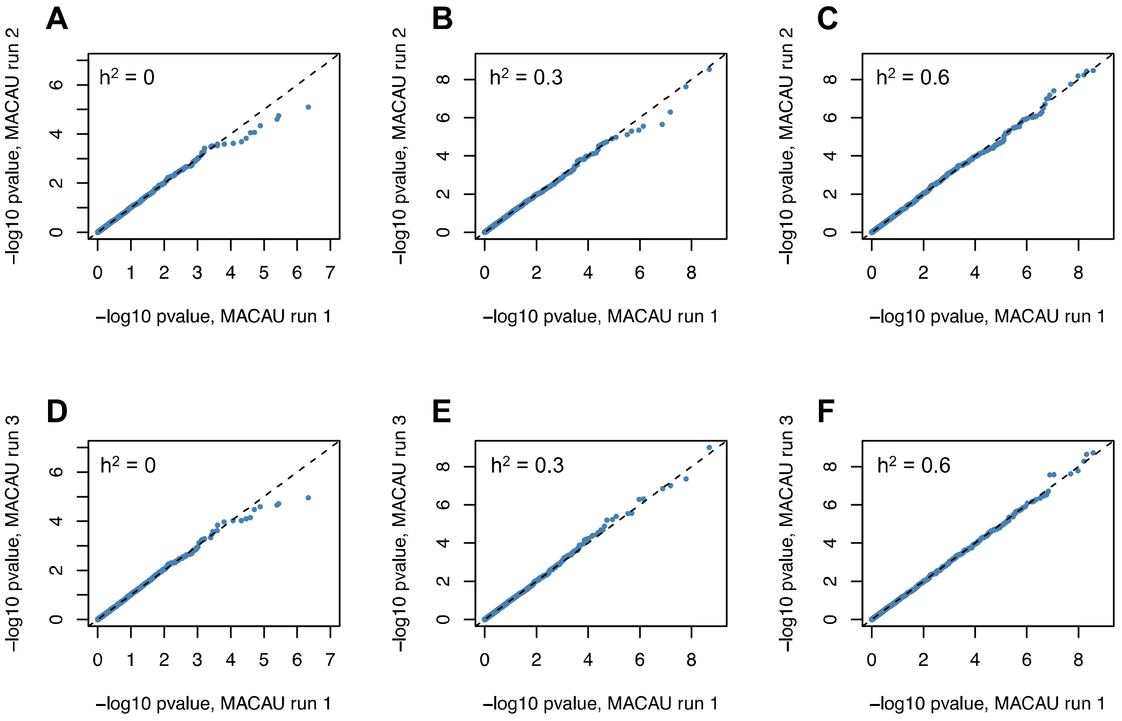
Figure 3. Age-associated CpG sites identified by MACAU in the baboon RRBS data. (A) The number of age-associated CpG sites detected at a given empirical FDR. The binomial model cannot detect age-associated sites at a false discovery rate below 0.20 and is consequently not shown. (B) For age-associated sites detected by MACAU (at a 10% FDR), the proportion of sites that gain or lose methylation with age is shown by genomic region. Positive = DNA methylation levels increase with age; Negative = DNA methylation levels decrease with age. (C) Age-associated CpG sites detected using MACAU (10% FDR) are more likely to fall near genes that are expressed in whole blood, compared to the background set of CpG sites near genes (**$p < 10^{-10}$). Further, age-associated CpG sites are more likely to occur near genes that are differentially expressed (DE) with age, compared to CpG



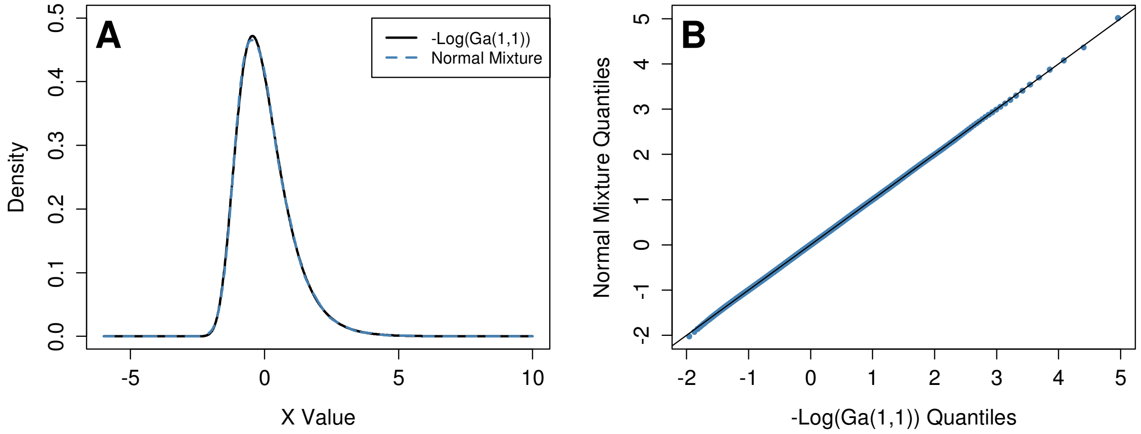
Supplementary Figure 1. In a real WGBS dataset (from *Arabidopsis*) and a real RRBS dataset (from yellow baboons), coverage varies widely across CpG sites and individuals. For each CpG site, we calculated the mean site-specific coverage across individuals, as well as the standard deviation of coverage values for those sites. The distribution of these average coverage values (A, C) and coverage standard deviation values (B, D) are shown for the *Arabidopis* WGBS dataset (A-B, in green) and the baboon RRBS dataset (C-D, in blue). The x-axes are plotted on a $\log_{10}$ scale.



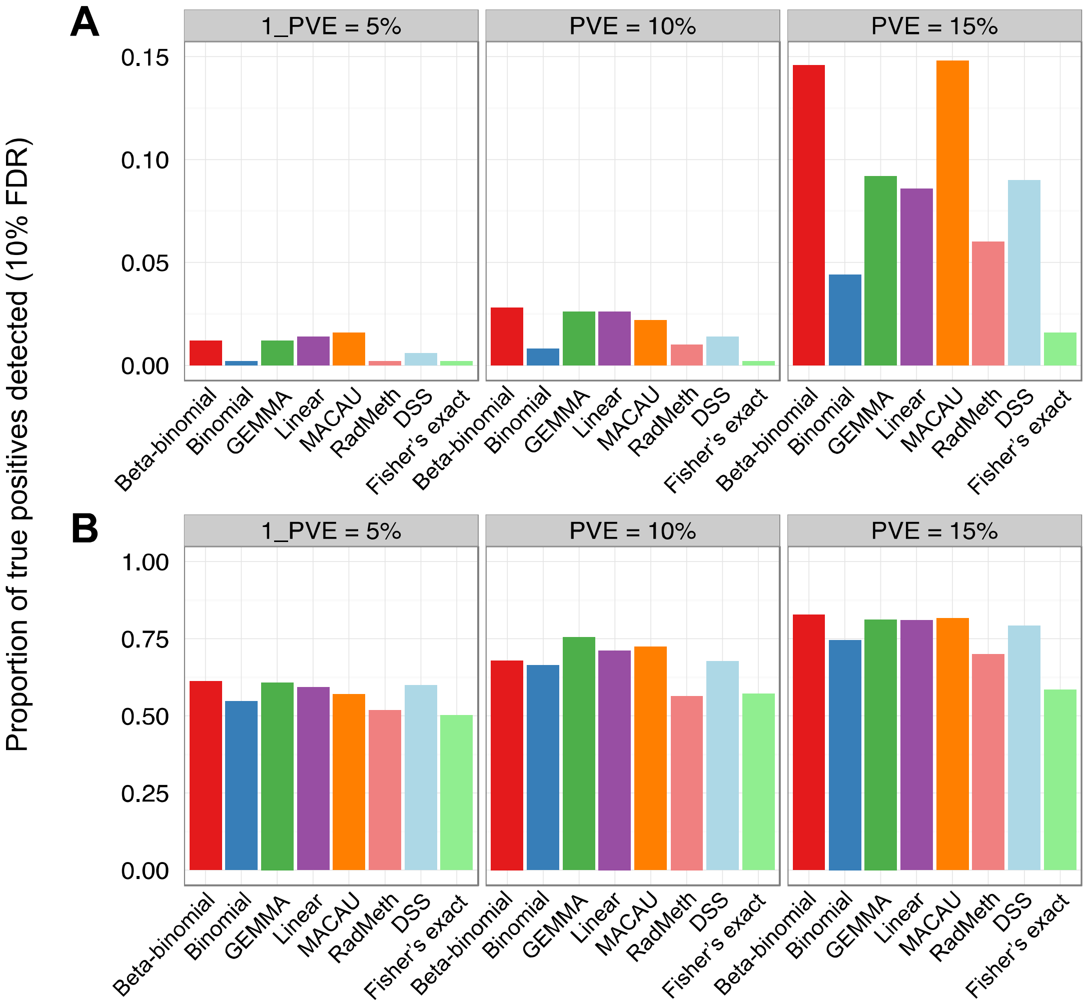
Supplementary Figure 2. MACAU p-values are consistent across runs. QQ-plots comparing the p-value distributions for 3 independent runs of MACAU on the same data sets, with different simulated heritability values (Panels A, D - h^2 = 0; Panels B, E - h^2 = 0.3; Panels C, F - h^2 = 0.6). Pairwise correlations between each independent run were R > 0.95 for h^2 = 0:,R > 0.97 for h^2 = 0.3; and R > 0.98 for h^2 = 0.6. Distributions shown are for analyses of simulated secondary dormancy effects on DNA methylation levels in the *Arabidopsis* data set (4000 sites, n=24 accessions).



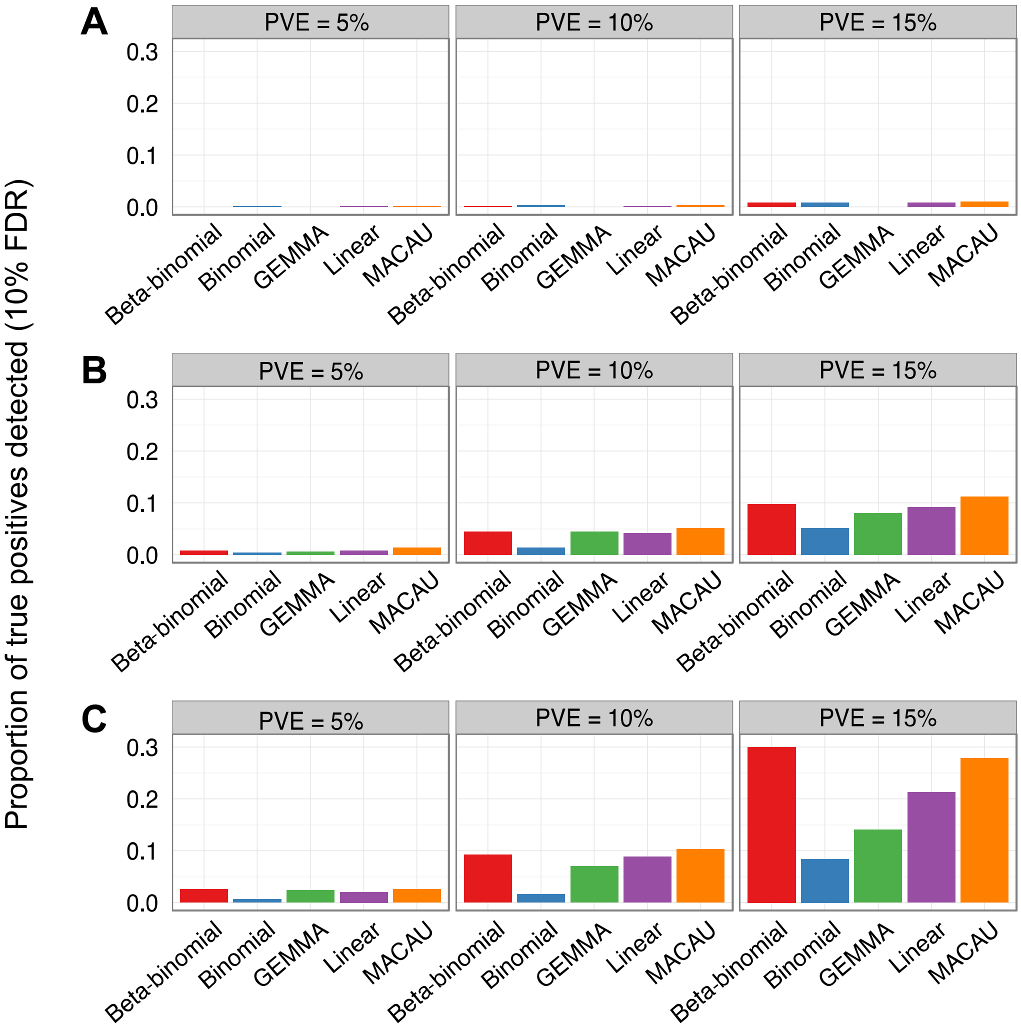
Supplementary Figure 3. The normal mixture provides an accurate approximation to the negative log gamma distribution. (A) Density plot and (B) quantile-quantile plots demonstrating that the normal mixture approximation approximates –log(Ga(r, 1)) well even in the most difficult case when r=1.



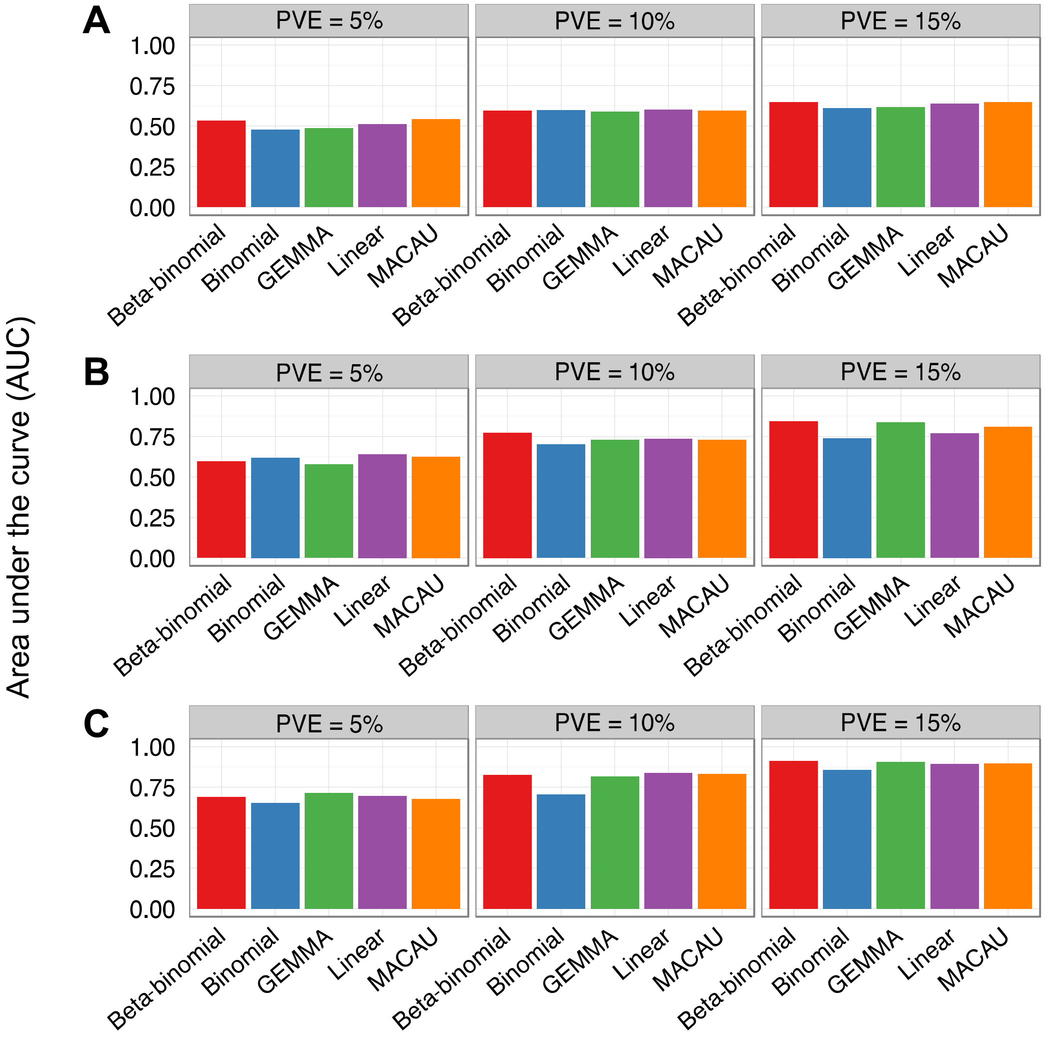
Supplementary Figure 4. Comparisons between methods when DNA methylation levels are not heritable, and the predictor variable is binarized. To include methods that can only analyze categorical differences in DNA methylation levels between two groups, we binarized age values in our simulated RRBS datasets (individuals below median age = young versus individuals above median age = old). We compared the ability of each method to detect true positives at a 10% FDR using simulated datasets (n = 5000 sites including 500 true positives and 4500 true negatives; percent variance explained by age varies as noted in the figure headings). For all simulations shown below, h^2 was set to 0. (A) Results for simulations with n = 50 individuals and (B) n = 80 indivdiuals are plotted below.



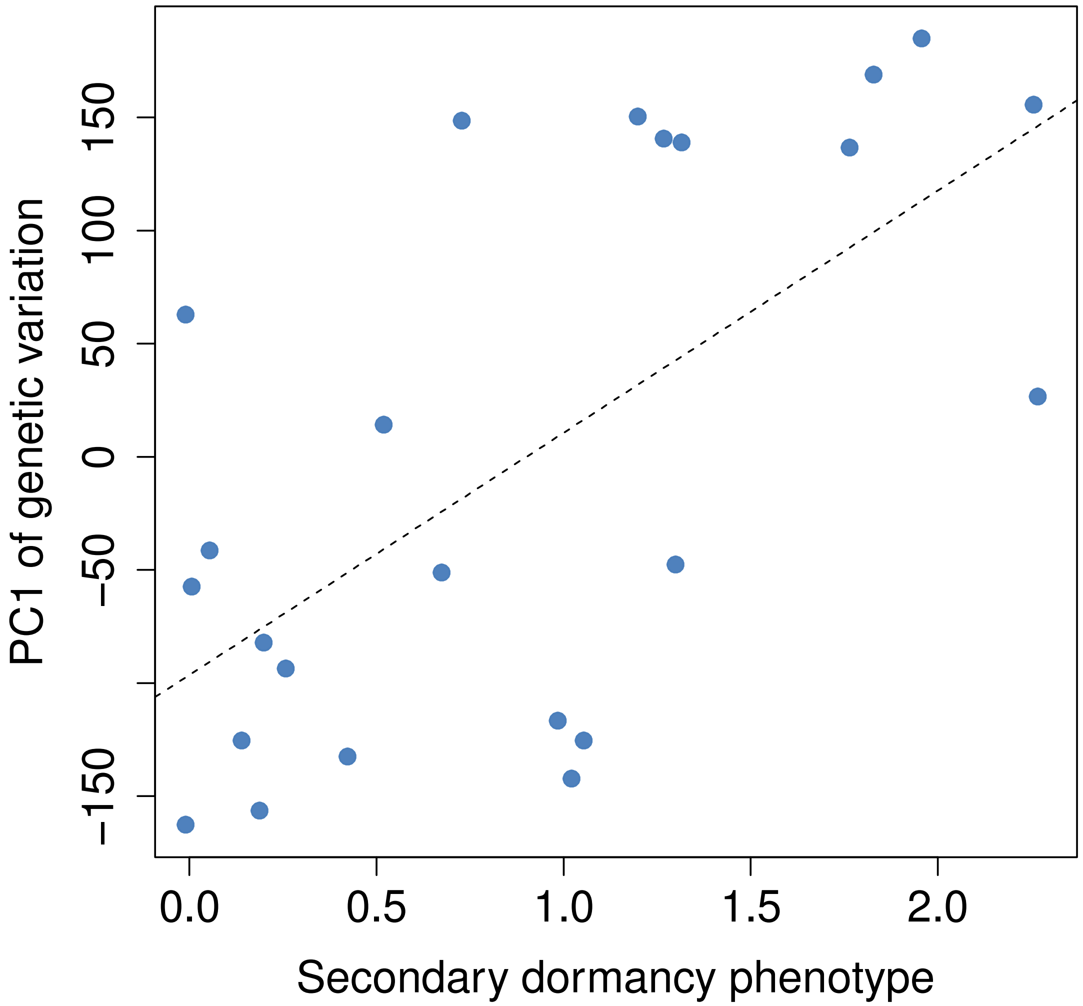
Supplementary Figure 5. Comparison across methods when DNA methylation levels are not heritable. We compared the ability of methods that work on continuous predictor variables to detect true positives at a 10% FDR using simulated data sets (n = 5000 sites including 500 true positives and 4500 true negatives; percent variance explained by age varies as noted in the figure headings). For all simulations shown below, h^2 was set to 0. (A) Results for simulations with n=20 individuals; (B) with n=50 individuals; and (C) with n=80 individuals.



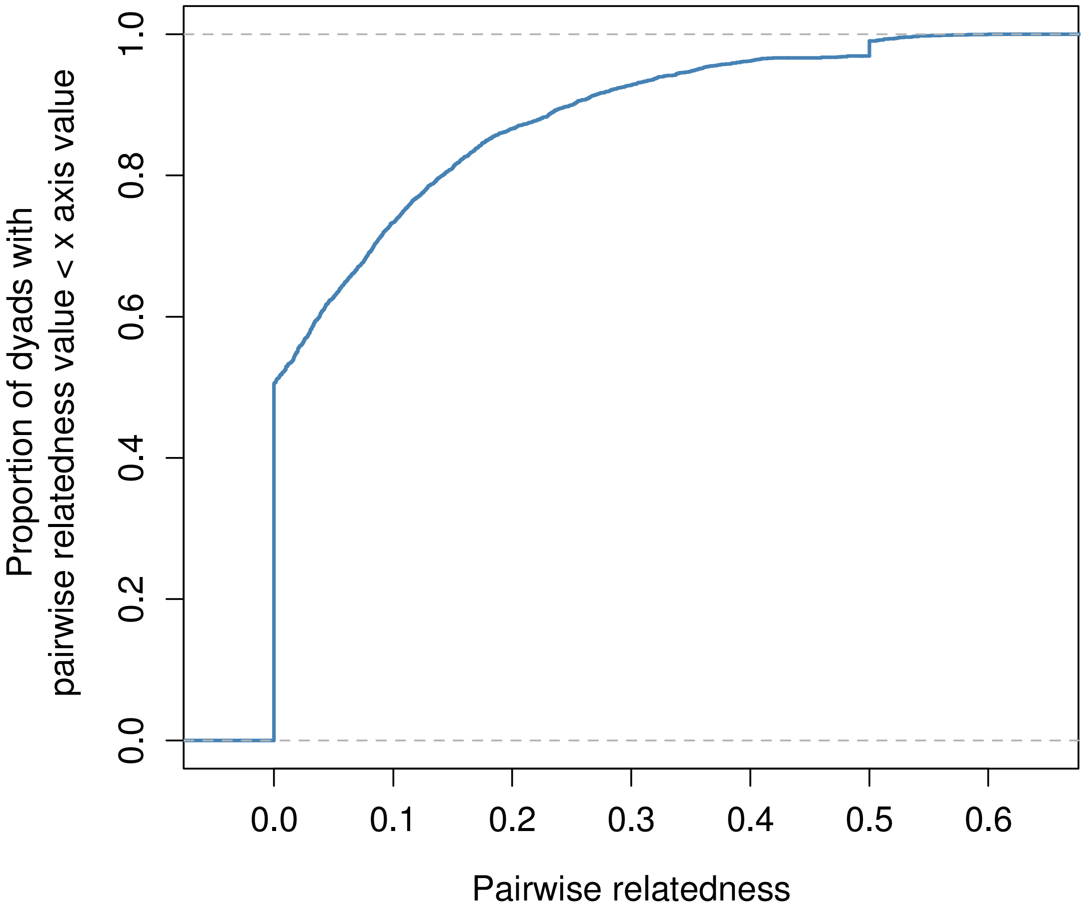
Supplementary Figure 6. ROC comparison across methods when DNA methylation levels are not heritable. Simulation parameters and sample sizes as in Supplementary Figure 3. Here, we show area under the curve for a receiver operating characteristic on on the y-axis instead of true positive detection rate. Visualized this way, the methods look more equivalent than using an FDR method because, AUC is based on true positive-false positive trade-offs across a range of p-value thresholds; methods can thus consequently yield high AUCs even when they harbor little power to detect true positives at FDR thresholds that are frequently used in practice.



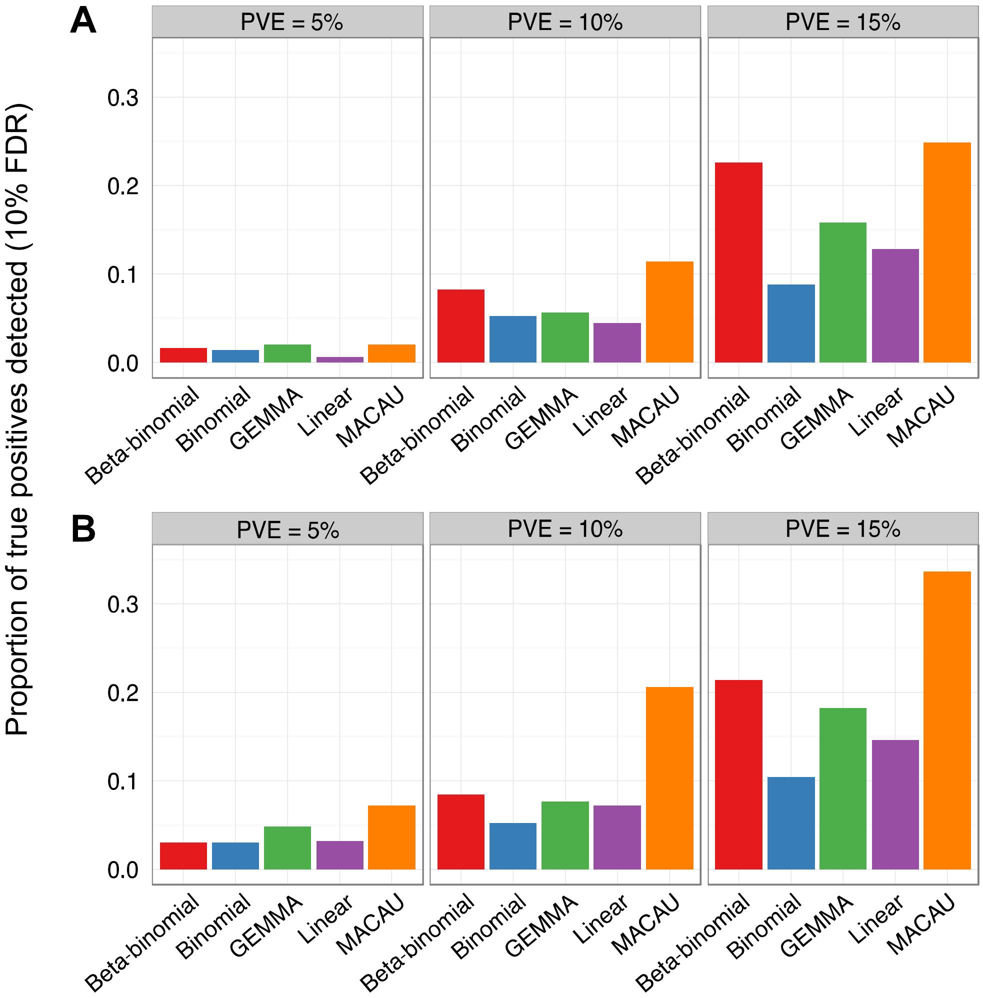
Supplementary Figure 7. Secondary dormancy is correlated with population structure in the *Arabidopsis* WGBS dataset. Principal components analysis on 188,093 genotyped sites with minor allele frequency >5% reveals that genetic background is correlated with secondary dormancy values. The correlation between the secondary dormancy phenotype values and the first principal component of the genetic relatedness matrix is $R^2 = 0.38$, p = 7.84×10^{-4} (n = 24). The first principal component (PC1) explains 8.5% of the genetic variance in the data set.



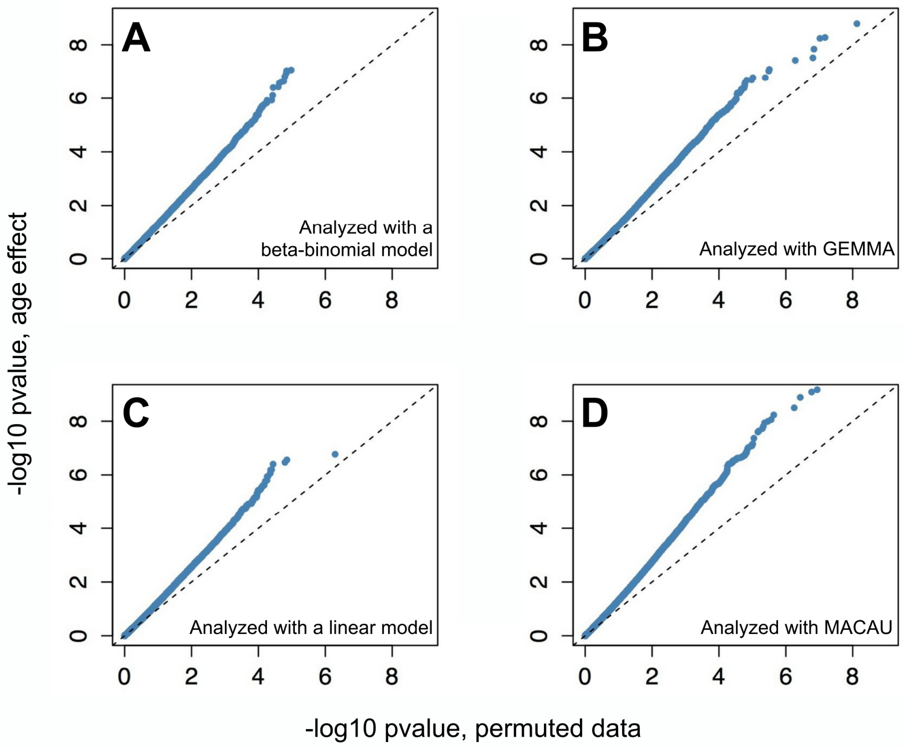
Supplementary Figure 8. Distribution of pairwise relatedness values for baboons (n=80) from a single social group, used in simulations. Approximately half of the individuals are unrelated, while a small proportion (~10%) are highly related (i.e., related at the level of half siblings or higher, r = 0.25).



Supplementary Figure 9. MACAU provides increased power to detect age-associated sites when DNA methylation levels are heritable. We simulated age effects on DNA methylation levels, in presence of genetic effects (panel A, $h^2 = 0.3$; panel B, $h^2 = 0.6$) across a range of effect sizes. The proportion of true positives detected at a 10% empirical FDR is plotted for each method and simulated dataset.



Supplementary Figure 10. Distribution of p-values from four different methods for the real RRBS data. QQ-plots comparing the p-value distributions for (i) a model testing for effects of age on DNA methylation levels in real RRBS data, plotted on the y-axis; and (ii) the same model when the age values were permuted across individuals, plotted on the x-axis.



Supplementary Figure 11. Distribution of sites covered in the baboon RRBS dataset (n = 433, 871 CpG sites). (A) Absolute number of sites analyzed for a given genomic region. See *Materials and Methods* for information on how we defined each genomic region. (B) Proportion of total annotated features in the baboon genome for which a least one CpG site was analyzed in this data set.

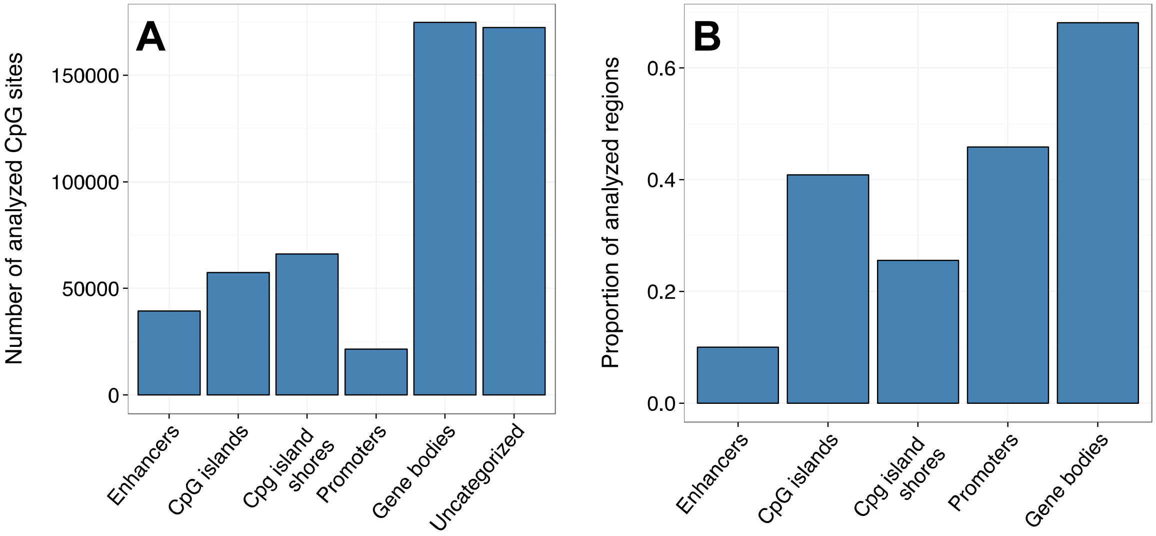


Table S1. Normal Mixture Approximations to -log(Ga(r, 1)) for r in [1, 5]

		k									
r		1	2	3	4	5	6	7	8	9	10
1	w_{rk}	0.2924	0.2828	0.1625	0.09697	0.08053	0.05949	0.01882	0.005167	0.001194	0.0001863
	$\frac{m_{rk}+\Psi(r)}{\sqrt{\Psi'(r)}}$	-0.8141	0.02214	0.8407	-0.7554	-0.1850	1.753	2.747	3.857	5.168	6.964
	$\sigma_r^2 / \Psi'(r)$	0.1904									
2	w_{rk}	0.3960	0.3893	0.1640	0.04299	0.007059	0.0006391				
	$\frac{m_{rk}+\Psi(r)}{\sqrt{\Psi'(r)}}$	-0.7848	0.06357	1.023	2.126	3.460	5.249				
	$\sigma_r^2 / \Psi'(r)$	0.2892									
3	w_{rk}	0.4862	0.3882	0.1093	0.01527	0.001014					
	$\frac{m_{rk}+\Psi(r)}{\sqrt{\Psi'(r)}}$	-0.6673	0.3084	1.443	2.779	4.496					
	$\sigma_r^2 / \Psi'(r)$	0.3804									
4	w_{rk}	0.5861	0.3503	0.06010	0.003509						
	$\frac{m_{rk}+\Psi(r)}{\sqrt{\Psi'(r)}}$	-0.5422	0.5501	1.873	3.578						
	$\sigma_r^2 / \Psi'(r)$	0.4660									
5	w_{rk}	0.5617	0.3696	0.06518	0.003533						
	$\frac{m_{rk}+\Psi(r)}{\sqrt{\Psi'(r)}}$	-0.5582	0.5013	1.784	3.439						
	$\sigma_r^2 / \Psi'(r)$	0.4829									

Table S2. Normal Mixture Approximations to -log(Ga(r, 1)) for r in [6, 170]

		k		
r		1	2	3
6-15	w_{rk}	$\frac{-0.6583 + 0.07464r + 0.1884r^2}{-0.03083 - 0.2930r + 0.3067r^2}$	$\frac{1.586 - 0.7519r + 0.3535r^2}{0.2643 + 0.1614r + 0.9698r^2}$	$\frac{0.01348 + 0.001274r - 0.00003837r^2}{0.8799 + 0.04313r - 0.001552r^2}$
	$\frac{m_{rk} + \Psi(r)}{\sqrt{\Psi'(r)}}$	$\frac{-0.3696 - 0.006706r - 0.009308r^2}{1.034 + 0.003362r + 0.02403r^2}$	$\frac{-0.8303 + 0.3906r + 0.09007r^2}{0.1318 - 0.09864r + 0.1682r^2}$	$\frac{-1.183 + 0.03989r + 0.4559r^2}{1.262 - 0.7045r + 0.2549r^2}$
	$\sigma_r^2 / \Psi'(r)$	$\frac{0.06108 + 0.6634r + 0.08889r^2}{0.3702 + 1.319r + 0.1145r^2}$		
16-170	w_{rk}	$\frac{0.6928 + 0.03790r + 0.00007142r^2}{0.7754 + 0.04535r + 0.00008905r^2}$	$\frac{0.8263 + 0.1529r + 0.001124r^2}{8.827 + 0.9978r + 0.006043r^2}$	
	$\frac{m_{rk} + \Psi(r)}{\sqrt{\Psi'(r)}}$	$\frac{-0.8917 - 0.1855r - 0.0009084r^2}{4.192 + 0.9940r + 0.007033r^2}$	$\frac{1.076 + 0.07260r + 0.0002470r^2}{0.5983 + 0.07564r + 0.0004561r^2}$	
	$\sigma_r^2 / \Psi'(r)$	$\frac{0.5995 + 0.03782r + 0.00001488r^2}{0.8664 + 0.04284r + 0.00001132r^2}$		

Table S3. Computation times for each method on the two real datasets. Computation was performed on a single core of an Intel Xeon L5420 2.50 GHz processor. *n* = number of individuals; *m* = number of sites.

Method	Software	Computation Time *Arabidopsis* (*n*=24, *m*=830,676)	 Baboon (*n*=50, *m*=433,871)
Linear model	R (lm)	0.55 min	0.44 min
Linear mixed model	GEMMA	1.3 min	1.2 min
Binomial model	R (glm)	71 min	51 min
Beta binomial model	R (self-implemented)	2 d	4.5 d
Binomial mixed model	MACAU	9.5 h	11 h
Binomial mixed model	MCMCglmm	12 d	19 d

Table S4. Baboon RRBS dataset sample characteristics and read mapping summary

Individual	Sex	Age of sampled individual	Bisulfite conversion rate	Age of blood sample	Total reads generated (in millions)	Uniquely mapped reads (in millions)	Proportion of uniquely mapped reads
AMB_01	M	11.29	0.9850	8.39	37.023210	25.099657	0.678
AMB_02	F	10.05	0.9994	6.37	33.071988	22.819672	0.690
AMB_03	M	7.67	0.9842	20.20	24.088246	16.943184	0.703
AMB_04	M	5.40	0.9988	25.16	14.728885	10.457508	0.710
AMB_05	M	18.01	0.9849	4.13	51.051990	35.687494	0.699
AMB_06	M	6.39	0.9847	25.21	21.887490	14.799792	0.676
AMB_07	M	6.85	0.9840	25.13	14.934174	10.012808	0.670
AMB_08	M	7.92	0.9988	25.21	32.611582	22.532321	0.691
AMB_09	M	5.16	0.9994	25.14	14.676613	10.609725	0.723
AMB_10	M	6.25	0.9837	25.13	35.170196	23.063683	0.656
AMB_11	F	14.56	0.9995	25.16	18.718679	13.103075	0.700
AMB_12	M	3.98	0.9837	6.29	26.055629	17.659530	0.678
AMB_13	M	6.01	0.9840	25.07	24.439863	16.309385	0.667
AMB_14	M	3.76	0.9989	25.13	20.659507	14.072821	0.681
AMB_15	F	9.53	0.9989	6.30	9.586029	7.285382	0.760
AMB_16	F	7.84	0.9994	25.17	18.432235	12.718242	0.690
AMB_17	M	11.01	0.9990	7.42	18.548701	12.902723	0.696
AMB_18	M	15.79	0.9990	20.20	36.644760	25.192966	0.687
AMB_19	M	3.04	0.9990	21.16	31.059312	21.320848	0.686
AMB_20	M	4.50	0.9990	25.09	29.389242	20.757701	0.706
AMB_21	F	6.71	0.9995	25.20	28.665765	19.779378	0.690
AMB_22	F	5.23	0.9994	25.20	16.783514	12.084130	0.720
AMB_23	M	9.79	0.9963	25.16	11.771241	8.083930	0.687
AMB_24	M	4.27	0.9987	25.13	24.482993	16.747343	0.684
AMB_25	M	6.00	0.9986	8.34	71.814200	42.517354	0.592
AMB_26	M	1.76	0.9987	20.88	15.461068	10.783672	0.697
AMB_27	M	5.98	0.9987	25.16	31.122370	21.156449	0.680
AMB_28	M	8.29	0.9980	25.09	35.575292	24.679908	0.694
AMB_29	M	4.79	0.9981	25.18	35.878244	25.526024	0.711
AMB_30	M	14.01	0.9980	25.20	15.382161	10.708392	0.696
AMB_31	M	2.90	0.9980	24.25	34.859844	24.044579	0.690
AMB_32	M	14.30	0.9980	24.28	21.899784	16.168539	0.738
AMB_33	F	5.03	0.9988	25.13	20.592762	14.620861	0.710
AMB_34	F	6.13	0.9963	25.08	39.120891	27.384624	0.700
AMB_35	F	3.96	0.9994	6.30	19.535813	13.870427	0.710
AMB_36	M	6.76	0.9978	5.92	39.790846	27.010700	0.679
AMB_37	M	6.11	0.9978	22.64	41.870572	29.168979	0.697

AMB_38	M	14.01	0.9978	25.15	23.945218	18.062503	0.754
AMB_39	F	8.10	0.9994	25.18	19.088828	13.553068	0.710
AMB_40	F	4.97	0.9995	25.22	22.715148	15.673452	0.690
AMB_41	F	3.49	0.9988	24.30	37.164581	26.015207	0.700
AMB_42	M	18.01	0.9977	24.25	30.003417	21.484400	0.716
AMB_43	F	4.69	0.9994	24.30	27.103481	18.972437	0.700
AMB_44	M	5.80	0.9990	25.22	23.952634	16.974279	0.709
AMB_45	F	16.44	0.9995	25.19	16.226065	11.682767	0.720
AMB_46	F	4.01	0.9964	25.21	53.669349	37.031851	0.690
AMB_47	M	3.64	0.9990	25.13	30.674203	20.747151	0.676
AMB_48	M	10.62	0.9991	25.22	37.266333	26.407920	0.709
AMB_49	M	11.85	0.9987	25.16	29.499525	20.155265	0.683
AMB_50	M	6.72	0.9988	23.27	79.784041	54.078860	0.678
Mean		7.79	0.9963	21.28	28.969570	19.970459	0.695
Standard deviation		4.19	0.0052	7.08	13.732295	8.938434	0.025

Text S1: Detailed Methods

1 Binomial Mixed Model

To detect differentially methylated sites, we model each potential target of DNA methylation one site at a time. For each site, we consider the following binomial mixed model (BMM):

$$y_i \sim \text{Bin}(r_i, \pi_i), \tag{1}$$

where r_i is the total read count for ith individual; y_i is the methylated read count for that individual, constrained to be an integer value less than or equal to r_i; and π_i is an unknown parameter that represents the true proportion of methylated reads for the individual at the site. We use a logit link to model π_i as a linear function of parameters:

$$\text{logit}(\pi_i) = \log(\lambda_i) = \mathbf{w}_i^T \boldsymbol{\alpha} + x_i\beta + g_i + e_i, \tag{2}$$

$$\mathbf{g} = c(g_1, \cdots, g_n)^T \sim \text{MVN}(0, \sigma^2 h^2 \mathbf{K}), \tag{3}$$

$$\mathbf{e} = c(e_1, \cdots, e_n)^T \sim \text{MVN}(0, \sigma^2(1-h^2)\mathbf{I}_{n\times n}), \tag{4}$$

where logit denotes a logistic transformation $\text{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$; $\lambda_i = \frac{\pi_i}{1-\pi_i}$ is the odds; $\mathbf{w}_i$ is a c-vector of covariates including an intercept and $\boldsymbol{\alpha}$ is a c-vector of corresponding coefficients; x_i is the predictor of interest and β is its coefficient; $\mathbf{g}$ is an n-vector of genetic random effects that model correlation due to population structure or individual relatedness; $\mathbf{e}$ is an n-vector of environmental residual errors that model independent variation; $\mathbf{K}$ is a known n by n relatedness matrix that can be calculated based on a pedigree or genotype data and that has been standardized to ensure $tr(\mathbf{K})/n = 1$ (this ensures that h^2 lies between 0 and 1, and can be interpreted as heritability, see [1]); $\mathbf{I}$ is an n by n identity matrix; $\sigma^2 h^2$ is the genetic variance component; $\sigma^2(1-h^2)$ is the environmental variance component; h^2 is the heritability of the logit transformed methylation proportion (i.e. $\text{logit}(\pi)$); and MVN denotes the multivariate normal distribution.

The binomial mixed model proposed here belongs to the generalized linear mixed model family [2]. Both $\mathbf{g}$ and $\mathbf{e}$ model over-dispersion, the increased variance in the data that is not explained by the binomial model. However, they model different aspects of over-dispersion: $\mathbf{e}$ models the variation that is due to independent environmental noise (a known problem in data sets based on sequencing reads), while $\mathbf{g}$ models the variation that is explained by kinship or population structure. Effectively, our model improves and generalizes the previous beta binomial model by introducing this extra $\mathbf{g}$ term to model individual relatedness due to kinship, population structure, or stratification.

2 Inference Method Overview

We are interested in testing the null hypothesis $H_0 : \beta = 0$. This requires obtaining the maximum likelihood estimate $\hat{\beta}$ from the model. Unlike its linear counter-part, obtaining the estimate of β from the binomial mixed model is not a trivial task, as the joint likelihood consists of an n-dimensional integral that cannot be solved analytically [2]. Previous frequentist approaches to address this problem include direct numerical integration using Gauss-Hermite quadrature [3], or Laplace approximation that is applied to the likelihood function [4] or the quasi-likelihood function [5–8]. However, both numerical integration and analytic approximation do not scale well with the increasing dimension of the integral, which unfortunately equals the sample size in our model. Even a second order Laplace approximation yields a biased estimate and overly narrow confidence interval, especially when the uncertainty in the variance component estimate is large [9–13]. Therefore, frequentist approaches for estimation and inference in the binomial mixed model remain notoriously difficult and is still an active area of research [14].

In contrast to the frequentist methods, Markov chain Monte Carlo (MCMC)-based Bayesian approaches provide an appearing alternative [11]. Bayesian methods naturally account for the uncertainty in the variance component estimates and can achieve arbitrary inference accuracy if the chain is allowed to run long enough. Despite these attractive theoretical features, however, constructing an efficient MCMC algorithm for practical problems is not easy. Previous MCMC approaches for generalized linear mixed models either require a normal approximation to the likelihood function that diminishes its gains over the frequentist methods [15,16], or use n-steps of Metropolis–Hastings algorithm to sample the n-dimensional latent rate variables where efficient proposal distributions for all of them can be hard to construct [17,18]. To improve upon these previous approaches, a new MCMC algorithm [19–21] has been recently developed based on auxiliary variable representation of the binomial distribution [22]. By introducing latent variables to replace the observed count data, the algorithm makes sampling and computation relatively straightforward.

Therefore, we rely on this particular form of MCMC in the present study. Our main contribution is to further develop an accurate approximation to the distribution of these latent variables, where the approximation form is specifically designed to allow us to adapt recent mixed model innovations [23–26] that substantially reduce the computational burden. By using a mean-normal mixture approximation to the negative log gamma distribution, our approach reduces the per-MCMC iteration computational complexity from $O(n^3)$ to $O(n^2)$, where n is the sample size. This modification allows the binomial mixed model to be efficiently applied to hundreds of individuals and millions of methylation sites.

Although we use MCMC for posterior sampling, our primary goal is not to perform a Bayesian analysis by producing Bayes factors for model comparison (although this is an interesting area to explore in the future). Rather, our goal is to use MCMC as a convenient and accurate tool to obtain the marginal likelihood of β that is otherwise infeasible or inaccurate to obtain under various frequentist approaches. Under asymptotics, both the likelihood function and the marginal posterior distribution for β will be approximately normal [27]. Since the likelihood function is simply the difference between the posterior and the prior, once we have obtained the posterior mean and standard deviation of β and paired these values to their prior counter-parts, we can easily obtain the approximate likelihood function and compute the approximate maximum likelihood estimate $\hat{\beta}$ and its standard error $se(\hat{\beta})$ using the method of moments. We can then construct approximate Wald test statistics and p values for hypothesis testing.

In the present study, we use flat priors for all nuisance parameters $(\boldsymbol{\alpha}, \sigma^2, h^2)$, or $p(\boldsymbol{\alpha}) \sim 1$, $p(\sigma^2) \sim 1$ and $p(h^2) \sim 1$. For the parameter of interest, β, we could also use a flat prior, in which case the posterior would be the likelihood. For computational stability reasons, however, we use a relatively informative prior, $\beta \sim N(0, \sigma_b^2)$ instead. A relatively informative prior restricts the sampling space when the likelihood is not informative, allowing efficient posterior sampling. Since we rely on the difference betwen the posterior and the prior for approximate inference, the choice of prior for β does not influence the eventual results. In the present study, we set $\sigma_b^2 = 1$.

Applications to real data confirm that this procedure produces well-calibrated p-values (Figure 1), suggesting that a few dozen samples are large enough to ensure asymptotic behavior. Moreover, although our approach is inherently stochastic – because the posterior mean and standard deviation of β may be slightly different for different chains – we show that a thousand MCMC iterations per site is large enough to produce stable estimates of the test statistics and p values (Figure S2).

3 The MACAU Algorithm

Below, we describe the MACAU algorithm, for Mixed model Association for Count data via data AUgmentation, in detail.

3.1 Data Augmentation

To bypass the difficult likelihood function that results from the count nature of the data, we introduce continuous auxiliary variables to replace y_i. For ith individual, observing y_i methylated reads out of r_i total reads is equivalent to observing a sequence of r_i binary read indicators $(y_{i1}, \cdots, y_{ir_i})$, where $y_{ij} = 1$ indicates that the jth read is a methylated read and $y_{ij} = 0$ indicates otherwise. Obviously, $y_i = \sum_{j=1}^{r_i} y_{ij}$. We can view each y_{ij} as a random variable generated from a logistic regression model with mean $\log(\lambda_i)$. We further introduce a continuous latent variable u_{ij} [19,20], often referred to as a utility [22], such that

$$u_{ij} = \log(\lambda_i) + \epsilon^1_{ij}, \quad \epsilon^1_{ij} \sim \mathrm{EV}(0,1), \tag{5}$$

where EV(0, 1) denotes a standard type-1 extreme value distribution with density function $e^{-x}e^{-e^{-x}}$. Then

$$y_{ij} = \begin{cases} 1, & \text{if } u_{ij} > \epsilon^0_{ij}, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\epsilon^0_{ij} \sim \mathrm{EV}(0,1)$. The above two equations come from the fact that the difference between two type-1 extreme value distributed random variables follows a logistic distribution, and a random variable that follows a logistic distribution serves as a liability variable for a logistic regression [22].

The attractive feature of introducing u_{ij} is that, conditional on all u_{ij}, the posterior of $(\boldsymbol{\alpha}, \beta, \sigma^2, h^2)$ no longer depends on the observed methylated read indicator y_{ij}, hence removing the non-linearity constraint that comes with the binomial aspect of our model. Applying the relationship between the EV distribution and the exponential distribution, we have $e^{-u_{ij}} \sim \mathrm{Exp}(\lambda_i)$ and $e^{-\epsilon^0_{ij}} \sim \mathrm{Exp}(1)$, where Exp denotes the exponential distribution. This relationship allows us to easily sample u_{ij} conditional on λ_i and y_{ij} based on the convenient exponential distribution rather than the more difficult EV distribution, as $e^{-u_{ij}} \sim \mathrm{Exp}(1+\lambda_i)$ if $y_{ij} = 1$ and $e^{-u_{ij}} \sim \mathrm{Exp}(1+\lambda_i) + \mathrm{Exp}(\lambda_i)$ if $y_{ij} = 0$.

An undesirable feature of the above approach, however, is that we have to work with a much larger latent space of u_{ij} than the original n observations of y_i. This drawback can be mitigated by combining all exponentiated negative latent utilities together [21], by introducing a new latent variable

$$z_i = -\log(\sum_{j=1}^{r_i} e^{-u_{ij}}) = \log(\lambda_i) + \epsilon_i, \tag{7}$$

where $\epsilon_i = -\log(\sum_{j=1}^{r_i} e^{-\epsilon^1_{ij}})$ follows a negative log gamma distribution, $-\log(\mathrm{Ga}(r_i, 1))$; Ga denotes a gamma distribution with the two parameters representing shape and rate, respectively. This is because a gamma random variable is a summation of independent exponential random variables with a same rate parameter.

Using the latent variable z_i instead of u_{ij} reduces the size of the latent space back to the observed space. Conditional on z_i, we again do not need to use y_i, allowing us to bypass the count feature of the observed data in the algorithm.

3.2 Normal Mixture Approximation

To further circumvent the difficulty introduced by the non-normality of ϵ_i, we follow previous ideas [20,21] to approximate the non-normal distribution by using a mixture of normals. Importantly, we take advantage of recent innovations in efficient mixed model algorithms [23–26] by using a mean mixture of normals where each normal distribution has a different mean but share the same variance.

Specifically, for every possible integer value of r, we identify a normal approximation in the form of $\sum_{k=1}^{k_r} w_{rk}\mathrm{N}(m_{rk}, s^2_r)$, to the negative log gamma distribution $-\log(\mathrm{Ga}(r,1))$. Because the mean ($-\Psi(r)$, where Ψ denotes a digamma function) and the variance ($\Psi'(r)$, where Ψ' denotes a trigamma function) of the negative log gamma distribution is a function of r, to ensure approximation stability we work on the standardized version of the negative log gamma distribution, by centering with the mean and standardizing with the standard deviation. Then, we estimate the number of components k_r, the weights w_{rk}, the means m_{rk} and the variance s_r^2 via the Nelder-Mead algorithm by minimizing the Kullback–Leibler (KL) divergence between the two distributions. These parameter estimates ensure that the KL divergence is smaller than 0.0005, so that the difference between the approximate and the exact distributions are ignorable in practice. Because the negative log gamma distribution asymptotically approximates a normal distribution, the approximation becomes easier for larger r. Therefore, we can use increasingly smaller number of normal components for accurate approximation.

For small values of r ($r \in [1, 5]$), we provide detailed parameter values in Table S1. For median values of r ($r \in [6, 169]$), we no longer need to store parameters for every r. Instead, we can interpolate the weight, mean and variance estimates across the range of r using rational functions without loss of accuracy. These functions are provided in the Table S2. For large values of r ($r \in [170, \infty)$, we use a single normal distribution $\mathrm{N}(0, \Psi'(r))$ for approximation. The mean normal mixture approximations are accurate. Even in the most difficult case where $r = 1$, we only observe small difference between the approximate and the exact distributions (Figure S3).

3.3 Detailed Sampling Steps and Efficient Computation

Now we are ready to describe the detailed MCMC algorithm. Here, with the normal mixture approximation, we have

$$z_i = \log(\lambda_i) + \epsilon_i = \mathbf{w}_i^T\boldsymbol{\alpha} + x_i\beta_i + g_i + e_i + \epsilon_i, \quad \epsilon_i \sim \sum_{k=1}^{k_{r_i}} w_{r_ik}\mathrm{N}(m_{r_ik}, s_{r_i}^2). \tag{8}$$

We introduce a vector of latent indicators $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_n)$, where each $\gamma_i \in (1, \cdots, k_{r_i})$ indicates which normal component the corresponding ϵ_i is from. Conditional on z_i and $(\boldsymbol{\alpha}, \beta, g_i, e_i)$, we have

$$P(\gamma_i = k|z_i, \boldsymbol{\alpha}, \beta, g_i, e_i) \propto w_{r_ik}\Phi(z_i - \log(\lambda_i) - m_{r_ik}, \sigma_{r_i}^2), \tag{9}$$

where $k \in (1, \cdots, k_{r_i})$ and Φ denotes the normal density function. Conditional on $\boldsymbol{\gamma}$, we can integrate out $\boldsymbol{\alpha}$, β, $\mathbf{g}$, $\mathbf{e}$ and $\boldsymbol{\epsilon}$ analytically to obtain the marginal distribution of σ^2 and h^2,

$$P(\sigma^2, h^2|\mathbf{z}, \boldsymbol{\gamma}) \propto |\mathbf{H}|^{-\frac{1}{2}}|\mathbf{W}^T\mathbf{H}^{-1}\mathbf{W}|^{-\frac{1}{2}}|\sigma_b^2\mathbf{x}^T\mathbf{P}_w\mathbf{x} + 1|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{z}-\mathbf{m}_{\boldsymbol{\gamma}})^T\mathbf{P}_x(\mathbf{z}-\mathbf{m}_{\boldsymbol{\gamma}})}, \tag{10}$$

where $\mathbf{z} = (z_1, \cdots, z_n)^T$, $\mathbf{m}_{\boldsymbol{\gamma}} = (m_{r_1\gamma_1}, \cdots, m_{r_n\gamma_n})^T$, $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_n)^T$, $\mathbf{D}_\mathbf{r}$ is an n by n diagonal matrix with iith element $\sigma_{r_i}^2$, $\mathbf{V} = h^2\mathbf{K} + (1-h^2)\mathbf{I}$, $\mathbf{H} = \sigma^2\mathbf{V} + \mathbf{D}_\mathbf{r}$, $\mathbf{P}_w = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{W}^T(\mathbf{W}^T\mathbf{H}^{-1}\mathbf{W})^{-1}\mathbf{W}\mathbf{H}^{-1}$ and $\mathbf{P}_x = \mathbf{P}_w - \mathbf{P}_w\mathbf{x}(\mathbf{x}^T\mathbf{P}_w\mathbf{x} + \sigma_b^{-2})^{-1}\mathbf{x}^T\mathbf{P}_w$.

We can use the Metropolis–Hastings (MH) algorithm to obtain posterior samples for σ^2 and h^2 jointly. Afterwards, we can obtain posterior samples for $\boldsymbol{\alpha}$, β and $\mathbf{g} + \mathbf{e}$ in turn,

$$P(\beta|\mathbf{z}, \boldsymbol{\gamma}, \sigma_g^2, \sigma_e^2) \sim \mathrm{N}((\mathbf{x}^T\mathbf{P}_w\mathbf{x} + \sigma_b^{-2})^{-1}\mathbf{x}^T\mathbf{P}_w(\mathbf{z} - \mathbf{m}_{\boldsymbol{\gamma}}), (\mathbf{x}^T\mathbf{P}_w\mathbf{x} + \sigma_b^{-2})^{-1}), \tag{11}$$

$$P(\boldsymbol{\alpha}|\mathbf{z}, \boldsymbol{\gamma}, \beta, \sigma_g^2, \sigma_e^2) \sim \mathrm{MVN}((\mathbf{W}^T\mathbf{H}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{H}^{-1}(\mathbf{z} - \mathbf{m}_{\boldsymbol{\gamma}} - \mathbf{x}\beta), (\mathbf{W}^T\mathbf{H}^{-1}\mathbf{W})^{-1}), \tag{12}$$

$$P(\mathbf{g} + \mathbf{e}|\mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \beta, \sigma^2, h^2) \sim \mathrm{MVN}(\sigma^2\mathbf{V}\mathbf{H}^{-1}(\mathbf{z} - \mathbf{m}_{\boldsymbol{\gamma}} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{x}\beta), \sigma^2\mathbf{V}\mathbf{H}^{-1}\mathbf{D}_\mathbf{r}). \tag{13}$$

Finally, conditional on y_i and λ_i, the posterior of z_i is easy to sample. By using the relationship between the gamma distribution and the exponential distribution, we have

$$z_i|y_i, \lambda_i \sim \mathrm{Ga}(r_i, 1 + \lambda_i) + \mathrm{Ga}(y_i, \lambda_i). \tag{14}$$

The most computationally expensive part of the algorithm is the MH step: a naive approach to evaluate $P(\sigma^2, h^2|z_i, \gamma_i)$ would involve cubic operations. Our mean normal mixture approximation allows us to evaluate this marginal likelihood efficiently as we can apply here the mixed model innovations developed recently [23–26]. This is because given the observed data, $\mathbf{D_r}$ is a fixed diagonal matrix where the elements do not depend on a $\boldsymbol{\gamma}$ that changes in every MCMC iteration. Therefore, for a given matrix $\mathbf{V}$, we can perform an eigen-decomposition on $\mathbf{D_r}^{-\frac{1}{2}}\mathbf{V}\mathbf{D_r}^{-\frac{1}{2}} = \mathbf{UDU}^T$. This allows us to decompose $\mathbf{H} = \sigma^2\mathbf{V} + \mathbf{D_r} = \mathbf{D_r^{\frac{1}{2}}}\mathbf{U}(\sigma^2\mathbf{D} + \mathbf{I})\mathbf{U}^T\mathbf{D_r^{\frac{1}{2}}}$. Afterwards, we can transform the latent variables and other covariates to obtain $\mathbf{D_r^{\frac{1}{2}}}\mathbf{U}(\mathbf{z} - \mathbf{m}_{\boldsymbol{\gamma}})$, $\mathbf{D_r^{\frac{1}{2}}}\mathbf{UW}$ and $\mathbf{D_r^{\frac{1}{2}}}\mathbf{Ux}$. This procedure avoids any cubic operations later on in the MCMC steps. Therefore, with the mean normal mixture approximation, we only need to perform eigen-decompositions at the beginning of the MCMC. Afterwards, each Gibbs step only requires quadratic operations (transformation of $\mathbf{z} - \mathbf{m}_{\boldsymbol{\gamma}}$). In practice, because $\mathbf{V}$ is a function of h^2, we assign a discrete uniform prior for h^2 and evaluate the eigen-decompositions on every discrete values of h^2. In the present study, we found that using either 10 or 100 discrete values of h^2 yields almost identical results (and we present the analyses results for the formal in the main text), suggesting that a fine grid for h^2 is not necessary because of our small sample size. Finally, for all analyses in the present study, we ran 1100 Gibbs sampling iterations with the first 100 as burn-in. In each Gibbs iteration, after sampling the latent variables $\mathbf{z}$ and the latent indicators $\boldsymbol{\gamma}$, we further ran 10 MH steps before continuing the Gibbs iterations.

4 Parameter Estimation and p Value Computation

Denote $\bar{\beta}$ as the posterior mean and σ^2_β as the posterior variance. Since both the likelihood and the posterior follow normal distributions asymptotically, and because we also use a normal distribution as the prior distribution, we can easily obtain the approximate maximum likelihood estimate and its standard error by the method of moments, or

$$\hat{\beta} = \sigma_b^2\bar{\beta}/(\sigma_b^2 - \sigma_\beta^2), \tag{15}$$

$$se(\hat{\beta}) = \sigma_b\sigma_\beta/\sqrt{\sigma_b^2 - \sigma_\beta^2}. \tag{16}$$

The condition $\sigma_b^2 > \sigma_\beta^2$ is guaranteed by asymptotics. In rare cases, however, this condition may not be satisfied because of the limited MCMC sampling iterations in practice. This may be particularly concerning for sites where the likelihood function is not informative. Arguably, these non-informative sites are the ones that we do not want to perform analysis on in the first place. Therefore, this condition gives us a natural way to perform post-filtering. In the software implementation, we do not analyze sites where $\sigma_\beta^2 \geq c\sigma_b^2$ for a user defined threshold c $(c \leq 1)$. We use $c = 0.95$ throughout the present study. This post-filtering step, however, has minimal influence on the results, as only a few dozen sites, out of half a million, are filtered out in each analysis.

References

1. Zhou X, Carbonetto P, Stephens M (2013) Polygenic modelling with Bayesian sparse linear mixed models. PLoS Genetics 9: e1003264.

2. McCulloch CE, Searle SR, Neuhaus JM (2008) Generalized, Linear, and Mixed Models. New York, NY, USA: Wiley-Interscience.

3. Pinheiro JC, Chao EC (2006) Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics 15: 58-81.

4. Raudenbush SW, Yang ML, Yosef M (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics 9: 141-157.

5. Goldstein H (1991) Nonlinear multilevel models with an application to discrete response data. Biometrika 78: 45-51.

6. Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88: 9-25.

7. Breslow NE, Lin X (1995) Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika 82: 81-91.

8. Lin X, Breslow NE (1996) Bias correction in generalised linear mixed models with multiple components of dispersion. Journal of the American Statistical Association 91: 1007-1016.

9. Goldstein H, Rasbash J (1996) Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society Series A 159: 505-513.

10. Rodriguez G, Goldman N (2001) Improved estimation procedures for multilevel models with binary response: A case-study. Journal of the Royal Statistical Society Series A 164: 339-355.

11. Browne WJ, Draper D (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis 3: 473-514.

12. Jang W, Lim J (2009) A numerical study of PQL estimation biases in generalized linear mixed models under heterogeneity of random effects. Communications in Statistics - Simulation and Computation 38: 692-702.

13. Fong Y, Rue H, Wakefield J (2010) Bayesian inference for generalized linear mixed models. Biostatistics 11: 397-412.

14. Tuerlinckx F, Rijmen F, Verbeke G, Boeck PD (2006) Statistical inference in generalized linear mixed models: A review. British Journal of Mathematical and Statistical Psychology 59: 225-255.

15. Zeger SL, Karim MR (1991) Generalized linear models with random eects: A Gibbs sampling approach. Journal of the American Statistical Association 86: 79-86.

16. Karim MR, Zeger SL (1992) Generalized linear models with random effects: salamander mating revisited. Biometrics 48: 631-644.

17. Clayton DG (1996) Generalized linear mixed models. In: Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (editors), Markov Chain Monte Carlo in Practice. London, UK: Chapman and Hall.

18. Gamerman D (1997) Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing 7: 57-68.

19. Scott SL (2011) Data augmentation, frequentistic estimation, and the Bayesian analysis of multinomial logit models. Statistical Papers 52: 87-109.

20. Fruhwirth-Schnatter S, Fruhwirth R (2007) Auxiliary mixture sampling with applications to logistic models. Computational Statistics and Data Analysis 51: 3509-3528.

21. Fruhwirth-Schnatter S, Fruhwirth R, Held L, Rue H (2009) Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Statistics and Computing 19: 479-492.

22. McFadden D (1974) Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. (Ed.), Frontiers of Econometrics. New York, NY, USA: Academic Press.

23. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

24. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, et al. (2011) FaST linear mixed models for genome-wide association studies. Nature Methods 8: 833-835.

25. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 44: 821-824.

26. Pirinen M, Donnelly P, Spencer CCA (2013) Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Annals of Applied Statistics 7: 369-390.

27. Schwartz L (1965) On Bayes procedures. Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete 4: 10-26.