HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2017 September 30.

Published in final edited form as:

Biometrics. 2017 September ; 73(3): 1018–1028. doi:10.1111/biom.12649.

Bayesian Genome- and Epigenome-wide Association Studies with Gene Level Dependence

E.F. Lock[*] and

Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, U.S.A

D.B. Dunson[*]

Department of Statistical Science, Duke University, Durham, NC 27708, U.S.A

Summary

High-throughput genetic and epigenetic data are often screened for associations with an observed phenotype. For example, one may wish to test hundreds of thousands of genetic variants, or DNA methylation sites, for an association with disease status. These genomic variables can naturally be grouped by the gene they encode, among other criteria. However, standard practice in such applications is independent screening with a universal correction for multiplicity. We propose a Bayesian approach in which the prior probability of an association for a given genomic variable depends on its gene, and the gene-specific probabilities are modeled nonparametrically. This hierarchical model allows for appropriate gene and genome-wide multiplicity adjustments, and can be incorporated into a variety of Bayesian association screening methodologies with negligible increase in computational complexity. We describe an application to screening for differences in DNA methylation between lower grade glioma and glioblastoma multiforme tumor samples from The Cancer Genome Atlas. Software is available via the package BayesianScreening for R: github.com/lockEF/BayesianScreening.

Keywords

Bayesian; DNA methylation; Genetic association study; Multiple testing; Nonparametric Bayes

1. Introduction

Several technologies that are used for genomic research measure data that are high-throughput and genome-wide. These data may be genetic or epigenetic. Technologies that measure genetic data include single nucleotide polymorphism (SNP) arrays, whole-exome sequencing, and whole-genome sequencing; technologies that measure epigenetic data include DNA methylation bisulphite arrays or bisulphite sequencing, and chromatin immunoprecipitation sequencing (ChIP-seq). These technologies all measure hundreds of

[*]elock@umn.edu

[*]dunson@duke.edu

7. Supplementary Materials

Web Appendix A referenced in Sections 2 and 5, Web Appendices B, C, and D referenced in Section 4 and an Excel spreadsheet with results for additional simulations referenced in Section 4, are available with this paper at the Biometrics website on Wiley Online Library..

thousands of variables, each of which can be mapped to a location on the genome. In this article we use the general term “marker” to refer to any such variable.

A recurring objective in genomic research is to test each marker for an association with a given phenotypic trait, such as disease status. These are commonly conducted in a frequentist framework, where a p-value for the null hypothesis of no association is calculated independently for each marker. Several thousand such studies have been conducted for genetic associations alone (Welter et al., 2014). While these studies have revealed several important biomarkers, they have also been criticized for lack of power and lack of reproducibility (Visscher et al., 2012). The reliance on p-values and binary conclusions may be partly responsible for these criticisms. P-values are a poor proxy for our degree of confidence that a true association exists, because they depend on the power of the test (Stephens and Balding, 2009). Furthermore, standard corrections for multiple comparisons that control the family-wise error rate or false discovery rate for a single study typically require exorbitant effect sizes, leaving most associated markers undetected (Park et al., 2010).

As an alternative to frequentist-based approaches, several methodologies have been developed to screen for genome-wide associations in a fully Bayesian framework (for a review see Stephens and Balding (2009)), and these are increasingly used in practice. Bayesian approaches directly compute the posterior probability that a marker is associated with a given trait, under a full probabilistic model for both the null and alternative hypotheses. This provides a natural framework for meta-analyses that combine results from multiple studies (Verzilli et al., 2008; Wen et al., 2014), and for borrowing information across multiple related markers within a single study to compute more well-informed and accurate weights of evidence in the form of posterior probabilities. Bayesian techniques that combine multiple related tests need not treat the null and alternative hypotheses asymmetrically; this is in contrast to frequentist approaches to the multiple comparisons problem that typically require larger effect sizes for the alternative as the number of tests grows.

Despite their potential flexibility for borrowing information, standard practice for Bayesian genome-wide testing is to screen each marker independently. This involves specifying the prior odds for association, which is multiplied by the Bayes factor at each marker (Stephens and Balding, 2009; Wakefield, 2009; Xu et al., 2012). Alternatively, the prior probability of association at each marker can be treated as unknown (with, for example, a $Beta(a, b)$ prior distribution) and inferred during posterior computation (Scott et al., 2010; Lock and Dunson, 2015). However, this still relies on the over-simplified premise that the probability of association is the same for all markers.

We propose a computationally scalable and widely applicable approach to inferring null probabilities that depend on the genomic location of each marker. Specifically, we describe an approach in which the prior probability of association for a given marker depends on the gene it encodes. The gene-specific probabilities are modeled with a nonparametric distribution that allows for appropriate genome-wide adjustments for multiplicity. We

demonstrate how this approach can dramatically improve posterior accuracy and interpretation when there is gene-level dependence among tests.

We apply our approach to an epigenome-wide association study of cancerous brain tumors that develop from astrocyte cells. We use DNA methylation data from the Illumina HumanMethylation450 array to compare methylation profiles between lower-grade astrocytoma and glioblastoma multiforme samples from The Cancer Genome Atlas. These data include methylation measurements at 294, 093 genomic sites that map to 24, 358 different genes. We apply our gene-level dependence model in conjunction with a previously described method for screening for differential distribution between groups in methylation array data based on shared kernels (Lock and Dunson, 2015). Our analysis reveals systematic differences in methylation distribution at a large number of genomic sites, and the proportion of sites with differential methylation varies substantially between genes.

1.1 Gene-wise Association Tests

Many methods have been developed that combine multiple markers within a single gene to test for an association at the gene level. For example, there is a rich literature on methods that aggregate genetic variants within a gene, via a direct sum or a regression model, to obtain a p-value for the null hypothesis that the gene has no association with the given phenotype (Pan et al., 2014; Wu et al., 2011; Liu et al., 2010). Similarly, there are methods that combine methylation markers within a given gene (or region) to obtain a composite p-value (Wang et al., 2012). These methods can substantially increase power, as many markers within a gene have a weak association that cannot be detected independently (Wojcik et al., 2015), and also reduce the number of overall tests for multiplicity correction. However, aggregating at the gene level may miss important marker-specific effects; for example, different mutations within the same gene can have very different phenotypic consequences (Rowntree and Harris, 2003).

In a Bayesian framework, Wilson et al. (2010) describe a genome-wide model for the association of genetic markers with an observed phenotype, in which the Bayes factor for model inclusion can be computed at the marker or gene level. In their implementation each marker has the same prior probability of association p genome-wide, with hyperprior $p \sim Beta(a, b)$; a gene is considered associated with the observed phenotype if any marker within the gene is associated. Alternatively, Ruklisa et al. (2015) describe a class of Bayesian approaches to rare variant association testing in which the prior probability that a given marker is associated depends on the gene it encodes. For their approach the gene-specific probabilities are estimated independently based on training data, with no borrowing of information across the genes. Nevertheless, they illustrate that gene-specific probabilities outperform genome-wide approaches. Our proposed approach is a flexible compromise between genome-wide and gene-specific priors for marker associations.

2. Model

Here we describe our hierarchical model for gene-specific probabilities in general, to convey its applicability to a wide variety of data types and Bayesian models for association. Data are collected for M genetic or epigenetic markers from N individuals, where each marker maps

to one of G genes. Let M_g be the number of markers that map to gene g, so that $M=\sum_{g=1}^{G} M_g$. Let X_{gmn} denote data for marker m in gene g ($m \in 1, \ldots, M_g$) for individual n, and let Y_n define a phenotypic response for individual n. Let $H_{0,gm}$ define a probabilistic model of no association with Y for marker m in gene g, and $H_{a,gm}$ define the alternative model of association. This framework is illustrated in Example 2.1.

Example 2.1—Assume markers represent SNPs and X_{gmn} is a binary indicator denoting the presence/absence of a minor allele at SNP m in gene g for sample n, Y_n is a binary response indicating disease status (affected or unaffected), and $\lambda_{a,gm}$ and $\lambda_{u,gm}$ represent the rate of minor allele presence in gene g and marker m among affected and unaffected individuals, respectively. A simple model (Balding, 2006) specifies $H_{0,gm}: \lambda_{a,gm} = \lambda_{u,gm} = \lambda_{gm}$ where λ_{gm} has a Uniform(0, 1) prior, and $H_{a,gm}: \lambda_{a,gm} \neq \lambda_{u,gm}$ where $\lambda_{a,gm}$ and $\lambda_{u,gm}$ have independent Uniform(0, 1) priors. Marginalizing over $\lambda_{a,gm}$ and $\lambda_{u,gm}$, the likelihoods under the null and alternative hypothesis are

$$P(X,Y|H_{0,gm}) = \beta\left(1+s^a_{1,gm}+s^u_{1,gm}, 1+s^a_{0,gm}+s^u_{0,gm}\right)$$

and

$$P(X,Y|H_{a,gm}) = \beta\left(1+s^a_{1,gm}, 1+s^a_{0,gm}\right)\ \beta\left(1+s^u_{1,gm}, 1+s^u_{0,gm}\right),$$

where β is the beta function, $s^a_{0,gm}$ and $s^a_{1,gm}$ give the number of affected individuals without and with the minor allele, respectively, and $s^u_{0,gm}$ and $s^u_{1,gm}$ are defined similarly for unaffected individuals.

The approach that follows is general and may be used regardless of the specific form of the likelihood under $H_{0,gm}$ and $H_{a,gm}$. See Stephens and Balding (2009) for a review of Bayesian models for genetic association studies. Details specific to a methylation screening application with continuous data are given in Section 5 and Web Appendix A. The posterior probability of the null for the given marker is

$$P(H_{0,gm}|X,Y) = \frac{P(H_{0,gm})P(X,Y|H_{0,gm})}{P(H_{0,gm})P(X,Y|H_{0,gm}) + P(H_{a,gm})P(X,Y|H_{a,gm})}.$$

Under our proposed model, prior probabilities are equal within a gene:

$$p_g = P(H_{0,gm}) \text{ for } m = 1, \ldots, M_g.$$

We use a nonparametric hyperprior to infer the gene-level prior probabilities $\{p_g\}_{g=1}^{G}$ and borrow information across the genes. Specifically, the distribution of the $p_g's$ is given a Dirichlet process prior (Ferguson, 1973) with a Beta(a, b) base distribution and concentration parameter α: $p_g \sim P$ where $P \sim$ DP(Beta(a, b), α). Under this framework, each p_g is drawn from a theoretically infinite number of realizations θ_h from Beta(a, b), with corresponding probability weights π_h:

$$p_g = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h},$$

where δ_{θ_h} is a point mass at θ_h. A consequence of this model is clustering of the genes, as values θ_h with larger weights π_h will correspond to the probability for several genes. This clustering property is useful for interpretation (e.g., to identify gene sets) but our primary motivation for using the Dirichlet process is to provide a sufficiently robust and flexible hierarchical distribution for the $p_g's$.

The concentration parameter α controls the dispersion of the weights π_h and, hence, influences the sizes of the gene clusters. As $\alpha \to 0$ a single realization will correspond to all genes (e.g., $p_1 = \ldots = p_G = \theta_1$); hence, the limit is equivalent to a genome-wide correction for multiplicity. As $\alpha \to \infty$ each gene will have its own realization (e.g., $p_1 = \theta_1$, $p_2 = \theta_2$, …); hence, the limit is equivalent to a separate, independently estimated probability for each gene. In practice we find that fixing α as a small positive value, such as $\alpha = 1$, allows for sufficient posterior flexibility between these two extremes.

It is also informative to consider the choice of a, b in the Beta base distribution. In applications where a Beta(a, b) distribution is used for a shared prior probability, fixing $b = 1$ is common (Scott et al., 2010). Choosing $a = \lambda M - 1$ provides a natural multiplicity adjustment, as the expected number of associated markers under the prior model is then $1/\lambda$ regardless of the number of markers M (Wilson et al., 2010). This result extends to our context, as $E(p_g) = a/(a + b)$ and therefore the expected number of associated markers under the prior is

$$\sum_{g=1}^{G} M_g E(p_g) = M \cdot \frac{a}{a+b}.$$

However, philosophically there may be little reason for the probability of association at each marker to be negatively effected by the number of markers measured. In practice we find that a simple uniform base distribution Beta(1, 1), while liberal as an *a priori* model, allows for substantial posterior flexibility and still performs well as a multiplicity correction under a global null. Alternatively, the beta hyperparameters can be determined subjectively, or empirically, from related association studies.

The parameter p_g should not be interpreted as the overall probability of association for gene g. Rather, it can be viewed as the inferred proportion of locations within the gene that are associated. This is one approach to prioritize genes, but more importantly the $p_g's$ can improve the accuracy of posterior inference at the marker level.

3. Inference

Here we describe a general Gibbs sampling scheme to compute the full posterior under the gene-level prior model specified above. This estimation approach is informative, illustrating how the marker parameters, gene parameters, and global parameters relate to each other. Fundamentally, the algorithm proceeds by sampling from the posterior of each marker, then updating the gene-specific probabilities and their corresponding Dirichlet process parameters.

We use the constructive stick-breaking representation of the Dirichlet process (Sethuraman, 1994) to sample from its full conditional distribution. That is, the probability weights $\boldsymbol{\pi}_h$ are generated by $\boldsymbol{\pi}_h = V_h \prod_{l<h}(1 - V_l)$, where $V_h \overset{iid}{\sim} \text{Beta}(1, \alpha)$. In practice we truncate the infinite mixture by a large integer H, and perform blocked Gibbs sampling (Ishwaran and James, 2001). Thus, letting C_g define the cluster index for gene g ($p_g = \theta_{C_g}$), $C_g \in \{1, \ldots, H\}$. The weights $\boldsymbol{\pi}_h$ usually decrease quickly to very small values, and thus the effect of truncation is negligible.

Assuming the marginal likelihoods under the null and alternative models can be computed for each marker, sampling from the full conditionals proceeds as follows:

1. Designate null markers $H_{0,gm}$ for $g = 1, \ldots, G$, $m = 1, \ldots, M_g$. The conditional probability of the null, $P(H_{0,gm} \mid X, Y, p_g)$, is

$$\frac{p_g P(X, Y|H_{0,gm})}{p_g P(X, Y|H_{0,gm}) + (1-p_g) P(X, Y|H_{a,gm})}.$$

2. Allocate indices C_g for $g = 1, \ldots, G$:

$$P\left(C_g = h|\theta_h, \{H_{0,gm}\}_{m=1}^{M_g}\right) \propto \pi_h \theta_h^{S_g} (1-\theta_h)^{M_g - S_g}$$

for $h = 1, \ldots, H$, where S_g is the number of null markers in gene g, $S_g = \sum_{m=1}^{M_g} (H_{0,gm})$.

3. Update the weights $\boldsymbol{\pi}_h$ for $h = 1, \ldots, H$. First, draw the stick-breaking weights $V_1, \ldots, V_{H-1}$. The full conditional distribution of V_h is $\text{Beta}\left(1 + \sum_{g=1}^{G} (C_g = h), \alpha + \sum_{g=1}^{G} (C_g > h)\right)$, with $V_H = 1$. Then set $\boldsymbol{\pi}_h = V_h \prod_{l<h}(1 - V_l)$ for $h = 1, \ldots, H$.

4. Update the atoms θ_h for $h = 1, \ldots, H$. The full conditional distribution of θ_h is Beta $(a + \tilde{S}_h, b + \tilde{M}_h - \tilde{S}_h)$, where $\tilde{M}_h$ is the total number of markers in genes allocated to cluster h, and $\tilde{S}_h$ is the number of null markers:

$$\tilde{M}_h = \sum_{\{g:C_g=h\}} M_g, \quad \tilde{S}_h = \sum_{\{g:C_g=h\}} S_g.$$

Set $p_g = \theta_{C_g}$ for $g = 1, \ldots, G$.

Point estimates for the gene-level probabilities p_g and marker posterior probabilities $P(H_{0,gm} \mid X, Y)$ can be obtained by averaging their draws over the sampling iterations.

For some association models the likelihoods under the null and alternative hypotheses in sampling step (1) may not be feasible to compute directly. In Example 2.1 we integrate over the model parameters $\lambda_{a,gm}$ and $\lambda_{u,gm}$ to obtain the marginal likelihoods under $H_{0,gm}$ and $H_{a,gm}$, but for more complex models this integration may not be analytically tractable. If not, additional sampling steps can be incorporated to update model-specific parameters for each marker under $H_{0,gm}$ and $H_{a,gm}$, and then condition on these parameters in step (1). Such an approach is needed for the association model used in the two-group methylation screening scenario described in Section 5, and the additional sampling steps are given in Web Appendix A.

4. Simulation Study

Here we present a simulation study to illustrate the advantages of our hierarchical model for gene-level probabilities. We compare (1) our hierarchical approach for inferring marginal probabilities of the null at each marker with (2) separate estimation, in which a probability is inferred independently for each gene, and shared by all markers for that gene, (3) joint estimation, in which a probability is inferred globally and shared by all markers, and (4) simple estimation, in which the prior is fixed at 0.5 for all markers.

For simplicity, here we consider the setting of Example 2.1. We simulate data for two groups, each with 80 individuals ($N = 160$). For null markers, binary values are simulated under a common probability for both groups, where this probability is drawn from a uniform distribution. For alternative markers, binary values are simulated under a different probability for each group, where these probabilities are drawn independently from a uniform distribution. The Bayes factor for the null over the alternative for a given marker is then

$$\frac{\beta(1+s_1+s_2, 1+160-s_1-s_2)}{\beta(1+s_1, 1+80-s_1)\beta(1+s_2, 1+80-s_2)},$$

where β defines the beta function, and s_1 and s_2 are the number of individuals for which the marker is present in groups 1 and 2, respectively. This is analogous to the prospective Bayes factor for SNP association testing introduced in Balding (2006). Data are simulated for $G =$

500 genes, where the number of markers within a gene M_g is drawn from {2, 3, …, 20} with equal probability. We again simulate gene-level probabilities for a global null hypothesis, a bimodal scenario where markers in 20% of genes are alternative and the other 80% are null, and where gene-level probabilities are generated from a Beta(1, 0.2) distribution.

We consider three different scenarios with dramatically different assumption on the distribution of null and alternative markers across the genes. For each scenario, we show the inferred distribution of the gene-specific probabilities p_g under the four methods considered. We also compute the expected overall error in classifying null and alternative markers, as the average misclassification probability over all markers.

First, we simulate data where the null is true for all markers, to illustrate how the four methods perform as a multiplicity adjustment. Results are shown in Figure 1A. The simple model with fixed prior probability of 0.5 performs relatively poorly; in this and other simulations the average error in classifying markers independently is approximately 20%. The joint and hierarchical models have negligible error, as they both borrow information globally to enforce appropriately high prior probabilities of the null. The model with separately inferred priors for each gene does not perform as well, as its shift toward the null is relatively weak, especially for those genes with a small number of markers.

Second, we simulate data from a bimodal distribution in which the majority of genes (80%) are null for all markers, but for a subgroup of genes (20%) the alternative is true for all markers. Results are shown in Figure 1B. In this case the hierarchical model performs well, as it identifies both modes and allocates the appropriate genes to each mode. The separate model performs better than the joint model, as the joint model does not account for the heterogeneity in the genes. However, the separate model is not competitive with the hierarchical model, as again the gene-specific probabilities have substantial uncertainty and do not shrink toward the two modes if they are estimated independently.

Third, we simulate the gene-specific probabilities from a Beta(1, 0.2) distribution, which has a majority of its mass near 1 (corresponding to genes in which the vast majority of markers are null) but a long left tail. Results are shown in Figure 1C. The joint and separate models perform similarly, as the joint model ignores the gene heterogeneity and the separate model exaggerates gene heterogeneity. The hierarchical model serves as a flexible compromise between the two extremes, and closely approximates the true gene-specific probabilities.

To compare the Bayesian methods above with frequentist methods for multiple hypothesis testing, we compute a p-value for the null using Fisher’s exact test at each marker. We consider different multiplicity adjustments for these p-values, including (5) separate false-discovery rate (FDR) corrections for each gene, using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995), (6) an overall FDR correction for all markers, and (7) a two-step hierarchical hypothesis testing framework (Li and Ghosh, 2014) that uses the Hochberg (Hochberg, 1988) and Benjamini-Hochberg methods. The latter method controls the *overall FDR* while allowing for dependence within sets of hypotheses. The overall FDR is defined as the rate of hypothesis sets that contain at least one false discovery. In our context, a set corresponds to markers within a gene.

P-values and Bayesian posterior probabilities have fundamental differences in philosophy and interpretation, and are not directly comparable. Nonetheless, for illustration we compare the various approaches above by considering standard thresholds on the posterior probability, p-value or FDR that are used to classify markers as null or alternative. For Bayesian methods we use 0.5 as a threshold on the posterior probability, and for the frequentist methods we use a significance threshold of $\alpha = 0.05$. The resulting misclassification rates under each simulation scenario are shown in Table 1A. Under the null model the hierarchical Bayesian approach gives very low error, similar to overall FDR corrections. Under the two scenarios with alternative markers the Bayesian hierarchical model performs substantially better than frequentist multiplicity corrections, which are overly conservative. For a more direct comparison to frequentist FDR methods we apply the Bayesian conditional FDR (Newton et al., 2004) to posterior probabilities from methods (1)–(4), specifying a desired FDR of 0.05. The results are given in Table 1B; the hierarchical Bayesian approach has the greatest power while maintaining the desired FDR. The two-step FDR tends to be the most conservative, as controlling the marker-wise FDR is often not sufficient to control the overall gene-wise FDR.

Additional simulation details are available in the supplementary material. Web Appendix B gives an ROC curve for the various methods to further assess classification accuracy, Web Appendix C gives results under different hyperparameter choices, and Web Appendix D gives a simulation for continuous data analogous to the application in Section 5. A spreadsheet available online gives results for 99 additional simulated datasets with varying sample size, number of genes, and distribution of gene-level probabilities. These results demonstrate the robustness of the hierarchical model.

5. Application to LGG-GBM Methylation

We implement our hierarchical gene-level prior model in a screen for differences in DNA methylation between lower grade gliomas (LGG) and glioblastoma multiforme (GBM) tumor samples that develop from astrocyte cells in the brain. Methylation is an epigenetic phenomenon that occurs at cytosine-phosphate-guanine (CpG) dinucleotide sites in the genome. Methylation is thought to play a significant role in LGG pathogenesis (TCGA Research Network, 2015) and GBM pathogenesis (TCGA Research Network, 2013), but the differences between the two tumor classes have not been well-characterized on a genome-wide scale. Both tumors are heterogenous and typically fatal, but a more complete understanding of their molecular differences is important, as LGGs often progress to GBMs and GBM patients have a much shorter survival time.

We use data from the Illumina HumanMethylation450 array, for 128 astrocyte derived LGG samples and 130 GBM samples ($N = 258$), from The Cancer Genome Atlas. Measured CpG sites that have any missing data are removed, as are sites that map to intergenic regions. After filtering, 294, 093 CpG sites remain, that map to 24, 358 distinct genes. The number of sites in a gene ranges broadly from 1 to 1017. Methylation measurement at each site are continuous between 0 (no methylation) and 1 (fully methylated across all cells in the tumor sample). The distribution of measurements at each site is commonly multimodal or skewed

and therefore not well characterized by parametric distributions (see, e.g., histograms in Figure 3).

Several computational methods have been developed to screen for differential methylation levels between groups, based on array data (Jaffe et al., 2012; Maksimovic et al., 2015) or sequencing data (Sun et al., 2014; Feng et al., 2014; Wu et al., 2015). However, the focus on differential methylation levels between groups may miss other important differences between group distributions; for example, certain genomic regions have been shown to exhibit more variability in methylation, and hence greater epigenetic instability, among cancer cells than among normal cells (Hansen et al., 2011). Therefore, we use a more flexible test for difference in the distribution of methylation measurements between GBM and LGG patients at each CpG site using a model with shared kernels previously described in Lock and Dunson (2015). The essential details of this model are given in Section 5.1, and its application with hierarchical gene-level priors is described in Section 5.2.

5.1 Shared kernel model for association

The shared kernel testing model is described in detail in Lock and Dunson (2015), where it is implemented on comparatively sparse methylation data (≈ 20, 000 sites) with a global prior and shown to compare favorably to frequentist and Bayesian parametric and non-parametric alternatives. Briefly, the distribution of methylation measurements at each CpG site is modeled as a mixture of normal kernels $F_1, \ldots, F_K$, truncated between 0 and 1, each with a different mean and variance. The number of kernels is determined by out-of-sample cross validation of the log posterior density. For the present application this yields $K = 8$ kernels that span the measurement range from 0 to 1. The kernels are shared across CpG sites, and thus capture shared patterns of multi-modality and skewness that are typical in methylation array data.

Under the null the kernel mixing weights at each site are the same between two groups (e.g., LGG and GBM cancer), and under the alternative they are different. Specifically, let $\Pi_{gm}^{(0)} = (\pi_{gm1}^{(0)}, \ldots, \pi_{gmK}^{(0)})$ be the kernel probability weights that define the generative distribution for gene g and site m for group 0. Let $\Pi_{gm}^{(1)}$ similarly define the kernel probability weights for group 1. Then,

$$X_{gmn} \sim \sum_{k=1}^{K} \pi_{gmk}^{(Y_n)} F_k,$$

where X_{gmn} gives the methylation level for site m in gene g for sample n, and $Y_n \in \{0, 1\}$ indicates the group membership of sample n. for sample n in group 0, Let $\Pi_{gm}^{(1)}$ similarly define the kernel probability weights for group 1. Under the null model $H_{0,gm}$, the mixing weights are the same for both groups: $\Pi_{gm}^{(0)} = \Pi_{gm}^{(1)}$. The kernel weights are assumed to be generated from a Dirichlet(λ) distribution, where λ is a hyper-parameter that is inferred

during the kernel estimation stage and fixed. Under $H_{a,gm}$, $\Pi_{gm}^{(0)}$ and $\Pi_{gm}^{(1)}$ are considered independent realizations from Dirichlet(λ).

This provides a robust and consistent framework for testing differential distribution, and can identify important differences that are not captured by simply comparing mean methylation levels. Furthermore, the method facilitates interpretation by modeling the full distribution, with uncertainty, for each class.

5.2 Results

We incorporate hierarchical gene-level priors within the shared kernel testing model, and compute the full posterior via Gibbs sampling. The full marginal likelihood under $H_{0,gm}$ and $H_{a,gm}$ is not analytically tractable, and therefore the general approach of Section 3 must be extended to sample other model parameters. Details regarding posterior computation are given in Web Appendix A.

The estimated gene-level probabilities are shown in Figure 2. Their distribution resembles that in the simulation shown in Figure 1C. For the majority of genes the distribution between the two groups is inferred to be equal at most sites ($p_g \approx 1$). However, there is a substantial left tail, corresponding to genes in which a large number of sites are inferred to differ between the two groups. For illustration we focus on one such gene, BST2, which has 9 measured CpG sites and an estimated gene-level probability of $p_g = 0.198$. We select BST2 because it has been considered as a tangible target for immunotherapy in the treatment of GBM (Etcheverry et al., 2010), an independent comparison of GBM and normal samples found differences in BST2 methylation that correlate with gene expression (Wainwright et al., 2011), and BST2 methylation may play a role in the pathonogenesis of other cancers (Mahauad-Fernandez et al., 2014). Figure 3 shows the genomic location and posterior probability of group equality for the nine CpG sites in BST2, as well as group histograms and posterior densities for methylation at three sites.

Given the large number of CpGs and corresponding genes with differential methylation distribution, we also investigate differences at a macro level. Figure 4 shows the site means and standard deviations within each group, for those sites with a posterior probability of a difference greater than 0.01 (24.6% of all CpGs). Mean methylation levels at these sites are generally greater in the LGG samples than the GBM samples; this is concordant with findings in a smaller comparison of 1536 CpG sites in 807 genes (Laffaire et al., 2011). The distribution of standard deviations is more curious, as LGG samples show a larger number of sites with either very high variability or very low variability in comparison to the distribution for GBM.

5.3 Validation

To assess the appropriateness of the hierarchical gene-level model, we consider the agreement of estimated gene-level probabilities and marker-level posteriors under cross validation. Specifically, we randomly select 10, 000 CpGs to leave out, and compute gene-level probabilities using the remaining 284, 093 CpGs. For each left out CpG, we measure the Kullback-Leibler divergence of its estimated gene-level probability under the reduced

data from its CpG-level posterior probability under the full data. This can be interpreted as the gain of information or degree of “surprise” between a CpG’s posterior probability and its gene-level prior (Lindley, 1956). We repeat this process using a separately estimated prior probability for each gene, a single inferred prior probability, and a prior probability of 0.5 (corresponding to the separate, joint and simple models in the Simulation Study). The hierarchical model yields the greatest agreement, with a mean Kullback-Leibler divergence of 0.451; the separate model has a divergence of 0.482, the joint model 0.560, and the simple model 0.654.

We also conduct two permutation studies, to further assess the appropriateness and flexibility of our gene-level model. First, we randomly permute the gene labels for each marker, so that there is no true gene-level dependence. The subsequent posterior means for the gene level prior probabilities p_g are shown in the top-left panel of Figure 5. The estimates converge appropriately to a single global probability near 0.72, in sharp contrast to the relatively dispersed estimates using the true data in Figure 2 of the main article. Second, we randomly permute the class labels but maintain the true gene labels, to generate a dataset with a global null but gene-level dependence. The subsequent posterior estimates are shown in the top-right panel of Figure 5, and cluster very close to 1. In fact, all 294, 093 estimated site-specific posterior probabilities of the null are greater than 0.5. Together, these results demonstrate that the hierarchical gene-level model appropriately shrinks gene-level priors toward a global pattern.

The estimated hierarchical gene-level probabilities closely approximate a single joint prior probability in Figure 5, where a joint prior is appropriate, but not for the true data. Thus, we conclude that a single joint model is an over-simplification for these data. We also compare the hierarchical gene-level prior probabilities under permutation with separate, independently estimated gene-level probabilities. The separately estimated probabilities are shown in the bottom row of Figure 5. These have a lot of variability under both permutation scenarios, illustrating how independent consideration of the genes sacrifices accuracy by exaggerating gene effects.

6. Discussion

Borrowing information and incorporating prior knowledge in a principled and computationally feasible way is an important challenge for Bayesian genome- and epigenome-wide screening methods. Here we present a flexible and generally applicable hierarchical model for inferring gene-specific probabilities, which may be extended in several ways. Under our prior all markers within a gene have an equivalent probability of association. The incorporation of other marker-level information, such as gene promoter status for DNA methylation (Weber et al., 2005) or functional annotation (e.g., synonymous vs. non-synonymous) for genotype markers (Kichaev et al., 2014; Ruklisa et al., 2015), may improve posterior precision and interpretation. For example, a Dirichlet process model may be used for gene-level intercepts within a regression model for marker probabilities that includes additional prior covariates (Lewinger et al., 2007). Incorporating additional gene-level prior information, such as allowing greater dependence within known functional gene networks (Zhang et al., 2014), is also a promising direction of future work.

Our focus is on association testing, illustrated with a simple likelihood model for binary markers and a shared kernel model for DNA methylation data. Other models for association may be used within the same framework, such as a regression approach that accounts for population stratification or other potential confounders. Moreover, markers that are statistically associated with a given phenotype may not affect the phenotype directly, especially if markers are correlated (e.g., linkage disequilibrium in genetic data). Our gene-level model and other prior information can also be used in the context of model inclusion probabilities within a multi-marker regression approach (Wilson et al., 2010; Zhang et al., 2014; Duan and Thomas, 2013), to select markers that have novel predictive power for a given phenotype and are therefore more likely to be causal. We have described a generally applicable approach to posterior computation in which sampling from the gene-level prior is incorporated into the sampling scheme for the specified association model. Computational scalability depends on the association model used. For example, in the methylation application of Section 5 less than 1% of computing time is spent on the draws for the gene-level prior parameters. Gibbs sampling for a high-dimensional multi-marker regression approach can be computationally challenging because the dependence of the markers results in slow mixing; alternatively, markers identified via association testing may subsequently be included as phenotype predictors in a second stage model (Yazdani and Dunson, 2015).

We consider hypotheses that are grouped by markers within a gene, but there are similar scenarios in other areas of genomics research. For example, when screening multiple genes for a phenotypic association (e.g., via microarray or RNA-seq data) the genes can be partitioned into groups based on pathways or other prior information. Frequentist methods that provide appropriate type I error control over genes and gene sets have been developed (Benjamini and Heller, 2008; Heller et al., 2009; Li and Ghosh, 2014), and these methods can be generalized to other problems that involve testing hypotheses over multiple sets. Broadly, our proposed model defines a general prior for multiple hypothesis testing within a Bayesian framework when the hypotheses can be partitioned into sets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Allison Ashley-Koch and Dr. Sandeep Dave for fruitful scientific discussions that motivated this work, and we thank the AE and three anonymous reviewers for helpful suggestions. This work was supported by the National Institute of Environmental Health Sciences (NIEHS) [R01-ES017436] and National Institutes of Health National Center for Advancing Translational Sciences (NIH/NCATS) [ULI RR033183 & KL2 RR0333182].

References

Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews Genetics. 2006; 7:781–791.

Benjamini Y, Heller R. Screening for partial conjunction hypotheses. Biometrics. 2008; 64:1215–1222. [PubMed: 18261164]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B. 1995; 57:289–300.

Duan L, Thomas DC. A Bayesian hierarchical model for relating multiple snps within multiple genes to disease risk. International Journal of Genomics. 2013; 2013

Etcheverry A, Aubry M, De Tayrac M, Vauleon E, Boniface R, Guenot F, Saikali S, Hamlat A, Riffaud L, Menei P, et al. DNA methylation in glioblastoma: impact on gene expression and clinical outcome. BMC Genomics. 2010; 11:701. [PubMed: 21156036]

Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Research. 2014; 42:e69. [PubMed: 24561809]

Ferguson TS. A Bayesian analysis of some nonparametric problems. The Annals of Statistics. 1973; 1:209–230.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, et al. Increased methylation variation in epigenetic domains across cancer types. Nature Genetics. 2011; 43:768–775. [PubMed: 21706001]

Heller R, Manduchi E, Grant GR, Ewens WJ. A flexible two-stage procedure for identifying gene sets that are differentially expressed. Bioinformatics. 2009; 25:1019–1025. [PubMed: 19213738]

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988; 75:800–802.

Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association. 2001; 96:161–173.

Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International Journal of Epidemiology. 2012; 41:200–209. [PubMed: 22422453]

Kichaev G, Yang W, Lindstrom S, Hormozdiari F, Eskin E, Price A, Kraft P, Pasaniuc B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genetics. 2014; 10:e1004722. [PubMed: 25357204]

Laffaire J, Everhard S, Idbaih A, Criniere E, Marie Y, de Reynies A, Schiappa R, Mokhtari K, Hoang-Xuan K, Sanson M, et al. Methylation profiling identifies 2 groups of gliomas according to their tumorigenesis. Neuro-Oncology. 2011; 13:84–98. [PubMed: 20926426]

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genetic Epidemiology. 2007; 31:871–882. [PubMed: 17654612]

Li Y, Ghosh D. A two-step hierarchical hypothesis set testing framework, with applications to gene expression data on ordered categories. BMC Bioinformatics. 2014; 15:108. [PubMed: 24731138]

Lindley DV. On a measure of the information provided by an experiment. The Annals of Mathematical Statistics. 1956:986–1005.

Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, et al. A versatile gene-based test for genome-wide association studies. The American Journal of Human Genetics. 2010; 87:139–145. [PubMed: 20598278]

Lock EF, Dunson DB. Shared kernel Bayesian screening. Biometrika. 2015; 102:829–842. [PubMed: 27046939]

Mahauad-Fernandez WD, Borcherding NC, Zhang W, Okeoma CM. Bone marrow stromal antigen 2 (BST-2) DNA is demethylated in breast tumors and breast cancer cells. PloS One. 2014; 10:e0123931.

Maksimovic J, Gagnon-Bartsch JA, Speed TP, Oshlack A. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. Nucleic acids research. 2015; 43:e106–e106. [PubMed: 25990733]

Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5:155–176. [PubMed: 15054023]

Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014; 197:1081–1095. [PubMed: 24831820]

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics. 2010; 42:570–575. [PubMed: 20562874]

Rowntree RK, Harris A. The phenotypic consequences of cftr mutations. Annals of Human Genetics. 2003; 67:471–485. [PubMed: 12940920]

Ruklisa D, Ware JS, Walsh R, Balding DJ, Cook SA. Bayesian models for syndrome-and gene-specific probabilities of novel variant pathogenicity. Genome Medicine. 2015; 7:120. [PubMed: 26589591]

Scott JG, Berger JO, et al. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics. 2010; 38:2587–2619.

Sethuraman J. A constructive definition of Dirichlet priors. Statistica Sinica. 1994; 4:639–650.

Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009; 10:681–690.

Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, Li W. MOABS: model based analysis of bisulfite sequencing data. Genome Biology. 2014; 15:38.

TCGA Research Network. The somatic genomic landscape of glioblastoma. Cell. 2013; 155:462–477. [PubMed: 24120142]

TCGA Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New England Journal of Medicine. 2015; 372:2481–2498. [PubMed: 26061751]

Verzilli C, Shah T, Casas JP, Chapman J, Sandhu M, Debenham SL, Boekholdt MS, Khaw KT, Wareham NJ, Judson R, et al. Bayesian meta-analysis of genetic association studies with different sets of markers. The American Journal of Human Genetics. 2008; 82:859–872. [PubMed: 18394581]

Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. The American Journal of Human Genetics. 2012; 90:7–24. [PubMed: 22243964]

Wainwright DA, Balyasnikova IV, Han Y, Lesniak MS. The expression of BST2 in human and experimental mouse brain tumors. Experimental and Molecular Pathology. 2011; 91:440–446. [PubMed: 21565182]

Wakefield J. Bayes factors for genome-wide association studies: comparison with p-values. Genetic Epidemiology. 2009; 33:79–86. [PubMed: 18642345]

Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S. IMA: an R package for high-throughput analysis of illumina's 450k infinium methylation data. Bioinformatics. 2012; 28:729–730. [PubMed: 22253290]

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schuebeler D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature Genetics. 2005; 37:853–862. [PubMed: 16007088]

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. The nhgri GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014; 42:D1001–D1006. [PubMed: 24316577]

Wen X, Stephens M, et al. Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions. The Annals of Applied Statistics. 2014; 8:176–203. [PubMed: 26413181]

Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM. Bayesian model search and multilevel inference for SNP association studies. The Annals of Applied Statistics. 2010; 4:1342–1364. [PubMed: 21179394]

Wojcik GL, Kao WL, Duggal P. Relative performance of gene-and pathway-level methods as secondary analyses for genome-wide association studies. BMC Genetics. 2015; 16:34. [PubMed: 25887572]

Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Research. 2015; 43:e141. [PubMed: 26184873]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics. 2011; 89:82–93. [PubMed: 21737059]

Xu J, Yuan A, Zheng G. Bayes factor based on the trend test incorporating hardy–weinberg disequilibrium: more power to detect genetic association. Annals of Human Genetics. 2012; 76:301–311. [PubMed: 22607017]

Yazdani A, Dunson DB. A hybrid Bayesian approach for genome-wide association studies on related individuals. Bioinformatics. 2015; 31:3890–3896. [PubMed: 26323717]

Zhang X, Xue F, Liu H, Zhu D, Peng B, Wiemels JL, Yang X. Integrative Bayesian variable selection with gene-based informative priors for genome-wide association studies. BMC Genetics. 2014; 15:130. [PubMed: 25491445]

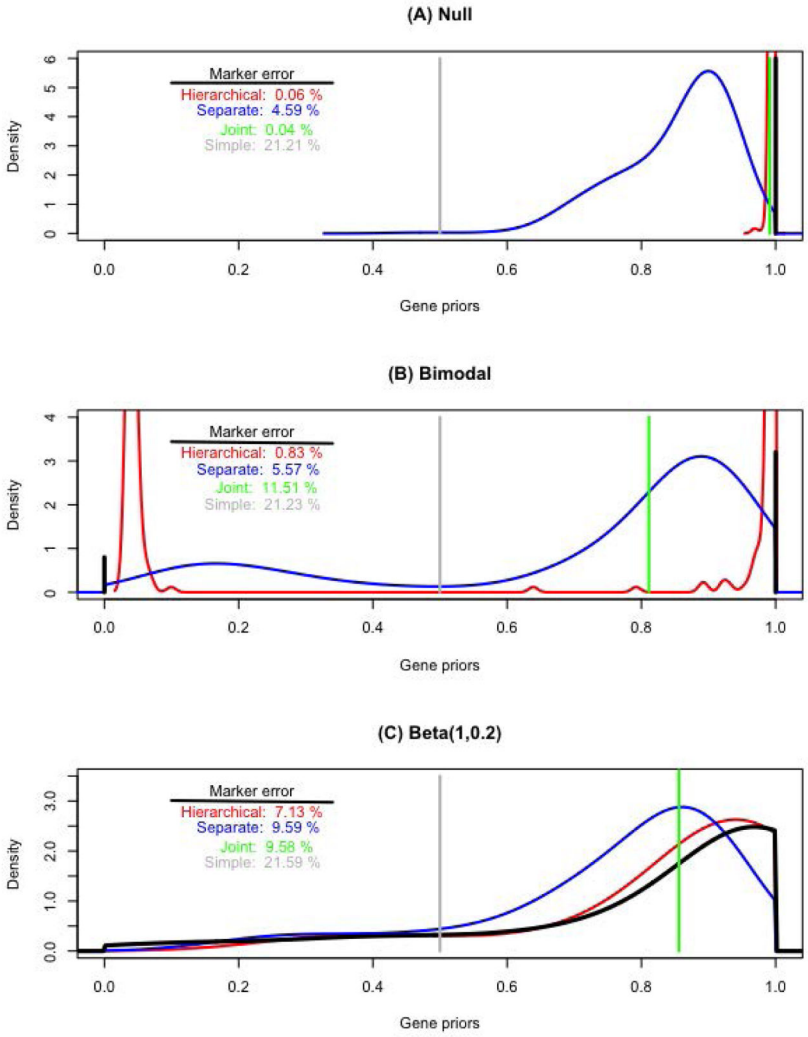

Figure 1.

Comparison of four approaches to inferring prior probabilities, under three simulation scenarios (A,B, and C). Kernel density estimates of the resulting gene-specific probabilities are shown for continuous distributions; discrete distributions are shown by vertical lines. The distribution of the true gene-specific probabilities is colored black. The expected overall error in classifying null and alternative markers is also shown for each method and each simulation scenario.

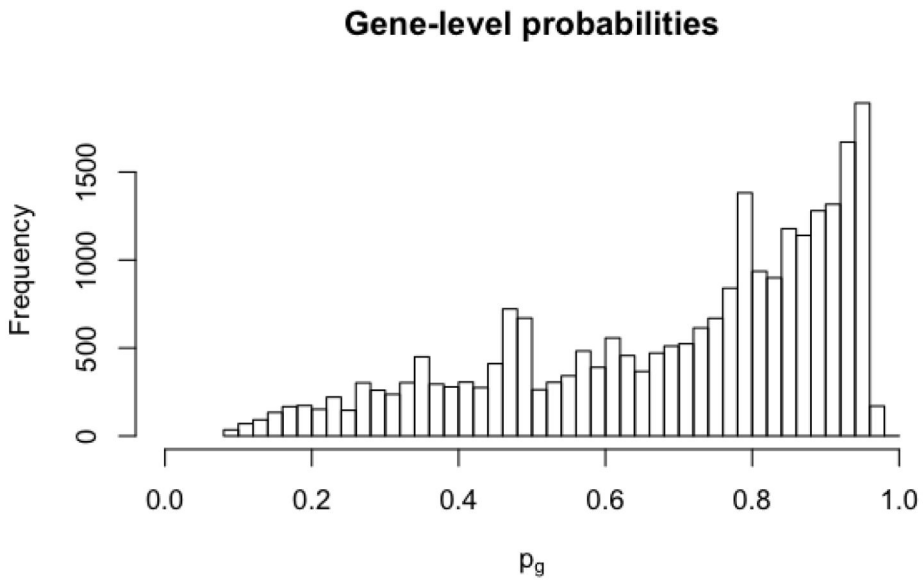

Figure 2.
Histogram of estimated gene-specific probabilities for no association, p_g.

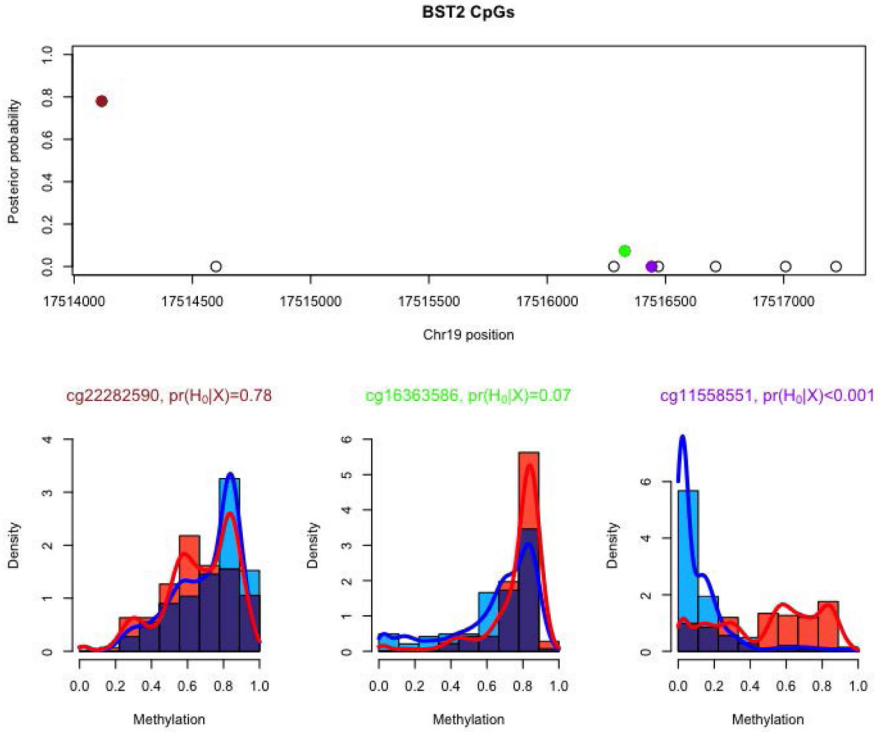

Figure 3.

The top panel shows the estimated posterior probability of an association with GBM-LGG status for the 9 CpG sites measured in the gene BST2, with their corresponding genomic location. The lower panel shows the estimated densities for the GBM (blue) and LGG (red) groups for three sites; histograms of each group are shown, and their overlap is colored purple.

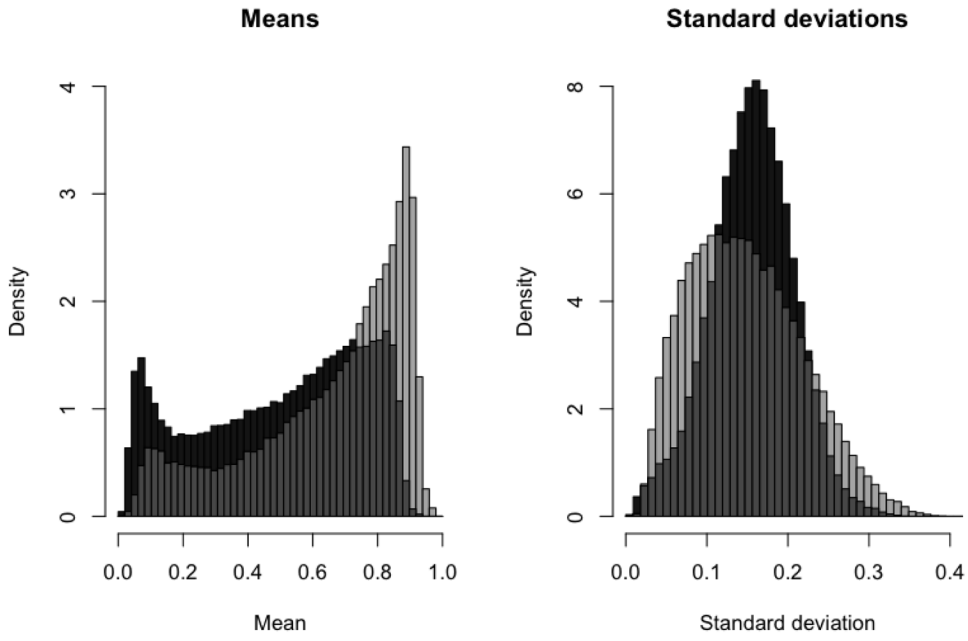

Figure 4.

Histograms of summary statistics for every CpG site with a posterior probability of the null less than 0.01, computed separately for the GBM (dark gray) and LGG (light gray) groups. Overlap between the two histograms is colored a neutral gray. The left panel show the site means, the right shows the site standard deviations.

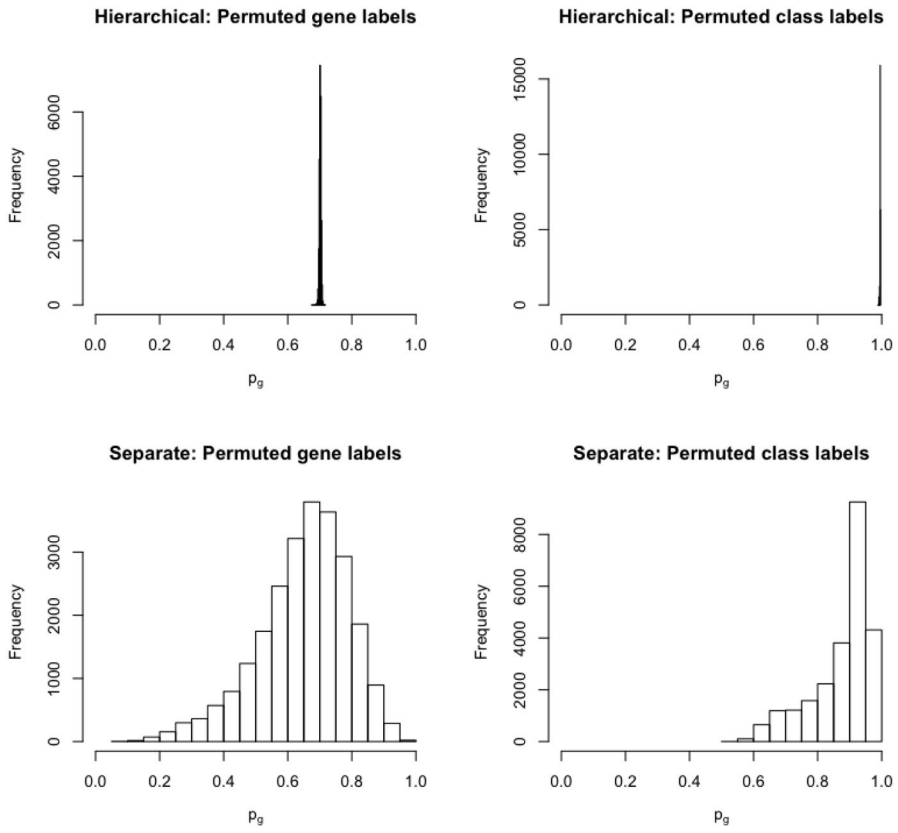

Figure 5.

Histogram of hierarchically estimated and separately estimated gene-level priors after randomly permuting gene labels (left column), and after randomly permuting class labels (right column).

Table 1

A: Error in classifying null and alternative markers, using 0.5 as a threshold for the posterior probability for Bayesian methods, and 0.05 as an FDR or P-value threshold for frequentist methods using Fisher's exact test. The average over 50 replicate simulations is shown, and the resulting standard error is less than 0.05% for all cells. B: false discovery rate and true positive rate (FDR/TPR) using Bayesian or frequentist methods with specified FDR=0.05.

A: Classification	**Null**	**Bimodal**	**Beta (1, 0.2)**
Bayesian			
(1) Hierarchical	0.01%	0.07%	3.92%
(2) Separate	0.50%	1.68%	4.87%
(3) Joint	0.01%	6.89%	5.34%
(4) Simple	5.07%	8.72%	7.82%
Frequentist			
(5) Separate FDR	0.23%	5.64%	5.41%
(6) Overall FDR	0.01%	7.15%	5.78%
(7) Two-step FDR	0.01%	7.73%	6.30%
No correction	2.91%	7.50%	6.49%

B: FDR	Bimodal FDR/TPR	Beta(1,0.2) FDR/TPR
Bayesian		
(1) Hierarchical	3.70%/99.9%	3.73%/76.0%
(2) Separate	1.12%/90.9%	6.72%/76.6%
(3) Joint	5.74%/69.3%	6.34%/67.0%
(4) Simple	13.0%/74.2%	15.6%/73.4%
Frequentist		
(5) Separate FDR	0.89%/73.3%	2.79%/69.2%
(6) Overall FDR	2.36%/65.9%	2.24%/64.0%
(7) Two-step FDR	0.30%/62.3%	0.98%/61.9%