Robust Bayesian FDR Control Using Bayes Factors, with Applications to Multi-tissue eQTL Discovery

Xiaoquan Wen

Abstract Motivated by the genomic application of expression quantitative trait loci (eQTL) mapping, we propose a new procedure to perform simultaneous testing of multiple hypotheses using Bayes factors as input test statistics. One of the most significant features of this method is its robustness in controlling the targeted false discovery rate even under misspecifications of parametric alternative models. Moreover, the proposed procedure is highly computationally efficient, which is ideal for treating both complex system and big data in genomic applications. We discuss the theoretical properties of the new procedure and demonstrate its power and computational efficiency in applications of single-tissue and multi-tissue eQTL mapping.

1 Introduction

With the emergence of high-throughput experimental technologies, false discovery rate (FDR) control [2] has become a standard statistical practice in many genomic applications. Despite the importance of the FDR control procedures, the overall power in a multiple testing problem is largely determined by the test statistics employed. In many application areas, the practitioners are facing increasingly complex systems in which many factors must be accounted for, and their relationships are typically non-trivial. Consequently, it is extremely difficult to describe these systems mathematically without parametric modeling, and it is even more difficult to determine powerful testing statistics. Bayesian models have exhibited some great advantages in such complex settings and have become increasingly popular. Within the Bayesian hierarchical modeling framework, Bayesian FDR control is conceptually straightforward and easy to
implement in a sound decision-theoretic framework, as illustrated by [11,17,18,32]. Nevertheless, its practical usage is often hindered primarily by two factors: first, its lack of robustness due to the heavy reliance on parametric assumptions in alternative model specifications [19] and second, its expensive computational cost.

This paper addresses both issues and proposes a novel Bayesian FDR control procedure motivated by the genomic application of expression quantitative trait loci (eQTL) mapping, where tens of thousands association tests are simultaneously performed. In the context of eQTL mapping, parametric models are particularly preferred because they provide necessary flexibility to (i) control unknown confounding/batch effects; (ii) incorporate prior knowledge of relevant genomic annotations; and (iii) handle complex data structure, e.g., in multi-tissue eQTL mapping. Those parametric models for eQTL mapping often contain hierarchical structures with natural Bayesian interpretations [8]. As a unique feature, our proposed procedure directly works with Bayes factors [14], which are often considered to be the natural test statistics within the Bayesian framework due to the likelihood principle. Most significantly, our proposed procedure controls the target FDR level even when the alternative models are misspecified, which is a property that is currently missing from most available Bayesian and parametric model-based multiple testing approaches.

In theory, it is always possible to convert a Bayes factor to a corresponding $p$ value (by treating the Bayes factor as a regular test statistic) and to apply the $p$ value-based FDR control procedures [2,23–27], a strategy known as the Bayes/Non-Bayes compromise [10]. However, in practice, the conversion often relies on permutation procedures, and their computational cost can be extremely high when treating complex models and/or large data sets, therefore not suitable to apply to the genomic data at the genome-wide scale. In fact, the computational efficiency is a general challenge for any multiple testing procedure that relies on permutation $p$ values. In comparison, for a wide range of parametric models, it has been demonstrated that Bayes factors can be computed or accurately approximated in a highly efficient manner [5,8,12–15,20–22,30,31]. Based on these results, the computational resources required by our proposed procedure for controlling FDR are trivial, which makes it ideal for treating both complex models and large data in genomic applications.

The remainder of the paper is organized as follows. We first provide the necessary background regarding FDR controls within the Bayesian framework in the Sect. 2, and then proceed to describe our novel Bayesian FDR control method in the Sect. 3. In the Sects. 4 and 5, we illustrate the proposed approach in applications of single-tissue and multi-tissue eQTL mapping, respectively. Finally, we summarize our findings and discuss future directions in the Sect. 6.

2 Notation and Setup

Throughout this paper, we consider multiple testing problems under a general hierarchical mixture model setting, which is generally assumed in both Bayesian and frequentist frameworks [7,9,17,18,24]. Consider a set of $m$ hypotheses for simultaneous testing. Let $Y_i$ denote the data collected for the $i$-th test, and let the complete collection of the observed data from all $m$ tests be denoted by $\mathcal{Y} := \{Y_1, \ldots, Y_m\}$.
Suppose that each $Y_i$ is generated from either a null model or some alternative model, namely,

$$Y_i \sim p_i^{Z_i},$$  \hspace{1cm} (2.1)

where $Z_i$ is a latent indicator that denotes the true generating distribution. Note that we do not require all the hypotheses to share the same null or alternative models. The vector of the latent indicators, $\mathbf{Z} := (Z_1, \ldots, Z_m)$, is of primary interest for inference, and the commonality of all the tests lies on the assumption

$$Z_i \mid \pi_0 \overset{i.i.d.}{\sim} \text{Bernoulli} \left(1 - \pi_0\right), \ i = 1, \ldots, m.$$  \hspace{1cm} (2.2)

In particular, the parameter $\pi_0$ represents the proportion of the data generated from the respective null hypothesis and is typically unknown. Finally, we assume that the tests are mutually independent conditional on $\pi_0$, i.e.,

$$\Pr(\mathbf{Z} \mid \mathbf{Y}, \pi_0) = \prod_{i=1}^{m} \Pr(Z_i \mid \mathbf{Y}, \pi_0) = \prod_{i=1}^{m} \Pr(Z_i \mid Y_i, \pi_0).$$  \hspace{1cm} (2.3)

Under the above hierarchical mixture model, the simultaneous hypothesis testing can be framed as a joint decision problem with respect to $\mathbf{Z}$. Let $\delta_i$ denote a decision (0 or 1) on $Z_i$ based on all the observed data, and define $D := \sum_{i=1}^{m} \delta_i$. Following the formulation of [17], we define the false discovery proportion (FDP) as the proportion of false discoveries among total discoveries, i.e.,

$$\text{FDP} := \frac{\sum_{i=1}^{m} \delta_i \left(1 - Z_i\right)}{D \lor 1}.$$  \hspace{1cm} (2.4)

where the conditional expectation is taken with respect to $\mathbf{Z}$, and $v_i := \Pr(Z_i = 1 \mid \mathbf{Y})$. Moreover, the frequentist control of the false discovery rate focuses on the quantity

$$\text{FDR} := \mathbb{E} \left(\text{FDP} \mid \mathbf{Y}\right) = \frac{\sum_{i=1}^{m} \delta_i \left(1 - v_i\right)}{D \lor 1},$$  \hspace{1cm} (2.5)

where the additional expectation is taken with respect to $\mathbf{Y}$ over (hypothetically) repeated experiments. It is important to note that controlling the Bayesian FDR is a sufficient but not necessary condition to control the corresponding frequentist FDR; thus the Bayesian FDR control is more stringent in theory. For measuring type II errors in a multiple testing setting, the false non-discovery proportion (FNP), the Bayesian false non-discovery rate (FNR) and the frequentist FNR are defined in a similar fashion.
Bayesian FDR control procedure is carried out by evaluating the posterior probability $v_i$ for each test, i.e.,

$$v_i = \int \Pr(Z_i = 1 \mid \mathcal{Y}, \pi_0) p(\pi_0 \mid \mathcal{Y}) d \pi_0. \quad (2.6)$$

In particular, the data from all tests jointly affect $p(\pi_0 \mid \mathcal{Y})$ and thereby impact the posterior distribution of each individual $Z_i$. However, conditional on $\pi_0$, $\Pr(Z_i = 1 \mid \mathcal{Y}, \pi_0)$ can be computed independently, i.e.,

$$\Pr(Z_i = 1 \mid \mathcal{Y}, \pi_0) = \Pr(Z_i = 1 \mid Y_i, \pi_0) = \frac{(1 - \pi_0) BF_i}{\pi_0 + (1 - \pi_0) BF_i}, \quad (2.7)$$

where $BF_i$ is the null-based Bayes factor [15] defined by

$$BF_i := \frac{p(Y_i \mid Z_i = 1)}{p(Y_i \mid Z_i = 0)} = \frac{p^1_i(Y_i)}{p^0_i(Y_i)}. \quad (2.8)$$

Particularly, in computing $BF_i$, all the potential nuisance parameters in the parametric models (null or alternative) are integrated out with respect to their prior distributions [14]. As demonstrated by [18] and [17], the Bayesian FDR control is based on the following natural decision rule,

$$\delta^*_i(t) = I (v_i > t), \quad (2.9)$$

i.e., the null hypothesis is rejected if the posterior probability that the observed data is generated from the alternative is high. For a predefined FDR level $\alpha$, the threshold $t_\alpha$ in (2.9) is determined by

$$t_\alpha = \arg \min_t \left( \sum_{i=1}^m \delta^*_i(t)(1 - v_i) \right) \leq \alpha \quad (2.10)$$

Furthermore, [16] demonstrated that the decision rule (2.9) is optimal in the sense that it minimizes the corresponding FNR while controlling for the FDR.

### 3 Robust Control of Bayesian FDR

Although the Bayesian FDR control is conceptually straightforward, its practical performance is susceptible to alternative model misspecifications. In comparison, the $p$ value-based frequentist FDR control procedures demand only adequate behavior of $p$ values under the null models and generally ensure targeted FDR control levels, regardless of the distribution of $p$ values under the assumed alternatives. This robustness property against type I error is often desirable in the hypothesis testing context.

In this section, we provide a detailed discussion of this topic and provide intuitions and theoretical justifications for a new Bayesian FDR control procedure.
should be noted that our notion of “model misspecification” includes the cases of prior misspecification in Bayesian settings.

3.1 General Approach

The specifications of parametric (null and alternative) models impact Bayesian FDR control in the following two aspects.

First, the computation of the Bayes factors relies on the model assumptions of $p_0^i$ and $p_1^i$. In hypothesis testing, the “correctness” of Bayes factors is formally characterized by a concept known as consistency in model comparisons, i.e., with sufficiently large sample size, Bayes factor $\rightarrow 0$ if the data is truly generated from the null model and $\rightarrow \infty$, if the data is generated from the alternative model. Intuitively, to prevent inflation of the predefined FDR level, which concerns the behavior of the testing procedure under the null, the consistency of the Bayes factors is only required under the null models. Such a condition is easy to satisfy (and also easy to check) as long as the null model is correctly specified. In contrast, when the alternative model is misspecified (which is almost certain in practice), the Bayes factor might not be consistent under the alternative. Nonetheless, as we will demonstrate later, it does not inflate the FDR but only affects the power/FNR given the true $\pi_0$. We also note that the assumption regarding the consistency of the Bayes factors under null models is analogous to the uniformity assumption of the null $p$ values, i.e., both are valid only if the null models are correctly specified. In summary, we expect that the null models are correctly specified and allow for imperfect specification of the alternative models. Nevertheless to ensure decent power in testing, it is preferred that the Bayes factor is far > 1 under the alternative scenario and especially in an asymptotic scenario. Importantly, the consistency property of the Bayes factors under the null models implies that the Bayes factors from the alternative models are stochastically dominant over the Bayes factors from the null models.

Second, the inference of $p(\pi_0 \mid \mathcal{Y})$ from the mixture model is sensitive to the alternative model specifications and therefore affects the Bayesian FDR control: if $\pi_0$ is systematically under-estimated, the predefined Bayesian FDR level is inflated; conversely, overestimating $\pi_0$ results in a conservative procedure that loses power but maintains the desired FDR level. To avoid underestimation of $\pi_0$ in all circumstances, we attempt to find a robust upper-bound estimator for $\pi_0$, denoted by $\hat{\pi}_0(\mathcal{Y})$, such that under the sampling distribution of $\mathcal{Y}$ and for all possible values of $\pi_0$,

$$\Pr\left(\hat{\pi}_0(\mathcal{Y}) \geq \pi_0 \mid \pi_0\right) \rightarrow 1,$$

as the number of the true null tests is sufficiently large.

(3.1)

The conditional probability expression emphasizes that the upper bound property of $\hat{\pi}_0(\mathcal{Y})$ holds true for arbitrary $\pi_0$ values. Given (3.1), for arbitrary prior distributions on $\pi_0$, it follows that

$$\Pr(\pi_0 \leq \hat{\pi}_0(\mathcal{Y}) \mid \mathcal{Y}) \rightarrow 1,$$

(3.2)
i.e., \( \hat{\pi}_0 \) forms a probability upper bound of the posterior distribution of \( \pi_0 \), whose corresponding prior distribution can be arbitrary.

The key to our approach is to ensure condition (3.1) even if the alternative models are misspecified. (It should be clear that such an upper bound does exist: for example, the trivial bound \( \hat{\pi}_0(\mathcal{Y}) \equiv 1 \) satisfies this requirement; however, it is too conservative to be useful.) In this paper, we rely on results from asymptotic theories to find the desired \( \hat{\pi}_0 \) estimates while utilizing virtually no assumptions from the alternative models. As a consequence, the resulting \( \hat{\pi}_0 \) may not be optimal in terms of “precision” (with respect to the truth), but it certainly has great appeal in guarding the predefined FDR level against alternative model misspecifications.

Based on the above discussion, we now define a new quantity

\[
\hat{v}_i := \Pr(Z_i = 1 \mid Y_i, \hat{\pi}_0) = \frac{(1 - \hat{\pi}_0)BF_i}{\hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i},
\]

which is consistently more conservative than the underlying true \( v_i \). We propose a conservative Bayesian FDR control procedure that essentially replaces \( v_i \) with \( \hat{v}_i \) in (2.9). Specifically, we denote the new decision rule \( \delta_i^\dagger(t) := I(\hat{v}_i > t) \) and \( D^\dagger(t) := \sum_{i=1}^m \delta_i^\dagger(t) \).

With a stronger assumption that assumes sufficiently large sample size for each test, we are able to show the following theoretical result for the proposed approach:

**Proposition 1** Assuming that the following conditions are satisfied,

1. the Bayes factors are consistent under the true null models.
2. there exists an upper-bound point estimator \( \hat{\pi}_0(\mathcal{Y}) \) such that
   \[
   \Pr(\pi_0 \leq \hat{\pi}_0(\mathcal{Y}) \mid \pi_0) \to 1, \forall \pi_0 \in (0, 1),
   \]
   as the number of the true null tests is sufficiently large.

The decision rule \( \delta_i^\dagger(t) := I(\hat{v}_i > t) \) controls the frequentist FDR at level \( \alpha \) with the rejection threshold

\[
 t_{\alpha}^\dagger = \arg \min_i \left( \frac{\sum_{i=1}^m \delta_i^\dagger(t)(1 - \hat{v}_i)}{D^\dagger(t) \lor 1} \leq \alpha \right),
\]

as the sample size per test and the number of the true null tests are sufficiently large. Furthermore, the Bayesian FDR can be consistently controlled in the sense that

\[
\Pr(\hat{\text{FDR}} \leq \alpha) \to 1.
\]

The proof is based on the above discussion, and the full details are given in the Appendix 1. Proposition 1 specifies two asymptotic conditions: both the number of true null tests and the sample size per test are required to be large. The large number of the true null tests ensures the probabilistic nature of the upper-bound property of \( \hat{\pi}_0 \). This condition is also implicitly required for \( p \) value-based FDR control procedures.
to guarantee the uniform distribution of \( p \) values from the true null tests. (This point is demonstrated by the connection between the QBF procedure and the Storey’s \( q \)-value procedure in the Sect. 3.2.2). It is important to note that this condition does not impose any restriction on the proportion of the true null tests, i.e., \( \pi_0 \). The condition for sample size per test is used for proving the stringent Bayesian FDR control criteria. Although we believe that the result of the Proposition 1 holds true even without this particular large sample requirement (especially for controlling frequentist FDR), the proof becomes extremely difficult, and our attempt is unsuccessful. Nevertheless, in practice, we observe that the stated decision rule works well for very modest sample sizes: illustrations are given in the simulation study and the real data application where the sample sizes are both quite limited.

### 3.2 Upper-Bound Estimation of \( \pi_0 \) Using Bayes Factors

In this section, we describe two methods that can robustly estimate \( \hat{\pi}_0 \) that satisfy Assumption 2 in the Proposition 1 using only Bayes factors. The two methods share a simple intuition based on the stochastic orderings: for Bayes factors with reasonable powers, the smaller values are more likely generated from the null models. Theoretically, they both rely on the weak law of large numbers (WLLN) to ensure the upper-bound property stated in (3.1). Note that both approaches require the number of the simultaneous tests is sufficiently large, but impose no restrictions on sample sizes in each test. R code code implementing the described computational procedures is made freely available at [https://github.com/xqwen/bfdr](https://github.com/xqwen/bfdr).

#### 3.2.1 Sample Mean-Based Estimator

We first introduce an estimation procedure for \( \hat{\pi}_0 \) using the sample mean of the observed Bayes factors. This procedure, named the EBF-procedure, is based on the following Lemma:

**Lemma 1** A well-defined Bayes factor satisfies

\[
\mathbb{E}(BF \mid H_0) = 1.
\]

The proof of the Lemma is trivial from the definition of the Bayes factor.

Intuitively, under some suitable conditions and by the large sample theory, the Lemma 1 implies that the sample means of the Bayes factors from the true null models converge to 1. On the other hand, the Bayes factors from the true alternative models with reasonable powers should be, on average, \( \geq 1 \) (i.e., favoring the alternative over the null models). Therefore, the sample mean of the observed Bayes factors carries information regarding the mixture percentage. We give a detailed description of the EBF procedure below:

1. Sort all observed Bayes factors in ascending order,

\[
BF(1), BF(2), \ldots, BF(m)
\]
2. Find

\[ d_0 = \arg \max_d \left( \frac{1}{d} \sum_{i=1}^{d} \text{BF}(i) < 1 \right) \]  \hspace{1cm} (3.4)

3. Estimate \( \hat{\pi}_0 \) by

\[ \hat{\pi}_0 = \frac{d_0}{m} \]  \hspace{1cm} (3.5)

Essentially, this procedure searches for the largest proportion of tests with mean Bayes factors \(<1\). We demonstrate in Lemma 2 and Proposition 2 that the EBF procedure produces the upper-bound estimate of \( \pi_0 \) required by the Proposition 1.

**Lemma 2** In the case \( \pi_0 = 1 \), the EBF procedure estimates \( \hat{\pi}_0 \xrightarrow{P} 1 \), provided that the number of tests is sufficiently large.

The proof is simply based on the law of large numbers, and the full details are given in the Appendix 2.

**Proposition 2** For a mixture of null and alternative models, the estimate by the EBF procedure satisfies

\[ \Pr(\hat{\pi}_0 \geq \pi_0 | \pi_0) \rightarrow 1, \]

provided that the number of the null tests is sufficiently large.

The Proposition 2 is trivially followed from Lemma 2, and the detailed argument is provided in the Appendix 3.

The EBF procedure also allows the study of some of the general properties of Bayesian FDR control. Here, we show one interesting example: if a single Bayes factor exceeds a certain threshold, the corresponding null hypothesis can be rejected with respect to a targeted FDR level without the explicit evaluation of \( \hat{\pi}_0 \). We call this threshold the automatic rejection threshold of Bayes factors in multiple testing. We have the following corollary concerning one such automatic rejection threshold:

**Corollary 1** Under the settings stated in Proposition 1 and at the FDR level \( \alpha \), if a Bayes factor exceeds \( m/\alpha \), the EBF procedure automatically rejects the corresponding null hypothesis.

The proof is trivially based on the estimate of \( \hat{\pi}_0 \) by the EBF procedure and Eq. (2.7); we give the details in the Appendix 4. Most interestingly, the automatic rejection threshold, \( m/\alpha \), clearly resembles Bonferroni corrections for \( p \) values.
3.2.2 Sample Quantile-Based Estimator

Alternatively, we can estimate $\hat{\pi}_0$ using the sample quantile information based on the distribution of Bayes factors under the null models. Let $F^0_i(x)$ and $F^1_i(x)$ denote the cumulative distribution functions of $BF_i$ under the null and the true (unknown) alternative models, respectively. Suppose that

$$F^0_i(q_{i},\gamma) = \gamma,$$

i.e., $q_{i,\gamma}$ is the $\gamma$-quantile of $BF_i$ under the null model. Consider the sample mean of the independent indicators $I(BF_i \leq q_{i,\gamma})$, for $i = 1, \ldots, m$. By Chebyshev’s weak law of large numbers, it follows that

$$\frac{1}{m} \sum_{i=1}^{m} I(BF_i \leq q_{i,\gamma}) \xrightarrow{p} \pi_0 \gamma + \frac{1 - \pi_0}{m} \sum_{i=1}^{m} F^1_i(q_{i,\gamma}).$$

(3.7)

Note that if $\gamma$ is small and/or $F^0_i(x)$ and $F^1_i(x)$ are well separated, the contribution from the $F^1_i(q_{i,\gamma})$ terms should also be small. This is because Bayes factors computed using data from the alternative models are stochastically dominant over the Bayes factors evaluated using data generated from the null models, a property implied by the consistency of the Bayes factors under the null models. Based on (3.7), we propose an estimating algorithm, named the QBF procedure, to estimate $\hat{\pi}_0$. We summarize the details of the QBF procedure and its theoretical properties in the following proposition:

**Proposition 3** For a given $\gamma$, the QBF-procedure estimates

$$\hat{\pi}_0 = \frac{\sum_{i=1}^{m} I(BF_i \leq q_{i,\gamma})}{m \gamma}.$$  

(3.8)

As the number of the null tests is sufficiently large, it follows that

$$\Pr(\hat{\pi}_0 \geq \pi_0 \mid \pi_0) \to 1.$$

The proof follows trivially from (3.7) by ignoring the contribution from the term $\frac{1 - \pi_0}{m} \sum_{i=1}^{m} F^1_i(q_{i,\gamma})$ on the right-hand side.

It should be noted that the QBF-procedure is connected with what was proposed by [27]: treating $BF_i$ as a test statistic and denoting its $p$ value by $p_i$; under the null model, it follows that

$$BF_i \leq q_{i,\gamma} \iff p_i > 1 - \gamma.$$  

(3.9)

This yields an equivalent representation of (3.8):

$$\hat{\pi}_0 = \frac{\sum_{i=1}^{m} I(p_i > 1 - \gamma)}{m \gamma},$$  

(3.10)

which is the exact formula proposed in [27].
Based on (3.7), it can be concluded that the QBF procedure generates an upper-bound estimate of $\pi_0$ using any $\gamma \in (0, 1)$ asymptotically. However, for Bayes factors with reasonable power, taking smaller $\gamma$ values yields relatively tighter upper bounds on $\pi_0$ (because of the relatively small contribution from $F_i^1(q_{i,\gamma})$). However if $\gamma$ is too small, the estimator $\hat{\pi}_0$ shown in (3.8) has undesired large variance for finite $m$. To strike a balance, we simply set $\gamma = 0.5$ for demonstrations in this paper. ([26,27] considered the choice of $\gamma$ as a problem of bias-variance tradeoff and offered an elegant statistical solution to find an “optimal” $\hat{\pi}_0$ by using multiple $\gamma$ values. It should be noted that their solution also works in our case.)

When the quantiles of the Bayes factors under the null models are not directly available, we can always employ a permutation procedure to estimate them from the samples at hand. According to (3.9), the permutation procedure is no different statistically than computing permutation $p$ values. However, because the quantiles preferred by the QBF procedures are typically not in the tail area, they can be estimated stably with fewer permutations (than is required for accurately assessing permutation $p$ values). Therefore, even in this case, the QBF procedure is still computationally efficient. We will fully illustrate this point in our numerical studies.

### 3.2.3 Comparison of EBF and QBF Procedures

One of the practical differences between the QBF and EBF procedures lies in the relative conservativeness of the resulting $\hat{\pi}_0$ estimates. We observe that the QBF procedure with $\gamma = 0.5$ typically yields a tighter upper bound than the EBF procedure. This might be explained by the choice of parameter $\gamma$ in the QBF procedure: smaller $\gamma$ values yield tighter upper-bound estimates. Although the EBF procedure does not explicitly use the quantile information, computing a sample mean from the null distribution of the Bayes factors requires examining $BF_i \leq q_{i,\gamma}$ for $\gamma \to 1$, i.e., nearly the full distribution is used, and this is seemingly analogous to applying the QBF procedure with a large $\gamma$ value.

Once the Bayes factors are computed, the computational cost for the EBF procedure is trivial in all cases. However, for the QBF procedure, in many situations in which the quantiles of the Bayes factors under the null models are not directly computable, there is an additional cost of permutations for quantile estimations.

### 4 Simulation Study: Single-Tissue eQTL Discovery

In this section, we demonstrate the proposed robust Bayesian FDR control method in the context of eQTL mapping, which aims to identify genetic variants that are associated with gene expression levels. Specifically, we consider a problem of genome-wide eQTL mapping in a single tissue for which we can compare the proposed Bayesian approach with the gold-standard frequentist approach based on permutation $p$ values.

Particularly, we consider the expression levels of $L$ different genes across the genome are measured. For each gene $i$, we consider $k_i$ candidate SNPs, which are typically located in the cis region of the target gene, and model its expression level $y_i$ in the sample using the following multiple linear regression model,
where \( \beta_{ij} \) denote the genotypes of the \( j \)-th candidate genetic variant in gene \( i \), and the \( \beta_{ij} \)s are typically correlated. Based on the linear model, the problem of eQTL discovery can be framed as testing the following hypothesis for each gene \( i \):

\[
H_0 : \beta_{ij} = 0, \forall j \quad \text{vs.} \quad H_1 : \text{some } \beta_{ij} \neq 0,
\]

and we perform the simultaneous hypothesis testing for all \( L \) genes. In practice, this QTL discovery procedure for identifying genes harboring causal associated SNPs in their cis regions is referred to as eGene discovery.

### 4.1 Simulation Scheme

We take the real genotype data from a published study [1] in which \( n = 85 \) unrelated individual samples are genotyped genome-wide. We select 10,000 genes from the data set, where each gene contains 40–120 cis-SNPs. Within each gene, genotypes of different genetic variants are typically correlated at various levels due to linkage disequilibrium (LD). For each gene, we simulate its phenotype vector according to the linear model (4.1). More specifically, we fix the intercept term to \( \mu_i = 1 \) and residual error variance to \( \sigma_i^2 = 1 \), and we again assume that it is known. Under the alternative scenario, for each gene, we randomly sample one to five associated SNPs, and simulate their genetic effect sizes by drawing from the distribution \( \beta_{ij} \sim N(0, 0.62) \). For each gene, with probabilities \( \pi_0 \) and \( 1 - \pi_0 \), we simulate under the null and the alternative scenarios, respectively.

#### 4.1.1 Methods for Comparison

The testing problem encountered in this simulation setting is challenging, primarily because there is no straightforward test statistic available for the desired gene-level test. To obtain a valid gene-level \( p \) value for gene \( i \), the standard frequentist approach for eQTL mapping performs \( k_i \) single SNP association tests and obtain \( p \) values based on the resulting \( t \)-statistics. Subsequently, the minimum of the \( k_i \) \( p \) values is regarded as the test statistic for the gene \( i \). (We shall term this test statistic as the min-\( p \) statistic). Intuitively, the min-\( p \) statistic is most powerful if there is only a single associated SNP within the candidates [4]. Note that the min-\( p \) statistic itself is no longer a valid \( p \) value, as it does not follow a uniform distribution under the null. Furthermore, because of the complex correlation structures among the genotypes, the null distribution of the min-\( p \) statistic is generally unknown. In practice, this issue is resolved by permuting the individual labels of the phenotype vectors and assessing a permutation \( p \) value of the min-\( p \) statistic, \( p_{minp} \), for each gene. These gene-level \( p \) values are then plugged into either the Storey procedure [23,24] or the Benjamini–Hochberg (B–H) [2] procedure for multiple testing control.
For Bayesian hypothesis testing, we apply a simple Bayesian model proposed by [22]. Briefly, this model explicitly assumes that, under the alternative scenario, exactly one genetic variant is associated with the expression level, which is similar to the implicit assumption made by the standard frequentist approach. Furthermore, it assumes that every candidate genetic variant is equally likely to be associated \textit{a priori}. Under these assumptions, for gene $i$, there are exactly $k_i$ possible alternative single SNP association models. For each alternative model with assumed causal SNP $i_j$, a Bayes factor, $BF_{i_j}$, can be analytically computed as a simple function of the $t$-statistic, $\hat{\beta}_{i_j}/se(\hat{\beta}_{i_j})$, by fitting the corresponding simple linear regression models [3,22,29]. We then compute a gene-level Bayes factor by averaging over all single SNP association models, i.e.,

$$BF_i = \frac{1}{k_i} \sum_{j} BF_{i_j}. \quad (4.2)$$

Note that $BF_i$ is a well-defined Bayes factor, and can be directly used in the EBF and the QBF procedure.

We apply both the EBF and the QBF procedures to analyze the simulated eQTL data. Because the null distribution of the Bayes factor is not available for the QBF procedure, we use permutations to estimate the median of the null Bayes factor distribution for each gene. In addition, we also obtain a permutation $p$ value by treating $BF_i$ as a gene-level test statistic, which is denoted by $p_{bf}$, and use it for the B–H and the Storey procedures.

### 4.2 Simulation Results

We simulate data for $\pi_0 = 0.15, \ldots, 0.95$ and create 20 independent replicates for each $\pi_0$ value. We apply the Storey and B–H procedures using both $p_{\text{min}}$ and $p_{bf}$, and use the gene-level Bayes factors for the EBF and QBF procedures.

We first examine the $\hat{\pi}_0$ estimates from the EBF, QBF, and Storey procedures. The estimates from the QBF and the Storey procedures are based on permutations. We vary the number of permutations from 100 to 500 to 5000 for $\gamma = 0.5$. We average the $\hat{\pi}_0$ estimates over replicates and summarize the results in Table 1. The $\hat{\pi}_0$ estimates obtained by the Storey procedure based on $p_{bf}$ are almost identical to the QBF estimates with the same permutation repeats. To avoid the redundancy, they are not shown. All the estimates satisfy the upper-bound requirement. The EBF procedure consistently yields the most conservative estimate. The results from the QBF and Storey procedures based on $p_{\text{min}}$ are mostly similar, although the QBF estimates are slightly less conservative. One important observation is that for all three methods, the $\hat{\pi}_0$ estimates have very little variation according to the number of permutations, which is mostly expected for the reason discussed in Sect. 3.2.2.

We then proceed to examine the realized false discovery and false non-discovery rates obtained by the B–H, Storey, QBF, and EBF procedures. The results are summarized in Tables 2 and 3. Because the estimates of $\hat{\pi}_0$ from the QBF procedure have only small variation for different numbers of permutations, its realized FDP and FNP results are
Table 1  \( \hat{\pi}_0 \) estimates by the Storey, QBF, and EBF procedures in the simulation study

| \( \pi_0 \) | \( \hat{\pi}_0 \) Storey (\( p_{\min} \)) | QBF | EBF |
|----------|------------------|-----|-----|
|          | 100              | 500 | 5000|     |
| 0.95     | 0.978            | 0.970| 0.974| 0.977| 0.969| 0.973| 0.982|
| 0.85     | 0.914            | 0.910| 0.905| 0.911| 0.907| 0.903| 0.943|
| 0.75     | 0.849            | 0.842| 0.846| 0.842| 0.845| 0.838| 0.901|
| 0.65     | 0.781            | 0.784| 0.787| 0.772| 0.775| 0.779| 0.857|
| 0.55     | 0.726            | 0.723| 0.723| 0.718| 0.714| 0.713| 0.813|
| 0.45     | 0.659            | 0.660| 0.657| 0.647| 0.650| 0.647| 0.766|
| 0.35     | 0.602            | 0.600| 0.596| 0.590| 0.587| 0.584| 0.718|
| 0.25     | 0.532            | 0.535| 0.530| 0.517| 0.520| 0.514| 0.668|
| 0.15     | 0.471            | 0.468| 0.466| 0.455| 0.452| 0.451| 0.616|

Each entry indicates the average estimates for different \( \pi_0 \) values taken over 20 simulated data sets. For the Storey and the QBF procedures, the estimates are based on the permutation procedures; the results obtained for different permutation numbers are also shown. All estimates satisfy the upper bound requirement, and they are not very sensitive to the number of permutations.

Table 2  Realized FDPs yielded by the B–H, Storey, QBF, and EBF procedures in simulation study II

| \( \pi_0 \) | \( \text{B–H} (p_{\min}) \) | \( \text{B–H} (p_{bf}) \) | \( \text{Storey} (p_{\min}) \) | \( \text{Storey} (p_{bf}) \) | QBF | EBF |
|----------|------------------|------------------|------------------|------------------|-----|-----|
|          | 500              | 5000             | 500              | 5000             |     |     |
| 0.95     | 0.000†           | 0.046            | 0.000†           | 0.043            | 0.000†| 0.043| 0.043| 0.050| 0.035|
| 0.85     | 0.033            | 0.043            | 0.031            | 0.043            | 0.033| 0.047| 0.031| 0.048| 0.039| 0.028|
| 0.75     | 0.031            | 0.036            | 0.029            | 0.036            | 0.042| 0.042| 0.041| 0.043| 0.031| 0.021|
| 0.65     | 0.028            | 0.032            | 0.026            | 0.032            | 0.038| 0.041| 0.040| 0.042| 0.029| 0.020|
| 0.55     | 0.025            | 0.028            | 0.025            | 0.028            | 0.038| 0.039| 0.038| 0.038| 0.025| 0.017|
| 0.45     | 0.021            | 0.022            | 0.021            | 0.022            | 0.032| 0.034| 0.033| 0.034| 0.022| 0.014|
| 0.35     | 0.018            | 0.018            | 0.017            | 0.018            | 0.029| 0.030| 0.029| 0.030| 0.020| 0.013|
| 0.25     | 0.012            | 0.013            | 0.011            | 0.012            | 0.023| 0.024| 0.024| 0.024| 0.016| 0.009|
| 0.15     | 0.007            | 0.008            | 0.007            | 0.008            | 0.016| 0.016| 0.016| 0.017| 0.011| 0.006|

Each entry is obtained by averaging over the 20 simulated data sets. The entries annotated by “†” indicate that the corresponding method fails to reject any hypothesis.

virtually invariant with respect to the permutation numbers; therefore, we only report the results using \( \hat{\pi}_0 \) estimated by 100 permutations. Contrary to the Bayesian procedures, the number of permutations directly impacts the precision of the permutation \( p \) values and consequently alters the outcomes of the \( p \) value based FDR control procedures. Such impacts can be severe especially when the value of \( \pi_0 \) is close to 1. In our experiment, when \( \pi_0 = 0.95 \), with 500 permutations, neither the Storey nor the B–H procedure is able to reject a single null hypothesis. The issue seems to be resolved by
Table 3  Realized FNPs obtained by the B–H, Storey, QBF, and EBF procedures in simulation study II

| $\pi_0$ | FNP | B–H ($p_{\min}$) 500 | B–H ($p_{bf}$) 500 | Storey ($p_{\min}$) 500 | Storey ($p_{bf}$) 500 | QBF 500 | EBF 500 |
|---------|-----|----------------------|---------------------|-------------------------|-------------------------|--------|--------|
| 0.95    | 0.050† | 0.034                | 0.050†              | 0.034                   | 0.050†                   | 0.034  | 0.034  |
| 0.85    | 0.104 | 0.101                | 0.103               | 0.100                   | 0.103                   | 0.099  | 0.103  |
| 0.75    | 0.173 | 0.169                | 0.171               | 0.167                   | 0.169                   | 0.165  | 0.172  |
| 0.65    | 0.247 | 0.243                | 0.244               | 0.240                   | 0.242                   | 0.236  | 0.241  |
| 0.55    | 0.329 | 0.325                | 0.325               | 0.321                   | 0.321                   | 0.314  | 0.330  |
| 0.45    | 0.417 | 0.414                | 0.411               | 0.409                   | 0.408                   | 0.406  | 0.409  |
| 0.35    | 0.516 | 0.515                | 0.511               | 0.509                   | 0.505                   | 0.503  | 0.507  | 0.517  |
| 0.25    | 0.631 | 0.629                | 0.626               | 0.623                   | 0.616                   | 0.615  | 0.618  |
| 0.15    | 0.761 | 0.760                | 0.757               | 0.755                   | 0.746                   | 0.745  | 0.739  |
| Each entry is obtained by averaging over the 20 simulated data sets. The entries annotated by “†” indicate that the corresponding method fails to reject any hypothesis and the FNP achieves the maximum value.

Table 4  Computational time measured for various FDR control procedures in simulation study II

| EBF   | QBF (100) | B–H/Storey (500) | B–H/Storey (5000) |
|-------|-----------|------------------|-------------------|
| 0m43s | 3m34s     | 11m56s           | 114m34s           |

Numbers in parenthesis indicate the number of permutations performed for each procedure.

performing more permutations; however, we have to pay a steep price in computations even in this simple setting.

Overall, all the methods maintain the desired level of FDR control, and their powers are quite comparable (when the accurate $p$ values are provided to the Storey and B–H procedure). More specifically, we note that

1. Results from the B–H and Storey procedures based on $p_{bf}$ provide benchmarks to compare the frequentist and the Bayesian FDR controls in this setting. The Bayesian FDR controls are expected to be more stringent, but the differences reflected by the realized FNPs are mostly small (especially when $\pi_0$ is close to 1).

2. The two test statistics (the Bayes factor and min-$p$) display very similar powers in gene-level testing, although the Bayes factor performs slightly better.

Although all the procedures yield extremely similar statistical performance, they have striking differences in computational efficiency. We benchmark the running time of a complete analysis of a single simulated data set using various procedures. All of the procedures tested are implemented in the C++ programming language and run on a machine with an 8-core Intel Xeon 2.13 GHz CPU. We present the results in Table 4.

Given required $p$ values or Bayes factors, the computations for various FDR control procedures are mostly trivial, the main difference in computational time is primarily attributed to the cost in permutations.
5 Real Data Application: Multi-tissue eQTL Discovery

In this section, we demonstrate the proposed statistical approach in a real data application of mapping eQTL across multiple tissues. The aim of the multi-tissue eQTL mapping is to investigate the commonality and the specificity of the eQTLs in different tissue environments. Comparing to the single-tissue eQTL study described in the simulation study, expression levels for each gene are collected in multiple tissue/cell types from the same set of unrelated individuals, and a joint eQTL analysis across all tissues is required. Within each tissue, we assume the same statistical model as the linear model (4.1); however, across tissues, the residual errors are generally assumed to be correlated [30].

Our primary goal is still to identify genes harboring causal eQTLs (i.e., eGenes): let \( \beta_{ijt} \) denote the association coefficient between expressions of gene \( i \) and its \( j \)-th genetic variant in tissue \( t \). For each gene \( i \), we are interested in testing the following hypothesis:

\[
H_0 : \beta_{ijt} = 0, \forall j, t \quad \text{vs.} \quad H_1 : \text{some } \beta_{ijt} \neq 0. \quad (5.1)
\]

Clearly, the problem is even more challenging when considering the potential association patterns across multiple tissues under the alternative scenario. First, multi-tissue data allow identifying association signals that might be modest but consistent across tissues. To achieve this identification, it requires a framework analogous to a meta-analysis that jointly models potential associations across tissue types. Secondly, it is known that a proportion of eQTLs are only active in specific tissue types (i.e., \( \beta_{ijt} \neq 0 \) only for some \( t \)). To account for this, we must explicitly address the configurations of possible association activities across multiple tissues.

In [6], those authors used a naive gene-level test statistic \( T_i = \min_{j,t} p_{ijt} \), where each \( p_{ijt} \) is the \( p \) value from a single association test for each gene-genetic variant pair in each tissue type. The \( p \) value of the overall test statistic, \( T_i \), must be evaluated in a very complicated permutation scheme to account for intra-individual correlations of gene expression levels and the LD in genetic variants. Apparently, this naive test statistic does not fully capture the alternative scenarios discussed above and is expected to lose power. In brief, to the best of our knowledge, there is no frequentist approach that can produce appropriate \( p \) values to test (5.1).

Recently, [8] proposed a full parametric Bayesian model to explicitly describe all possible alternative scenarios and derived a Bayes factor to efficiently test eGenes, taking advantage of Bayesian model averaging. More specifically, for eQTL data collect from \( T \) different tissues, they consider all \( 2^T - 1 \) possible configurations of association activities for each potential associated SNP (e.g., a configuration \((1, 1, ..., 1)\) indicates the eQTL is active in all tissues). For each given configuration, a Bayes factor can be computed analytically accounting for the similarity and heterogeneity of the genetic effect sizes (\( \beta_{ijt} \)'s) of each active association in different tissues [31]. The probability weights on each type of the total \((2^T - 1)\) non-zero configuration model are then estimated using an EM algorithm by pooling information across all available gene-SNP data across the genome. Given the resulting weights on configurations and the “(at most) one causal association per cis region” assumption, a weighted Bayes factor is
evaluated using Bayesian model averaging. Note that the weighted Bayes factor itself is a well-defined Bayes factor and can be plugged into the proposed EBF and QBF procedures. [8] convincingly demonstrated that the resulting Bayes factor is much more powerful than the naive test statistic \( T_i = \min_{j,t} p_{ijt} \). Nevertheless, to apply a multiple testing procedure for identifying eGenes, they had to rely on permutation procedures to obtain \( p \) values by treating the resulting Bayes factors as the frequentist test statistics.

We apply both the EBF and the QBF procedures to a real multiple tissue eQTL data set that was originally published in [6] and re-analyzed by [8]. In the experiment, 75 unrelated Western-European individuals are genotyped. Expression levels from this set of individuals were measured genome-wide in primary fibroblasts, Epstein-Barr virus-immortalized B cells (LCLs), and T cells. The expression data went through quality control and normalization steps (by [6]), and a subset of 5012 genes that were deemed highly expressed in all three cell types is selected for eQTL analysis.

Based on the Bayes factor derived in [8], the QBF (in which \( \hat{\pi}_0 \) is estimated with only 100 permutations) and EBF procedures identify 1002 and 927 eGenes, respectively, at a Bayesian FDR level of 5%. In comparison, [8] identified 1012 eGenes by applying the Storey procedure using the permutation \( p \) values of the Bayes factors (with 10,000 permutations) also at an FDR level of 5%. Although the three methods yield similar results, the EBF and the QBF procedures achieve dramatic computational time reduction by avoiding large amounts of permutations. Finally, [8] noted that Storey’s procedure based on permutation \( p \) values of the naive \( T_i \) statistics (with 10,000 permutations) only identified 627 eGenes. Among those eGenes, 592 genes (or 94%) are also identified by the EBF and QBF approaches.

6 Discussion

We have introduced a Bayesian FDR control procedure with Bayes factors that is robust to misspecifications of alternative models. This feature should provide peace of mind for practitioners who are attempting parametric Bayesian models in multiple hypothesis testing. Nevertheless, within our framework, the model specification still dictates the overall performance, e.g., a badly designed alternative model would have very little power and would therefore be useless. Our central message throughout this paper has been that various FDR control procedures have little practical difference if the same or similar test statistics are applied; however, our proposed procedure encourages well-designed parametric modeling approaches to obtain more powerful test statistics.

Another advantage of our approach is its computational efficiency, especially when compared with the alternative method’s reliance on permutation \( p \) values. However, this statement assumes the Bayes factors can be easily evaluated. Many previous and ongoing works have been devoted to efficient evaluation and approximation of Bayes factors in various model systems. In particular, [12] demonstrated that almost all the well-known frequentist test statistics (e.g., \( z \), \( F \), and \( \chi^2 \) statistics) can be “translated” into corresponding Bayes factors analytically, and [30] demonstrated that Bayes
factors can be analytically approximated in a general linear model system that includes multiple linear regression, multivariate linear regression models, and many others as special cases.

Our proposed procedure bears certain similarities to the local FDR procedure [7, 28]. However, it should be noted that the two procedures differ significantly in the following two aspects. First, our procedure utilizes Bayes factors as test statistics whereas local FDR procedure relies on \(z\)-statistics. We note that in complex parametric modeling settings, \(z\)-statistic is not always straightforwardly defined (e.g., in our eQTL mapping examples). Second, the local FDR procedure is an empirical Bayes approach: it plugs in the MLE of \(\pi_0\) instead of performing full Bayesian inference on \(\pi_0\). As discussed in this paper, this approach can be susceptible to model misspecifications. In contrast, our proposed approach is robust to model specification, and philosophically is fully Bayesian by viewing \(\hat{\pi}_0\) as a conservative approximation (by a point mass) to the posterior distribution of \(\pi_0\).

Acknowledgements We thank Debashis Ghosh, Matthew Stephens, and Timothee Flutre for their fruitful discussion and feedbacks. We are grateful for the insightful comments from the two anonymous reviewers. This work is supported by the NIH Grant R01 MH101825 (PI: M.Stephens).

Appendix 1: Proof of Proposition 1

Proof Suppose that the true generative distributions of \(Y_i\) under the null and alternative model are given by \(p_{i0}^*\) and \(p_{i1}^*\), respectively. As sample size \(n_i\) is sufficiently large for each test, the true Bayes factor, \(BF_i^*\), has the following properties

\[
\begin{align*}
BF_i^* &\xrightarrow{P} 0, \quad \text{if } Y_i \sim p_{i0}^*, \\
\text{and} &
BF_i^* &\xrightarrow{P} \infty, \quad \text{if } Y_i \sim p_{i1}^*. 
\end{align*}
\]

(6.1)

Assumption 1, in contrast, only requires that the assumed Bayes factors satisfy

\[
BF_i \xrightarrow{P} 0, \quad \text{if } Y_i \sim p_{i0}^*.
\]

(6.2)

Under the conditions stated in Assumption 2, \(\Pr(\hat{\pi}_0 \geq \pi_0 \mid \pi_0) \to 1\) implies that \(\Pr(\hat{\pi}_0 \geq \pi_0 \mid Y) \to 1\) for an arbitrary prior distribution on \(\pi_0\). Let \(\epsilon := \Pr(\hat{\pi}_0 < \pi_0 \mid Y)\). Then, it follows that

\[
\epsilon \xrightarrow{P} 0, \quad \text{as the number of null tests is sufficiently large.}
\]

(6.3)

Importantly, the large number of the null tests is required for ensuring the upper-bound property of \(\hat{\pi}_0\).
Consequently,

\[
\Pr(Z_i = 0 \mid \mathcal{Y}) = \int \frac{\pi_0}{\pi_0 + (1 - \pi_0)BF_i^*} p(\pi_0 \mid Y_i) \, d\pi_0 \\
= \int_{\pi_0 \leq \hat{\pi}_0} \frac{\pi_0}{\pi_0 + (1 - \pi_0)BF_i^*} p(\pi_0 \mid Y_i) \, d\pi_0 \\
+ \int_{\pi_0 > \hat{\pi}_0} \frac{\pi_0}{\pi_0 + (1 - \pi_0)BF_i^*} p(\pi_0 \mid Y_i) \, d\pi_0 \\
\leq \hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i^* + \epsilon
\]  

(6.4)

By (6.1), as \(n_i \to \infty\),

\[
\frac{\hat{\pi}_0}{\hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i^*} \xrightarrow{p} I\{Y_i \sim p_i^{0*}\} + 0 \cdot I\{Y_i \sim p_i^{1*}\}
\]  

(6.5)

whereas by (6.2),

\[
\frac{\hat{\pi}_0}{\hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i^*} \xrightarrow{p} 1, \quad \text{if} \quad Y_i \sim p_i^{0*},
\]

\[
\frac{\hat{\pi}_0}{\hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i^*} \geq 0, \quad \text{if} \quad Y_i \sim p_i^{1*}.
\]  

(6.6)

Hence,

\[
\lim_{n \to \infty} \Pr\left(\frac{\hat{\pi}_0}{\hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i^*} \geq \frac{\hat{\pi}_0}{\hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i^*}\right) = 1,
\]  

(6.7)

and by (6.3) and (6.4), each individual test satisfies

\[
\lim_{n_i \to \infty} \Pr ((1 - \hat{v}_i) \geq \Pr(Z_i = 0 \mid Y_i)) = 1
\]  

(6.8)

The decision rule stated in the Proposition 1 yields a true Bayesian FDR

\[
FDR = \sum_{i=1}^{m} \delta_i^+ \Pr(Z_i = 0 \mid Y_i) / D^+ \lor 1.
\]  

(6.9)

By (6.8), it is clear that

\[
\lim_{n \to \infty} \Pr\left(\frac{\text{FDR} \leq \sum_{i=1}^{m} \delta_i^+(1 - \hat{v}_i)}{D^+ \lor 1}\right) = 1.
\]  

(6.10)
Therefore, the Bayesian FDR can be consistently controlled. Furthermore, controlling 
\[ \frac{\sum_{i=1}^{m} \delta_i (1 - \hat{v}_i)}{D^\top \vee 1} \leq \alpha \] ensures that 
\[ \text{FDR} = E(\text{FDR}) \leq E \left( \frac{\sum_{i=1}^{m} \delta_i (1 - \hat{v}_i)}{D^\top \vee 1} \right) \leq \alpha, \] (6.11)
because \( \bar{\text{FDR}} \) is obviously bounded.

**Appendix 2: Proof of Lemma 2**

**Proof** Let \((\text{BF}(1), \ldots, \text{BF}(m))\) denote the order statistics from the \(m\) Bayes factors that are all generated from the respective null hypotheses. Let \(M_j := \frac{1}{j} \sum_{i=1}^{j} \text{BF}(i)\) denote the partial sample mean computed by the EBF procedure. Note that the sequence \(M_1, M_2, \ldots\) is monotonically non-decreasing. Furthermore, by the law of large numbers and the result of Lemma 1, it follows that 
\[ M_m \xrightarrow{p} 1, \] (6.12)
for sufficiently large \(m\).

In the case that \(d_0 < m\), it must be true that 
\[ 1 \leq M_{d_0+1} \leq M_m \xrightarrow{p} 1. \]

Because the truncated sample mean from the order statistics \((\text{BF}(d_0+1), \ldots, \text{BF}(m))\) converges to a quantity that is strictly >1, their contribution to the overall sample mean, \(M_m\), must be negligible, i.e., 
\[ \frac{m - d_0}{m} \left( \frac{1}{m - d_0} \sum_{j=d_0+1}^{m} \text{BF}(j) \right) \xrightarrow{p} 0. \]

Taken together, we have shown that 
\[ \hat{\pi}_0 = \frac{d_0}{m} \xrightarrow{p} 1. \] (6.13)

**Appendix 3: Proof of Proposition 2**

**Proof** In the general mixture case, let \(\mathcal{S}^0\) denote the subset of Bayes factors whose data are generated from the null models. Based on Lemma 2, applying the EBF procedure on \(\mathcal{S}^0\) results in an estimate
\[
\frac{d_{S_0}}{|S_0|} \xrightarrow{\mathcal{D}} 1, \quad (6.14)
\]

where \(|S_0|\) denotes the cardinality of \(S_0\). In the mixed samples, (6.14) suggests that

\[
M_{d_{S_0}} = \frac{1}{d_{S_0}} \sum_{j=1}^{d_{S_0}} BF(j) \leq 1. \quad (6.15)
\]

The LHS should be strictly < 1 if there exists small values of Bayes factors from the alternative models.

Because the EBF procedure finds the largest subset whose sample mean is < 1, it must hold true that

\[
\Pr(d_0 \geq d_{S_0} | \pi_0) \to 1, \quad (6.16)
\]

and thus we conclude that

\[
\Pr(\hat{\pi}_0 \geq \pi_0 | \pi_0) \to 1. \quad (6.17)
\]

\hfill \Box

**Appendix 4: Proof of Corollary 1**

**Proof** Given a predefined FDR level \(\alpha\), the rejection threshold on the estimated false discovery probabilities is given by

\[
t^\ast_\alpha = \arg \min_t \left( \frac{\sum_{i=1}^{m} \delta_i^\ast(t)(1 - \hat{\nu}_i)}{D^\ast(t) \lor 1} \leq \alpha \right).
\]

Equivalently when there is at least one rejection, the above rejection threshold implies that the rejection set \(\Omega := \{i : \hat{\nu}_i > t^\ast_\alpha\}\) is the largest set such that

\[
\sum_{i \in \Omega} \Pr(Z_i = 0 | Y_i, \hat{\pi}_0) / ||\Omega|| \leq \alpha,
\]

where \(\Pr(Z_i = 0 | Y_i, \hat{\pi}_0) = (1 - \hat{\nu}_i)\) and \(||\Omega|| = D^\ast(t^\ast_\alpha)\) denotes the cardinality of the set \(\Omega\). That is, the average estimated false rejection probability in the rejection set should be \(\leq \alpha\). Consequently, it implies that if the \(i\)-th test yields \(\Pr(Z_i = 0 | Y_i, \hat{\pi}_0) \leq \alpha\), it must be included in the rejection set (because \(\Omega\) is the largest such set). Therefore, to prove the corollary, it is sufficient to show that \(\Pr(Z_i = 0 | Y_i, \hat{\pi}_0) \leq \alpha\), provided that \(BF_i \geq \frac{m}{\alpha}\).
Applying the EBF procedure, a single Bayes factor with the value exceeding $m$ leads to

$$\hat{\pi}_0 \leq 1 - \frac{1}{m}.$$  

(6.18)

This implies

$$\Pr(Z_i = 0 \mid Y_i, \hat{\pi}_0) \leq \frac{1 - \frac{1}{m}}{(1 - \frac{1}{m}) + \frac{1}{m}BF_i} \leq \frac{m - 1}{m + (m - 1)\alpha} \cdot \alpha < \alpha$$  

(6.19)

References

1. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y (2012) Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci USA 109(4):1204–1209
2. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57(1):289–300
3. Cox DR, Hinkley D (1979) Theoretical statistics. Chapman & Hall, London
4. De la Cruz O, Wen X, Ke B, Song M, Nicolae DL (2010) Gene, region and pathway level analyses in whole-genome studies. Genet Epidemiol 34(3):222–231
5. DiCiccio TJ, Kass RE, Raftery A, Wasserman L (1997) Computing Bayes factors by combining simulation and asymptotic approximations. J Am Stat Assoc 92(439):903–915
6. Dimas AS, Deutsch S, Stranger BE, Montgomery SB et al (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325(5945):1246–1250
7. Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical bayes analysis of a microarray experiment. J Am Stat Assoc 96(456):1151–1160
8. Flutre T, Wen X, Pritchard J, Stephens M (2013) A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet 9(5):e1003486
9. Genovese C, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Stat 32(3):1035–1061
10. Good I (1992) The bayes/non-bayes compromise: a brief review. J Am Stat Assoc 87(419):597–606
11. Ji Y, Lu Y, Mills GB (2008) Bayesian models based on test statistics for multiple hypothesis testing problems. Bioinformatics 24(7):943–949
12. Johnson VE (2005) Bayes factors based on test statistics. J R Stat Soc Ser B 67(5):689–701
13. Johnson VE (2008) Properties of Bayes factors based on test statistics. Scand J Stat 35(2):354–368
14. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795
15. Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g priors for Bayesian variable selection. J Am Stat Assoc 103(481):410–423
16. Müller P, Parmigiani G, Robert C, Rousseau J (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. J Am Stat Assoc 99(468):990–1001
17. Müller P, Parmigiani G, Rice K (2006) FDR and Bayesian multiple comparisons rules. In: Bayesian statistics 8, vol 0. Oxford University Press, p 349–370
18. Newton MA, Noueiry A, Sarkar D, Ahlquist P (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 5(2):155–76
19. Oppe n-Rhein R, Strimmer K (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Stat Appl Genet Mol Biol 6
20. Raftery AE (1996) Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 83(2):251–266

Springer
21. Saville BR, Herring AH (2009) Testing random effects in the linear mixed model using approximate bayes factors. Biometrics 65(2):369–376
22. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3(7):e114
23. Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64(3):479–498
24. Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31(6):2013–2035
25. Storey JD (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. J R Stat Soc Ser B 69(3):347–368
26. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100(16):9440–9445
27. Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66(1):187–205
28. Sun W, Cai TT (2007) Oracle and adaptive compound decision rules for false discovery rate control. J Am Stat Assoc 102(479):901–912
29. Wakefield J (2009) Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol 33(1):79–86
30. Wen X (2014) Bayesian model selection in complex linear systems, as illustrated in genetic association studies. Biometrics 70(1):73–83
31. Wen X, Stephens M (2014) Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. Ann Appl Stat 8(1):176–203
32. Whittemore AS (2007) A Bayesian false discovery rate for multiple testing. J Appl Stat 34(1):1–9