A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies

Jon Wakefield*

In light of the vast amounts of genomic data that are now being generated, we propose a new measure, the Bayesian false-discovery probability (BFDP), for assessing the noteworthiness of an observed association. BFDP shares the ease of calculation of the recently proposed false-positive report probability (FPRP) but uses more information, has a noteworthy threshold defined naturally in terms of the costs of false discovery and nondiscovery, and has a sound methodological foundation. In addition, in a multiple-testing situation, it is straightforward to estimate the expected numbers of false discoveries and false nondiscoveries. We provide an in-depth discussion of FPRP, including a comparison with the $q$ value, and examine the empirical behavior of these measures, along with BFDP, via simulation. Finally, we use BFDP to assess the association between 131 single-nucleotide polymorphisms and lung cancer in a case-control study.

With the advent of new genotyping and other molecular biology technologies, there has been a huge increase in the quantities of data that are available for analysis; this has focused attention on the manner by which associations are reported in the epidemiology literature and, in particular, on methods by which the number of false positives can be controlled without missing too many scientifically interesting associations. Motivated by candidate-gene and genomewide association studies, we consider here the reporting of associations and the problem of multiple-hypothesis testing.

Despite numerous protestations to the contrary (for a particularly clear exposition, see the work of Goodman), a common error is to view $P$ values as the probability of the null hypothesis given the observed statistic, when, in fact, they give the probability of the statistic given the hypothesis. To assess the probability of the hypothesis given the data, a Bayesian approach is needed, and this requires the specification of the prior probability of the hypothesis and the probability of the data under specified alternatives. This motivates a move away from $P$ values, and here we suggest an approach based on an approximate Bayes factor that we call the “Bayesian false-discovery probability” (BFDP). Recently, Wacholder et al. introduced the false-positive reporting probability (FPRP) as a means to assess whether the strength of an association was “noteworthy,” a terminology we use here interchangeably with “discovery,” the latter term being common in the literature on multiple-hypothesis testing. FPRP was introduced as a criteria, “to help investigators, editors, and readers of research articles to protect themselves from overinterpreting statistically significant findings that are not likely to signify a true association; this endeavor seems extremely useful, and we introduce BFDP to satisfy the same objective. FPRP has generated a lot of interest, and we discuss its relationship to BFDP and to the related but subtly different $q$ value. The difference is that FPRP uses the observed significance region, whereas the $q$ value has a fixed region, which allows the false-discovery rate (FDR) to be controlled, a property not inherited by FPRP, making it difficult to calibrate. We outline how BFDP may be used for design and, in a multiple-hypothesis-testing context, how BFDP provides the expected numbers of false and missed discoveries.

**Methods**

*Reporting a Hypothesis as Noteworthy via Bayesian Decision Theory*

Under frequentist inference, the null hypothesis $H_0$ is viewed as nonrandom; so, to calculate the probability of $H_0$, one must take a Bayesian standpoint; alternatives to $H_0$ must also be considered, to define the sample space of hypotheses. Let $y = (y_1, ..., y_n)$ denote the observed data and $H_1$ the alternative hypothesis. The application of Bayes’s theorem gives the probability of the hypothesis $H_0$ given data $y$ as

$$
Pr(H_0 | y, H_0 \cup H_1) = \frac{p(y | H_0) Pr(H_0 | H_0 \cup H_1)}{p(y | H_0 \cup H_1)},
$$

where

$$
p(y | H_0 \cup H_1) = p(y | H_0) Pr(H_0 | H_0 \cup H_1) + p(y | H_1) Pr(H_1 | H_0 \cup H_1)
$$

is the probability of the data averaged over $H_0$ and $H_1$. $Pr(H_0 | H_0 \cup H_1)$ is the prior probability that $H_0$ is true (given that one of $H_0$ and $H_1$ is true), and $Pr(H_1 | H_0 \cup H_1) = 1 - Pr(H_0 | H_0 \cup H_1)$ is the prior on the alternative hypothesis. It is clear from...
equation (1) that we are calculating the probability of the null given that either \( H_0 \) or \( H_1 \) is true. Hence, we are calculating the "relative truth"; \( H_0 \) may provide a poor fit to the data, but so may \( H_1 \).

Although recognizing that we are conditioning on either \( H_0 \) or \( H_1 \), being true is of crucial importance, we now suppress this in our notation, for brevity, and write:

\[
\Pr(H_0 | y) = \frac{p(y | H_0) \pi_0}{p(y | H_0) \pi_0 + p(y | H_1) (1 - \pi_0)} \tag{2}
\]

where \( \pi_0 = \Pr(H_0) \) is the prior on the null. We rewrite equation (2) as

\[
\Pr(H_0 | y) = \frac{\text{BF} \times \text{PO}}{\text{BF} \times \text{PO} + 1},
\]

where

\[
\text{BF} = \frac{p(y | H_0)}{p(y | H_1)}
\]

is the Bayes factor and

\[
\text{PO} = \frac{\pi_0}{1 - \pi_0}
\]

is the prior odds of no association. The use of the Bayes factors as a summary of the evidence contained in the observed data has been advocated in both a medical context\(^{11}\) and a genetic epidemiology context.\(^{12}\) Often, the alternative hypothesis will be indexed by a continuous parameter. For example, for the case of \( H_0; \theta = 0 \) versus \( H_1; \theta 
eq 0 \), we have \( p(y | H_1) = |\mu(\theta)| \pi(\theta) d\theta \), where \( \pi(\theta) \) is the prior on \( -\infty < \theta < \infty \).

In terms of making a decision as to which one of \( H_0 \) and \( H_1 \) to report (given that we will report one of them), the Bayesian decision theoretic solution is to assign costs to the consequences of making a decision, given the truth of \( H_0 \) or \( H_1 \). Specifically, let \( C(\delta, H) \) be the cost associated with decision \( \delta \) when the truth is \( H \). Table 1 gives the four costs, two of which are zero; \( C_a \) is the cost of a false discovery (we decide to report an association as noteworthy when, in fact, the null is true), and \( C_n \) is the cost of a false nondiscovery (we decide to call an association nonnoteworthy). The ratio of costs, \( C_a/C_n \), will determine the expected number of these that are false discoveries and false nondiscoveries. The ratio of costs, \( C_a/C_n \), for candidate-gene studies is likely to be lower than that for genomewide association studies, since, in the latter, we wish to produce a "long list" of candidates for future investigation, whereas, in the former, the follow-up of noteworthy candidates is more expensive. If a large study attempts to be definitive, then the cost of a false discovery will be greater, and a more stringent (i.e., lower) threshold for \( \Pr(H_1 | y) \) will result. Wacholder et al.\(^{7} \) provide an excellent discussion of issues regarding the decision process in a variety of contexts.

### Table 1. Costs Corresponding to Decision \( \delta \)

<table>
<thead>
<tr>
<th>Truth</th>
<th>( \delta = 0 ) (Nonnoteworthy)</th>
<th>( \delta = 1 ) (Noteworthy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H = H_0 )</td>
<td>0</td>
<td>( C_a )</td>
</tr>
<tr>
<td>( H = H_1 )</td>
<td>( C_n )</td>
<td>0</td>
</tr>
</tbody>
</table>

Note.—\( C_a \) is the cost of a false discovery, and \( C_n \) is the cost of a false nondiscovery.

**Simple Null Hypothesis with Composite Alternative and Adjustment for Confounding**

We now consider a specific situation in which disease risk, \( p \), is modeled via the logistic regression

\[
\logit p = x' \gamma + z \theta ,
\]

with \( \gamma \), a \( c \times 1 \) vector of log relative risks corresponding to confounders, and \( \theta \), the log relative risk of interest; \( x \) is a \( c \times 1 \) vector of confounders, and \( z \) is the value of an "exposure." The null and alternative hypotheses are \( H_0; \theta = 0 \) and \( H_1; \theta 
eq 0 \), with \( \gamma \) unspecified under each. For concreteness, suppose that the association between risk and a particular SNP is of interest and assume a dominant model, so that \( z \) is an indicator of being heterozygous or homozygous for the mutant allele and \( e^\theta \) is the associated relative risk.

The calculation of the Bayes factor in equation (3) requires the complete specification of both the data-generating mechanism (the likelihood), the prior distribution for all parameters of the model, and the calculation of multidimensional integrals. In general, each of these steps is difficult (appendix B contains details of these specifications). Instead, suppose a logistic regression produces an estimate, \( \hat{\theta} \), with associated standard error \( \hat{\sigma} \). We then (i) summarize the information in the likelihood concerning the parameter of interest \( \theta \), using \( \hat{\theta} \) and its asymptotic distribution \( N(\theta, V) \), and (ii) consider a prior for \( \theta \) only, rather than a joint specification for \( \theta \) and \( \gamma \). We assume this prior is \( \theta \sim N(0, W) \); it is natural to assume that the prior is centered at 0, corresponding to the null of no association (the probability of any specific value, in particular the null value \( \theta = 0 \), is zero, and the prior on the null value is \( \pi_0 \)). Whereas \( \pi_0 = 1 - \pi_0 \), the prior probability on
the existence of an association, $W$ describes the size of the strength of the association, conditional on the existence of one.

Appendix B shows that, with (i) and (ii), the Bayes factor $p(y|H_o)/p(y|H_1)$ is replaced by $p(\hat{\theta}|H_0)/p(\hat{\theta}|H_1)$, and we obtain the approximate Bayes factor

$$ABF = \frac{p(\hat{\theta}|H_0)}{p(\hat{\theta}|H_1)} = \sqrt{\frac{V + W}{V}} \exp \left[ -\frac{\hat{\theta}^2}{2} \times \frac{W}{V(V + W)} \right]$$

$$= \frac{1}{1 - r} \exp \left( \frac{Z^2}{2} \right),$$

where $Z = \hat{\theta}/\sqrt{V}$ is the usual $Z$ statistic and the shrinkage factor

$$r = \frac{W}{V + W}$$

is the ratio of the prior variance to the total variance. It is important to note that the Bayes factor depends not only on $Z$ but also on the power through $V$, which itself depends on the minor-allele frequency (MAF) and on the size of the effect. We define the BFDP as follows:

$$\text{BFDP} = \frac{\text{ABF} \times \text{PO}}{\text{ABF} \times \text{PO} + 1}.$$

The name “BFDP” reflects the fact that, if we report an association as noteworthy, this is the probability of the null, and, therefore, the probability of a false discovery. The “Bayesian” label emphasizes that we are taking a model-based approach, so the probability is conditional on the model and, in particular, on the assumed $\pi_v$. Given that BFDP is an approximation of $\text{Pr}(H_o|y)$, we should report the association as noteworthy if expression (4) is satisfied.

In addition to reporting BFDP, it is also informative to give point and interval estimates for $\theta$. Under the model given above, and given an association, the approximate posterior for the relative risk is given by

$$e^{\theta} | \hat{\theta} \sim \text{lognormal}(\hat{\theta}, rV),$$

so that the posterior median is $\exp(\hat{\theta})$. Hence, the maximum-likelihood estimator (MLE) $\exp(\hat{\theta})$ undergoes shrinkage toward the prior mean of 1, with the amount of shrinkage given by equation (6). As the sample size increases, $V \to 0$ and $r \to 1$, so that the posterior concentrates around the MLE. Shrinkage is desirable in situations in which the power is low (i.e., $V$ is large), since it reduces the reporting of chance associations based on few data. Hierarchical models also produce this behavior and have been advocated in an epidemiological context.13,14 Under expression (7), a 100(1 - $\alpha$)% credible interval for the relative risk is given by $\exp[\hat{\theta} \pm \Phi^{-1}(\alpha/2), \sqrt{V}]$, where $\Phi(\cdot)$ is the distribution function of a standard normal random variable.

Prior Choice

Recall that we assume the prior $\theta \sim N(0, W)$. It is relatively straightforward to pick the prior variance $W$ under the alternative by specifying that the relative risk $e^{\theta}$ lies in $[1/e_0, e_1]$ with probability $1 - \epsilon$. This leads to

$$W = \left[ \frac{\theta_0}{\Phi^{-1}(1 - \frac{\epsilon}{2})} \right]^2.$$

For example, we might assume that, with probability 0.95, the relative risk lies between 2/3 and 3/2; this corresponds to the choices $\epsilon = 0.05$ and $\theta_0 = \log(1.5)$, which give, via equation (8), $W = 0.21^2$. In practice, we can examine the sensitivity to $W$ by considering a range of values. We should be wary of setting $W$ to be very large, however, since $BF \to \infty$ as $W \to \infty$, so that the null is chosen for very large values of the prior variance.11 In practice, this is not an issue in the context considered here, since one can choose sensible values as an upper bound for the relative risk a priori.

In the context of a genomewide association study, we may wish to specify the prior variance (and hence the expected effect size) conditional on the MAF associated with the SNP whose relative risk we are estimating, although we do not pursue such a specification here. The allelic spectrum has been discussed by a number of authors.6

A graphical representation of BFDP is provided in figure 1a. In this hypothetical situation, we suppose that a logistic regression has produced an estimate of $\hat{\theta} = \log(1.316)$, along with associated SE of $\sqrt{V} = 0.102$ (the latter depending on the estimate and on the MAF), and we assume a prior variance of $W = [\log(2)/1.96]^2 = 0.35^2$. These values result in the two normal curves in figure 1, $N(0, 0.10^2)$ under $H_0$ and $N(0, 0.10^2 + 0.35^2)$ under $H_1$; $p(\hat{\theta}|H_0)$ and $p(\hat{\theta}|H_1)$ are indicated, and the ratio of these values provides an approximate Bayes factor of 0.11, so $\theta$ is 1/0.11 = 9 times more likely under $H_1$ than under $H_0$.

Design by Use of BFDP

We now describe how BFDP may be used to assess whether a study is likely to provide a noteworthy association. Recall that the data are entered into the BFDP via the approximate Bayes factor, which is a simple function of $\hat{\theta}$ and $V$, so we need only these quantities to calculate BFDP, along with the prior specifications $W$ and $\pi_v$. For illustration, we consider a dominant model with $h$, the probability of being heterozygous or homozygous for the mutant allele (we refer to this as “exposed”); $g_o$, the probability of disease given two copies of the wild-type (referred to as “unexposed”); and $g_1 = g_o \times e^{\theta}$, the probability of disease given one or two mutant alleles. Under this scenario, the expected frequencies by exposure and disease status are

$$r_{oo} = \text{Pr}(\text{unexposed}|\text{control}) = \frac{(1 - g_o)(1 - h)}{(1 - g_o)(1 - h) + (1 - g_o)h},$$

$$r_{o1} = \text{Pr}(\text{unexposed}|\text{case}) = \frac{g_o(1 - h)}{g_o(1 - h) + g_o'h},$$

where $r_{oo}$ and $r_{o1}$ are the expected frequencies by exposure and disease status.
Figure 1. Graphical representation of BFDP (a) and FPRP (b). For BFDP, the approximate Bayes factor is the ratio of indicated densities, whereas, for FPRP, the dark- and light-shaded areas represent $\hat{a}$ and $1 - \hat{\beta}$, respectively.

\[ r_{10} = \Pr(\text{exposed} \mid \text{control}) = \frac{(1 - g_0)h}{(1 - g_0)(1 - h) + (1 - g_0)h}, \]

and

\[ r_{11} = \Pr(\text{exposed} \mid \text{case}) = \frac{g_1h}{g_0(1 - h) + g_1h}, \]

and, asymptotically, we have

\[ V = \frac{1}{n_0} \left( \frac{1}{r_{10}} + \frac{1}{r_{11}} \right) + \frac{1}{n_1} \left( \frac{1}{r_{10}} + \frac{1}{r_{11}} \right), \]

where $n_0$ and $n_1$ are the numbers of controls and cases. For any particular set of design parameters, there is a distribution of BFDP, since the observed numbers of exposed and unexposed in the case and control groups vary across simulations. To illustrate, figure 2 shows the distribution of BFDP for data simulated with $\pi_0 = 0.99$, $\theta = \log(1.5)$, $n_0 = n_1 = 1,000$, $g_0 = 0.001$, and $W = [\log(2)/1.96]^2$, for $h$ of 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. For these frequencies, the Bayesian “powers” of achieving a noteworthy BFDP at a level of 0.8 (which corresponds to a false non-discovery being four times as costly as a false discovery) are 0%, 13%, 42%, 76%, 88%, 92%, and 92%, respectively.

We next calculate the sample size required to obtain a BFDP <0.8, with probability 0.8 (the Bayesian power)—that is,

\[ \Pr, \left[ \Pr \left( H_0 \mid \hat{\theta}(\mathbf{y}) \right) \leq 0.8 \right] \geq 0.8. \] (9)
Figure 2. Distribution of BFDP by allele frequency for 1,000 cases and 1,000 controls with a relative risk of 1.5 and with \( p = 0.99 \). The solid line connects the median BFDP at each frequency. The dotted line corresponds to a noteworthy threshold of 0.8; the powers to achieve this level (i.e., for the BFDP to fall below this threshold) at the given frequencies are 0%, 13%, 42%, 76%, 88%, 92%, and 92%.

Figure 3. Number of cases required to obtain a BFDP of <0.8 with probability 0.8, as a function of the frequency of one or two mutant allele copies and \( \theta \), with a relative risk of 1.5 (under the assumption of an equal number of cases and controls).

is equal to equation (10). Writing

\[
\text{FPRP} = \frac{\hat{a}_{1-\beta} \times \pi_0}{\hat{a}_{1-\beta} \times \pi_0 + 1} = \frac{\text{FF} \times \text{PO}}{\text{FF} \times \text{PO} + 1}
\]

we see that the “frequentist factor” \( \text{FF} = \hat{a}_{1-\beta} \) is taking the role of the Bayes factor. An alternative derivation of FPRP occurs if we take \( H_0; \theta = 0 \) versus \( H_1; \theta \neq 0 \) but assume a point mass prior at \( \theta = \theta_i \). In this case, \( \Pr(T|H_i) = \Pr(T|\theta_i) \), and we again obtain equation (10).

Figure 1b graphically illustrates the calculation of FPRP for the same \( \theta \) and \( \nu \) used for the calculation of the approximate Bayes factor. The power is calculated at \( \theta_i = \log(1.5) \), and \( \hat{a} \) and \( 1 - \beta \) are shown as the dark- and light-shaded areas, respectively. In this example, \( \hat{a}/(1 - \beta) \), so that the data, \( T \), are 150 times more likely under the alternative than under the null, a much stronger conclusion than that reached using the approximate Bayes factor.

Although FPRP is a Bayesian procedure, when compared with BFDP, it differs in each of the likelihood, prior, and decision-rule choices.

The likelihood.—Information is being lost by considering \( T = \)

Insight into FPRP may be gained by considering a formal Bayesian approach. Assume that the “data” consist of \( T \) as defined above, and consider the point null and alternative hypotheses \( H_0; \theta = 0 \) and \( H_1; \theta = \theta_i \). Then,

\[
\text{FPRP} = \Pr(H_0 | T) = \frac{\Pr(T | H_0) \pi_0}{\Pr(T | H_0) \pi_0 + \Pr(T | H_1)(1 - \pi_0)}
\]
\[ |\hat{\theta}| > \hat{\theta} \text{ (a censored observation), rather than \hat{\theta} itself. Given that we are using the observed } P \text{ value, this does not seem appropriate, since the observation is not larger than a particular value, it is exactly that value. An intermediary approach between FPRP and BFDP, which considers only a single alternative but does not censor the observation, would be to use}

\[
Pr(H_0|\hat{\theta}) = \frac{LR \times PO}{1 + LR \times PO},
\]

where

\[
LR = \frac{\phi[\hat{\theta}]}{\phi[\hat{\theta} - \theta_1]},
\]

is the likelihood ratio and \( \phi(\cdot) \) is the density function of a standard normal random variable. Note that

\[
Pr(H_0|\hat{\theta}) > \hat{\theta} < Pr(H_0|\hat{\theta}), \quad (11)
\]

so that taking the censored observation gives a lower bound on the posterior probability of the null associated with the observed value of the estimate. In their appendix, Wacholder et al. state, “the FPRP value is the lowest FPRP value at which a test would yield a noteworthy finding.” Although expression (11) would suggest that FPRP < BFDP, the inequality in (11) is for a fixed prior, and it is possible for FPRP to be larger than BFDP if \( \theta_1 \) is close to 1 (since, then, the power is low and FPRP is relatively larger). Finally, the use of a two-sided \( P \) value is not consistent with the evaluation of power at a point (one-sided) alternative.

The prior.—Assuming that \( \theta = \theta_1 \), as a prior is overly restrictive, BFDP averages over all alternatives \( \theta \neq 0 \).

The decision rule.—The earlier discussion of decision theory reveals that it is helpful to base the definition of “noteworthiness” on the costs of false discovery and nondiscovery. For FPRP, the decision is with respect to the specific alternative, \( \theta_1 \), which is more difficult to assess than the cost for the general alternative. Of the values suggested by Wacholder et al., \( \theta \) is 0.2, which corresponds to the belief that the cost of a false declaration that \( \theta = \theta_1 \) is four times larger than the cost of calling nonnoteworthy a true association of strength \( \theta = \theta_1 \). This will often not reflect the aims of a genomewide association study because, as noted, we would rather have a longer list than a shorter list of SNPs. For a candidate-gene study, a lower threshold is more plausible.

### Multiple-Hypothesis Testing

We now consider how BFDP may be used for multiple-hypothesis testing, a situation not explicitly considered by Wacholder et al.\(^{23}\) Table 2 gives the possible outcomes when \( m \) tests are performed; \( m_0 \) is the true number of null hypotheses, and \( k \) the number of tests that are classified as noteworthy. There is now a large body of literature on how to perform multiple tests.\(^{9,10,24-28}\) It has been recognized that estimating and characterizing procedures according to the expected numbers of false discoveries and nondiscoveries (given by \( E[V] \) and \( E[T] \) in table 2) or their rates (\( E[V/k] \) and \( E[T/(m - k)] \)) is far more relevant than controlling familywise error rates, as is done, for example, by the Bonferroni method (which controls the probability that \( V \geq 1 \)). The latter produces a relatively small \( k \) and is therefore conservative, which can lead to substantial power loss and the missing of many true associations.

Appendix C shows that, if \( C_\beta/C_\alpha \) describes the common ratio of costs of false nondiscovery to costs of false discovery across tests, then we should report associations for which

\[
Pr(H_0|\hat{\theta}) < \frac{C_\beta/C_\alpha}{1 + C_\beta/C_\alpha},
\]

so that, as with a single test, we have an intuitive rule that is defined in terms of the posterior probability of the null. Let \( k \) be the number of such associations. We rank the associations in order of increasing BFDP, \( Pr(H_0|\hat{\theta}) \), \( j = 1, \ldots, m \), so that \( j = 1, \ldots, k \) represent the noteworthy tests. The expected numbers of true nonnoteworthy, false noteworthy, false nonnoteworthy, and true noteworthy tests are given in table 3. The expected number of false discoveries is the sum of the null probabilities over those hypotheses we report, whereas the expected number of false nondiscoveries is the sum of the probabilities on the alternative over the nonnoteworthy tests. The expected number of false noteworthy tests divided by the number of noteworthy tests has been proposed\(^{22}\) in conjunction with FPRP, to “control the Bayes FDR”; one can evaluate this quantity with any Bayesian approach, however, not just with the specific likelihood and prior setup considered in FPRP.

#### q Values

Recently, there has been a great deal of interest in the \( q \)-value method,\(^{39}\) which, in a setting of multiple-hypothesis testing and for a fixed critical region, \( \Gamma \), provides an estimate of the FDR (or, more precisely, the positive FDR, which is the FDR conditioned on at least one noteworthy test). The \( q \) value corresponding to \( \Gamma \) is

\[
Pr(H_0|\hat{\theta} \in \Gamma) = \frac{\alpha(\Gamma)p_0}{\alpha(\Gamma)p_0 + [1 - \beta(\Gamma)](1 - p_0)} = \alpha(\Gamma) \times \frac{p_0}{Pr(\hat{\theta} \in \Gamma)}, \quad (12)
\]

with the term \( \alpha(\Gamma)p_0/Pr(\hat{\theta} \in \Gamma) \) estimated from the totality of \( P \) values. If all hypotheses are null, then the distribution of \( P \) values is uniform, so departures from uniformity inform on the fraction of nulls, \( p_0 \), whereas the complete distribution gives an estimate of the denominator in (12).

For a given region, \( \Gamma \), the average FDR is controlled at level \( Pr(H_0|\hat{\theta} \in \Gamma) \). Suppose a SNP has a \( q \) value of \( q_\beta \) this does not
mean that this SNP is false positive with probability \( q_0 \). The false-positive probability for this SNP could be much higher, because \( q_0 \) is the average proportion of false-positive SNPs that would occur if we call the SNP noteworthy, and this collection contains SNPs that are more noteworthy. The ranking of \( P \) and \( q \) values is identical, since the denominators of all tests are equal (unlike BFDP and FPRP, which explicitly use the powers of each test in the calculation of their respective denominators).

In microarray experiments (to which \( q \) values have been extensively applied), the empirical estimation of \( \pi_0 \) is reliable because a large number of tests are nonnull. Estimation is more difficult in genomewide association studies because the true number of associations is a tiny fraction of the total number of tests performed. It may also be argued that the false-nondiscovery rate is more relevant for genomewide association studies, since missing a true association is more costly than making a false discovery; however, the false-nondiscovery rate still requires an estimate of \( \pi_0 \).

Summary of FPRP

FPRP is a Bayesian procedure that takes as data the observed tail area and assumes a point prior under the alternative. FPRP ignores information by conditioning on a tail area and is difficult to calibrate from a Bayesian perspective, because the ratio of costs is with respect to a specific alternative. Because FPRP uses critical regions defined by the observed \( P \) values, the regions are not constant across tests, so frequentist FDR is not controlled; such control is provided by the \( q \)-value method.

Results

Empirical Comparison of FPRP and BFDP

We consider a hypothetical, genomewide association case-control study in which the association between disease and \( m = 100,000 \) SNPs, with alleles \( A \) and \( a \) at each candidate marker, is evaluated for 3,000 cases and 3,000 controls. To simulate data, we assume that the disease risk, \( p \), is given by the logistic regression model

\[
\logit p = \alpha + \beta o ,
\]

where \( e^\beta = 0.002 \) is the baseline risk and \( z = 0 \), 0.5, and 1 corresponds to 0, 1, and 2 copies of the mutant allele, respectively. Hence, we assume an additive genetic model, with \( e^\beta \) corresponding to the relative risk associated with two copies of the mutant allele. We take penetrances, from equation (13), to be given by

\[
f_0 = \Pr(\text{case}|aa) = 0.002 ,
\]

\[
f_i = \Pr(\text{case}|Aa) = e^{0.2} \times f_0 ,
\]

\[
f_s = \Pr(\text{case}|AA) = e^\beta \times f_0 ,
\]

and then evaluate the probabilities of \( aa, Aa \), and \( AA \), given case and control status, by use of Bayes’s theorem. Across all \( m \) SNPs, we randomly generate the MAF from a uniform distribution on \( [0.05, 0.50] \). We assume that \( m - m_0 = 100 \) SNPs are associated with disease and generate the log relative risks from a beta distribution with parameters 1 and 3, scaled to lie between log(1.1) and log(1.5); hence, we have an L-shaped distribution of effect sizes, with the relative risks more likely to be closer to 1.1 than to 1.5. The remaining \( m_0 = 99,900 \) SNPs are not associated with disease. We take \( \theta_i = \log(1.5) \) for FPRP and \( W = \log(1.5)/1.96 \) for BFDP.

A major conclusion from this simulation is that, even with 3,000 cases and 3,000 controls, there is very low power for detection of SNPs with low MAFs and/or log relative risks close to 0. With ratios of cost of false nondiscovery to cost of false discovery of 4:1, 10:1, 20:1, and 50:1, the BFDP thresholds of noteworthiness are 0.80, 0.91, 0.95, and 0.98 and yield only 9, 11, 15, and 20 true associations, respectively. Figure 4 plots \( \theta \) against MAF for the 100 true associations for these four thresholds and indicates the noteworthy and nonnoteworthy SNPs (blackened and nonblackened circles, respectively). The inability to detect SNPs with low MAF and/or relative risks close to 1 is apparent.

As discussed above, FPRP and BFDP are not directly comparable; so, rather than present side-by-side results, we discuss each in turn. Figure 5 gives the numbers of true nonnoteworthy, false noteworthy, false nonnoteworthy, and true noteworthy tests (i.e., the quantities \( U, V, T, \) and \( S \) in table 2) for BFDP as a function of the chosen threshold. The expected posterior estimates of each of these quantities (calculated using the formulas in table 3) are displayed as dashed lines, and the true numbers of null and nonnull associations appear as dotted lines. The problem with nondiscoveries is apparent in figure 5d—even for thresholds very close to 1, the majority of true nonnull associations go undetected. Figure 5b shows that the number of false discoveries increases dramatically as the threshold approaches 1. The only comfort from this simulation is that the expected numbers of true and false noteworthy and nonnoteworthy tests are very accurate, so at least we have some indication of the reliability of the results. A caveat, however, is that these numbers were calculated using the true \( \pi_0 \), and are highly sensitive to this value. The prior on \( \theta \) was \( N(0,W) \), however, which is quite different from the distribution from which the effects of the nonnull SNPs were generated.

---

Table 3. Expected Numbers of Tests That Are True or False and Nonnoteworthy or Noteworthy When \( m \) Tests Are Performed and \( k \) Tests Are Called Noteworthy

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Nonnoteworthy ((m - k))</th>
<th>Noteworthy ((k))</th>
<th>Total ((m))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 )</td>
<td>( \text{Expected True} ) &amp; ( \text{Expected False} ) &amp; ( m_0 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H_1 )</td>
<td>( \sum_{j=1}^{m} \Pr(H_0</td>
<td>\hat{\theta}<em>j) ) &amp; ( \sum</em>{j=1}^{m} \Pr(H_1</td>
<td>\hat{\theta}_j) ) &amp; ( m - m_0 )</td>
</tr>
</tbody>
</table>

---

29. [Add note or reference here if needed.]
Figure 4. Log relative risk versus MAF for the 100 nonnull SNPs in the simulated data. Blackened and nonblackened circles represent SNPs called noteworthy and nonnoteworthy, respectively, by BFDP at the stated threshold.
Figure 5. Operating characteristics of BFDP. Panels a, b, c, and d correspond to possibilities U, V, T, and S in table 2, respectively. In each panel, the solid lines represent the numbers of true nonnoteworthy, false nonnoteworthy, false nonnoteworthy, and true noteworthy tests at each threshold for BFDP. The dashed lines are the posterior expected numbers of tests that are true nonnoteworthy, false nonnoteworthy, false nonnoteworthy, and true noteworthy. The dotted lines represent the ideal outcome that a perfect test would achieve.
The results for FPRP are displayed in figure 6, on the same scale as figure 5. The overall behavior of BFDP and FPRP is the same, as can be seen in figure 7a, in which we see that the rankings are generally quite similar (with a few exceptions), but FPRP produces posterior null estimates that are much smaller than those produced by BFDP. This is because, from expression (11), FPRP is a lower bound on the posterior probability corresponding to the observed estimates. This inequality is for a fixed prior, and the priors for FPRP and BFDP are different, but the dominant difference between the two approaches for the priors chosen here is because of conditioning on point estimates (for BFDP) versus conditioning on tail areas (for FPRP). The estimates of the expected numbers of true and false noteworthy and nonnoteworthy tests under FPRP (the dashed lines) are not useful. In the first row of figure 6, noteworthy and nonnoteworthy tests under FPRP (the dashed lines) fall beneath the solid lines. Similarly, in the second row, the probabilities of \( H_0 \) are bounded above. In both cases, the rankings are generally quite similar (with a few exceptions), but FPRP produces posterior null estimates that are much smaller than those produced by BFDP. This is because, from expression (11), FPRP is a lower bound on the posterior probability corresponding to the observed estimates. This inequality is for a fixed prior, and the priors for FPRP and BFDP are different, but the dominant difference between the two approaches for the priors chosen here is because of conditioning on point estimates (for BFDP) versus conditioning on tail areas (for FPRP). The estimates of the expected numbers of true and false noteworthy and nonnoteworthy tests under FPRP (the dashed lines) are not useful. In the first row of figure 6, noteworthy and nonnoteworthy tests under FPRP (the dashed lines) fall beneath the solid lines. Similarly, in the second row, the probabilities of \( H_0 \) are bounded above. In each case, the bounds are not tight and so are not practically useful.

Using the \( P \) values from the \( m = 100,000 \) tests, we implement the \( q \)-value approach that controls FDR; \( q_0 \) was estimated as 1, reflecting the difficulty in estimating a proportion very close to 1. Table 4 gives the numbers of true and false nonnoteworthy and noteworthy results as a function of the selected FDR level. The final column shows that the procedure does control FDR, although, as with BFDP and FPRP, the number of missed true associations is large. Figure 7d plots the \( q \) values versus BFDP; the latter are always larger, which reflects expression (11) to a large extent. \( q \) values are a lower bound because they are evaluated by conditioning on a tail area. Table 5 provides a tabulation of BFDP values, although we stress that BFDP and \( q \) values are not directly comparable because they are packaging the information in the totality of tests in a different manner. BFDP does control FDR here, however.

Figure 7c and 7d plots the \( q \) values (which have the same rankings as the \( P \) values) against FPRP and BFDP, respectively, and we see that the rankings can be different because the power of each test is not used in the \( q \)-value calculation. For example, in figure 7c, there are two SNPs (highlighted with boxes) with \( q \) values of 0.26 but with quite different FPRP values of 0.02 and 0.21, corresponding to powers of 0.90 and 0.07, respectively. A striking example of this phenomenon in a candidate-gene study is presented in the next section.

### Lung Cancer Association Study

We now present an illustration of the calculation of BFDP in the context of a multicenter, case-control study with 2,250 lung cancer cases and 2,899 controls and examine the association with 131 SNPs. This study is described more fully by Hung et al.\(^{17}\); the results for the study are not yet published, so we do not reveal the SNP names here. To calculate BFDP, we take \( W = \log^2 \) and, for FPRP, \( \theta_0 = 1.5 \), with \( q_0 = 0.98 \) for both BFDP and FPRP, reflecting that, in this candidate-gene study, we believe that two or three SNPs might be associated with lung cancer. Under both BFDP and FPRP, the priors are such that we are implicitly assuming that the SNPs are independent (and, in particular, are not in linkage disequilibrium). In practice, the effect of ignoring the dependence will be a loss of efficiency in estimation. One approach to overcoming this problem is to specify a hierarchical model\(^{30}\), although this causes loss of the ease of implementation of BFDP.

We assume an additive genetic model; it would be straightforward to repeat this experiment with dominant or recessive genetic models or with a nonadditive model. Appendix D gives details of how BFDP may be calculated in the situation in which there are two or more relative risks. We fit 131 logistic-regression models, controlling for age, sex, cigarette-pack years, and country, and retain for the calculation of BFDP the estimates and SEs, \( \hat{\theta}_j, \hat{\gamma}_j \), \( j = 1, \ldots, 131 \).

We assume that the cost of a false nondiscovery is three times as great as the cost of a nondiscovery, which gives a cutoff value for BFDP of 0.75. We list the six most likely SNP associations under BFDP in table 6, along with a number of additional summaries. Use of the 0.75 cutoff gives one noteworthy SNP under BFDP, with \( \Pr(H_0|\hat{\theta}) = 0.80 \).

### Table 4. Summary of \( q \) Values for Simulated Data

<table>
<thead>
<tr>
<th>( q )</th>
<th>( k )</th>
<th>True and Nonnoteworthy</th>
<th>False and Nonnoteworthy</th>
<th>False and Noteworthy</th>
<th>True and Noteworthy</th>
<th>Empirical FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>2</td>
<td>99,900</td>
<td>0</td>
<td>98</td>
<td>2</td>
<td>.00</td>
</tr>
<tr>
<td>.10</td>
<td>4</td>
<td>99,900</td>
<td>0</td>
<td>96</td>
<td>4</td>
<td>.00</td>
</tr>
<tr>
<td>.20</td>
<td>4</td>
<td>99,900</td>
<td>0</td>
<td>96</td>
<td>4</td>
<td>.00</td>
</tr>
<tr>
<td>.30</td>
<td>8</td>
<td>99,898</td>
<td>2</td>
<td>94</td>
<td>6</td>
<td>.25</td>
</tr>
<tr>
<td>.40</td>
<td>8</td>
<td>99,898</td>
<td>2</td>
<td>94</td>
<td>6</td>
<td>.25</td>
</tr>
<tr>
<td>.50</td>
<td>8</td>
<td>99,898</td>
<td>2</td>
<td>94</td>
<td>6</td>
<td>.25</td>
</tr>
<tr>
<td>.60</td>
<td>23</td>
<td>99,886</td>
<td>14</td>
<td>91</td>
<td>9</td>
<td>.61</td>
</tr>
<tr>
<td>.70</td>
<td>28</td>
<td>99,882</td>
<td>18</td>
<td>90</td>
<td>10</td>
<td>.64</td>
</tr>
<tr>
<td>.80</td>
<td>37</td>
<td>99,874</td>
<td>26</td>
<td>89</td>
<td>11</td>
<td>.70</td>
</tr>
<tr>
<td>.90</td>
<td>160</td>
<td>99,760</td>
<td>140</td>
<td>80</td>
<td>20</td>
<td>.88</td>
</tr>
<tr>
<td>.95</td>
<td>199</td>
<td>99,721</td>
<td>179</td>
<td>80</td>
<td>20</td>
<td>.90</td>
</tr>
</tbody>
</table>

**Note.**—Columns 3, 4, 5, and 6 correspond to possibilities \( U, V, T, \) and \( S \) in table 2, respectively.
Figure 6. Operating characteristics of FPRP. Panels a, b, c, and d correspond to possibilities U, V, T, and S in Table 2, respectively. In each panel, the solid lines represent the numbers of true nonnoteworthy, false nonnoteworthy, false nonnoteworthy, and true noteworthy tests at each threshold for FPRP. The dashed lines are the posterior expected numbers of tests that are true nonnoteworthy, false nonnoteworthy, false nonnoteworthy, and true noteworthy. The dotted lines represent the ideal outcome that a perfect test would achieve.
Figure 7. Comparison between BFDP, FPRP, $P$, and $q$ for the 5,000 SNPs with the lowest values of BFDP. 

- **a**, BFDP versus FPRP. 
- **b**, $P$ values versus FPRP. 
- **c**, $q$ values versus FPRP. 
- **d**, $q$ values versus BFDP.
Table 5. Summary of BFDP Values for Simulated Data

<table>
<thead>
<tr>
<th>BFDP</th>
<th>True and Nonnoteworthy</th>
<th>False and Noteworthy</th>
<th>False and Nonnoteworthy</th>
<th>True and Noteworthy</th>
<th>Empirical FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>2</td>
<td>99,900</td>
<td>0</td>
<td>98</td>
<td>2</td>
</tr>
<tr>
<td>.10</td>
<td>2</td>
<td>99,900</td>
<td>0</td>
<td>98</td>
<td>2</td>
</tr>
<tr>
<td>.20</td>
<td>3</td>
<td>99,900</td>
<td>0</td>
<td>97</td>
<td>3</td>
</tr>
<tr>
<td>.30</td>
<td>3</td>
<td>99,900</td>
<td>0</td>
<td>97</td>
<td>3</td>
</tr>
<tr>
<td>.40</td>
<td>4</td>
<td>99,900</td>
<td>0</td>
<td>96</td>
<td>4</td>
</tr>
<tr>
<td>.50</td>
<td>7</td>
<td>99,899</td>
<td>1</td>
<td>94</td>
<td>6</td>
</tr>
<tr>
<td>.60</td>
<td>7</td>
<td>99,899</td>
<td>1</td>
<td>94</td>
<td>6</td>
</tr>
<tr>
<td>.70</td>
<td>8</td>
<td>99,888</td>
<td>2</td>
<td>94</td>
<td>6</td>
</tr>
<tr>
<td>.80</td>
<td>16</td>
<td>99,883</td>
<td>7</td>
<td>91</td>
<td>9</td>
</tr>
<tr>
<td>.90</td>
<td>32</td>
<td>99,873</td>
<td>21</td>
<td>89</td>
<td>11</td>
</tr>
<tr>
<td>.95</td>
<td>66</td>
<td>99,848</td>
<td>52</td>
<td>86</td>
<td>14</td>
</tr>
</tbody>
</table>

Note.—Columns 3, 4, 5, and 6 correspond to possibilities U, V, T, and S in table 2, respectively.

In this article, we have suggested a new measure for identifying noteworthy associations, BFDP, keeping the objectives of but refining the criteria for FPRP in the work of Wacholder et al.² Specifically, we advocate BFDP for reducing the number of “discoveries” that are reported but not replicated in subsequent investigations. Whereas FPRP is difficult to calibrate, a threshold for BFDP may be chosen that explicitly considers the costs of false discovery and false nondiscovery.

BFDP can be applied using an estimate and SE or with a confidence interval. Published articles that report such summaries can therefore be critically interpreted. An Excel spreadsheet and an R function for calculation of BFDP are available online at J.W.’s Web site.

We have developed BFDP, using the model-based asymptotic distribution of the MLE, but we can replace V with any estimate that is thought to be appropriate—for example, a sandwich estimator or one that allows for over-dispersion to account for population heterogeneity.²⁹
Boxplots of BFDP (a–c) and FPRP (d–f). $\pi_0$ values are shown above each panel. For panels a–c, each set of boxplots corresponds to $\theta_p = 1.50, 1.88, 2.25, 2.63, \text{ and } 3.00$ (the 97.5% point of the prior); for panels d–f, we evaluate the power at $\theta_1 = 1.50, 1.88, 2.25, 2.63, \text{ and } 3.00$. The dashed lines in panels a–c denote the thresholds that correspond to false nondiscovery being three times as costly as false discovery, so that SNPs with BFDP values below the line are noteworthy.
Within a logistic-regression framework, observed race/ethnicity may be explicitly modeled with additional parameters to control for heterogeneity; simulations have demonstrated that logistic regression can control population substructure. We have concentrated on SNPs, but BFDP may also be applied to insertions/deletions or copy-number changes.

Different studies that report the required summaries can be combined to produce the totality of evidence of a particular association. For example, for two studies with estimates \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \):

\[
\Pr(H_0 | \hat{\theta}_1, \hat{\theta}_2) = \frac{\mathrm{ABF}(\hat{\theta}_1, \hat{\theta}_2) \times \pi_\theta(1 - \pi_\theta)}{\mathrm{ABF}(\hat{\theta}_1, \hat{\theta}_2) \times \pi_\theta(1 - \pi_\theta) + 1}, \tag{14}
\]

where \( \mathrm{ABF}(\hat{\theta}_1, \hat{\theta}_2) = \mathrm{ABF}(\hat{\theta}_1) \times \mathrm{ABF}(\hat{\theta}_2 | \hat{\theta}_1) \),

\[
\mathrm{ABF}(\hat{\theta}_2 | \hat{\theta}_1) = \frac{p(\hat{\theta}_2 | H_1)}{p(\hat{\theta}_2 | H_0)}.
\]

and \( p(\hat{\theta}_2 | H_0, \hat{\theta}_1) = E_{\theta | \hat{\theta}_1}[p(\hat{\theta}_2 | \theta)] \) is the density for \( \hat{\theta}_2 \) averaged over the posterior for \( \theta \) given \( \hat{\theta}_1 \) (and is available in closed form). Expression (14) may be used for replication studies. In such a context, even if two studies provide small Bayes factors (and therefore strong evidence of an association), this must still be combined with the prior odds on the null, \( \pi_\theta/(1 - \pi_\theta) \); if the latter is large, then the overall evidence may still be inconclusive.

It is becoming increasingly common to perform genomewide scans in multiple stages. When specifying the costs may change across stages, since, early on, we do not wish to lose SNPs that might be associated with disease, whereas, in the final stage, a more stringent cutoff is desirable. It is straightforward to perform such a procedure with BFDP. Currently, it is common to rank \( P \) values and then select a set of SNPs for the next phase on the basis of the smallest \( P \) values. We would advocate ranking via the approximate Bayes factor instead. This will, in general, provide a different ordering, since the powers are not constant across SNPs (because they depend on the allele frequencies and the strength of the association), which is not accounted for by the \( P \) value. In the lung cancer example, we saw a situation in which the smallest \( P \) value did not correspond to a noteworthy SNP.

If one is willing to relax ease of computation in the multiple-testing scenario, then one may model the totality of estimates \( \hat{\theta}_i, i = 1, \ldots, m \), as arising from a mixture of two distributions. The proportion of null tests, \( \pi_\omega \), may be estimated from the data, as may the densities of null and nonnull estimates, although, with a small expected number of nonnull associations, estimation is likely to be sensitive to the prior for \( \pi_\omega \). A number of models and implementation strategies have been suggested for simultaneous inference using data from multiple tests. Two of the easiest to implement are empirical Bayes and full Bayes with importance sampling. Each of these procedures may be used in the context considered here, by replacing the likelihood with the asymptotic distribution of the MLE. A great advantage in using the asymptotic distribution is that it results in a model that is very conducive to analytic examination and provides straightforward computation. In the lung cancer example, we did not attempt hierarchical modeling because the number of SNPs was relatively small and the a priori fraction of nonnull associations was also small, so that reliable estimation of the nonnull density would not be feasible.

**Acknowledgments**

I thank Rayjean Hung, for the use of the lung cancer data, and two referees, for constructive comments. I also thank James McKay and Martyn Plummer for many helpful discussions.

**Appendix A**

**Minimizing Expected Loss**

With respect to table 1, the posterior expected cost associated with decision \( \delta \) is

\[
E[C(\delta, H)] = C(\delta, H_0) \Pr(H_0 | y) + C(\delta, H_1) \Pr(H_1 | y),
\]

so that, for the two possible decisions (nonnoteworthy or noteworthy), the expected costs are

\[
E[C(\delta = 0, H)] = 0 \times \Pr(H_0 | y) + C_\delta \times \Pr(H_1 | y)
\]

\[
E[C(\delta = 1, H)] = C_\alpha \times \Pr(H_1 | y) + 0 \times \Pr(H_1 | y),
\]

and we should choose \( \delta = 1 \) if \( C_\delta \times \Pr(H_1 | y) \geq C_\alpha \times \Pr(H_0 | y) \)—that is, if

\[
\Pr(H_1 | y) \geq \frac{C_\alpha}{C_\alpha + C_\delta} = \frac{1}{1 + \frac{C_\delta}{C_\alpha}}.
\]
Appendix B

The Approximate Bayes Factor

We derive an approximation of the Bayes factor $Pr(y|H_0)/Pr(y|H_1)$ for logistic-regression model (5) and $n$ individuals. The likelihood under $H_1$ is given by

$$Pr(y|\gamma, \theta) = \prod_{i=1}^{n} Pr(y_i|\gamma, \theta) = \prod_{i=1}^{n} p_i(1 - p_i)^{1 - \gamma},$$

whereas, under $H_0$, we set $\theta = 0$. We have

$$Pr(y|H_1) = \int_{\gamma} \int_{\theta} \prod_{i=1}^{n} p_i(\gamma, \theta)^{\gamma}[1 - p_i(\gamma, \theta)]^{1 - \gamma} \pi(\gamma, \theta) d\gamma d\theta \quad \text{(B1)}$$

and

$$Pr(y|H_0) = \int_{\gamma} \prod_{i=1}^{n} p_i(\gamma, \theta = 0)^{\gamma}[1 - p_i(\gamma, \theta = 0)]^{1 - \gamma} \pi(\gamma, \theta = 0) d\gamma,$$

where $\pi(\gamma, \theta)$ is the prior. The integrals in (B1) and (B2) are analytically intractable; thus, some form of approximation or simulation technique is required.

To calculate an approximation of the Bayes factor, we first replace the likelihood $Pr(y|\gamma, \theta)$ by the asymptotic distribution $p(\hat{\gamma}, \hat{\theta} | \gamma, \theta)$, where $\hat{\gamma}, \hat{\theta}$ is the MLE. This approximation will be accurate given the sample sizes in typical genomewide-association and candidate-gene studies, unless the MAF is very small. We write $\beta = (\gamma, \theta)$ and assume normal priors for $\gamma$ and $\theta$:

$$\hat{\beta} | \beta \sim N(\beta, V), \quad \beta \sim N(m, W).$$

Straightforward algebra yields the predictive distribution $\hat{\beta} | H_1 \sim N(m_v, V + W)$. A similar derivation is available under $H_0$, allowing a closed form for the Bayes factor. This approach requires specification of the joint prior $\pi(\gamma, \theta)$; we adopt a simpler strategy here.

Suppose that $\hat{\gamma}$ and $\hat{\theta}$ are independent—that is,

$$\begin{bmatrix} \hat{\gamma} \\ \hat{\theta} \end{bmatrix} \sim N_{c+1}\left(\begin{bmatrix} \gamma \\ \theta \end{bmatrix}, \begin{bmatrix} I_c & 0 \\ 0 & I_1 \end{bmatrix}\right), \quad \text{(B3)}$$

where $I_c$ is the $c \times c$ expected information associated with $\gamma$, $I_1$ is the expected information associated with $\theta$, and $0$ is a $c \times 1$ vector containing zeros. Further, suppose that the prior factorizes as $\pi(\gamma, \theta) = \pi(\gamma) \times \pi(\theta)$. Under these circumstances, it is straightforward to show that, under $H_1$,

$$p(\hat{\gamma}, \hat{\theta} | H_1) = p(\hat{\gamma} | H_1) \times p(\hat{\theta} | H_1),$$

where $p(\hat{\gamma} | H_1) = \int p(\hat{\gamma} | \gamma) \pi(\gamma) d\gamma$ and, under $H_0$,

$$p(\hat{\gamma}, \hat{\theta} | H_0) = p(\hat{\gamma} | H_0) \times p(\hat{\theta} | H_0).$$

An approximate Bayes factor is given by the ratio of these quantities and depends on only $\theta$:

$$ABF = \frac{p(\hat{\theta} | H_0)}{p(\hat{\theta} | \theta) \pi(\theta) d\theta} = \frac{p(\hat{\theta} | H_0)}{p(\theta | H_1)}. \quad \text{(B4)}$$

In genome association studies, the independence assumption in (B3) is likely to be reasonable, since genetic information will often be at least approximately independent of environmental information.

The above derivation means that, effectively, we need to consider only the sampling distribution of the MLE,
\[ \hat{\theta} | \theta \sim N(\theta, V), \] and the prior for \( \theta \sim N(0, W) \), where we have assumed the prior mean is zero. These choices result in the approximate Bayes factor (B4) being the ratio of the prior predictive densities:

\[ \hat{\theta} | H_1 \sim N(0, V + W), \hat{\theta} | H_0 \sim N(0, V), \]

to give

\[ ABF = \left( \frac{V + W}{V} \right)^{1/2} \exp \left( -\frac{\hat{\theta}^2}{2V(V+W)} \right) = \frac{1}{1-r} \exp \left( -\frac{Z^2}{2} \right), \]

where \( Z = \hat{\theta}/\sqrt{V} \) is the usual \( Z \) statistic and \( r = W/(V + W) \) is a shrinkage factor. This form is used within BFDP throughout this article.

To investigate the accuracy of the approximation, we simulated case-control data with varying numbers of cases and controls, MAFs, and relative risks. In all cases, we placed an \( N(0,1) \) prior on the intercept. The exact Bayes factor was calculated using both a rejection algorithm and importance sampling. Figure B1 displays the approximate Bayes factors plotted against the exact Bayes factors for a subset of the simulations. In each of the scenarios, 50 data sets were generated, and, in all cases, the approximation is accurate.

**Appendix C**

**Multiple Testing**

Suppose we wish to perform \( m \) tests with common costs \( C_p \) and \( C_a \) for each test. The aim is to define a rule for deciding which of the \( m \) null hypotheses we will flag as noteworthy and to determine the operating characteristics, in terms of false discovery and false nondiscovery, for this rule. The cost associated with a particular set of decisions \( \delta_1, \ldots, \delta_m \) is

\[ E[\{L(\delta_1, \ldots, \delta_m),(H_{i_1}, \ldots,H_{i_m})\}] = C_p \sum_{j=1}^{m} \delta_j \Pr(H_{i_j}|y_j) + C_a \sum_{j=1}^{m} (1 - \delta_j) \Pr(H_{i_j}|y_j), \]

where \( \delta_j = 0 \) or \( 1 \), according to whether we call test \( j \) nonnoteworthy or noteworthy. In this case, Muller et al. show that we should report association \( j \) as noteworthy if

\[ \Pr(H_{i_j}|y_j) \geq \frac{1}{1 + C_p/C_a}. \quad (C1) \]

Without loss of generality, assume the tests are ranked from smallest to largest in terms of \( \Pr(H_{i_j}|y_j), j = 1, \ldots, m \), and that the first \( k \) tests, \( 0 \leq k \leq m \), are deemed worthy of reporting as noteworthy according to rule (C1). The posterior expected numbers of true and false nonnoteworthy tests are given in table 3; note that the sum of these two quantities is \( m - k \), the number of tests not deemed noteworthy. Similarly, the expected numbers of false and true noteworthy tests are given in table 3, and the sum of these two quantities is \( k \), the number of tests called noteworthy.

**Appendix D**

**General BFDP**

Suppose we have the logistic regression model

\[ \logit p = \mathbf{x}^t \gamma + z_1 \theta_1 + z_2 \theta_2, \]

where, for example, \( z_1 \) and \( z_2 \) may be indicators of one or two copies of a mutant allele, with \( e^{x_1} \) and \( e^{x_2} \) representing the associated relative risks. A classical testing procedure would compare \( H_{i_1}: \theta_1 = \theta_2 = 0 \) and \( H_{i_2}: \theta_1 \neq 0, \theta_2 \neq 0 \), by use of a likelihood-ratio statistic with 2 df. The approximate Bayes factor is only slightly more complicated than in the single exposure case and is based on

\[ \hat{\theta} | \theta \sim N_{i_1}(\theta, V), \theta \sim N_{i_2}(0, W), \]
Figure B1. Approximate Bayes factor versus exact Bayes factor as a function of the number of controls, $n_0$; the number of cases, $n_1$; and the log relative risk, $\theta$. The MAF is 0.05, and there are 50 simulated data sets in each plot. In panel a, $n_0 = n_1 = 250$ and $\theta = \log(1.0)$. In panel b, $n_0 = n_1 = 250$ and $\theta = \log(1.3)$. In panel c, $n_0 = n_1 = 500$ and $\theta = \log(1.0)$. In panel d, $n_0 = n_1 = 500$ and $\theta = \log(1.3)$. 
where \( \theta = (\theta_1, \theta_2) \), \( \mathbf{V} \) is the asymptotic variance-covariance matrix of the estimate \( \hat{\theta} \), and \( \mathbf{W} \) is the \( 2 \times 2 \) variance-covariance matrix of the prior. This leads to

\[
\hat{\theta} | H_0 \sim (2\pi)^{-1/2} | \mathbf{V} |^{-1/2} \exp \left( -\frac{\hat{\theta}^T \mathbf{V}^{-1} \hat{\theta}}{2} \right)
\]

\[
\hat{\theta} | H_1 \sim (2\pi)^{-1/2} | \mathbf{V} + \mathbf{W} |^{-1/2} \exp \left( -\frac{\hat{\theta}^T (\mathbf{V} + \mathbf{W})^{-1} \hat{\theta}}{2} \right),
\]

to give the approximate Bayes factor

\[
\text{ABF} = \frac{p(\hat{\theta} | H_0)}{p(\hat{\theta} | H_1)} = | \mathbf{V} |^{-1/2} | \mathbf{V} + \mathbf{W} |^{-1/2} \exp \left( \frac{\hat{\theta}^T (\mathbf{V}^{-1} + (\mathbf{V} + \mathbf{W})^{-1}) \hat{\theta}}{2} \right).
\]

In specifying

\[
\mathbf{W} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix},
\]

it is sensible to choose \( W_{11} \leq W_{22} \), since we expect the effect of a single copy of the mutant allele to be no greater than the effect of two copies: \( W_{12} = W_{21} = \rho \sqrt{W_{11} W_{22}} \), and \( \rho \) may be specified on the basis of linkage-disequilibrium information.

### Web Resource

The URL for data presented herein is as follows:


### References

22. Whittemore A (2007) A Bayesian false discovery rate for mul-
tiple testing. J Appl Stat 34:1–9
a new paradigm? Cancer Epidemiol Biomarkers Prev 14:1359
false discovery rates for microarrays. Genet Epidemiol 23:70–
86
multiple testing. In: Bernardo JM, Bayarri MJ, Berger JO,
Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian
statistics 7: proceedings of the Seventh Valencia International
nomewide studies. PNAS 100:9440–9445
sample size for multiple testing: the case of gene expression
28. Tyrer J, Pharoah PDP, Easton DF (2006) The admixture max-
imum likelihood test: a novel experiment-wise test of asso-
ciation between disease and multiple SNPs. Genet Epidemiol
30:636–643
disequilibrium: genetic structure and spatial relations. Am J
Hum Genet 72:351–363
protects against population structure in genetic association
studies. Genome Res 16:290–296
32. Lowe CE, Cooper JD, Chapman JM, Barratt BJ, Twells RCJ,
Green EA, Savage DA, Guja C, Ionescu-Tirgoviste C, Tuomi-
genes using hSNPs: a staged approach. Genes Immun 5:301–
305
Bayes analysis of a microarray experiment. J Am Stat Assoc
96:1151–1160
34. Scott JG, Berger JO (2006) An exploration of aspects of Bayes-
35. Pauler DK, Wakefield JC, Kass RE (1999) Bayes factors and
approximations for variance component models. J Am Stat
Assoc 94:1242–1253