Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2008;37:641–653
doi:10.1093/ije/dym257

Reporting and interpretation in genome-wide association studies

Jon Wakefield

Accepted 4 December 2007

Background In the context of genome-wide association studies we critique a number of methods that have been suggested for flagging associations for further investigation.

Methods The *P*-value is by far the most commonly used measure, but requires careful calibration when the *a priori* probability of an association is small, and discards information by not considering the power associated with each test. The *q*-value is a frequentist method by which the false discovery rate (FDR) may be controlled.

Results We advocate the use of the Bayes factor as a summary of the information in the data with respect to the comparison of the null and alternative hypotheses, and describe a recently-proposed approach to the calculation of the Bayes factor that is easily implemented. The combination of data across studies is straightforward using the Bayes factor approach, as are power calculations.

Conclusions The Bayes factor and the *q*-value provide complementary information and when used in addition to the *P*-value may be used to reduce the number of reported findings that are subsequently not reproduced.

Keywords Bayes theorem, epidemiologic methods, genetic polymorphism, testing

Recent technological advances allow the simultaneous interrogation of huge numbers of pieces of genetic information. We concentrate on genome-wide association studies (GWAS)[1,2] in which single nucleotide polymorphisms (SNPs) are measured on sets of cases and controls over several stages. There are a number of standard platforms containing so-called tagSNPs that have been selected to capture common polymorphisms by exploiting linkage disequilibrium between SNPs.[3] As a typical example, Sladek *et al.*[4] recently reported a two-stage GWAS. At the first stage genotypes were obtained for 392 935 SNPs in 1363 type 2 diabetes cases and controls; these numbers represent the samples sizes after quality control checks on the genotyping, and removal of subjects who exhibited admixture or other inconsistencies. In a second stage the associations between disease and 57 SNPs were investigated in 2617 cases and 2894 controls, and eight were deemed significant after a Bonferroni correction had been applied in response to the multiple tests performed. A number of high profile GWASs have now been reported,[5–7] and many more will follow in the near-future.

This exciting development produces new challenges in terms of statistical analysis and interpretation.[8–11] Two key differences with conventional hypothesis testing situations, are the large number of tests that are performed, and the low *a priori* probability of a non-null association in each test. Historically, the usual situation was of a single experiment in which the prior probability of the alternative was not small—if this were not the case then a costly experiment would not be performed.

Given a set of tests from a GWAS we identify two important endeavors:

(i) Ranking the associations in order to determine a list of SNPs to carry forward to the next stage

Departments of Statistics and Biostatistics, University of Washington, Seattle, USA. E-mail: jonno@u.washington.edu

of study, when the size of the list has already been decided upon.

(ii) Calibrating inference to allow estimation of: the number of false discoveries and false non-discoveries, or the size of the list, or the probability of the null given the data for reported associations.

By far the most common measure used for flagging SNPs as 'noteworthy'[9] is the *P*-value. As we describe below, *P*-values are difficult to calibrate and there are various frequentist approaches for providing more interpretable measures, in particular via control of the false discovery rate (FDR). Alternatively, a Bayesian approach may be followed in which the probability of the null, given the data may be computed for each SNP; crucial to this approach is the calculation of the Bayes factor, which is the ratio of the probability of the data under the null to the probability of the data under the alternative. The Bayes factor was recently extensively used in the Wellcome Trust Case Control Consortium study[7] that investigated seven diseases using a common set of controls. The calculation of the Bayes factor requires specification of a prior distribution over all unknown parameters, and the evaluation of multi-dimensional integrals, and requires specialized software. To overcome these difficulties we concentrate on an asymptotic Bayes factor that has been recently proposed.[12]

Methods

Consider a typical GWAS in which for each SNP we wish to test H_0:$\theta = 0$ vs H_1:$\theta \neq 0$, in the context of a specified genetic model in which θ is the log odds ratio associated with exposure (for example, 1 or 2 copies of the mutant allele for a dominant genetic model). Further, assume we have a test statistic T with $E[T] = \theta$. For example, we may fit a logistic regression model (perhaps adjusting for matching or other variables) so that T is the maximum likelihood estimate of the log odds ratio. In large samples the statistic T is normally distributed with mean θ and standard error $\sqrt{V}$.

The interpretation of *P*-values

Before we see any data the α level of a two-sided test corresponding to T is $\alpha = \Pr(|T| > t_\alpha | H_0)$ and the power $1-\beta_\alpha = \Pr(|T| > t_\alpha | \theta)$ corresponding to this α may be calculated for different values of θ. Such *pre-data* inference is used for power calculations; α and β_α are frequentist probabilities with a long-run interpretation so that for a fixed critical region with threshold t_α, a proportion α of tests will be rejected using this rule when H_0 is true. Once the data are observed *post-data* inference is more relevant.[13] This has lead to the standard practice of quoting an *observed* significance level, or *P*-value, given by $p = \Pr(|T| > t_{\text{obs}} | H_0)$ where t_{obs} is the *observed* value of the test statistic. A critical issue is how to interpret this *P*-value; there are two common mis-interpretations. The first is to observe a *P*-value of 0.003 (say) and state: 'Under repeated sampling from the null we would have obtained this value, or a more extreme one, in only 0.3% of data sets'; this is incorrect since we have not observed 0.003 or a more extreme value, but rather exactly 0.003. With an *a priori fixed* critical region t_α it is correct to make such a statement, but once an observed significance level is quoted we have revised the critical region on the basis of the data and cannot appeal to long-run frequencies.

The second problem is the temptation to view the significance level as the probability of the null hypothesis given t_{obs}. Using Bayes theorem we have

$$\Pr(H_0|\text{data}) = \frac{p(\text{data}|H_0)\pi_0}{p(\text{data}|H_0)\pi_0 + p(\text{data}|H_1)(1-\pi_0)} \quad (1)$$

which depends on two quantities that are not used in the calculation of the *P*-value: the *prior* on H_0, π_0 and the power, $p(\text{data}\ |H_1)$, that is, the probability of the data under the alternative. Dividing both sides of (1) by $\Pr(H_1|\ \text{data})$ gives the posterior odds of no association:

$$\frac{\Pr(H_0|\text{data})}{\Pr(H_1|\text{data})} = \frac{p(\text{data}|H_0)}{p(\text{data}|H_1)} \times \frac{\pi_0}{1-\pi_0} \quad (2)$$

or, in words,

$$\text{Posterior Odds of } H_0 = \text{Bayes Factor} \times \text{Prior Odds of } H_0$$

so that the Bayes factor is an odds ratio corresponding to the posterior odds of the null divided by the prior odds of the null. The Bayes factor has been previously advocated as a measure of the evidence for an association in a GWAS.[7,12] When ranking associations we see, from (2), that if the prior odds $\pi_0/(1-\pi_0)$ are constant across SNPs then the ranks will be the same regardless of the specific value of π_0 taken. However, the rankings will change as a function of the power, $p(\text{data}\ |H_1)$, which varies across SNPs as a function of the minor allele frequency (MAF).

We now demonstrate the influence of the prior on the calibration of *P*-values. A lower bound for the probability of the null is given by:

$$\text{Posterior Odds of } H_0 > \{-e \times p \times \log p\} \times \text{Prior Odds of } H_0 \quad (3)$$

where $e = 2.7183$. The lower bound in (3) is valid for $p < 1/e = 0.368$, Sellke *et al.*[14] Figure 1 shows the lower bound on $\Pr(H_0|\ \text{data})$ as a function of the *P*-value for the four prior choices: $\pi_0 = 0.95$, 0.99, 0.999, 0.9999. For a *P*-value of 10^{-5} and $\pi_0 = 0.9999$ we have $\Pr(H_0|\text{data}) \geqslant 0.76$, so that there is at least a 76% chance that the null is true, even with such a small *P*-value. This bound is at first sight startling but some comfort is gathered by consideration of the situation in which the prior odds are one (so that we have equal prior weight on the null and on the alternative); *P*-values of 0.05 and 0.01 then give lower bounds on the null of 0.29 and 0.11, respectively. In addition

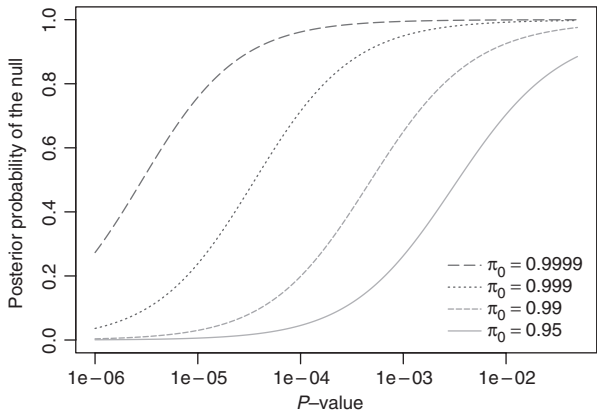


Figure 1 Lower bound on the posterior probability of the null, as a function of the *P*-value, and the prior on the null, π_0

Table 1 Possibilities when m tests are performed and k are called noteworthy

	Non-noteworthy	Noteworthy	
H_0	A	B	m_0
H_1	C	D	m_1
	$m-k$	k	m

to the low prior probabilities of an association in GWAS the other crucial aspect is that many hundreds of thousands of tests are being performed at once, and so by chance alone very small *P*-values will be observed. For example, if 500 000 SNPs are examined then even if the null is true for all tests we would still expect to see four *P*-values $<10^{-5}$.

To evaluate the probability of H_0 one must consider competing explanations for the data, i.e. the power under alternative hypotheses. It is important to consider power because although a small *P*-value suggests that the data are unlikely given H_0, they may also be unlikely under reasonable alternatives. From (2), we see that even if $p(\text{data}|H_0)$ is small, the Bayes factor may not be small if $p(\text{data}|H_1)$ is small also.

Control of FDR via *q*-values

The possible outcomes when m multiple-hypothesis tests are performed are given in Table 1; m_0 is the true number of nulls and is of course unknown; $\pi_0 = m_0/m$ is the proportion of nulls amongst all tests. The key issue is how to decide upon a criterion for calling an association noteworthy; with such a criterion, k is the number of tests called noteworthy. The number of false discoveries is B, and the number of false non-discoveries is C. In a GWAS we wish to make B and C as small as possible with D close to m_1.

Historically, the type I error (false discovery) was deemed the more important of the two types of error (false discovery and false non-discovery), which lead to the use of the Bonferroni correction, which controls the familywise error rate, that is the probability of making at least one type I error, $\Pr(B \geqslant 1)$—there is an implicit prior assumption that the probability that *all* tests are null is not small.[15] If we believe that all tests could be null then aiming to make the number of false positives zero is justifiable. In the context of a GWAS the use of Bonferroni will often be an overly conservative procedure since, at least in early stages of genome-wide investigations, one is more concerned with avoiding missed associations, and making some false discoveries is not too high a cost to pay in order to achieve more true hits. By overly protecting against false discoveries one loses power in detecting real associations. A second issue is that the usual Bonferroni correction was derived for independent tests, and in a GWAS there is dependence amongst the tests due to linkage disequilibrium, and correlated tests lead to an overly conservative procedure.[16]

More recently, Benjamini and Hochberg[17] suggested a powerful and simple method for controlling the frequentist expected FDR, that is the proportion of rejected tests that are truly null: $\mathrm{E}[B/k]$. Subsequently, Storey and colleagues[18,19] have advocated the use of *q*-values. Suppose we reject all tests for which $|T| > t_{\text{fix}}$ for a fixed threshold t_{fix}. Then the probability of the null for tests that fall within this critical region is

$$q(t_{\text{fix}}) = \Pr(H_0 | |T| > t_{\text{fix}}) = \frac{\alpha(t_{\text{fix}})\pi_0}{\Pr(|T| > t_{\text{fix}})} \qquad (4)$$

where $\Pr(|T| > t_{\text{fix}}) = \alpha(t_{\text{fix}})\pi_0 + [1-\beta(t_{\text{fix}})](1-\pi_0)$ is the probability of a rejection and $\alpha(t_{\text{fix}})$ is the α level corresponding to t_{fix}. Hence for a rule defined by t_{fix}, $q(t_{\text{fix}})$ is the probability of a false discovery, and Storey[19] shows that such a rule applied to multiple tests controls the (frequentist) FDR at level $q(t_{\text{fix}})$.

For a particular SNP one can take $t_{\text{fix}} = t_{\text{obs}}$, where t_{obs} is the observed statistic. Then we obtain the *q*-value $q(t_{\text{obs}})$ where $\alpha(t_{\text{obs}}) = p$. Hence if we have a rule that just calls this SNP, and all SNPs with a more extreme statistic, noteworthy, then the FDR is controlled at level $q(t_{\text{obs}})$; because this threshold includes *more* noteworthy SNPs (for which the probability of H_0 is lower) the probability that this SNP is a false positive may be much higher than the FDR, however.

To evaluate *q*-values for each SNP in practice it would appear from (4) that we need an *a priori* estimate of π_0. However, we may write

$$\Pr(H_0 | |T| > t_{\text{obs}}) = p \times \frac{\pi_0}{\Pr(|T| > t_{\text{obs}})}$$

and Storey[19] shows that the second term can be estimated from the totality of *P*-values, which removes the need to specify π_0. Intuitively, under the null, the distribution of *P*-values is uniform and so when we are in a multiple-hypothesis testing situation we can use the departure of the distribution

of all P-values from uniformity to estimate π_0, an empirical approach that has much appeal.

The false non-discovery rate (FNR) is defined as $E[C/(m-k)]$ and is the expected proportion of non-noteworthy tests that are truly non-null. However, in a GWAS, the number of non-noteworthy tests, $m-k$, will be very large (and close to m); hence, even if the majority of true associations are missed, C will still be relatively small and so $E[C/(m-k)]$ will also be close to zero and difficult to accurately estimate. The ratio of the non-null associations missed C/m_1 (i.e. 1–sensitivity) is clearly of interest, but difficult to estimate since both C and m_1 are unobserved.

The false positive report probability

In response to the large proportion of false positives generated by the reporting of P-values in genetic association studies, Wacholder and colleagues,[9] in a wide-ranging and seminal article, introduced the false probability report probability (FPRP):

$$\Pr(H_0|\text{data}) = \text{FPRP} = \frac{p \times \pi_0}{p \times \pi_0 + \text{power} \times (1-\pi_0)} \quad (5)$$

where the 'data' are given by $|T| > t_{\text{obs}}$ and the power $= \Pr(\text{data} \mid \theta_1)$ is evaluated at a pre-specified θ_1, and for $|T| > t_{\text{obs}}$. If we rewrite (5) as

$$\text{Posterior Odds of } H_0 \text{ given } \{p, \text{power}\} = \frac{p}{\text{power}} \times \text{Prior Odds of } H_0$$

it is clear that the evidence in the data to support H_0 are summarized in terms of the ratio p/power, which again illustrates that when a set of tests differ in their power the rankings of P-values and FPRP will differ also; for fixed P-value FPRP gives more weight to H_1 when the power is high. The functional form of (5) is familiar to epidemiologists; the baseline (prior) odds of the event H_0 is revised in light of the odds ratio p/power, to give the posterior odds. FPRP lies somewhere between a Bayesian and a frequentist approach since a Bayesian calculation is carried out using frequentist reporting statistics; the 'data' correspond to p and the power, the latter is calculated at the simple alternative $H_1{:}\theta = \theta_1$, with a prior point mass of $1-\pi_0$ at this value.

FPRP has a number of drawbacks[12] which we now briefly describe, in order to motivate an alternative that we describe in the next section. Information is being lost by considering $|T| > t_{\text{obs}}$ only, rather than conditioning on the exact value observed, t_{obs}; it can be shown that $\Pr(H_0 \| T| > t_{\text{obs}}) \leqslant \Pr(H_0 | T = t_{\text{obs}})$ so that FPRP is a lower bound on the probability of H_0. It is inconsistent to consider a two-sided P-value and the power corresponding to a one-sided alternative. When one knows the side of the null to which the estimate falls then a single tail area is appropriate. With respect to frequentist properties FPRP does not provide control of FDR because a variable threshold for T is used which does not permit long-run frequencies to be calculated—in particular the FDR is not controlled by FPRP. Finally, it would be desirable to consider a range of values for the alternative θ, rather than a single value θ_1.

Table 2 Costs of making the two types of error, C_{FD} is the cost of a false discovery, and C_{FND} the cost of a false non-discovery

		Decision	
		Not Noteworthy	Noteworthy
Truth	H_0	0	C_{FD}
	H_1	C_{FND}	0

The Bayesian false discovery probability

For the ranking of associations we have seen that for a Bayesian approach with a constant prior odds across SNPs we need only consider the Bayes factor, and not the absolute value of $\Pr(H_0|\text{data})$. For the second endeavor of calibration the posterior probability of the null is required, and we describe a Bayesian decision theory approach to the choice of which of H_0 or H_1 to report. This requires the costs of false non-discovery and false discovery to be specified, Table 2 gives the costs of making the two types of error.

The decision theory solution is to report H_1 if the

$$\text{Posterior Odds of } H_0 < \frac{C_{\text{FND}}}{C_{\text{FD}}} \quad (6)$$

so that we only need to consider the ratio of costs $C_{\text{FND}}/C_{\text{FD}}$. If the costs are equal then we should report an association as noteworthy if the posterior odds on H_0 is <1; if $C_{\text{FND}}/C_{\text{FD}} = 4$, so that missing a discovery is four times as costly as reporting a null association, then an association should be called noteworthy if the posterior odds on H_0 is <4, i.e. if the posterior probability of H_1, $\Pr(H_1|\text{data})$, is >0.2. We now discuss Bayesian error measures that are closely related to FDR and FNR. For a single test:

- If we call a hypothesis *noteworthy* then $\Pr(H_0|\text{data})$ is the probability of a *false discovery*.
- If we call a hypothesis *not noteworthy* then $\Pr(H_1|\text{data})$ is the probability of a *false non-discovery*.

In a multiple-hypothesis testing situation, we can sum $\Pr(H_0|\text{data})$ over all associations that are called noteworthy to give the expected number of false discoveries; summing $\Pr(H_1|\text{data})$ over all associations called non-noteworthy gives the expected number of false non-discoveries.

The data appear in the posterior odds through the Bayes factor, which is given by $p(\text{data}|H_0)/p(\text{data}|H_1)$, and is the ratio of the probabilities of the data under H_0 and H_1. For FPRP the denominator (power) was evaluated at a single alternative, θ_1. An alternative approach is to place a prior on

plausible values of θ. The denominator of the Bayes factor is then given by

$$p(\text{data}|H_1) = \int p(\text{data}|\theta) \times g(\theta)\text{d}\theta$$

which is the power as a function of θ, averaged over the prior, $g(\theta)$.

To evaluate the Bayes factor in general requires the specification of the prior over *all* unknown parameters, and the calculation of multi-dimensional integrals. An approximate Bayes factor that removes these difficulties, and avoids the drawbacks of FPRP has been recently developed,[12] and takes as data the estimate of the log odds ratio, $\widehat{\theta}$, with associated standard error $\sqrt{V}$. The asymptotic distribution of the estimator is $N(\theta, V)$, where θ is the true value, and this distribution provides the likelihood in the evaluation of the Bayes factor. As prior a normal distribution centered on zero and with variance W is taken—this reflects the expected distribution of the sizes of effects over all non-null SNPs. This combination gives the *approximate Bayes factor* (ABF):

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

where $Z = \widehat{\theta}/\sqrt{V}$ is the usual Z statistic, and $r = W/(V+W)$. Hence we see that the Bayes factor depends on both the Z statistic and the power through V (which depends on the MAF and the sample size). All that is required data-wise to calculate ABF is a confidence interval on the parameter of interest, and we provide a number of illustrations in the Examples from the Literature section. The posterior odds is given by

$$\text{Posterior Odds of } H_0 \text{ given } \widehat{\theta} = \text{ABF} \times \text{Prior Odds of } H_0$$

To choose W we may specify a range of relative risks that we believe is *a priori* plausible. For example, if we believe that there is a 95% chance that the relative risks lie between 2/3 and 1.5 then the standard deviation of the prior is $\sqrt{W} = \log(1.5)/1.96$ (equation (3), is a lower bound on the posterior odds of H_0 over all W, Sellke *et al.*[14]).

If we pick the prior variance $W = K \times V$ (where V is the asymptotic variance of $\widehat{\theta}$ and $K > 0$ is a constant) then ABF is given by

$$\text{ABF} = \sqrt{1+K}\exp\left(-\frac{Z^2}{2}\frac{K}{1+K}\right)$$

which depends on the data only through Z. Hence, for this prior, rankings based on ABF and the P-value will be identical.[20] Under this prior, larger effect sizes are anticipated (in a very specific way) when the MAF is low and/or the sample size is small (since in this case the variance V is large). While I would not suggest that this prior should be used, since it is not likely to reflect carefully considered prior opinion, it does reveal a prior that is implicitly consistent with the p-value approach and so can explain observed differences between rankings based on p-values and Bayes factors. Further discussion is given elsewhere.[20]

The posterior probability of the null is given by

$$\Pr(H_0|\widehat{\theta}) = \frac{\text{ABF} \times \text{Prior Odds}}{1 + \text{ABF} \times \text{Prior Odds}}$$

which was called the Bayesian false discovery probability (BFDP) by Wakefield.[12] In general the Bayes factor is a measure of the evidence in the data for one scientific hypothesis (H_0) compared with another (H_1), and a number of authors have suggested that 'a rough descriptive statement about standards of evidence in scientific investigation'[21] may be presented in terms of $-\log_{10}$BF. It turns out that, although the *rankings* of the approximate Bayes factors and P-values will in general differ (apart from under the prior $W = V \times K$), if we treat ABF as a statistic and evaluate the frequentist P-value associated with this statistic then they are identical to P-values obtained using the Wald statistic $Z = \widehat{\theta}/\sqrt{V}$. This is because for fixed V the approximate Bayes factor is simply a transformation of Z^2 and so the lower tail of the distribution of the Bayes factor (lower ABF, more evidence for the alternative) corresponds exactly to the upper tail of a chi-squared (from which the P-value is calculated), Appendix 1 contains details.

The fact that ABF simply depends on Z^2 and V allows the expected number of tests falling beyond $-\log_{10}$BF thresholds under the null to be easily calculated, given a set of MAFs and sample sizes (which jointly determine the distribution of V). Hence evidential guidelines may be based on the frequentist properties of the Bayes factor by comparing the observed number falling beyond thresholds of $-\log_{10}$BF with those expected under the null, a point that we illustrate in the Operating Characteristics via Simulation section. Similar ideas have appeared recently in the genetics literature.[22] We emphasize that although the P-values corresponding to Z and ABF are identical, the frequency distribution of ABF across SNPs will differ according to the MAFs of the SNPs under consideration.

The simple form of ABF also means that power calculations are straightforward.[20] If we decide to call a SNP noteworthy if the posterior odds of H_0 drop below the ratio of costs of false non-discovery to false discovery, call this C, then the power to detect a relative risk of RR_1 is given by

$$\begin{aligned}&\Pr\{\text{ABF}(W,Z,V) \times \pi_0/(1-\pi_0) < C|\text{RR}_1\}\\ &= \Pr\left\{Z^2 \geq -\frac{2}{r}\log\left[C\frac{1-\pi_0}{\pi_0}\sqrt{1-r}\right]|\text{RR}_1\right\}\end{aligned}$$

and under H_1 Z^2 is a non-central χ^2 random variable with a single degree of freedom and non-centrality parameter $(\log \text{RR}_1)^2/V$. For example, Figures 2a and b illustrate the powers to detect a relative risk of 1.5 for sample sizes of 1000 and 2000 and various

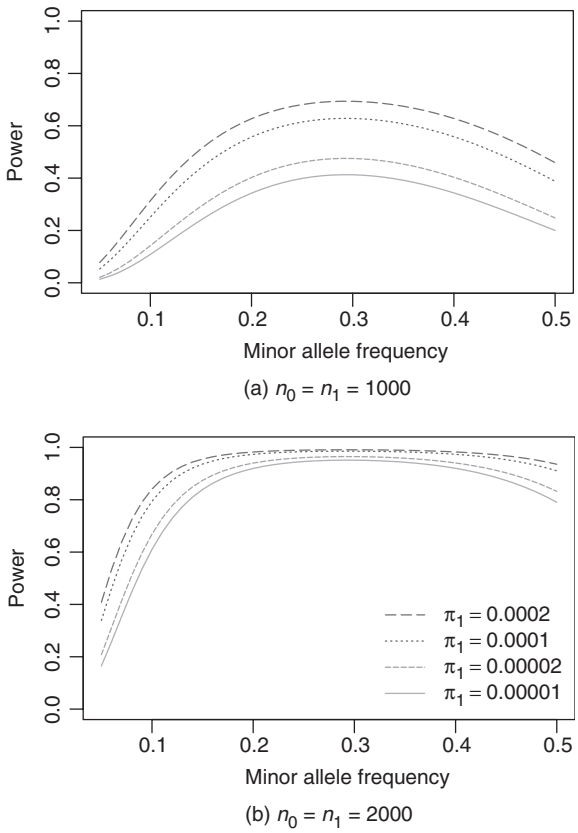


Figure 2 Power to detect a relative risk of 1.5, as a function of MAF and π_1, the probability of a non-null association. The genetic model is dominant, and the ratio of costs of false non-discovery to false discovery is 10 so that the null is rejected if the posterior probability of the alternative is >0.09

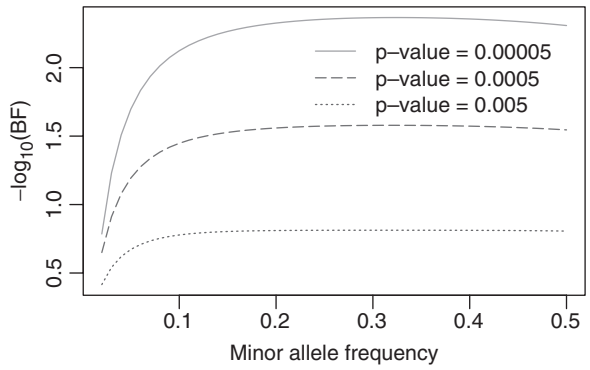


Figure 3 Evidence in favour of the alternative vs the null for three different *P*-values, as a function of the MAF

choices of π_1 (the prior probability of an association), under a dominant genetic model and with a ratio of costs $C = 10$ (so that false non-discovery is 10 times worse than false discovery). The 97.5% point of the lognormal prior on the effect size is 2 (which determines the prior variance W). The effect of both sample size and MAF on the variance of the estimator (and hence the power) is apparent.

Given the massive multiple hypothesis testing carried out in genome-wide scans, replication is essential.[23] Combination of data across studies (assuming that the effect is constant across studies) to produce a Bayes factor summarizing both sets of data is straightforward since

$$\mathrm{ABF}(\widehat{\theta_1}, \widehat{\theta_2}) = \mathrm{ABF}(\widehat{\theta_1}) \times \mathrm{ABF}(\widehat{\theta_2}|\widehat{\theta_1}) \qquad (7)$$

where $\mathrm{ABF}(\widehat{\theta_2}|\widehat{\theta_1}) = p(\widehat{\theta_2}|H_0)/p(\widehat{\theta_2}|\widehat{\theta_1}, H_1)$ and $p(\widehat{\theta_2}|\widehat{\theta_1}, H_1) = \mathrm{E}_{\theta|\widehat{\theta_1}}[p(\widehat{\theta_2}|\theta)]$ which is available in a simple form, Appendix 2 gives details. The last expression simply shows that when we evaluate the probability of the data $\widehat{\theta_2}$ under the alternative we average over the posterior for θ given $\widehat{\theta_1}$; this contrasts with the evaluation of the probability for $\widehat{\theta_1}$ under the alternative for which we average over the prior for θ, i.e. $p(\widehat{\theta_1}|H_1) = \mathrm{E}_\theta[p(\widehat{\theta_1}|\theta)]$.

We now turn to the thorny issue of choice of π_0. As more genome-wide association studies are carried out lower bounds on $\pi_1 = 1 - \pi_0$ will be obtained from the confirmed 'hits'—it is a lower bound since clearly many non-null SNPs for which we have a low power of detection will be missed. In a GWAS the proportion of true non-null signals is likely to be small, and so estimation of π_0 using the empirical distribution of the totality of *P*-values is likely to be difficult. However, if an estimate of $\pi_0 < 1$ is obtained using the *q*-values methodology then this may be used as a non-subjective 'empirical' prior. We emphasize that π_1 is the proportion of non-null associations in the data, and not the proportion we think we have the power to detect.

We now illustrate how power is not considered when a *P*-value is calculated. In Figure 3 each curve corresponds to a fixed *P*-value and the vertical axis measures the evidence in favour of the alternative, $-\log_{10}\mathrm{BF}$, so that a value of 2 means that the data are 100 times more likely under the alternative than under the null. On the horizontal axis we have the minor allele frequency (MAF), which drives the power. We assume a dominant genetic model and take a prior that assumes that the odds ratio is <1.5 with probability 0.975 and, crucially, takes the effect size to be independent of the MAF. We concentrate on the curve labelled $P = 0.00005$. For a MAF close to 0.05 (low power) the Bayesian evidence in favour of the alternative is small because to obtain such a small *P*-value requires a large $\widehat{\theta}$ which is unlikely under the prior. The *P*-value provides more evidence because the implicit prior on the effect size ($W = K \times V$) places more probability on larger effect sizes at lower MAFs. As the MAF increases the power also increases and under the Bayes factor approach the evidence in favour of the alternative consequently increases also. For a MAF close to 0.5 we have strong power and the evidence starts to decrease, in contrast to *P*-values for which it is well known that the null will be rejected

for large sample sizes, even if $e^{\widehat{\theta}}$ only differs from unity by a small amount. The reason for the discrepancy is that although the data may be highly unlikely the null, the data may also be unlikely under the alternative also and so the relative evidence is reduced (under the *P*-value approach there is no alternative hypothesis). This behaviour is also discussed by Spiegelhalter *et al.*[24] We stress, however, that for MAFs between 0.15 and 0.50 there is little practical difference between rankings based on *P*-values and Bayes factors here.

Operating characteristics via simulation

We carry out a simulation study in which there are 3000 cases and 3000 controls and assume that 317 000 SNPs are to be examined, of which 100 are truly associated with disease. We take a linear additive model on the logistic scale[25] with θ the log relative risk associated with two copies of the mutant allele. We generate the log relative risks for the 100 SNPs from a beta distribution with parameters 1 and 3 scaled to lie between log(1.1) and log(1.5), and then with probability 0.5 change the sign (so that in expectation there is a 50% chance of a detrimental or protective effect). The relative risks are assumed independent of the MAFs, and for the latter we assume for all SNPs a uniform distribution between 0.05 and 0.50. The blue and red filled circles in each panel of Figure 4 show the distribution of the non-null log relative risks plotted against the MAF.

We calculate the ABF based on $\widehat{\theta}, V$ obtained from 317 000 logistic regression models fitted to each SNP, and a prior that assumes, independently of the MAF, that the odds ratios lie between 2/3 and 1.5 with probability 0.95. The four panels of Figure 4 show the number of SNPs called as noteworthy (blue circles) using BFDP with different thresholds, the number missed (red circles), and the number of false discoveries (green circles, with points jittered in the vertical direction for clarity). The four thresholds correspond to illustrative ratios of costs, $C_{\mathrm{FND}}/C_{\mathrm{FD}}$ of 4:1, 20:1, 50:1, 100:1. We see the diminishing returns in setting higher and higher thresholds with the FDR increasing dramatically as the threshold increases. To emphasize the difficulty in detecting non-null SNPs when the power is low we have used the true π_0 in the calculation of BFDP, which corresponds to the best possible scenario. In general choosing the ratio of costs is not straightforward though replication studies will clearly have ratios that are lower since we would like to see the posterior probability of the null being small, more discussion is available elsewhere.[9–12]

Figure 5 shows the number of SNPs that we need to call noteworthy to obtain a specified number of true 'hits'. The dashed line is the line of $y = x$ and a perfect procedure would follow this line. We see that the signal is only strong for the first few SNPs (the two most noteworthy SNPs under ABF and the *P*-value are true associations, the third is not) and

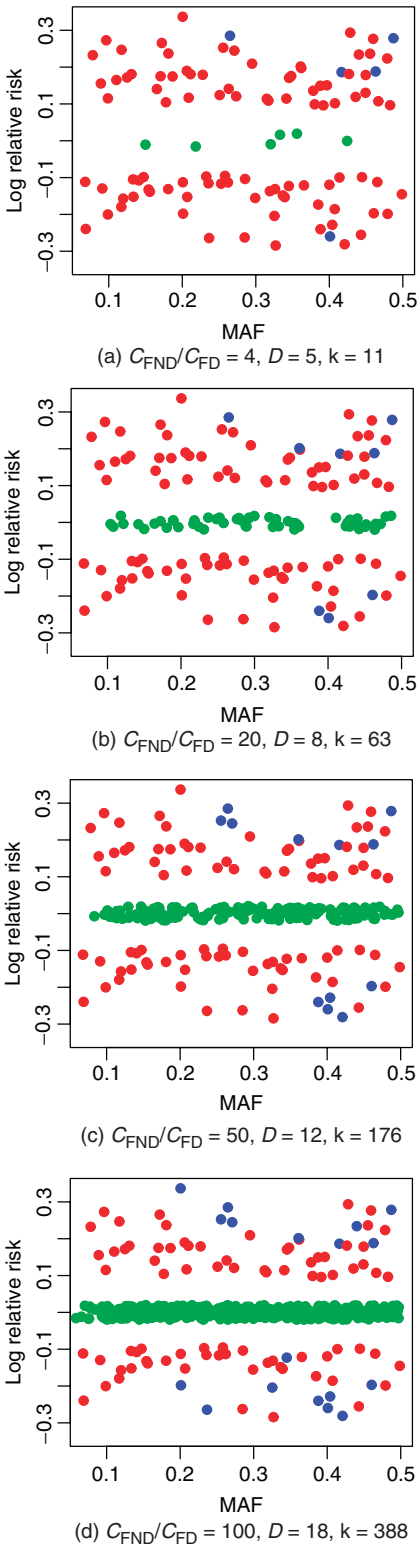


Figure 4 Discoveries (blue circles), non-discoveries (red circles) and false discoveries (green circles) using BFDP for four different thresholds corresponding to ratio of costs of false non-discovery to false discovery of 4:1, 20:1, 50:1, 100:1 in panels (a), (b), (c), (d). $C_{\mathrm{FND}}/C_{\mathrm{FD}}$ is the ratio of costs, D the number of true discoveries, and k the total number of SNPs called noteworthy

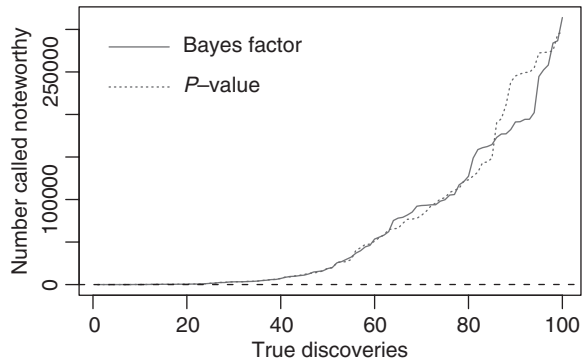


Figure 5 Number of SNPs called noteworthy in order to detect a specified number of true discoveries, with noteworthiness based on *P*-values and Bayes factors. The dashed line is the line of equality and shows that after the first few hits the curves move increasingly away from the dashed line demonstrating that the false discovery rate increases rapidly as the length of the list is increased

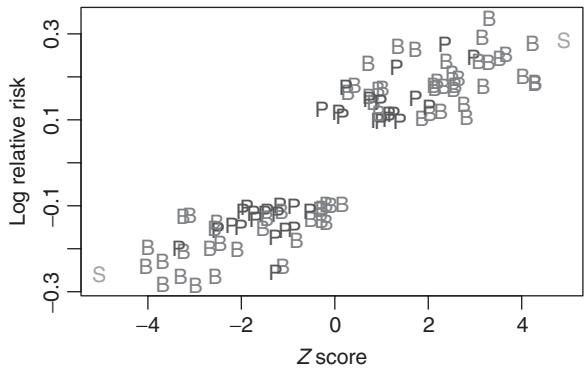


Figure 6 True log relative risks versus *Z*-scores for the 100 non-null SNPs; the 63 points marked 'B' had lower rankings on the Bayes factor list, while the 35 marked 'P' had lower rankings on the *P*-value lists; the two SNPs marked 'S' had identical rankings (and were the first two found)

early in the list we need to call an increasing number of SNPs noteworthy in order to flag the true non-null associations. To discover the final few signals the list must include virtually all of the SNPs. Figure 6 shows the SNPs with lower rankings on the Bayes factor list (marked 'B', 63 points) or on the *P*-value list (marked 'P', 35 points), with the first two SNPs (marked 'S') being equally ranked. We see that the majority of SNPs for which *P*-values performed better had true log relative risks close to 1 and so would need very large sample sizes to be reproducible. The explanation for *P*-values ranking low power alternatives earlier is the implicit *P*-value prior; for two SNPs with the same *Z*-score, the one with the greater power will provide more evidence against the null under the Bayes factor approach. This implicit prior also explains why here the Bayes factors are superior overall in terms of flagging associations earlier—the data were generated with effect size independent of MAF.

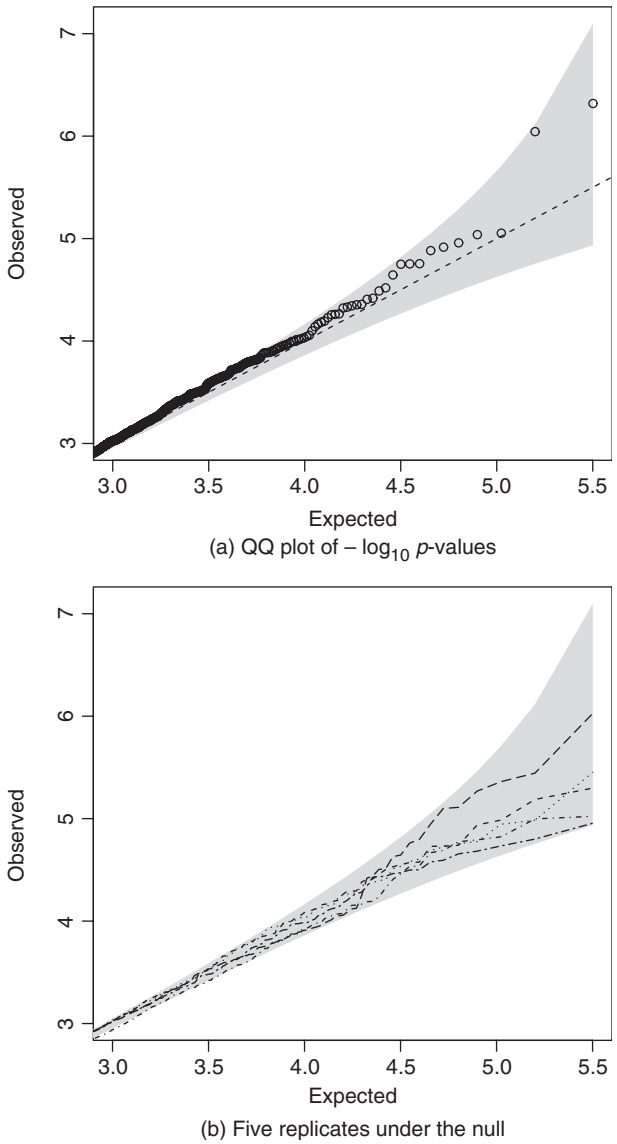


Figure 7 QQ plot of $-\log_{10}$ *P*-values

Figure 7a gives the QQ plot of $-\log_{10}$ *P*-values; as already noted *P*-values based on the statistic ABF are identical to the *P*-values based on the Wald statistic *Z*. The shaded areas are pointwise 95% confidence intervals.[26] Such plots are difficult to interpret due to sampling variability in the upper tail and the dependency in the plotted points. For clarity we have only plotted points that are greater than 3 (the region on interest). We see that only two of the points are distinct from the remainder. To aid in interpretation, Figure 7b gives five realizations under the null, and the dependency and sampling variability is apparent.

Table 3 gives the expected number of tests falling within different bands under the null, along with the observed number. Informally, we would conclude that the top two SNPs appear to be real hits while approximately four of the next nine hits are real. This table differs from that based on *P*-values since the MAFs of

Table 3 Strengths of evidence and observed and expected numbers of Bayes factor statistics falling within evidential bands

Bayes Factor	$-\log_{10}$BF	Expected	Observed	$\frac{\text{Observed}}{\text{Expected}}$
<0.0001	>4	0.3	2	6.30
0.0001–0.001	3–4	5.2	9	1.74
0.001–0.01	2–3	89.0	108	1.21
0.01–0.1	1–2	1703.2	1736	1.02
0.1–0.32	0.5–1	8070.4	8164	1.01

the 317K SNPs in this dataset are explicitly considered (in other words, Table 3 accounts for power). Figure 8 gives a number of summaries of the q-value method when applied to the simulated data. The proportion of non-null tests was empirically estimated as 0.003 (the true proportion is $100/317\,000 = 0.0003$) by the q-value method.

Figure 8a plots q-values against P-values and illustrates that most of the q-values are close to 1. In Figure 8b we plot the expected number of false discoveries, as calculated via the q-value and BFDP methods (both based on $\widehat{\pi}_0$ estimate from the q-value method) versus the number of true discoveries between 1 and 50. The expected number of false discoveries for BFDP is the sum of the posterior probabilities of the null over all SNPs called noteworthy, and for the q-value approach it is q times the number of SNPs called noteworthy at that threshold. Two features are apparent: first the expected number of false discoveries increases rapidly with the number of true discoveries and second, the two methods give very similar estimates. In Figure 8c we plot q-values vs BFDP (with the latter calculated using the q-value estimate of π_0) and see a reasonable amount of agreement though the q-values tend to be smaller since, as noted, they are a lower bound on the posterior probability of the null.

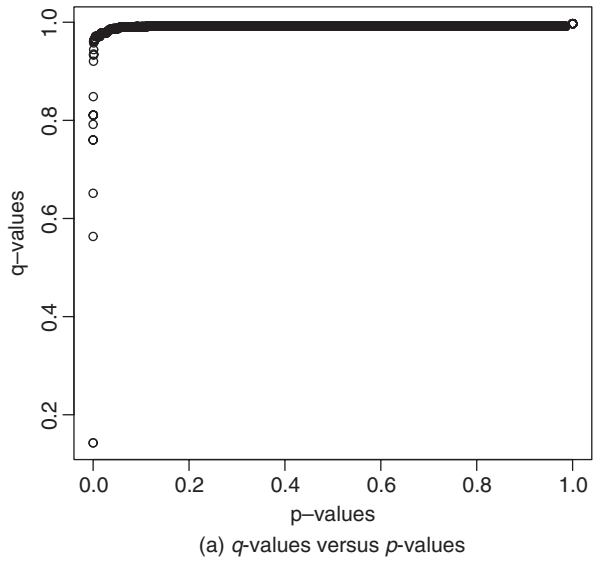
Examples from the Literature

Table 4 gives point estimates of odds ratios and confidence intervals (CIs) for SNP rs9939609 from a GWAS for Type II diabetes.[6] Bayes factors and BFDP are calculated under three prior distributions with proportions of non-null SNPs of 1/5000, 1/10000 and 1/50 000. The estimate (CI) in the first row of the table corresponds to an association found in 1924 type 2 diabetes patients[6] when compared to 2938 controls (490 032 SNPs were examined in total). There is strong evidence of a non-null association for this FTO gene variant, which manifests itself in very small probabilities of the null under all three priors. In a second stage this association was examined in 3757 type 2 diabetes cases and 5346 controls and in the second line of the table we see a greatly reduced relative risk estimate, and the three posterior probabilities of the null for these data alone are all >0.9. However, combining the Bayes factors using equation (8) in Appendix 2 we obtain a combined $-\log_{10}$BF of 13.8, greater than the sum of the two individual contributions (which is 10) because the estimates and confidence intervals are in broad agreement. Hence the data are overwhelmingly in favour of the alternative so that even with a prior of 1/50 000 the posterior probability of the null is 7.6×10^{-10}. For summarizing inference under the alternative the (2.5%, 50%, 97.5%) points of the prior are (0.67, 1, 1.5), being refined to (1.17,1.26,1.36) after the first stage data and finally to (1.15,1.21,1.27) using both stages of data. The posterior interval after stage 1 is virtually identical to the asymptotic CI in Table 4 because the variance of $\widehat{\theta}_1$ is so small compared to the prior variance, W (the shrinkage factor $r = 0.97$ showing that the prior is dominated by the data). The summary of the association is of a relative risk increase of 21%.

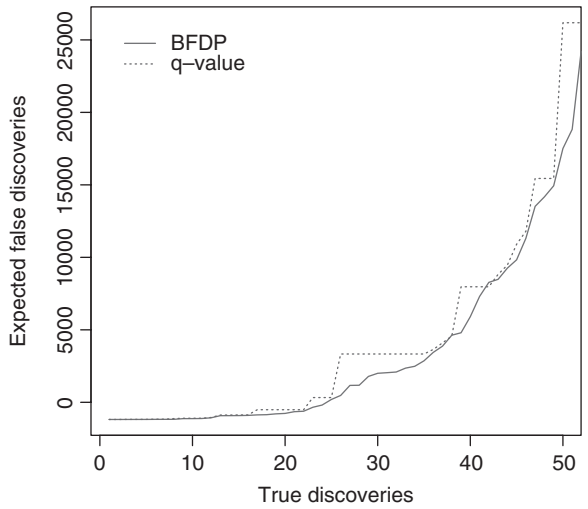
Table S5 of the supplementary table of Sladek *et al.*[4] gives the genotype counts for cases and controls for 43 SNPs that passed the first stage selection cut-off. For illustration for SNP rs7913837 we fitted a logistic regression model using a risk model that is linear (on the logistic scale) in the number of mutant alleles. We then calculated the Bayes factor, and BFDP using the resultant relative risk estimate and asymptotic variance. The latter was multiplied by the estimated genomic control inflation factor[27] of 1.1233. This illustrates that the asymptotic distribution that is used in the ABF calculation can incorporate additional information. Under a prior that assumes a narrower range of risks, (2/3,1.5) with probability 0.95, the evidence for a non-null association is not strong, Table 4, last line. Figure 9 illustrates the sensitivity of BFDP to the prior on effect size, for three different values of π_1, the probability of a non-null association. Under prior effect sizes that give more weight to larger values of the odds ratio we see greater evidence of an association. The lower bounds on the posterior probability of the null, given by equation (3) are also indicated as dashed lines. We see that beyond an upper value of around 3 there is little sensitivity in the Bayes factor. This figure indicates that care must be taken in the choice of prior distribution. We note that in the second stage of the study the relative risk estimate was much smaller (1.45 for two mutant alleles).

Conclusions

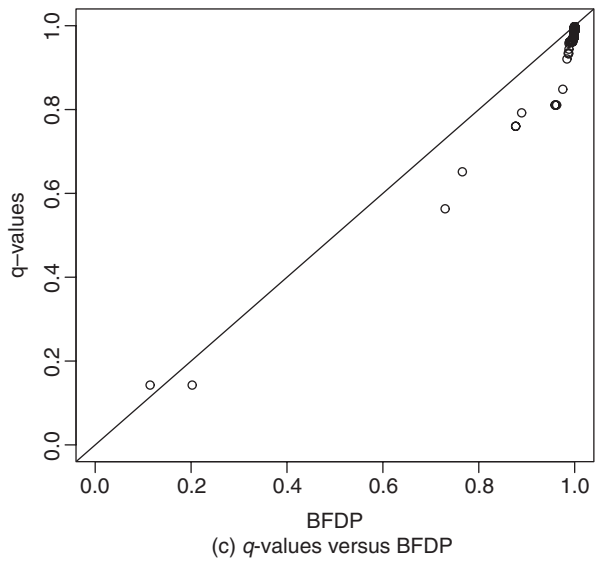
We have discussed the interpretation of P-values in GWAS and shown that small P-values have to be taken in the context of low prior probabilities of an association and the multiple-hypothesis tests that have been carried out, as previously argued by Wacholder *et al.*[9] In terms of reporting, P-values are useful in that their null distribution is known to be uniform, but they do not consider power. We have shown that they implicitly correspond to a particular



(a) *q*-values versus *p*-values



(b) Excepted false discoveries versus true discoveries



(c) *q*-values versus BFDP

Figure 8 BFDP, *P*- and *q*-value summaries

prior relationship between the MAF and the strength of association. The *q*-value explicitly estimates the proportion of non-null tests using the totality of *P*-values, and provides an estimate of the FDR for any fixed threshold, but in GWASs the proportion of non-null associations is small and more experience of its use in this context is required.

A refinement of FPRP, BFDP has been described here and elsewhere,[12] and has the advantage of only requiring a confidence interval for its calculation. Treating the distribution of the statistic as the data also provides flexibility and allows, for example, over-dispersion (genomic control) to be simply incorporated by multiplying the variance of the odds ratio by the overdispersion factor. Treating the asymptotic Bayes factor as a statistic one may evaluate its frequentist properties and it turns out that the *P*-values associated with the ABF are identical to those for the conventional Wald statistic. We stress, however, that the rankings of ABF and *P*-values will differ in general, since the former takes into account the power.

We have presented BFDP in its simplest form, and a number of extensions are currently being explored. We may allow the variance on the size of the effect, W, to depend on the MAF to exploit the common perception that larger detrimental effects may occur with rarer minor allele frequencies. We have assumed a fixed threshold across all SNPs (corresponding to fixed costs) but we may wish for the costs (and therefore the threshold) to depend on the MAF, with greater costs associated with more common alleles, since these will have a greater attributable risk. The ratio of costs will clearly depend on the phase of the study and on the sample size. Since all that is required for the calculation of ABF is a point estimate/standard error the approach may used with designs other than the case-control, for example survival endpoints in a case-cohort study. The design must also be acknowledged in the analysis phase for other outcome-dependent sampling schemes such as two-phase sampling. The use of Bayes factors based on test statistics has been previously advocated as a robust and theoretically sound strategy.[28,29] The asymptotic Bayes factor described here may also be used for model averaging over different genetic models, which has been advocated elsewhere.[30]

Replacing confidence intervals with *P*-values does not overcome the problems of reporting when the prior probability of an association is low. The posterior distribution for the relative risk of an association *given* an association (i.e. H_1) is lognormal with parameters $r\widehat{\theta}$ and $r\widehat{\theta}$. Without assuming an association the posterior consists of a point mass of BFDP at RR = 1 and the remaining 1–BFDP is the area under the lognormal distribution.

Throughout we have used the term noteworthy, following Wacholder *et al.*[9] but these tests may be alternatively labelled as 'anomalous' recognizing that the flagged associations may be due to errors in the

Table 4 Frequentist and Bayesian summaries for reported SNPs. The 97.5% point of the prior for the odds ratio was set at 1.5

SNP[REF]	Est	95% C.I.	*P*-value	$-\log_{10}$ BF	BFDP with Prior: 1/5000	1/10 000	1/50 000
rs9939609[6]	1.27	1.16–1.37	6.4×10^{-10}	7.28	0.00026	0.00052	0.0026
rs9939609[6]	1.15	1.09–1.23	4.6×10^{-5}	2.72	0.905	0.950	0.990
rs7913837[4]	2.20	1.57–3.07	4.0×10^{-6}	2.55	0.933	0.965	0.993

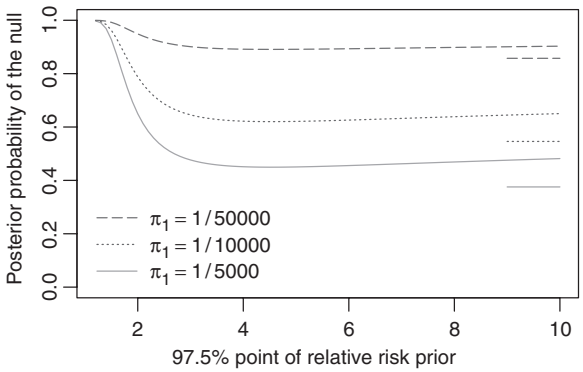


Figure 9 Sensitivity of BFDP (the posterior probability of the null) to the upper 97.5% point of the prior on the odds ratio, for three different priors for an association, π_1. The shorter lines under each of the main lines represent the theoretical lower bound on BFDP, over all choices of prior variance *W*

data such as differential genotyping errors. Software to evaluate approximate Bayes factors and posterior moments is available from the website: `http//faculty.washington.edu/jonno/cv.html`.

Returning to the endeavors highlighted in the introduction:

(i) To rank associations the Bayes factor provides an alternative to the *P*-value which accounts for power. Bayes factor and *P*-values will often provide very similar rankings, with differences only for SNPs with low MAFs, and the extent of the differences depending on the association in the prior between size of effect and MAF. We would recommend close examination of any discrepancies between SNPs that appear in one but not both highly-rank lists.

(ii) To calibrate inference/decide upon the list length for further investigation, the *q*-value and BFDP may be used to estimate FDR or the probability of the null given the data. BFDP may also be used to interpret reported associations, though the absolute values are highly dependent upon an appropriate choice of π_0, the prior on the null. Careful consideration of the prior should also be taken, both in terms of the sizes of effect anticipated, and whether effect size is likely to depend on MAF.

Acknowledgements

This work was partially supported by grant 1 U01–HG004446–01 from the National Institutes of Health. I would also like to thank David Balding and John Storey for providing helpful comments on an earlier draft.

KEY MESSAGES

- Extreme caution is required in the use of *P*-values in a genome-wide association study due to the low *a priori* probability of any association being non-null, and the large number of tests being performed.
- The Bayes factor provides an appealing alternative to the *P*-value for deciding on the noteworthiness of an association, though care should be taken in the specification of a prior on the effect size.
- Since the interpretation of *P*-values depends crucially on the power associated with the test, which depends in turn on the sample size and minor allele frequency, a single universal *P*-value noteworthy threshold is not generally appropriate.

References

[1] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;**6:**95–108.

[2] Wang WYS, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;**6:**109–18.

[3] Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphims for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004;**74:** 106–20.

[4] Sladek R, Rocheleau G, Ring J *et al*. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;**445:**881–85.

[5] Easton DF, Pooley KA, Dunning AM *et al*. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007;**447:**1–9.

[6] Frayling TM, Timpson NJ, Weedon MN *et al*. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007;**316:**889–94.

[7] The Wellcome Trust Case Control Consortium. Genome-wide association study between 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;**447:**661–78.

[8] Colhoun HM, McKeigue PM, Davey-Smith G. Problems of reporting genetic associations with complex outcomes. *The Lancet* 2003;**361:**865–72.

[9] Wacholder S, Chanock S, Garcia-Closas M, El-ghormli L, Rothman N. Assessing the probability that a postitive report is false: an approach for molecular epidmiology studies. *J Nat Cancer Inst* 2004;**96:**434–42.

[10] Thomas DC, Clayton DG. Betting odds and genetic associations. *J Nat Cancer Inst* 2004;**96:**421–23.

[11] Ioannidis JPA. Why most published research findings are false. *PLoS* 2005;**2:**696–701.

[12] Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* 2007;**81:**208–27.

[13] Goodman SN. *p* values, hypothesis tests and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993;**137:**485–96.

[14] Sellke T, Bayarri MJ, Berger JO. Calibration of *p* values for testing precise null hypotheses. *Am Stat* 2001;**55:** 62–71.

[15] Westfall PH, Johnson WO, Utts JM. A Bayesian perspective on the bonferroni adjustment. *Biometrika* 1995;**84:**419–27.

[16] Nyholt DR. A simple correction for multiple testing for single nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74:**765–69.

[17] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc, Ser B* 1995;**57:**289–300.

[18] Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Nat Acad Sci* 2003;**100:** 9440–45.

[19] Storey JD. The positive false discovery rate: A Bayesian interpretation and the *q*-value. *Ann Stat* 2003;**31:**2013–35.

[20] Wakefield JC. Bayes Factors for Genome-Wide Association Studies. Comparison with *p*-values and Power Calculations. *Submitted*, 2007.

[21] Kass R, Raftery A. Bayes factors. *J Am Stat Assoc* 1995;**90:**773–95.

[22] Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantative traits. *PLOS Genet* 2007;**3:**1296–1308.

[23] NCI-NHGRI Working Group on Replication in Association Studies. Replicating genotype-phenotype associations. *Nature* 2007;**447:**655–60.

[24] Spiegelhalter DJ, Abrams K, Myles JP. *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. Chichester: Wiley, 2004.

[25] Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997;**53:**1253–61.

[26] Stirling WD. Enhancements to aid interpretation of probability plots. *The Statistician* 1982;**31:**211–20.

[27] Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;**55:**997–1004.

[28] Johnson VE. Bayes factors based on test statistics. *J Royal Statis Soc, Ser B* 2005;**67:**689–701.

[29] Johnson VE. Properties of Bayes factors based on test statistics. *Scand J Stat* 2007. Published on-line, October 31st, 2007.

[30] Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;**39:** 906–13.

Appendix 1

Let $S = -\log_{10}$ BF denote the log to the base 10 of the approximate Bayes factor. The latter is a function of Z^2, which is χ^2_1 under the null, and the standard error $\sqrt{V}$ which differs between SNPs. To evaluate the expected numbers of S that exceed a threshold s_0 we note that for fixed V:

$$\Pr(S \geq s_0|V) = \Pr\left(Z^2 \geq \frac{-2\log_{10}\{\sqrt{1-r}/10^{s_0}\}}{r}|V\right)$$

where $r = W/(V+W)$. Across all SNPs we take the expectation over the distribution of V:

$$\Pr(S \geq s_0) = \mathrm{E}_V[\Pr(S \geq s_0|V)]$$

so that we simply have the average of χ^2_1 tail errors.

For evaluating the *P*-values we examine the tail areas for each SNP *conditional* on the variance V and so the *P*-values are identical to those obtained for the *P*-values based on the Wald statistic Z.

Appendix 2

Suppose we have results from two independent studies and that for a particular SNP, $\widehat{\theta}_1$ has distribution $N(\theta, V_1)$, and $\widehat{\theta}_2$ has distribution $N(\theta, V_2)$, where we have assumed a common log odds ratio θ is being estimated. After seeing the first stage data only, the posterior distribution $\theta|\widehat{\theta}_1$ has mean and variance

$$\mu_1 = \mathrm{E}[\theta|\widehat{\theta}_1] = r\widehat{\theta}_1$$
$$\sigma_1^2 = \mathrm{var}(\theta|\widehat{\theta}_1) = rV_1$$

where $r = W/(V_1+W)$. After seeing both sets of data the posterior distribution $\theta|\widehat{\theta}_1, \widehat{\theta}_2$ has mean and variance

$$\mu_2 = E[\theta|\widehat{\theta}_1, \widehat{\theta}_2] = R\widehat{\theta}_1 V_2 + R\widehat{\theta}_2 V_1$$
$$\sigma_2^2 = \mathrm{var}(\theta|\widehat{\theta}_1, \widehat{\theta}_2) = RV_1V_2$$

where $R = W/(V_1W + V_2W + V_1V_2)$. For both stages a 95% posterior credible interval for the relative risk e^{θ} is given by

$$\exp(\mu \pm 1.96 \times \sigma)$$

with substitution of the appropriate μ, σ.

The Bayes factor summarizing the information with respect to H_0 and H_1 in the two studies is given by:

$$\mathrm{ABF}(\widehat{\theta}_1, \widehat{\theta}_2) = \sqrt{\frac{W}{RV_1V_2}} \times \exp\left\{-\frac{1}{2}\left(Z_1^2RV_2 + 2Z_1Z_2R\sqrt{V_1V_2} + Z_2^2RV_1\right)\right\}$$

where $Z_1 = \widehat{\theta}_1/\sqrt{V_1}$ and $Z_2 = \widehat{\theta}_2/\sqrt{V_2}$ are the usual Z statistics. Note that if the first and third terms in the exponent are large then the Bayes factor will be small and will favour the alternative; if Z_1 and Z_2 are of the same sign then the second term will also suggest the alternative, but if they are of opposite sign then the evidence in favour of H_0 will increase as we would expect. Care should be taken in examining summary measures only since two small Bayes factors (or P-values) may be associated with effects in opposite directions, which obviously does not correspond to strong evidence of the alternative; the above combined Bayes factor automatically penalizes such a situation.