



UW Biostatistics Working Paper Series

7-11-2007

Reporting and Interpretation in Genome-Wide Association Studies

Jon Wakefield

University of Washington, jonno@u.washington.edu

Suggested Citation

Wakefield, Jon, "Reporting and Interpretation in Genome-Wide Association Studies" (July 2007). *UW Biostatistics Working Paper Series*. Working Paper 311.

<http://biostats.bepress.com/uwbiostat/paper311>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

Reporting and Interpretation in Genome-Wide Association Studies

Jon Wakefield

International Agency for Research on Cancer, Lyon, France

Departments of Statistics and Biostatistics, University of Washington, Seattle, USA

Summary

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

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 SNPs³. 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 1,363 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 2,617 cases and 2,894 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 reported⁵⁻⁷, and many more will follow in the near-future.

This exciting development produces new challenges in terms of statistical analysis and interpretation⁸⁻¹¹. 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:

1. Ranking the associations in order to determine a list of SNPs to carry forward to the next stage of study, when the size of the list has already been decided upon.
2. Calibrating inference to allow, for example, the number of false discoveries and false non-discoveries, the size of the list to be estimated, or the probability of the null given the data, to be estimated 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 study⁷ 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, an approximate Bayes factor was recently proposed¹², and it on this quantity that we concentrate upon.

Methods

Consider a typical investigation in which for each SNP we wish to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, where θ is the log odds ratio for which 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) with T the estimate of the log odds ratio; in large samples 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 relevant¹³. This has led 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 in only 0.3% of data sets”; this is incorrect since we have chosen to report the *exact* cut-off. 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. Rearrangement of (1) 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, which varies across SNPs as a function of the 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:

$$\text{Posterior Odds of } H_0 > \{-2.72 \times p \times \log p\} \times \text{Prior Odds of } H_0 \quad (3)$$

which is valid for $p < 0.37$, 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 five prior choices: $\pi_0 = 0.9, 0.95, 0.99, 0.999, 0.9999$. For a p -value of 10^{-5} and $\pi_0 = 0.9999$ we have $\Pr(H_0 | \text{data}) \geq 0.76$, 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 1 (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 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; if 500,000 SNPs are examined, for example, then even if the null is true for all tests we would still expect to see four p -values less than 10^{-5} .

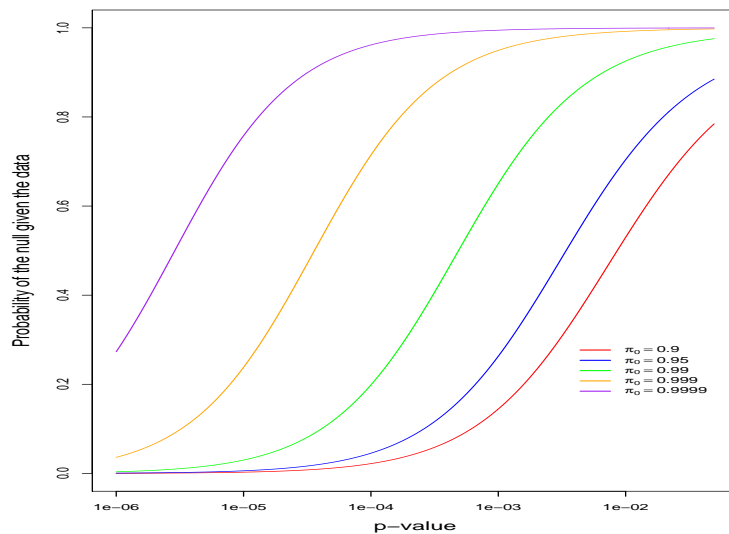


Figure 1: Lower bound on the posterior probability of the null, as a function of the p -value, and the prior on the null.

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 posterior odds of H_0 may be large if $p(\text{data} | H_1)$ is small also. If we have high power then there is stronger evidence in the data in favor of the alternative, when compared with the situation in which we have low power; hence the Bayes factor is quantifying what is intuitively sensible. We return to this issue subsequently.

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 criteria for calling an association noteworthy; with such a criteria, k is the number of tests called noteworthy. The number of false discoveries is V , and the number of false non-discoveries is T . In a GWAS we wish to make V and T as small as possible so that S is 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, $\Pr(V \geq 1)$ – there is an implicit prior assumption that the probability that *all* tests are null is not small¹⁵, since 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 missing associations, and making some false discoveries is not to high a cost to pay in order to get more true hits. By overly protecting against false discoveries one loses power in detecting real associations.

	Non-Noteworthy	Noteworthy	
H_0	U	V	m_0
H_1	T	S	m_1
	$m - k$	k	m

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

More recently, Benjamini and Hochberg¹⁶ suggested a powerful and simple method for controlling the frequentist expected FDR, that is the proportion of rejected tests that are truly null: $E[\frac{V}{k}]$. Subsequently, Storey and colleagues^{17;18} have advocated the use of q -values, a refinement that provide a means of calibrating p -values in terms of the FDR. Specifically, 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

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

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 Storey¹⁸ 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 have the q -value $q(t_{\text{obs}})$ where $\alpha(t_{\text{fix}}) = p$. 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 Storey¹⁸ 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 approach that has much appeal.

The false non-discovery rate (FNR) is defined as $E[\frac{T}{m-k}]$ 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, T will still be small and so $E[\frac{T}{m-k}]$ will also be close to zero and difficult to accurately estimate. The ratio of the non-null associations missed $\frac{T}{m_1}$ (i.e. 1-sensitivity) is clearly of interest, but difficult to estimate since T and m_1 are both 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 colleagues⁹, 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 = $\Pr(\text{data} | \theta_1)$ 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 $\frac{p}{\text{power}}$, which again illustrates that when a set of tests differ in their power the rankings of p -values and FPRP will differ also, with FPRP giving more weight to H_1

when the power is high. The functional form of (5) is familiar to epidemiologists; the baseline odds of the event H_0 , and is revised in light of the odds ratio $\frac{p}{\text{power}}$, 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 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 $\Pr(H_0| |T| > t_{\text{obs}}) \leq \Pr(H_0| T = t_{\text{obs}})$ 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, once one knows the side then a single tail area is appropriate. With respect to frequentist properties FPRP does not provide control of FDR because a variable threshold is used which does not allow long-run frequencies to be calculated – in particular 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 .

The Bayesian False Discovery Probability

For the ranking of associations we have seen that following 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 the latter 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.

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

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.

The decision theory solution is to report H_1 if the

$$\text{Posterior Odds of } H_0 < \frac{C_{\text{FND}}}{C_{\text{FD}}}. \tag{6}$$

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 less than 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 less than 4, i.e. if the posterior probability of H_1 , $\Pr(H_1 | \text{data})$, is greater than 0.2. We now discuss 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 was evaluated under a single alternative, an alternative approach is to place a prior on plausible values of θ . The denominator of the Bayes factor is then given by

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

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 developed¹², and takes as data the estimate of the log odds ratio, $\hat{\theta}$. The asymptotic distribution of the estimator is $N(\theta, V)$, where θ is the true value and \sqrt{V} is the standard error of this estimator, provides the likelihood in the evaluation of the Bayes factor. As prior we take a normal distribution centered on zero and with variance W – 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 $Z = \hat{\theta}/\sqrt{V}$ 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 later. The posterior odds is given by

$$\text{Posterior Odds of } H_0 \text{ Given } \hat{\theta} = \text{ABF} \times \text{Prior Odds of } H_0$$

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 $\Pr(H_0 | \text{data})$ over all W , Sellke et al.¹⁴). The Bayesian false discovery probability (BFDP) is given by

$$\text{BFDP} = \frac{\text{ABF} \times \text{Prior Odds}}{1 + \text{ABF} \times \text{Prior Odds}}$$

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”¹⁹ may be presented in terms of $-\log_{10}\text{BF}$. It turns out that, although the *rankings* of the approximate Bayes factors and p -values will in general differ, 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 $Z = \hat{\theta}/\sqrt{V}$ (Appendix 1 contains details). The latter follows because for fixed V the approximate Bayes factor is simply a transformation of Z . This fact allows the expected numbers falling beyond $-\log_{10}\text{BF}$ thresholds to be easily calculated. Hence evidential guidelines may be based on the frequentist properties of the Bayes factor by comparing the observed number falling beyond thresholds of $-\log_{10}\text{BF}$ with those expected under the null, a point that we illustrate in the simulations section. Similar ideas have appeared recently in the genetics literature²⁰. 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.

Given V the ABF is a simple function of Z which means that power calculations are straightforward. If we decide to call a SNP noteworthy if the posterior odds of H_0 drop below C , then the power to detect a relative risk of RR_1 is given by

$$\Pr \{ \text{ABF}(W, Z, V) \times \pi_0 / (1 - \pi_0) < C | \text{RR}_1 \} = \Pr \{ Z^2 \geq g(C, \pi_0, W, V) | \text{RR}_1 \}$$

and under H_1 Z^2 is a non-central χ^2 random variable. For example, Figure 2 illustrates for sample sizes of 1000 and 2000, under a dominant genetic model. 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 essential²¹. Combination of data across studies (assuming that the effect is

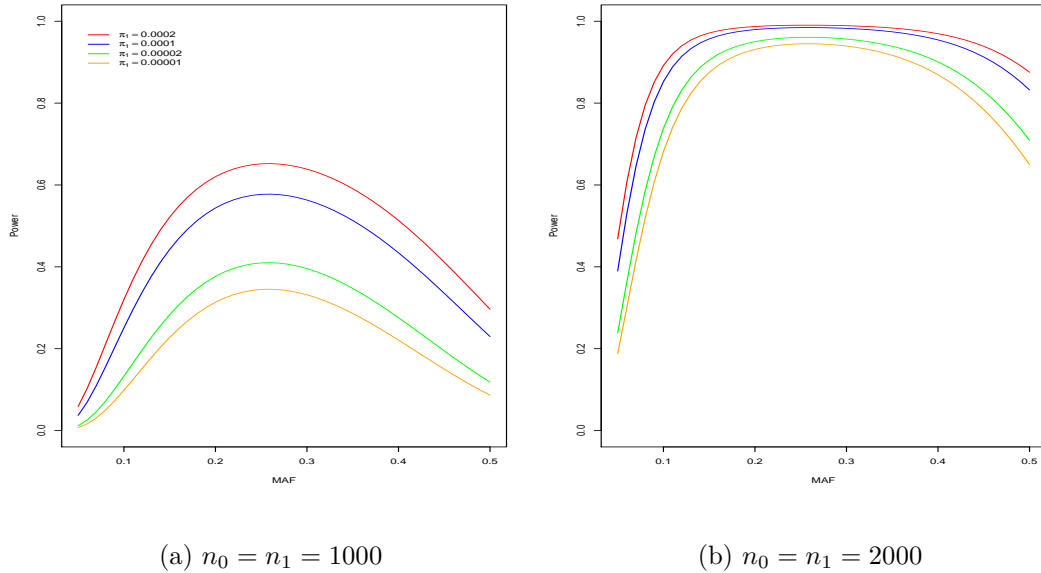


Figure 2: Power to detect a relative risk of 1.5, as a function of MAF and π_1 , the probability of a non-null association. The genetic model is dominant, and the ratio of costs is 10 so that the null is rejected if the posterior probability of the alternative is greater than 0.09.

constant across studies) to produce a Bayes factor summarizing both sets of data is straightforward since

$$\text{ABF}(\hat{\theta}_1, \hat{\theta}_2) = \text{ABF}(\hat{\theta}_1) \times \text{ABF}(\hat{\theta}_2 | \hat{\theta}_1) \tag{7}$$

where $\text{ABF}(\hat{\theta}_2 | \hat{\theta}_1) = p(\hat{\theta}_2 | H_0) / p(\hat{\theta}_2 | \hat{\theta}_1, H_1)$ and $p(\hat{\theta}_2 | \hat{\theta}_1, H_1) = E_{\theta | \hat{\theta}_1} [p(\hat{\theta}_2 | \theta)]$ 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 $\hat{\theta}_2$ under the alternative we average over the posterior for θ given $\hat{\theta}_1$; this contrasts with the evaluation of the probability for $\hat{\theta}_1$ under the alternative for which we average over the prior for θ .

We now turn to the thorny issue of choice of π_0 . As more genome-wide association studies are carried out lower bound 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. If an estimate of π_0 less than 1 is obtained using the q -values methodology then this may be used as a non-subjective reference point.

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 favor of the alternative ($-\log_{10}\text{BF}$), so that a value of 2 means that the data are 100 times more likely under the alternative than the null. On the horizontal axis we

have the minor allele frequency (MAF), which drives the power. We assume a model in which we have a recessive genetic model and assume that the odds ratio is less than 1.5 with probability 0.975. We concentrate on the curve labeled $p = 0.00005$. For MAF close to 0.05 (low power) the evidence in favor of the alternative is small because to obtain such a small p -value requires a large $\hat{\theta}$ which is unlikely under the prior. As the MAF increases the power also increases and the evidence in favor of the alternative consequently increases also. For MAF close to 0.5 we have strong power the evidence starts to decrease, in contrast to p -values for which it is well-known that the null be rejected for large sample sizes, even if $e^{\hat{\theta}}$ only differs from unity by a small amount. This behavior is also discussed by Spiegelhalter et al.²².

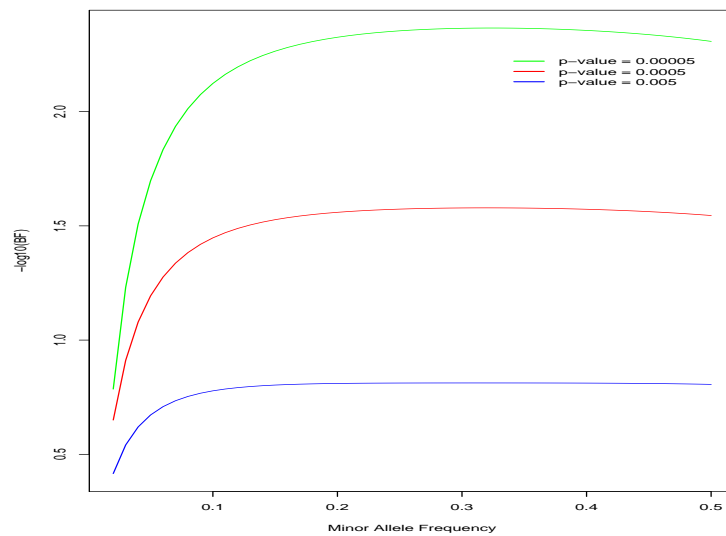


Figure 3: Evidence in favor of the alternative versus the null for three different p -values, as a function of MAF.

Operating Characteristics via Simulation

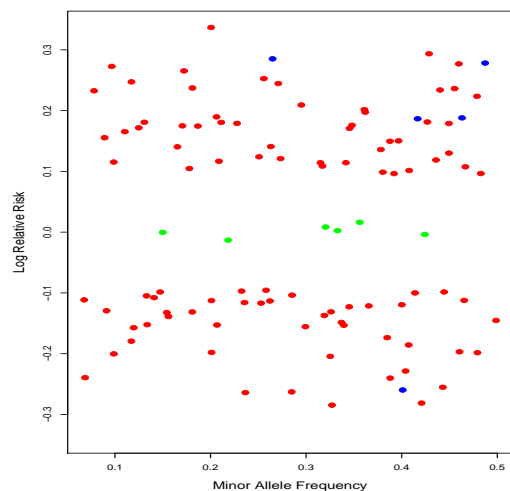
We carry out a simulation study in which there are 3,000 cases and 3,000 controls. We assume that 317,000 SNPs are to be examined of which 100 are truly associated with disease. We take a linear additive model²³ with θ the log relative risk associated with two copies of the mutant allele. We generate the log relative risks for these 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 MAF. The four panels of this figure 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 ratios of costs, $C_{\text{FND}}/C_{\text{FD}}$ 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.

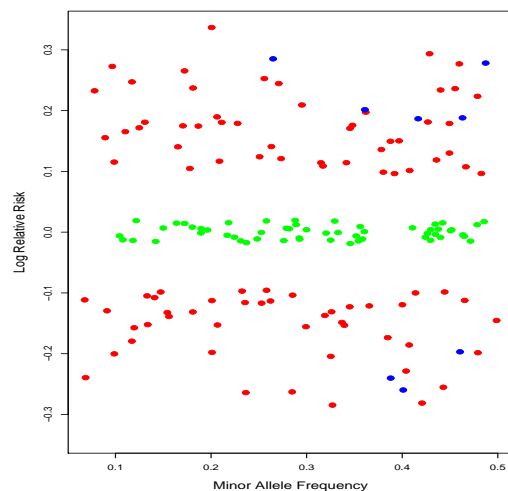
Figure 5 shows the number of SNPs that we need to call noteworthy to obtain a specified number of “hits”. The dashed line is the line of $y = x$ and a perfect procedure would follow this line. We see that the signal is only strong for the first few SNPs (the two most noteworthy SNPs under ABF and the p -value are true associations, the third is not) and 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 median position on the list for non-null SNPs for which the Bayes factor gave the better ranking is 5,123, whereas the median on the p -value list is 65,588 again illustrating that the SNPs that are picked up first by the p -value are ones that would not be passed to a second phase since the associated signal was very weak.

Figure 7 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 $100/317,000=0.0003$). Figure 7(a) plots the expected versus observed $-\log_{10} p$ -values and indicates an excess in the tail; p -values based on the statistic ABF are identical to the p -values based on the Wald statistic Z . Table 3 gives the expected number of tests falling within different bands under the null, along with the observed numbers. We would conclude that the top two SNPs appear to be real hits while approximately 4 of the next 9 hits are real. This table differs from that based on p -values since the MAFs of the 317K SNPs in this dataset are explicitly considered (in other words, Table 3 accounts for power).

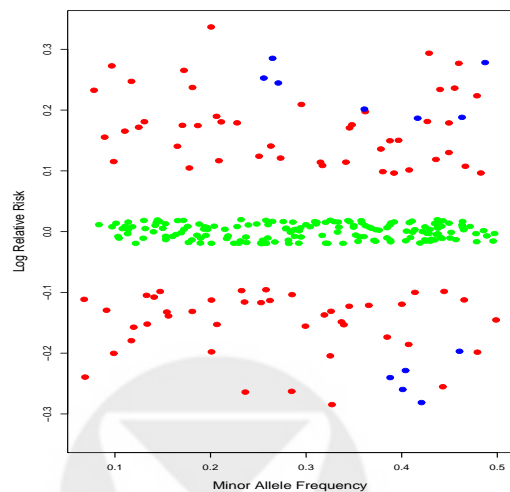
Figure 7(b) plots q -values against p -values and illustrates that most of the q -values are close to 1. The expected number of false discoveries is the q -value times the number of SNPs called noteworthy at that threshold, and goes up rapidly with the number of true discoveries (Figure 7(c)). Also plotted is the Bayesian estimate of the number of expected false discoveries, which shows similar behavior to the q -value at least for the first 50 hits (after this the q -value display anomalous behavior).



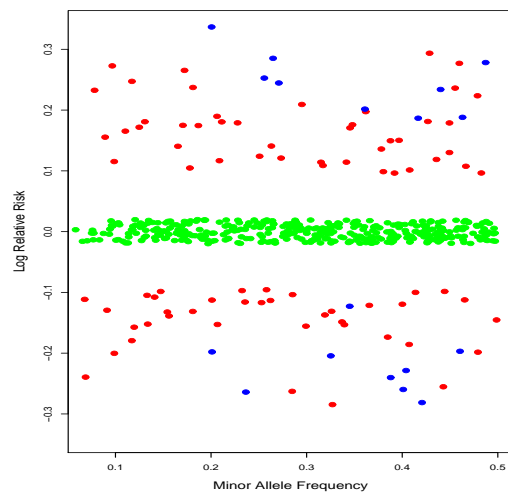
(a) $C_{\text{FND}}/C_{\text{FD}} = 4, S = 5, k = 11$



(b) $C_{\text{FND}}/C_{\text{FD}} = 20, S = 8, k = 63$



(c) $C_{\text{FND}}/C_{\text{FD}} = 50, S = 12, k = 176$



(d) $C_{\text{FND}}/C_{\text{FD}} = 100, S = 18, k = 388$

Figure 4: Discoveries (blue circles), non-discoveries (red circles) and false discoveries (green circles) using BFD 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). $C_{\text{FND}}/C_{\text{FD}}$ is the ratio of costs, S the number of true discoveries, and k the total number of SNPs called noteworthy.

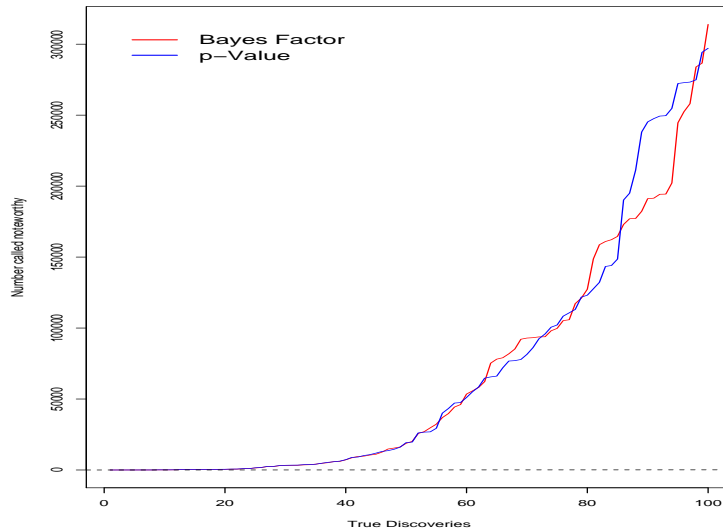


Figure 5: Number of SNPs called noteworthy in order to detect a specified number of true discoveries, with noteworthiness based on p -values and BFDP and FPRP. The dashed line is the line of equality and shows that after the first few hits the curve moves increasingly away from the dashed line demonstrating that the FDR increases rapidly as the length of the list is increased.

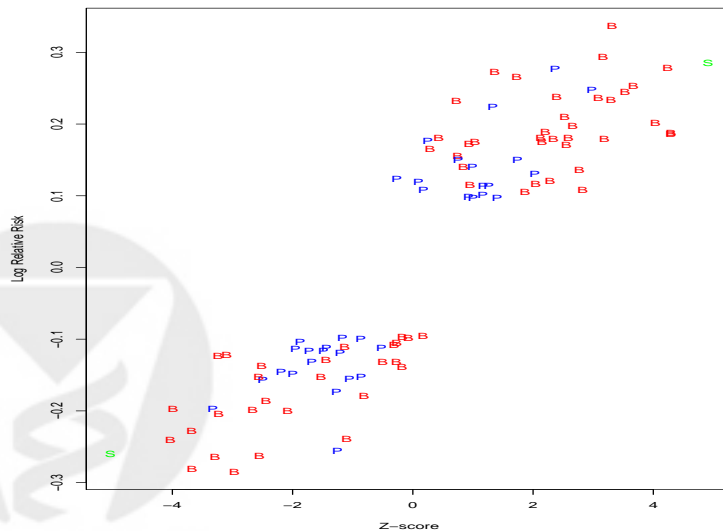
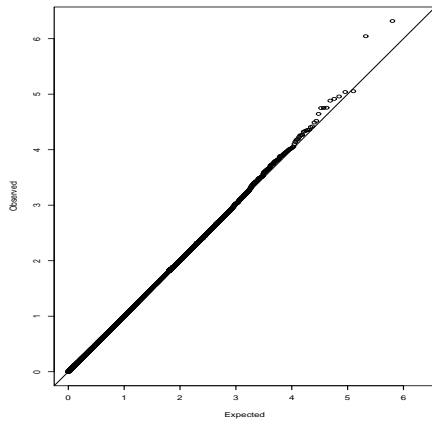
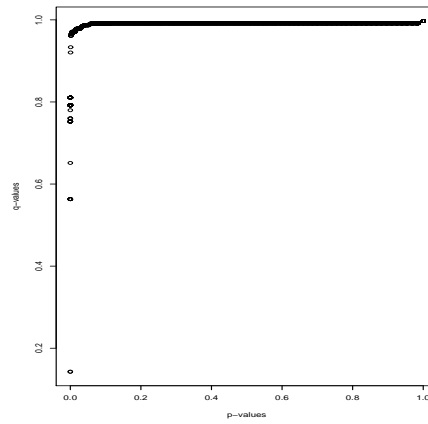


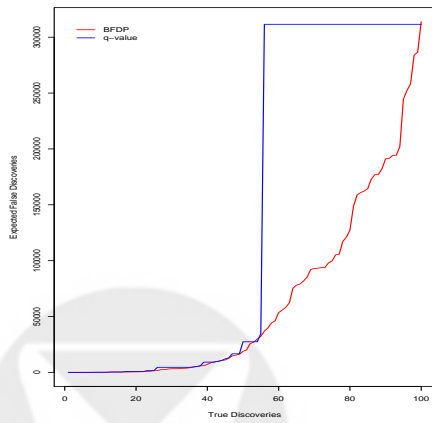
Figure 6: Log relative risks versus Z -scores for the 100 non-null SNPs; the 63 points marked “B” had lower rankings on the Bayes factor list, while the 35 marked “P” had lower rankings on the p -value lists; the two SNPs marked “S” had identical rankings (and were the first two found).



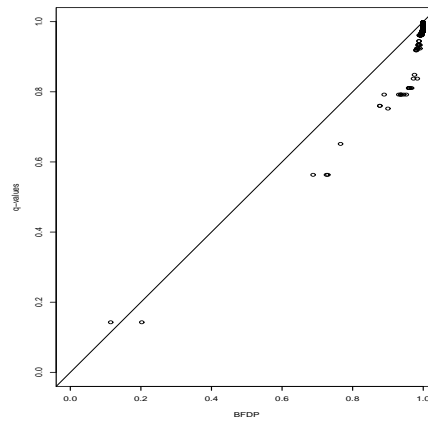
(a) QQ plot of $-\log_{10} p$ -values



(b) q -values versus p -values



(c) Expected false discoveries versus true discoveries



(d) q -values versus BFDp

Figure 7: BFDp, p - and q -value summaries.

Bayes Factor	$-\log_{10}\text{BF}$	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

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

Examples from the Literature

Table 4 gives point estimates of odds ratios and confidence intervals (CIs) from a number of genome-wide association studies that have appeared in the literature, and for which we have calculated Bayes factors and BFD_P under three prior distributions with proportions of non-null SNPs of 1/5,000, 1/10,000 and 1/50,000.

The estimate (CI) in the first row of the table corresponds to an association found in 1,924 type 2 diabetes patients⁶ when compared to 2,938 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 3,757 type 2 diabetes cases and 5,346 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 greater than 0.9. However, combining the Bayes factors using equation (8) in Appendix 2 we obtain a combined $-\log_{10}\text{BF}$ of 13.8. Hence the data are overwhelmingly in favor 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 (5%, 50%, 95%) 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 $\hat{\theta}_1$ 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%.

In the third and fourth rows of Table 4 we examine an estimate of 11.14 reported in 96 patients previously described with wet age-related macular degeneration, as compared to 130 age-matched controls (97,824 SNPs were examined in total). The small sample size means that the data are only 20 times more likely under the alternative as compared to the null, giving posterior probabilities for the null close to 1 under each prior, π_0 . The observed estimate is very unlikely under the assumed prior for the size of the effect

which predicts 95% of relative risks to lie between 2/3 and 1.5, and also contributes to the weak evidence here. Changing the prior to have 95% of the mass between 0.2 and 5 greatly increases the evidence though under the prior with the smallest probability of the alternative, there is still a posterior probability of 0.269 for the null. A prior with a wider range is more appropriate here since the case sample was enriched and so we would expect a greater effect.

The estimate and confidence interval for rs8051542 appear as the third SNP in Table 2 of Easton et al.⁵ and summarize the third stage of a GWAS. These data alone do not provide strong evidence of an association (51% posterior probability of the null under the most optimistic prior).

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 factor²⁴ of 1.1233. The last two lines of the table give the Bayes factor and BFDP for two different priors for the size of the relative risk. One prior assumes that with probability 0.95 the relative risk associated with 2 mutant copies is [2/3,1.5] and the other is [0.2,5]. Under the prior that assumes a narrower range of risks the evidence for a non-null association is not strong. In the second stage of the study the relative risk estimate was much smaller (1.45 for two mutant alleles).

SNP ^{REF}	Est	95% C.I.	<i>p</i> -value	−log ₁₀ BF	BFDP with Prior:		
					1/5,000	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
rs10490924 ²⁵	11.14	4.83–25.69	1.6 × 10 ^{−8}	1.28	0.996	0.998	1.00
rs10490924 ^{25*}	11.14	4.83–25.69	1.6 × 10 ^{−8}	5.13	0.036	0.069	0.269
rs8051542 ⁵	1.09	1.06–1.13	2.8 × 10 ^{−6}	3.68	0.511	0.677	0.913
rs7913837 ⁴	2.20	1.57–3.07	4.0 × 10 ^{−6}	2.55	0.933	0.965	0.993
rs7913837 ^{4*}	2.20	1.57–3.07	4.0 × 10 ^{−6}	3.74	0.477	0.646	0.901

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 apart from * for which the 97.5% point was set at 5.

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 multiple-hypothesis tests being 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 or the prior of an 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 the proportion of non-null associations is small in GWASs and more experience of its use in this context is required.

A refinement of FPRP, BFDP has been described here and elsewhere¹², 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, overdispersion (genomic control) to be simply incorporated by multiplying the variance of the odds ratio by the overdispersion factor. Treating the approximate 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. The use of Bayes factors based on test statistics has been previously advocated as a robust and theoretically sound strategy²⁶.

Replacing confidence intervals with p -values does not overcome the problems of reporting when the prior probability of an association is low, since confidence intervals assume that the null has been rejected. The posterior distribution for the relative risk of an association *given* an association (i.e. H_1) is lognormal with parameters $r\hat{\theta}$ and $r\hat{\theta}$. 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.

Throughout we have used the term noteworthy, following Wacholder et al.⁹, but these tests may be alternatively labelled as “anomalous” recognising that the flagged associations may be due to errors in the 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:

1. To rank associations the Bayes factor provides an alternative to the p -value which accounts for power.
2. 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. BFDP may also be used to interpret reported associations.

Appendix 1

Let $S = -\log_{10} \text{BF}$ denote the log to the base 10 of the approximate Bayes factor. The latter is a function of Z^2 , which is χ_1^2 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 have

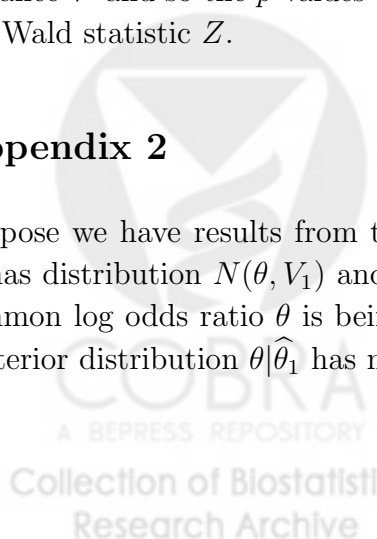
$$\Pr(S \geq s_0) = E_v [\Pr(S \geq s_0|V)]$$

so that we simply have the average of χ_1^2 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, $\hat{\theta}_1$ has distribution $N(\theta, V_1)$ and $\hat{\theta}_2$ 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 $\theta|\hat{\theta}_1$ has mean and variance


$$\begin{aligned} \mu_1 &= E[\theta|\hat{\theta}_1] = r\hat{\theta}_1 \\ \sigma_1^2 &= \text{var}(\theta|\hat{\theta}_1) = rV_1 \end{aligned}$$

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

$$\begin{aligned}\mu_2 &= E[\theta|\hat{\theta}_1, \hat{\theta}_2] = R\hat{\theta}_1V_2 + R\hat{\theta}_2V_1 \\ \sigma_2^2 &= \text{var}(\theta|\hat{\theta}_1, \hat{\theta}_2) = RV_1V_2\end{aligned}$$

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:

$$\text{ABF}(\hat{\theta}_1, \hat{\theta}_2) = \sqrt{\frac{W}{RV_1V_2}} \exp\left\{-\frac{1}{2}\left(Z_1^2RV_2 + 2Z_1Z_2R\sqrt{V_1V_2} + Z_2^2RV_1\right)\right\} \quad (8)$$

where $Z_1 = \hat{\theta}_1/\sqrt{V_1}$ and $Z_2 = \hat{\theta}_2/\sqrt{V_2}$ 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 favor 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 favor 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.

References

- [1] J.N. Hirschhorn and M.J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6:95–108, 2005.
- [2] W.Y.S. Wang, B.J. Barratt, D.G. Clayton, and J.A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6:109–118, 2005.
- [3] C.S. Carlson, M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D.A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, 74:106–120, 2004.
- [4] R. Sladek, G. Rocheleau, J. Ring, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, and B. Balkau et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

- [5] D.F. Easton, K.A. Pooley, A.M. Dunning, P.D.P. Pharoah, D. Thompson, D.G. Ballinger, J.P. Struewing, J. Morrison, H. Field, and R. Luben. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1–9, 2007.
- [6] T.M. Frayling, N.J. Timpson, M.N. Weedon, E. Zeggini, R.M. Freathy, and C.M. Lindgren et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316:889–894, 2007.
- [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*, 447:661–678, 2007.
- [8] H.M. Colhoun, P.M. McKeigue, and G. Davey-Smith. Problems of reporting genetic associations with complex outcomes. *The Lancet*, 361:865–872, 2003.
- [9] S. Wacholder, S. Chanock, M. Garcia-Closas, L. El-ghormli, and N. Rothman. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96:434–442, 2004.
- [10] D.C. Thomas and D.G. Clayton. Betting odds and genetic associations. *Journal of the National Cancer Institute*, 96:421–423, 2004.
- [11] J.P.A. Ioannidis. Why most published research findings are false. *Public Library of Science Medicine*, 2:696–701, 2005.
- [12] J. Wakefield. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, 2007. To appear.
- [13] S.N. Goodman. p values, hypothesis tests and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137:485–496, 1993.
- [14] T. Sellke, M.J. Bayarri, and J.O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55:62–71, 2001.
- [15] P.H. Westfall, W.O. Johnson, and J.M. Utts. A Bayesian perspective on the bonferroni adjustment. *Biometrika*, 84:419–427, 1995.
- [16] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.
- [17] J.D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Science*, 100:9440–9445, 2003.

- [18] J.D. Storey. The positive false discovery rate: A Bayesian interpretation and the q -value. *The Annals of Statistics*, 31:2013–2035, 2003.
- [19] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.
- [20] B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *Public Library of Science Genetics*, 2007. To appear.
- [21] NCI-NHGRI Working Group on Replication in Association Studies. Replicating genotype-phenotype associations. *Nature*, 447:655–660, 2007.
- [22] D J Spiegelhalter, K Abrams, and J P Myles. *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. Wiley, Chichester, 2004.
- [23] P.D. Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, 53:1253–1261, 1997.
- [24] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.
- [25] A. DeWan, M. Liu, S. Hartman, S. Shao-Mon Zhang, D.T. Liu, C. Zhao, P.O.S Tam, W.M. Chan, D.S.C Lam, M. Snyder, C. Barnstable, C.P. Pang, and J. Hoh. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*, 314:989–992, 2006.
- [26] V.E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society, Series B*, 67:689–701, 2005.

