Bayes Factors for Genome-Wide Association Studies: Comparison with *P*-values

Jon Wakefield*

Departments of Statistics and Biostatistics, University of Washington, Box 357232, Seattle, Washington

The Bayes factor is a summary measure that provides an alternative to the *P*-value for the ranking of associations, or the flagging of associations as "significant". We describe an approximate Bayes factor that is straightforward to use and is appropriate when sample sizes are large. We consider various choices of the prior on the effect size, including those that allow effect size to vary with the minor allele frequency (MAF) of the marker. An important contribution is the description of a specific prior that gives identical rankings between Bayes factors and *P*-values, providing a link between the two approaches, and allowing the implications of the use of *P*-values to be more easily understood. As a summary measure of noteworthiness *P*-values are difficult to calibrate since their interpretation depends on MAF and, crucially, on sample size. A consequence is that a consistent decision-making procedure using *P*-values requires a threshold for significance that reduces with sample size, contrary to common practice. *Genet. Epidemiol.* 33:79–86, 2009. © 2008 Wiley-Liss, Inc.

Key words: Bonferroni correction; *P*-values; prior distributions; strength of evidence

*Correspondence to: Jon Wakefield, Departments of Statistics and Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232. E-mail: jonno@u.washington.edu

Received 15 February 2008; Revised 23 April 2008; Accepted 10 June 2008
Published online 18 July 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20359

INTRODUCTION

Genome-wide association studies (GWASs) are an exciting prospect for discovering genetic variants that are detrimental or protective for human disease [Hirschhorn and Daly, 2005; Wang et al., 2005], and a number of important results have already been reported [DeWan et al., 2006; Sladek et al., 2007; Easton et al., 2007; Wellcome Trust Case Control Consortium, 2007].

The most common summary measure for inference regarding a single locus in a GWAS is the *P*-value; *P*-values have a number of well-documented drawbacks, however [Sterne and Smith, 2001; Goodman, 1999; Wacholder et al., 2004]. An acute problem is the difficulty in choosing a significance threshold; the commonest approach is to use a Bonferroni correction to control the family-wise error rate at 5%. Leaving aside for the moment the arbitrariness of the 5% threshold, a serious problem with the use of the *P*-value is the irrelevance of power, which is a function of sample size and minor allele frequency (MAF), in the threshold decision. The use of a single threshold regardless of sample size implicitly implies that the ratio of costs of type I to type II errors varies with sample size. Specifically, consider two situations with low power in the first and high power in the second; if the *P*-value threshold is the same in both situations then one is accepting that the cost of a type II error is higher in the second situation. Later, we give a precise formulation of this argument.

The Bayes factor, defined as the ratio of the probability of the data under the null and alternative hypotheses, provides an alternative to the *P*-value for assessing the consistency of a set of data with a null hypothesis, as compared to an alternative. Bayes factors have been previously discussed in a genome-wide context [Wakefield, 2007; Marchini et al., 2007; Wellcome Trust Case Control Consortium, 2007], and also suggested for other genetic settings [Servin and Stephens, 2007]. The more widespread use of the Bayes factor has been hampered by the need for prior distributions to be specified for all of the unknown parameters in the model, and the need to evaluate multidimensional integrals, a complex computational task. In this paper, we provide a more rigorous derivation of a recently proposed asymptotic Bayes factor [Wakefield, 2007], that avoids each of these requirements. An important additional contribution is the description of a specific prior on the effect size that leads to identical rankings between SNPs based on the *P*-value and on the asymptotic Bayes factor, providing an important conceptual link between the two and allowing a Bayesian interpretation of the *P*-value. In particular, the *P*-value implicitly assumes a prior in which larger effect sizes are anticipated at lower MAFs.

METHODS

BAYES FACTORS

In a GWAS, a common approach [Balding, 2006] is to fit the logistic model:

$$\frac{p_j}{1-p_j} = \exp(\alpha + \theta x_j), \tag{1}$$

where p_j is the probability of disease for an individual with $j = 0, 1, 2$ copies of the minor allele at a particular SNP, and x_j is a variable that depends on the assumed genetic model. For $j = 0, 1, 2$ copies we have $x_j = 0, 1, 1$ for a dominant genetic model, $x_j = 0, 0, 1$ for a recessive genetic model, and $x_j = 0, 1, 2$ for a multiplicative genetic model [Sasieni, 1997]. The general two degree of freedom model can also be considered, but for simplicity of explanation we concentrate on genetic models in which there is a single parameter of interest, θ. Model (1) may be easily extended to include matching and other variables for which adjustment is required, as detailed in the appendix.

Under a rare disease assumption, the relative risk corresponding to departure from the null model is given by $\text{RR} = \exp(\theta)$, and we wish to compare the null hypothesis $H_0 : \text{RR} = 1$ with the alternative $H_1 : \text{RR} \neq 1$. As described elsewhere [Wakefield, 2008], there are two endeavors that may be carried out in the context of a GWAS. The first is to *rank* markers in terms of association, to provide a list of those that should be carried through to a next phase. The second is *calibration* of inference to make a final decision as to whether to call the marker "significant", i.e. associated with disease, or not. Each of these tasks may be carried out using the Bayes factor (BF) given by

$$\text{BF} = \frac{\Pr(\boldsymbol{y}|H_0)}{\Pr(\boldsymbol{y}|H_1)},$$

where $\boldsymbol{y}$ is the observed data, and corresponds to a vector of binary indicators when disease status is the phenotype. If the Bayes factor equals 1 then the data are equally likely under the null and the alternative, and the smaller/larger the value the Bayes factor attains, the more/less the alternative is favored. For a measure of "significance" the posterior odds on H_0 are required:

$$\text{posterior odds on } H_0 \ = \text{BF} \times \ \text{prior odds on } H_0,$$

where the prior odds on H_0 are given by $\pi_0/(1-\pi_0)$, with $\pi_0 = \Pr(H_0)$ the prior probability of the null. Ranking may be carried out directly on the basis of the Bayes factor if the prior odds of no association is constant across all SNPs, since it is the relative value rather than the absolute value that is relevant.

The Bayes factor requires the specification of a prior distribution for all unknown parameters, and for logistic regression models it is computationally expensive to evaluate, which has lead to the search for simple approximations. Previously, [Wellcome Trust Case Control Consortium, 2007], Bayes factors were calculated using the Laplace approximation [Kass and Raftery, 1995]. This approximation can be difficult to implement, however, since a search for the maximum of the multidimensional posterior is required for each association. Below we describe an alternative asymptotic Bayes factor that is based on the output from a simple logistic regression analysis; the only data input required for calculation is a confidence interval for the parameter of interest θ (or equivalently an estimate and standard error). Maximization of a binomial likelihood is required when a logistic regression model is fitted, but this operation is routinely carried out by all statistical packages. The Bayes factor we describe has a simple closed form, which offers a number of additional benefits including ease of power calculations, and straightforward combination of evidence across studies.

Let $\widehat{\theta}$ and $\sqrt{V}$ represent the maximum likelihood estimate (MLE) and standard error from a logistic regression analysis for a particular SNP; V depends on the genetic model, the case and control sample sizes, n_1 and n_0, and on the frequency of individuals with each genotype (equation (8) gives a specific form). Asymptotically, that is as n_0 and n_1 increase, the MLE $\widehat{\theta}$ has the normal distribution $N(\theta, V)$. Combining this "likelihood" with a normal prior, $N(0, W)$, on the log relative risk, θ, gives the asymptotic Bayes factor

$$\text{ABF} = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{z^2}{2}\frac{W}{(V+W)}\right), \qquad (2)$$

where $z^2 = \widehat{\theta}^2/V$ is the usual Wald statistic. High/low values of the asymptotic Bayes factor occur when z^2 is small/large and correspond to evidence for/against the null hypothesis. The appendix provides a rigorous derivation of the ABF.

A great advantage of the Bayes factor, (2), is that is depends on readily available summaries only, though if the sample sizes are not large there may be some loss in efficiency if $\{\widehat{\theta}, V\}$ do not summarize all the information contained in the full data concerning θ.

The crucial difference between inference based on the ABF and on the P-value calculated from the Wald statistic, which equals $2\{1 - \Phi(|z|)\}$ (where $\Phi(\cdot)$ is the distribution function of a standard normal random variable), is that ABF depends, in addition to z, on the power through the asymptotic variance V. The relationship between ABF and V is not monotonic. For fixed z (i.e. fixed P-value) and fixed prior variance W, we examine the behavior of ABF as a function of V. The Bayes factor compares the evidence between H_0 and H_1, assuming that one of them is true. Under the null $\widehat{\theta} \sim N(0, V)$ while under the alternative $\widehat{\theta} \sim N(0, V + W)$ and the Bayes factor is the ratio of these two densities evaluated at the observed $\widehat{\theta}$. Figure 1 illustrates the behavior of the evidence for the alternative (1/ABF) against V for $z = 4$ and $W = 0.21^2$ (corresponding to a 95% belief that the relative risk is less than 1.5). For low values of V (high power), the evidence for H_1 is not strong since although the data (the z-score) are unlikely under H_0, they

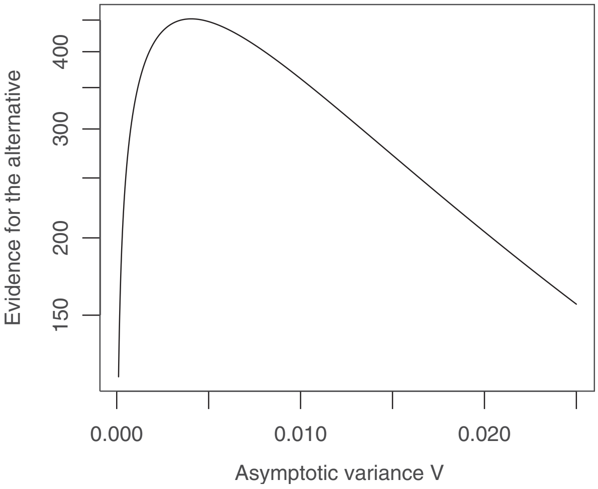

Fig. 1. Evidence in favor of the alternative as a function of the asymptotic variance of the estimator, *V*, for fixed *P*-value (i.e. fixed *z*). Small and large values of *V* correspond to high and low power, respectively; *V* is proportional to the reciprocal of the sample size.

are unlikely under H_1 also—this behavior contrasts with the P-value under which very small departures from H_0 provide small P-values when the power is high. The evidence for H_1 increases rapidly as the power decreases, to a maximum at $V = W/(z^2-1)$. Beyond this point, there is a decrease in the evidence for H_1 since the power is not sufficient to provide strong evidence.

We briefly examine the asymptotic properties of ABF. For simplicity, assume that $n_0 = n_1 = n$ and let $\widehat{\theta}_n$ be the MLE based on case and control samples of size n. The asymptotic variance $V_n = F/n$ where F depends on the genotype frequencies and the genetic model but not on n. We have $\sqrt{n}(\widehat{\theta}_n - \theta) \to N(0, F)$ as $n \to \infty$, by the properties of MLEs and

$$\begin{aligned}\log \mathrm{ABF}_n &= \frac{1}{2}\log\left(1+\frac{Wn}{F}\right) - \frac{n\widehat{\theta}_n^2}{2F}\times\frac{W}{F/n+W}\\ &= \log\left(1+\frac{Wn}{F}\right) - \frac{\{\sqrt{n}(\widehat{\theta}_n-\theta)+\sqrt{n}\theta\}^2}{2F}\\ &\quad\times\frac{W}{F/n+W}.\end{aligned}$$

When $\theta = 0$, $\log \mathrm{ABF}_n \to \infty$ as $n \to \infty$ and when $\theta \neq 0$, $\log \mathrm{ABF}_n \to -\infty$ so that ABF is consistent under both the null and the alternative—the correct model is chosen with probability 1 as the sample sizes increase.

When data from two studies (or phases) are available, with estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ and standard errors $\sqrt{V_1}$ and $\sqrt{V_2}$, the Bayes factor for the combined evidence is given by

$$\begin{aligned}&\mathrm{ABF}(\widehat{\theta}_1, \widehat{\theta}_2)\\ &= \sqrt{\frac{W}{RV_1V_2}}\exp\left\{-\frac{1}{2}(z_1^2RV_2 + 2z_1z_2R\sqrt{V_1V_2} + z_2^2RV_1)\right\},\end{aligned} \tag{3}$$

where $R = W/(V_1W + V_2W + V_1V_2)$ and z_1 and z_2 are the z-statistics arising from the two studies. Strong evidence in favor of the alternative requires large $|z|$ statistics *of the same sign*. It is straightforward to extend this formula to three or more studies.

For both single and multiple studies, and with $\mathrm{PO} = \pi_0/(1-\pi_0)$ the prior odds on the null, the posterior probability of H_0 is given by the *Bayesian False Discovery Probability* [Wakefield, 2007]

$$\mathrm{BFDP} = \frac{\mathrm{ABF}\times\mathrm{PO}}{1+\mathrm{ABF}\times\mathrm{PO}}. \tag{4}$$

PRIOR SPECIFICATION

The approximate Bayes factor sidesteps the need for specification of a prior on the nuisance parameters, but requires a prior for the log relative risk of interest, θ, which is taken as normally distributed with mean 0 and variance W. The latter variance is the single specification that is needed, and we describe three particular choices.

Effect-MAF independence: The simplest choice is to take the variance, W, as independent of the MAF. The prior distribution of the relative risk, $\exp(\theta)$, is lognormal and we may specify an upper value RR_u, above which we believe that relative risks will occur with low probability. If the prior probability of a relative risk above RR_u is q we obtain $W = \{\log \mathrm{RR}_\mathrm{u}/\Phi^{-1}(1-q)\}^2$. For example, for a 5% chance that relative risks are above 2, $\mathrm{RR}_\mathrm{u} = 2, q = 0.05$, and $W = 0.42^2$.

Effect-MAF dependence: It has been argued that larger genetic effects will be associated with smaller MAFs [for discussion, see Wang et al., 2005], in which case the variance W should depend on the MAF. Selection would imply that large detrimental relative risks should not occur for common variants. A simple form that can mimic this behavior is

$$W(M) = \delta_0 \exp(-\delta_1 \times M), \tag{5}$$

where M is the MAF, and the parameters $\delta_0, \delta_1 > 0$ are chosen a priori. For example, one may set the upper bound on the prior for the relative risk at two values of the MAF and then solve for δ_0, δ_1. Specifically, let M_lo and M_hi be the rare and non-rare MAFs at which we specify relative risks of $\mathrm{RR}_\mathrm{u}^\mathrm{lo} > \mathrm{RR}_\mathrm{u}^\mathrm{hi}$, above each of which we believe relative risks will lie with probability q. The variances at the rare and non-rare variants are

$$\begin{aligned}W_\mathrm{lo} &= \{\log(\mathrm{RR}_\mathrm{u}^\mathrm{lo}/\Phi^{-1}(1-q)\}^2,\\ W_\mathrm{hi} &= \{\log(\mathrm{RR}_\mathrm{u}^\mathrm{hi}/\Phi^{-1}(1-q)\}^2,\end{aligned}$$

which may be solved to give

$$\begin{aligned}\delta_1 &= \frac{\log(W_\mathrm{lo}) - \log(W_\mathrm{hi})}{M_\mathrm{hi} - M_\mathrm{lo}},\\ \delta_0 &= W_\mathrm{lo}\exp(\delta_1 \times M_\mathrm{lo}).\end{aligned}$$

An implicit P-value prior: In general, both ranking and significance of SNPs will differ when assessed using Bayes factors and P-values, and it is of great interest to see when the approaches can be reconciled. This unification occurs when the Bayes factor depends on the data only through z^2 in a monotonic fashion, since this is the only function of the data that determines the P-value. This is achieved if we eliminate V from ABF and occurs if we take the variance to be proportional to the asymptotic variance of the MLE:

$$W(M, n_0, n_1) = K \times V(M, n_0, n_1), \tag{6}$$

where K does not depend on the data (and, in particular, does not depend on n) and for simplicity, we have assumed that the marker in Hardy-Weinberg equilibrium so that the genotype frequencies are a function of the MAF, M. We have emphasized that the variance of the estimator, and hence the prior variance, depend on n_0, n_1, and M. This prior gives

$$\mathrm{ABF} = \sqrt{1+K}\exp\left(-\frac{z^2}{2}\frac{K}{(1+K)}\right). \tag{7}$$

We want K to be independent of the data because we want a Bayes factor that depends on the z score (and therefore the P-value) only. Such a prior was discussed with respect to the use of P-values in the context of a normal model by Cox and Hinkley [1974, pp 395–399]. Under the P-value prior, (6):

$$\begin{aligned}\log \mathrm{ABF}_n &= \frac{1}{2}\log(1+K)\\ &\quad - \frac{\{\sqrt{n}(\widehat{\theta}_n-\theta)+\sqrt{n}\theta\}^2}{2F}\times\frac{K}{1+K},\end{aligned}$$

which tends to $\frac{1}{2}\log(1+K)$ when $\theta = 0$ and not ∞ as is desirable—the P-value ABF is inconsistent under the null. This is a consequence of the fact that for a fixed P-value threshold p_T the null will be incorrectly rejected a proportion

p_T of the time under repeated sampling (which is intuitively why we need the P-value threshold to decrease with increasing sample size, all other things being equal).

The implicit P-value prior (6) gives stronger belief in a small effect size when the sample size is large and/or the MAF is not rare (since in both cases V is small). The dependence on the sample sizes n_0 and n_1 is alarming and does not make sense in the genome-wide context (in contrast to a designed experiment in which one would pick larger sample sizes when the expected effect was small, behavior that would be reflected in the prior also). For ranking, n_0 and n_1 will be constant across SNPs (give or take missing values) and so this aspect of the prior is not troubling.

The prior association between effect size and MAF is in the direction expected (larger effects at rarer MAFs) and is determined in a very specific manner by the dependence of V on the MAF. A convenient form for the asymptotic variance of $\widehat{\theta}$ is available from a score test [Slager and Schaid, 2001], and equals the variance estimate of the MLE from a logistic regression model based on the information, where the latter is evaluated at the null. For simplicity, we assume Hardy-Weinberg equilibrium to give:

$$V = \frac{n_0 + n_1}{\begin{array}{c} n_0 n_1[(1-M)^2 x_0^2 + 2M(1-M)x_1 + M^2 x_2^2 \\ -\{(1-M)^2 x_0 + 2M(1-M)x_1 + M^2 x_2\}^2] \end{array}} \tag{8}$$

where M is the MAF, and x_0, x_1, and x_2 depend on the genetic model (examples of which were given above). The variance of the P-value prior (6) is not a simple function of the MAF, and so we graphically illustrate the shape, plotting against the comparison prior, (5). In Figure 2, we plot the priors for MAFs of 0.10, 0.30, and 0.50; the P-value priors are drawn as the solid lines while the comparison priors are the dashed lines. For the comparison prior δ_0 and δ_1 were chosen to give 95% points of 2.5 and 1.2 at MAFs of 0.05 and 0.50; K was chosen to approximately match the two priors at MAFs of 0.10 and 0.30. The variance of the comparison prior can decrease rapidly with increasing MAF (with large values of δ_1), while the P-value prior exhibits a more gradual decrease.

To rectify the undesirable dependence of the prior on sample size, while retaining the effect-MAF relationship implied by the P-value, one can take $W(M) = K^\star \times F(M)$, where $F(M)$ is given in (8) and does not depend on sample size. For simplicity assume $n = n_0 = n_1$; under the new prior:

$$\text{ABF} = \sqrt{1 + nK^\star} \exp\left[-\frac{z^2}{2}\frac{nK^\star}{(1+nK^\star)}\right] \tag{9}$$

$$= \sqrt{1 + nK^\star} \exp\left[-\frac{1}{2}\Phi^{-1}\left(1 - \frac{p}{2}\right)^2 \frac{nK^\star}{1+nK^\star}\right], \tag{10}$$

which depends on n, in contrast to the form derived from the prior $W = K \times V$. Hence two ABFs calculated from (10) with the same P-value, but with different sample sizes will provide different evidence for/against the null. This consequence is not pertinent to ranking, since rankings are based on comparisons with fixed sample sizes. Good [1992] reviews Bayesian and frequentist approaches to hypothesis testing and derives a "$\sqrt{n}$" rule for standardizing P-values; an expression is derived for a Bayes factor [Good, 1992, equation (4)] that has the same asymptotic behavior as (10).

THE SPECIFICATION OF P-VALUE THRESHOLDS

We have shown that rankings with P-values and Bayes factors based on (6) will be identical, but for calibration the two approaches are more difficult to unify. As summarized elsewhere [Wakefield, 2007] the Bayesian decision theory approach to calibration is to specify the costs of false non-discovery C_{FND} and false discovery C_{FD}, define $R = C_{\text{FND}}/C_{\text{FD}}$, and then flag the SNP as "significant" if the posterior odds on H_0 drop below the ratio R. Hence an association will be called noteworthy if

$$\text{ABF} \times \text{PO} < \text{R} \tag{11}$$

so that there are three elements to the decision problem: the ratio of the probabilities of the data under null and alternative, ABF, the prior odds on H_0, PO, and the ratio of costs, R.

It is enlightening to use a formal decision theory approach to pick a P-value threshold for calling an association noteworthy, based on (10), and assuming that the prior odds, PO, and the ratio of costs, R, do not depend on sample size or MAF. Substituting (9) into (11) and rearranging, one finds that the z^2 threshold is

$$z_T^2 = \frac{2(1+K^\star n)}{K^\star n} \log\left(\frac{\text{PO}}{R}\sqrt{1+K^\star n}\right). \tag{12}$$

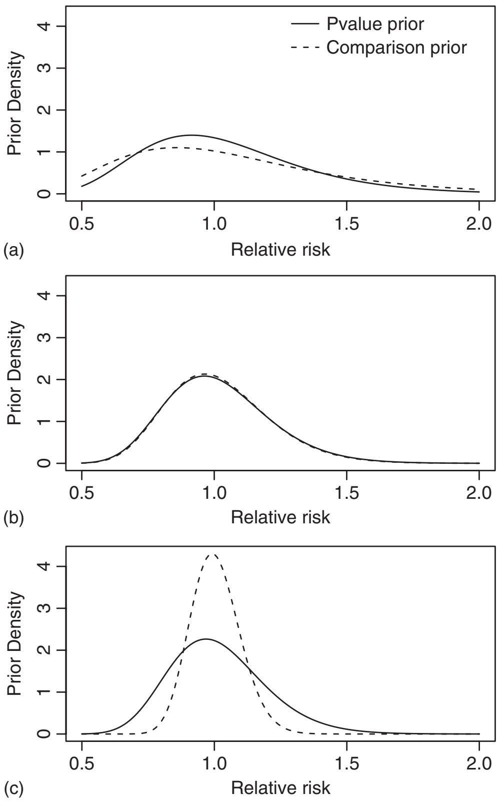

Fig. 2. Implicit P-value prior (solid lines) with prior variance on the log relative risk $W(M) = K \times V(M)$ (where M is the MAF, $n_0 = n_1 = 1{,}000$), as compared to the prior $W(M) = \delta_0 \exp(-\delta_1 \times \text{M})$ (dashed lines); δ_0 and δ_1 are chosen so that at a MAF of 0.05 the 95% point of the prior is 2.5, and at 0.50 it is 1.2. (a) MAF = 0.10; (b) MAF = 0.30; (c) MAF = 0.50.

The P-value threshold that should be used is therefore

$$p_\mathrm{T} = 2[1 - \Phi^{-1}(z_\mathrm{T})] \tag{13}$$

From (12) we see that the z^2 threshold increases (so that the P-value threshold decreases) as the prior odds of no association increase or as the ratio of costs of false discovery to false non-discovery increase, both as expected. It is difficult to make general statements about the dependence of the threshold on sample size since the formula is a complex function of n, as we saw in Figure 1.

To evaluate particular values of the threshold one must pick a value of $K^\star$ to calibrate the prior. If we take $K^\star = 1$ then this prior is equivalent to the unit-information prior [Kass and Wasserman, 1995] that has been suggested as a "reference prior" for Bayes factors; this prior is not appealing in the GWAS setting in which substantive prior opinion on the size of possible effects exists. In what follows, for illustration, we choose a dominant model, evaluate the asymptotic variance at the null, and choose $K^\star$ so that the 95% points of the prior on the relative risk are 2.5 and 1.7 at MAFs of 0.05 and 0.50, respectively. Figure 3 shows the P-value threshold (on a log 10 scale) as a function of n, and for particular values of π_1, the prior on the alternative, and the ratio of costs R (neither of which, recall, are assumed to depend on sample size or MAF). The curves are not horizontal, which shows that it is important to make thresholds depend on sample size if a P-value approach is taken. The non-monotonic behavior is due to the complex relationship between the ABF and power (Fig. 1). Requiring a smaller P-value for larger sample size (larger power) has been frequently advocated [Wacholder et al., 2004; Wellcome Trust Case Control Consortium, 2007]. From a decision theory perspective, taking a common threshold across all sample sizes is only consistent with one or more of the prior variance on the effect size, the prior odds on an association, or the ratio of costs, changing with n, and in a very stylized way. We stress that we are not advocating the use of the rule (13) in practice, as we would rather use a Bayes factor with the prior on the effect size reflecting our actual beliefs.

MULTIPLE TESTING

An important observation is that the threshold rule, (10), does not depend on the number of SNP associations to be considered, m, though interpretation, and, in particular, the expected number of false discoveries, will depend on this number. The small P-value thresholds shown in Figure 3 occur in large part due to the low a priori probability of a non-null hypothesis. Hence the threshold rule we have derived is in direct contrast to the Bonferroni threshold which is based on the number of tests considered, with the sample size being irrelevant. For example, Risch and Merikangas [1996] argue that when testing 10^6 associations a Bonferroni correction would suggest a level of 5×10^{-8}, while Dahlman et al. [2002] suggest P-values in the range 10^{-7}–10^{-8} if 250–500 K tests are carried out. In spite of the limitations of the Bonferroni correction, including the relevance of controlling the probability of a single type I error, and questions over the relevant number of tests over which to control [Colhoun et al., 2003; Balding, 2006], it remains the most common method of adjustment in GWASs [DeWan et al., 2006; Frayling et al., 2007].

If one believes that all of the nulls might be true then the Bonferrroni correction is more meaningful, and a similar adjustment can be obtained via a Bayesian approach [Westfall et al., 1995]. Specifically, suppose that when one specifies the prior over the set of m null associations one wishes to control the prior probability of all the nulls being true. The simplest way of achieving this is to take the prior on each null, independently, as $\Pi_0^{1/m}$ to give a prior probability of all nulls being true as Π_0. In contrast, if one specifies independent priors on each null to be π_0, the induced prior on all nulls being true is π_0^m. Under the latter, if $\pi_0 = 1 - 1/100,000$ (so that we believe that, on average, 1 in 100 K SNPs are non-null), the prior on all 500 K SNPs being null is 0.0067, i.e. very unlikely. If this prior truly reflects ones beliefs then controlling the family-wise error rate (as is achieved by the Bonferonni correction) is an unappealing criterion.

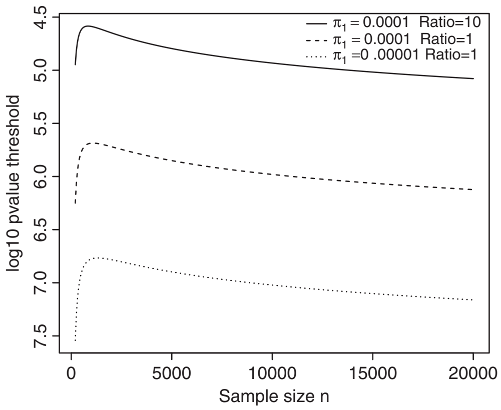

Fig. 3. P-value threshold at which to call an association noteworthy versus cases and control sample sizes n, and for different values of the prior on the alternative π_1, and ratios of costs of false non-discovery to false discovery. This is under a particular prior on effect size (choice of $K^\star$ in equation (12)) such that the 95% points of the prior on the relative risk are 2.5 and 1.7 at MAFs of 0.05 and 0.50, respectively.

TABLE I. P-value thresholds (to the $\log_{10}$) for $\Pi_0 = 0.9999$

	Number of tests, m				
	1	1,000	100,000	317,000	500,000
Bonferroni threshold	−6.17	−9.17	−11.17	−11.67	−11.87
Power prior threshold	−6.17	−9.27	−11.32	−11.83	−12.03

Note: For the Bonferroni threshold we take the P-value threshold corresponding to $m = 1$, and divide by m. For the power prior we take $\pi_0 = \Pi_0^{1/m}$.

We illustrate using the asymptotic Bayes factor (9). A Bonferonni threshold is obtained by taking the P-value threshold corresponding to $\pi_0 = \Pi_0$ for a single test, and then dividing this threshold by m. This may be compared with the threshold arising from (9) with the prior for each test taken as $\pi_0 = \Pi_0^{1/m}$, which we refer to as the power prior. Table I gives thresholds based on $\Pi_0 = 0.9999$, calculated for $n = 1,000$. For a single test, the threshold under this prior is 6.8×10^{-7}. We see that there is a close correspondence between the Bonferonni and power prior thresholds. The message here is that the Bonferroni correction has a Bayesian justification, but only for a very extreme prior that will often be inappropriate in a GWAS.

DISCUSSION

We have considered the use of Bayes factors in GWASs and have illustrated that the use of the *P*-value may be equated with a particular prior on the relative risk parameter. We formally state the main result of the paper.

Result: Let p_{ik} denote the probability of disease for individual i and marker k and consider the logistic regression model:

$$\frac{p_{ik}}{1-p_{ik}} = \exp(\boldsymbol{\alpha} \boldsymbol{z}_i + \theta_k x_{ik}), \qquad (14)$$

where $\boldsymbol{z}_i$ is a $1 \times p$ vector of confounders with associated parameters $\boldsymbol{\alpha}$, and x_{ik} is the score for marker k, and is a function of the genetic model and the number of minor alleles of this marker possessed by individual i, $i = 1, \ldots, n$, $k = 1, \ldots, m$. The null and alternative hypotheses of interest are $H_{0k} : \theta_k = 0$ and $H_{1k} : \theta_k \neq 0$. Let $\widehat{\theta}_k$ be the MLE for θ_k with associated standard error $\sqrt{V_k}$, and Wald statistic *P*-value, $p_k = 2\{1 - \Phi(|z_k|)\}$ where $z_k = \widehat{\theta}_k / \sqrt{V_k}$. Consider a Bayesian analysis based on $\widehat{\theta}_k, V_k$ with prior $\theta_k \sim_{\text{ind}} N(0, W_k)$, $k = 1, \ldots, m$. Under the prior $W_k = K \times V_k$, where K does not upon k, the Bayes factor that measures the evidence in favor of H_{0k} as compared to H_{1k} is

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

Under the above specifications, the rankings of p_k and ABF_k will be identical. The rankings are also identical if K depends on n but n is constant across markers.

The implicit *P*-value prior depends on the MAF in a qualitatively reasonable way, with stronger effects anticipated at lower MAFs. The relationship between effect size and MAF is not strong, however (as illustrated in Fig. 2), and lists of top-ranked SNPs from *P*-value and a Bayes factor approach with prior independence between effect size and MAF, will often be similar, with differences only for SNPs with very low MAFs.

For final inference the use of the *P*-value is problematic, however, since its interpretation depends on sample size. We have shown, using a decision theory approach that it is not optimal to take a single *P*-value threshold for all sample sizes. Rather, if the prior odds and ratio of costs of false non-discovery and false discovery do not change with sample size, the approach that is already informally taken, based on experience and examination of interval estimates. A threshold that is independent of n leads to a procedure that is inconsistent, in that the correct model is not chosen with probability 1 as the sample size increases.

With a formal Bayesian approach to testing, one may allow the ratio of costs and the prior odds to depend on sample size and MAF also. For example, as data collection proceeds the cost of false discovery will increase relative to the cost of false non-discovery (early on we would like a long list), and such behavior can be formalized with a Bayesian decision theory approach. We may also want the ratio of the costs of type II to type I errors to depend on the MAF. For example, from a public health standpoint it could be argued that the ratio should be higher for a more common variant (since the population attributable risk is higher for such a variant).

Andrews [1994], in a wide-ranging article, showed the relationship between Bayes factors and Wald, likelihood ratio, and score statistics, under more general priors; the normal prior discussed above is a special case which is practically convenient. See also Efron and Gous [2001] and Johnson [2005, 2008]; the latter derive properties of Bayes factors based on test statistics.

For calculation, the Bayes factor requires just a point estimate/standard error, or a confidence interval. R code is available at: http://faculty.washington.edu/jonno/cv.html

The asymptotic Bayes factor can be used in any situation in which an estimate and standard error are available. A number of authors have considered regression using imputed unmeasured SNPs [Marchini et al., 2007; Servin and Stephens, 2007]. When data on such SNPs are analyzed the uncertainty in genotype must be acknowledged; and there are now a number of packages that allow valid inference to be made [Sinnwell and Schaid, 2005]. The implemented methods use weighted logistic regression model, with the weights given by the posterior probabilities of the genotypes, and a generalized estimating equation (with the clusters being the individuals), to account for the repeated observations on each individual [French et al., 2006]. The use of estimates and standard errors from such approaches results in a Bayes factor that is adjusted for measurement error in the genotype.

Recently, two-phase sampling has been suggested as a method by which efficiency may be gained in a genetic epidemiology study [Chatterjee and Chen, 2007]. Specifically, at phase 1 inexpensive covariates may be measured on all individuals, with genetic and exposure information only gathered on an informative set of individuals at phase 2. The selection mechanism is carefully chosen to maximize information, but depends on the outcome and so must be accounted for in the estimation scheme. In other situations, survival data may be the endpoint of interest so that, for example, the Cox model may be the appropriate analysis tool. In both of these cases, a valid estimate and standard error is produced by the relevant software, and these can be used to evaluate the Bayes factor described here, so long as the sample size is large.

ACKNOWLEDGMENTS

This work was partially supported by grant 1 U01–HG004446–01 from the National Institutes of Health. The author would like to thank two anonymous referees for constructive comments that aided in clarification of the material.

REFERENCES

Andrews DWK. 1994. The large-scale correspondence between. Econometrica 62:1207–1232.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781–791.

Chatterjee N, Chen YH. 2007. Maximum likelihood inference on a mixed conditionally and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. J R Stat Soc Ser B 69:123–142.

Colhoun HM, McKeigue PM, Davey-Smith G. 2003. Problems of reporting genetic associations with complex outcomes. Lancet 361:865–872.

Cox DR, Hinkley DV. 1974. Theoretical Statistics. London: Chapman & Hall.

Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SCL, Allahabadia A, Franklyn JA, Tuomilehto J, Tuomilehto-Wolf E, Cucca F, Guja C, Ionescur-Tirgoviste C, Stevens H, Carr P, Nutland S, McKinney P, Shield JP, Wang W, Cordell HJ, Walker N, Todd JA, Concannon P. 2002. Parameters for reliable results in genetic association studies in common disease. Nat Genet 30:149–150.

DeWan A, Liu M, Hartman S, Zhang, ShaoMon S, Liu DT, Zhao C, Tam POS, Chan WM, Lam DSC, Snyder M, Barnstable C, Pang CP, Hoh J. 2006. HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 314:989–992.

Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1–9.

Efron B, Gous A. 2001. Scales of evidence for model selection: Fisher versus Bayes with discussion. In: Lahiri P, editor. Model Selection, Institute of Mathematical Statistics Lecture Notes, Monograph Series. p 208–256.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894.

French B, Lumley T, Monks SA, Rice KM, Hindorff LA, Reiner AP, Psaty BM. 2006. Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol 30:485–494.

Good IJ. 1992. The Bayes/non-Bayes compromise: A brief review. J Am Stat Assoc 87:597–606.

Goodman SN. 1999. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 130:995–1004.

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

Johnson VE. 2005. Bayes factors based on test statistics. J R Stat Soc Ser B 67:689–701.

Johnson VE. 2008. Properties of Bayes factors based on test statistics. Scand J Stat 35:354–368.

Kass R, Raftery A. 1995. Bayes factors. J Am Stat Assoc 90:773–795.

Kass RE, Vaidyanathan SK. 1992. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. J R Stat Soc Ser B 54:129–144.

Kass RE, Wasserman L. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90:928–934.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

Sasieni PD. 1997. From genotypes to genes: Doubling the sample size. Biometrics 53:1253–1261.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: Candidate regions and quantative traits. Public Libr Sci Genet 3:1296–1308.

Sinnwell JP, Schaid DJ. 2005. haplo.stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. Am J Hum Genet 70:425–434.

Sladek R, Rocheleau G, Ring J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885.

Slager SL, Schaid DJ. 2001. Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. Hum Hered 52:149–153.

Sterne JAC, Smith GD. 2001. Sifting the evidence—what's wrong with significance tests? Br Med J 322:226–231.

Wacholder S, Chanock S, Garcia-Closas M, El-ghormli L, Rothman N. 2004. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. J Nat Cancer Inst 96:434–442.

Wakefield J. 2007. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 81:208–227.

Wakefield JC. 2008. Reporting and interpretation in genome-wide association studies. Int J Epidemiol 37:641–653.

Wang WYS, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: Theoretical and practical concerns. Nat Rev Genet 6:109–118.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study between 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Westfall PH, Johnson WO, Utts JM. 1995. A Bayesian perspective on the Bonferroni adjustment. Biometrika 84:419–427.

APPENDIX

DERIVATION OF THE ASYMPTOTIC BAYES FACTOR

In a case-control study, we have binary disease indicators Y_i on $i = 1, \ldots, n_1$ cases, and $i = n_1 + 1, \ldots, n_1 + n_0$ controls. These data follow a binomial distribution with index 1 and probability p_i, the risk for individual i. We assume the logistic regression model:

$$\frac{p_i}{1 - p_i} = \exp(\boldsymbol{\alpha} \boldsymbol{z}_i + \theta x_i), \tag{15}$$

where $\boldsymbol{z}_i$ is a $1 \times p$ vector of confounders (which includes the intercept) with associated parameters $\boldsymbol{\alpha}$, and x_i depends on the genetic model and is a function of the number of mutant alleles possessed by individual i. The Bayes factor is given by $\Pr(\boldsymbol{y}|H_0)/\Pr(\boldsymbol{y}|H_1)$ where the numerator and denominator are integrals given, respectively, by

$$\Pr(\boldsymbol{y}|H_0) = \int \Pr(\boldsymbol{y}|\boldsymbol{\alpha}, \theta = 0)\pi(\boldsymbol{\alpha})\, d\boldsymbol{\alpha},$$

$$\Pr(\boldsymbol{y}|H_1) = \int \Pr(\boldsymbol{y}|\boldsymbol{\alpha}, \theta)\pi(\boldsymbol{\alpha}, \theta)\, d\boldsymbol{\alpha}\, d\theta,$$

where $\pi(\boldsymbol{\alpha}, \theta)$ is the prior over all parameters and $\pi(\boldsymbol{\alpha})$ is the prior over the nuisance parameters only. These integrals are analytically intractable for the binomial likelihood and the specification of multivariate priors is cumbersome. We follow a different approach and assume we are in an asymptotic situation in which n_0 and n_1 are "large", a situation that is almost always satisfied in GWASs. In this case, inference for the logistic model may be carried out on the basis of the asymptotic distribution:

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{\theta} \end{bmatrix} \sim N_{p+1}\left(\begin{bmatrix} \boldsymbol{\alpha} \\ \theta \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{00} & \boldsymbol{I}_{01} \\ \boldsymbol{I}_{01}^{\mathrm{T}} & \boldsymbol{I}_{11} \end{bmatrix}^{-1} \right), \tag{16}$$

where $\boldsymbol{I}_{00}$ is the $p \times p$ matrix expected information concerning $\boldsymbol{\alpha}$, I_{11} is the expected information concerning θ, and $\boldsymbol{I}_{01}$ is the $p \times 1$ vector of cross terms. In Wakefield [2007] it was assumed that $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\theta}$ were independent. Here we relax this assumption, reparameterize the model and consider $(\boldsymbol{\alpha}, \theta) \to (\boldsymbol{\beta}, \theta)$ where

$$\boldsymbol{\beta} = \boldsymbol{\alpha} + \frac{\boldsymbol{I}_{01}}{I_{00}}\theta,$$

which yields

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\theta}\end{bmatrix} \sim N_{p+1}\left(\begin{bmatrix}\boldsymbol{\beta}\\ \theta\end{bmatrix}, \begin{bmatrix}\boldsymbol{I}^{\star}_{00} & \mathbf{0}\\ \mathbf{0}^{\mathrm{T}} & I_{11}\end{bmatrix}^{-1}\right), \tag{17}$$

where $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\alpha}} + (\boldsymbol{I}_{01}/I_{00}) \times \widehat{\theta}$ and $\mathbf{0}$ is a $p \times 1$ vector of zeros. Hence, asymptotically, $p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|\boldsymbol{\beta}, \theta) = p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta}) \times p(\widehat{\theta}|\theta)$. We assume independent priors on $\boldsymbol{\beta}$ and θ, $\pi(\boldsymbol{\beta}, \theta) = \pi(\boldsymbol{\beta})\pi(\theta)$ and calculate the Bayes factor working with "data" $\{\widehat{\boldsymbol{\beta}}, \widehat{\theta}\}$. Under H_0:

$$\begin{aligned} p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|H_0) &= \int p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|\boldsymbol{\beta}, \theta = 0)\pi(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}\\ &= \int p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta} \times p(\widehat{\theta}|\theta = 0)\end{aligned}$$

and under H_1:

$$\begin{aligned} p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|H_1) &= \int\int p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}, \theta)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\theta\\ &= \int\int p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta})p(\widehat{\theta}|\theta)\pi(\boldsymbol{\beta})\pi(\theta)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\theta\\ &= \int p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta} \int p(\widehat{\theta}|\theta)\pi(\theta)\,\mathrm{d}\theta.\end{aligned}$$

Hence the Bayes factor based on $(\widehat{\boldsymbol{\beta}}, \widehat{\theta})$ is given by

$$\mathrm{ABF} = \frac{p(\widehat{\theta}|\theta = 0)}{\int p(\widehat{\theta}|\theta)\pi(\theta)\,\mathrm{d}\theta}. \tag{18}$$

The reparameterization trick works because of the assumption of independent priors on $\boldsymbol{\beta}$ and θ, which does not imply independent priors on $\boldsymbol{\alpha}$ and θ. We emphasize that we never need to specify the prior on $\boldsymbol{\beta}$, because terms involving $\boldsymbol{\beta}$ cancel in the Bayes factor calculation. Under the prior $\theta \sim N(0, W)$ the Bayes factor (18) becomes

$$\mathrm{ABF} = \sqrt{\frac{V + W}{V}} \exp\left(-\frac{z^2}{2}\frac{W}{(V + W)}\right),$$

where $V = I_{11}^{-1}$. The approach employed is similar to the "null orthogonality" reparameterization of Kass and Vaidyanathan [1992]. The reparameterization is also that which is used when the linear model:

$$Y_i = \alpha + x_i\theta + \varepsilon_i$$

is written as

$$Y_i = \beta + (x_i - \overline{x})\theta + \epsilon_i$$

which, of course, yields uncorrelated least squares estimators $\widehat{\beta}, \widehat{\theta}$.