HHS Public Access

Author manuscript

Pharmacogenomics J. Author manuscript; available in PMC 2015 February 01.

Published in final edited form as:

Pharmacogenomics J. 2014 August ; 14(4): 309–315. doi:10.1038/tpj.2013.44.

Detecting Signals in Pharmacogenomic Genome-Wide Association Studies

Jon Wakefield, PhD[1,2], **Veronika Skrivankova, MS**[2], **Fang-Chi Hsu, PhD**[3], **Michele Sale, PhD**[4], and **Patrick Heagerty, PhD**[2]

[1] Department of Statistics, University of Washington, Box 357232, Seattle, WA 98195, jonno@uw.edu, Telephone: 206-616-9388, Fax: 206-543-3279.

[2] Department of Biostatistics, University of Washington, Seattle, WA

[3] Department of Biostatistical Sciences, Wake Forest University, Winston-Salem, NC

[4] Robert M. Berne Cardiovascular Research Center, University of Virginia, VA

Abstract

In one common pharmacogenomic scenario, outcome measures are compared for treated and untreated subjects across genotype defined subgroups. The key question is whether treatment benefit (or harm) is particularly strong in certain subgroups, and therefore statistical analysis focuses on the interaction between treatment and genotype. However, genome-wide analysis in such scenarios requires careful statistical thought since, in addition to the usual problems of multiple testing, the marker-defined sample sizes, and therefore power, vary across the individual genotypes being evaluated. The variability in power means the usual practice of using a common *p*-value threshold across tests has difficulties. The reason is that the use of a fixed threshold, with variable power, implies that the costs of type I and type II errors are varying across tests in a manner which is implicit rather than dictated by the analyst. In this paper we discuss this problem and describe an easily implementable solution based on Bayes factors. We pay particular attention to the specification of priors, which is not a straightforward task. The methods are illustrated using data from a randomized controlled clinical trial in which homocysteine levels are compared in individuals receiving low and high doses of folate supplements and across marker subgroups. The method we describe is implemented in the R computing environment with code available from http://faculty.washington.edu/jonno/cv.html.

Keywords

Bayes factors; Bonferroni correction; Significance threshold

Conflict of Interest

None of the authors has a conflict of interest.

Supplementary Material

Supplementary information is available at The Pharmacogenomics Journal's website.

Introduction

The key statistical question in pharmacogenomic studies, in which genetic marker-defined subgroups are examined to see if they respond well or poorly to treatment, is how to flag a signal as “significant”. We describe two recent examples of case-control pharmacogenomic studies, in order to outline the usual current approach to the decision problem. In a recent study [1], epilepsy patients were treated with lamotrigine or related drugs. The cases had either Stevens-Johnson syndrome (SJS) or toxic epidermal necrolysis, both of which are potentially life threatening adverse drug reactions characterized by skin blistering. After quality control, 837,070 SNPs were examined using logistic regression. The authors quote 5×10^{-8} as the genome-wide significance level that was sought (corresponding to a Bonferroni correction that controls the family-wise error rate) but no *p*-values below 10^{-6} were found. Various case and control populations were examined but the number of cases in the different analyses performed was around 100 and so the power was low. In a second example, a candidate gene approach was reported in which cholesterol levels were examined across SNP defined subgroups in 554 statin users [2]. Again, a Bonferroni correction was applied to determine the *p*-value significance threshold. Two SNPs achieved the required level, with one being confirmed in a replication study.

We focus on the situation in which the phenotype is quantitative, though the methods are applicable generally. In the simplest situation, two treatments are randomly assigned to *n* subjects, with a quantitative summary measure (for example, a change in a biomarker) being subsequently measured. Information on *J* markers (e.g. SNPs) for each of the *n* study participants may then be used to define subgroups of interest. For example, in a recessive model, for each diallelic SNP, the two comparison subgroups of interest would be those that possess zero copies of the minor allele (subgroup 1) and those who possess one or two copies of the minor allele (subgroup 2). A standard analysis fits a linear model with treatment and marker main effects, and a treatment by marker interaction. The null hypothesis of interest is that the interaction parameter is zero, i.e., that individuals in each of the two marker-defined subgroups respond equally to treatment. The key point is that, across all SNPs, *J* interaction parameters are examined and the power to detect non-zero interactions will typically vary hugely over this collection, because the subgroup sizes are a direct function of the minor allele frequency (MAF). Conventional approaches to testing control a measure such as the family-wise error rate (FWER) by taking a *fixed p*-value threshold across all *J* tests. A key problem with this strategy is that the differential power (and therefore type II error) across tests, combined with a fixed type I error implies that the *costs* of the two types of error are varying across tests; one would prefer an approach in which both the type I and type II errors rate go to zero as the information increases. Intuitively, when the power is close to 1, one can afford to reduce the type I error rate, even if there is a corresponding decrease in power. The use of *p*-values in pharmacogenomics has been criticized previously [3] with a Bayesian approach being advocated as a possible solution. In this paper we describe a simple procedure with the desired characteristics, based on Bayes factors. Related but distinct approaches have previously been suggested [4–8].

The issue of variable power across multiple tests is common to all testing situations, but is acute in pharmacogenomic situations since the subset sizes are highly variable across tests.

The difficulties associated with fixed threshold rules has been previously pointed out in the context of case-control genome-wide association studies (GWAS) [9].

Materials and Methods

The VISP Trial

In this paper we analyze data from the Vitamin Intervention for Stroke Prevention (VISP) trial which is an NIH-funded, multi-center, double-blind, randomized, controlled clinical trial [10]. The aim of this trial was to determine whether a daily intake of high dose folic acid and vitamins B6 and B12 was associated with cardiovascular endpoints. One individual was removed since their data were considered outlying; this individual's data and the implications of retaining their data are discussed in the supplementary material. We examine data on $n = 1,670$ individuals of European ancestry, with 837 randomized to the high dose and 833 to the low dose. After quality control procedures, 803, 122 SNPs were available for analysis. The outcome is the intermediate variable homocysteine level with high levels in blood being associated with cardiovascular disease. In the VISP trial, levels were measured longitudinally but for simplicity we take as outcome the difference between the baseline and the first post-baseline measurements. The average change was −0.37 μmol/L in the low dose group versus −2.36 μmol/L in the high dose group, to give a difference of −1.99 μmol/L ($p < 2 \times 10^{-16}$) between the treatment groups. In this paper we take as our objective the examination of the treatment effect by marker, in order to determine whether genetic markers can identify subgroups with exceptionally strong or weak treatment responses.

Researchers can apply for access to the VISP genetic at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000343.v3.p1

Frequentist Boundaries

In this section we describe frequentist approaches to multiple hypothesis testing and begin by introducing some notation in the context of a quantitative trait and a pair of treatment groups. Let Y_i and $T_i = 0/1$ represent the response and treatment indicator for individual i, $i = 1, ..., n$, and let $M_i = 0/1$ be a marker indicator for individual i and for a generic SNP. In the VISP primary analysis we assume a recessive genetic model with $M_i = 0/1$ corresponding to 0/1 or 2 copies of the minor allele. The choice of a recessive model is made for illustration only, and the fundamental approach and modeling issues we discuss exist for any choice (in the supplementary materials we report on the fitting of an additive model for the VISP data). We assume that J SNPs will be examined. To characterize treatment effects for a generic marker we use the model

$$Y_i = \alpha + x_i\phi + T_i\beta + M_i\gamma + T_i \times M_i\Delta + \epsilon_i \quad (1)$$

with independent error terms ε_i having variance σ^2. In model (1), x_i corresponds to individual-level covariates that we wish to adjust for (such as age and gender), β is the main effect of treatment and Δ is the interaction parameter of interest, with the null hypothesis $H_0 : \Delta = 0$ being compared to $H_1 : \Delta \neq 0$. Specifically, the interaction parameter Δ contrasts the treatment effect when $M_i = 1$ with the treatment effect when $M_i = 0$, and a non-zero interaction implies an association between the marker and treatment response. Hypothesis

testing may be based on the Z statistic $Z=\hat{\Delta}/\sqrt{V}$, with V the estimated asymptotic variance of the MLE. The observed p-value based on a Z statistic is $p = \Pr(|Z| > z_{obs}|H_0)$. A small p-value can arise because H_0 is true but we were "unlucky", or because H_0 is not true. As we shall argue, deciding between these explanations depends on the power of the test, which in turn should depend on the standard error $\sqrt{V}$.

In the context of this paper, since J tests are performed, the multiple testing aspect must be considered when determining a p-value threshold from a frequentist perspective. A criterion that has received attention in the pharmacogenomics literature [1,2] is control of the family-wise error rate (FWER), which is the probability of one or more false discoveries (type I errors). The Bonferroni correction controls the FWER at a level of α by taking the individual p-value threshold as α/J. Universal recommendations have been suggested for setting α in the GWAS literature [11,12]. However, these recommendations are independent of power considerations such as sample size and MAF. For example, an early study [13] used Bonferroni with 96 cases and 50 controls and 103,611 SNPs (to give a threshold of $0.05/103,611 = 4.8 \times 10^{-7}$) while a much larger study [14] used Bonferroni with 17,513 cases and 4,533 controls and 311,524 SNPs (to give a threshold of $0.05/311,524 = 1.6 \times 10^{-7}$). More recently [15–17], the idea of a universal recommendation has been critically discussed, with a Bayesian approach being suggested as an alternative.

There are two primary problems with the use of the Bonferroni approach. First, controlling the FWER is not appropriate when we do not expect all of the nulls to be true. Second, as described above, a p-value threshold that is common to all tests, regardless of the power of each of the tests, is not sensible. Simply said, the type II error is going to zero with increasing sample size, so why not the type I error? In frequentist inference, *consistency* is sought, which means that procedures are favored if the correct answer (the true hypothesis in a testing setting) is recovered with increasing probability as the sample size increases. But the use of a constant threshold leads to an *inconsistent* procedure since, by construction, the type I error rate is not going to zero with increasing n. This is of course true in the usual frequentist hypothesis testing situations.

Other frequentist criteria can be used as an alternative to FWER. An intuitive measure is the expected number of false discoveries (EFD). If J_0 is the true number of null signals amongst the J tests we have: $\text{EFD} = J_0 \times \alpha \leq J \times \alpha$. The latter inequality will be practically useful in situations in which we expect J_0 to be close to J. Methods that control the false discovery rate (FDR) are also popular [18–20]. While EFD and FDR are more reasonable criteria than the FWER, one still needs to specify a threshold size, and this should decrease as the sample size increases. Current practice is again to ignore power when setting a p-value threshold. But, as with the FWER, the threshold should also go to zero with increasing sample size, so that both type I and II errors go to zero. However, there are no prescriptions available. We now describe a method for determining a significance boundary, based on a Bayesian formulation that provides a solution with the desired characteristics.

Bayesian Decision Boundaries

A Bayesian assessment of the evidential content of the data with respect to the two hypotheses is provided by 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: BF = p(data $|H_0$)/p(data $|H_1$). Large values of the Bayes factor favor the null, and values close to zero favor the alternative. The Bayes factor has been recently advocated as a measure of evidence in a number of genetic contexts [15, 16, 21–23]. To convert the evidence of the data contained within the Bayes factor into posterior probabilities on each of the two hypotheses one needs to specify a prior on these hypotheses. Let π_0 represent the prior on the null being true, so that PO = $\pi_0/(1-\pi_0)$ is the prior odds on the null. Via Bayes theorem we can then evaluate the posterior probability that the null is true as

$$Pr\,(H_0|\text{data}) = \frac{BF \times PO}{1+BF \times PO}, \quad (2)$$

with Pr(H_1|data) = 1 – Pr(H_0|data). As an example, if the Bayes factor is 1/4 and the prior odds are 1/2, then the posterior probability on the null is $\frac{1/4\times 1/2}{1+1/4\times 1/2}=\frac{1}{9}$. In order to pick a posterior probability threshold at which to declare significance we may go a step further and appeal to decision theory. In a Bayesian decision theory approach utilities (or costs) are placed on all combinations of actions (choice of whether or not to reject) and "truths" (null hypothesis correct or not). The principle of minimization of expected costs is then followed [24]. With respect to Table 1, suppose R = C_{II}/C_I is the ratio of the costs of type II to type I errors. Historically, a type I error has been viewed as more harmful than a type II error, but this slant is not always desirable. A Bayesian decision theory approach chooses H_1 if the posterior odds on H_0 falls below R which, from (2), occurs if

$$BF \times PO < \text{R}. \quad (3)$$

In general, the use of Bayes factors faces two large practical hurdles. First, one must specify prior distributions over all of the unknown parameters in the model, which can be a challenging task. For example, with respect to (1), priors would be required for α, φ, β, γ, Δ and σ. Second, the computation of the numerator and denominator of the Bayes factor requires evaluation of integrals whose dimensions are equal to the number of parameters contained in the model under H_0 and H_1, respectively. Such integration is not trivial and must be carried out J times in the multiple testing context. To overcome these hurdles, one suggestion is to replace the likelihood arising from the original data by the sampling distribution of the estimator of the parameter of interest [15]. The idea is to take the "data" as $\hat{\Delta}$ in which case we have a likelihood $p\left(\hat{\Delta}|\Delta\right)$. In large samples (such as those that are typically in GWAS) this distribution will be normal and, effectively, the available information in the data concerning the parameter of interest has been summarized in the sampling distribution of the estimator. Similar approaches have a long history, in particular in the context of clinical trials [25].

The model for a generic marker with interaction parameter Δ is

$$\text{Likelihood:} \quad \hat{\Delta}|\Delta \sim N(\Delta, V) \quad (4)$$

$$\text{Prior:} \quad \Delta \sim N(0, W). \quad (5)$$

where $\sqrt{W}$ is the prior standard deviation, so that a 95% prior interval on the size of the interaction is $\pm 1.96\sqrt{W}$. The Bayes factor is relatively insensitive to the choice of *W* [26]. In terms of the likelihood, only a confidence interval is required for specification, since this interval may be used to find $\hat{\Delta}$ and *V* for use in (4).

The model (4) and (5) leads to a very simple form of Bayes factor for the marker, as shown elsewhere [15]:

$$BF = \sqrt{\frac{V+W}{V}} exp\left(-\frac{Z^2}{2}\frac{W}{V+W}\right)$$

where *Z* is the *Z* statistic. A crucial observation is that the evidence is based on the *Z* score *and* on the standard error $\sqrt{V}$, where the latter in turn depends on the subgroup sample sizes.

One may view Bayes factors as a mechanism by which *Z*-score boundaries can be calculated as a function of the standard error $\sqrt{V}$. From rearrangement of (3) the Bayesian Z^2 score threshold for rejection is:

$$Z^2 > z_{\mathrm{B}}^2 = \left(\frac{V+W}{W}\right)\left\{log\left(\frac{V+W}{V}\right) + 2 \quad log\left(\frac{PO}{\mathrm{R}}\right)\right\} \quad (6)$$

to give a threshold which is an explicit function of *V*, R and PO. Notice that this Bayesian derivation makes no mention of type I and type II errors though one may calculate these errors from (6) if one wishes to evaluate the frequentist properties that correspond to particular choices of PO, R and *W*. An example is given in the Results section below.

We make the following observations on the Bayesian boundary (6). If the prior odds on the null (PO) increases, the threshold increases so that we require *more* evidence to overcome the initial scepticism. If the cost of type II to type I errors, R, increases, then the threshold decreases (to give a more liberal rule) and we require less evidence from the data. Beyond a certain point, as *V* decreases, the type I error decreases to zero. Specifically, let *n* denote an appropriate measure of sample size and write $V = k/n$, where *k* is not a function of *n*. Then, as $n \to \infty$, from (6)

$$z_{\mathrm{B}}^2 \to \underbrace{log\left(1 + \frac{nW}{k}\right)}_{\to\infty} + 2\, log\left(\frac{PO}{\mathrm{R}}\right),$$

so that the boundary increases, and so the type I error tends to zero. It can also be shown [26] that the Bayes factor tends to zero under the alternative as $n \to \infty$ [27]. We discuss the behavior of the Bayesian boundary relative to a boundary that is *constant* with respect to *n*. Such a constant boundary may be obtained through the usual implementations of FWER, EFD or FDR procedures. The Bayesian approach implicitly trades type I and type II errors through the specification of R. This is in contrast to the usual frequentist approach in which a type I error is fixed and the power is then determined. Figure 1 illustrates this behavior of the Bayesian threshold, given by the square root of (6), for a range of standard errors that are consistent with the VISP trial. The critical thresholds were taken from (6) with $\pi_1 = 0.0001$, so that with $J = 803,122$ tests we would expect around $J \times \pi_1 = 80$ signals. These signals will not reflect 80 different causal variants since typically multiple SNPs will tag each causal variant. For reference, the Bonferroni threshold for a FWER of 20% (corresponding to a *p*-value of $0.20/803122 = 2.5 \times 10^{-7}$) is indicated as a horizontal line. For *small n* (large standard error) the Bayesian approach increasingly requires greater evidence because of the low power. For *large n* (small standard error) the Bayesian approach requires more evidence because the power is high and so some of the small type II error is traded with type I error to give a more conservative procedure.

Results

A Priori Operating Characteristics

We now return to the VISP data and begin by examining the operating characteristics of potential Bayesian decision boundaries, based on different priors. This procedure may be carried out *before* the data are analyzed, since it allows one to choose a threshold rule aided by consideration of type I and type II error probabilities. For ranges of the standard error consistent with the VISP data (Figure S1 of the supporting material), we calculated the type I errors with the *Z*-score boundaries given by (6). We assume R = 1 (equal costs of type I and type II errors) and $\pi_1 = 0.001, 0.0001, 0.0001$. For $J = 803,122$ SNPs this corresponds to expecting 803, 80 and 8 non-null interactions, respectively. Figure 2 plots various useful operating characteristics. The negative log base 10 type I error probabilities are plotted as a function of the standard error in the upper left panel of Figure 2. We see the expected U-shaped behavior of the probabilities corresponding to the Bayesian decision boundaries (as discussed in the last section and as seen in Figure 1). We next evaluated the power to detect a drop of 5 units and the resultant curves are in the upper right panel of Figure 2. As in the upper left panel, the $\pi_1 = 0.0001$ (sceptical) prior is that which most closely mimics the FWER curve, at least for standard errors between 0.5 and 2.

We now transform the type I error and power into more intuitive quantities, the expected number of false discoveries (EFD) and the expected number of true discoveries (ETD). To determine the EFD and ETD we require specification of the number of null and non-null signals, which we label as J_0 and J_1, respectively (so that $J = J_0 + J_1$). We take the true number of signals as $J_1 = 50$ so that there are $J_0 = 803,072$ null signals. Then EFD = $J_0 \times \alpha$ and ETD = $J_1 \times (1 - \beta)$ where α and β are the type I and type II errors. We emphasize that in a GWAS in which the fraction of non-null associations is close to zero, the ETD is highly sensitive to the choice of J_1 (in contrast to EFD, which is insensitive). The resultant plots of

EFD and ETD are in the bottom row of Figure 2. Care is required in the interpretation of the ETD plot since at each standard error it is assumed that *all* of the signals have this standard error. We also stress that when we specify a prior π_1, the ETD is the true proportion of non-null signals, not the proportion of signals that we have the power to detect. This should be borne in mind when we actually analyze the data and consider the proportion of signals we are missing; we return to this point in the Discussion. The most liberal prior of $\pi_1 = 0.001$ produces a large number of type I errors (around 20 for standard errors in the mid-range) and might be judged to give unacceptably poor performance. The most sceptical prior is more conservative than Bonferroni (with a FWER of 20%) and the prior with $\pi_1 = 0.0001$ is a compromise for this choice of J_1. For example, for a standard error of 1, around 2 false discoveries would be expected (as in the lower left panel) but with around 10 more true signals being detected (as seen in the lower right panel), which seems a reasonable trade-off. Note, however, that if we think the number of true signals is smaller than $J_1 = 50$ then the number of true signals will fall proportionally. For example, at a standard error of 1, if $J_1 = 5$ then we would only expect to detect a single additional signal, when compared to the use of Bonferroni. Armed with this information we move to an analysis of the VISP data.

VISP Analysis

We fitted model (1) to the $J = 803,122$ SNPs with gender and age (by quintile) being included and corresponding to the x_i term. The genetic subgroups are defined as having at least one copy of the minor allele as compared to two copies of the major allele. The number in the former subgroup ranges between 21 and 1,564 across SNPs. The standard error of $\hat{\Delta}$ is $\sigma\sqrt{1/n_{00}+1/n_{01}+1/n_{10}+1/n_{011}}$ where n_{tm} is the number of individuals in treatment group t, $t = 0/1$ and marker subgroup m, $m = 0/1$ for a generic SNP. Hence, the standard error is driven by the smallest subgroup. To emphasize, the same 833 low dose and 837 high dose responses are used in each of the J comparisons, but they are differentially distributed in the four treatment × marker cells across the J comparisons.

Figure 3 plots the Z-scores versus the standard error, along with boundary corresponding to a FWER of 20%, and Bayes boundaries with $\pi_1 = 0.001, 0.0001, 0.00001$ and ratio of costs R = 1 (to give a posterior probability threshold of 0.5). We chose W to give a 95% prior interval for the interactions Δ of ±10. The curvature in the three Bayes boundaries acknowledges the variable power. For both the most conservative prior and the Bonferroni approach (with a FWER of 20%, which gives a p-value threshold of 2.5×10^{-7}) two SNPs are flagged. With a FWER of 5% the Bonferroni threshold is 6.2×10^{-8} and results in a single SNP being deemed significant. With the more optimistic prior of $\pi_1 = 0.0001$, a further two signals are flagged (and these are not significant using Bonferroni). There are few strong signals for these data, however. Table S1 in the supplementary materials contains more details on each of the top SNPs.

Figure 4 plots the posterior probabilities of the alternative hypothesis (with $\pi_1 = 0.0001$) versus chromosomal position (this is similar to a Manhattan plot in which $-\log_{10}$ p-values are plotted against position). The 3 SNPs that fall outside of the boundary in Figure 3 are highlighted. The strongest signal is for SNP rs3736238 on chromosome 17. For this SNP there are 42 individuals in the $M = 1$ subgroup, of which 24 and 18 are in the low and high

dose groups, respectively. The interaction effect was $\hat{\Delta} = -6.7$ (so that we have an enhanced effect), with s.e. $\left(\hat{\Delta}\right) = 1.4$ to give a *Z* score of –4.8 and a *p*-value of 1.5×10^{-8}. The reciprocal of the Bayes factor (evidence in favor of the alternative) is 1.1×10^{6}. The posterior probability on the alternative is 0.99 so that under this prior there is strong evidence to conclude the interaction is real. The probability of this signal being a false discovery is 0.01 *under* our assumed prior.

To illustrate the sensitivity of the conclusions to the prior on the effect size, reducing the 95% prior interval on Δ to ±3 gives $1/BF = 1.6 \times 10^{4}$ and a posterior probability on the alternative of 0.38. Figure S2 of the supporting material gives the behavior of the *Z*-score threshold, as shown previously as Figure 1, under the revised prior on Δ. Hence, for this SNP, the significance is greatly reduced because the observed size of effect was –6.7, which is very unlikely under the revised prior. If we change the prior on the alternative to $\pi_1 = 0.00001$, then the posterior probability on the alternative is reduced to 0.91 (so that the probability that this is a false discovery is 0.09). Figure S4 of the supporting material contains the $\pi_1 = 0.00001$ version of Figure 4.

The most significant SNP, rs3736238, is in gene flotillin 2 (FLOT2) on chromosome 17 and is located in an exonic splicing enhancer and thus may regulate splicing processes and subsequent mRNA stability. The amino acid substitution at amino acid position 279 of FLOT2 arises from the missense SNP rs3736238 (F-SNP database http://compbio.cs.queensu.ca/F-SNP). Hence, this SNP could affect the protein secondary structure or function. FLOT2 encodes a caveolae-associate protein, which may function in neuronal signaling (www.genecards.org). A recent study [28] showed that the DHHC5 protein palmitoylates FLOT2 in response to neuronal differentiation signals. It may also be related to the progression of multiple types of cancers and metastasis formation [29]. Currently the function of FLOT2 with respect to homocysteine levels or stroke related diseases is unknown but is worthy of further research. Given that there is no obvious compelling functional argument for the association with this SNP, and the not completely conclusive evidence arising from the posterior probability (which is highly sensitive to π_1) we would recommend that further work, preferably at the molecular level, be carried out for confirmation. The MAF is 1.3% for this SNP and so the enhanced effect will only be seen in a small group of individuals.

The second signal is rs16893296 on chromosome 6 with a posterior probability on the alternative of 0.96. point estimate and standard error are –4.6 and 0.85 with a *p*-value of 7.1×10^{-8}. The nearby genes are LOC442160 and LOC442161. This variant is in a weak DNaseI site, but has a low conservation score, which is fairly weak evidence for being “functional”. The third significant SNP rs1739317 (posterior probability on the alternative of 0.81 and a *p*-value of 4.0×10^{-7}) is located on chromosome 6 with nearby genes C6orf32 and LOC134997. The evidence that this SNP is functional is weak since this variant is also in a weak DNaseI site with a low conservation score. The mechanisms for the associations between this SNP and homocysteine levels are unknown. Hence, to validate the results, again replication studies are needed.

On Figure 4, lines of significance corresponding to ratios of cost equal to 1 and 3 are drawn at posterior probabilities of 0.5 and 0.25, respectively. An additional signal is called significant if we apply the more liberal 0.25 boundary. Further details for all four flagged SNPs are contained in the supporting materials.

Figure 5 shows that the *p*-values and Bayes factors differ in their rankings due to the differing sample sizes/standard errors. The points are color-coded by the size of the standard error and we see that the points with larger standard errors are consistently ranked as giving greater evidence for the alternative under the Bayesian approach. This behavior occurs here because of the association between the Z^2 boundary and the standard error for these priors, as shown in Figure 1. Specifically, the majority of the signals occur in that portion of the latter curve in which the Bayes boundary lies below the FWER boundary. Figure S3 of the supporting material shows an example in which distinctly different behavior occurs.

Discussion

In this section we expand on the practical difficulties of implementing the method described here, and in particular on the choice of prior distributions. We also report some additional analyses that address the sensitivity of our conclusions to various model choices.

The posterior probability of the alternative is highly dependent on the choice of prior on the null π_0, and a sensitivity analysis is always warranted. Ideally, rather than fix π_0 as we have done, one would estimate of π_0 from the totality of data (i.e. over all J SNPs), but this is difficult because in a GWAS the proportion of *detectable* null signals is typically very close to 1; there may be many thousands of small but non-zero effects, but the power to detect these signals is low, with usual sample sizes. In other contexts, such as the analysis of gene expression data [20], the data can be used to estimate π_0 more reliably. If the same prior on the null is used for all the tests, the rankings based on the Bayes factor will remain the same as the ranking based on posterior probabilities. However, calibrating the Bayes factors to the probability scale requires prior probabilities. Within a sensitivity exercise one may include an analysis in which any available information on particular SNPs may be included. Recently, the utility of including prior information, in this case based on a search of medical abstracts, has been demonstrated [30].

In our VISP analysis, we chose the value $\pi_1 = 0.0001$ by examining frequentist summaries before the real data analysis was performed. We define π_1 as the proportion of SNPs that would be associated with the disease, if the power were 1. This proportion includes tag SNPs and so we would not expect the number of non-null signals to be binomial (since the signals do not correspond to independent Bernoulli random variables, which will induce overdispersion) though under our chosen prior, one would expect $J \times \pi_1 = 80$ non-null interactions. After the data are analyzed we can, for those SNPs declared as null (i.e. all but 3 SNPs in the VISP trial), sum up the posterior probabilities of being non-null, and this gives the expected number of false non-discoveries (where the expectation is over the posterior distribution). For the VISP data, this expected number is 24.6 so that we are missing a large number of signals, with lack of power being a major issue. For the three significant signals, at the 0.5 threshold, the probabilities of the null being true are 0.01, 0.04

and 0.19, so that the expected number of false discoveries is 0.24. Taking the threshold of significance as 0.25 gives an additional SNP as being declared significant. The sum of the posterior probabilities of the null is 0.98 in this case and so, under this prior, we would expect one of the reports signals to be a false discovery. The supporting material contains a discussion of the results under the more conservative $\pi_1 = 0.00001$ prior.

A related interesting exercise is to simulate the distribution of observed effect sizes under our assumed priors (on both the proportion of non-null signals and the effect sizes), using the observed distribution of standard errors. The distribution of effect sizes is $N(\Delta, V + W)$ for the non-null signals and $N(0, V)$ for the null signals. We can then evaluate the power, and hence determine the number of signals we would expect to detect given our prior assumptions. For the VISP data, with a proportion of non-null signals $\pi_1 = 0.0001$, R= 1 and 95% range for the effect sizes of ±10, we would expect to see 52 true positives and one false positive. Given we only observed three non-null signals, this implies that either the range of effect sizes (as defined through W) was too wide or, more probably, that our estimate of π_1 was optimistic. Repeating this exercise with $\pi_1 = 0.00001$ gives 5 true positives and close to 0 false positives, which is more consistent with that which was observed.

The posterior probability (and the Z-score threshold) is equally sensitive to R as to π_1, as one can see from the symmetry in (6). The form of the latter suggests that all we need to do is to fix PO/R. As mentioned above, in the VISP analysis we selected π_1 by examining the frequentist operating characteristics. An alternative method [26] for obtaining PO/R is to specify a value for the Z^2 boundary, z^2_{B}, at a particular V (for example, at a MAF and sample size that one is familiar with) and then solve for $U = \log(\mathrm{PO/R})$ via

$$\hat{U} = \frac{z^2_{\mathrm{B}} \times W}{2\,(V+W)} - \frac{1}{2} log\left(\frac{V+W}{V}\right).$$

With this value of PO/R one can then proceed to use (6) across the observed range of standard errors.

Finally, we describe some alternative analyses of the VISP data and model extensions.

Alternative Genetic Model

In the analyses of the paper we assume a recessive genetic model. As an alternative, in the supplementary material we report the results from the fitting of an additive model. For these data the three most significant SNPs were the same under both the recessive and additive models and the posterior probabilities of non-null association were very similar.

Constant Variance Assumption

It may be important to allow different variances in each of the treatment groups [6,7] to improve power and we have carried out an additional analysis to address this issue. With respect to model (1) we allowed for different variances in the two treatment groups, and derived the asymptotic variance $\hat{V}$ under this model. In addition, we carried out an analysis using sandwich estimation. The results are reported in the supplementary material; for these

data, the conclusions show only small changes. This is not surprising here, since it is known that for equal sized groups the two-sample *t* test gives the same statistic [31].

Joint Modeling of Main Effect and Interaction

Another important modeling choice that has been considered previously [4–8] is the joint relationship between the main effect of genotype and the interaction. There are two components to this modeling: one may jointly model the probabilities of the *events* "main effect is zero" and "interaction effect is zero", and one may also jointly model the *sizes* of the main and interaction effects. Here we have chosen, primarily for simplicity, to model the effects separately in the sense that the existence or size of a main effect has no impact on our prior on the existence or size of an interaction. In other approaches [5] interactions are not allowed unless main effects are present. In our context we would not wish to make this strong assumption since the marker effect may only exist amongst the treated, that is, with respect to model (1) there may be situations in which $\gamma = 0$ and $\Delta \neq 0$. In other situations one may wish to model the dependence between main and interaction effects' in the supplementary materials we describe a more complex model in which one may encode more refined prior beliefs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This was research funded by the GARNET grant NHGRI U01 HG005157.

References

1. Shen Y, Nicoletti P, Floratos A, Pirmohamed M, Molokhia M, Geppetti P, et al. Genome-wide association study of serious blistering skin rash caused by drugs. The Pharmacogenomics Journal. 2012; 12:96–104. [PubMed: 21221126]
2. de Keyser CE, Eijgelsheim M, Hofman A, Sijbrands EJG, van der Zee AHM, van Duijin CM, et al. Single nucleotide polymorphisms in genes that are associated with a modified response to statin therapy: the Rotterdam Study. The Pharmacogenomics Journal. 2012; 11:72–80. [PubMed: 20195290]
3. Bacanu SA, Whittaker JC, Nelson MR. How informative is a negative finding in a small pharmacogenetic study? The Pharmacogenetics Journal. 2012; 12:93–95.
4. Barber JM, Mangravite LM, Hyde CL, Smith DICJD, McCarty CA, Li X, et al. Genome-Wide Association of Lipid-Lowering Response to Statins in Combined Study Populations. PLoS One. 2010; 5:e9763. [PubMed: 20339536]
5. Maranville JC, Luca F, Richards AL, Wen X, Witonsky DB, Baxter S, et al. Interactions between Glucocorticoid Treatment and Cis-Regulatory Polymorphisms Contribute to Cellular Response Phenotypes. PLoS Genetics. 2011:e1002162. [PubMed: 21750684]
6. Wen X, Stephens M. Bayesian methods for genetic association analyses with heterogeneous subgroups: from meta-analysis to gene-environment interactions. Annals of Applied Statistcs. 2013 Under revision.
7. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genetics. 2013; 9:e1003486. [PubMed: 23671422]

8. Mangravite LM, Engelhardt BE, Medin MW, Smith JD, Brown CD, Chasman DI, et al. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. Nature. 2013 Available online, March 2013.
9. Wakefield JC. Reporting and Interpretation in Genome-Wide Association Studies. International Journal of Epidemiology. 2008; 37:641–653. [PubMed: 18270206]
10. Spence JD, Howard VJ, Chambless LE, Malinow MR, Pettigrew LC, Stampfer M, et al. Vita-min Intervention for Stroke Prevention (VISP) trial: rationale and design. Neuroepidemiology. 2001; 16:16–25. [PubMed: 11174041]
11. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273:1516–1517. [PubMed: 8801636]
12. Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SCL, et al. Parameters for reliable results in genetic association studies in common disease. Nature Genetics. 2002; 30:149–150. [PubMed: 11799396]
13. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science. 2005; 308:385–389. [PubMed: 15761122]
14. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature Genetics. 2007; 39:865–869. [PubMed: 17529974]
15. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. American Journal of Human Genetics. 2007; 81:208–227. [PubMed: 17668372]
16. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
17. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics. 2008; 9:356–369.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289–300.
19. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B. 2002; 64:479–498.
20. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Science. 2003; 100:9440–9445.
21. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics. 2007; 39:906–913. [PubMed: 17572673]
22. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantative traits. PLoS Genetics. 2007; 3:1296–1308.
23. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009; 10:681–690.
24. Pamigiani, G.; Inoue, L. Decision Theory: Principles and Approaches. John Wiley and Sons; 2009.
25. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). Journal of the Royal Statistical Society, Series A. 1994; 157:357–416.
26. Wakefield JC. Commentary: Genome-wide significance thresholds via Bayes factors. International Journal of Epidemiology. 2012; 42:286–291. [PubMed: 22345299]
27. Wakefield J. Bayes factors for genome-wide association studies: comparison with P -values. Genetic Epidemiology. 2009; 33:79–86. [PubMed: 18642345]
28. Li Y, Martin BR, Cravatt BF, Hofmann SL. DHHC5 protein palmitoylates flotillin-2 and is rapidly degraded on induction of neuronal differentiation in cultured cells. The Journal of Biological Chemistry. 2012; 287:523–530. [PubMed: 22081607]
29. Berger T, Ueda T, Arpaia E, Chio II, Shirdel EA, Jurisica I, et al. Flotillin-2 deficiency leads to reduced lung metastases in a mouse breast cancer mode. Oncogene. 2012 Epub Ahead of Print.

30. Johansson M, Roberts A, Chen D, Li Y, Delahaye-Sourdeix M, Aswani N, et al. Using prior information from the medical literature in GWAS of oral cancer identifies novel susceptibility variant on chromosome 4 – the AdAPT method. PLoS One. 2012; 7

31. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annual Review of Public Health. 2002; 23:151–169.

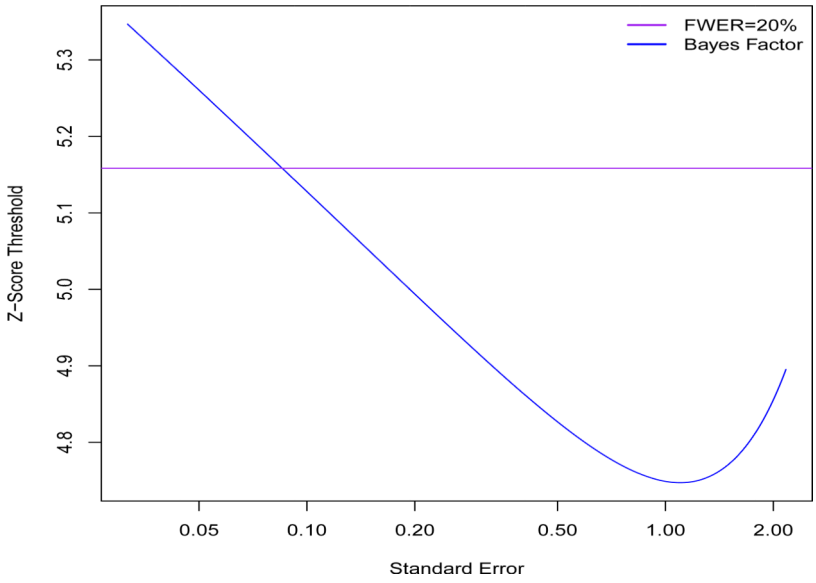

Figure 1.

Bayesian *Z*-score threshold as a function of the standard error; FWER is the family-wise error rate which is set at 0.2; here we take the number of tests as $J = 803,122$. The Bayesian threshold is based on a prior on the alternative of 0.0001, a ratio of costs of type II to type I errors of R = 1 and a prior standard deviation on the interaction effect size of $\sqrt{W}=5.1$; this prior gives a 95% interval of (−10,10).

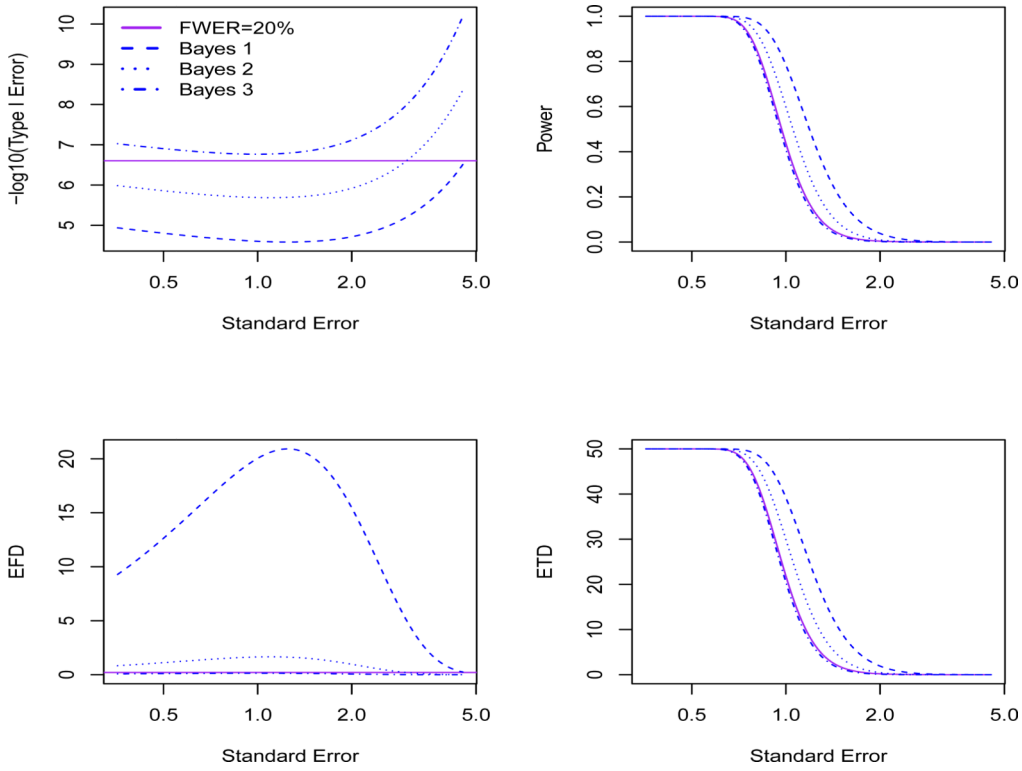

Figure 2.

Operating characteristics of prospective Bayesian decision boundaries. A comparison with a Bonferroni correction corresponding to a family-wise error rate of 0.2 is also included. For all Bayesian boundaries R = 1 and “Bayes 1”, “Bayes 2”, “Bayes 3” correspond to priors of π_1 = 0.001, 0.0001, 0.00001. In the top right panel the power to detect a drop of 5 is plotted. The expected number of false discoveries (EFD) is plotted in the lower left panel and the expected number of true discoveries (ETD) in the lower right. It is assumed that there are J_1 = 50 non-null associations.

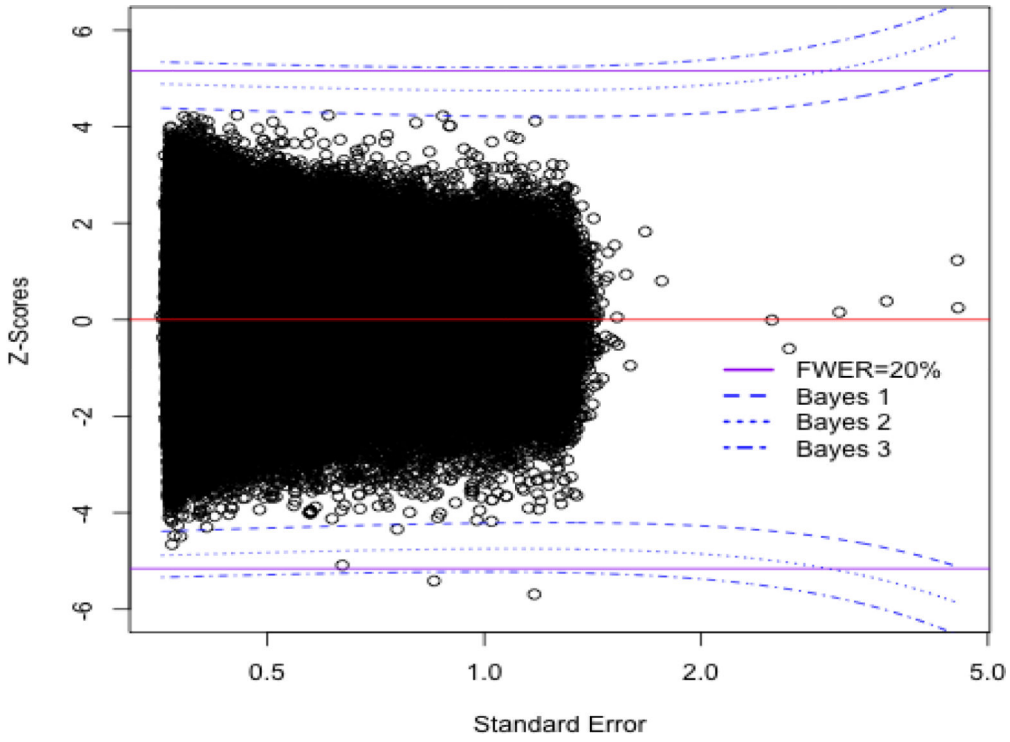

Figure 3.

Z-score threshold as a function of the standard error for the VISP data, ratio of costs of type II to type I errors R = 1 and varying priors on the alternative of π_1 = 0.001, 0.0001, 0.00001 (to give Bayes 1, Bayes 2, Bayes 3 boundaries, indicated in blue). The purple lines correspond to the Bonferroni correction for a family-wise error rate of 20%.

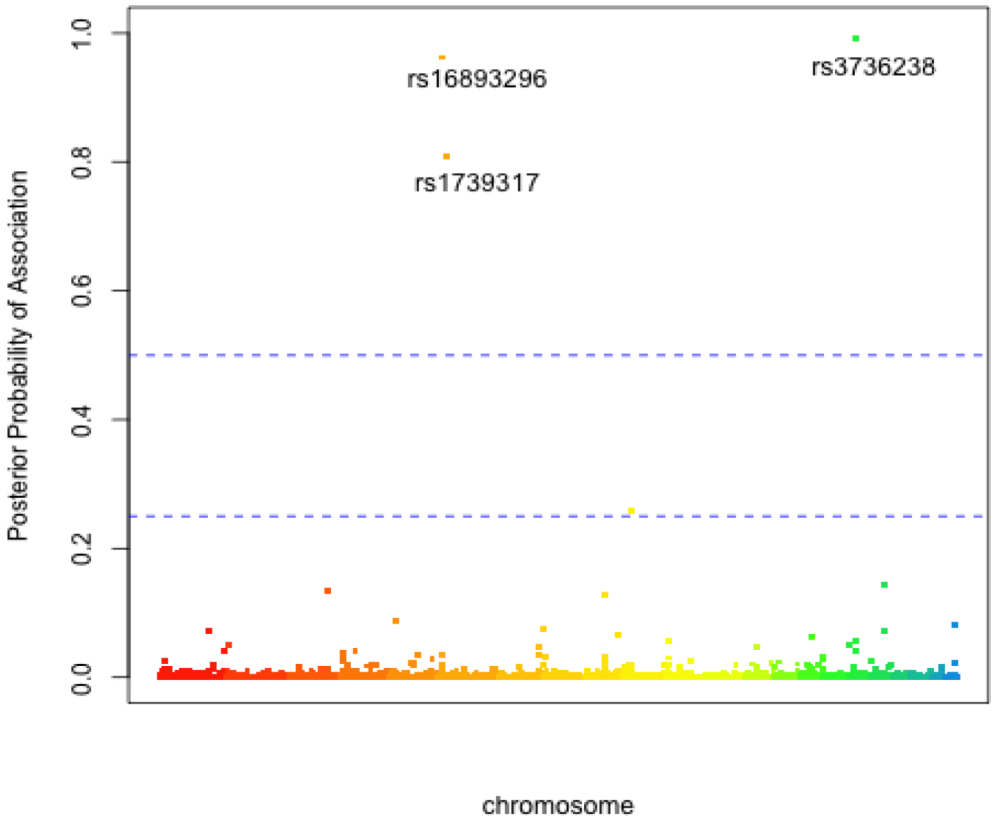

Figure 4.

Posterior probability on the alternative plotted versus genomic position for the VISP data. Each point corresponds to a marker subgroup so that points closer to 1 have a greater probability of corresponding to a significant interaction. The prior on the alternative is π_1 = 0.0001. The horizontal dashed blue line at 0.5 corresponds to a threshold with R = 1 (equal costs of type I and type II errors) while the line at 0.25 corresponds to R = 3 (type II errors being three times worse than type I errors). SNP labels are attached to the three SNPs which we would declare as significant under the R = 1 rule.

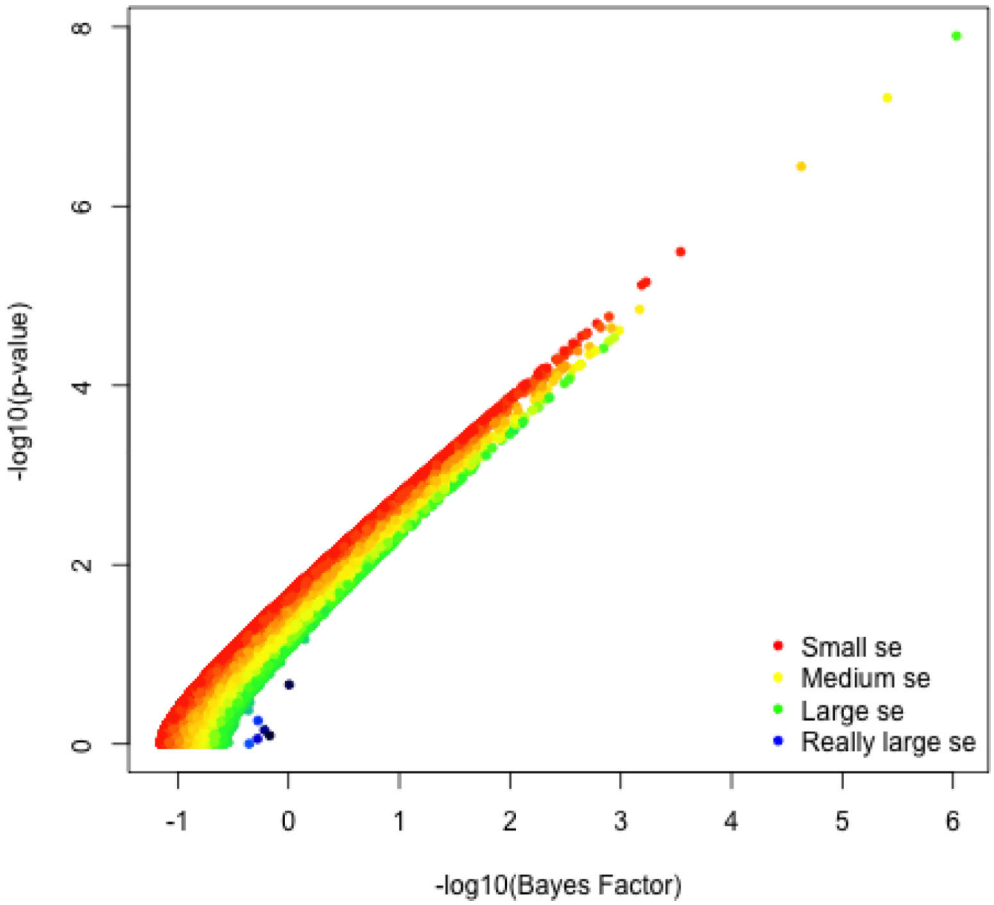

Figure 5.

$-\log_{10}$BFs versus $-\log_{10} p$-values, color-coded by standard error with $W = 10$. The signals with larger standard errors are generally more significant under the Bayesian approach than under the approach based on Z-scores only.

Table 1

2×2 table of losses for the case of two hypotheses H_0 and H_1; C_{I} and C_{II} are the costs of type I and type II errors, respectively.

		Truth	
		H_0	H_1
Decision	H_0	0	C_{11}
	H_1	C_1	0