ORIGINAL ARTICLE

Detecting signals in pharmacogenomic genome-wide association studies

J Wakefield1,2, V Skrivankova2, F-C Hsu3, M Sale4 and P Heagerty2

In a 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 the statistical analysis focuses on the interaction between treatment and genotype. However, genome-wide analysis in such scenarios requires careful statistical thought as, 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 that 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 vary across tests in a manner that 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.

The Pharmacogenomics Journal (2014) 14, 309–315; doi:10.1038/tpj.2013.44; published online 7 January 2014

Keywords: Bayes factors; Bonferroni correction; significance threshold

INTRODUCTION

The key statistical question in pharmacogenomic studies, in which genetic marker-defined subgroups are examined to see whether 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 single nucleotide polymorphisms (SNPs) were examined using logistic regression. The authors quote $5 \times 10^{-8}$ 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, although 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 (for example, 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 di allelic SNP, the two comparison subgroups of interest would be those who 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—that is, those 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 nonzero interactions will typically vary hugely over this collection, because the subgroup sizes are a direct function of the minor allele frequency. 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 errors vary across tests; one would prefer an approach in which both the type I and type II error rates 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 previously3 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 approaches have previously been suggested.4–8

1Department of Statistics, University of Washington, Seattle, WA, USA; 2Department of Biostatistics, University of Washington, Seattle, WA, USA; 3Department of Biostatistical Sciences, Wake Forest University, Winston-Salem, NC, USA and 4Robert M. Berne Cardiovascular Research Center, University of Virginia, Charlottesville, VA, USA. Correspondence: Dr J Wakefield, Department of Statistics, University of Washington, Box 357232, Seattle, WA 98195, USA.
E-mail: jonno@uw.edu
Received 20 February 2013; revised 2 November 2013; accepted 12 November 2013; published online 7 January 2014
The issue of variable power across multiple tests is common to all testing situations, but is acute in pharmacogenomic situations because the subset sizes are highly variable across tests. The difficulties associated with fixed threshold rules have 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 trial (VISP) trial, which is an EU-funded, multicenter, double-blinded, 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 end points. One individual was removed as the participant’s data were considered outlying; this individual’s data and the implications of retaining the data are discussed in the Supplementary Materials. We examined data from n = 167642 SNPs of European ancestry, with 837 randomized to the high dose and 833 to the low dose. After quality control procedures, 803122 SNPs were available for analysis. The outcome is the intermediate variable homocysteine level in the VISP trial, levels were measured longitudinally but, for simplicity, we use the baseline measurements. The average change was 0.37 μmol l−1 in the low-dose group versus −2.36 μmol l−1 in the high-dose group, to give a difference of −1.99 μmol l−1 (P < 2 × 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 Yi and Ti = 0/1 represent the response and treatment indicator for individual i = 1,...,n, and let Mi = 0/1 be a marker for individual i. The choice of a recessive model is made for illustration only, and the fundamental aspect must be considered when determining a threshold. However, as with the FWER, the threshold should also tend toward zero with increasing sample size, so that both type I and II errors tend 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|H0)/p(data|H1). 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 p0 represent the prior on the null being true, so that PO = p0/(1 − p0) is the prior odds on the null. Via Bayes theorem, we can then evaluate the posterior probability that the null is true as

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

with \(BF = \frac{p(\text{data}|H_0)}{p(\text{data}|H_1)}\). 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 1/3. 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 \(R = C_0/C_1\) is the ratio of the costs of type I to type II errors. Historically, a type I error has been viewed as more harmful than a type II error, but this

| Decision | H0 | H1 |
|----------|----|----|
| H0       | 0  | C1 |
| H1       | C0 | 0  |

Table 1. 2 × 2 table of losses for the case of two hypotheses H0 and H1; C1 and C0 are the costs of type I and type II errors, respectively
slant is not always desirable. A Bayesian decision theory approach chooses $H_1$ if the posterior odds on $H_0$ fall below $R$, which, from model (2), occurs if

$$BF \propto PO < R$$

(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 model (1), priors would be required for $\alpha, \phi, \beta, \gamma, \Delta$ and $\sigma$. 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. The idea is to take the ‘data’ as $\Delta$, in which case we have a likelihood $p(\Delta | \Lambda)$. 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, particularly in the context of clinical trials.

The model for a generic marker with interaction parameter $\Delta$ is

$$\text{Likelihood: } \Delta | \Lambda \sim N(0,W)$$

(4)

$$\text{Prior: } \sqrt{W} \sim N(0,W)$$

(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 $W$. In terms of the likelihood, only a confidence interval is required for specification, as this interval may be used to find $\Delta$ and $V$ for use in model (4).

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

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

where $Z$ is the $Z$ statistic. A crucial observation is that the evidence is based on the $Z$ score and on the s.e. $\sqrt{V}$, 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 s.e. $\sqrt{V}$. From rearrangement of model (3) the Bayesian $Z^*$ score threshold for rejection is:

$$Z^* > Z_\theta = \left( \frac{V+W}{W} \right) \left\{ \log \frac{V+W}{V} + 2 \log \left( \frac{PO}{R} \right) \right\}$$

(6)

to give a threshold that is an explicit function of $V, R$ and $PO$. Notice that this Bayesian derivation makes no mention of type I and type II errors, although 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 model (6). If the prior odds on the null ($PO$) increases, the threshold increases so that we require more evidence to overcome the initial skepticism. 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 also decreases to zero. Specifically, let $n$ denote an appropriate measure of sample size and write $V = kn$, where $k$ is not a function of $n$. Then, as $n \to \infty$, from (6)

$$Z_\theta \to \log \left( 1 + \frac{nW}{k} \right) + 2 \log \left( \frac{PO}{R} \right)$$

so that the boundary increases, and so the type I error tends to zero. It can also be shown$^{27}$ that the Bayes factor tends to zero under the alternative as $n \to \infty$. 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 determined. Figure 1 illustrates this behavior of the Bayesian threshold, given by the square root of (6), for a range of s.e. values that are consistent with the VISP trial. The critical thresholds were taken from model (6) with $\pi_1 = 0.0001$, so that with $J = 803122$ tests we would expect around $J \times \pi_1 = 80$ signals. These signals will not reflect 80 different causal variants, as typically multiple SNPs will tag each causal variant. For reference, the Bonferroni threshold for a FWER of 20% (corresponding to a $p$-value of $0.02/803122 = 2.4 \times 10^{-8}$) is indicated as a horizontal line. For small $n$ (large s.e.) the Bayesian approach increasingly requires greater evidence because of the low power. For large $n$ (small s.e.) 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, as it allows one to choose a threshold rule aided by consideration of type I and type II error probabilities. For ranges of the s.e. consistent with the VISP data (Supplementary 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 $J = 803122$ 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 s.e. 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 five 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 s.e. values between 0.5 and 2.

We now transform the type I error and power into more intuitive quantities, the EFD and the expected number of true
highly sensitive to the choice of the fraction of non-null associations is close to zero, the ETD 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, as at each s.e. it is assumed that all of the signals have this s.e.. We also stress that when we specify a prior $\pi_1$, 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 section.

The most liberal prior of $\pi_1 = 0.001$ produces a large number of type I errors (around 20 for s.e. in the mid-range) and might be judged to give unacceptably poor performance. The most skeptical 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 s.e. = 1, around two 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 s.e. = 1, if $J_1 = 5$ then we would only expect to detect a single additional signal when compared with the use of Bonferroni. Armed with this information we move to an analysis of the VISP data.

VISP analysis

We fitted model model (1) to the $J = 803\,122$ SNPs with gender and age (by quintile) being included and corresponding to the $x_i$ term in model (1). The genetic subgroups are defined as having at least one copy of the minor allele as compared with two copies of the major allele. The number in the former subgroup ranges between 21 and 1564 across SNPs. The s.e. of $\Lambda$ is $\sigma \sqrt{1/n_{t0} + 1/n_{t1} + 1/n_{t01} + 1/n_{t11}}$ where $n_{t0}, n_{t1}, n_{t01}, n_{t11}$ are the number of individuals in treatment group $t$, $t = 0/1$, and marker subgroup $m$, $m = 0/1$, for a generic SNP. Hence, the s.e. 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 $\times$ marker cells across the $J$ comparisons.

Figure 3 plots the $Z$ scores versus the s.e., 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 $\Delta \pm 10$. 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 $2.5 \times 10^{-7}$) two SNPs are flagged. With a FWER of 5% the Bonferroni threshold is $6.2 \times 10^{-8}$ and results in a single SNP being deemed significant. With the more optimistic prior of $\pi_1 = 0.0001$, a further signal is flagged (and is not significant using Bonferroni). There are few strong signals for these data, however. Supplementary 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 three SNPs that fall outside of the $\pi_1 = 0.0001$ 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 $\Delta = -6.7$ (so that we have an enhanced effect), with

![Figure 2](image-url)
Figure 3. Z-score threshold as a function of the s.e. for the VISP data, ratio of costs of type II to type I errors $R = 1$ and varying priors on the alternative of $\pi_1 = 0.001, 0.0001, 0.00001$ (to give Bayes 1, Bayes 2, Bayes 3 boundaries). The horizontal lines correspond to the Bonferroni correction for a FWER 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 $\pi_1 = 0.0001$. The horizontal dashed line at 0.5 corresponds to a threshold with $R = 1$ (equal costs of type I and type II errors), whereas 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 Bonferroni correction for a FWER of 20%.

To conclude that 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 $\Delta$ to $\pm \Delta$ gives $1/BF = 1.6 \times 10^4$ and a posterior probability on the alternative of 0.38. Supplementary 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 $\Delta$. 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). Supplementary 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 caveola-asssociate protein, which may function in neuronal signaling (www.geneCards.org). A recent study 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 malignant transformation. 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 $\pi_1$), we would recommend that further work, preferably at the molecular level, be carried out for confirmation. The minor allele frequency is 1.3% for this SNP and so the enhanced effect will be seen only in a small group of individuals.

The second signal is rs16893296 on chromosome 6 with a posterior probability on the alternative of 0.96. The point estimate and s.e. are $-4.6$ and 0.85 with a $P$-value of $7.1 \times 10^{-8}$. The nearby genes are LOC442160 and LOC442161. This variant is in a weak DNaasel 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 $4.0 \times 10^{-7}$) is located on chromosome 6 with nearby genes G6orf32 and LOC134997. The evidence that this SNP is functional is weak as this variant is also in a weak DNaasel 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 S5 shows that the $P$-values and Bayes factors differ in their rankings because of the differing sample sizes or s.e. values. The points are color coded by the size of the s.e. and we see that the points with larger s.e. values are consistently ranked as giving greater evidence for the alternative under the Bayesian approach. This behavior occurs because of the association between the $Z^2$ boundary and the s.e. 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. Supplementary Figure S3 of the supporting material shows an example in which distinctly different behavior occurs.
number is 24.6 and hence we are potentially 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 EFD 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 four reported 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 S.E. values. 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 \( \pm 10 \), we would expect to see 52 true positives and 1 false positive. Given that the fact that we observed only 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 \( \pi_1 \) was optimistic. Repeating this exercise with \( \pi_0 = 0.00001 \) gives five true positives and close to zero 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 \( \pi_1 \), as one can see from the symmetry in (6). The form of the latter suggests that all we need to do is fix \( PO/R \).

As mentioned above, in the VISP analysis we selected \( \pi_1 \) by examining the frequentist operating characteristics. An alternative method\(^{25}\) for obtaining \( PO/R \) is to specify a value for the \( Z^2 \) boundary, \( z_{\beta}^2 \), at a particular \( V \) (for example, at a minor allele frequency and sample size that one is familiar with) and then solve for \( U = \log(PO/R) \) via

\[
U = \frac{z_{\beta}^2 \times W}{2(V + W)} - \frac{1}{2} \log\left(\frac{V + W}{V}\right)
\]

with this value of \( PO/R \) one can then proceed to use (6) across the observed range of S.E. values.

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 Materials 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\(^6\) to improve power, and we have carried out additional analyses to address this issue. With respect to model model (1) we allowed for different variances in the two treatment groups and derived the asymptotic variance \( V \) under this model. In addition, we carried out an analysis using sandwich estimation. The results are reported in the Supplementary Materials; for these data, the conclusions show only small changes. This is not surprising here, as 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\(^{25}\) is the joint relationship between the main effect of

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 \( \pi_0 \) and a sensitivity analysis is always warranted. Ideally, rather than fix \( \pi_0 \) as we have done, one would preferably estimate \( \pi_0 \) from the totality of data (that is, 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 nonzero effects, but the power to detect these signals is low with the usual sample sizes. In other contexts, such as the analysis of gene expression data, the observed range of S.E. values. The distribution of effect sizes, this implies that either the range of effect sizes (as defined through \( W \)) was too wide or, more probably, that our estimate of \( \pi_1 \) was optimistic. Repeating this exercise with \( \pi_0 = 0.00001 \) gives five true positives and close to zero 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 \( \pi_1 \), as one can see from the symmetry in (6). The form of the latter suggests that all we need to do is fix \( PO/R \).

As mentioned above, in the VISP analysis we selected \( \pi_1 \) by examining the frequentist operating characteristics. An alternative method\(^{25}\) for obtaining \( PO/R \) is to specify a value for the \( Z^2 \) boundary, \( z_{\beta}^2 \), at a particular \( V \) (for example, at a minor allele frequency and sample size that one is familiar with) and then solve for \( U = \log(PO/R) \) via

\[
U = \frac{z_{\beta}^2 \times W}{2(V + W)} - \frac{1}{2} \log\left(\frac{V + W}{V}\right)
\]

with this value of \( PO/R \) one can then proceed to use (6) across the observed range of S.E. values.

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 Materials 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\(^6\) to improve power, and we have carried out additional analyses to address this issue. With respect to model model (1) we allowed for different variances in the two treatment groups and derived the asymptotic variance \( V \) under this model. In addition, we carried out an analysis using sandwich estimation. The results are reported in the Supplementary Materials; for these data, the conclusions show only small changes. This is not surprising here, as 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\(^{25}\) 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,1 interactions are not allowed unless main effects are present. In our context we would not wish to make this strong assumption as the marker effect may exist only among the treated—that is, with respect to model (1) there will be situations in which γ = 0 and Δ ≠ 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.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

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

REFERENCES
1 Shen Y, Nicoletti P, Floratos A, Pirmohamed M, Molokhia M, Gepetti P et al. Genome-wide association study of serious blistering skin rash caused by drugs. Pharmacogenomics J 2012; 12: 96–104.
2 de Keyser CE, Eigelschheim M, Hofman A, Sijbrands EJG, van der Zee AHM, van Duijn CM et al. Single nucleotide polymorphisms in genes that are associated with Alzheimer’s disease: the Rotterdam Study. Pharmacogenomics J 2012; 11: 72–80.
3 Bacanu SA, Whitaker JC, Nelson MR. How informative is a negative finding in a small pharma-cogenetic study? Pharmacogenetics J 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 Genetics 2011; 7: e1002162.
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; 7: e1002162.
6 Wen X, Stephens M. Bayesian methods for genetic association analyses with heterogeneous subgroups: from meta-analysis to gene-environment interactions. Annals of Applied Statistics 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.
8 Mangravite LM, Engelhardt BE, Medin NW, Smith JD, Brown CD, Chasman DI et al. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. Nature 2013; 7: e36888 Available online, March 2013.
9 Wakefield JC. Reporting and interpretation in genome-wide association studies. Int J Epidemiol 2008; 37: 641–653.
10 Spence JD, Howard VJ, Chambless LE, Malinow MR, Pettigrew LC, Stampfer M et al. Vitamin Intervention for Stroke Prevention (VISP) trial: rationale and design. Neuroepidemiology 2001; 16: 16–25.
11 Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996; 273: 1516–1517.
12 Dahlman I, Eaves JA, Kosoy R, Morrison VA, Heward J, Gough SCL et al. Parameters for reliable results in genetic association studies in common disease. Nat Genet 2002; 30: 149–150.
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.
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. Nat Genet 2007; 39: 865–869.
15 Wakefield JA. Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 2007; 81: 208–227.
16 Welcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature 2007; 447: 661–678.
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. Nat Rev Genet 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. Proc Natl Acad Sci 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. Nat Genet 2007; 39: 905–913.
22 Seivin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genetics 2007; 3: 1296–1308.
23 Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet 2009; 10: 681–690.
24 Pamigiani G, Inoue L. Decision Theory: Principles and Approaches. John Wiley and Sons: Chichester, UK, 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. Int J Epidemiol 2012; 41: 286–291.
27 Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol 2009; 33: 79–86.
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. J Biol Chem 2012; 287: 523–530.
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 model. Oncogene 2012; 32: 4989–4994.
30 Johansson M, Roberts A, Chen D, Li Y, Delahaye-Sourdex M, Awuni 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: e36888.
31 Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002; 23: 151–169.

Supplementary Information accompanies the paper on the The Pharmacogenomics Journal website (http://www.nature.com/tpj)