Quantifying posterior effect size distribution of susceptibility loci by common summary statistics

February 28, 2020

Olga A. Vsevolozhskaya,[1] Dmitri V. Zaykin[2,*]

[1]Biostatistics Department, University of Kentucky, Lexington, KY, 40536, USA

[2]Biostatistics and Computational Biology, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, 27709, USA

*Correspondence: Dmitri V. Zaykin, Senior Investigator, Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, P.O. Box 12233, Research Triangle Park, NC 27709, USA. Tel.: +1 984-287-3694; Email address: dmitri.zaykin@nih.gov

Abstract

Testing millions of SNPs in genetic association studies has become standard routine for disease gene discovery. In light of recent re-evaluation of statistical practice, it has been suggested that P-values are unfit as summaries of statistical evidence. Despite this criticism, P-values contain information that can be utilized to address the concerns about their flaws. We present a new method for utilizing evidence summarized by P-values for estimating odds ratio (OR) based on its approximate posterior distribution. In our method, only P-values, sample size, and standard deviation for ln(OR) are needed as summaries of data, accompanied by a suitable prior distribution for ln(OR) that can assume any shape. The parameter of interest, ln(OR), is the only parameter with a specified prior distribution, hence our model is a mix of classical and Bayesian approaches. We show that our method retains the main advantages of the Bayesian approach: it yields direct probability statements about hypotheses for OR and is resistant to biases caused by selection of top-scoring SNPs. Our method enjoys greater flexibility than similarly inspired methods in the assumed distribution for the summary statistic and in the form of the prior for the parameter of interest. We illustrate our method by presenting interval estimates of effect size for reported genetic associations with lung cancer. Although we focus on OR, the method is not limited to this particular measure of effect size and can be used broadly for assessing reliability of findings in studies testing multiple predictors.

Keywords: approximate Bayes methods; strength of evidence; prior distributions; P-values

1 Introduction

Modern human genetic studies routinely examine a very large number of potential associations of genetic variants with health-related outcomes. In the majority of studies, results are filtered by P-values with the most promising results selected based on a significance threshold, adjusted to accommodate the number of tests in a study. Partly due to their widespread use, P-values have been at the center of replicability crisis (Wasserstein and Lazar, 2016; Johnson, 2013; Greenland et al., 2016). Cumming (2008) and Halsey et al. (2015) demonstrated that P-values are 'fickle' in that they can vary greatly between replicates even when statistical power is high. P-values also do not measure confidence regarding a hypothesis, even though the operational use of P-values is to make decisions about rejecting/accepting the null hypothesis (Lazzeroni et al., 2014, 2016). Although P-values are poorly suited for what they are used for in practice, they are efficient summaries of data. For example, Lin and Zeng (2010) and Zaykin (2011) have shown that meta-analysis of summary statistics (e.g., P-values) can be statistically nearly as efficient as joint analysis of individual participant data. Thus, P-values contain information that can be used to judge the degree of uncertainty about hypotheses and parameters. The Bayesian framework gives researchers the ability to estimate the probability of their hypothesis directly. There is a number of successful proposals where information summarized by P-values is used for Bayesian or approximately Bayesian inference. The false discovery rate (FDR) is perhaps the most well known example. The FDR is defined as an estimated proportion of false discoveries among test results that are deemed to be "discoveries," based on statistical criteria. The FDR has distinct Bayesian characteristics because the posterior probability of a hypothesis, averaged over discoveries, is the proportion of false discoveries estimated for a given study. The original FDR method by Benjamini and Hochberg is not a truly Bayesian approach, nor was it intended to be one by design, because the FDR controls the rate of erroneous discoveries out of all discoveries made as a long run average, taken over experiments with or without discoveries. When statistical power is high (and hence assuming that not all of study's null hypotheses are true), Benjamini-Hochberg's FDR can be viewed as a proportion of false findings estimated concerning a specific study (Benjamini and Hochberg, 1995; Zaykin et al., 2000). That is, conveniently, as sample size increases, Benjamini-Hochberg's FDR approaches the Bayesian posterior proportion of true null hypotheses among rejected hypotheses. Likewise,

based solely on P-values are Empirical Bayes counterparts of the Benjamini-Hochberg FDR, such as Storey's q-value (Storey, 2002) and the local FDR (Efron et al., 2001). These two methods preserve the ranking of P-values.

The Empirical Bayes methods became popular in applications such as the analysis of differential gene expression. Empirical Bayes approaches for genome-wide association studies are starting to emerge (Spencer et al., 2016). Among the challenges is sparsity as well as weakness of the true signals in the genome, causing the false discovery rate variance to be relatively high compared to the declared target rate (Dudbridge et al., 2006). Current knowledge of genetics of complex diseases may allow one to specify plausible prior parameters explicitly. For example, it would be reasonable to assume that there is a very small *a priori* chance that a randomly picked SNP would carry OR value greater than 3, and that probability of OR value being at least 2 is also small, but definitely larger. Further, one can place with high confidence a large prior probability for the OR to be around 1. Thus, a biologically realistic prior distribution for an effect size can be constructed based on existing prior knowledge.

A simple method that accommodates explicit prior assumptions is the False Positive Report Probability (FPRP) by Wacholder et al. (2004). The method received great attention and has been cited over 1,500 times at the time of this writing, in part due to its simplicity. In the FPRP, the prior proportion of non-associated SNPs, $\Pr(H_0)$, should be specified, as well as a prior value for OR. That is, a single OR value is used to approximate the actual OR distribution across associated variants. Given these two assumed values, power can be estimated for a given SNP. Then, FPRP/(1-FPRP) can be calculated as the ratio of P-value over the computed power times the prior odds for H_0. The FPRP method has a number of shortcomings. Among the main ones is the assumption that a single value of OR is representative of the OR distribution, and the use of cumulative probabilities (P-value and power), which leads to an undesirable property that the FPRP can never be greater than the assumed prior probability of H_0 (Lucke, 2009; Wakefield, 2008).

Wakefield (2007) proposed a Bayesian method that is similar in simplicity to FPRP but remedies its shortcomings. Wakefield's Bayesian measure of the probability of false discovery and his approximate Bayes factor (ABF) utilize information contained in a test statistic for ln(OR) to obtain an approximate posterior effect size distribution for a given SNP, and the corresponding Bayes

factor. To approximate the posterior effect size distribution for real signals, the ABF method assumes that the effect size distribution across the genome, measured by the logarithm of odds ratio, can be approximated by a normal distribution centered at zero. Large values of the variance parameter, W, of that prior distribution imply large spread of ln(OR) values, and therefore, high probability of encountering variants with relatively large effect size. Non-associated variants in ABF are modeled as having zero effect size. Alternatively, if for non-associated SNPs we assume a normal distribution with nearly zero variance, then the prior for both real and non-associated variants becomes a mixture of two normal distributions, both centered at zero. Thus, the prior distribution in ABF is controlled by two parameters, the mixture proportion of non-associated SNPs, $\Pr(H_0)$, and the variance of the effect size distribution for associated SNPs, W.

The ABF method and its extensions (Spencer et al., 2015) are "approximate Bayes" methods, because the posterior calculation, which is based on a Z-statistic for ln(OR), assumes the prior distribution for the parameter of interest only, i.e., ln(OR). In contrast, a fully Bayesian model would have to include a joint prior distribution for covariates of the model along with ln(OR). It can be difficult to specify a realistic joint prior for all of the parameters, and the ABF finds a middle ground in which approximate, yet surprisingly precise posterior estimates for ln(OR) are obtained by utilizing its prior distribution and by plugging in frequentist estimates for any additional parameters. The ABF has desirable statistical properties, including explicit drawing on power through dependence on the standard error. Note that P-values are calculated assuming H_0, and under this assumption they are not a function of the sample size. This property of ABF leads to a generally different ranking than ranking by P-values and facilitates discovery of real signals (Wakefield, 2008).

A great practical advantage of the ABF is that only ln(OR), the corresponding normal Z-statistic, and the standard error of ln(OR) are needed as summaries of data. In this paper, we propose a method that similarly requires only these summaries of data but is not limited to OR as a measure of effect size and offers much greater flexibility in terms of the prior distribution for ln(OR). Similarly to ABF, our Flexible Approximate Bayes (FAB) approach improves precision of effect size estimates for multiple SNPs tested within a study and makes use of summary statistics to estimate the posterior distribution for the parameter of interest. However, the core novelty of our approach is the ability for a researcher to use any arbitrary prior effect size distribution, e.g., based

on previously published empirical results. Our method includes ABF as a special case when the conjugate normal model is assumed, therefore it shares ABF's advantageous statistical properties demonstrated previously by Wakefield (Wakefield, 2008, 2007, 2009). The ability to specify an arbitrary prior distribution in our method is important, because a well-specified prior distribution for effect sizes in the genome makes the posterior effect size estimates resilient to bias due to selection of top-ranking results. For example, one can take a SNP with the smallest P-value in a GWAS and compute the posterior probability of H_0. In certain cases, notably in models with correctly specified and non-hierarchical priors, this posterior probability, as well as an estimate of the effect size, is unaffected by multiple testing or by selection of top ranking results (Senn, 2008; Dawid, 1994; Efron, 2011, 2010; Berry and Hochberg, 1999; Scott et al., 2010). In contrast, the usual estimates, e.g., the largest odds ratios observed in a study, tend to be over-estimating the actual values – a phenomenon known as the winner's curse. Comparing to P-value adjustments in multiple testing, this advantage of the Bayesian approach is more than a simple convenience, because it is not always straightforward to define the amount of testing that a P-value needs to be adjusted for, and even a properly adjusted P-value still lacks probabilistic interpretation in terms of degree of confidence that the result is not spurious (Wasserstein and Lazar, 2016).

Robustness of posterior estimates to selection bias holds only when the effect size distribution is specified correctly (Kuo et al., 2015; Senn, 2008; Dawid, 1994; Efron, 2011; Scott et al., 2010). One concern is that the normal prior distribution assumed in ABF may not always provide sufficient flexibility. For example, one may prefer an asymmetric prior distribution, because without randomization of which allele is being tested at each SNP, the symmetry implies that susceptibility and protective effects are equally likely. Further, expecting that the bulk of truly associated SNPs in genome-wide studies carry small effect sizes, it is desirable to be able to accommodate distributions where there are occasional moderate or large effect sizes, while a sizable part of the density is closer to zero than what a normal distribution can provide. Moreover, methods have been emerging for estimation of disease-specific effect size distributions from GWAS and replication studies (Park et al., 2011, 2010). These methods allow one to utilize effect size distributions reported in a tabulated way, where each range of the effect size would be accompanied by its estimated frequency in the genome. Such empirically estimated distributions are not necessarily expected to follow any standard or symmetric form.

Via application of the FAB approach to simulated and real data, we found that FAB produces nearly unbiased effect size estimates under various forms of selection of the top-scoring results. Selection is not only expected to generate bias in classical effect estimates but this bias is difficult to correct for. While many adjustments for the winner's curse bias have been proposed, there can be no unbiased estimate for the top-ranking hits (Bowden and Dudbridge, 2009) unless external information, which takes the form of a prior distribution in Bayesian approaches, is utilized. Our method incorporates this prior information in a flexible way and shares simplicity of ABF and FPRP, requiring only summary statistics for its implementation.

2 Methods

Consider test statistics whose distribution under the alternative hypothesis is driven by a single parameter (i.e., the standardized effect size). These statistics span the majority of tests used for genetic association analyses, including normal, chi-squared, Student's t, and F densities. For instance, a standard Z-score, $Z = \sqrt{n} \times \hat{\delta}$, satisfies this property, where $\delta = \mu/\sigma$ is the standardized effect size. If the outcome is a case/control classification with a binary predictor (exposure), the test statistic can be expressed as:

$$\begin{aligned} Z &= \sqrt{n} \times \hat{\delta} = \sqrt{n} \times \frac{\ln(\widehat{\mathrm{OR}})}{\sqrt{\mathrm{Var}(\ln(\widehat{\mathrm{OR}}))}} \\ &= \sqrt{n} \times \frac{\ln(\widehat{\mathrm{OR}})}{\sqrt{\frac{1}{w}\frac{1}{\hat{p}_1(1-\hat{p}_1)} + \frac{1}{1-w}\frac{1}{\hat{p}_2(1-\hat{p}_2)}}} \\ &= \sqrt{n} \times \frac{\hat{\mu}}{\hat{\sigma}}, \end{aligned} \tag{1}$$

where w is the proportion of cases in the sample, $\hat{p}_1$ is the frequency of exposure among cases and $\hat{p}_2$ is that among the controls. For these statistics, deviations from the null hypothesis can be expressed by a single parameter γ. For example, under the alternative hypothesis, the standard score and the statistic in Eq. (1) will be distributed as $Z \sim N(\gamma, 1)$, where $\gamma = \sqrt{n} \times \delta$ is the non-centrality parameter.

Here, we propose new Flexible Approximate Bayes (FAB) approach for calculating the posterior effect size distribution for a SNP based on the observed data as summarized by a P-value or an

association test statistic. FAB is developed as a univariate method and currently does not allow simultaneous evaluation of effect sizes for several SNPs. The approach is "flexible" because it requires only a specification of an arbitrary prior distribution on the parameter of interest (e.g., ln(OR)), which is not limited to parametric families and can be empirically derived as a table of plausible values of the effect sizes and the respective probabilities of their occurrence. FAB does not require access to raw data, and instead, it accommodates common test statistics as summaries of effect size, including χ^2-, F-, t-, and Z-statistics. Posterior distribution for the parameter of interest is obtained using the following steps: (1) specify the effect size distribution for the parameter (μ) as a binned histogram of values $(\mu_1, \mu_2, \ldots, \mu_B)$ with corresponding bar heights $(\Pr(\mu_1), \Pr(\mu_2), \ldots, \Pr(\mu_B))$ and transform it to the prior distribution for the non-centrality parameter by plugging in an estimate of the standard error of μ, i.e., $\hat{\gamma}_i = \sqrt{n}\ \mu_i/\hat{\sigma},\ i = 1, \ldots, B$; (2) using the Bayes theorem, calculate the posterior non-centrality distribution given the association test statistic as the summary of data, $\Pr(\gamma_i \mid \text{P-value})$, and transform it to the approximate posterior distribution of the parameter of interest as $\mu_i = \gamma_i \cdot \hat{\sigma}/\sqrt{n}$. In the next two sections, we detail the rationale behind these steps.

2.1 Prior Choice

To implement FAB, we start by supplying some prior (external) information regarding possible values of the effect size for μ=ln(OR), with their respective probabilities. For example, one may consider testing whether a SNP has an effect size that is at most $|\mu^0|$ in magnitude and define the null hypothesis $H_0 : \mu \leq |\mu^0|$ with the alternative hypothesis specified as $H_A : \mu > |\mu^0|$. Then, the proportion of SNPs with effect sizes smaller than $|\mu^0|$ can be regarded as the prior probability of H_0, $\pi = \Pr(H_0)$, and $\Pr(H_A) = 1 - \pi$. This specification makes a sharp distinction between effect sizes that are large enough to be considered genuine and correspond to H_A, and a set of smaller effect sizes that correspond to H_0, and can be represented by a mixture of two distributions with the mixture proportion π. In our approach, continuous prior distributions are approximated by discretization. For example, assuming the bin size being equal to s, and a continuous distribution with the cumulative function (CDF) $F(x)$, one can obtain a table of discretized values (for positive values of x) by using values $s/2, s + s/2, 2s + s/2, ...$ with the respective weights given by $F(s) - F(0), F(2s) - F(s), ...$, stopping when the CDF value becomes sufficiently close to zero. The negative

part of the discretized distribution is obtained equivalently, and finally, the binned probabilities are normalized to add up to one. Similarly to Wakefield (2007), we can assume the prior $\mu \sim N(0, W)$, where μ is the log odds ratio. To choose W, we may specify a range of odds ratio values that we believe to be plausible, e.g., if we believe that there is a 95% chance that ORs lie between 1/2 and 2, then the standard deviation of the prior would be $\sqrt{W} = \ln(2)/1.96$. For squared values of effect sizes, e.g., $\ln^2(\text{OR})$, such distribution may be L-shaped, with a sizable spike around zero that reflects a preponderance of signals carrying small effects. In terms of genome wide association scans, such distribution reflects prior knowledge that a randomly chosen SNP has a small effect size with high probability.

For non-parametric prior densities, the prior information for μ should be supplied in a tabulated manner outlining possible values with the corresponding frequencies (see our analysis of lung cancer data for a detailed discussion of this scenario). Under these models, instead of specifying priors for the dichotomous scenario as π and $1 - \pi$, we would partition the prior distribution of μ into a histogram with a finite set of bin values $\mu_1, \mu_2, \ldots, \mu_B$ and the corresponding prior probabilities, $\Pr(\mu_i)$. The choice for the number of bins B is explored in our simulation results section. Finally, to obtain the prior distribution for the non-centrality parameter under the alternative hypothesis, we would divide bin values by the estimated standard error of the effect size, $\hat{\gamma}_i = \sqrt{n} \cdot \mu_i / \hat{\sigma}$, and leave the corresponding bar heights unchanged, i.e., $\Pr(\mu_i) = \Pr(\gamma_i), i = 1, \ldots, B$.

Notice subtle implications of specifying the prior for μ in a discrete fashion via a histogram. On the one hand, it allows FAB to be flexible and not to be restricted to a parametric prior distributions for the effect size. For instance, FAB is not restricted to the typical default choice of the Normal distribution for the causal SNP effect size prior (Wakefield, 2007), Laplace (Walters et al., 2019) or Normal-gamma (Alenazi et al., 2019) prior. On the other hand, FAB includes the parametric scenario as a special case. For example, if one assumes a normal prior density for the log odds ratio associated with exposure (e.g., 1 or 2 copies of the mutant allele), $\mu \sim N(0, W)$, and discretizes the bell-shaped density into a histogram with sufficient number of bins, the results of FAB and AFB approaches are going to be equal (within rounding errors).

2.2 Posterior density via FAB

The application of Bayes's theorem gives the posterior probability for the non-centrality parameter γ_i given data as summarized by a P-value as:

$$\Pr(\gamma_i \mid \text{P-value}) = \frac{\Pr(\mu_i) f(Z = z \mid \hat{\gamma}_i)}{\sum_{j=1}^{B} \Pr(\mu_j) f(Z = z \mid \hat{\gamma}_j)}, \tag{2}$$

where $f(\cdot)$ is the test statistic density with the non-centrality parameter $\hat{\gamma}_i = \sqrt{n}\ \mu_i/\hat{\sigma},\ i = 1, \ldots, B$, and z is the observed value of the test statistic Z, obtained from the one-sided P-value, p, by applying the inverse normal CDF, $\Phi^{-1}(\cdot)$, transformation: $z = \Phi^{-1}(1-p)$. Two-sided P-value can be utilized in two ways. Using the one-degree of freedom chi-square test, as an example, (e.g. squared values of Z in Eq. A1 of Appendix) two-sided P-values can be transformed to Z via the inverse chi-square CDF, $\Psi^{-1}(\cdot)$, as $\text{sign}(\ln(\text{OR}))\sqrt{\Psi^{-1}(1-p)}$. Alternatively, P-value can be converted to the chi-square statistic as $X^2 = \Psi^{-1}(1-p)$ and the prior distribution, such as a squared normal or a gamma distribution, can be placed on $\log(\text{OR})^2$ directly. In the later case, $f(\cdot)$ in Eq. 2 becomes a non-central chi-square density. The next steps in obtaining the posterior distribution and the posterior estimate for log(OR) are illustrated by Eqs. A2 and A3 (Supplementary Appendix).

Once all posterior probabilities corresponding to each value of γ_i are obtained, the values $\delta_i = \gamma_i/\sqrt{n}$ form a binned histogram of the posterior distribution for the standardized effect size. With this posterior distribution, one can construct $(1-\alpha)\%$ credible intervals, make probabilistic statements, and estimate the standardized effect size as the posterior mean:

$$E(\delta \mid \text{P-value}) = \sum_{i=1}^{B} \delta_i\ \Pr(\delta_i \mid \text{P-value}). \tag{3}$$

Next, posterior distribution for the standardized parameter can be converted to an approximate posterior distribution of the parameter itself, by plugging in the sample standard deviation, e.g., $\mu = \delta \cdot \hat{\sigma}$. With our focus on odds ratio, disease risk p is related to exposure and other variables via logistic model, $\text{logit}(p) = \mathbf{x}'\boldsymbol{\beta} + v\mu$, where $\mathbf{x}'$ and $\boldsymbol{\beta}$ are vectors of covariates and their respective coefficients, and μ is ln(OR) with the corresponding exposure value, v. Following Wakefield (2007), we consider a prior distribution for μ only, rather than a joint prior on all parameters.

3 Results

3.1 Simulation study and methods comparisons

We first assessed the performance of the proposed method through a simulation study, with the following objectives: (1) to show that the approximate posterior distribution for ln(OR) can be accurately extracted from P-value, and (2) to show that the resulting posterior distribution is robust to selection bias in large scale experiments. The FAB approach allows one to obtain effectively exact posterior distribution for the standardized effect size (in a model without covariates) and an approximated posterior distribution for the effect size itself.

In a special case of the conjugate normal model with zero-centered normal prior, FAB is expected to approach ABF as the number of bins of the discretized normal prior increases. Although in this case we expect posterior estimates obtained by the two methods to be the same, it is of interest to confirm their resistance to the winner's curse. Further, it is critical for a posterior approximation to be robust to various forms of selection. For example, we want to make sure that the effect size is not going to be overestimated if a predictor with the maximum observed OR or the minimum P-value is selected from a multiple testing experiment. One of the reviewers of this paper pointed out that precise unbiasedness is not really an issue in practice, and as long as the order of bias magnitude is low, one can still use biased results for accurate power calculations in planning a replication. Here, we show that when prior information is specified correctly, posterior estimates are nearly unaffected by bias due to selection. The key advantage of the method compared to P-values is that posterior estimates (e.g., of odds ratios) and probabilities of hypotheses show robustness to bias due to multiple testing.

To generate data under the normal prior model (where ABF and FAB are expected to be equivalent), we assumed that among 20,000 SNPs there would be about one with an odds ratio over two, and from that we determined the prior variance $W = 0.03174$. Thus, the prior distribution for the effect size was $\ln(\text{OR}) \sim N(0, W)$. For the implementation of the FAB approach purposes, we discretized this distribution to 3,387 bins (this value is obtained by truncating the prior normal distribution at its 10^{-6} and $1 - 10^{-6}$ quantiles and using the bins of length 5×10^{-4}). Further details of data generation step are given in Appendix (Simulation Study Setup section).

For each simulation, we performed a total of L tests and selected top-ranking SNPs based on

one of the "selection rules," such as P-value threshold. After calculating the estimated $\ln(\widehat{\text{OR}})$ and the posterior expectations for the selected SNP, we averaged the results across simulations and obtained (1) the expected values of the true log odds ratio that gave rise to the selected test statistic, $E(\ln(\text{OR}))$, (2) the corresponding average of posterior expectations, and (3) the averaged frequentist estimate for the log of odds ratio, $E(\ln(\widehat{\text{OR}}))$.

Table 1 presents the results of SNP selection based on either the minimum P-value or the maximum ln(OR) absolute value. These types of selection are expected to generate bias in classical effect estimates (the winner's curse phenomenon), although some methods were developed to alleviate this issue (e.g., Zhong and Prentice (2008)). As Table 1 clearly indicates, ABF and FAB posterior estimates are almost identical and are unbiased with respect to the true expected value $E(\ln(\text{OR}))$, computed as the average of the actual values of ln(OR) across simulations. This average is not equal to zero (the mean of the prior distribution) due to selection of top-ranking results, as reflected by the first column of Table 1. Further, the last column of Table 1 indicates that the simple average over the maximum of frequentist estimates of ln(ORs), $E(\ln(\widehat{\text{OR}}))$, is highly biased for small sample sizes ($n = 500$) and moderately biased for larger sample sample sizes ($n = 5000$). Thus, relative to the classical approach, Bayesian approaches require substantially smaller sample sizes to produce estimates of the effect size of the top-ranked SNPs that are close to the true values.

A different type of selection is when SNPs are selected based on a P-value magnitude. For example, "P ~ 0.05" in the first column of the Table 2 denotes the type of selection where P-values that fall in a narrow interval were retained, that is, $0.025 < P < 0.075$. This approximates the scenario when a researcher has obtained a particular P-value (e.g., 0.05) and the question is, what bias would be expected in the estimates of ln(OR) over repeated experiments that yield a similar P-value.

The second part of the table shows a thresholding type of selection, that is, selection of SNPs with P-values that are smaller than a given threshold. On the one hand, average of frequentist estimates of ln(OR) across simulations, $E(\ln(\widehat{\text{OR}}))$, shows bias due to selection – a consequence of the winner's curse. On the other hand, ABF and FAB approaches produce unbiased values that are expectedly very similar to one another. No correction for multiple testing was done to either the ABF or the FAB method, nor it is needed, because the prior distribution already properly accommodates the high frequency of SNPs carrying very small values of ln(OR).

Based on posterior distribution of ln(OR), one can also obtain credible interval estimates, as provided in Table 3. It is interesting to compare posterior intervals to the classical confidence intervals (CI). When viewed as Bayesian posterior intervals, CI's incorporate the implicit assumption that the variance of the prior distribution is so large that any value of ln(OR) is equally likely. The influence of this assumption on the interval endpoints decreases with the sample size, n, and for that reason it may be incorrectly assumed that genetic studies with thousands of observations would yield intervals which endpoints can be interpreted as probabilistic bounds for ln(OR). However, due to sparsity of SNPs carrying large effect sizes as well as to selection of top-ranking results, CIs can remain substantially biased even with large n. In contrast, coverage of the ABF and the FAB methods is around the declared 80%. The type of selection shown in the table is P-value-based. Similar results were obtained for selections based on large values of odds ratios instead of P-values (results not shown).

One of the features of the proposed FAB approach is its flexibility with regard to what shape the prior distribution can take. To illustrate this feature, we used an asymmetric prior distribution shown in Figure 1 to obtain simulation results for the top SNPs selected based on the minimum P-value in experiments with L =1,000, 50,000, and 10^5 tests. The results are summarized in Table 4 and show that posterior expectations calculated with the FAB approach closely match the true average values of the effect size, while estimated values from the logistic regression overestimate the effect size due to the winner's curse. Posterior expectations based on ABF in Table 4 with the variance of the normal prior, $N(0, W)$, given by the variance of the non-normal prior distribution also tend to overestimate the effect size due to the incorrect prior specification but the bias is smaller in magnitude than that of the frequentist estimate.

Next, we checked the spread of FAB posterior expectations relative to the true value of ln(OR), given the asymmetric prior. Figure 2 shows contour plots of the FAB posterior estimates for ln(OR) against the actual values of ln(OR). The classical estimates, $\ln(\widehat{\mathrm{OR}})$ from the equation for the Z-statistic were also included. Each point on the graphs was selected based a SNP with the largest ln(OR) out of 10,000 tests. As expected, compared to the FAB method, the classical estimate exhibit the winner's curse, i.e., upward bias, as well as a lower correlation with the true value of $\ln(\widehat{\mathrm{OR}})$ (i.e., plots for the classical estimates appear higher on the graphs and they are less elliptical than those for the FAB estimates). The two types of estimates become more similar as sample size

increases.

Finally, we checked the effect of the number of bins B on the posterior mean, median and coverage estimation. Data for Table 5 have been simulated in the same way as for the first row of Table 1, that is, assuming the normal prior distribution for ln(OR), $N(0, W)$ and selection of the smallest P-value from $1,000$ Z-tests using the sample size of 500. Table 6 assumed a two-sided test instead, based on the P-value obtained from the squared Z-statistic. Thus, the statistic now is asymptotically a one degree of freedom chi-square and the posterior computation is carried out using the non-central chi-square (rather than normal) density. The prior distribution in this case was obtained by squaring the $N(0, W)$ prior, resulting in the prior distribution for $\ln(\text{OR})^2$ being a scaled chi-square, $W \times \chi^2_{(1)}$. Tables 5 and 6 summarize the results and indicate that $B > 100$ provides satisfactory coverage for posterior intervals based on a Z-statistic (Table 5), and $B > 150$ provides satisfactory coverage for posterior intervals based on the χ^2-statistic (Table 6). The point posterior estimates are more robust than the interval estimates to mis-specification of the continuous prior distribution due to discretization, and remain close to the true values when the number of bins is as low as 50.

In summary, our results reassure the validity of the proposed approximate Bayesian approach, where the prior distribution is specified only for the parameter of interest. For example, a potential concern could have been that the odds ratio alone is not sufficient to describe a population in terms of parameters that affect disease risk. Risk of disease given allele, $\Pr(D|A)$, and allele frequency, $\Pr(A)$, are the two other parameters that also have respective distributions across SNPs. Thus, "simulated reality" in our experiments was derived from three random distributions rather than from a single one for ln(OR). In contrast, during the analysis of simulated data by our method, a prior distribution is specified for the ln(OR) only. The three random parameters (ln(OR), $\Pr(D|A)$, and $\Pr(A)$) together determine prevalence of disease due to SNP as well as allele frequencies among affected and unaffected individuals. Furthermore, standard deviation for ln(OR) is also random, as a function of these parameters. Therefore, $\Pr(D|A)$ and $\Pr(A)$ would have to be part of a prior distribution in a fully Bayesian model. Nevertheless, our simulation studies confirm validity of the proposed method as judged by nearly unbiased point and interval posterior estimates in the presence of potentially bias-inducing selection of top-scoring P-values or odd ratios.

3.2 Lung cancer association study

We applied the FAB approach in re-analysis of four rare candidate variants reported by Jin et al (Jin et al., 2015) in a study of low-frequency genetic associations with lung cancer. To calculate posterior effect sizes for candidate susceptibility variants, we used their reported P-values from the genome-wide discovery stage (Jin et al., 2015) and prior effect size distribution for cancers, as reported by Park et al (Park et al., 2010). Park et al. provided twenty ORs with the corresponding estimated number of genetic variants carrying these effect sizes, which we used as prior frequencies of encountering such effect sizes in the genome. However, the direct use of twenty reported ORs leads to a limiting assumption that there are only twenty distinct effect sizes. Moreover, it puts strict bounds on the minimum and the maximum possible effect size values (the minimum log(OR) value reported by Park et al was -0.19 (with 0.006 frequency) and the maximum was 0.23 (with 0.004 frequency)). Therefore, we smoothed a twenty-bin histogram and obtained a continuous prior density with ∼80% of observations falling within -0.19 - 0.23 interval. We further discretized this continuous distribution using 3,000 bins and employed the resulting prior for the analysis. The exact details of prior construction, as well as R script for replication analysis are available on the web resource that accompanies this article.

Genetic variant with the smallest discovery P-value reported by Jin et al (Jin et al., 2015), was rs200847762 within the *FKBPL* gene (OR = 0.21, P = 1.84×10^{-6}), Table 7. This variant replicated in first of the two replication studies (OR = 0.19, 95% Confidence Interval: (0.08 - 0.46)). In the second replication study, the estimated OR was consistent with the discovered effect direction, 0.66, but the confidence interval included one (0.34 - 1.28). Application of the FAB approach produced the posterior estimate of OR for the discovery study equal to 0.57, with the 95% Credible Interval (0.33 - 0.92), suggesting a genuine causal signal within this region (we also report posterior estimates for the median of OR, which are close in values to the estimates of the posterior mean). This signal was further suggested to be in association with breast cancer among Chinese women by Zhou et al (Zhou et al., 2017). The therapeutic and diagnostic potential of *FKBPL* to targeting tumours was discussed by Robson and James (Robson and James, 2012) and McKeen et al (McKeen et al., 2011). Genetic variant with the second smallest P-value in Jin et al. discovery study was rs2298090 within the *HIST1H1E* gene (P = 6.16×10^{-5}). This signal also

failed to replicate at the second stage of their study, OR = 0.67, 95% CI (0.44 - 1.02). The FAB posterior mean estimate was OR = 0.66 with the 95% interval (0.42 - 1.01). Finally, rs9469031 within the *BAT2* gene (also known as *PRRC1A*) and rs6141383 within the *BPIFB1* gene (both discovery P-values $\approx 10^{-4}$) were replicated in the second stage of the original study by Jin et al (Jin et al., 2015). The FAB estimates suggest lower effect size magnitudes than what is reported in the discovery stage. Specifically, for rs9469031 candidate SNP, OR = 0.68, with 95% posterior interval (0.49 - 0.94); for rs6141383, OR = 1.42, with 95% posterior interval (0.99 - 2.01). The posterior OR estimates tend to be closer to the replication values than to the possibly inflated values found during the discovery stage. Note that in all cases posterior intervals are wider than the 95% confidence intervals for the discovery stage, because Bayesian intervals account for sparsity of real associations via the prior distribution specified for OR and thus do not need adjustment for multiple testing.

4 Discussion

Ongoing "replicability crisis" put P-values at the center of controversy. Although P-values are often misused and misinterpreted, they are routinely available and reported. In this article, we suggested a new method to compute an approximate posterior distribution for ln(OR), in part based on P-value or the usual test statistic for testing significance of odds ratio. This posterior distribution can be used in versatile ways. One can estimate posterior probability of genetic association with disease, obtain point estimates and test interval hypotheses about probabilistic ranges for the effect size. The method is unaffected by the winner's curse and by multiple testing, provided one can realistically describe the effect size distribution across SNPs in the genome.

Our method is "approximate" in reference to its Bayesian part, because the prior distribution is specified only for effect size, the parameter of interest. Wakefield's ABF (Wakefield, 2007) is approximate in the same sense as our method. In other words, approximation is a compromise between Bayesian and classical approaches. In a fully Bayesian model, complete joint prior distribution for all parameters (including coefficients for covariates) should be specified, leading to increased computational complexity, and to difficulties in specification of prior distributions with realistic dependencies between parameters. The approximate Bayesian methods enable one to uti-

lize our basic understanding about the distribution of effect size (such as relative risk or odds ratio) across genetic variants in the genome and to construct posterior estimates from simple and readily available summary statistics. Through simulation studies, we confirmed effectiveness of this "compromise" strategy, where we place a prior distribution on the effect size only and demonstrated that the proposed method is resistant to potential bias due to selection of top-ranking P-values and odds ratios and that it gives accurate point and interval estimates of true effect size.

The proposed FAB approach is general in the sense that it can be used with any prior effect size distribution and with statistics other than the normal Z-score for ln(OR). Some examples are given in the link to our software below, including a simulation R script for testing validity of posterior inference for the mean difference between two groups, using the two-sample T-test statistic as a summary of data. Parameters of the prior distribution in our method, e.g., prior variance for the distribution of log(OR), do not need to be specified directly, because the prior distribution can be handcrafted as a table of values with their respective frequencies. This can be done with relative ease as methods emerge for estimating effect size distribution from genome-wide association data (Chen et al., 2013; Park et al., 2010). Such empirically estimated distributions are not necessarily expected to follow any symmetric or standard form. Thus, more flexibility is needed, and the FAB method allows researchers to incorporate empirically estimated distributions directly, without having to choose a prior from one of the standard distributions. That being said, any standard distribution can be easily discretized and used as a prior in applications of our method. Another obvious application is re-evaluation of published findings. Access to individual records may be difficult, but summary statistics needed for approximate posterior inference are more available. Posterior distributions for standardized effect sizes are particularly simple and expected to be accurate when such distribution is derived from a test statistic whose non-centrality is a function of standardized effect size.

Web Resources

The URL for software referenced in this article is available at:

`https://github.com/dmitri-zaykin/Mix_Bayes`

Acknowledgements

This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Conflict of Interest

We have no conflicts of interest to declare.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

Alenazi, A. A., Cox, A., Juarez, M., Lin, W.-Y., and Walters, K. (2019). Bayesian variable selection using partially observed categorical prior information in fine-mapping association studies. Genetic epidemiology, 43(6):690–703.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300.

Berry, D. A. and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. Journal of Statistical Planning and Inference, 82(1-2):215–227.

Bowden, J. and Dudbridge, F. (2009). Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. Genet Epidemiol, 33(5):406–418.

Chen, D., Jiang, X., Akula, N., Shugart, Y., Wendland, J., Steele, C., Kassem, L., Park, J., Chatterjee, N., Jamain, S., et al. (2013). Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. Mol. Psychiatry, 18(2):195–205.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect. Psychol. Sci., 3(4):286–300.

Dawid, A. P. (1994). Selection paradoxes of Bayesian inference. Lecture Notes-Monograph Series, pages 211–220.

Dudbridge, F., Gusnanto, A., and Koeleman, B. P. (2006). Detecting multiple associations in genome-wide studies. Human Genomics, 2(5):310.

Efron, B. (2010). The future of indirect evidence. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(2):145.

Efron, B. (2011). Tweedie's formula and selection bias. Journal of the American Statistical Association, 106(496):1602–1614.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. J Am Statist Assoc, 96(456):1151–1160.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol., pages 1–14.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle P value generates irreproducible results. Nat. Methods., 12(3):179–185.

Jin, G., Zhu, M., Yin, R., Shen, W., Liu, J., Sun, J., Wang, C., Dai, J., Ma, H., Wu, C., et al. (2015). Low-frequency coding variants at 6p21.33 and 20q11.21 are associated with lung cancer risk in Chinese populations. Am. J. Hum. Genet., 96(5):832–840.

Johnson, V. E. (2013). Revised standards for statistical evidence. Proc. Natl. Acad. Sci., 110(48):19313–19317.

Kuo, C.-L., Vsevolozhskaya, O. A., and Zaykin, D. V. (2015). Assessing the probability that a finding is genuine for large-scale genetic association studies. PLOS ONE, 10(5):e0124107.

Lazzeroni, L., Lu, Y., and Belitskaya-Levy, I. (2014). P-values in genomics: apparent precision masks high uncertainty. Mol. Psychiatry, 19(12):1336–1340.

Lazzeroni, L. C., Lu, Y., and Belitskaya-Lévy, I. (2016). Solutions for quantifying P-value uncertainty and replication power. Nat. Methods, 13(2):107–108.

Lin, D. and Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 34(1):60–66.

Lucke, J. (2009). A critique of the false-positive report probability. Genet Epidemiol, 33(2):145–150.

McKeen, H. D., Brennan, D. J., Hegarty, S., Lanigan, F., Jirström, K., Byrne, C., Yakkundi, A., McCarthy, H. O., Gallagher, W. M., and Robson, T. (2011). The emerging role of FK506-binding proteins as cancer biomarkers: a focus on FKBPL. In Meeting on Signalling and Human Disease, volume 39, pages 663–668. Portland Press.

Park, J.-H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., Chanock, S. J., Fraumeni, J. F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proceedings of the National Academy of Sciences, 108(44):18026–18031.

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet., 42(7):570–575.

Robson, T. and James, I. F. (2012). The therapeutic and diagnostic potential of FKBPL; a novel anticancer protein. Drug Discovery Today, 17(11):544–548.

Scott, J. G., Berger, J. O., et al. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics, 38(5):2587–2619.

Senn, S. (2008). A note concerning a selection "paradox" of Dawid's. The American Statistician, 62(3):206–210.

Spencer, A. V., Cox, A., Lin, W.-Y., Easton, D. F., Michailidou, K., and Walters, K. (2015). Novel Bayes factors that capture expert uncertainty in prior density specification in genetic association studies. Genet Epidemiol, 39(4):239–248.

Spencer, A. V., Cox, A., Lin, W.-Y., Easton, D. F., Michailidou, K., and Walters, K. (2016). Incorporating functional genomic information in genetic association studies using an empirical Bayes approach. Genetic epidemiology, 40(3):176–187.

Storey, J. (2002). A direct approach to false discovery rates. J Royal Stat Soc B, 64(3):479–498.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Canc Inst, 96(6):434–442.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am. J. Hum. Genet., 81(2):208–227.

Wakefield, J. (2008). Reporting and interpretation in genome-wide association studies. Int J Epidemiol, 37(3):641–53.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol, 33(1):79–86.

Walters, K., Cox, A., and Yaacob, H. (2019). Using GWAS top hits to inform priors in Bayesian fine-mapping association studies. Genetic epidemiology, 43(6):675–689.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. Am. Stat., 70(2):129–133.

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. Journal of evolutionary biology, 24(8):1836–1841.

Zaykin, D. V., Young, S. S., and Westfall, P. H. (2000). Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller et al. Genetics, 154(4):1917–1918.

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics, 9(4):621–634.

Zhou, W., Jiang, Y., Zhu, M., Hang, D., Chen, J., Zhou, J., Dai, J., Ma, H., Hu, Z., Jin, G., et al. (2017). Low-frequency nonsynonymous variants in FKBPL and ARPC1B genes are associated with breast cancer risk in Chinese women. Molecular carcinogenesis, 56(2):774–780.

Tables

Table 1: **Summary of $\ln(\mathbf{OR})$ expectations out of L tests**

Number of tests, L	n	True value	Posterior estimate		Frequentist estimate
		$E(\ln(\mathrm{OR}))$	ABF	FAB	$E(\ln(\widehat{\mathrm{OR}}))$
Select min P					
$1,000$	500	0.38	0.38	0.38	0.85
	1,000	0.45	0.45	0.44	0.71
	5,000	0.53	0.53	0.53	0.59
1×10^5	500	0.53	0.51	0.51	1.12
	1,000	0.61	0.60	0.60	0.95
	5,000	0.72	0.72	0.71	0.90
Select max OR					
$1,000$	500	0.33	0.31	0.31	0.95
	1,000	0.42	0.41	0.41	0.77
	5,000	0.53	0.54	0.54	0.62
1×10^5	500	0.37	0.32	0.32	1.40
	1,000	0.51	0.48	0.48	1.09
	5,000	0.73	0.73	0.71	0.84

Table 2: **Expectations of $\ln(\mathbf{OR})$ under different types of P-value selection**

Type of selection	n	True value	Posterior estimate		Frequentist estimate
		$E(\ln(\mathrm{OR}))$	ABF	FAB	$E(\ln(\widehat{\mathrm{OR}}))$
$\mathrm{P} \sim 0.05$	1,000	0.17	0.17	0.17	0.29
$\mathrm{P} \sim 0.001$	1,000	0.29	0.29	0.29	0.47
$\mathrm{P} \sim 1 \times 10^{-5}$	1,000	0.38	0.38	0.38	0.62
$\mathrm{P} \sim 1 \times 10^{-7}$	1,000	0.46	0.46	0.46	0.73
$\mathrm{P} < 0.05$	1,000	0.24	0.24	0.24	0.39
$\mathrm{P} < 0.001$	1,000	0.34	0.33	0.33	0.54
$\mathrm{P} < 1 \times 10^{-5}$	1,000	0.42	0.42	0.42	0.68
$\mathrm{P} < 1 \times 10^{-7}$	1,000	0.50	0.49	0.49	0.78

Table 3: **Empirical coverage of the two credible intervals and a confidence interval at the 80% nominal level**

Type of Selection	n	80% Empirical Coverage		
		ABF	FAB	CI
P ~ 0.05	1,000	79.90%	79.82%	72.70%
P ~ 0.001	1,000	80.39%	80.33%	50.13%
P $\sim 1 \times 10^{-7}$	1,000	80.29%	79.94%	20.60%
P < 0.05	1,000	79.85%	79.78%	60.12%
P < 0.001	1,000	80.22%	80.11%	41.13%
P $< 1 \times 10^{-7}$	1,000	80.37%	79.40%	16.56%
min P, 1K tests	500	80.43%	80.12%	7.91%
	5,000	79.82%	78.99%	60.43%
min P, 10K tests	500	80.27%	79.26%	2.38%
	5,000	80.76%	79.98%	55.28%

Table 4: **Summary of $\ln(\mathbf{OR})$ expectations for the minimum P-value out of L tests with a non-conjugate prior distribution**

Number of tests, L	n	$E(\ln(\text{OR}))$	FAB	ABF	$E(\ln(\widehat{\text{OR}}))$
$1,000$	500	0.67	0.67	0.92	1.08
	1,000	0.74	0.73	0.89	0.96
	5,000	0.81	0.81	0.85	0.86
$50,000$	500	0.82	0.81	1.15	1.34
	1,000	0.91	0.90	1.11	1.20
	5,000	1.00	1.00	1.05	1.07
1×10^5	500	0.84	0.83	1.19	1.39
	1,000	0.92	0.92	1.14	1.23
	5,000	1.02	1.02	1.08	1.09

Table 5: **Point and 80%-interval estimates for log(OR) as a function of the number of bins.**

Number of bins	log(OR), posterior median	log(OR), posterior mean	Coverage
3000	0.37	0.38	0.80
2139	0.37	0.38	0.80
1596	0.37	0.38	0.80
1191	0.37	0.38	0.80
889	0.37	0.38	0.80
663	0.37	0.37	0.80
495	0.37	0.38	0.80
369	0.37	0.37	0.80
275	0.37	0.37	0.80
205	0.37	0.37	0.80
153	0.37	0.37	0.80
114	0.37	0.37	0.79
85	0.37	0.37	0.79
63	0.37	0.37	0.78
47	0.36	0.37	0.78
35	0.36	0.36	0.75
26	0.36	0.36	0.73
19	0.35	0.35	0.74
14	0.35	0.35	0.72
11	0.34	0.35	0.68
8	0.34	0.34	0.55
6	0.33	0.33	0.61
4	0.23	0.23	0
3	0.13	0.13	0

True mean value of log(OR) across simulations is 0.38.
True median value of log(OR) across simulations is 0.38.
Frequentist estimate, averaged across simulations, is 0.85.

Table 6: **Point and 80%-interval estimates for log(OR)^2 as a function of the number of bins.**

Number of bins	log(OR), posterior median	log(OR), posterior mean	Coverage
3000	0.17	0.18	0.80
2310	0.18	0.18	0.80
1724	0.17	0.18	0.80
1286	0.17	0.18	0.80
960	0.17	0.18	0.79
716	0.17	0.18	0.79
534	0.17	0.18	0.79
399	0.17	0.18	0.79
297	0.17	0.18	0.79
222	0.17	0.18	0.79
165	0.17	0.18	0.79
123	0.17	0.18	0.78
92	0.17	0.18	0.78
68	0.17	0.18	0.78
51	0.17	0.18	0.78
38	0.17	0.18	0.77
28	0.17	0.18	0.76
21	0.17	0.18	0.75
15	0.17	0.18	0.73
11	0.17	0.18	0.69
8	0.17	0.18	0.64
6	0.18	0.18	0.58
4	0.18	0.19	0.41
3	0.20	0.20	0.26

True mean value of log(OR) across simulations is 0.18.
True median value of log(OR) across simulations is 0.17.
Frequentist estimate, averaged across simulations, is 0.81.

Table 7: **Reanalysis of variants associated with lung cancer risk**

Chr	Gene	Variant ID	Stage	Case	Con	MAF		OR (95% CI)	P-value
						Case	Con		
6p21.33	*BAT2*	rs9469031	Disc	1,341[a]	1,962[a]	0.019	0.036	0.52 (0.37-0.73)	$1.54^{a}\times10^{-4}$
			Rep I	1,114	1,245	0.024	0.042	0.61 (0.44-0.83)	–
			Rep II	3,508	3,631	0.011	0.018	0.62 (0.46-0.84)	–
			Post (mean)	–	–	–	–	0.68 (0.49-0.94)	–
			Post (med)	–	–	–	–	0.68 (0.49-0.94)	–
6p21.33	*FKBPL*	rs200847762	Disc	1,341[a]	1,982[a]	0.004	0.19	0.21 (0.11-0.39)	$1.84^{a}\times10^{-6}$
			Rep I	1,110	1,239	0.003	0.014	0.19 (0.08-0.46)	–
			Rep II	3,581	3,666	0.002	0.003	0.66 (0.34-1.28)	–
			Post (mean)	–	–	–	–	0.57 (0.33-0.92)	–
			Post (med)	–	–	–	–	0.57 (0.35-0.95)	–
6p22.2	*HIST1H1E*	rs2298090	Disc	1,341[a]	1,982[a]	0.006	0.020	0.32 (0.19-0.56)	$6.16^{a}\times10^{-5}$
			Rep I	1,101	1,235	0.013	0.024	0.56 (0.36-0.87)	–
			Rep II	3,429	3,551	0.006	0.008	0.67 (0.44-1.02)	–
			Post (mean)	–	–	–	–	0.66 (0.42-1.01)	–
			Post (med)	–	–	–	–	0.66 (0.41-1.00)	–
20q11.21	*BPIFB1*	rs6141383	Disc	1,341[a]	1,982[a]	0.025	0.012	2.00 (1.36-2.95)	$4.80^{a}\times10^{-4}$
			Rep I	1,110	1,240	0.022	0.013	1.68 (1.07-2.63)	–
			Rep II	3,507	3,455	0.019	0.013	1.64 (1.24-2.17)	–
			Post (mean)	–	–	–	–	1.42 (0.99-2.01)	–
			Post (med)	–	–	–	–	1.42 (1.00-2.03)	–

Abbreviations are as follows. Chr: chromosomal region; MAF: minor allele frequency; CI: confidence interval; Disc: discovery; Rep: replication; Post: posterior; Con: Control; med: median.

[a] Summary values from the discovery study that were used to calculate posterior expectations and 95% credible intervals.

Figure Legends

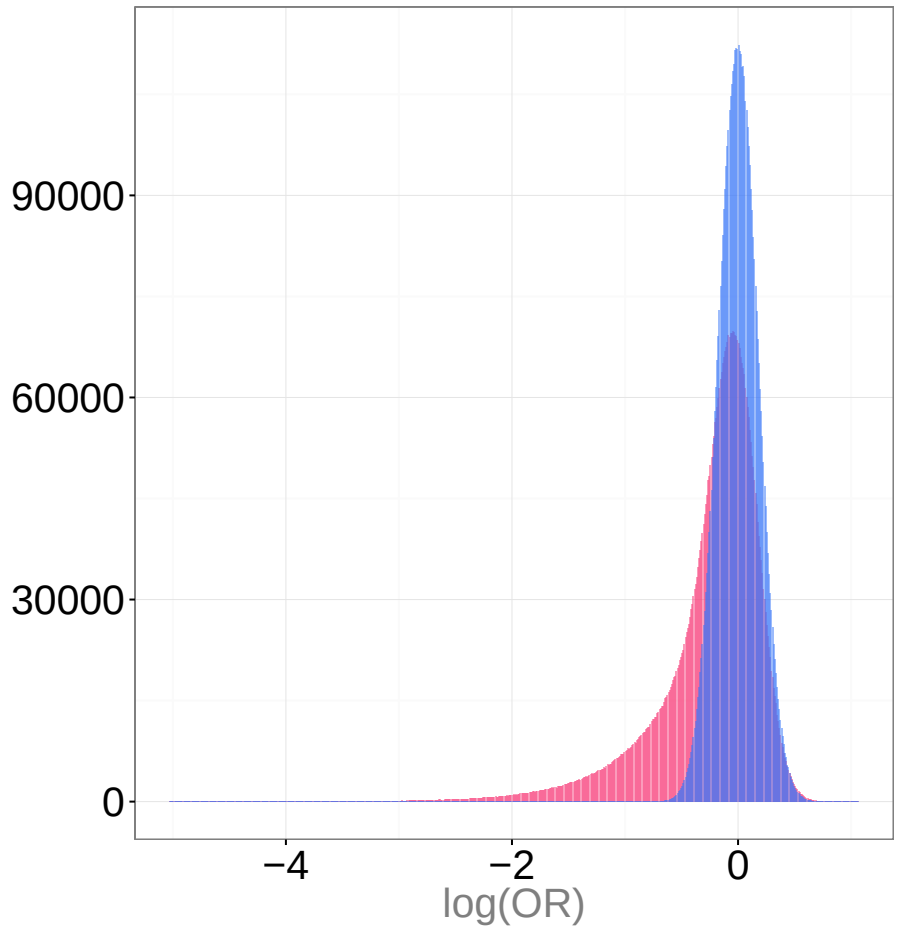

Figure 1: **Conjugate Normal and Non-conjugate Skewed Prior Distributions for log(OR).** Symmetric density is the normal prior with the variance set to approximate the occurrence of OR>2 at about 1 in 20,000 SNPs. The asymmetric density represents non-conjugate skewed prior. The graph plots the two distributions on the same scale with the overlap shown in purple.

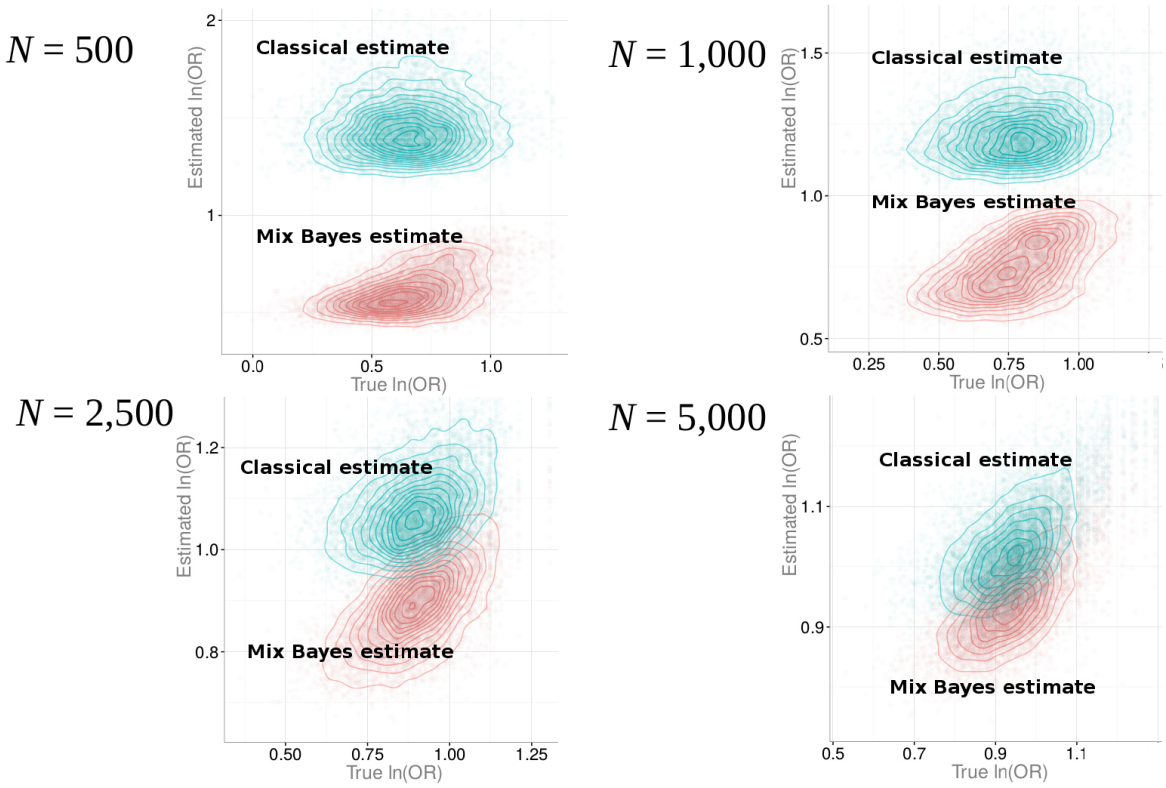

Figure 2: **Simulation scatter plots.** Each point in the graphs is an association result selected on the basis of a SNP with the minimum P-value out of 10,000 tests.

Supplementary Appendix

ABF: A special case of FAB

Although our proposed FAB approach is very general in the sense that it can deal with any arbitrary prior effect size distribution, additional insights into its behaviour may be gained by showing its equivalence to the approximate Bayes factor (Wakefield, 2007) under the conjugate normal model. Specifically, Wakefield (Wakefield, 2007) showed that if disease risk, p, is modeled via logistic regression, then assuming a zero-centered normal prior distribution for $\mu = \ln(\text{OR}) \sim N(0, W)$ and given $Z = \hat{\mu}/V$, where V is the standard error of $\hat{\mu}$, the approximate posterior distribution for OR is

$$\begin{aligned} e^{\mu} \mid \hat{\mu} &\sim \text{lognormal}(r\hat{\mu}, rV), \\ \text{where } r &= \frac{W}{W+V}. \end{aligned} \tag{1}$$

The logistic model includes the parameter of interest, $\mu = \ln(\text{OR})$, with the corresponding exposure value, v, as well as vectors of covariates and their respective coefficients, $\mathbf{x}'$ and $\boldsymbol{\beta}$:

$$\text{logit}(p) = \mathbf{x}'\boldsymbol{\beta} + v\mu, \tag{2}$$

but the prior distribution is specified for μ only, thus the posterior distribution for μ is approximate. An estimator for ln(OR) can be obtained as an approximate posterior expectation of ln(OR), which is simply $r\hat{\mu}$.

The ABF method received much attention from the genetics community, in part due to simplicity of its calculation: the Bayes factor and the approximate posterior distribution for ln(OR) are computed based on the usual Z-statistic for ln(OR). The variance parameter W of a normal prior distribution for ln(OR) in ABF reflects the spread of the ln(OR) across SNPs in the genome.

Simulation Study Setup

Given a population value of ln(OR) for a given SNP drawn from a prior distribution, one also needs to draw two other random population values, the probability of disease given a risk allele,

$\Pr(D|A)$, and the allele frequency, $\Pr(A)$. Under Hardy-Weinberg equilibrium, $\Pr(D|A)$ is defined as an average over the risks for the genotypes that contain the allele A, i.e., AA and $A\bar{A}$, and when the second allele in a genotype is also A with probability $\Pr(A)$, and $\bar{A}$ otherwise. The simulation setup can be summarized as follows:

1. For each SNP, draw ln(OR) from the prior distribution.

2. Draw population values of $\Pr(D \mid A) \sim \text{Beta}(0.9, 2)$ and $\Pr(A) \sim \text{Uniform}(0.1, 0.9)$. The Beta(0.9, 2) distribution for the absolute risk associated with the allele A is skewed toward zero. It was chosen to model predominantly weak signals. For example, with this distribution, $\Pr(D \mid A) < 10\%$ is about 25% and $\Pr(D \mid A) < 25\%$ is about 50%.

3. Other population values can now be computed as follows:

$$\Pr(D \mid \bar{A}) = \left[1 - \text{OR}(1 - \Pr(D \mid A)^{-1})\right]^{-1}$$

$$w = \Pr(D) = \Pr(D \mid A)\Pr(A) + \Pr(D \mid \bar{A})(1 - \Pr(A))$$

$$p_1 = \Pr(A \mid D) = \Pr(D \mid A)\Pr(A)/w$$

$$p_2 = (1 - \Pr(D \mid A))\Pr(A)/(1 - w)$$

4. Data generating step. Draw two binomial samples of alleles for case and control individuals (balanced design),

$$x_{11} \sim \text{Bin}(n_D, p_1), \quad x_{12} = n_D - x_{11}$$

$$x_{21} \sim \text{Bin}(n_{\bar{D}}, p_2), \quad x_{22} = n_{\bar{D}} - x_{21}$$

$$n_D = n_{\bar{D}}$$

The total sample size in this model is the number of alleles (twice the number of individuals). Under the assumed Hardy-Weinberg equilibrium, analysis of allele counts approximates analysis of genotype counts by the usual trend test (Sasieni, 1997). Next, add 0.5 to the four counts, $n_{ij} = x_{ij} + 0.5$.

	Allele	
Disease status	A	$\bar{A}$
D	$n_{11} = n_D\hat{p}_1$	$n_{12} = n_D(1-\hat{p}_1)$
$\bar{D}$	$n_{21} = n_{\bar{D}}\hat{p}_2$	$n_{12} = n_{\bar{D}}(1-\hat{p}_2)$

Based on the counts, compute Z-statistic as:

$$Z = \frac{\ln(n_{11}n_{22}n_{12}^{-1}n_{21}^{-1})}{\sqrt{n_{11}^{-1}+n_{12}^{-1}+n_{21}^{-1}+n_{22}^{-1}}} = \sqrt{n} \times \frac{\ln(\widehat{\mathrm{OR}})}{\hat{\sigma}}, \tag{A1}$$

where $n = n_{11} + n_{12} + n_{21} + n_{22}$ and

$$\hat{\sigma} = \sqrt{\frac{1}{\hat{w}}\frac{1}{\hat{p}_1(1-\hat{p}_1)} + \frac{1}{1-\hat{w}}\frac{1}{\hat{p}_2(1-\hat{p}_2)}}. \tag{A2}$$

Sample proportion of cases, $\hat{w}$, is $1/2$ by design, due to $n_D = n_{\bar{D}}$.

5. Use Equ. (2), to obtain the posterior probabilities and Equ. (3) to obtain the posterior expectation δ. The posterior estimate for ln(OR) is given by

$$\widehat{\ln(\mathrm{OR})} = \hat{\sigma}\,\delta/\sqrt{n}. \tag{A3}$$

References

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, pages 1253–1261.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.*, 81(2):208–227.