Article:
Walters, K. orcid.org/0000-0002-5718-5734, Cox, A. orcid.org/0000-0002-5138-1099 and Yaacob, H. (2019) Using GWAS top hits to inform priors in Bayesian fine-mapping association studies. Genetic Epidemiology, 43 (6). pp. 675-689. ISSN 0741-0395

https://doi.org/10.1002/gepi.22212

Using GWAS top hits to inform priors in Bayesian fine-mapping association studies

Kevin Walters[1*] Angela Cox[2], Hannuun Yaacob [1],

[1] School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

[2] Department of Oncology, Sheffield Cancer Research Centre, University of Sheffield Medical School, Sheffield, UK

[*] E-mail: k.walters@sheffield.ac.uk, Telephone: +44 (0)114 2223720

Abstract

The default causal SNP effect size prior in Bayesian fine-mapping studies is usually the Normal distribution. This choice is often based on computational convenience, rather than evidence that it is the most suitable prior distribution. The choice of prior is important because previous studies have shown considerable sensitivity of causal SNP Bayes factors to the form of the prior. In some well-studied diseases there are now considerable numbers of GWAS top hits along with estimates of the number of yet-to-be-discovered causal SNPs. We show how the effect sizes of the top hits and estimates of the number of yet-to-be-discovered causal SNPs can be used to choose between the Laplace and Normal priors, to estimate the prior parameters and to quantify the uncertainty in this estimation. The methodology can readily be applied to other priors. We show that the top hits available from breast cancer GWAS provide overwhelming support for the Laplace over the Normal prior which has important consequences for variant prioritisation. This work in this paper enables practitioners to derive more objective priors than are currently being employed and could lead to prioritisation of different variants.

Data Sharing Statement

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

Introduction

Fine-mapping is the process of assigning probabilities of causation to candidate variants identified through genome-wide association studies (GWAS). The ultimate aim is to identify the target gene whose function is altered by the variants, taking into account the functional annotation of the individual variants [Spain and Barrett, 2015]. The probabilities used are typically posterior probabilities of causal association. Calculating posterior probabilities of causal association require specification of the causal SNP effect size prior distribution in a Bayesian analysis. This paper attempts to show how GWAS top hits can be used objectively to inform both the parametric form of the effect size prior distribution and its parameters.

There are now many fine-mapping methods that require explicit specification of a prior on the effect sizes. These include single SNP approaches (Wakefield [2009], Spencer et al. [2015], Spencer et al. [2016]) and multi-SNP methods (Benner et al. [2016]; Boggis et al. [2016]; Pereira et al. [2017]; Guan and Stephens [2011]; Pickrell [2014]; Chen et al. [2015]). The effect size prior is sometimes taken to be normally distributed, $\beta \sim N(0, w)$, for some value of w. For example, $w = 0.1$ in Pickrell [2014], whilst w is partly user-supplied and partly determined by case/control numbers in both FINEMAP [Benner et al., 2016] and CAVIARBF [Chen et al., 2015]). In other approaches a scale mixture of Normals is used for the effect size prior. For example BIMBAM [Servin and Stephens, 2007] assumes a prior of the form $\beta \sim N(0, v/\tau)$, where v is a user-supplied number and τ has a Gamma distribution (meaning that τ can be integrated out of the joint distribution of (β, τ)). Other scale mixture distributions are used in Boggis et al. [2016], Pereira et al. [2017] and Hoggart et al. [2008]. In methods where the value of w is specified, the justification is usually based on subjective belief about the likely effect size distribution. In the scale mixture of Normal approaches, the scale mixture distribution is often chosen for computational convenience. It should be noted that the Laplace prior we are advocating can be represented as a scale mixture of normals if the the mixture distribution is Exponential with a specific parameter. We feel that a prior distribution based on observed GWAS top hits is a more objective choice than some of those that are currently used.

Spencer et al. [2014] considered the Normal distribution for the effect size and demonstrated that causal SNP Bayes factor ranks are highly sensitive to changes in the prior effect size variance

even in large studies. It seems reasonable to assume that these ranks will also be sensitive to the choice of effect size prior distribution. Thus it is highly desirable that putative disease-specific causal SNP effect sizes that have been identified in GWAS to date are used to inform the effect size prior distribution in Bayesian fine-mapping studies.

As of 2015, 148 genome-wide significant SNPs (top hits) had so far been identified in breast cancer [Fachal and Dunning, 2015] with a further 65 identified more recently [Michailidou et al., 2017]. We base our analysis on the 148 top hits. These top hits do not necessarily represent actual causal SNPs, but the effect size of the causal SNP tagged by a top hit is likely to be very close to the effect size of the given top hit. All the top hits so far identified have absolute log odds ratios above $\log(1.02)$. This reflects the lack of statistical power to detect causal SNPs with lower effect sizes, even for common causal SNPs. This detection threshold needs to be taken into account in any analysis to determine suitable priors. Another important consideration is the number of SNPs with smaller effect sizes that are yet-to-discovered. For example in breast cancer it is estimated that there are 1168 additional loci that harbour disease associated SNPs [Michailidou et al., 2013]. This additional information also provides information about the appropriateness of particular prior effect size distributions.

In this paper we consider how both GWAS top hits and the hypothesized number of yet-to-be-discovered SNPs (YTBDS) can be used to inform the effect size prior for use in either fine-mapping studies or indeed in GWAS. Additionally we allow for uncertainty in the number of YTBDS. We start by describing the two approaches based on different likelihoods. We then consider the coverage of confidence intervals for the parameter of the prior distribution using several approaches. We demonstrate how to choose between the priors using Bayesian model selection and show how to use all available information relating to GWAS top hits to form a prior on the effect size distribution parameters that can then be used in Bayesian fine-mapping studies. We consider two priors: the Normal and Laplace priors. The Normal distribution is widely used in Bayesian fine-mapping studies [Wakefield, 2008] but the Laplace distribution, which also crucially gives a closed form Bayes factor when used with a normal likelihood [Wakefield, 2009], provides a better fit to our breast cancer top hits. This work has important implications for the choice of prior in future Bayesian fine-mapping studies.

It is important to be clear about the difference between the (non-truncated, non-censored) prior

on the effect size that is used in a Bayesian fine-mapping study and the (truncated or censored) prior on the effect size we use in this paper, soley to estimate the prior parameters. For example, consider a Normal $N(0, w)$ distribution as a prior on the effect size distribution used in a fine-mapping study. We are interested in using the GWAS top hits to estimate w using either truncated or censored distributions for the effect sizes. We further determine whether w can justifiably be assigned a single value or whether there is sufficient uncertainty that it needs to be assigned a probability distribution. To do this we might put a prior on w, perhaps using expert elicitation techniques, use the GWAS top hits as our 'data' and hence derive a posterior distribution for w. This posterior distribution for w could then be used as a prior for w in a Bayesian fine-mapping study. If the posterior is sufficiently concentrated then the posterior mode could be used as a plug in estimate of the prior parameter.

Materials and Methods

A Normal and Laplace probability distribution are shown in Figure 1 where the variance of the Laplace and Normal random variables is fixed at 7×10^{-3}. To be clear about parameterisations, we specify the probability density function (PDF) for the Normal and Laplace random variables in Equations (1) and (2).

$$Y \sim N(0, w), \quad f_Y(y) = \frac{1}{\sqrt{2\pi w}} \exp\left(\frac{-y^2}{2w}\right) \quad w > 0,\ y \in \mathbb{R} \tag{1}$$

$$Y \sim La(\lambda), \quad f_Y(y) = \frac{\lambda}{2} \exp\left(-\lambda \mid y \mid\right) \quad \lambda > 0,\ y \in \mathbb{R} \tag{2}$$

In this paper we are concerned with choosing a suitable prior for the random variable, denoted by X, representing the effect size (log odds ratio). This prior can then be used as a more objective prior in fine-mapping studies. it is important, at this point, to distinguish between random variables X and Y. Y has a standard Laplace or Normal probability distribution whereas X may have a truncated distribution on some subset of $\mathbb{R}$ or some values of X may be censored. We will discuss approaches to using the GWAS top hits to specify this prior. As can be seen in Figure 1 the Laplace prior might be a better choice than the Normal distribution when there is a large proportion of very small odds ratios along with a small number of relatively large ones.

Maximum likelihood approaches 1 - truncated random variables

Since very small log odds ratios are still mostly undetected due to a lack of power even in GWAS conducted by large consortia, it seems reasonable to assume that the top hits come from a particular subset of the real line: $X \in (-\infty, -c] \bigcup [c, \infty)$ for some positive log odds ratio c (eg $c = \log(1.02)$). Then the PDF of X is

$$f_X(x;\theta) = \frac{f_Y(x;\theta)}{2\int_{-\infty}^{-c} f_Y(x;\theta)\,dx}\mathbb{I}(\mid x \mid > c) \tag{3}$$

where Y is a non-truncated random variable, as in Equations (1) and (2), θ is its parameter (w or λ) and $\mathbb{I}(z)$ is 1 if z is true and is zero otherwise. If we let $F(.)$ represent the cumulative distribution function (CDF), then the log-likelihood of $x_1, \ldots, x_n$ (ignoring any constants) is given by

$$l(\theta; x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \log(f_Y(x_i;\theta)) - n\log(F_Y(-c;\theta)) \tag{4}$$

The truncated Laplace maximum likelihood estimate (MLE) is $\hat{\lambda} = n/(\sum_{i=1}^{n} |x_i| - nc)$ whilst for the truncated Normal, calculating the MLE requires numerical approaches (eg the `optimize` function in `R`).

Maximum likelihood approaches 2 - censored random variables

It may be that there are estimates of the number of YTBDS for some diseases. For example in breast cancer, Fachal et al. [Fachal and Dunning, 2015] estimated that there may be 1000 or more YTBDS. We can use this information to inform the MLE. We assume that in addition to the observed top hits we also have an additional m YTBDS. We make the further assumption that these have low effect size and the concomitant lack of power justifies why they have yet to be discovered. Of course they could be very rare SNPs with larger effect sizes but very rare SNPs are not the target of GWAS, since there is never likely to be enough power to detect them even in large consortia studies. We further assume that all m YTBDS have log odds ratios in a symmetric interval around zero $[-c, c]$ where $c > 0$ and is likely to be small. This is akin to standard interval censoring except that we observe values in the tails, rather than the central region of the distribution. We take c to be the minimum absolute log odds ratio in the breast cancer top hits which is $\log(1.02)$, but note

that log(1.02) is technically an upper bound for c. The log-likelihood for a single observation x (log-odds ratio) is given by

$$l(x;\theta) = \begin{cases} log f_Y(x;\theta) & \text{if } \mid x \mid \geq c \\ log(2F_Y(c;\theta) - 1) & \text{otherwise} \end{cases} \tag{5}$$

and the log-likelihood based on observations $x_1, \ldots, x_n$ and m YTBDS is given by

$$l(\theta; x_1, x_2, ..., x_n, m) = \sum_{i=1}^{n} \log(f_Y(x_i;\theta)) + m \log(2F_Y(c\,;\theta) - 1) \tag{6}$$

Analytic solutions are not available for the MLEs for this likelihood so we use the `optimize` function in `R` to find the MLE.

Estimating the uncertainty of the MLE

It is appealing to use the MLE as plug-in parameter estimate in the prior in a Bayesian fine-mapping analysis. In order to assess whether this is appropriate we need to consider the uncertainty in the MLE. Large uncertainties would make it inappropriate to use the MLE as a plug-in estimate. We quantify the uncertainty using 95% confidence intervals for the parameter estimates. We calculate confidence intervals for the parameters in three ways: using observed information and an asymptotic normal approximation for the MLE, using expected (Fisher) information and an asymptotic normal approximation for the MLE; using a likelihood interval based on the likelihood ratio test.

The observed information has been shown, based on both empirical and theoretical considerations, to have better coverage properties [Efron and Hinkley, 1978] but this study mostly focussed on the shape rather than the scale parameter so it is not clear in the cases we consider which has better properties in practice. The observed information is simpler to calculate since it avoids the need for integration but the expected information is available in closed for our random variables (derivations are provided in the Appendix). We consider confidence intervals for four random variables: the truncated Laplace and Normal random variables and the censored Laplace and Normal random variables.

Observed and expected information

The observed information can be used to derive the standard error of the MLE. If $l(\theta; x_1, \ldots, x_n)$ is the log likelihood and θ^* is the MLE, then the observed information that the data gives about θ is

$$J(\theta) = -\frac{d^2}{d\theta^2} l(\theta; x_1, \ldots, x_n)\Big|_{\theta=\theta^*} \tag{7}$$

Calculating the observed information is a straightforward exercise in differentiation for all our cases. The expected information about θ in one observation is given by

$$I_1(\theta) = \mathbb{E}_{X_i|\theta}\left(\frac{d}{d\theta} l(\theta; X_i)\right)^2 = -\mathbb{E}_{X_i|\theta}\left(\frac{d^2}{d\theta^2} l(\theta; X_i)\right) \tag{8}$$

For n independent and identically distributed random variables it follows that $I_n(\theta) = nI_1(\theta)$. Calculating the expected information is straightforward for the truncated random variables. For the censored random variables, we adopt the notation and modify the approach of Gertsbakh [Gertsbakh, 1995].

Let Y be a random variable as defined in either Equations (1) or (2), x be an observed top hit, $h(x,\theta) = \frac{d^2 log f_Y(x;\,\theta)}{d\theta^2}$ and $q(\theta) = \frac{d^2 log(2F_Y(c;\,\theta)-1)}{d\theta^2}$. After noting that $Y|Y \geq c$ has density $f_Y(x;\theta)/(2F_Y(c;\theta)-1)$, by conditioning on whether Y is censored is not, we have

$$\mathbb{E}_{X|\theta}\left(\frac{d^2 l(X,\theta)}{d\theta^2}\right) = \int_{-\infty}^{-c} h(x;\theta) f_Y(x;\theta)\, dx + \int_{c}^{\infty} h(x;\theta) f_Y(x;\theta)\, dx + (2F_Y(c;\theta)-1) q(\theta) \tag{9}$$

$$= \int_{-\infty}^{-c} f''_{\theta\theta} - \left(f'_\theta\right)^2/f\, dx + \int_{c}^{\infty} f''_{\theta\theta} - \left(f'_\theta\right)^2/f\, dx + 2\left(F''_{\theta\theta} - \frac{2F'_\theta}{(2F-1)}\right) \tag{10}$$

where $f = f_Y(x;\theta)$, $f'_\theta = \frac{df_Y(x;\theta)}{d\theta}$, $f''_{\theta\theta} = \frac{d^2 f_Y(x;\theta)}{d\theta^2}$, $F = F_Y(c;\theta)$, $F'_\theta = \frac{dF_Y(c;\theta)}{d\theta}$, $F''_{\theta\theta} = \frac{d^2 F_Y(c;\theta)}{d\theta^2}$. Since $2F''_{\theta\theta} = -\int_{-\infty}^{-c} f''_{\theta\theta}\, dx - \int_{c}^{\infty} f''_{\theta\theta}\, dx$ we have that

$$\mathbb{E}_{X|\theta}\left(\frac{-d^2 l(X,\theta)}{d\theta^2}\right) = \int_{-\infty}^{-c} \left(f'_\theta\right)^2/f\, dx + \int_{c}^{\infty} \left(f'_\theta\right)^2/f\, dx + \frac{4F'_\theta}{(2F-1)} \tag{11}$$

We calculate the confidence interval in the usual way assuming that the MLE is asymptotically normal with an expectation equal to the true parameter and variance equal to the reciprocal of the

information (observed or expected). The calculations for the observed and expected information in all cases are given in the Appendix.

Likelihood interval

The likelihood interval is based the standard likelihood ratio test. Assume the data consists either of just the top hits $x_1, \ldots, x_n$ or the top hits plus the number of YTBDS, θ is the parameter of interest, $\hat{\theta}_n$ is the MLE of θ which depends on the sample size n, and l is the log-likelihood. Then standard statistical theory states that, asymptotically, minus twice the difference in the log likelihood evaluated at the true value of the parameter and the MLE has an asymptotic χ^2_1 distribution. Therefore we can define a 95% likelihood confidence interval as $\{\theta : l(\theta; \text{data}) > l(\widehat{\theta}_n; \text{data}) - 1.92\}$.

Model selection

We have so far been concentrating on estimating parameters for various Laplace and Normal random variables. Perhaps the key question from a practitioner's perspective is how to choose between competing prior distributions. One approach is simply to compare the empirical and theoretical CDFs (see Figures 2 and 3) but to obtain a quantitative measure we adopt a Bayesian model framework. Let M_1 represent the Laplace prior and M_2 represent the Normal prior. We will consider the truncated and censored cases separately. Let $Pr(M_1) = p$, $Pr(M_2) = 1 - p$ and $\boldsymbol{x} = x_1 \ldots, , x_n$. Then Bayes theorem gives

$$Pr(M_1|\boldsymbol{x}) = \frac{pf(\boldsymbol{x}|M_1)}{pf(\boldsymbol{x}|M_1) + (1-p)f(\boldsymbol{x}|M_2)} \tag{12}$$

where

$$f(\boldsymbol{x}|M_i) = \int_\theta \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta \tag{13}$$

and θ is the parameter for model M_i, $\pi(\theta)$ is the prior on θ (assigned before the top hits are observed) and $f(\boldsymbol{x}|\theta)$ is the likelihood which can be obtained from Equations (4) and (6). Here p is the prior probability of the Laplace model. The top hits data is used to update this prior

probability to give the posterior probability of the Laplace model. The posterior probability is then a function of p and the marginal likelihoods of the Laplace and Gaussian models. So prior ambivalence about the Laplace versus Gaussian model would be represented by $p = 1/2$. We assume a convenient and flexible Ga(η, ϕ) prior for both w and λ. The Gamma PDF is given in Equation (14) and this parameterisation is used throughout the paper. We show in the next section how the hyperparameters η and ϕ could be elicited. For the truncated Laplace we can derive $f(\boldsymbol{x}|M_1) = (\phi/\kappa)^{\eta}(2\kappa)^{-n}\Gamma(n+\eta)/\Gamma(\eta)$. For the censored Laplace we have

$$f(\lambda) = \frac{\phi^{\eta}}{\Gamma(\eta)}\lambda^{\eta-1}\exp(-\lambda\phi) \tag{14}$$

$$f(\boldsymbol{x}, m|\lambda) = [1-\exp(-\lambda c)]^m \left(\frac{\lambda}{2}\right)^n \exp(-\lambda\sum_{i=1}^{n}|x_i|) \tag{15}$$

$$= \left(\frac{\lambda}{2}\right)^n \sum_{i=0}^{m}\binom{m}{i}(-1)^i \exp\left(-\lambda\left[\sum_{i=1}^{n}|x_j| + ci\right]\right) \tag{16}$$

With $h_i = \sum_{j=1}^{n}|x_j| + \phi + ci$ the marginal likelihood can be shown to be

$$f(\boldsymbol{x}, m) = \frac{\phi^{\eta}}{2^n}\frac{\Gamma(n+\eta)}{\Gamma(\eta)}\sum_{i=0}^{m}\frac{(-1)^i}{h_i^{n+\eta}}\binom{m}{i} \tag{17}$$

For the marginal likelihoods involving a Normal prior we need to use a numerical procedure, eg Monte Carlo integration to evaluate the integral. The Jeffrey's prior is appealing as a prior for scale parameters, and for the truncated Laplace case is easy to derive but is improper ($\pi(\lambda) \propto 1/\lambda$) so isn't suitable for marginal likelihood calculations. The Jeffrey's prior doesn't have a closed-form for the censored Laplace case so we cannot easily use Jeffrey's priors for model selection in either case.

Eliciting priors on λ and w

We put a Ga(η, ϕ) prior on λ and w and elicit from an expert, two values of the log odds ratios that enable us to specify values of the hyperparameters η and ϕ. Let Y be a random variable as defined in either Equations (1) or (2) and Q represent its 75^{th} percentile (so Q is the median of the positive log odds ratio). For $Y \sim$ La(λ) we have that $Q = ln2/\lambda$ whilst for $Y \sim N(0, w)$ we can show that $Q = \sqrt{w}\Phi^{-1}(0.75)$ where Φ is the CDF of the standard Normal distribution. We

then elicit values x_l and x_u such that $Pr(Q > x_u) = 0.05$ and $Pr(0 < Q < x_l) = 0.05$; x_l and x_u represent the expert-elicited end points of the central probability region of the 75^{th} percentile. Once values of x_l and x_u have been elicited we can obtain λ_l and λ_u via $\lambda_l = ln2/x_u$ and $\lambda_u = ln2/x_l$, and w_l and w_u via $w_l = (x_l/\Phi^{-1}(0.75))^2$ and $w_u = (x_u/\Phi^{-1}(0.75))^2$. To find the hyperparameters η and ϕ for the prior on λ we solve the simultaneous equations $Pr(\lambda < \lambda_l; \lambda \sim \text{Ga}(\eta, \phi)) = 0.05$ and $Pr(\lambda > \lambda_u; \lambda \sim \text{Ga}(\eta, \phi)) = 0.05$ for η and ϕ. We use the `optim` function in `R` to do this. We do something similar to find η and ϕ for the Ga(η, ϕ) prior on w. An alternative approach to elicitation is simply to use the asymptotic distribution of the MLE as the prior. In this case the prior would be a Normal distribution with mean equal to the MLE and variance equal to the MLE variance (often provided by the optimising software used). We provide more details of this in the results section.

Simulation study

We simulated genotype data between base pairs 201666128 and 201866128 around the CASP8 region on chromosome 2 using Hapgen2 [Su et al., 2011]. We considered two scenarios, both with a single causal SNP. We simulated 100 data sets for each scenario. In both scenarios we simulated and analysed the data assuming an additive risk model on the log odds scale. In the first scenario we simulated 10,000 cases and 10,000 controls with a causal SNP MAF of 0.448 and a simulated odds ratio of 1.06. In the second scenario we simulated 20,000 cases and 20,000 controls with a causal SNP MAF of 0.115 and a simulated odds ratio of 1.03. Both of these are relatively small odds ratios but it is at these smaller odds ratios that we might expect our priors to give a superior rank to the causal SNP, since the effect size prior variance is estimated assuming many YTBDS with small odds ratios. We calculated the Wakefield Bayes factors with three values of the effect size prior variance w. The first value is $w = 0.0051$ which is based on the top hits data assuming there are 100 YTBDS (see Table 2); the second value is $w = 0.1$ which is the default value used in Pickrell [2014]; the third value is $w = 0.04$ which is the default value used in SNPTEST2 [Marchini and Howie, 2010]. For each simulated data set we calculated the mean of the true positive rate at a set of given false positive rates and plot the results using an ROC curve.

Results

Simulated coverage using the three confidence interval approximations

We define the unobserved region probability (URP) as $F_Y(c) - F_Y(-c)$ for some value c and where $F(.)$ is the CDF. A small URP implies we observe top hit values close to zero (as for the breast cancer top hits) whilst a large URP means we only observe values in the tails of the distribution. The latter might represent GWAS studies still using relatively small numbers of cases and controls and consequently little power to detect small odds ratios. We assess the coverage (proportion of confidence intervals containing the true parameter value) when $c = \log(1.02)$ in two cases: $w = 0.025$, $\lambda = 5$ and $w = 0.00036$ and $\lambda = 61$. The first (second) case corresponds to a URP of 0.1 (0.7). The coverage probability should be 0.95 for a nominal 95% confidence interval. For the truncated random variables we simply simulate n realisations of X. For the censored random variables we simulate n realisations of X and then determine the number of YTBDS by solving $n/m = 2F_Y(-c)/2F_Y(c) - 1$ for m.

The results are shown in Table 1 using 10,000 simulated data sets for each scenario. These results show that the confidence intervals based on truncated random variables generally have better coverage properties (coverage probability closer to 0.95) than those based on censored random variables, for all sample sizes n. For the truncated Laplace the interval based on the likelihood has good coverage at all values of n whereas the coverage based on the observed and expected information improves as n increases. For the truncated Normal, when the URP is small, the coverage is generally close to 0.95 for the likelihood interval for all n and close to 0.95 for the observed and expected information approach when n is large. For the truncated Normal, when the URP is 0.7, the coverage properties are considerably worse.

For the censored Laplace the coverage appears close to that expected but is consistently too high (around 0.96). When the URP is small, the coverage based on the likelihood interval is close to 0.95 for the censored Normal random variable, but not when the URP is larger. When the URP is small, the coverage based on the expected information is close to 0.95 for large n for the censored Normal but intervals based on the observed information have coverage close to 1 in most cases.

These results suggest that the URP is particularly important in determining the coverage properties of the intervals used, particularly for the censored cases and particularly for the censored

Normal case. Likelihood intervals are generally better, in terms of coverage, than those based on information. For the censored Normal case the expected information sometimes performs better, in terms of coverage, than the observed information. This is contrary to observations made in other studies that don't consider truncated or censored random variables [Efron and Hinkley, 1978].

Graphical assessment of the fit of the top hits to the priors

Using the 148 top hits as the observations and the log-likelihood in Equation (4) we obtain MLEs $\widehat{w} = 0.00693$, $\widehat{\lambda} = 18.31$ for the truncated random variables. Figure 2 shows the CDF of the Normal and Laplace priors with these MLEs as well as the empirical CDF of the 148 top hits. The Laplace prior visually seems to fit the top hits better than the Normal prior. The poor fit around $x = 0$ is simply due to slight imbalance in the observed proportion of positive log odds ratios.

We use Equation (6) to estimate λ and w for the censored random variables with various values of the number YTBDS. Figure 3 shows the CDF of the two possible prior distributions using the MLEs from Equation (6) as well as the empirical CDF of the 148 top hits. Again we see a better fit of the top hits to the Laplace prior than to the Normal prior.

MLEs based on the top hits

Table 2 shoes the MLE and the 95% confidence interval for the parameters for the truncated and censored Laplace and Normal random variables using the 148 breast cancer top hits. For the censored case the number of YTBDS was varied between 100 and 1000. We used different methods for constructing the intervals depending on which method gave the most reliable results in Table 1. It is clear from Table 2 that the parameter estimates are very sensitive to the estimated number of YTBDS. Generally the number of YTBDS is not known with much certainty. The maximum likelihood approaches considered do not easily allow this uncertainty to be incorporated in to parameter estimation. As a result we advocate a Bayesian approach that naturally accommodates this uncertatinty.

Results of the simulation experiment

In the Materials and Methods section we described a simulation study to assess the causal SNP ranks, based on Wakefield Bayes factors, using our prior based on the top hits data and two other

values of the effect size prior variance. The results are shown in the ROC curves in Figure 4. For these parameter settings, our method of informing the effect size prior variance using the top hits data gives superior performance compared to the two default values used in Pickrell [2014] and SNPTEST2 [Marchini and Howie, 2010].

Using the top hits to choose between competing priors for the effect sizes

Suppose we use the previously described elicitation approach and elicit $x_l = \log(1.005)$ and $x_u = \log(1.12)$ representing prior belief that the log odds ratios of breast cancer causal SNPs are relatively small. These values give $\lambda \sim \text{Ga}(1.49, 0.028)$ and $w \sim \text{Ga}(0.564, 73.5)$. Using these priors we consider the effect of sample size on the posterior probabilities of the models by taking the n top hits with largest absolute values (the n values furthest from 0). We take c as the minimum of the absolute values of the elements of these n values. Smaller numbers of top hits are what you might expect in GWAS studies of smaller size and in diseases less well studied than breast cancer. In these cases there will be fewer genome-wide significant GWAS top hits and larger effect sizes will be discovered first. We consider $n = 10, 50, 100$ and 148 where 148 is the total number of available breast cancer top hits. We also considered what happens in the case where the elicited values of the 75^{th} percentile of the log odds ratio are $x_l = \log(1.05)$ and $x_u = \log(1.20)$ representing prior belief that the log odds ratios of breast cancer causal SNPs are relatively large. The corresponding priors on λ and w are shown in Figure 5. The posterior probabilities of the laplace model, using truncated random variables, are shown in Table 3.

Regardless of whether the prior belief is for relatively small or large log odds ratios we find overwhelming evidence in favour of the Laplace model with $n \geq 50$ with posterior probabilities exceeding 0.99 in all cases. For $n = 10$ the posterior probabilities drops to around 0.60 but still show more support for the Laplace than the Normal prior.

Table 3 shows the posterior probabilities of the laplace model, using censored random variables. We see that the number of YTBDS has more of an impact on the posterior probability of the Laplace model than the number of observed top hits. When there are 10 YTBDS there is very strong support for the Normal prior, with 50 YTBDS the evidence is equivocal with generally more support for the Laplace prior. When there are 100 or more YTBDS the support for the Laplace is very strong. We don't consider $m > 100$ since the posterior probability in nearly all cases is close

to one.

Prior construction using GWAS top hits for the Laplace prior

The numbers in Table 3 show overwhelming support for the Laplace prior when $n = 148$ and m is fifty or more, which is believed to be the case in breast cancer ($m \approx 1000$). As a result we only consider the Laplace prior hereafter. We assume a $\lambda \sim Ga(\eta = 1.5, \phi = 0.028)$ prior that has strong support for smaller odds ratios (as used in Table 3). Using Equation (3) we get a truncated Laplace likelihood based on n top hits $x_1, \ldots, x_n$ which is proportional to $\lambda^n \exp(-\lambda(\sum |x_i| - nc))$. The posterior distribution of λ is therefore a gamma distribution $\lambda|x_1, \ldots, x_n \sim \text{Ga}(n+\eta, \sum |x_i| - nc + \phi)$. Figure 6(a) shows the posterior distribution for $n = 25, 75$ and 148. The posterior mode here is $(n + \eta - 1)/(\sum |x_i| - nc + \phi)$ which, with $\eta = 1.488$ and $\phi = 0.0285$, is very close to the MLE given by $n/(\sum |x_i| - nc)$ for all values of n considered. Similarly the MLE variance given by $n/(\sum_{i=1}^{n} |x_i| - nc)^2$ and posterior variance given by $(n+\eta)/(\sum_{i=1}^{n} |x_i| - nc + \phi)^2$ are quite similar for all n so, assuming a truncated random variable for the log odds ratio, a sensible approach is simply to use the Normal distribution of the MLE as a prior for λ in fine-mapping (even when n is small). Alternatively, given the relatively small posterior variance it is also reasonable to use the MLE as a plug in estimate for λ in the Bayes factor calculation (whilst checking the sensitivity of the Bayes factor to using other plausible values of λ indicated by the posterior).

Using knowledge of the number of YTBDS to inform the prior

Assuming a $\lambda \sim Ga(\eta = 1.5, \phi = 0.028)$ prior we can derive a closed form expression for the posterior distribution. The prior and likelihood are given in Equations (14) - (16) and the marginal likelihood in Equation (17) and so, after a little algebra, we find that the posterior probability density is given by

$$f(\lambda|x_1, \ldots, x_n, m) = \sum_{i=0}^{m} w_i f_{Q_i}(\lambda) \tag{18}$$

where $Q_i \sim \text{Ga}(n + \eta, h_i)$, $h_i = \sum_{j=1}^{n} |x_j| + \phi + ci$ and

$$w_i = \frac{\binom{m}{i} \frac{(-1)^i}{h_i^{n+\eta}}}{\sum_{i=0}^{m} \binom{m}{i} \frac{(-1)^i}{h_i^{n+\eta}}} \tag{19}$$

The posterior distribution is an alternating sign mixture of gamma random variables.

Figure 6(b) shows the posterior distributions obtained with varying number of YTBDS (m=100, 250, 500 and 1000). The 95% credible intervals in Figure 6(b) and the 95% confidence intervals based on the MLEs in Table 2 are very similar so we conclude that the prior is highly uninformative. Again we could use the asymptotic distribution of the MLE rather than the posterior distribution as a fine-mapping prior for the effect size. However this is only possible if there is certainty in the number of YTBDS. This is quite unlikely and so we need to be able to take account of the uncertainty in the number of YTBDS. If M is the random variable representing the number of YTBDS then using the notation of Equations (18) and (19) we have

$$f(\lambda|x_1,\ldots,x_n) = \sum_m Pr(M=m) \sum_{i=0}^{m} w_i f_{Q_i}(\lambda). \tag{20}$$

In breast cancer we believe there may be around 1000 additional loci but there is considerable uncertainty in this number so we choose two different Beta-Binomial(α, β) priors on M with support $m \in [0, 2000]$ with parameters $\alpha = 11, \beta = 4$ giving a negatively skewed prior and $\alpha = 7, \beta = 7$ giving a prior for m symmetric around $m = 1000$. These two priors for m are shown in Figure 7(a). Figure 7(b) shows both the prior distribution on λ and the posterior distributions on λ for both of the BetaBinomial prior distributions on m. How these derived posteriors might be used in practice again depends on the variance. A small variance implies that using a point estimate for λ may be justified, with some checks on the sensitivity to small changes. The variance observed in Figure 6(b) suggests that using the posterior mode, mean or median in breast cancer may not be appropriate. Ideally we would take the results of Figure 6(b) as the prior for the parameter in a Bayesian fine-mapping study, but a Normal or Gamma distribution approximation may also be a sensible compromise.

R libraries

The R code developed to to implement the methods in this paper is available for download from http://www.kevinwalters.staff.shef.ac.uk/

Discussion

In this paper we show how GWAS top hits can be used, in addition to estimates of the number of YTBDS if they are available, to choose between competing causal SNP effect size prior distributions. We show how likelihood-based approaches can be used to estimate the parameters of these effect size distributions and how the uncertainty in the parameter estimates can be quantified. We also demonstrate how the uncertainty in the number of YTBDS can be accommodated. We show that for the breast cancer GWAS top hits, the Laplace distribution convincingly provides a better fit compared with the Normal distribution which is the default prior distribution in Bayesian fine-mapping association studies.

Even with 148 breast cancer GWAS top hits there is still considerable uncertainty in the posterior distribution of λ which in part results from the uncertainty put on the number of YTBDS. It is clear that the estimates of λ can change substantially when the number of YTBDS is taken into account. This suggest that caution needs to be applied when using an estimate of the number of YTBDS and that the uncertainty in the number of YTBDS must be fully taken into account. At the very least this paper shows how GWAS top hits can be used to obtain MLEs of the scale parameter of the effect size prior along with a confidence interval.

In current fine-mapping studies the parameters (λ or w) are either given default values or subjective elicitation is used to find the value that best fits the expert's beliefs. Our framework provides a way to use the elicitation process to instead elicit a prior distribution on λ and then update this distribution using the top hits to give a posterior distribution that could then be used as a fine-mapping prior.

In a univariate fine-mapping study it is straightforward to calculate the Bayes factor with a prior on the parameter of the effect size distribution. This could be done simply by Monte Carlo integration, which has well understood and guaranteed asymptotic convergence properties. In a multivariate model selection approach that enumerates over possible models, such as that used in FINEMAP [Benner et al., 2016], putting a prior on the parameter of the effect size distribution would lead to intractable marginal likelihoods. Calculating these marginal likelihoods by say Monte Carlo integration is not computationally feasible due to the number of possible models that need to be enumerated. In this case the best that could possibly be done is to use the posterior mode or

median as a point estimate of the hyperparameter. A sensitivity analysis should then be undertaken to assess the effect of small changes in the parameter value on the posterior probability of causal association for each SNP.

We have derived closed form expressions for the posterior distributions on λ for the Laplace prior but these do not exist for the Normal prior. Instead it would be necessary to do some sort of posterior sampling, for example via Metropolis-Hastings. This is another clear advantage of the Laplace prior over the Normal prior on the effect sizes.

A potential weakness of this approach is the assumption that GWAS top hits are all causal SNPs or in strong linkage disequilibrium with causal SNPs. To us this seems a reasonable assumption. If not, it seems reasonable to assume that the probability that a top hit is not causal is likely to the same for each top hit, and not related to the top hit effect size. So if the truely causal SNPs are a subset of the top hits then at least they are a random subset.

The aim of this paper is to bring to the attention of practitioners the fact that genome-wide significant GWAS top hits, even in very small number, can be highly informative about the scale parameters of the causal SNP effect size prior distribution which are currently often specified in ad hoc ways. Using more objective data-driven approaches described in this paper ameliorate the criticisms about using completely subjective elicited priors. Additional information about the number of YTBDS can also provide much information about the prior parameters but the uncertainties in the estimates of the number of YTBDS must be carefully taken into account.

Acknowledgements

Some of this was carried out as part of a PhD funded by the Ministry of Education Malaysia (Higher Education) and University of Malaya. We thank three anonymous reviewers whose helpful comments enabled us to improve the quality of the manuscript.

References

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). Finemap: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501.

Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.*, 3:e114.

Boggis, E., Milo, M., and Walters, K. (2016). equips: eqtl analysis using informed partitioning of snps–a fully bayesian approach. *Genet Epidemiol*, 40(4):273–283.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., and Schaid, D. J. (2015). Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65:457–483.

Fachal, L. and Dunning, A. M. (2015). From candidate gene studies to gwas and post-gwas analyses in breast cancer. *Curr Opin Genet Dev*, 30:32–41.

Gertsbakh, I. B. (1995). On the fisher information in type-i censored and quantal response data. *Stat Probabil Lett*, 23:297–306.

Guan, Y., Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat*, 5(3):1780–1815.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7):e1000130.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*, 45(4):353–361.

Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemacon, D., Soucy, P., Glubb, D., and Rostamianfar, A. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94.

Owen, D. (1980). A table of normal integrals. *Comm Statist Simulation Comput*, 9(4):389–419.

Pereira, M., Thompson, J. R., Weichenberger, C. X., Thomas, D. C., and Minelli, C. (2017). Inclusion of biological knowledge in a bayesian shrinkage model for joint estimation of snp effects. *Genet Epidemiol*, 41(4):320–331.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, 94(4):559–573.

Spain, S. L., Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Hum Mol Genet*, 24(R1):R111–119.

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. *Genet Epidemiol*, 40(3):176–187.

Spencer, A. V., Cox, A., and Walters, K. (2014). Comparing the efficacy of snp filtering methods for identifying a single causal snp in a known association region. *Ann Hum Genet*, 78(1):50–61.

Su, Z., Marchini, J., and Donnelly, P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305.

Wakefield, J. (2008). Reporting and interpretation in genome-wide association studies. *Int J Epidemiol*, 37(3):641–653.

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

Appendix

Information for the truncated Laplace distribution

From Equations (2) and (4) the log-likelihood is

$$l(\lambda; x_1, \ldots, x_n) = n \log\lambda + \lambda \left(nc - \sum_{i=1}^{n} |x_i| \right) \tag{21}$$

and both the observed and expected information are therefore given by $I_n(\lambda) = J(\lambda) = n\lambda^{-2}$

Information for the truncated Normal distribution

Let $v = c/\sqrt{(}w)$, $\phi(.)$ represent the standard Normal PDF and $\Phi(.)$ represent the standard Normal CDF. From Equation (4) the log-likelihood based on a single observation x_i is

$$l(w; x_i) = -\frac{1}{2}\log(w) - \log\Phi(-v) - \frac{x_i^2}{2w}$$

and hence

$$\frac{d^2 l(w; x_i)}{dw^2} = \frac{1}{2w^2} - \frac{d^2 \log\Phi(-v)}{dw^2} - \frac{x_i^2}{w^3}. \tag{22}$$

It follows, from the second fundamental theorem of calculus, that

$$\frac{d^2}{dw^2}\Phi(-v) = \frac{v^5(v^2-3)\phi(v)}{4c^4}. \tag{23}$$

Therefore

$$\frac{d^2}{dw^2}\log\Phi(-v) = \frac{v^5\phi(v)}{4c^4\Phi^2(-v)}\left((v^2-3)\Phi(-v) - v\phi(v)\right). \tag{24}$$

Substituting Equation (24) into Equation (22) we obtain

$$J(w) = \frac{n}{4}\left(\frac{v}{c}\right)^4 \left[\frac{v\phi(v)}{\Phi^2(-v)}\left((v^2-3)\Phi(-v) - v\phi(v)\right) - 2\right] + \left(\frac{v}{c}\right)^6 \sum_{i=1}^{n} x_i^2. \tag{25}$$

Using Equation (22), the expected information contained in a single observation x_i is

$$I_1(w) = -\frac{1}{2w^2} + \frac{d^2}{dw^2}\log\Phi(-v) + \frac{1}{w^3}\mathbb{E}\big(X_i^2\big). \tag{26}$$

By symmetry we have

$$\mathbb{E}\left(X_i^2\right) = 2\int_c^\infty x_i^2 f_X(x_i; w)\, dx_i = w\left[1 + \frac{v\phi(v)}{\Phi(-v)}\right] \tag{27}$$

where $f_X(x_i; w)$ is given in Equation (3) and we have used the standard result for $\int u^2\phi(u)du$ [Owen, 1980]. Substituting Equation (24) and (27) into Equation (26) we find

$$I_n(w) = \frac{n}{4}\left(\frac{v}{c}\right)^4\left[\frac{v\phi(v)}{\Phi^2(-v)}\left((v^2-3)\Phi(-v) - v\phi(v)\right) + 4\left(1 + \frac{v\phi(v)}{\Phi(-v)}\right) - 2\right]. \tag{28}$$

Information for the censored Laplace distribution

Using Equation (6) direct differentiation gives

$$J(\lambda) = \frac{n}{\lambda^2} + \frac{mc^2 exp(-\lambda c)}{(1 - exp(-\lambda c))^2} \tag{29}$$

Using Equation (11), we need expressions for $f_Y^{'}(x;\lambda)^2/f_Y(x;\lambda)$ and $F_Y^{'}(c;\lambda)$ where $f_Y(x;\lambda)$ is given in Equation (2). Basic calculation gives

$$f_Y^{'}(x;\lambda)^2/f_Y(x;\lambda) = \frac{1}{2\lambda}(1 - \lambda \mid x \mid)^2 \exp(-\lambda \mid x \mid) \tag{30}$$

and

$$\int_c^\infty f_Y^{'}(x;\lambda)^2/f_Y(x;\lambda)\, dx = \int_{-\infty}^{-c} f_Y^{'}(x;\lambda)^2/f_Y(x;\lambda)\, dx = \left(\frac{1+\lambda^2c^2}{2\lambda^2}\right)\exp(-\lambda c). \tag{31}$$

If $c > 0$ we have $F_Y(c;\lambda) = 1 - 1/2\exp(-\lambda c)$ and

$$\frac{4F_Y(c;\lambda)^{'2}}{2F_Y(c;\lambda) - 1} = \frac{c^2\exp(-\lambda c)}{\exp(\lambda c) - 1} \tag{32}$$

Putting these into Equation (11) gives

$$I_{n+m}(\lambda) = (n+m)\left(\frac{exp(-\lambda c)}{\lambda^2} + \frac{c^2}{exp(\lambda c) - 1}\right) \tag{33}$$

Information for the censored Normal distribution

Assuming that $x_1, \ldots, x_n$ are observed and there are m censored observations, Equation (6) gives the log likelihood as

$$l(w; x_1, \ldots, x_n, m) = -\frac{n}{2}\log(w) + m\log\left[2\Phi(v) - 1\right] - \frac{\sum_{i=1}^{n} x_i^2}{2w} \tag{34}$$

Direct differentiation gives, after much algebra, that

$$J(w) = \left(\frac{v}{c}\right)^6 \sum_{i=1}^{n} x_i^2 - n\left(\frac{v}{c}\right)^4 + \frac{mv^5\phi(v)}{2c^4(2\Phi(v)-1)^2}\left[2v\phi(v) + (v^2-3)(2\Phi(v)-1)\right] \tag{35}$$

Using Equation (11), we need expressions for $f_Y^{'}(x;w)^2/f_Y(x;w)$ and $F_Y^{'}(c;w)$. Direct calculation gives

$$f_Y^{'}(x;w)^2/f_Y(x;w) = \frac{(x^2-w)^2}{4w^4} f_Y(x;w) \tag{36}$$

and then using the standard result for $\int u^4\phi(u)du$ [Owen, 1980] we find

$$\int_c^{\infty} f_Y^{'}(x;w)^2/f_Y(x;w)\, dx = \int_{-\infty}^{-c} f_Y^{'}(x;w)^2/f_Y(x;w)\, dx = \frac{1}{4w^2}\left[2 - 2\Phi(v) + v(1+v^2)\phi(v)\right] \tag{37}$$

With $F_Y(c;w) = \Phi(v)$ we get

$$\frac{4F_Y(c;w)^{'2}}{2F_Y(c;w)-1} = \frac{v^2\phi^2(v)}{w^2(2\Phi(v)-1)}. \tag{38}$$

Putting these into Equation (11) we get

$$I_{n+m}(w) = (n+m)\left(\frac{v}{c}\right)^4\left[1 - \Phi(v) + \frac{v(1+v^2)\phi(v)}{2} + \frac{v^2\phi^2(v)}{2\Phi(v)-1}\right] \tag{39}$$

Figure legends

Figure 1: Probability density of a Normal and Laplace random variable both with a variance of 0.007 and an expectation of zero.

Figure 2: Empirical CDF of the top hits and the theoretical CDFs of the Normal and Laplace distributions with the parameters taken to be the maximum likelihood estimates from a likelihood based on truncated random variables.

Figure 3: Empirical CDF of the top hits and theoretical CDFs of the Normal and Laplace distributions with the parameter taken to be the maximum likelihood estimates from a likelihood based on censored random variables for various values of the number of YTBDS (m).

Figure 4: ROC curves comparing the ranking performance of Wakefield Bayes factors with a $\beta \sim N(0, w)$ effect size prior for three values of w: $w = 0.0051$, based on the top hits data assuming there are 100 YTBDS; $w = 0.1$, the default value used in Pickrell [2014]; $w = 0.04$, the default value used in SNPTEST2 [Marchini and Howie, 2010]. The top row shows the whole false positive range. The bottom row shows false positive values below 0.2. The left hand plots show the results for a single causal SNP with a MAF of 0.448 and an odds ratio of 1.03 with 20,000 cases and 20,000 controls. The right hand plots show the results for a single causal SNP with a MAF of 0.115 and an odds ratio of 1.06 with 10,000 cases and 10,000 controls. A hundred data sets are used in each scenario.

Figure 5 main: Gamma(η, ϕ) prior distributions on λ and w used in model comparison. In both plots two priors are shown, one representing strong support for small odds ratios, the other for strong support for larger odds ratios.

Figure 5a: Gamma prior distributions on λ.

Figure 5b: Gamma prior distributions on w.

Figure 6 main: Posterior distributions on λ with likelihoods based on a) truncated random variables and b) censored random variables. The prior is $\lambda \sim Ga(\eta = 1.5, \phi = 0.028)$ in both plots.

Figure 6a: Posterior distributions on λ as functions of the number of top hits available (n).

Figure 6b: Posterior distributions on λ by number of YTBDS (m) using the information from all $n = 148$ top hits.

Figure 7 main: Various priors on m and λ and posteriors on λ based on different priors on m.

Figure 7a: Two Beta-Binomial priors on m allowing uncertainty around the most likely value of $m = 1000$.

Figure 7b: The elicited Gamma prior on λ and two posteriors on λ with different Beta-Binomial priors on m.

Tables

		URP = 0.1				URP = 0.7			
		Truncated		Censored		Truncated		Censored	
		La(λ)	N(0,w)	La(λ)	N(0,w)	La(λ)	N(0,w)	La(λ)	N(0,w)
n=200	Observed	0.950	0.945	0.960	1.000	0.950	0.975	0.960	0.917
	Expected	0.950	0.945	0.960	0.952	0.950	0.978	0.960	0.909
	Likelihood	0.949	0.949	0.960	0.956	0.949	0.914	0.960	0.997
n=50	Observed	0.951	0.940	0.961	1.000	0.950	0.970	0.961	0.999
	Expected	0.951	0.931	0.962	0.939	0.950	0.968	0.961	0.996
	Likelihood	0.950	0.950	0.961	0.956	0.950	0.887	0.960	0.980
n=10	Observed	0.953	0.914	0.962	0.999	0.961	0.898	0.963	1.000
	Expected	0.955	0.872	0.963	0.883	0.961	0.861	0.963	0.997
	Likelihood	0.948	0.947	0.961	0.955	0.947	0.945	0.962	0.999

Table 1. Coverage of 95% confidence intervals for Normal and Laplace parameters for truncated and censored random variables where $c = \log(1.02)$ and the number of top hits is 10, 50 and 200. Three methods for constructing the intervals are considered based on: observed information, expected information and the likelihood interval. In columns 3 - 6, $w = 0.025$ and $\lambda = 5$ which give a small unobserved region probability (URP) of 0.1. In columns 7 - 10, $w = 0.00036$ and $\lambda = 61$ which gives a large URP of 0.7

		Laplace(λ)		N(0,w)	
Type	number of yet-to-be discovered SNPs	$\hat{\lambda}$	95% CI	$\hat{w}$ $\times 10^{-5}$	95% CI $\times 10^{-5}$
Truncated	NA	18.3	(15.4, 21.3)	692	(559,867)
Censored	100	20.8	(18.2, 23.4)	510	(430,615)
Censored	250	30.1	(27.1, 33.0)	326	(286,374)
Censored	500	42.4	(39.1, 45.6)	202	(183,226)
Censored	1000	60.5	(56.8, 64.2)	121	(113,130)

Table 2. MLEs and 95% confidence intervals for Normal and Laplace parameters for truncated and censored random variables based on the 148 breast cancer top hits and variable numbers of YTBDS for the censored case.

Posterior probability of Laplace(λ) prior								
	Prior support for smaller effect sizes ($x_l = 1.005,\ x_u = 1.120$)				Prior support for larger effect sizes ($x_l = 1.050,\ x_u = 1.200$)			
Number of top hits	truncated likelihood	censored likelihood			truncated likelihood	censored likelihood		
		$m = 10$	$m = 50$	$m = 100$		$m = 10$	$m = 50$	$m = 100$
148	1.0	1.8×10^{-3}	1.0	1.0	1.0	4.8×10^{-3}	1.0	1.0
100	1.0	9.9×10^{-5}	0.31	1.0	1.0	4.3×10^{-4}	0.50	1.0
50	1.0	1.4×10^{-3}	0.51	1.0	0.99	5.5×10^{-3}	0.74	1.0
10	0.64	0.12	0.60	0.95	0.60	0.19	0.78	0.96

Table 3. Posterior probabilities, using the censored likelihood, for the Laplace prior where the number of top hits (n) varies between 10 and 148, the number of YTBDS (m) varies between 10 and 100. Two scenarios are considered: prior belief in smaller log odds ratios and prior belief in larger log odds ratios.

Figures

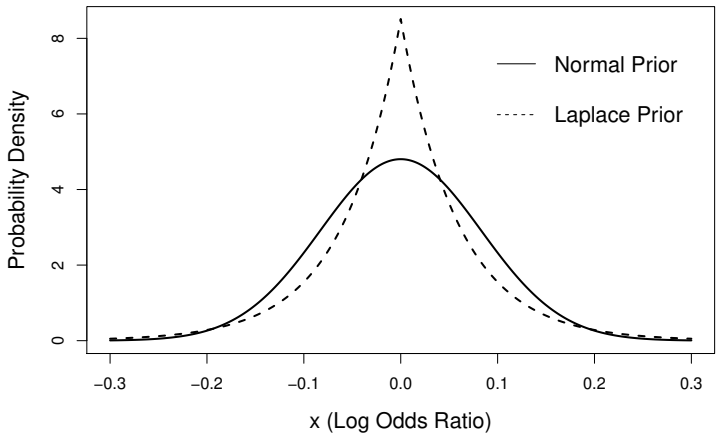


Figure 1

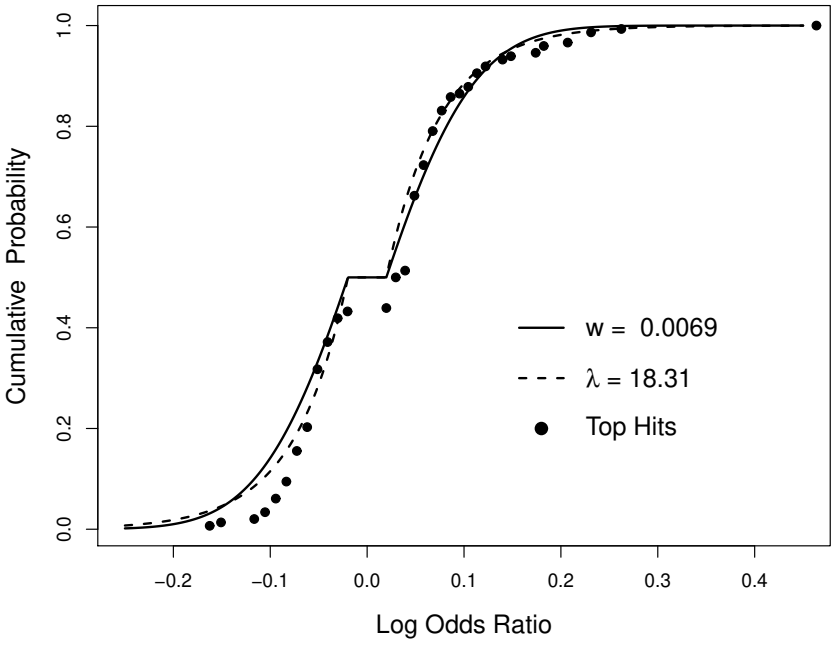


Figure 2

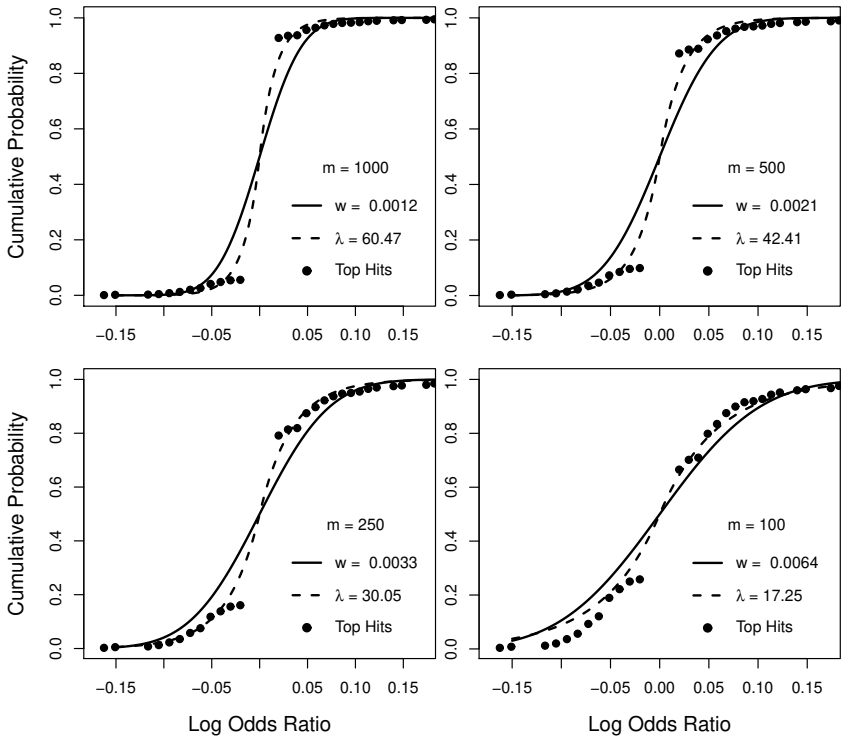


Figure 3

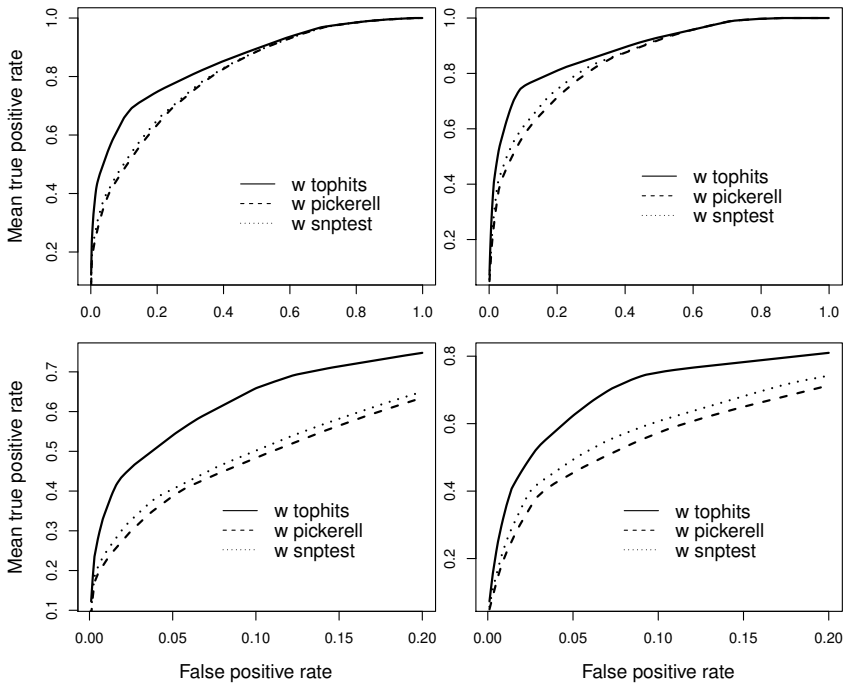
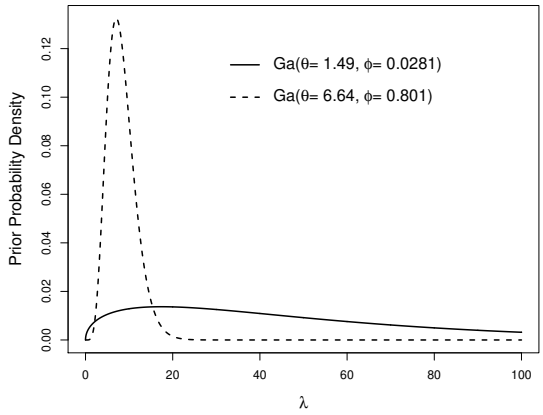
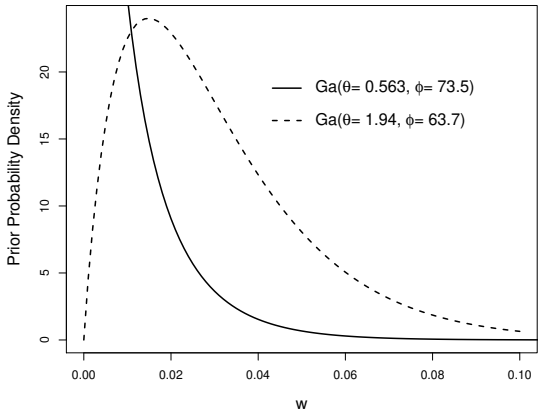


Figure 4

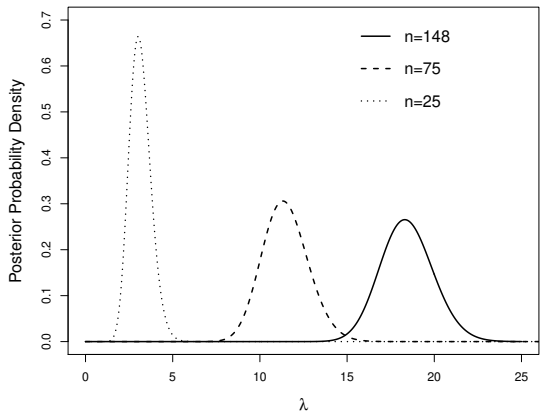


(a)

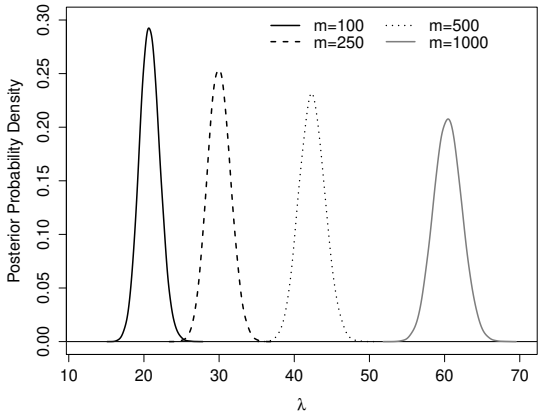


(b)

Figure 5

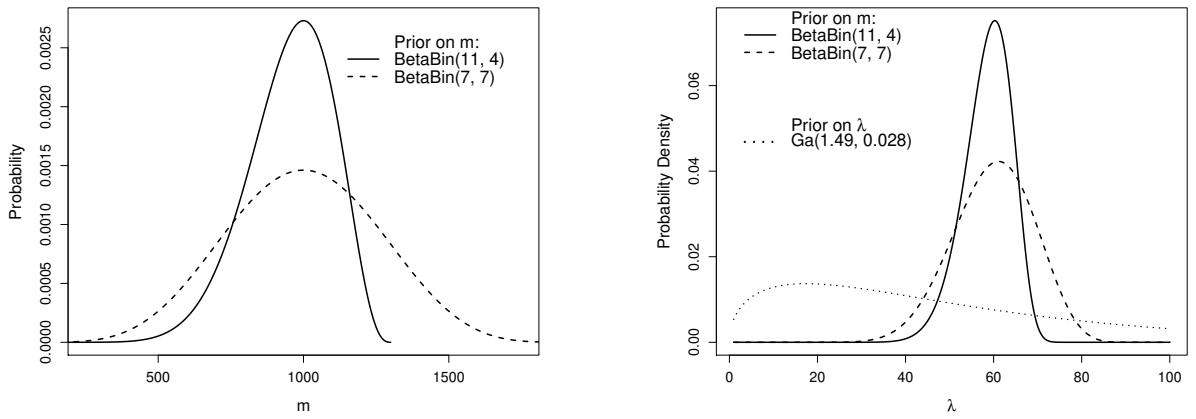


(a)



(b)

Figure 6


Probability
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
500
1000
1500
m
Prior on m:
BetaBin(11, 4)
BetaBin(7, 7)
Probability Density
0.00
0.02
0.04
0.06
0
20
40
60
80
100
λ
Prior on m:
BetaBin(11, 4)
BetaBin(7, 7)
Prior on λ
Ga(1.49, 0.028)

(a) (b)

Figure 7