This is a repository copy of *Bayesian variable selection using partially observed categorical prior information in fine-mapping association studies*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/145782/

Version: Accepted Version

Article:
Alenazi, A., Cox, A. orcid.org/0000-0002-5138-1099, Juarez, M. et al. (2 more authors) (2019) Bayesian variable selection using partially observed categorical prior information in fine-mapping association studies. Genetic Epidemiology. ISSN 0741-0395

https://doi.org/10.1002/gepi.22213

This is the peer reviewed version of the following article: Alenazi, AA, Cox, A, Juarez, M, Lin, W-Y, Walters, K. Bayesian variable selection using partially observed categorical prior information in fine-mapping association studies. Genet. Epidemiol. 2019, which has been published in final form at https://doi.org/10.1002/gepi.22213. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Bayesian variable selection using partially observed categorical prior information in fine-mapping association studies

Abdulaziz A. Alenazi[1,4], Angela Cox[2], Miguel Juarez[1], Wei-Yu Lin[2,3], Kevin Walters[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

[3] Northern Institute for Cancer Research, University of Newcastle, Newcastle, UK

[4] Department of Mathematics, Northern Border University, Arar, Saudi Arabia

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

Abstract

Several methods have been proposed to allow functional genomic information to inform prior distributions in Bayesian fine-mapping case-control association studies. None of these methods allow the inclusion of partially-observed functional genomic information. We use functional significance scores that combine information across multiple bioinformatics sources to inform our effect size prior distributions. These scores are not available for all SNPs but by partitioning SNPs into naturally occurring functional significance score groups, we show how missing functional significance scores can easily be accommodated via finite mixtures of elicited priors. Most current approaches adopt a formal Bayesian variable selection approach and either limit the number of causal SNPs allowed or use approximations to avoid the need to explore the vast parameter space. We focus instead on achieving differential shrinkage of the effect sizes through prior scale mixtures of normals and use marginal posterior probability intervals to select candidate causal SNPs. We show via a simulation study how this approach can improve localisation of the causal SNPs compared to existing mutli-SNP fine-mapping methods. We also apply our approach to fine-mapping a region around the *CASP8* gene using the iCOGS consortium breast cancer SNP data.

Data Sharing Statement

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

Introduction

Genome-Wide Association Studies have been successful in identifying thousands of regions harbouring SNPs associated with complex disease traits. Fine-mapping of these regions is required to identify the actual causal SNP(s) within these regions. Early Bayesian methods focussed on univariate approaches (Wakefield [2009]; Spencer et al. [2015]; Spencer et al. [2014]) but there are now many methods that model the joint effects of SNPs within a region. Most methods use a classical Bayesian model selection approach by introducing an indicator vector representing which SNPs have non-zero effect sizes. Effect sizes are then estimated conditional on them being non-zero. Paintor [Kichaev et al., 2014a] employs an EM algorithm approach and requires enumeration over all indicator vectors, restricting the number of causal SNPS at each locus considered to be no more than three. CaviarBF [Chen et al., 2015] similarly enumerates over all indicator vectors to calculate Bayes factors to compare models and to estimate posterior probabilities of inclusion. Because it performs an exhaustive search it also limits the number of causal SNPs in the model to be no more than five. To overcome the limitation on the number of causal SNPs allowed in the model, FINEMAP [Benner et al., 2016] uses an efficient shotgun stochastic search algorithm to search over the space of indicator vectors which have non-negligible probability. HyperLasso [Hoggart et al., 2008] uses penalized regression with a Normal-Exponential-Gamma prior and searches for the posterior mode, which either leads to inclusion or exclusion of SNPs from the final model, and does not quantify the uncertainty in the SNPs selected. Classical Bayesian variable selection with a large number of predictors and even a modest number of causal SNPs has to overcome the problem of a vast model space, either by limiting the number of causal SNPs in the model or by only considering configurations with high probability.

There are also fine-mapping methods that do not utilize the indicator vector approach and hence avoid having to deal with the vast model space. These shrinkage-based methods do not perform formal variable selection but instead rely on shrinking non-causal SNP effect sizes to values very close to zero. The posterior credible interval or some statistic derived from it is then used to rank the SNPs or weigh up the posterior evidence for the effect size being non-zero. Boggis et al. [2016] successfully used this kind of approach in eQTL mapping with a continuous response and Pereira et al. [2017] used it in fine-mapping with a binary response. We adopt the framework of these

shrinkage based approaches in this paper.

There have also been many methods suggested to include functional genomic information in fine-mapping studies. Some are designed for scenarios with at most a single SNP (Spencer et al. [2016]; Pickrell [2014]) whilst most now allow the presence of multiple causal SNPs by modelling the joint effects of SNPs (Pereira et al. [2017]; Kichaev et al. [2014a]; Quintana and Conti [2013]). There are often restrictions on the type of functional genomic information that can be incorporated in the model and analysis; Kichaev et al. [2014a] and Pereira et al. [2017] for example assume that the annotations are binary. We chose to use functional significance (FS) scores [Lee and Shatkay, 2009] to inform our effect size prior. These are normalised scores (FS $\in$ [0, 1]) which combine information from 16 publicly available databases. The FS score relating to a SNP is a measure of the deleterious effect of that particular SNP, as measured by these 16 different sources.

To the best of our knowledge the only approach to include functional genomic information that has dealt with the issue of partially observed functional information is Spencer et al. [2016] who simply replaced any unobserved value with 0 which was the greatest lower bound of the observed values for each annotation. This lack of consideration of partially observed functional data is surprising given the potentially sparse functional information that is available for many SNPs. It may be that authors are carefully choosing the type of functional information to include to avoid this problem. If the annotations used to inform priors are those chosen for completeness of data rather than disease-specific informativeness then there could be a possible loss of valuable information.

In this paper we propose a differential shrinkage approach to fine mapping that uses the normal gamma prior [Griffin et al., 2010] to induce shrinkage. We utilise 148 effect size estimates from breast cancer GWAS to calibrate the prior shrinkage and outline a coherent approach to incorporating partially observed functional information in the multi-causal SNP setting. We employ a fully Bayesian approach to address the uncertainly in the posterior parameter estimates. We assume that genotype data is available but our method could readily be extended to include GWAS summary statistics, as is done in Benner et al. [2016], Kichaev et al. [2014a] and Chen et al. [2015]. This approach could be applied to any univariate partially observed grouped prior information, such as a single categorical variable from the ENCODE database [ENCODE, 2011]. We apply our differential shrinkage method to fine mapping data in the CASP8 region between base positions 201,500,074 and 202,569,992. The data we used is from the iCOGS array developed by the Collaborative

Oncological Gene-Environment Study (COGS) Consortium [Michailidou et al., 2013]. The data consists of 1733 variants (501 genotyped, 1232 imputed using Impute2 [Marchini and Howie, 2010]) on 46,450 breast cancer cases and 42,600 controls.

Materials and Methods

Asymptotic Gaussian likelihood

Since we have case-control data the natural risk model is based on a generalised linear model (GLM) with a logit link function. Let $\boldsymbol{\beta}$ represent a p-vector of the regression coefficients in our logistic model and α represent the intercept. With $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$ we use the asymptotic approximation that $\hat{\boldsymbol{\phi}} \mid \boldsymbol{\phi} \sim \mathrm{N}(\boldsymbol{\phi}, \boldsymbol{V})$, where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimate of $\boldsymbol{\phi}$ and $\boldsymbol{V}$ is the variance-covariance matrix estimated from a GLM that includes all the SNPs. This means that that we can replace our logistic likelihood for the binary response with a Gaussian one for the maximum likelihood estimates (MLEs), which makes the MCMC faster and easier to implement. In order to avoid problems with exact multi-collinearity we remove SNPs (columns of X) that have identical genotypes to another SNP so that X is a full-rank matrix.

Normal-gamma prior

Consider our asymptotic approximation for the MLEs

$$\hat{\boldsymbol{\phi}} \mid \boldsymbol{\phi} \sim N(\boldsymbol{\phi}, \boldsymbol{V}), \tag{1}$$

with $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$. The scale mixture of normals (see for example Andrews and Mallows [1974], West [1987]) is a convenient representation of the prior distribution for the regression coefficients ($\boldsymbol{\beta}$) needed in a Bayesian framework. The prior distribution for β_i (the ith element of $\boldsymbol{\beta}$) can be expressed as

$$\pi(\beta_i) = \int N(\beta_i \mid 0, \psi_i)\, dF(\psi_i) \quad \text{for} \quad i = 1, ..., p, \tag{2}$$

where $F(.)$ represents a mixing distribution and ψ_i is its idiosyncratic prior variance. The random variables representing the prior effect sizes of different SNPs are assumed to be mutually

independent. Griffin et al. [2010] proposed

$$\beta_i|\psi_i \sim N\left(0, \psi_i\right) \text{ and } \psi_i \sim Ga(.), \tag{3}$$

so that each β_i has a potentially different variance ψ_i a priori. This structure gives us the so-called normal-gamma (NG) prior. Specifically, the prior for ψ_i is

$$\pi\left(\psi_i|\lambda, \gamma^2\right) \sim Ga\left(\lambda, \frac{1}{2\gamma^2}\right), \tag{4}$$

where λ and γ^2 are hyperparameters. This allows sufficient flexibility in the effect size variances. This prior seems appropriate for fine-mapping where most coefficients are zero or close to zero but we want to allow for the possibility of large effect sizes. The law of total variance using Equations (3) and (4) gives $\text{var}\left(\beta_i|\lambda, \gamma^2\right) = 2\lambda\gamma^2$. Putting a suitable inverse gamma distribution on the prior variance ($2\lambda\gamma^2 \sim IG\left(2, M\right)$) gives $\mathbb{E}\left(\text{var}\left(\beta_i|\lambda, \gamma^2\right)\right) = \mathbb{E}\left(2\lambda\gamma^2\right) = M$. M should be suitably chosen to reflect the effect sizes expected in fine-mapping studies.

Prior structure

Griffin et al. [2010] suggested a hierarchical form of prior that shrinks larger effect sizes less than smaller effect sizes (differential shrinkage). We modify this prior as follows

$$\begin{aligned}
\alpha &\sim N\left(0, 0.1\right), && (5)\\
\beta_i \mid \psi_i &\sim N\left(0, \psi_i\right), && (6)\\
\psi_i \mid \lambda, \gamma^{-2} &\sim Ga\left(\lambda, \frac{1}{2\gamma^2}\right), && (7)\\
\gamma^{-2} \mid \lambda &\sim Ga\left(2, \frac{M}{2\lambda}\right), && (8)\\
\lambda &\sim \text{Ex}(\kappa), && (9)
\end{aligned}$$

where the prior of $\gamma^{-2}|\lambda$ comes from $2\lambda\gamma^2 \sim IG\left(2, M\right)$.

Estimating the hyper parameter κ

Our hierarchical prior model has $\lambda \sim \text{Exp}(\kappa)$. Since our particular interest is in breast cancer, we estimate κ using the log-odds ratios from GWAS top hits in breast cancer (Fachal and Dunning [2015]). Because the top hits effect sizes are only GWAS estimates we decided to partition them into five groups to reduce sensitivity to individual values and find the value of κ that 'best fits', in a particular sense, the empirical cumulative distribution function (ecdf). Because of the approximate symmetry in the ecdf of the log-odds ratio we chose to use only odds ratios greater than one; specifically we used $\{\beta_1, \beta_2, \beta_3, \beta_4\} = \{\log(1.05), \log(1.08), \log(1.10), \log(1.20)\}$. We find the optimal value of κ (κ^*) as

$$\kappa^* = \underset{\kappa}{\text{argmin}} \left\{ \sum_{i=1}^{4} \left(F_i(\beta_i) - \hat{F}_i(\beta_i) \right)^2 \right\} \tag{10}$$

where $F_i(\beta_i)$ is the cumulative distribution function (cdf) of a random variable with a NG distribution and $\widehat{F}_i(\beta_i)$ is the ecdf, both evaluated at β_i. In addition to the 148 breast cancer top hits, Michailidou et al. [2013] estimated there were approximately an additional 1000 yet-to-be-discovered causal SNPs in breast cancer (although the true number is not known with much certainty). In calculating $\widehat{F}_i(\beta_i)$ we took account of the number of yet-to-be-discovered causal SNPs. We assumed all yet-to-be-discovered causal SNPs had $|\log(\text{OR})| < \log(1.05)$ since there is little power to detect odds ratios in this range for GWAS sample sizes currently used. Because there is no closed form for the cdf of a random variable with a NG distribution we used Sequential Monte Carlo for each value of κ considered (from 0.5 to 200 in increments of 0.5). Because of the uncertainty in the number of yet-to-be-discovered SNPs we considered four possible values (1000, 500, 200 or 100) which gave κ in the range $13 < \kappa^* < 143$. The number of yet-to-be-discovered SNPs that gives the best fit to the ecdf (as measured by our criterion) is 1000 which is consistent with the results of [Michailidou et al., 2013]. We assume this value in the rest of the paper and set $\kappa = 143$ which is the estimate corresponding to 1000 yet-to-be-discovered SNPs.

Incorporating Partially Observed Prior Functional Information in the Prior

Griffin et al. [2010] proposed a prior for the hyper-parameters λ and $\gamma^{-2}|\lambda$ to provide variability

around the Lasso prior ($\lambda = 1$). They used a single parameter for λ and γ^{-2} for all predictors. We modify the structure of the NG prior by allowing λ and γ^{-2} to take four different values based on four different classes of the FS scores [Lee and Shatkay, 2009]. FS scores naturally divide into four groups: FS > 0.5, FS $= 0.5$, FS < 0.5, and missing FS (FS = NA). SNPs with FS > 0.5 are most likely to be deleterious, SNPs with FS < 0.5 are unlikely to be deleterious, the evidence for SNPs with FS $= 0.5$ is unclear, and the deleterious effects of SNPs with missing FS scores (FS=NA) are unknown.

We modify the distribution of $\mathrm{Var}\left(\beta_i|\lambda,\gamma^2\right)\left(=2\lambda\gamma^2\right)$ for each group as follows

$$\text{Group 1 (G1); FS} > 0.5 : 2\lambda_1\gamma_1^2 \sim IG\left(2, M_1\right), \tag{11}$$

$$\text{Group 2 (G2); FS} = 0.5 : 2\lambda_2\gamma_2^2 \mid w \sim w \times IG\left(2, M_1\right) + (1-w) \times IG\left(2, M_2\right), \tag{12}$$

$$\text{Group 3 (G3); FS} < 0.5 : 2\lambda_3\gamma_3^2 \sim IG\left(2, M_2\right), \tag{13}$$

$$\text{Group 4 (G4); FS} = \text{NA} : 2\lambda_4\gamma_4^2 \mid h \sim h \times IG\left(2, M_1\right) + (1-h) \times IG\left(2, M_2\right), \tag{14}$$

where $M_2 < M_1$ are constants to be determined and w and h are the mixture parameters for Groups 2 and 4 and are assigned Beta(2, 2) and Beta(1, 4) distributions respectively. Beta distributions are used for computational convenience. This modification ensures that

$$\underset{i\in\mathrm{G1}}{\mathbb{E}}\left(\mathrm{Var}\left(\beta_i|\lambda_1,\gamma_1^2\right)\right) > \underset{i\in\mathrm{G2}}{\mathbb{E}}\left(\mathrm{Var}\left(\beta_i|\lambda_2,\gamma_2^2\right)\right) > \underset{i\in\mathrm{G4}}{\mathbb{E}}\left(\mathrm{Var}\left(\beta_i|\lambda_4,\gamma_4^2\right)\right) > \underset{i\in\mathrm{G3}}{\mathbb{E}}\left(\mathrm{Var}\left(\beta_i|\lambda_3,\gamma_3^2\right)\right)$$

since $M_1 > wM_1 + (1-w)M_2 > hM_1 + (1-h)M_2 > M_2$ for $w, h \in [0, 1]$ and $M_2 < M_1$. The idea of incorporating the FS score into the prior is to decrease the effect size shrinkage in Group 1 and 2 SNPs (high FS scores) relative to the effect size shrinkage of Group 3 SNPs (low FS scores) with the shrinkage of Group 4 SNP effect sizes somewhere in between. The aim is to select values of M_1 and M_2 that lead to appropriate differential shrinkage. Too much differential shrinkage means that the prior almost completely dictates the relative posterior effect sizes. With too much differential shrinkage, a causal SNP allocated to Group 3 or 4 would have little chance of having a posterior credible interval that didn't include zero.

We need to determine suitable choices of M_1 and M_2. Since the expectation of a random variable with an $IG(2, M)$ distribution is M we set M_1 to be the variance of the breast cancer top

hits, which was 0.01 approximately. This is the expectation of the prior variance for SNPs whose effect size we want to shrink by a relatively small amount. We considered various values of M_2 and assessed, in a univariate analysis, the relative shrinkage (under the two values of M) of the posterior expectation of the breast cancer top hits effect sizes. We calculate the relative univariate shrinkage factor (SF) as follows

$$\mathrm{SF} = 1 - \frac{\mathbb{E}_2\left(\beta \mid \hat{\beta}\right)}{\mathbb{E}_1\left(\beta \mid \hat{\beta}\right)}, \tag{15}$$

where $\mathbb{E}_i\left(\beta \mid \hat{\beta}\right)$ represents the posterior expectation of a parameter with a NG prior that has an expected prior variance of M_i. A value of SF close to 0 implies similar posterior expectations under both values of M_1 and M_2. Values close to 1 imply high levels of differential shrinkage. The posterior mean for SNP i can be approximated using

$$\mathbb{E}\left(\beta_i \mid \hat{\beta}_i\right) \approx \frac{\sum_k \beta_k f_R\left(R = \hat{\beta}_i\right)}{\sum_k f_R\left(R = \hat{\beta}_i\right)}. \tag{16}$$

where β_k is sampled from the NG prior using sequential Monte Carlo sampling, $R \sim N\left(\beta_k, \frac{1}{nf_i(1-f_i)}\right)$ [Slager S. L., 2001], n is the number of cases (or controls), and f_i is the minor allele frequency (MAF) of breast cancer top hit i.

Figure 1 shows the relative univariate shrinkage factor calculated using $M_1 = 0.01$ and $M_2 = 0.001$ with 16000 cases and 16000 controls . In Figure 1 more than two thirds of the effect sizes have SFs more than 0.1 and the SF does not exceed 0.5. We consider these SFs to be large enough to allow differential shrinkage but not so large that effect sizes in the high shrinkage groups will be shrunk too close to zero. Figure 2 shows the prior of the log-odds ratio (β) for $M_1 = 0.01$ and $M_2 = 0.001$ with λ fixed at its expected value of 1/148 and γ^2 fixed such that the expected prior variance $2\lambda\gamma^2$ is equal to M_1 or M_2.

Implementing the MCMC

We use MCMC to sample from the posterior distribution. Derivations of the full conditional distributions are given in the Appendix. We investigated possible posterior summaries (mean,

median, credible interval) to rank the SNPs and found that high probability marginal posterior credible intervals gave the best AUC in most scenarios considered. Therefore we only report results based on the marginal posterior credible interval. In order to derive ROC curves in simulated data when the causal SNP is known we need a quantitative, and preferably continuous, measure for each SNP to apply thresholds to. We use the maximum detection credible interval size defined as

$$1 - 2 \times \min \left\{ Pr\left(\beta_i \mid \hat{\beta}_i\right) > 0, Pr\left(\beta_i \mid \hat{\beta}_i\right) < 0 \right\}$$

where $\hat{\beta}_i$ is the maximum likelihood estimate of β_i. For the simulated data we ran the MCMC for 20,000 iterations with a burn-in of 2,000. Because of the highly correlated SNPs our chains showed high levels of autocorrelation so we thinned all our chains by 50 to remove the autocorrelation to an acceptable level. Standard convergence checks suggested the chains had converged.

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]. Genotypes for 16000 cases and 16000 controls were simulated. SNPs with MAF less than 0.01 were removed. This left around 280 SNPs in the fine mapping region. The simulation study was designed to have very high levels of correlation between the SNPs, odds ratio likely to be encountered in fine-mapping studies and large sample sizes as these are increasingly common as consortia pool data to increase statistical power. In all four simulation scenarios we simulate 2 causal SNPs. The odds ratio of the common and rarer causal SNPs is always 1.08 and 1.13 respectively. The scenarios simulate different patterns of linkage disequilibrium (LD) between the causal SNPs, MAFs of the causal SNPs. We aimed to choose causal SNPs that gave a nonimal marginal power of 80% for each causal SNP at the lower sample size. However sometimes it was not possible to identify two causal SNPs with the exact desired MAF and LD and so there is some variation in power around the target 80%. VanLiere and Rosenberg [2008] showed that if the MAFs of two SNPs satisfy $\mathrm{MAF}_1 > \mathrm{MAF}_2$ then the maximum r^2 value between two SNPs is given by

$$r^2_{\max} = \frac{(1 - \mathrm{MAF}_1) \times \mathrm{MAF}_2}{(1 - \mathrm{MAF}_2) \times \mathrm{MAF}_1}, \tag{17}$$

We report $r^2_{\max}$ along with the other simulation parameters in Table 1. We simulated 10 replicate data sets for each scenario considered.

We considered what effect the placement of the two causal SNPs into different prior groups would have. With two causal SNPs they are either both placed in the same FS score prior group or in different groups. When determining the percentages of SNPs in each of the four FS score functional groups we used the same percentages as those observed for the SNPs in the iCOGS data. These percentages are approximately 2, 5, 31 and 62% for the FS > 0.5, FS$= 0.5$, FS< 0.5 and FS=NA groups respectively.

We initially considered all four possible cases when placing the two causal SNPs in the same FS score prior group, that is both SNPs in each of the four FS score groups separately. However the results when both were in Group 1 were quite similar to those when they were in Group 2 and similarly for Groups 3 and 4. As a consequence, when considering both causal SNPs in the same group we only present results for the cases where both were in Group 1 (FS > 0.5) or both were in Group 3 (FS < 0.5). These two groups represent the extreme shrinkage groups.

When placing the two causal SNPs in different groups the number of possibilities is much higher than when they go into the same prior group so we present results only relating to Scenario 4 in Table 1. This is the moderate LD case. The four cases we considered for LD Scenario 4 are presented in Table 2. In the iCOGS data 62% of FS scores are missing so in all four scenarios of Table 2 we assign one of the two causal SNPs to the NA FS score group. In Scenarios 1, 2 and 4 the rare causal SNP resides in the NA FS score group whilst the common causal SNP resides in one of the remaining three groups. We also wanted to see what effect interchanging the groups of the rare and common causal SNPs would have, so Scenario 3 swaps the FS score groups of the causal SNPs specified in Scenario 1.

We want to compare our results to those obtained using other competitive approaches. Previous results [Benner et al., 2016] indicate that FINEMAP performs very well compared to other recently proposed multivariate approaches so we choose to compare the performance of our NG method only with that of FINEMAP, assuming that other methods are unlikely to perform substantially better. We used the default values of FINEMAP except that we allowed the maximum number of causal SNPS to be both two and five. Five is the default whilst two is the true number of SNPs in our simulations. We wanted to see whether FINEMAP would perform better when the maximum

number of SNPs allowed in FINEMAP was equal to the true number of SNPs in the simulations.

In the next section we present results for the scenarios in Tables 1 and 2. We also present results from an analysis of the iCOGS consortium [Michailidou et al., 2013] breast cancer case-control SNP data for a region of the CASPASE 8 gene using our methods and describe how the inclusion of functional information affects the SNPs selected.

The iCOGS data

The iCOGs array [Michailidou et al., 2013] was designed by the Collaborative Oncological Gene-environment Study (COGS) for fine-mapping studies. The array includes SNPs that are highly associated with breast, ovarian and prostate cancer. We focus on the region of Chromosome 2 between base positions 201500074 and 202569992 that includes the *CASP8* gene. We apply the methods on a total of 1733 SNPs with 46450 cases and 42600 controls. Those 1733 SNPs come from two sources: 501 SNPs are selected out of the 585 SNPs from the Breast Cancer Association Consortium that passed quality control checks, 1232 SNPs were imputed using IMPUTE2 [Marchini and Howie, 2010]. The SNPs were considered to be imputed successfully if their imputation accuracy exceeded 90% [Lin et al., 2015].

Results

Causal SNPs in the same FS score group

Figure 3 shows the results of using the standard NG, the NG with both causal SNPs in the FS>0.5 group, the NG with both causal SNPs in the FS<0.5 group and FINEMAP with both two and five as the maximum number of allowed causal SNPs. The results for all four scenarios in Table 1 are presented. The first observation is that the performance of FINEMAP is more variable than that of the NG. In Scenario 3, the high LD case, FINEMAP performs very poorly, especially when there are allowed to be up to five causal SNPs. In Scenario 4, the moderate LD case, FINEMAP has a lower area under the curve than the standard NG but reaches a TPR of 1 quicker than the NG. In Scenarios 1 and 2, the low LD case, the performance of FINEMAP improves substantially with at least one of the FINEMAP models giving the best performance. In Scenario 1 the choice of a maximum of two causal SNPs in FINEMAP is superior to all other methods. In Scenario 2

the choice of a maximum of five causal SNPs is competitive with the NG approaches (and reaches a TPR of one much quicker than the other approaches) but allowing now more than two causal SNPs leads to a very poor performance.

The results for the NG approaches all follow the same pattern. The standard NG (without using functional information) is outperformed by the case when both causal SNPs are placed in the FS<0.5 group, but the best performance comes from placing both causal SNPs in the FS>0.5 group. The observation that the standard NG performs less well than when both causal SNPs are placed in the FS<0.5 group (Group 3) is surprising. Group 3 is subject to more prior shrinkage a priori than the standard NG. Presumably the information in the likelihood here dominates the prior shrinkage. An obvious question at this stage is whether is a consequence of both causal SNPs residing in the same prior group.

Causal SNPs in different FS score group

Figure 4 shows the results of placing the two causal SNPs in different prior groups (see Table 2) for LD Scenario 4 (see Table 1). One of the lines in each sub-figure in Figure 4 represents the results from one of the scenarios in Table 2. In each figure we also show the results for the standard NG, for the modified NG when placing both causal SNPs in the FS>0.5 group (representing the best case scenario), and for FINEMAP with a maximum of two causal SNPs (since Figure 3(d) shows little difference in the FINEMAP ROC curve with a maximum of two or five causal SNPs allowed).

We see that in all cases placing the two causal SNPs in different prior groups gives a performance between the performance of the standard NG and the NG with both causal SNPs in Group 1. This seems particularly surprising in Scenario B since in this scenario the causal SNPs are in groups in which the prior variance is a mixture distribution (Groups 2 and 4) and there is only one causal SNP in each of these groups to inform the posterior of the mixture parameters w and h (see Equations (12) and (14)). It appears that allowing prior groups to have prior effect size variances that are mixture distributions may afford enough flexibility that even a single signal in a group can increase the posterior variance sufficiently (relative to groups with no causal SNPs) to boost its rank.

Detection of SNPs within an LD block

A major challenge in fine-mapping is identifying causal variants that are in very high LD with several non-causal SNPs. To investigate how using functional information might aid this identification we identified SNPs in high LD (with $r^2 \geq 0.8$) with a simulated causal SNP and allocated these high LD SNPs to specific functional groups. We used the simulation parameters in Scenario 4 in Table 1. The causal variant is SNP 46. We consider three different rather extreme scenarios of allocation to groups and compare it to the not using any functional information. The scenarios are shown in Table 3. In each scenario all SNPs in the LD block are allocated to either group 1 (FS > 0.5) or 4 (FS = NA). We show the mean rank of each SNP (rank 1 is high) averaged over 10 simulated datasets. In scenario LD1 no functional groups are used. In scenario LD2 (best case scenario) we place the causal SNP in group 1 and the other high LD SNPs in group 4. In scenario LD3 (worst case scenario) we place the non-causal SNP in highest LD with the causal SNP into group 1 and the other high LD SNPs (including the causal one) into group 4. In scenario LD4 we place the causal SNP and the SNP most correlated with it into group 1, and the rest of the high LD SNPs in group 4. In all scenarios the SNPs outside the LD block are randomly placed into FS score groups according to the distribution of the FS score groups observed in the iCOGS data.

Table 3 shows that the standard normal gamma gives a mean rank for the causal SNP of 12 (there are approximately 280 SNPs being ranked). In the best case scenario (LD2) this jumps to a mean rank of 4 and in the worst case scenario (LD3) barely drops compared to the standard normal gamma, yielding a mean rank of 14. In scenario LD4 where the causal SNP and the one most correlated with it are placed in the same functional group, the causal SNP rank is 6 compared to 41 for the SNP most correlated with it. All these scenarios indicate that grouping SNPs may lead to higher ranks for the causal SNP(s) even when they are placed in groups with high shrinkage.

Detection of SNPs in the iCOGS data

We compare the SNPs selected by the standard normal gamma (NG), the NG prior using FS score groups (NGFS), and FINEMAP. We compare the top 20 SNPs by rank for each method. We use the same value of $M_1 = 0.01$ and $M_2 = 0.001$ as used in our simulation experiment. In NGFS, each SNP is placed into its true FS score group. The results of the analysis are shown in Figure 5.

There is strong agreement between the NG and NGFS but only one SNP (rs2540050) is selected in the top 20 by all methods (ranked 8th, 5th and 15th in NG, NGFS and FINEMAP resperctively). This SNP has a MAF of 0.08 and the estimated effect size (odds ratio) is 1.017, derived from the FINEMAP output. It is interesting to note that rs2540050 was not in the list of likely candidate causal SNP in either of the two univariate analyses undertaken in this region previously (Spencer et al. [2015]; Spencer et al. [2016]) even though alternative functional information was used in the [Spencer et al., 2016] analysis.

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 have shown how prior categorical SNP functional genomic information can be used to inform an absolutely continuous prior in Bayesian fine mapping studies where some SNPs cannot be assigned to a prior category. This approach could be applied to any Bayesian variable selection problem where a single variable, which is either categorical or can be made categorical by binning, is to be used to assign explanatory variables to prior groups. It could also be applied to studies in which multiple variables are to be used to assign explanatory variables to prior groups provided each explanatory variable can be assigned to a single category according to some decision process.

Our approach is to specify an absolutely continuous prior on the effect size and exploit differential shrinkage to shrink the effect sizes of non-causal SNPs more than those of causal SNPs. It contrasts with the spike and slab approach which explicitly allows for a non-zero effect size in the prior spike. The spike and slab approach, whilst seemingly providing a more natural framework for model selection, comes with its own challenging computational issues: how to search over the potentially vast model space. The two approaches to this are either to enumerate over all models containing a small number of causal SNPS (Kichaev et al. [2014a]; Chen et al. [2015]) or to employ a stochastic search algorithm [Benner et al., 2016]. We have shown that using shrinkage alone,

without formal model selection via an indicator vector, does compete with these approaches in terms of causal SNP ranks and allows functional genomic information to be incorporated easily into the effect size prior.

This shrinkage-only based approach does however raise some issues that do not arise in the standard Bayesian variable selection approach. The first is the choice of the prior variances for the different shrinkage groups. This equates to the choice of M_1 and M_2 in our Inverse-Gamma distributions for the prior variances. More work needs to be done in using either expert elicitation or objective approaches to estimating these key quantities. For example, using genome wide information on the effect sizes of disease-specific validated SNPs stratified by FS score group would provide objective estimates of the prior variances and hence the values of M_1 and M_2. However, there are currently not enough disease-specific validated SNPs with available FS scores to be able to do this. An additional drawback of the shrinkage-only approach is that the MCMC can be slow on several thousand SNPs because of the matrix inversion needed in each iteration. However, using the Cholesky decomposition speeds up this process considerably. This inversion also necessitates the removal of perfectly correlated SNPs.

One of the obvious attractions of using functional genomic information is that in can be used to disentangle the true signal from correlated signals in a block of SNPs in high LD. We have shown that this is indeed the case when using the Normal Gamma prior with different functional groups. We demonstrate that even when a causal SNP is in a functional group with a priori low probability of being causal the signal does not diminish significantly compared to its correlated neighbours, and that the ranks of causal SNPs in functional groups with high a priori probability of being causal are clearly boosted when functional genomic information is used.

Acknowledgements

This work was carried out as part of a PhD project funded by the Northern Border University, Saudi Arabia. We would like to thank all studies in the Breast Cancer Association Consortium (BCAC) and the Collaborative Oncological Gene-environment Study (COGS) for the use of their data. The COGs study would not have been possible without the contributions of the following: Per Hall (COGS); Paul Pharoah, Manjeet K. Bolla, Qin Wang (BCAC), Andrew Berchuck (OCAC), Ros-

alind A. Eeles, Douglas F. Easton, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL), Georgia Chenevix-Trench, Antonis Antoniou, Lesley McGuffog, Fergus Couch and Ken Offit (CIMBA), Joe Dennis, Alison M. Dunning, Andrew Lee, and Ed Dicks, Craig Luccarini and the staff of the Centre for Genetic Epidemiology Laboratory, Javier Benitez, Anna Gonzalez-Neira and the staff of the CNIO genotyping unit, Jacques Simard and Daniel C. Tessier, Francois Bacot, Daniel Vincent, Sylvie LaBoissire and Frederic Robidoux and the staff of the McGill University and Gnome Qubec Innovation Centre, Stig E. Bojesen, Sune F. Nielsen, Borge G. Nordestgaard, and the staff of the Copenhagen DNA laboratory, and Julie M. Cunningham, Sharon A. Windebank, Christopher A. Hilker, Jeffrey Meyer and the staff of Mayo Clinic Genotyping Core Facility. Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The authors declare no conflicts of interest.

References

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *J Royal Stat Soc Series B*, 36(1):99–102.

Benner, C., Spencer, C. C. A., 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.

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.

ENCODE (2011). A user's guide to the encyclopedia of dna elements (encode). *PLoS biology*, 9(4):e1001046.

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.

Griffin, J. E., Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

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.

Kichaev, G., Yang, W-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014a). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*, 10(10):e1004722.

Lee, P. H. and Shatkay, H. (2009). An integrative scoring system for ranking snps by their potential deleterious effects. *Bioinformatics*, 25(8):1048–1055.

Lin, W-Y., Camp, N. J., Ghoussaini, M., Beesley, J., Michailidou, K., Hopper, J. L., Apicella, C., Southey, M. C., Stone, J., Schmidt, M. K., et al. (2015). Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Human Molecular Genetics*, 24(1):285–298.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 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.

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.

Quintana, M. A. and Conti, D. V. (2013). Integrative variable selection via bayesian model uncertainty. *Statistics in Medicine*, 32(28):4938–4953.

Slager S. L., Schaid, D. J. (2001). Case-control studies of genetic markers: power and sample size approximations for armitage's test for trend. *Hum Hered*, 52(3):149–153.

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.

VanLiere, J. M. and Rosenberg, N. A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theoretical population biology*, 74(1):130–137.

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

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.

Appendix: MCMC Full Conditional Distributions

In this section, we will calculate the full conditional distributions of each parameter. Let β_{ij} and ψ_{ij} represent the effect size and prior variance of the ith SNP in FS score group j respectively. Further let $\boldsymbol{\beta}$ represent the vector of all effect sizes across all groups, α represent the intercept and p_j represent the number of SNPs in Group j.

Full conditional distributions for α and $\boldsymbol{\beta}$

The full conditional distribution for $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$ is an application of the standard result for the conditional distribution of a random variable with a product of independent normals prior and multivariate normal likelihood (see for example Griffin et al. [2010]).

$$f(\boldsymbol{\phi} \mid \boldsymbol{\Lambda}, \hat{\boldsymbol{\phi}}) \sim \boldsymbol{N}\left(\left(\boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}}, \boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right), \tag{18}$$

where $\boldsymbol{\Lambda}^{-1} = \text{diag}\,(0.1, \psi_{11}, \ldots, \psi_{p_1 1}, \psi_{12}, \ldots, \psi_{p_2 2}, \psi_{13}, \ldots, \psi_{p_3 3}, \psi_{14}, \ldots, \psi_{p_4 4})$.

Full conditional distributions for ψ_{ij}

The full conditional distribution for ψ_{ij} is conditional on the FS score group-specific values of λ_j and γ_j^{-2} The full conditional comes from the product of prior distributions in Equations (6) and (7) as follows

$$\begin{aligned} f\left(\psi_{ij} \mid \beta_{ij}, \lambda_j, \gamma_j^{-2}\right) &\propto \frac{1}{\sqrt{\psi_{ij}}} \exp\left(-\frac{\beta_{ij}^2}{2\psi_{ij}}\right) (\psi_{ij})^{\lambda_j - 1} \exp\left(-\frac{\psi_{ij}}{2\gamma_j^2}\right) && (19) \\ &\propto (\psi_{ij})^{(\lambda_j - \frac{1}{2}) - 1} \exp\left\{-\frac{1}{2}\left(\gamma_j^{-2}\psi_{ij} + \frac{\beta_{ij}^2}{2\psi_{ij}}\right)\right\}. && (20) \end{aligned}$$

The expression is the kernel of a generalised inverse Gaussian (GIG) distribution with $\lambda_j - \frac{1}{2}, \gamma_j^{-2}$ and β_{ij}^2.

Full conditional distributions for λ_j

We show in detail how to derive the full conditional distribution for λ_1. The remaining full conditional distributions for λ_2, λ_3 and λ_4 are similar. Because of the multi-modal posterior we follow

the approach of Griffin et al. [2010] and ensure the same prior variance is used in the Metropolis-Hastings ratio. As a result we need to consider all terms in the joint distribution containing either λ_1 or γ_1^{-2}. We use Equations (7) - (9) to obtain

$$f\left(\lambda_1 \mid \psi_{11}, \ldots, \psi_{p_1 1}, \gamma_1^{-2}\right) \propto$$
$$\frac{\left(\frac{1}{2\gamma_1^2}\right)^{p_1\lambda_1}}{\Gamma(\lambda_1)^{p_1}} \left(\prod_{i=1}^{p_1}\psi_{i1}\right)^{\lambda_1-1} \exp\left(-\sum_{i=1}^{p_1}\frac{\psi_{i1}}{2\gamma_1^2}\right)\left(\frac{1}{2\lambda_1\gamma_1^2}\right)\left(\frac{1}{\lambda_1}\right)\exp\left(-\frac{M_1}{2\lambda_1}\gamma_1^{-2}\right)\pi\left(\lambda_1\right).$$

With $2\lambda_1'\gamma_1'^2 = 2\lambda_1\gamma_1^2$ we derive the acceptance probability for λ_1' as

$$\min\left\{1, \frac{\lambda_1\pi\left(\lambda_1'\right)}{\lambda_1'\pi\left(\lambda_1\right)}\left(\frac{\Gamma\left(\lambda_1\right)}{\Gamma\left(\lambda_1'\right)}\right)^{p_1}\exp\left(\left(\frac{1}{2\gamma_1^2}-\frac{1}{2\gamma_1'^2}\right)\sum_{i=1}^{p_1}\psi_{i1}\right)\frac{\left(2\gamma_1^2\right)^{p_1\lambda_1}}{\left(2\gamma_1'^2\right)^{p_1\lambda_1'}}\left(\prod_{i=1}^{p_1}\psi_{i1}\right)^{\lambda_1'-\lambda_1}\right\}. \quad (21)$$

An identical approach for $j = 3$ gives the same as Equation (21) with λ_1, γ_1^{-2}, ψ_{i1} and p_1 replaced with λ_3, γ_3^{-2}, ψ_{i3} and p_3. For $j = 2$ we use Equation (12) to derive

$$\gamma_2^{-2} \mid \lambda_2, w \sim wGa\left(2, \frac{M_1}{2\lambda_2}\right) + (1-w)Ga\left(2, \frac{M_2}{2\lambda_2}\right)$$

Applying the same approach to that used in deriving Equation (21) we find, after some algebra, that with $2\lambda_1'\gamma_1'^2 = 2\lambda_1\gamma_1^2$ the acceptance probability is the same as Equation (21) with λ_1, γ_1^{-2}, ψ_{i1} and p_1 replaced with λ_2, γ_2^{-2}, ψ_{i2} and p_2. For $j = 4$ we simply replace λ_1, γ_1^{-2}, ψ_{i1} and p_1 with λ_4, γ_4^{-2}, ψ_{i4} and p_4 in Equation (21).

Full conditional distributions for γ_j^{-2}

Let $S_j = p_j\lambda_j + 2$ and $A_{jk} = \frac{1}{2}\left(\frac{M_k}{\lambda_j} + \sum_{i=1}^{p_j}\psi_{ij}\right)$. The full conditional distribution of γ_1^{-2} is

$$\begin{aligned} f\left(\gamma_1^{-2} \mid \psi_{11}, \ldots, \psi_{p_1 1}, \lambda_1\right) &\propto \left(\gamma_1^{-2}\right)^{p_1\lambda_1}\exp\left(-\frac{\sum_{i=1}^{p_1}\psi_{i1}}{2}\gamma_1^{-2}\right)\times\gamma_1^{-2}\exp\left(-\frac{M_1}{2\lambda_1}\gamma_1^{-2}\right) \\ &\propto \left(\gamma_1^{-2}\right)^{S_1-1}\exp\left(-A_{11}\gamma_1^{-2}\right). \end{aligned} \quad (22)$$

which is the kernel of a gamma distribution with shape S_1 and rate A_{11}. A similar argument for $j = 3$ gives the full conditional distribution of γ_3^{-2} as a gamma distribution with shape S_3 and scale

A_{32}. For $(j=2)$ the full conditional distribution can be calculated as follows

$$\begin{aligned}
f\left(\gamma_2^{-2} \mid \ldots\right) &\propto \left(\gamma_2^{-2}\right)^{p_2\lambda_2} \exp\left(-\frac{\sum_{i=1}^{p_2}\psi_{i2}}{2}\gamma_2^{-2}\right)\left\{wM_1^2\gamma_2^{-2}\exp\left(-\frac{M_1}{2\lambda_2}\gamma_2^{-2}\right)\right\} \\
&+ \left(\gamma_2^{-2}\right)^{p_2\lambda_2} \exp\left(-\frac{\sum_{i=1}^{p_2}\psi_{i2}}{2}\gamma_2^{-2}\right)\left\{(1-w)\,M_2^2\gamma_2^{-2}\exp\left(-\frac{M_2}{2\lambda_2}\gamma_2^{-2}\right)\right\} \\
&\propto wM_1^2\left(\gamma_2^{-2}\right)^{S_2-1}\exp\left(A_{21}\gamma_2^{-2}\right) + (1-w)\,M_2^2\left(\gamma_2^{-2}\right)^{S_2-1}\exp\left(A_{22}\gamma_2^{-2}\right) \\
&\propto \frac{wM_1^2}{A_{21}^{S_2}}\left[\frac{A_{21}^{S_2}}{\Gamma\left(S_2\right)}\left(\gamma_2^{-2}\right)^{S_2-1}\exp\left(A_{21}\gamma_2^{-2}\right)\right] \\
&+ \frac{(1-w)\,M_2^2}{A_{22}^{S_2}}\left[\frac{A_{22}^{S_2}}{\Gamma\left(S_2\right)}\left(\gamma_2^{-2}\right)^{S_2-1}\exp\left(A_{22}\gamma_2^{-2}\right)\right]
\end{aligned}$$

So it follows that

$$\begin{aligned}
\gamma_2^{-2} \mid \psi_{i2}, \lambda_2, w &\sim A_{22}^{S_2}wM_1^2\; Ga\left(S_2, A_{21}\right) + A_{21}^{S_2}\left(1-w\right)M_2^2\; Ga\left(S_2, A_{22}\right) \\
&\sim \Delta_1\; Ga\left(S_2, A_{21}\right) + \left(1-\Delta_1\right)\; Ga\left(S_2, A_{22}\right),
\end{aligned}$$

where $\Delta_1 = \left(\frac{A_{22}^{S_2}wM_1^2 + A_{21}^{S_2}(1-w)M_2^2}{A_{22}^{S_2}wM_1^2}\right)^{-1} = \left(1 + \left(\frac{1-w}{w}\right)\left(\frac{M_2}{M_1}\right)^2\left(\frac{A_{21}}{A_{22}}\right)^{S_2}\right)^{-1}$.

An identical approach gives $\gamma_4^{-2} \mid \psi_{i4}, \lambda_4, h \sim \Delta_2\; Ga\left(S_4, A_{41}\right) + \left(1-\Delta_2\right)\; Ga\left(S_4, A_{42}\right)$, where

$\Delta_2 = \left(1 + \left(\frac{1-h}{h}\right)\left(\frac{M_2}{M_1}\right)^2\left(\frac{A_{41}}{A_{42}}\right)^{S_4}\right)^{-1}$.

Full conditional distributions for w and h

Let E_k represent the probability density $f_{X_k|\lambda_2}\left(X_k = \gamma_2^{-2}\right)$, where $X_k|\lambda_2 \sim Ga\left(2, \frac{M_k}{2\lambda_2}\right)$ and G_k represent the probability density $f_{Y_k|\lambda_4}\left(Y_k = \gamma_4^{-2}\right)$, where $Y_k|\lambda_4 \sim Ga\left(2, \frac{M_k}{2\lambda_4}\right)$. Then

$$\begin{aligned}
f\left(w \mid \lambda_2, \gamma_2^{-2}\right) &= \Big[wE_1 + (1-w)\,E_2\Big]\pi\left(w\right) \\
&\propto w^{3-1}\left(1-w\right)^{2-1}E_1 + w^{2-1}\left(1-w\right)^{3-1}E_2
\end{aligned}$$

Let $\Delta = E_1/\left(E_1+E_2\right)$, then the full conditional distribution for w is given by

$$w \mid \lambda_2, \gamma_2^{-2} \sim \Delta\,\mathrm{Beta}\left(3,2\right) + \left(1-\Delta\right)\mathrm{Beta}\left(2,3\right).$$

With $\Omega = G_1/\left(G_1 + 4G_2\right)$, a similar argument gives

$$h \mid \lambda_4, \gamma_4^{-2} \sim \Omega \operatorname{Beta}(2,4) + (1-\Omega) \operatorname{Beta}(1,5),$$

Figure Legends

Figure 1: Histogram of the relative univariate shrinkage factors of the breast cancer top hits calculated using $M_1 = 0.01$ and $M_2 = 0.001$ with 16,000 cases and 16,000 controls.

Figure 2: The log prior $\pi(\beta)$ for $M_1 = 0.01$ and $M_2 = 0.001$ with $\lambda = 1/148$ and γ^2 fixed such that the expected prior variance $2\lambda\gamma^2$ is equal to M_1 or M_2.

Figure 3 main: ROC curves for FINEMAP (with both two and five as the maximum number of allowed causal SNPs) and approaches using the maximum posterior credible interval size with a NG prior hierarchy. Both the standard NG (with $M = 0.01$) and the NG prior with different prior shrinkage groups ($M_1 = 0.01$ and $M_2 = 0.001$) are shown. The methods are applied to 10 simulated datasets from Hapgen2 with the two causal SNPs having odds ratios of 1.08 and 1.13. The MAFs and LD between the causal SNPs are given in Table 1. In the modified NG prior approach both causal SNPs are always in the same functional group; either both in Group1 (FS > 0.5) or Group3 (FS < 0.5). Each dataset has 16000 cases and 16000 controls. The numbers of SNPs in groups 1 to 4 are 8, 15, 89, 170 respectively.

Figure 4 main: ROC curves for FINEMAP (with both two and five as the maximum number of allowed causal SNPs) and approaches using the maximum posterior credible interval size with a NG prior hierarchy. Both the standard NG (with $M = 0.01$) and the NG prior with different prior shrinkage groups ($M_1 = 0.01$ and $M_2 = 0.001$) are shown. The methods are applied to 10 simulated datasets from Hapgen2 with the two causal SNPs having odds ratios of 1.08 and 1.13. Each dataset has 16000 cases and 16000 controls. The r^2 value is 34% of its maximum possible based on the allele frequencies. The two causal SNPs are either placed in the same FS scores group or in different FS scores groups. The numbers of SNPs in FS score groups 1 to 4 are 6, 15, 89 and 177 respectively.

Figure 5: A venn diagram showing the overlap of the top 20 SNPs selected in an analysis of the iCOGs data using the standard NG prior (NG), the NG prior using FS score groups (NGFS), and FINEMAP. In NGFS, each SNP is placed into its true FS score group. The iCOGs data has 1733 genotpyed and imputed SNPs and 46450 cases and 42600 controls.

	Scenario			
	1	2	3	4
$r^2/r^2_{\max}$ between causal SNPs	0.03	8×10^{-5}	0.66	0.34
Causal SNP 1 MAF	0.28	0.3	0.28	0.31
Causal SNP 2 MAF	0.09	0.06	0.07	0.1
Causal SNP 1 power	0.77	0.8	0.77	0.82
Causal SNP 2 power	0.79	0.4	0.55	0.87

Table 1. Simulation parameters for the two causal SNPs. Genotypes were simulated for 16000 cases and 16000 controls. The odds ratios of the common (SNP 1) and rarer (SNP 2) causal SNPs are 1.08 and 1.13 respectively.

Scenario	Common Causal SNP Group	Rare Causal SNP Group
A	Group 1 (FS > 0.5)	Group 4 (FS $= NA$)
B	Group 2 (FS $= 0.5$)	Group 4 (FS $= NA$)
C	Group 4 (FS $= NA$)	Group 1 (FS > 0.5)
D	Group 3 (FS < 0.5)	Group 4 (FS $= NA$)

Table 2. Four Scenarios for placing the common and rare causal SNPs into different FS score groups.

Scenario	SNP	25	30	34	46	47	48	58	65	71
	r^2 with SNP 46	0.83	0.81	0.83	1	0.9	0.84	0.8	0.82	0.82
LD1	Mean Rank	172	112	132	12	76	98	215	158	164
LD2	Functional Group	4	4	4	1	4	4	4	4	4
	Mean Rank	132	96	110	4	65	91	198	207	218
LD3	Functional Group	4	4	4	4	1	4	4	4	4
	Mean Rank	134	90	113	14	44	101	204	195	196
LD4	Functional Group	4	4	4	1	1	4	4	4	1
	Mean Rank	154	110	129	6	41	113	206	140	193

Table 3. The mean SNP rank over 10 simulated datasets for Scenario 4 in Table 1 for SNPs in high LD ($r^2 \geq 0.8$) with the causal SNP (SNP 46). SNPs in the LD block are placed in different functional groups in each scenario. In scenario LD1 no functional groups are used.

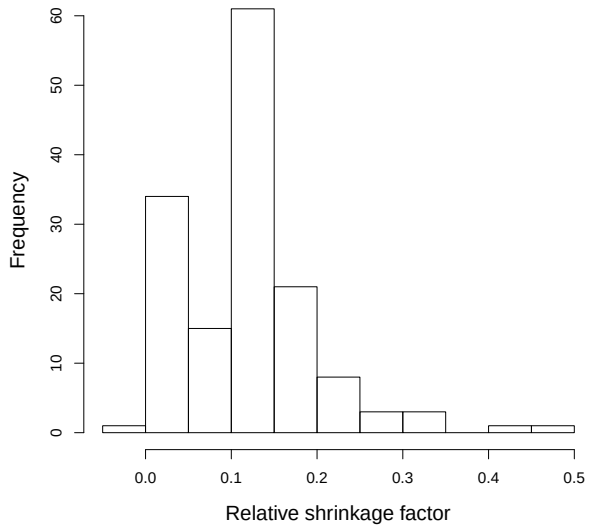


Figure 1

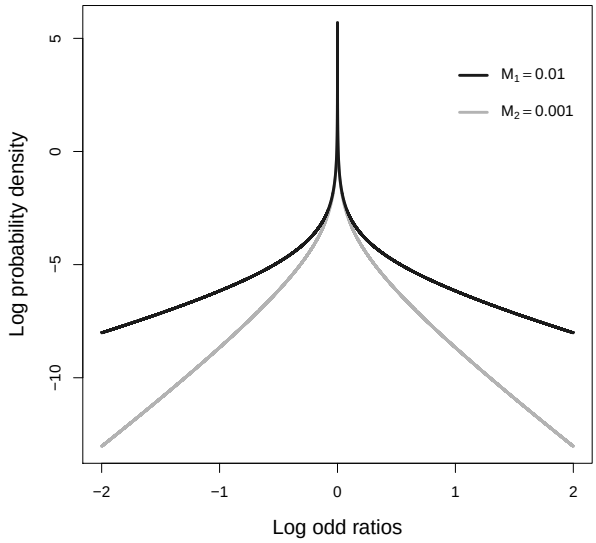
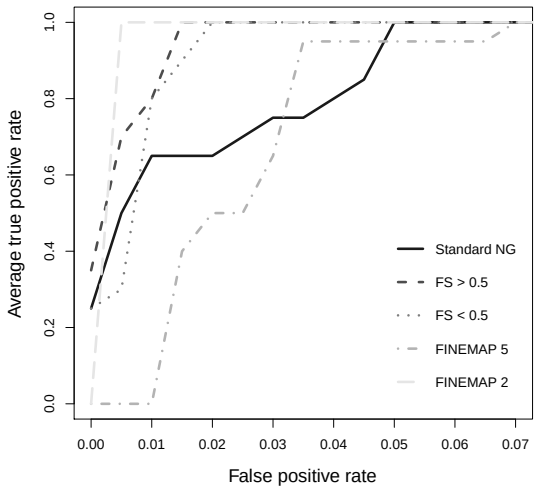
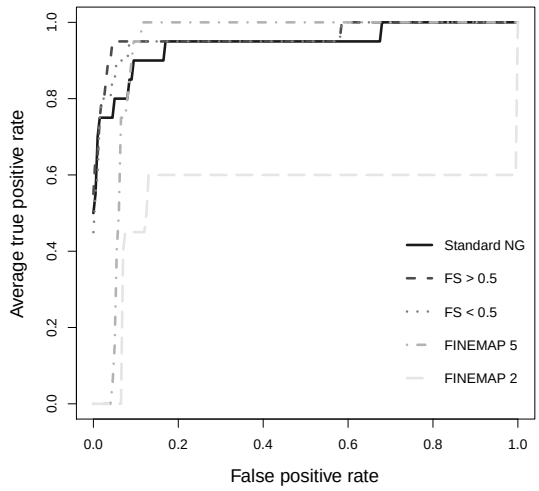

Log probability density
5
0
−5
−10
−2
−1
0
1
2
Log odd ratios
M1 = 0.01
M2 = 0.001

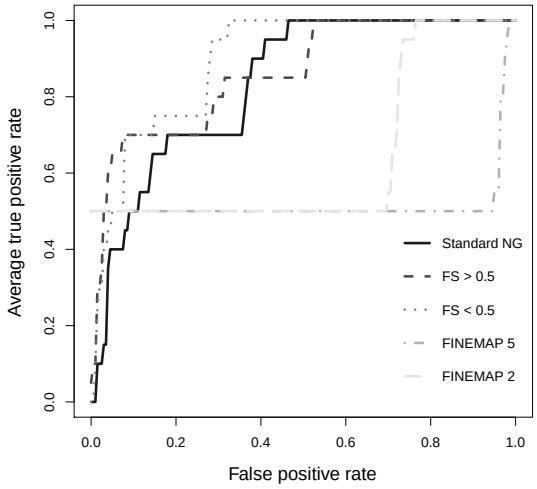
Figure 2



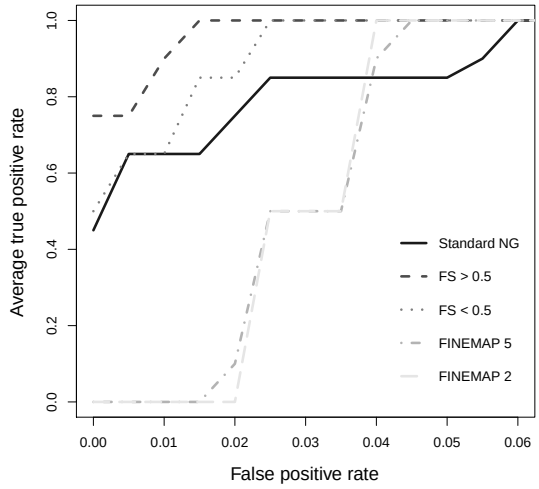
(a) Scenario 1 (Low LD).



(b) Scenario 2 (Low LD).

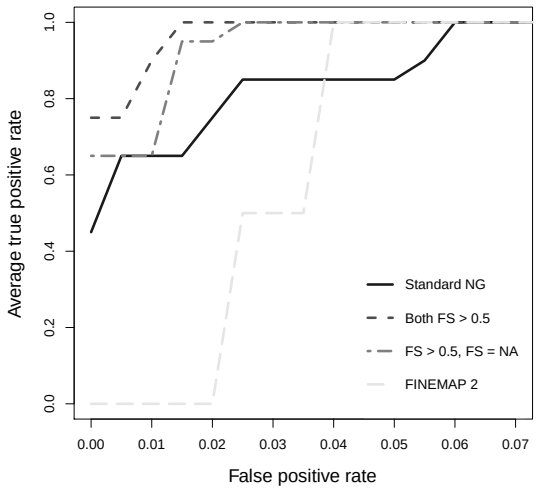


(c) Scenario 3 (High LD).

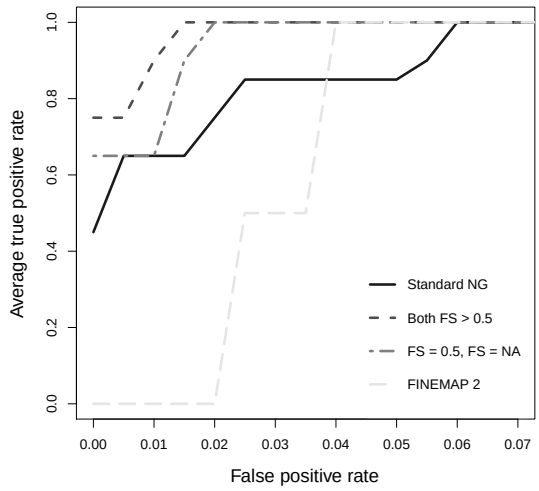


(d) Scenario 4 (Moderate LD).

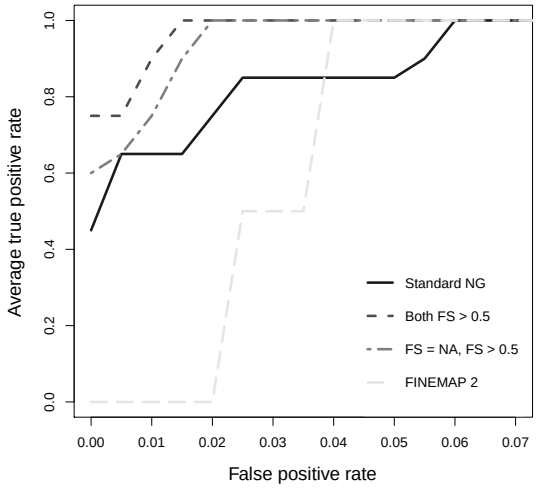
Figure 3



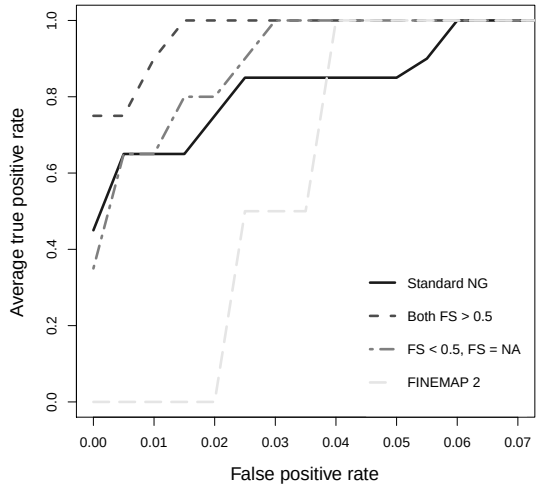
(a) Scenario A



(b) Scenario B



(c) Scenario C



(d) Scenario D

Figure 4

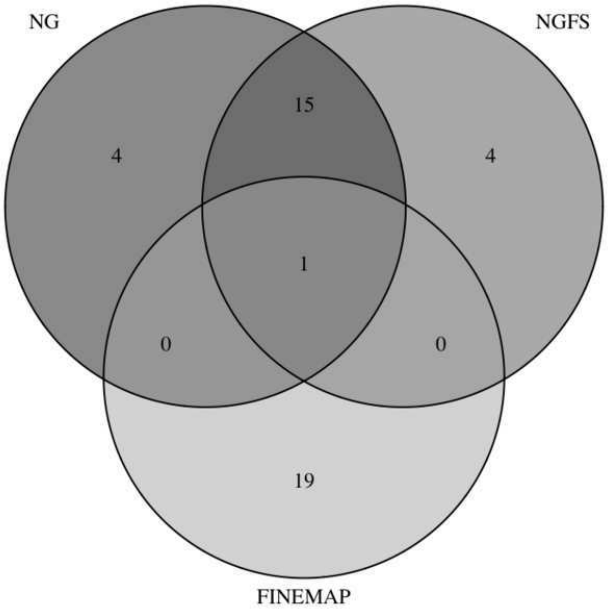


Figure 5