Inclusion of biological knowledge in a Bayesian shrinkage model for joint estimation of SNP effects

Miguel Pereira,[1] John R. Thompson,[2] Christian X. Weichenberger,[3] Duncan C. Thomas[4] and Cosetta Minelli[1]

[1]National Heart and Lung Institute, Imperial College London, London, United Kingdom

[2]Department of Health Sciences, University of Leicester, Leicester, United Kingdom

[3]Center for Biomedicine, European Academy of Bolzano/Bozen (EURAC), Bolzano, Italy (Affiliated to the University of Lübeck, Lübeck, Germany)

[4]Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, University of Southern California

Running title: Biological control of shrinkage

Correspondence to: Miguel Pereira, National Heart and Lung Institute, Imperial College London, Emmanuel Kaye Building, 1b Manresa Road, SW3 6LR London, United Kingdom; E-mail: miguel.pereira14@imperial.ac.uk; Tel: 0207-594-7938

Abstract

With the aim of improving detection of novel single-nucleotide polymorphisms (SNPs) in genetic association studies, we propose a method of including prior biological information in a Bayesian shrinkage model that jointly estimates SNP effects. We assume that the SNP effects follow a normal distribution centered at zero with variance controlled by a shrinkage hyperparameter. We use biological information to define the amount of shrinkage applied on the SNP effects distribution, so that the effects of SNPs with more biological support are less shrunk towards zero, thus being more likely detected.

The performance of the method was tested in a simulation study (1000 datasets, 500 subjects with ~200SNPs in 10 linkage disequilibrium (LD) blocks) using a continuous and a binary outcome. It was further tested in an empirical example on body mass index (continuous) and overweight (binary) in a dataset of 1,829 subjects and 2,614 SNPs from 30 blocks. Biological knowledge was retrieved using the bioinformatics tool Dintor which queried various databases.

The joint Bayesian model with inclusion of prior information outperformed the standard analysis: in the simulation study, the mean ranking of the true LD block was 2.8 for the Bayesian model vs. 3.6 for the standard analysis of individual SNPs; in the empirical example, the mean ranking of the six true blocks was 8.5 vs. 9.3 in the standard analysis. These results suggest that our method is more powerful than the standard analysis. We expect its performance to improve further as more biological information about SNPs becomes available.

Keywords: genetic association studies; prior knowledge; Bayesian model; shrinkage

Introduction

Genome-wide association (GWA) studies have identified many single-nucleotide polymorphisms (SNPs) associated with complex diseases and traits (Visscher, Brown, McCarthy, & Yang, 2012). However, the estimated effect sizes of these variants are frequently small and, overall, the identified SNPs explain only a small proportion of the predicted heritability of most diseases (Manolio, 2010).

In classical GWA analyses, SNP effects are estimated individually and their p-values are adjusted for multiple testing since thousands or millions variants are studied. SNPs are typically analyzed individually (hereafter referred to as 'standard analysis') because of the $p >> n$ problem: the number of parameters, p, far exceeds the number of individuals, n, which precludes the joint analysis of all SNPs in a standard regression model. Notwithstanding this, for complex traits with many associated genetic variants, the simultaneous analysis of all SNPs would be more powerful and provide better estimates because it combines the information across multiple variants (Cho et al., 2010). A variety of penalization (or regularization) methods has been proposed to address the issue of having more variables than observations (Hastie, Tibshirani, & Friedman, 2009). These methods impose constrains on the regression coefficients by means of a penalized objective function subject to a tuning parameter that "shrinks" the coefficients towards zero while allowing their estimation. However, these methods are usually computationally intensive when applied to high dimensional settings, which is the case of GWA studies.

Another way of improving the yield of genetic variants identified is to incorporate prior knowledge about the SNPs into the analysis of GWA studies (Cantor, Lange, & Sinsheimer, 2010). The rationale behind this method is that SNP selection can be further improved by giving more weight to SNPs that have a biological role, while penalizing SNPs with no apparent function. Bayesian approaches represent the obvious statistical choice for the inclusion of external information and it provides tremendous flexibility in the types of evidence that can be incorporated. Many Bayesian methods developed for GWA analysis emphasize that they allow the integration of prior information but few publications have focused on the practicalities of this problem. To our knowledge, the work performed so far has focused on including prior information

either at the level of the SNP effects (Fridley et al., 2010; Vannucci & Stingo, 2010; Stingo, Chen, Tadesse, & Vannucci, 2011; Spencer et al., 2016) or at the level of the SNP's probability of association with the outcome (Lewinger, Conti, Baurley, Triche, & Thomas, 2007; Thompson et al., 2013).

The advantages of the joint estimation of SNP effects and the inclusion of prior biological information can be combined by using the Bayesian counterparts of the frequentist regularization methods with informative priors to reflect available knowledge. Bayesian regularization methods have a hierarchical formulation that sets a prior distribution on the SNP effects and models the dispersion parameter of that distribution using hyperpriors that induce shrinkage (also known as shrinkage priors) (Gianola, 2013). Usually, uninformative shrinkage priors are used and the data defines the amount of shrinkage that is applied across all SNPs (Gianola, 2013).

Here we propose a method to include prior biological information into a Bayesian model that jointly estimates SNP effects by using biological information to modulate shrinkage. We consider a model where the SNP effects follow a normal distribution and the variance is modelled by a shrinkage hyperprior (Gelman, 2006). The goal is to use shrinkage to constrain the variances of the effects of SNPs with little biological support so that their effect estimates are shrunk towards zero (no effect). In contrast, SNPs with a lot of biological support are assigned a different shrinkage parameter value so that the SNP effects are less constrained. This translates to the estimation of larger SNP effects and a higher likelihood of association with the outcome.

To implement this approach of including biological information into a Bayesian model by modulating shrinkage, we first address the problem of understanding the behavior and performance of a Bayesian shrinkage model when differential shrinkage is applied. We then address the problem of translating biological information into differential shrinkage that will have an impact in SNP detection. Finally, we propose a measure to define the set of shrinkage parameter values that yield the best performance. The choice of the shrinkage parameter is a general problem in Bayesian shrinkage models, especially when it cannot be based on the data at hand unlike in maximum likelihood and Empirical Bayes approaches. It is

even more complex in the context of differential shrinkage because it implies defining more than one shrinkage parameter value.

Our approach differs from the Bayesian Adaptive Lasso method (Leng, Tran, & Nott, 2010), which also considers variable-specific shrinkage. In the Bayesian Adaptive Lasso, the shrinkage parameters are chosen using empirical Bayes estimation by maximum likelihood or a hierarchical Bayes approach, where the shrinkage parameters are assigned a hyperprior distribution and estimated alongside the data. In our approach, the shrinkage parameters are defined based strictly on external information which is translated into a gradient of shrinkage.

This approach also differs from other methods that include prior knowledge and estimate SNP effects. Here we use prior knowledge to create differential shrinkage and model the variance of the effects instead of modelling the SNP effects directly. This allows us to bypass the problem of modeling the direction of effect, which has been dealt before by using mixture models (Fridley et al., 2010) or by adding parameters to the model, such as vectors that indicate the direction of effect (Quintana et al., 2013). By modelling the variance of SNPs effects that are, *a priori*, centered at zero the direction of effect is strictly dependent on the data. This approach also simplifies the integration of external information as it avoids the need to specify the direction that each item of information will have on the SNP effect.

The remainder of this work is structured as follows: we first conduct a simulation study to test our approach under different hypothetical scenarios of differential shrinkage and under SNP-specific shrinkage according to simulated prior knowledge, where prior knowledge was simulated under the assumption that the truly associated SNPs tend to have more prior support. This ensures that failure to improve SNP detection would be due to poor model performance and not due to lack of relevance of the information used. We further test the method using an empirical example on Body Mass Index (BMI) with real data from The European Community Respiratory Health Survey (ECRHS) (Burney, Luczynska, Chinn, & Jarvis, 1994). We create a dataset where we know which SNPs are truly associated with BMI and test the method after incorporating real biological information retrieved from various online repositories.

Materials and Methods

Bayesian Model

We considered an approach based on the model framework proposed by Yi *et al*. (Yi, Liu, Zhi, & Li, 2011), a family of Bayesian hierarchical Generalized Linear Models (BhGLM) with shrinkage priors that perform joint analysis of SNPs in a computationally efficient way through the application of the EM-IWLS algorithm (Yi & Ma, 2012). This is a modified version of the standard Iterative Weighted Least Squares (IWLS) algorithm (Green, 1984) that includes Expectation-Maximization (EM) steps (Dempster, Laird, & Rubin, 1977). It is based on the estimation of the marginal posterior mode and can very quickly run an analysis on thousands of SNPs (Yi & Ma, 2012). This is essential for our approach since our ultimate goal is to scale up to millions of SNPs as in GWA analyses.

The BhGLM model framework is very flexible and encompasses different models with different priors. For this work, we considered a normally distributed, continuous trait Y and the following model:

$$y_i = \alpha_0 + \sum_j \alpha_j x_{ij} + \varepsilon_i$$

where α_0 is the intercept, α_j is the effect of the j^{th} SNP, x_{ij} is the genotype of the j^{th} SNP in the i^{th} subject and ε_i is the random error. The Bayesian hierarchical formulation of the model is the following:

$$\alpha_j | \tau_{\alpha_j}^2 \sim N(\mu_j, \tau_{\alpha_j}^2)$$

$$\tau_{\alpha_j}^2 | s_{\alpha_j}^2 \sim Inv - \chi^2(1, s_{\alpha_j}^2)$$

where μ_j is the prior mean of the SNP effect, $\tau_{\alpha_j}^2$ is the variance of α_j and $s_{\alpha_j}^2$ a scale hyperparameter. This formulation considers that α_j follows a scale mixture of normal distributions with SNP-specific variance $\tau_{\alpha_j}^2$ (Yi & Ma, 2012) and corresponds to the hierarchical formulation of the Cauchy distribution for the SNP effects with location μ_j and scale s_{α_j}.

We hypothesize that prior knowledge can be included at the level of the scale hyperparameter $s^2_{\alpha_j}$ by defining its value according to external biological information. The rational of the method is the following: the scale hyperparameter controls the variance of the SNP effect ($\tau^2_{\alpha_j}$) and, consequently, a small $s^2_{\alpha_j}$ will be associated with smaller variances and estimates of the effect size, α_j, closer to the prior mean, μ_j. Conversely, a larger $s^2_{\alpha_j}$ will lead to larger variances and estimates of α_j that can deviate further from the mean according to support from the data. In our setting, we assume that the SNPs come from a scale-mixture of normal distributions with prior mean centred at zero and set $\mu_j = 0$. We study the behavior of the model under different values of $s^2_{\alpha_j}$ informed by prior knowledge in both a simulation study and an empirical example where we know which SNPs are associated with the outcome. In all our analyses, we group the SNPs in linkage disequilibrium (LD) blocks. Our main focus is to detect independent genomic regions rather than individual SNPs because the detected SNPs are not necessarily the causal ones but, more likely, markers in LD with the causal variant.

All analyses were performed using R statistical software version 3.2.2 (https://www.r-project.org/) and the freely available package BhGLM (http://www.ssg.uab.edu/bhglm/) which implements the EM-IWLS algorithm. Further details about the functions and specific parameters used can be found in the supplementary data (Text S2).

Simulation study

We simulated 1000 datasets each with 500 subjects, using the software GENOME (Liang, Zöllner, & Abecasis, 2007) to generate 10 independent LD blocks with ~20-30 SNPs each (~200-300 SNPs in total). After excluding SNPs with minor allele frequency lower than 5%, one SNP was randomly chosen to be the causal SNP and assigned an effect size $\alpha_{causal} = 0.15$. This effect was used to generate a continuous trait Y, so that $Y \sim N(\alpha_{causal} \cdot SNP, 1)$. The effect size was chosen based on the results of the standard

analysis, such that it was not so large that it ranked the causal region first in all the datasets, nor so small that it consistently ranked below the 50th percentile.

The behavior of the model was first studied under three scenarios: 1) constant $s^2_{\alpha_j}$ for all SNPs, 2) higher $s^2_{\alpha_j}$ for only the SNPs in the true LD block, and 3) higher $s^2_{\alpha_j}$ for the SNPs in one randomly chosen false LD block with intermediate $s^2_{\alpha_j}$ for the SNPs in the true LD block. Scenario 2) aims to demonstrate the performance of the model under ideal conditions and works as a proof-of-concept (Supplementary Table 1). Scenario 3) aims to demonstrate that the model is still able to detect the true SNPs even when they are not the ones whose effects are less shrunk towards zero, which is a possible real situation (Supplementary Table 2). In this scenario, one false LD block is randomly chosen and all its SNPs are assigned a higher $s^2_{\alpha_j}$ than the SNPs in the true LD block. All other false SNPs are assigned a lower $s^2_{\alpha_j}$. We set upper and lower bounds for $s^2_{\alpha_j}$ corresponding to the randomly chosen LD block and the other false LD blocks, respectively. We assign an intermediate value of $s^2_{\alpha_j}$ to the SNPs in the true LD block, calculated by averaging the predefined lower and upper bounds of $s^2_{\alpha_j}$.

Performance was assessed by the average ranking of the true LD block across all datasets, with the ranking of the LD block defined by the best ranking SNP in that block taking into account the Bayesian p-value. All results were compared with those of the standard analysis with SNPs ranked based on their frequentist p-values.

One question that arises is how to define the most appropriate shrinkage/scale parameter. The shrinkage parameter controls the variance of the estimated SNP effects and, ideally, the best parameter will shrink towards zero the SNPs in the false LD blocks while minimally shrinking the SNPs in the true LD block. We hypothesize that a good set of shrinkage parameters will lead to a large ratio between the variance around zero of the SNP effects of the true LD block and the variance around zero of the SNP effects of the false LD blocks. Since, in practice, the true LD block is unknown we base the parameter selection on the ratio of the variance about zero of the SNP effects of the top LD block (1st out of 10) over the variance

about zero of the SNP effects of the remaining blocks. We refer to these measures as the variance ratio of the true block and the variance ratio of the top block, respectively. We test the usefulness of these measures by comparing the variance ratios obtained with several sets of shrinkage parameters with the performance of the model.

Translation of prior knowledge into shrinkage

To create a link between prior knowledge and shrinkage, we first simulated prior knowledge by randomly assigning to each LD block an integer between 0 and 10, where 0 corresponds to absence of prior support and 10 corresponds to the maximum prior support. All the SNPs in a LD block were assigned the block's value of prior knowledge. The causal SNP was then sampled from the dataset using prior knowledge as frequency weights in a random drawing, so that SNPs with a value of prior knowledge of 10 would have more probability of being selected to be the causal SNP than SNPs with less prior knowledge.

We translated prior knowledge into shrinkage by setting a range of values for $s^2_{\alpha_j}$, with the lower bound $s^2_{\alpha_{j,0}}$ corresponding to prior knowledge=0 and the higher bound $s^2_{\alpha_{j,10}}$ corresponding to prior knowledge=10. Intermediate values of prior knowledge were interpolated linearly in-between.

We repeated the analyses using a binary outcome instead of a continuous one. In particular, we used our simulated datasets with a simulated a binary outcome to test model performance in the setting of a generalized linear model with a *logit* link function and a similar Bayesian hierarchical formulation. The details of the model and analyses can be found in the supplementary data (Text S1).

Empirical example

We tested our approach using an empirical example with data on Body Mass Index (BMI) from The European Community Respiratory Health Survey (ECRHS), a multicenter European study aimed at identifying factors associated with the world-wide increase in asthma prevalence (Burney et al., 1994). We

chose BMI because it is a widely studied phenotype with many genetic variants identified and replicated in multiple studies (Locke et al., 2015).

From the original ECRHS GWA study on 1,829 individuals and 2,588,592 SNPs, we constructed a subset with a smaller number of SNPs in order to reduce the computational time. This dataset included SNPs known to be associated with BMI ('true' SNPs) and randomly selected SNPs ('false' SNPs). The selection of 'true SNPs' was based on a recently published meta-analysis of GWA studies on BMI that identified 97 significant loci and constitutes the largest meta-analysis of BMI so far (Locke et al., 2015).

Using the data from ECRHS, we performed the standard regression analysis of the 97 SNPs along with all the SNPs in the same LD blocks (11,389 SNPs in total), and selected 6 SNPs using the following criteria: 3 significant SNPs (rs7138803, p=0.001; rs11057405, p=0.036; rs758747, p=0.049), 2 SNPs with p-values close to 0.05 (rs6567160, p=0.066, rs7243357, p=0.070) and 1 SNP with a p-value close to 0.1 (rs1558902, p=0.116). Additionally, twenty-four SNPs were randomly chosen from the 2,588,592 SNPs available in the ECRHS dataset and were used as controls in the analysis ('false' SNPs). In both instances, the SNPs were chosen so that no two SNPs mapped to the same LD block. All 30 SNPs (6 true and 24 random SNPs) were mapped to their corresponding LD blocks using Pos2LDBlock, a tool that maps SNPs to LD blocks based on *D'* (Taliun, Gamper, & Pattaro, 2014; Weichenberger et al., 2015). With this approach, we obtained a dataset of 2,614 SNPs and 1,829 individuals to which the model and our method were applied. In all analyses, both the outcome and the genotypes were standardized by subtracting the mean and diving by the standard deviation.

Retrieval of prior biological knowledge

We retrieved prior knowledge based on a set of 10 questions regarding biological characteristics of the SNPs (Figure 1) that have been previously associated with an increase in a SNPs probability of association with several outcomes (Minelli et al., 2013). All questions have binary 'Yes' or 'No' answer, corresponding

to a score of 1 or 0, respectively. The final amount of prior knowledge for each SNP was the sum of the scores obtained for the 10 questions. The prior knowledge was independent of the results of the meta-analysis of GWA studies on BMI (Locke et al., 2015) used to define the 'true SNPs', which excludes the possibility of knowledge contamination.

Most of the data retrieval was performed using the data integration framework 'Dintor', a bioinformatics tool suite that queries information from multiple online datasets and is designed for use in functional annotation of genomic and proteomic data (Weichenberger et al., 2015). Additionally, we obtained data from the following databases: Pfam Protein Families database (Finn et al., 2014), Mouse Genome Informatics (Eppig, Blake, Bult, Kadin, & Richardson, 2015) and Reactome (Croft et al., 2014; Milacic et al., 2012). Further details regarding the retrieval of prior information can be found in the supplementary data (Text S3).

Translation of prior knowledge into shrinkage

Similarly to the simulation study, we tested the performance of the model both when ignoring prior knowledge, therefore using only one shrinkage parameter value, and when incorporating prior knowledge, thus using different values of shrinkage parameters according to the level of prior support. We set the upper bound of shrinkage to correspond to the maximum score of prior knowledge obtained, and interpolated the shrinkage parameter values linearly in between. Performance of the model was evaluated based on the average ranking of the true LD blocks, with the ranking of the LD block defined by the best ranking SNP in that block, according to its Bayesian p-value. The variance ratio of the estimated effects of the top six blocks over the remaining blocks was also calculated to assess the accuracy of this measure in identifying the best set of parameters. We set the threshold at six because it corresponded to the real number of true LD blocks. These results were compared with the variance ratio of the true LD blocks over the false blocks and studied the association of these measures with model performance.

As for the simulation study, we repeated the analysis using a binary outcome, in this case, by dichotomizing BMI into normal (≤25kg/m^2) vs. overweight/obese (>25kg/m^2) and tested the performance of the model using our empirical dataset. Further details can be found in the supplementary data (Text S2).

Results

Simulation study

Table 1 and Figure 2a summarize the results obtained with different shrinkage parameter values without inclusion of prior knowledge, that is, assigning the same shrinkage parameter to all the SNPs in each dataset (scenario 1, constant $s_{\alpha_j}^2$ for all SNPs). The average ranking of true LD block obtained using different shrinkage parameters is compared with the standard analysis where the average ranking was ~3.6. Models 1.1 and 1.2 mimic the scenario of no shrinkage, since assigning a large scale parameter will result in large variances of the SNP effects. The true scenario of no shrinkage corresponds to $s_{\alpha_j}^2 = \infty$ which is the same as fitting a standard linear model with all the SNPs as predictors and, as expected, models 1.1 to 1.5 show little or no improvement comparing to random guessing. However, as $s_{\alpha_j}^2$ decreases and more shrinkage is applied, the ranking of the true LD block improves until it is comparable with the standard analysis in models 1.8 to 1.11. This suggests that equally shrinking all the SNP effects does not produce better results than the standard analysis in this particular scenario.

In all simulated datasets, the effect size for the causal SNP was set to 0.15 and, given the properties of the Cauchy distribution, setting $s_{\alpha_j} = 0.1$ would be a sensible value to estimate SNP coefficients in the range ± 0.4. Therefore, in order to detect SNPs with coefficients ~0.15 such as the causal SNP and the SNPs in the same LD block, a sensible value of $s_{\alpha_j}^2$ is $0.1^2 = 0.01$. Noticeably, in model 1.7 (Table 1), we observe the first improvement in performance of the model as $s_{\alpha_j}^2$ decreases. A stronger shrinkage as in model 1.8, produces better results as it induces more shrinkage on the false SNPs while maintaining a good range to

detect the true SNPs (± 0.04).

We calculated the variance ratios of the SNP effects of the top and true LD blocks to test the hypothesis that they predict model performance and can be used to determine the best set of shrinkage parameters in a real dataset. An increase in the variance ratio for the top block is associated with an improvement in performance of the model with the best ranking being achieved for model 8 (Table 1, $s^2_{\alpha_j} = 0.001$). This measure also remains constant when even stronger shrinkage is applied and is associated with stabilization in performance (Table 1, models 1.9 to 1.11). The same pattern of variation is observed for the variance ratio of the true blocks, which suggests that both these measures are associated with performance and that the variance ratio of the top block is a good proxy for the variance ratio of the true block (Figure 2a).

The performance of the model was also studied under the scenarios of higher $s^2_{\alpha_j}$ for only the SNPs in the true LD block (scenario 2) and higher $s^2_{\alpha_j}$ for the SNPs in one randomly chosen false LD block with intermediate $s^2_{\alpha_j}$ for the SNPs in the true LD block (scenario 3). In scenario 2, our approach markedly improved upon the standard analysis with a mean ranking of the true LD block of 1 in most of the models tested (Supplementary Table 1). Scenario 3 also showed improved performance with an average ranking of ~1.5 in the best model (vs. ~3.6 in the standard analysis, Supplementary Table 2). This later result suggests that our approach is able to detect the true LD block in a situation where it is not the block with more prior support, which is a likely case in a real study.

Translation of prior knowledge into shrinkage

Table 2 shows the shrinkage parameter values in terms of the lower bound for $s^2_{\alpha_j}$ (prior support=0) and the upper bound corresponding to support=10. Intermediate values were obtained by linear interpolation. These results suggest that the model can outperform the standard analysis when differential shrinkage is applied. Models 2.4, 2.8 and 2.12 all show a significant improvement with an average ranking of the true block of ~2.8 vs. ~3.6 in the standard approach (Table 2). All these models have the same upper bound of shrinkage

($s^2_{\alpha_j} = 0.001$) while the lower bound ranges from 10^{-4} to 10^{-6}. Even though there is a marginal improvement in average ranking with a decreasing lower bound, the results suggest that, in all these models, the SNP effects of the blocks with little prior knowledge are being effectively shrunk towards zero. Interestingly, the upper bound for these models is the same as the largest $s^2_{\alpha_j}$ (0.001) that yielded the best results in the scenario where shrinkage is constant for all SNPs (Table 1, model 1.8).

The analysis of the variance ratios for the top and true LD blocks shows an even stronger association between performance and these measures, with local maxima corresponding to the three best performing models (Figure 2b) and the absolute maximum corresponding to the best performing model overall (Table 2, model 2.12). The same pattern of variation is again observed between the variance ratio of the top and the true LD block which suggested that the variance ratio of the top block is a good proxy when differential shrinkage is applied.

The results of the simulation study using a binary outcome yielded very similar results to those obtained using a continuous outcome. The model did not improve upon the standard analysis when no differential shrinkage was applied with the model matching the performance of the standard SNP analysis when $\boldsymbol{s^2_{\alpha_j} < 0.01}$ (Models S3.8-S3.11, Supplementary Table 3). When prior information was included as differential shrinkage, the model was able to improve the ranking of the true LD blocks in comparison with the standard analysis: the best mean ranking of the true LD block obtained was 2.9 for the Bayesian model vs. 3.6 for the standard analysis (model S4.3, Supplementary Table 4). Additionally, the best models corresponded to maxima of the variance ratios for the top block and true blocks (Supplementary Table 3), which further supports the ability of these measures to identify the best set of parameters to use.

Empirical example

The retrieval of prior information showed that most SNPs did not contain any prior biological support (n=945), 873 SNPs had a prior knowledge score of 1 and 701, 111 and 14 had score of 2, 3 and 4,

respectively, with a score of 4 being the maximum amount of prior knowledge obtained. Therefore, the upper bound of shrinkage was set to correspond to a score of 4 in this example.

Similarly to the simulation study, the model can match but not outperform the standard analysis in the scenario of constant shrinkage when the model is applied to a set of real data with real prior biological information. (Table 3). In particular, models 3.9 to 3.12 (Table 3) show the same average ranking for the true SNPs as the standard analysis (~9.3). Model 3.8 shows a better, but not substantial, result with an average of ~9.2.

To validate the variance ratio of the top blocks as a measure to determine the best set of shrinkage parameters, we calculated the variance ratio of the top six blocks for both scenarios of constant and differential shrinkage and compared with the variance ratio of the six true LD blocks and the associated model performance. The pattern of variation of these measures is shown in Figure 3a. The same relationship between the variance ratios and performance is observed, with maxima of both types of variance ratios being associated with the best performance of the model (Table 3, models 3.8 to 3.11). It can be noted that we obtain very high variance ratios in models 3.1 and 3.2, which are higher than the ratios we obtain when stronger shrinkage is applied in models 3.9 to 3.11 (Table 3). However, the values used for the shrinkage parameter in models 3.1 and 3.2 ($s_{\alpha_j}^2 = 100{,}000$ and $10{,}000$, respectively) correspond to scenarios of almost no shrinkage which are similar to performing a standard linear regression in a model with too many covariates, as previously discussed.

For the scenario of differential shrinkage (Figure 3b), the variance ratio of both the top LD blocks and the true LD blocks was strongly related with performance, with the three local maxima corresponding to the three best performing models (Table 4, models 4.4, 4.8 and 4.12). The variance ratio of the true blocks achieves local maxima for similar models (models 4.3, 4.7 and 4.11) but not for the top performing ones. Additionally, the variance ratios for the top blocks show much higher maxima, which supports the idea that this is not only a good proxy but is a better measure than the variance ratio of the true blocks (Figure 3b).

Models 4.4, 4.8 and 4.12 were the top performing models with model 4 ranking the LD blocks, on average, one position higher than the standard analysis (~8.5 and ~9.3, respectively) and corresponding to the absolute maximum of the variance ratios for the top LD blocks. The shrinkage parameters of these models also correspond to the best performing parameter pairs in our simulation study with model 4.6 corresponding to model 2.4, the third best performing model in our simulations.

Similarly to the simulation study, the results obtained using a binary outcome showed improvement upon the standard anaylsis when prior information is included in the model. In the scenario of constant shrinkage applied to all SNPs (no inclusion of prior information) there was no improvement in the average ranking of the true LD block comparing to the standard analysis (Supplementary Table 5). We observed that, in this case, the model does not converge when $s_{\alpha_j}^2 \geq 10{,}000$ (models S.51 and S.52, Supplementary Table 5), which are approximations to the situation of no shrinkage. When prior information is included, we observed an improvement in the mean ranking of the 6 true blocks from ~8.2 in the standard SNP analysis to ~7.3 in the Bayesian model (model S6.13, Supplementary Table 6). The maximum values of the variance ratios of the top block and true blocks were again able to identify the best performing models (Supplementary Tables 5 and 6).

These results, alongside their simulation study counterparts, suggest that our approach is also applicable to a binary outcome; the margin of improvement is slightly lesser than with a continuous outcome, probably due to the loss of statistical power associated with binary outcomes.

To illustrate the efficiency of the algorithm, we obtained the computational time for each model and the overall computational time for both the simulated and empirical datasets. All analyses were run on an Intel® Core™ i7-4770 CPU (quad-core) with 3.40GHz. The computational time varied from 0.06 to 0.30 seconds for each individual simulated dataset. For the empirical example with 1,828 subjects and 2,614 SNPs, the computational time varied between 3.12 and 10.26 seconds to analyse using the different shrinkage parameter combinations.

Discussion

In this study, we use a Bayesian shrinkage model that jointly estimates SNP effects and induce SNP-specific shrinkage as a means to include external biological information. We show that differential SNP-specific shrinkage was able to outperform the standard analysis in detecting the LD blocks that contained the true SNPs. This contrasts with the fact that, without differential shrinkage, the model cannot outperform the standard analysis and performs poorly when little shrinkage is applied. When no shrinkage is applied, the model corresponds to a standard linear model with estimation performed by the least squares method, which is known to perform poorly in the presence of a large number of strongly correlated SNPs.

The translation of different levels of biological knowledge into shrinkage was able to improve the ranking of the LD block that contains the causal SNP both in the simulated and empirical examples. This suggests that this approach is useful in increasing the yield of detected SNPs. In the simulation study, we did not use real biological knowledge but rather use simulated knowledge that increases the chance of a SNPs being selected to be a true SNP. The improved performance observed demonstrated that our approach works when biological knowledge is relevant. In the empirical example, we use real biological information that has been suggested to increase the probability of a SNP being associated with an outcome (Minelli et al., 2013). In this case, the improved performance observed shows not only that the approach works, but also confirms that the type of prior knowledge used is relevant. It is important to note that the prior knowledge considered in this paper refers to biological characteristics of the SNPs, independently from whether the SNP has been already discovered in previous genetic association studies. Therefore, our approach will not be biased towards replicating previous findings and should work as well with novel SNP discoveries. Additionally, the performance of this approach is expected to improve as more information about the SNPs becomes available.

The model framework used in this study was previously applied to genetic association studies with the purpose of increasing power to detect rare variants (Yi et al., 2011). The SNPs were grouped in four groups according to MAF (frequent vs. rare variants) and whether the SNP was synonymous or non-synonymous.

The authors split shrinkage into three categories: weak shrinkage ($s_{\alpha_j}^2 > 1$), moderate shrinkage ($0.2 \leq s_{\alpha_j}^2 \leq 1$) and strong shrinkage ($s_{\alpha_j}^2 < 0.2$) and apply moderate shrinkage to their data. In our study, we find that much strong shrinkage is required to obtain optimal results with values significantly smaller than the upper bound of strong shrinkage (e.g. $s_{\alpha_j}^2 \leq 0.001$). Yi *et al*. (2011) use rare variants in the model while, in our study, we exclude all variants with a MAF<0.1. The authors center the mean of the SNP effects of the rare variants at 1 or at a functional score between 0.16 and 1. As expected, this inflates the estimated SNP effect and less shrinkage is required to detect these variants. Yi *et al*. also applied a new model where group effects are estimated along with the SNP effects. The overall SNP effect is the product of the estimated group effect with the estimated SNP effect, and this can also affect the value of the optimal shrinkage parameter. Although we are interested in genomic regions defined by LD, a form of group structure, we chose not to use this particular model setting due to its high false positive rate (personal communication with the author; the R function that implemented this scenario was removed from the *BhGLM* package when it was updated).

An important problem that is common to all shrinkage methods, Bayesian or not, is the selection of the best shrinkage parameter. In classical shrinkage methods, the best shrinkage parameter is usually estimated through cross-validation (Hastie et al., 2009). In the Bayesian framework, the shrinkage parameter is either estimated using the empirical Bayes method by marginal maximum likelihood or by assigning a hyperprior distribution to the parameter and estimating it along with the other model parameters (Park & Casella, 2008). We use a different approach to determine the best parameters by calculating the ratio of the variance of the estimated SNP effects around 0 in the top blocks over the remaining blocks. This is based on two assumptions: 1) the true LD blocks tend to have more prior knowledge and 2) the model with the best set of parameters will rank these blocks high. Our results show that the maximization of the estimated SNP effect variance ratio of the top ranking LD blocks over the remaining blocks is a good measure to define the best set of shrinkage parameters, particularly when differential shrinkage is applied. This measure allows the practical implementation and application to real examples with thousands of SNPs requiring

user-input only for the expected number of true LD blocks. For the calculation of the variance ratio, we used the first top-ranking block in each of the simulated datasets and the six top-ranking blocks in the empirical dataset, as these were the real number of true blocks in each example and it was known *a priori*. In a real dataset, the user has to choose the number of blocks that are expected to be true to decide which blocks are the 'top blocks' and calculate the variance ratio.

An alternative to maximizing the variance ratio of the top ranking LD blocks to determine the best set of shrinkage parameters is to use a simulation approach where some blocks would be randomly chosen to be the 'true blocks' and used to simulate an outcome similar to the real outcome being studied assuming certain a SNP effect. Performance would be evaluated for different shrinkage parameter values as performed in this study. The model would then run for several rounds with different 'true blocks' and different simulated outcomes. The shrinkage parameters would be determined by the best performing set in all rounds of simulation. This approach, however, is much more computationally demanding than testing different shrinkage parameter values and choosing the values that maximize the variance ratio of the top blocks. While we predict that this approach could be applicable to our datasets, our goal is to apply our method to large datasets with many thousands of SNPs and, ultimately, to GWA studies for which this simulation-based approach would become impractical.

Our method illustrates that it is possible to include prior biological knowledge in a Bayesian joint SNP analysis by modulating shrinkage. Nonetheless, it is not necessarily exclusive to this particular setting. It is potentially applicable to other high dimensional settings where Bayesian shrinkage methods are required and prior information is available, including models that perform variable selection. A recent study applied a similar hierarchical shrinkage model that performs variable selection to study SNP associated with expression Quantitative Trait *loci* (eQTLs), that is SNPs associated with changes in gene expression (Boggis et al., 2016). External biological information was used by grouping SNPs into functional groups and calculating a functional significance score. This score was incorporated in the model via the expectation of the marginal prior variance of the SNP effects, which is similar to our approach. The authors show that

the inclusion of prior information was able to increase SNP detection and that the model outperformed other similar Bayesian models that did not include external information. Even though this is work is applied to eQTLs, it illustrates that the modulation of shrinkage is a feasible approach to include external information in a Bayesian model like we advocate.

The assessment of the effect of prior knowledge in shrinkage has two different aspects: first, the effect of the modulation of the shrinkage parameter value *per se*, which is independent of prior knowledge and second, the translation of prior knowledge into differential shrinkage. In this manuscript, we explored both of these aspect but further work is required to fully understand the impact that prior information can have in this approach and to optimize its translation into shrinkage. Improvements in both these aspects will make better use of external information and are expected to further improve SNP detection.

Acknowledgements

This work was carried out as part of a PhD project funded by the National Heart and Lung Institute (NHLI) foundation. DCT has grant support from the NIH grants P01 CA196569, R01 ES019876, P30 ES07048, and P30 CA014089. The authors have no conflicts of interest to declare.

References

Boggis, E. M., Milo, M., & Walters, K. (2016). eQuIPS: eQTL Analysis Using Informed Partitioning of SNPs - A Fully Bayesian Approach. *Genetic Epidemiology*, 40(4), 273–28.

Burney, P. G., Luczynska, C., Chinn, S., & Jarvis, D. (1994). The European Community Respiratory Health Survey. *The European Respiratory Journal*, *7*(5), 954–60.

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American Journal of Human Genetics*, *86*(1), 6–22.

Cho, S., Kim, K., Kim, Y. J., Lee, J., Cho, Y. S., Lee, J., … Park, T. (2010). Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association

Analysis, 416–428.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., … D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42*(Database issue), D472-7.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., & Richardson, J. E. (2015). The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Research*, *43*(Database issue), D726-36.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222-30.

Fridley, B. L., Serie, D., Jenkins, G., White, K., Bamlet, W., Potter, J. D., & Goode, E. L. (2010). Bayesian mixture models for the incorporation of prior knowledge to inform genetic association studies. *Genetic Epidemiology*, *34*(5), 418–26.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*, *194*(3), 573–96.

Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives Author(s. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(2), 149–192.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). misc, Springer-Verlag New York.

Leng, C., Tran, M. N., & Nott, D. (2010). Bayesian Adaptive Lasso. Retrieved from http://arxiv.org/abs/1009.2300

Lewinger, J. P., Conti, D. V, Baurley, J. W., Triche, T. J., & Thomas, D. C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol*, *31*(8), 871–882.

Liang, L., Zöllner, S., & Abecasis, G. R. (2007). GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics (Oxford, England)*, *23*(12), 1565–7.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *The New England Journal of Medicine*, *363*(2), 166–76.

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., … Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, *4*(4), 1180–211.

Minelli, C., De Grandi, A., Weichenberger, C. X., Gogele, M., Modenese, M., Attia, J., … Thompson, J. R. (2013). Importance of Different Types of Prior Knowledge in Selecting Genome-Wide Findings for Follow-Up. *Genet Epidemiol*, *37*(2), 205–213.

Minelli, C., De Grandi, A., Weichenberger, C. X., Gögele, M., Modenese, M., Attia, J., … Thompson, J.

R. (2013). Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genetic Epidemiology*, *37*(2), 205–13.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Quintana, M. A., Schumacher, F. R., Casey, G., Bernstein, J. L., Li, L., & Conti, D. V. (2013). Incorporating prior biologic information for high-dimensional rare variant association studies. *Human Heredity*, *74*(3–4), 184–195.

Spencer, A. V., Cox, A., Lin, W.-Y., Easton, D. F., Michailidou, K., & Walters, K. (2016). Incorporating Functional Genomic Information in Genetic Association Studies Using an Empirical Bayes Approach. *Genetic Epidemiology*, 40(3):176-87.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., & Vannucci, M. (2011). Incorporating Biological Information Into Linear Models: A Bayesian Approach To The Selection Of Pathways And Genes. *The Annals of Applied Statistics*, *5*(3), 1978–2002.

Taliun, D., Gamper, J., & Pattaro, C. (2014). Efficient haplotype block recognition of very long and dense genetic sequences. *BMC Bioinformatics*, *15*(1), 10.

Thompson, J. R., Gögele, M., Weichenberger, C. X., Modenese, M., Attia, J., Barrett, J. H., … Minelli, C. (2013). SNP prioritization using a Bayesian probability of association. *Genetic Epidemiology*, *37*(2), 214–21.

Vannucci, M., & Stingo, F. (2010). Bayesian Models for Variable Selection that Incorporate Biological Information. *Bayesian Statistics*, *9*, 1–20.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24.

Weichenberger, C. X., Blankenburg, H., Palermo, A., D'Elia, Y., König, E., Bernstein, E., & Domingues, F. S. (2015). Dintor: functional annotation of genomic and proteomic data. *BMC Genomics*, *16*(1), 1081.

Yi, N., Liu, N., Zhi, D., & Li, J. (2011). Hierarchical generalized linear models for multiple groups of rare and common variants: Jointly estimating group and Individual-Variant effects. *PLoS Genetics*, *7*(12).

Yi, N., & Ma, S. (2012). Hierarchical shrinkage priors and model fitting for high-dimensional generalized linear models. *Statistical Applications in Genetics and Molecular Biology*, *11*(6).

Legends

Figure 1. List of external biological information questions included in the statistical model. *All genes within ±5kb of the SNP are considered.

Figure 2. Variance ratios for the estimated SNP effects in the simulation study: a. Constant shrinkage applied to all SNPs with the shrinkage parameter values represented on the x-axis; b. Range of shrinkage parameters with lower and upper bounds to shrinkage represented on the x-axis. The variance ratios of the top and true blocks are represented by the continuous black and light gray lines, respectively. The average ranking of the true block is represented by the dotted line with squares and its scale is on the right-hand side y-axis.

Figure 3. Variance ratios for the estimated SNP effects in the simulation study: a. Constant shrinkage applied to all SNPs with the scale parameter values represented on the x-axis; b. Range of shrinkage parameters with lower and upper bounds to shrinkage represented on the x-axis. The variance ratios of the top and true blocks are represented by the continuous black and light gray lines, respectively. The average ranking of the true block is represented by the dotted line with squares and its scale is on the right-hand side y-axis.

Tables

Table 1. Average ranking of the true LD block in the simulation study with a continuous outcome. Results are presented for the standard SNP analysis and Bayesian model with constant shrinkage across SNPs. Variance ratios for the top block and for the true block are also presented and the best performance is reached when the variance ratio is maximized.

Constant $s^2_{\alpha_j}$ for all the SNPs				
Model	**Scale ($s^2_{\alpha_j}$)**	**Variance ratio**		**Average Ranking**
		Top block	True block	
Standard analysis	-	2.201	1.501	3.624
1.1	10^5	2.740	1.147	5.210
1.2	10^4	2.740	1.147	5.210
1.3	5	2.574	1.135	5.209
1.4	2.5	2.351	1.118	5.231
1.5	1	2.021	1.108	5.183
1.6	0.1	2.361	1.374	4.865
1.7	0.01	2.424	1.755	3.673
1.8	0.001	2.455	1.814	3.623
1.9	10^{-4}	2.456	1.814	3.624
1.10	10^{-5}	2.456	1.813	3.624
1.11	10^{-6}	2.456	1.813	3.624

Table 2. Average ranking of the true LD block in the simulation study with a continuous outcome. Results are presented for the standard SNP analysis and Bayesian model with constant shrinkage across SNPs. Variance ratios for the top block and for the true block are also presented and the best performance is reached when the variance ratios are maximized.

Differential $s^2_{\alpha_j}$ according to prior information					
Model	**Scale ($s^2_{\alpha_j}$)**		**Variance ratio**		**Average Ranking**
	prior support = 0	prior support = 10	Top block	True block	
Standard analysis	-	-	2.201	1.501	3.624
2.1	10^{-4}	1	2.463	1.189	4.858
2.2	10^{-4}	0.1	2.629	1.350	4.739
2.3	10^{-4}	0.01	4.571	2.305	3.482
2.4	10^{-4}	0.001	4.723	2.654	2.835
2.5	10^{-5}	1	2.463	1.189	4.857
2.6	10^{-5}	0.1	2.627	1.351	4.737
2.7	10^{-5}	0.01	4.599	2.320	3.471
2.8	10^{-5}	0.001	5.111	2.726	2.806
2.9	10^{-5}	10^{-4}	2.608	1.820	3.366
2.10	10^{-6}	0.1	2.627	1.351	4.737
2.11	10^{-6}	0.01	4.609	2.320	3.472
2 12	10^{-6}	0.001	5.140	2.733	2.802
2 13	10^{-6}	10^{-4}	2.605	1.827	3.361

Table 3. Average ranking of the true LD block in the empirical example with ECRHS data and BMI as the outcome. Results are presented for the standard SNP analysis and Bayesian model with constant shrinkage across SNPs. Variance ratios for the top six blocks and for the true blocks are also presented and the best performance is reached when the variance ratio is maximized.

Constant $s^2_{\alpha_j}$ for all the SNPs				
Model	**Scale ($s^2_{\alpha_j}$)**	**Variance ratio**		**Average Ranking**
		Top 6 blocks	True blocks	
Standard analysis	-	4.527	2.471	9.333
3.1	10^5	5.220	2.971	14.500
3.2	10^4	7.194	2.977	14.833
3.3	5	2.208	0.399	17.000
3.4	2.5	2.512	0.748	15.500
3.5	1	2.301	1.137	11.333
3.6	0.1	1.305	1.662	10.500
3.7	0.01	2.508	2.147	12.500
3.8	0.001	4.531	2.450	9.167
3.9	10^{-4}	4.517	2.466	9.333
3.10	10^{-5}	4.514	2.465	9.333
3.11	10^{-6}	4.514	2.465	9.333

Table 4. Average ranking of the true LD blocks in the empirical example with ECRHS data and BMI as the outcome. Results are presented for the standard SNP analysis and Bayesian model when differential shrinkage is applied, with the lower and upper bounds for the shrinkage parameter values reported in columns 2 and 3, respectively. Variance ratios for the top six blocks and for the true blocks are also presented and the best performance is reached when the variance ratio is maximized.

Differential $s^2_{\alpha_j}$ according to prior information					
Model	**Scale ($s^2_{\alpha_j}$)**		**Variance ratio**		**Average Ranking**
	prior support = 0	prior support = 4	Top 6 blocks	True blocks	
Standard analysis	-	-	4.527	2.471	9.333
4.1	10^{-4}	1	2.442	2.060	10.833
4.2	10^{-4}	0.1	4.049	2.355	9.167
4.3	10^{-4}	0.01	7.259	2.927	8.833
4.4	10^{-4}	0.001	8.803	2.226	8.500
4.5	10^{-5}	1	2.442	2.060	10.833
4.6	10^{-5}	0.1	4.055	2.357	9.167
4.7	10^{-5}	0.01	7.343	2.947	8.833
4.8	10^{-5}	0.001	9.894	2.280	8.667
4.9	10^{-5}	10^{-4}	5.006	2.428	9.500
4.10	10^{-6}	0.1	4.056	2.357	9.167
4.11	10^{-6}	0.01	7.351	2.948	8.833
4.12	10^{-6}	0.001	9.988	2.288	8.667
4.13	10^{-6}	10^{-4}	4.903	2.454	9.500

Figures

Figure 1

Questions on prior biological information

1. SNP in a transcribed but not translated region?
2. SNP in a translated region but does not change the amino acid?
3. SNP changes the amino acid but not in functional protein domain?
4. SNP in a functional protein domain?
5. SNP in a regulatory region which is not transcribed?
6. SNP in a transcribed regulatory region?
7. SNP in a genomic region evolutionary conserved in vertebrates?
8. SNP in a gene (±5kb*) that has been associated with the same/closely related phenotype in functional models (animal or *in vitro* studies)?
9. SNP in a gene (±5kb*) which is highly expressed in a tissue relevant to the phenotype?
10. SNP in a gene (±5kb*) which shows gene/protein interactions relevant to the phenotype?

Figure 2

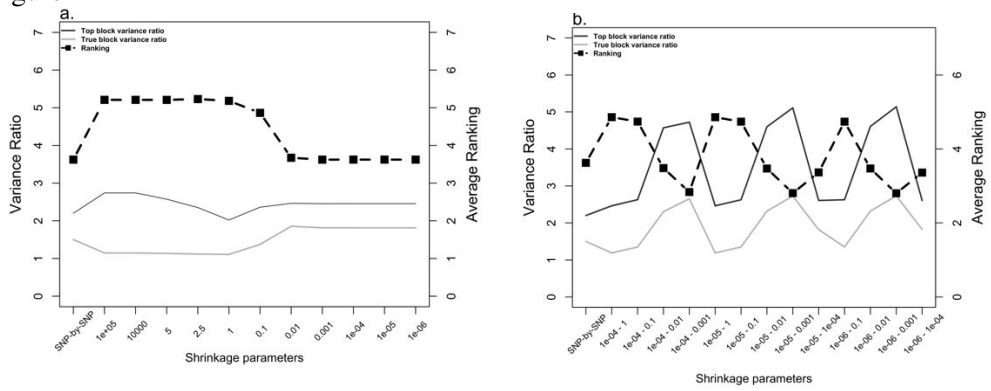
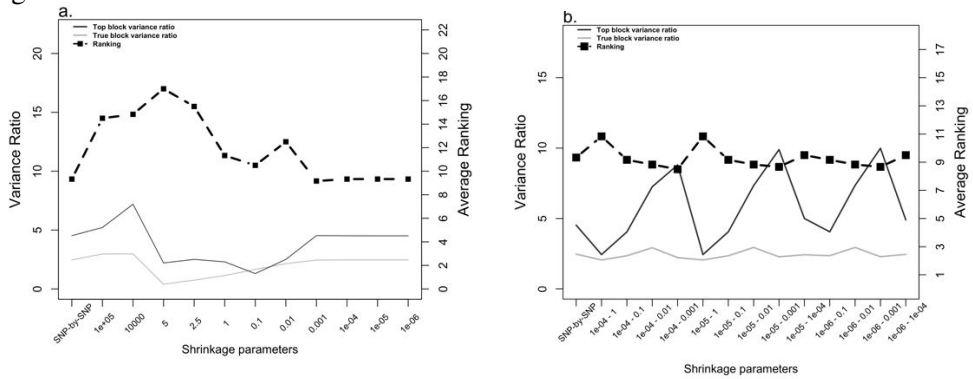


Figure 3



SUPPLEMENTARY MATERIAL

Inclusion of biological knowledge in a Bayesian shrinkage model for joint estimation of SNP effects

Pereira M, Thompson JR, Weichenberger CX, Thomas DC, Minelli C

Text S1. Statistical Analysis using the BhGLM package

We used the function $bglm()$ from the BhGLM package (http://www.ssg.uab.edu/bhglm/) and assigned different values to the function parameter `prior.scale` which corresponds to $s_{\alpha_j}^2$. The other function parameters were set as follows:

- `prior='t'`, indicates the hierarchical formulation of the model.
- `prior.mean=0`, prior mean of the SNP effects, $\mu_j = 0$.
- `mean.update=FALSE`, indicates that the prior mean is not updated at each iteration of the algorithm.
- `scale.update=FALSE`, indicates that $s_{\alpha_j}^2$ is not updated at each iteration of the model. It is fixed according to the biological information.

The `group` parameter was set so that each LD block constituted one group. This allows the algorithm to update the α_j parameters by group, which is more efficient and reduces computational time. The remaining parameters were set to default.

Text S2. Application of the approach to a binary outcome

Methods

Our approach was further tested in the scenario of a binary outcome, a type of outcome which is frequently the focus of genetic association studies. In this example, we consider a binomially distributed trait Y and use the following generalized linear model with a *logit* link function:

$$\eta_i = log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \sum_j \alpha_j x_{ij} + \varepsilon_i$$

where η_i is the linear predictor, p_i is the probability of having the trait Y, α_0 is the intercept, α_j is the effect of the j^{th} SNP, x_{ij} is the genotype of the j^{th} SNP in the i^{th} subject and ε_i is the random error. The Bayesian hierarchical formulation of the model is the following:

$$\alpha_j|\tau_{\alpha_j}^2 \sim N(\mu_j, \tau_{\alpha_j}^2)$$

$$\tau^2_{\alpha_j} | s^2_{\alpha_j} \sim Inv - \chi^2(4, s^2_{\alpha_j})$$

This hierarchical formulation is similar to the model used for the continuous outcome except that the scaled-inverse χ^2 distribution is slightly more informative with 4 degrees of freedom, to avoid extreme variance estimates.

Similarly to the continuous outcome, the performance of the model was assessed in a simulation study and in an empirical example. In both cases, the approach was tested under the scenario of constant shrinkage (no inclusion of prior information) and under the scenario of differential shrinkage based on prior information. Performance was assessed by the average ranking of the true LD block across the datasets, with the ranking of the LD block defined by the best ranking SNP in that block based on the Bayesian p-value. All results were compared to those of the standard SNP analysis with SNPs ranked based on their frequentist p-values. We also calculated the variance ratios for the top block and for the true block (top 6 blocks and true blocks in the empirical example) to assess the usefulness of these measures in the setting of a binary outcome.

Simulation study

We used the same simulated datasets derived using the software GENOME (Liang, Zöllner, & Abecasis, 2007) – 1000 datasets with 500 subjects, ~200-300 SNPs per dataset divided in 10 LD blocks. One SNP was randomly chosen to be the true SNP and assigned an effect size $\alpha_{causal} = 0.30$. This effect was used to generate a continuous trait H, so that $H \sim Bin(500, p)$, where $p = \frac{1}{1+\exp(-\alpha_{causal} \cdot SNP)}$.

The effect size was chosen based on the results of the standard SNP analysis and aimed to match the same average ranking obtained for the continuous outcome (Table 1, main manuscript).

Empirical example

We used the same dataset from ECRHS (Burney, et al., 1994) that we derived to test our approach with BMI as a continuous outcome and considered the same set of 6 true SNPs in 6 separate LD blocks out

of a total of 30 blocks. BMI was dichotomized as ≤25 vs. >25kg/m^2 (normal vs. overweight/obese) with 557 subjects in the overweight/obese category (n=1,829 subjects).

Text S3. Retrieval of prior biological information

Data retrieval was performed mainly using a local installation of the data integration framework 'Dintor', a bioinformatics tool suite that queries information from multiple online datasets and is designed for use in GWA and next-generation sequencing studies (Weichenberger et al., 2015). It includes 35 independent modules for retrieving and manipulating data, and performs operations such as SNP and gene mapping to genomic coordinates, integration of LD block information, conversion of gene and protein identifiers, orthology annotation and access to regulatory regions. Dintor can be accessed through an online GUI, however we have chosen to use the UNIX command line interface, which greatly facilitates data integration into our modelling pipeline.

Alongside Dintor, other databases/tools used were: Pfam Protein Families database (Finn et al., 2014, http://pfam.xfam.org) for questions 3 and 4, Mouse Genome Informatics (Eppig et al., 2015, http://www.informatics.jax.org) for question 8 and Reactome (Croft et al., 2014; Milacic et al., 2012, http://www.reactome.org) for question 10. The answers to the 10 questions were obtained in a semi-automatic way, as data from Pfam, Mouse Genome Informatics and Reactome need to be manually retrieved.

Supplementary Tables

Supplementary Table 1. Average ranking of the true LD block in the setting where the true LD block is assigned a larger $s^2_{\alpha_j}$ than the false blocks.

Higher $s^2_{\alpha_j}$ for all the SNPs in the true LD block					
Model	**Scale ($s^2_{\alpha_j}$)**		**Variance ratio**		**Average Ranking**
	prior support = 0	prior support = 10	Top block	True block	
Standard analysis	-	-	2.201	1.501	3.624
S1.1	10^{-4}	1	8293409736.738	8293409736.738	1.000
S1.2	10^{-4}	0.1	519970188.609	519970188.609	1.000
S1.3	10^{-4}	0.01	1263435.331	1263435.331	1.000
S1.4	10^{-4}	0.001	316.323	316.362	1.001
S1.5	10^{-5}	1	17608947561.657	17608947561.657	1.000
S1.6	10^{-5}	0.1	1104024286.997	1104024286.997	1.000
S1.7	10^{-5}	0.01	2682582.041	2682582.041	1.000
S1.8	10^{-5}	0.001	671.719	671.719	1.000
S1.9	10^{-5}	10^{-4}	3.765	3.851	2.484
S1.10	10^{-6}	0.1	1104024286.997	1104024286.997	1.000
S1.11	10^{-6}	0.01	2682582.041	2682582.041	1.000
S1.12	10^{-6}	0.001	671.719	671.719	1.000
S1.13	10^{-6}	10^{-4}	3.765	3.851	2.484

Supplementary Table 2. Average ranking of the true LD block in the setting where one false LD block has more prior knowledge than the true LD block. All the SNPs in one false LD block were randomly assigned a large $s^2_{\alpha_j}$, corresponding to maximum prior knowledge, and the SNPs in the true LD block were assigned an intermediate $s^2_{\alpha_j}$ corresponding to an intermediate prior knowledge score of 5. The remaining false SNPs were assigned a smaller $s^2_{\alpha_j}$ corresponding to absence of prior information.

Intermediate $s^2_{\alpha_j}$ for all the SNPs in the true LD block with high $s^2_{\alpha_j}$ for the SNPs in one false LD block						
Model	**Scale ($s^2_{\alpha_j}$)**			**Variance ratio**		**Average Ranking**
	prior support = 0	True LD block - prior support = 5	False LD block* - prior support = 10	Top block	True block	
Standard analysis	-	-	-	2.201	1.501	3.624
S2.1	10^{-4}	0.5	1	22.882	6.842	1.530
S2.2	10^{-4}	0.05	0.1	38.391	3.258	1.630
S2.3	10^{-4}	0.005	0.01	119.026	1.670	1.882
S2.4	10^{-4}	$5.5 \cdot 10^{-4}$	0.001	43.425	3.292	1.836
S2.5	10^{-5}	0.5	1	24.762	6.544	1.540
S2.6	10^{-5}	0.05	0.1	39.070	3.169	1.645
S2.7	10^{-5}	0.005	0.01	124.682	1.665	1.876
S2.8	10^{-5}	$5.5 \cdot 10^{-4}$	0.001	68.867	2.722	1.867
S2.9	10^{-5}	$5.5 \cdot 10^{-5}$	10^{-4}	18.101	3.110	3.003
S2.10	10^{-6}	0.05	0.1	39.738	3.162	1.638
S2.11	10^{-6}	0.005	0.01	125.424	1.522	1.902
S2.12	10^{-6}	$5.5 \cdot 10^{-4}$	0.001	62.545	2.714	1.862
S2.13	10^{-6}	$5.5 \cdot 10^{-5}$	10^{-4}	17.939	2.944	3.094

*Randomly chosen false LD block to which more prior knowledge was assigned.

Supplementary Table 3. Average ranking of the true LD block in the simulation study with a binary outcome. Results are presented for the standard SNP analysis and Bayesian model with constant shrinkage across SNPs. Variance ratios for the top block and for the true block are also presented and the best performance is reached when the variance ratio is maximized.

Constant $s^2_{\alpha_j}$ for all the SNPs				
Model	**Scale ($s^2_{\alpha_j}$)**	**Variance ratio**		**Average Ranking**
		Top block	True block	
Standard analysis	-	2.447	1.716	3.765
S3.1	10^5	2.779	1.232	5.357
S3.2	10^4	2.779	1.232	5.357
S3.3	5	2.308	1.127	5.341
S3.4	2.5	2.059	1.101	5.322
S3.5	1	1.776	1.113	5.246
S3.6	0.1	1.943	1.368	4.528
S3.7	0.01	2.371	1.706	3.773
S3.8	0.001	2.948	1.709	3.779
S3.9	10^{-4}	14.387	1.814	3.77
S3.10	10^{-5}	17.431	2.217	3.766
S3.11	10^{-6}	17.469	2.206	3.769

Supplementary Table 4. Average ranking of the true LD block in the simulation study with a binary outcome. Results are presented for the standard SNP analysis and Bayesian model with constant shrinkage across SNPs. Variance ratios for the top block and for the true block are also presented and the best performance is reached when the variance ratios are maximized.

Differential $s^2_{\alpha_j}$ according to prior information					
Model	**Scale ($s^2_{\alpha_j}$)**		**Variance ratio**		**Average Ranking**
	prior support = 0	prior support = 10	Top block	True block	
Standard analysis	-	-	2.447	1.716	3.765
S4.1	10^{-4}	1	2.728	1.929	4.439
S4.2	10^{-4}	0.1	4.447	3.168	3.355
S4.3	10^{-4}	0.01	8.528	5.135	2.932
S4.4	10^{-4}	0.001	7.343	4.078	2.957
S4.5	10^{-5}	1	2.729	1.929	4.439
S4.6	10^{-5}	0.1	4.450	3.171	3.353
S4.7	10^{-5}	0.01	8.639	5.185	2.937
S4.8	10^{-5}	0.001	8.086	4.429	2.945
S4.9	10^{-5}	10^{-4}	6.257	2.518	3.326
S4.10	10^{-6}	0.1	4.451	3.171	3.352
S4.11	10^{-6}	0.01	8.650	5.189	2.938
S4.12	10^{-6}	0.001	8.164	4.466	2.944
S4.13	10^{-6}	10^{-4}	6.437	2.596	3.298

Supplementary Table 5. Average ranking of the true LD block in the empirical example with ECRHS data and BMI as the outcome. Results are presented for the standard SNP analysis and Bayesian model with constant shrinkage across SNPs. Variance ratios for the top six blocks and for the true blocks are also presented and the best performance is reached when the variance ratio is maximized.

Constant $s^2_{\alpha_j}$ for all the SNPs				
	Scale ($s^2_{\alpha_j}$)	**Variance ratio**		**Average Ranking**
		Top 6 blocks	True blocks	
Standard analysis	-	4.226	2.316	8.167
S5.1	10^5	*	*	*
S5.2	10^4	*	*	*
S5.3	5	1.154	1.278	14.333
S5.4	2.5	1.167	1.307	14.167
S5.5	1	1.087	1.427	9.5
S5.6	0.1	1.222	1.445	13
S5.7	0.01	1.421	1.982	10.667
S5.8	0.001	4.291	2.335	8.167
S5.9	10^{-4}	4.309	2.349	8.167
S5.10	10^{-5}	4.309	2.349	8.333
S5.11	10^{-6}	4.310	2.350	8.167

*Models for which effect estimates and p-values cannot be obtained due to very low shrinkage.

Supplementary Table 6. Average ranking of the true LD blocks in the empirical example with ECRHS data and BMI as the outcome. Results are presented for the standard SNP analysis and Bayesian model when differential shrinkage is applied, with the lower and upper bounds for the shrinkage parameter values reported in columns 2 and 3, respectively. Variance ratios for the top six blocks and for the true blocks are also presented and the best performance is reached when the variance ratio is maximized.

Differential $s^2_{\alpha_j}$ according to prior information					
	Scale ($s^2_{\alpha_j}$)		**Variance ratio**		
	prior support = 0	prior support = 4	Top 6 blocks	True blocks	**Average Ranking**
Standard analysis	-	-	4.226	2.316	8.167
S6.1	10^{-4}	1	2.145	1.813	9.000
S6.2	10^{-4}	0.1	3.317	2.245	10.000
S6.3	10^{-4}	0.01	8.684	2.532	9.167
S6.4	10^{-4}	0.001	9.116	2.140	8.000
S6.5	10^{-5}	1	2.145	1.813	9.000
S6.6	10^{-5}	0.1	3.318	2.245	10.000
S6.7	10^{-5}	0.01	5.528	2.225	9.333
S6.8	10^{-5}	0.001	8.744	2.543	8.333
S6.9	10^{-5}	10^{-4}	9.667	2.178	7.667
S6.10	10^{-6}	0.1	3.318	2.246	10.000
S6.11	10^{-6}	0.01	5.459	2.231	9.333
S6.12	10^{-6}	0.001	8.750	2.544	8.500
S6.13	10^{-6}	10^{-4}	9.703	2.183	7.333

References

Burney, P. G., Luczynska, C., Chinn, S., & Jarvis, D. (1994). The European Community Respiratory Health Survey. *The European Respiratory Journal*, *7*(5), 954–60.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., … D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42*(Database issue), D472-7.

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., & Richardson, J. E. (2015). The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Research*, *43*(Database issue), D726-36.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222-30.

Liang, L., Zöllner, S., & Abecasis, G. R. (2007). GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics (Oxford, England)*, *23*(12), 1565–7.

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., … Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, *4*(4), 1180–211.

Weichenberger, C. X., Blankenburg, H., Palermo, A., D'Elia, Y., König, E., Bernstein, E., & Domingues, F. S. (2015). Dintor: functional annotation of genomic and proteomic data. *BMC Genomics*, *16*(1), 1081