# Pathways-driven Sparse Regression Identifies Pathways and Genes Associated with High-density Lipoprotein Cholesterol in Two Asian Cohorts

Matt Silver[1,2,*], Peng Chen[3], Ruoying Li[4], Ching-Yu Cheng[3,5,6], Tien-Yin Wong[5,6], E-Shyong Tai[3,4], Yik-Ying Teo[3,7,8,9,10], and Giovanni Montana[1]

[1]Statistics Section, Department of Mathematics, Imperial College London, UK
[2]MRC International Nutrition Group, London School of Hygiene and Tropical Medicine, UK
[3]Saw Swee Hock School of Public Health, National University of Singapore
[4]Yong Loo Lin School of Medicine, National University of Singapore
[5]Department of Ophthalmology, National University of Singapore
[6]Singapore Eye Research Institute, Singapore National Eye Center
[7]NUS Graduate School for Integrative Science and Engineering, National University of Singapore
[8]Life Sciences Institute, National University of Singapore
[9]Genome Institute of Singapore, Agency for Science, Technology and Research
[10]Department of Statistics and Applied Probability, National University of Singapore
[*]email: matt.silver@lshtm.ac.uk

## Abstract

Standard approaches to analysing data in genome-wide association studies (GWAS) ignore any potential functional relationships between genetic markers. In contrast gene pathways analysis uses prior information on functional structure within the genome to identify pathways associated with a trait of interest. In a second step, important single nucleotide polymorphisms (SNPs) or genes may be identified within associated pathways. The pathways approach is motivated by the fact that many genes do not act alone, but instead have effects that are likely to be mediated through their interaction in gene pathways. Where this is the case, pathways approaches may reveal aspects of a trait's genetic architecture that would otherwise be missed when considering SNPs in isolation. Most pathways methods begin by testing SNPs one at a time, and so fail to capitalise on the potential advantages inherent in a multi-SNP, joint modelling approach. Here we describe a dual-level, sparse regression model for the simultaneous identification of pathways, genes and SNPs associated with a quantitative trait. Our method takes account of various factors specific to the joint modelling of pathways with genome-wide data, including widespread correlation between genetic predictors, and the fact that variants may overlap multiple pathways. We use a resampling strategy that exploits finite sample variability to provide robust rankings for pathways, SNPs and genes. We test our method through simulation, and use it to perform pathways-driven SNP selection in a search for pathways, genes and SNPs associated with variation in serum high-density lipoprotein cholesterol (HDLC) levels in two separate GWAS cohorts of Asian adults. By comparing results from both cohorts we identify a number of candidate pathways including those associated with cardiomyopathy, and T cell receptor and PPAR signalling. Highlighted genes include those associated with the L-type calcium channel, adenylate cyclase, integrin, laminin, MAPK signalling and immune function. *Software implementing the methods described here, together with sample data is available at http://www2.imperial.ac.uk/~gmontana/psrrr.htm.*

# Introduction

Much attention continues to be focused on the problem of identifying SNPs and genes influencing a quantitative or dichotomous trait in genome wide scans [53]. Despite this, in many instances gene variants identified in GWAS have so far uncovered only a relatively small part of the known heritability of most common diseases [84]. Possible explanations include the presence of multiple SNPs with small effects, or of rare variants, which may be hard to detect using conventional approaches [84, 52, 28].

One potentially powerful approach to uncovering the genetic etiology of disease is motivated by the observation that in many cases disease states are likely to be driven by multiple genetic variants of small to moderate effect, mediated through their interaction in molecular networks or pathways, rather than by the effects of a few, highly penetrant mutations [63]. Where this assumption holds, the hope is that by considering the joint effects of variants acting in concert, pathways GWAS methods will reveal aspects of a disease's genetic architecture that would otherwise be missed when considering variants individually [86, 25]. In this section we describe a sparse regression method utilising prior information on gene pathways to identify putative causal pathways, along with the constituent SNPs and genes that may be driving pathways association.

Sparse modelling approaches are becoming increasingly popular for the analysis of genome wide datasets [65, 17, 3, 88]. Sparse regression models enable the joint modelling of large numbers of SNP predictors, and perform 'model selection' by highlighting small numbers of variants influencing the trait of interest. These models work by penalising or constraining the size of estimated regression coefficients. An interesting feature of these methods is that different sparsity patterns, that is different sets of genetic predictors having specified properties, can be obtained by varying the nature of this constraint. For example, the lasso [78] selects a subset of variants whose main effects best predict the response. Where predictors are highly correlated, the lasso tends to select one of a group of correlated predictors at random. In contrast, the elastic net [92] selects groups of correlated variables. Model selection may also be driven by external information, unrelated to any statistical properties of the data being analysed. For example, the fused lasso [80, 79] uses ordering information, such as the position of genomic features along a chromosome to select 'adjacent' features together.

Prior information on functional relationships between genetic predictors can also been used to drive the selection of groups of variables. In the present context, information mapping genes and SNPs to functional gene pathways has recently been used in sparse regression models for pathway selection. Chen et al. [15] describe a method that uses a combination of lasso and ridge regression to assess the significance of association between a candidate pathway and a dichotomous (case-control) phenotype, and apply this method in a study of colon cancer etiology. In contrast, Silver et al. [67] use group lasso penalised regression to select pathways associated with a multivariate, quantitative brain-imaging phenotype characteristic of structural change in the brains of patients with Alzheimer's disease.

In identifying pathways associated with a trait of interest, a natural follow-up question is to ask which SNPs and/or genes are driving pathway selection? We might further ask a related question: can the use of prior information on putative gene interactions within pathways increase power to identify causal SNPs, compared to alternative methods that disregard such information? One way to answer these questions is by conducting a two-stage analysis, in which we first identify important pathways, and then in a second step search for SNPs within selected pathways [20, 21]. There are however a number of problems with this approach.

Firstly, highlighted SNPs are then not necessarily those that were driving pathway selection in the first step of the analysis. Secondly, the implicit (and reasonable) assumption is that only a small number of SNPs in a pathway are driving pathway selection, so that ideally we would prefer a model that has this assumption built in. The above considerations point to the use of a 'dual-level' sparse regression model that imposes sparsity at both the pathway and SNP level. Such a model would perform *simultaneous* pathway and SNP selection, with the additional benefit of being simpler to implement.

A suitable sparse regression model enforcing the required dual-level sparsity is the sparse group lasso (SGL) [69]. SGL is a comparatively recent development in sparse modelling, and in simulations has been shown to accurately recover dual-level sparsity, in comparison to both the group lasso and lasso [26, 69]. SGL has been used for the identification of rare variants in a case-control study by grouping SNPs into genes [91]; for the identification of genomic regions whose copy number variations have an impact on RNA expression levels [59]; and to model geographical factors driving climate change [14]. SGL can be seen as fitting into a wider class of structured-sparsity inducing models that use prior information on relationships between predictors to enforce different sparsity patterns [90, 38, 41].

In the next section (Methods) we outline our method for sparse, pathways-driven SNP selection, and demonstrate through simulation that the incorporation of prior information mapping SNPs to gene pathways can boost the power to detect SNPs associated with a quantitative trait. In the following section (Results), we describe an application study, in which we investigate pathways, SNPs and genes associated with serum high-density lipoprotein cholesterol (HDLC) levels in two separate cohorts of Asian adults. HDLC refers to the cholesterol carried by small lipoprotein molecules, so called high density lipoproteins (HDLs). HDLs help remove the cholesterol aggregating in arteries, and are therefore protective against cardiovascular diseases [81]. Serum HDLC levels are genetically heritable ($h^2 = 0.485$) [57]. GWAS studies have now uncovered more than 100 HDLC associated loci (see www.genome.gov/gwastudies, Hindorff et al. [32]). However, considering serum lipids as a whole, variants so far identified account for only 25-30% of the genetic variance, highlighting the limited power of current methodologies to detect hidden genetic factors [76].

# Methods

This section is organised as follows. We begin by introducing the sparse group lasso (SGL) model for pathways-driven SNP selection, along with an efficient estimation algorithm, for the case of non-overlapping pathways. We then describe a simulation study illustrating superior group (pathway) and variant (SNP) selection performance in the case that the true supporting model is group-sparse. We continue by extending the previous model to the case of overlapping pathways. In principle, we can then solve this model using the estimation algorithm described for the non-overlapping case. However, we argue that this approach does not give us the outcome we require. For this reason we describe a modified estimation algorithm that assumes pathway independence, and demonstrate in a simulation study that this new algorithm is able to identify the correct SNPs and pathways with improved sensitivity and specificity. We complete this section with a description of a method to reduce bias in SNP and pathway selection, together with a subsampling procedure to rank SNPs and pathways in order of importance.

## The sparse group lasso model

We arrange the observed values for a univariate quantitative trait or phenotype, measured for $N$ unrelated individuals, in an $(N \times 1)$ response vector $\mathbf{y}$. We assume minor allele counts for $P$ SNPs are recorded for all individuals, and denote by $x_{ij}$ the minor allele count for SNP $j$ on individual $i$. These are arranged in an $(N \times P)$ genotype design matrix $\mathbf{X}$. Phenotype and genotype vectors are mean centred, and SNP genotypes are standardised to unit variance, so that $\sum_i x_{ij}^2 = 1$, for $j = 1, \dots, P$.

We assume that all $P$ SNPs may be mapped to $L$ groups or pathways, $\mathcal{G}_l \subset \{1, \dots, P\}$, $l = 1, \dots, L$, and begin by considering the case where pathways are disjoint or non-overlapping, so that $\mathcal{G}_l \cap \mathcal{G}_{l'} = \emptyset$ for any $l \neq l'$. We denote the vector of SNP regression coefficients by $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)$, and additionally denote the matrix containing all SNPs mapped to pathway $\mathcal{G}_l$ by $\mathbf{X}_l = (\mathbf{x}_{l_1}, \mathbf{x}_{l_2}, \dots, \mathbf{x}_{P_l})$, where $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{Nj})'$, is the column vector of observed SNP minor allele counts for SNP $j$, and $P_l$ is the number of SNPs in $\mathcal{G}_l$. We denote the corresponding vector of SNP coefficients by $\boldsymbol{\beta}_l = (\beta_{l_1}, \beta_{l_2}, \dots, \beta_{P_l})$.

In general, where $P$ is large, we expect only a small proportion of SNPs to be 'causal', in the sense that they exhibit phenotypic effects. A key assumption in pathways analysis is that these causal SNPs will tend to be enriched within a small set, $\mathcal{C} \subset \{1, \dots, L\}$, of causal pathways, with $|\mathcal{C}| \ll L$, where $|\mathcal{C}|$ denotes the size (cardinality) of $\mathcal{C}$. We denote the set of causal SNPs mapping to pathway $\mathcal{G}_l$ by $\mathcal{S}_l$, and make the further assumption that most SNPs in a causal pathway are non-causal, so that $|\mathcal{S}_l| < P_l$, where $|\mathcal{S}_l|$ denotes the size (cardinality) of $\mathcal{S}_l$. A suitable sparse regression model imposing the required, dual-level sparsity pattern is the sparse group lasso (SGL). We illustrate the resulting causal SNP sparsity pattern in Figure 1, and compare it to that generated by the group lasso (GL), a group-sparse model that we used previously in a sparse regression method to identify gene pathways [66, 67].

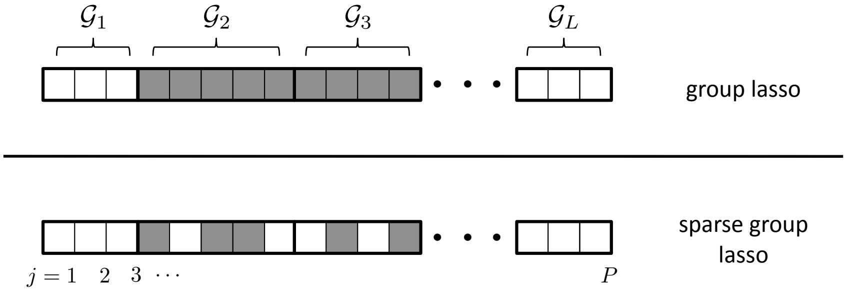


Figure 1: Sparsity patterns enforced by the group lasso and sparse group lasso. The set $\mathcal{S} \subset \{1, \dots, P\}$ of causal SNPs influencing the phenotype are represented by boxes that are shaded grey. Causal SNPs are assumed to occur within a set $\mathcal{C} \subset \{1, \dots, L\}$ of causal pathways, $\mathcal{G}_1, \dots, \mathcal{G}_L$. Here $\mathcal{C} = \{2, 3\}$. The group lasso enforces sparsity at the group or pathway level only, whereas the sparse group lasso additionally enforces sparsity at the SNP level.

With the SGL [69], sparse estimates for the SNP coefficient vector, $\boldsymbol{\beta}$ are given by

$$\hat{\boldsymbol{\beta}}^{SGL} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + (1 - \alpha)\lambda \sum_{l=1}^{L} w_l ||\boldsymbol{\beta}_l||_2 + \alpha\lambda ||\boldsymbol{\beta}||_1 \right\} \tag{1}$$

where $\lambda$ $(\lambda > 0)$ and $\alpha$ $(|\alpha| \leq 1)$ are parameters controlling sparsity, and $w_l$ is a pathway weighting parameter that may vary across pathways. (1) corresponds to an ordinary least squares (OLS) optimisation, but with two additional constraints on the coefficient vector, $\boldsymbol{\beta}$, that tend to shrink the size of $\boldsymbol{\beta}$, relative to OLS estimates. One constraint imposes a group lasso-type penalty on the size ($\ell_2$ norm) of $\boldsymbol{\beta}_l, l = 1, \ldots, L$. Depending on the values of $\lambda$, $\alpha$ and $w_l$, this penalty has the effect of setting multiple pathway SNP coefficient vectors, $\hat{\boldsymbol{\beta}}_l = \mathbf{0}$, thereby enforcing sparsity at the pathway level. Pathways with non-zero coefficient vectors form the set $\hat{\mathcal{C}}$ of 'selected' pathways, so that

$$\hat{\mathcal{C}}(\lambda, \alpha) = \{l : \hat{\boldsymbol{\beta}}_l \neq \mathbf{0}\}.$$

A second constraint imposes a lasso-type penalty on the size ($\ell_1$ norm) of $\boldsymbol{\beta}$. Depending on the values of $\lambda$ and $\alpha$, for a selected pathway $l \in \hat{\mathcal{C}}$, this penalty has the effect of setting multiple SNP coefficient vectors, $\hat{\beta}_j = 0, j \subset \mathcal{G}_l$, thereby enforcing sparsity at the SNP level within selected pathways. SNPs with non-zero coefficient vectors then form the set $\hat{\mathcal{S}}_l$ of selected SNPs in pathway $l$, so that

$$\hat{\mathcal{S}}_l(\lambda, \alpha) = \{j : \hat{\beta}_j \neq 0, j \in \mathcal{G}_l\}.$$

The set of all selected SNPs is given by

$$\hat{\mathcal{S}} = \bigcup_{l \in \hat{\mathcal{C}}} \hat{\mathcal{S}}_l.$$

The sparsity parameter $\lambda$ controls the degree of sparsity in $\boldsymbol{\beta}$, such that the number of pathways and SNPs selected by the model increases as $\lambda$ is reduced from a maximal value $\lambda_{max}$, above which $\hat{\boldsymbol{\beta}} = \mathbf{0}$. The parameter $\alpha$ controls how the sparsity constraint is distributed between the two penalties. When $\alpha = 0$, (1) reduces to the group lasso, so that sparsity is imposed only at the pathway level, and all SNPs within a selected pathway have non-zero coefficients. When $0 < \alpha < 1$, solutions exhibit dual-level sparsity, such that as $\alpha$ approaches 0 from above, greater sparsity at the group level is encouraged over sparsity at the SNP level. When $\alpha = 1$, (1) reverts to the lasso, so that pathway information is ignored.

## Model estimation

For the estimation of $\hat{\boldsymbol{\beta}}^{SGL}$ we proceed by noting that the optimisation (1) is convex, and (in the case of non-overlapping groups) that the penalty is block-separable, so that we can obtain a solution using block, or group-wise coordinate gradient descent (BCGD) [82]. A detailed derivation of the estimation algorithm is given in the accompanying supplementary information, Section A.

From (17) and (18), the criterion for selecting a pathway $l$ is given by

$$||S(\mathbf{X}_l' \hat{\mathbf{r}}_l, \alpha\lambda)||_2 > (1 - \alpha)\lambda w_l, \tag{2}$$

and the criterion for selecting SNP $j$ in selected pathway $l$ by

$$||X_j' \hat{\mathbf{r}}_{l,j}||_1 > \alpha\lambda, \tag{3}$$

where $\hat{\mathbf{r}}_l = \hat{\mathbf{r}}_l - \sum_{m \neq l} \mathbf{X}_l \hat{\boldsymbol{\beta}}_l$ and $\hat{\mathbf{r}}_{l,j} = \hat{\mathbf{r}}_l - \sum_{k \neq j} X_k \hat{\beta}_k$ are respectively the pathway and SNP partial residuals, obtained by regressing out the current estimated effects of all other pathways and SNPs respectively. The complete algorithm for SGL estimation using BCGD is presented in Box 1.

**Box 1** SGL-BCGD estimation algorithm

```
1. initialise $\boldsymbol{\beta} \leftarrow \mathbf{0}$.

2. repeat: [pathway loop]
       for pathway $l = 1, 2, \ldots, L$:
           if $||S(\mathbf{X}_l'\hat{\mathbf{r}}_l, \alpha\lambda)||_2 \leq (1-\alpha)\lambda w_l$
               $\boldsymbol{\beta}_l \leftarrow \mathbf{0}$
           else
               repeat: [SNP loop]
                   for $j = l_1, \ldots, l_{P_l}$:
                       if $\beta_j = 0$ :
                           Newton update $\beta_j^{**} \leftarrow \beta_j$ using (24) and (22)
                       else:
                           Newton update $\beta_j^{**} \leftarrow \beta_j$ using (21) and (22)
                       if $f(\boldsymbol{\beta}_l^{**}) > f(\boldsymbol{\beta}_l)$:
                           $\beta_j^{**} \leftarrow \frac{\beta_j^{**}+\beta_j}{2}$
                       $\beta_j \leftarrow \beta_j^{**}$
               until convergence of $\boldsymbol{\beta}_l$ [SNP loop]
   until convergence of $\boldsymbol{\beta}$ [pathway loop]

3. $\hat{\boldsymbol{\beta}}^{SGL} \leftarrow \boldsymbol{\beta}$
```

## SGL simulation study 1

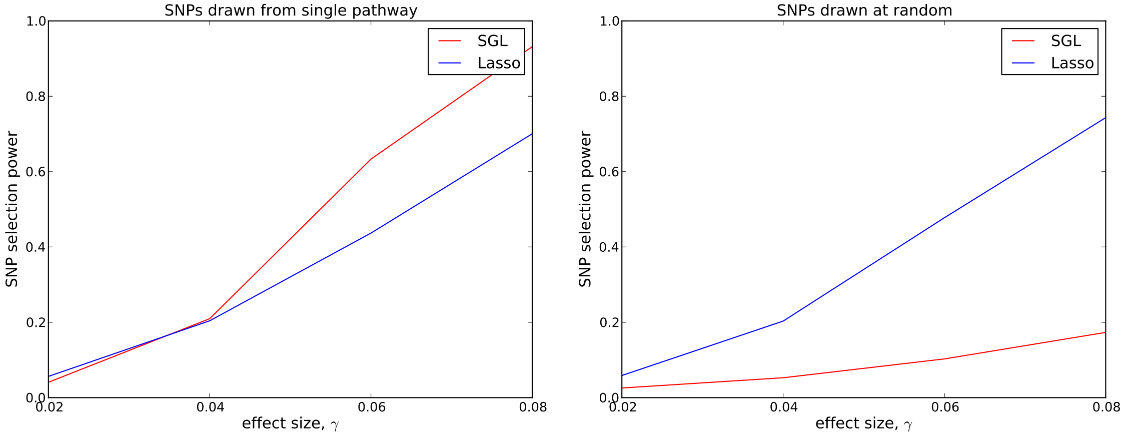
We test the hypothesis that where causal SNPs are enriched in a given pathway, pathway-driven SNP selection using SGL will outperform simple lasso selection that disregards pathway information in a simple simulation study. We simulate $P = 2500$ genetic markers for $N = 400$ individuals. Marker frequencies for each SNP are sampled independently from a multinomial distribution following a Hardy Weinberg equilibrium frequency distribution. SNP minor allele frequencies are sampled from a uniform distribution $\mathcal{U}[0.1, 0.5]$. SNPs are distributed equally between 50 non-overlapping pathways, each containing 50 SNPs.

We then test each competing method over 500 Monte Carlo (MC) simulations. At each simulation, a baseline univariate phenotype is sampled from $\mathcal{N}(10, 1)$. To generate genetic effects, we randomly select 5 SNPs from a single, randomly selected pathway $\mathcal{G}_l$, to form the set $\mathcal{S} \subset \mathcal{G}_l$ of causal SNPs. Genetic effects are then generated as described in Section C.

To enable a fair comparison between the two methods (SGL and lasso), we ensure that both methods select the same number of SNPs at each simulation. We do this by first obtaining the SGL solution, $\hat{\mathcal{S}}^{SGL}$, with $\lambda = 0.85\lambda_{max}$ and $\alpha = 0.8$, which ensures sparsity at both the pathway and SNP level. We use a uniform pathway weighting vector $\boldsymbol{w} = \mathbf{1}$. We then compute the lasso solution using coordinate descent over a range of values for the lasso regularisation penalty, $\lambda$, and choose the set

$$\hat{\mathcal{S}}^{lasso}(\lambda') \quad \text{such that} \quad |\hat{\mathcal{S}}^{lasso}(\lambda')| = |\hat{\mathcal{S}}^{SGL}|$$

where $|\hat{\mathcal{S}}^{SGL}|$ is the number of SNPs previously selected by SGL, and $|\hat{\mathcal{S}}^{lasso}(\lambda')|$ is the number of SNPs selected by the lasso with $\lambda = \lambda'$. We measure performance as the mean power to detect all 5 causal SNPs over 500 MC simulations, and test a range of genetic effect sizes



(a)

(b)

Figure 2: SGL vs Lasso. Comparison of power to detect 5 causal SNPs. Each data point represents mean power over 500 MC simulations. *Left:* Causal SNPs drawn from single causal pathway. *Right:* Causal SNPs drawn at random.

($\gamma$) (see C). In a follow up study, we compare the performance of the two methods in a scenario in which pathways information is uninformative. For this we repeat the previous simulations, but with 5 causal SNPs drawn at random from all 2500 SNPs, irrespective of pathway membership. Results are presented in Figure 2.

Referring to Figure 2, we see that where causal SNPs are concentrated in a single causal pathway (Figure 2 - left), SGL demonstrates greater power (and equivalently specificity, since the total number of selected SNPs is constant), compared with the lasso, above a particular effect size threshold (here $\gamma \approx 0.04$). Where pathway information is not important, that is causal SNPs are not enriched in any particular pathway (Figure 2 - right), SGL performs poorly.

To gain a deeper understanding of what is happening here, we also consider the power distributions across all 500 MC simulations corresponding to each point in the plots of Figure 2. These are illustrated in Figure 3. The top row of plots illustrates the case where causal SNPs are drawn from a single causal pathway. Here we see that there is a marked difference between the two distributions (SGL vs lasso). The lasso shows a smooth distribution in power, with mean power increasing with effect size. In contrast, with SGL the distribution is almost bimodal, with power typically either 0 or 1, depending on whether or not the correct causal pathway is selected. This serves as an illustration of the advantage of pathway-driven SNP selection for the detection of causal SNPs in the case that pathways are important. As previously found by Zhou et al. [91] in the context of rare variants and gene selection, the joint modelling of SNPs within groups gives rise to a relaxation of the penalty on individual SNPs within selected groups, relative to the lasso. This can enable the detection of SNPs with small effect size or low MAF that are missed by the lasso, which disregards pathways information and treats all SNPs equally. Finally, where causal SNPs are not enriched in a causal pathway (bottom row of Figure 3), as expected SGL performs poorly. In this case SGL will only select a SNP where the combined effects of constituent SNPs in a pathway are large enough to drive pathway selection.

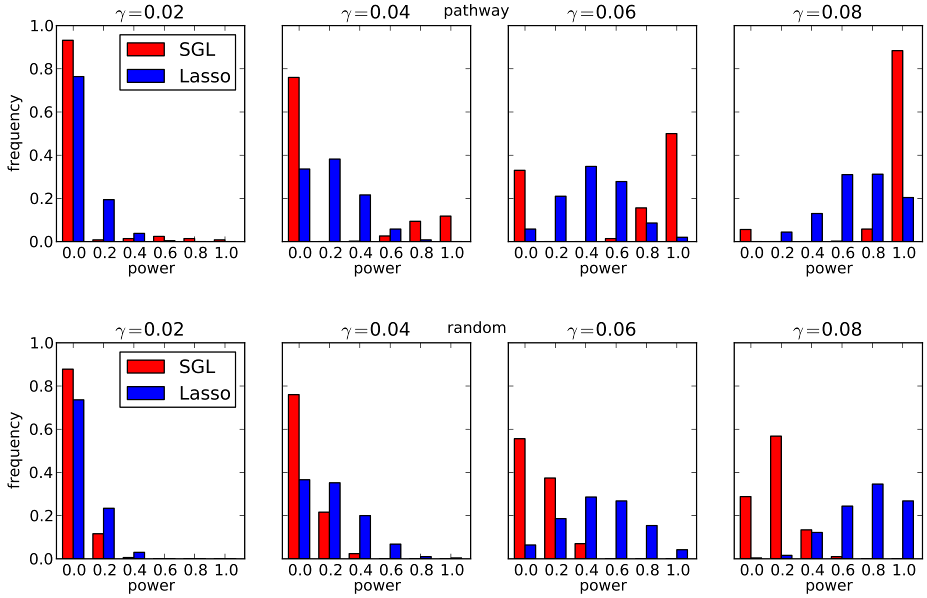


Figure 3: SGL vs Lasso. Distribution over 500 MC simulations of power to detect 5 causal SNPs. Each plot represents the power distribution at a single data point in Figure 2. The power distribution is discrete, since each method can identify $0, 1, 2, 3, 4$ or $5$ causal SNPs, with corresponding power $0, 0.2, 0.4, 0.6, 0.8$ or $1.0$. *Top row:* Causal SNPs drawn from single causal pathway. *Bottom row:* Causal SNPs drawn at random.

## The problem of overlapping pathways

The assumption that pathways are disjoint does not hold in practice, since genes and SNPs may map to multiple pathways (see Figure 7). This means that typically $\mathcal{G}_l \cap \mathcal{G}_{l'} \neq \emptyset$ for some $l \neq l'$. In the context of pathways-driven SNP selection using SGL, this has two important implications. Firstly, the optimisation (1) is no longer separable into groups (pathways), so that convergence using coordinate descent is no longer guaranteed [82]. Secondly, we wish to be able to select pathways independently, and the SGL model as previously described does not allow this . For example consider the case of an overlapping gene, that is a gene that maps to more than one pathway. If a SNP mapping to this gene is selected in one pathway, then it must be selected in each and every pathway containing the mapped gene, so that all pathways mapping to the gene are selected. We instead want to admit the possibility that the joint SNP effects in one pathway may be sufficient to allow pathway selection, while the joint effects in another pathway containing some of the same SNPs do not pass the threshold for pathway selection.

A solution to both these problems is obtained by duplicating SNP predictors in $\mathbf{X}$, so that SNPs belonging to more than one pathway can enter the model separately [39, 66]. The process works as follows. An expanded design matrix is formed from the column-wise concatenation of the $L, (N \times P_l)$ sub-matrices, $\mathbf{X}_l$, to form the expanded design matrix $\mathbf{X}^* = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L]$ of size $(N \times P^*)$, where $P^* = \sum_l P_l$. The corresponding $P^* \times 1$ parameter vector, $\boldsymbol{\beta}^*$, is formed by joining the $L, (P_l \times 1)$ pathway parameter vectors, $\boldsymbol{\beta}_l^*$, so that $\boldsymbol{\beta}^* = [\boldsymbol{\beta}_1^{*\prime}, \boldsymbol{\beta}_2^{*\prime}, \ldots, \boldsymbol{\beta}_L^{*\prime}]'$. Pathway mappings with SNP indices in the expanded variable space are reflected in updated

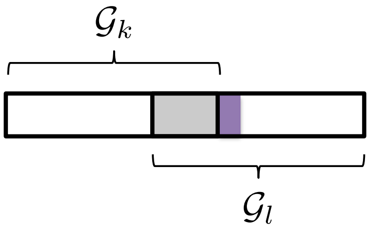


Figure 4: Two pathways with partially overlapping causal SNPs. Causal SNPs (marked in grey) in the set $\mathcal{S}_k$ overlap both pathways, so that $\mathcal{S}_k = \mathcal{G}_k \cap \mathcal{G}_l$. Additional causal SNPs, $\mathcal{S}_l \cap \setminus \mathcal{S}_k$, (marked in purple) occur in pathway $l$ only.

groups $\mathcal{G}_1^*, \ldots, \mathcal{G}_L^*$. The SGL estimator (1), adapted to account for overlapping groups, is then given by

$$\hat{\boldsymbol{\beta}}^{SGL*} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\mathbf{y} - \mathbf{X}^* \boldsymbol{\beta}^*||_2^2 + (1-\alpha)\lambda \sum_{l=1}^{L} w_l ||\boldsymbol{\beta}_l^*||_2 + \alpha\lambda ||\boldsymbol{\beta}^*||_1 \right\}. \tag{4}$$

With this overlap expansion, the model is then able to perform pathway and SNP selection in the way that we require, and the corresponding optimisation problem is amenable to solution using the BCGD estimation algorithm described in Box 1. However, for the purpose of pathways-driven SNP selection, the application of this algorithm presents a problem. This arises from the replication of overlapping SNP predictors in each group, $\mathbf{X}_l^*$, that they occur.

Consider for example the simple situation where there are two pathways, $\mathcal{G}_k^*, \mathcal{G}_l^*$, containing sets of causal SNPs $\mathcal{S}_k^* \subseteq \mathcal{G}_k^*$ and $\mathcal{S}_l^* \subseteq \mathcal{G}_l^*$ respectively. Here the $^*$ indicates that SNP indices refer to the expanded variable space. We begin by assuming that $\mathcal{S}_k^*$ and $\mathcal{S}_l^*$ contain the same SNPs, so that in the *unexpanded* variable space, $\mathcal{S}_k = \mathcal{S}_l$.

We then proceed with BCGD by first estimating $\boldsymbol{\beta}_k^*$. We assume that the correct SNPs are selected, so that $\{\hat{\beta}_j^* \neq 0 : j \in \mathcal{S}_k^*\}$, and $\hat{\beta}_j^* = 0$ otherwise. For the estimation of $\boldsymbol{\beta}_l^*$, the estimated effect $\sum_{j \in \mathcal{S}_k^*} X_j^* \hat{\beta}_j^*$, of these overlapping causal SNPs is removed from the regression, through its incorporation in the block residual $\hat{\mathbf{r}}_l^* = \mathbf{y} - \sum_{j \in \mathcal{S}_k^*} X_j^* \hat{\beta}_j^*$. Since no other causal SNPs exist in pathway $\mathcal{G}_l^*$, $\mathbf{X}_l^{*'} \hat{\mathbf{r}}_l^* = \mathbf{0}$, so that the criterion for pathway selection, $||S(\mathbf{X}_l^{*'} \hat{\mathbf{r}}_l^*, \alpha\lambda)||_2 > (1-\alpha)\lambda w_l$ (2) is not met. That is $\mathcal{G}_l^*$ is not selected.

Now consider the case where additional, non-overlapping causal SNPs, possibly with smaller effects, occur in $\mathcal{G}_l^*$, so that in the unexpanded variable space, $\mathcal{S}_k \subset \mathcal{S}_l$. In other words, causal SNPs are *partially overlapping* (see Fig 4)[1]. During BCGD pathway $\mathcal{G}_l^*$ is then less likely to be selected by the model, than would be the case if there were no overlapping SNPs, since once again the effects of overlapping causal SNPs, $\mathcal{S}_k \cap \mathcal{S}_l = \mathcal{S}_k$, are removed.

For pathways-driven SNP selection, we will argue that we instead require that SNPs are selected in each and every pathway whose joint SNP effects pass a revised pathway selection threshold, irrespective of overlaps between pathways. This is equivalent to the previous pathway selection criterion (2), but with the additional assumption that pathways are inde-

[1] This is the situation for example where multiple causal genes overlap both pathways, but one or more additional causal genes occur in $\mathcal{G}_l$.

pendent, in the sense that they do not compete in the model estimation process. We describe a revised estimation algorithm under the assumption of pathway independence below.

We justify the strong assumption of pathway independence with the following argument. In reality, we expect that multiple pathways may simultaneously influence the phenotype, and we also expect that many such pathways will overlap, for example through their containing one or more 'hub' genes, that overlap multiple pathways [46, 48]. By considering each pathway independently, we aim to maximise the sensitivity of our method to detect these variants and pathways. In contrast, without the independence assumption, a competitive estimation algorithm will tend to pick out one from each set of similar, overlapping pathways, and miss potentially causal pathways and variants as a consequence. We illustrate this idea in the simulation study in the following section. One potential concern is that by not allowing pathways to compete against each other, specificity may be reduced, since too many pathways and SNPs may be selected. We discuss the issue of specificity further in the context of results from the simulation study.

A detailed derivation of the SGL model estimation algorithm under the independence assumption is given in supplementary information, Section B. The main results are that the pathway (2) and SNP (3) selection criteria become

$$\begin{aligned} ||S(\mathbf{X}_l'\mathbf{y}, \alpha\lambda)||_2 &> (1-\alpha)\lambda w_l, \quad \text{and} \\ ||X_j'\mathbf{y}||_1 &> \alpha\lambda \end{aligned} \tag{5}$$

respectively. The key difference is that partial derivatives $\hat{\mathbf{r}}_l$ and $\hat{\mathbf{r}}_{l,j}$ are replaced by $\mathbf{y}$, that is each pathway is regressed against the phenotype vector $\mathbf{y}$. This means that there is no block coordinate descent stage in the estimation, so that the revised algorithm utilises only coordinate gradient descent within each selected pathway. For this reason we use the acronym SGL-CGD for the revised algorithm, and SGL-BCGD for the previous algorithm using block coordinate gradient descent. The new algorithm is described in Box 2.

Finally, we note that for SNP selection we are interested only in the set $\hat{\mathcal{S}}$ of selected SNPs in the unexpanded variable space, and not the set $\mathcal{S}^* = \{j^* : \beta_j^* \neq 0, j^* \in \{1, \ldots, P^*\}\}$. Since, under the independence assumption, the estimation of each $\boldsymbol{\beta}_l^*$ does not depend on the other estimates, $\boldsymbol{\beta}_k^*, k \neq l$, we do not need to record separate coefficient estimates for each pathway in which a SNP is selected. Instead we need only record the set $\hat{\mathcal{S}}_l, l \in \hat{\mathcal{C}}$ of SNPs selected in each selected pathway. This has a useful practical implication, since we can avoid the need for an expansion of $\mathbf{X}$ or $\boldsymbol{\beta}$, and simply form the complete set of selected SNPs as

$$\hat{\mathcal{S}} = \bigcup_{l \in \hat{\mathcal{C}}} \hat{\mathcal{S}}_l.$$

## SGL simulation study 2

We now explore some of the issues raised in the preceding section, specifically the potential impact on pathway and SNP selection power and specificity of treating the pathways as independent in the SGL estimation algorithm. We do this in a simulation study in which we simulate overlapping pathways. The simulation scheme is specifically designed to highlight differences in pathway and SNP selection with the independence assumption (using the SGL-CGD estimation algorithm in Box 2) and without it (using the standard SGL estimation algorithm in Box 1).

**Box 2** SGL-CGD estimation algorithm for overlapping pathways

```
1. initialise $\hat{\boldsymbol{\beta}}^* \leftarrow \mathbf{0}$.

2. for pathway $l = 1, 2, \ldots, L$:
       if $||S(\mathbf{X}_l^{*\prime}\mathbf{y}, \alpha\lambda)||_2 \leq (1-\alpha)\lambda w_l$
           $\hat{\boldsymbol{\beta}}_l^* \leftarrow \mathbf{0}$
       else
           repeat: [CGD (SNP) loop]
               for $j = l_1, \ldots, l_{P_l}$:
                   if $\hat{\beta}_j^* = 0$ :
                       Newton update $\hat{\beta}_j^{**} \leftarrow \hat{\beta}_j^*$ using (31) and (22)
                   else:
                       Newton update $\hat{\beta}_j^{**} \leftarrow \hat{\beta}_j^*$ using (30) and (22)
                   if $f(\boldsymbol{\beta}_l^{**}) > f(\boldsymbol{\beta}_l^*)$:
                       $\hat{\beta}_j^{**} \leftarrow \frac{\hat{\beta}_j^{**} + \hat{\beta}_j^*}{2}$
                   $\hat{\beta}_j^* \leftarrow \hat{\beta}_j^{**}$
           until convergence

3. $\hat{\boldsymbol{\beta}}^{SGL} \leftarrow \boldsymbol{\beta}^*$
```

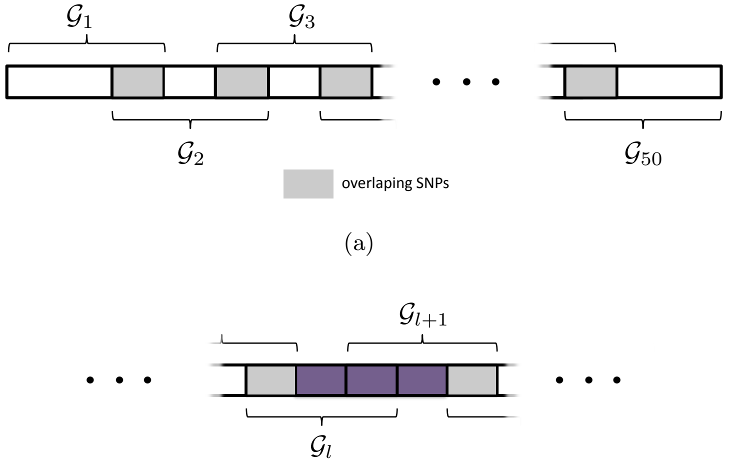
SNPs with variable MAF are simulated using the same procedure described in the previous simulation study, but this time SNPs are mapped to 50 *overlapping* pathways, each containing 30 SNPs. Each pathway overlaps any adjacent (by pathway index) pathway by 10 SNPs. This overlap scheme is illustrated in Figure 5 (a).

As before we consider a range of overall genetic effect sizes, $\gamma$. A total of 2000 MC simulations are conducted for each effect size. At MC simulation $z$, we randomly select two adjacent pathways, $\mathcal{G}_l, \mathcal{G}_{l+1}$ where $l \in \{1, \ldots, 49\}$. From these two pathways we randomly select 10 SNPs according to the scheme illustrated in Figure 5 (b). This ensures that causal SNPs overlap a minimum of 1, and a maximum of 2 pathways, with $\mathcal{S}_z \subset (\mathcal{G}_l \cap \backslash \mathcal{G}_{l-1}) \cup (\mathcal{G}_{l+1} \cap \backslash \mathcal{G}_{l+2})$. The true set of causal pathways, $\mathcal{C}$, is then given by $\{l\}$, $\{l+1\}$ or $\{l, l+1\}$ (although simulations where $|\mathcal{C}| = 1$ will be extremely rare). Genetic effects on the phenotype are generated as described previously (Section C).

SNP coefficients are estimated for each algorithm, SGL-BCGD and SGL-CGD, using the same regularisation with $\lambda = 0.85\lambda_{max}$ and $\alpha = 0.85$ for both.

The average number of pathways and SNPs selected by SGL-BCGD and SGL-CGD across all 2000 MC simulations is reported in Table 1. As expected, for both models, the number of selected variables (pathways or SNPs) increases with decreasing effect size, as the number of pathways close to the selection threshold set by $\lambda_{max}$ increases.

For each model, at MC simulation $z$ we record the pathway and SNP selection power, $|\hat{\mathcal{C}}_z \cap \mathcal{C}_z| / |\mathcal{C}_z|$ and $|\hat{\mathcal{S}}_z \cap \mathcal{S}_z| / |\mathcal{S}_z|$ respectively. Since the number of selected variables can vary slightly between the two models, we also record false positive rates (FPR) for pathway and



(a)

(b)

Figure 5: SGL Simulation Study with overlapping pathways. (a): Illustration of pathway overlap scheme. The are 30 SNPs in each pathway. Pathways $\mathcal{G}_l, (l = 1, \ldots, 50)$ overlap each adjacent pathway by 10 SNPs. (b): Causal SNPs from adjacent pathways, $l, l+1$ are randomly selected from the region marked in purple, ensuring that SNPs in $\mathcal{S}$ overlap a maximum of two pathways.

Table 1: Simulation Study 2: Mean number of pathways and SNPs selected by each model at each effect size, $\gamma$, across 2000 MC simulations.

| | | $\gamma$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.02 | 0.04 | 0.06 | 0.08 | 0.1 | 0.12 |
| pathways | SGL-CGD | 5.8 | 5.9 | 5.4 | 4.8 | 3.9 | 3.2 |
| | SGL-BCGD | 5.8 | 5.9 | 5.4 | 4.8 | 3.9 | 3.2 |
| SNPs | SGL-CGD | 26.6 | 27.0 | 24.8 | 22.2 | 18.5 | 15.3 |
| | SGL-BCGD | 28.8 | 29.3 | 26.7 | 23.6 | 19.4 | 15.8 |

SNP selection as $|\hat{\mathcal{C}}_z \cap \backslash \mathcal{C}_z|/|\hat{\mathcal{C}}_z|$ and $|\hat{\mathcal{S}}_z \cap \backslash \mathcal{S}_z|/|\hat{\mathcal{S}}_z|$ respectively.

The large possible variation in causal SNP distributions, causal SNP MAFs etc. make a comparison of mean power and FPR between the two methods somewhat unsatisfactory. For example, depending on effect size, a large number of simulations can have either very high, or very low pathway and SNP selection power, masking subtle differences in performance between the two methods. Since we are specifically interested in establishing the relative performance of the two methods, we instead illustrate the number of simulations at which one method outperforms the other across all 2000 MC simulations, and show this in Figure 6. In this figure, the number of simulations in which SGL-CGD outperforms SGL, i.e. where SGL-CGD power > SGL-BCGD power, or SGL-CGD FPR < SGL-BCGD FPR, are shown in green. Conversely, the number of simulations where SGL-BCGD outperforms SGL-CGD are shown in red.

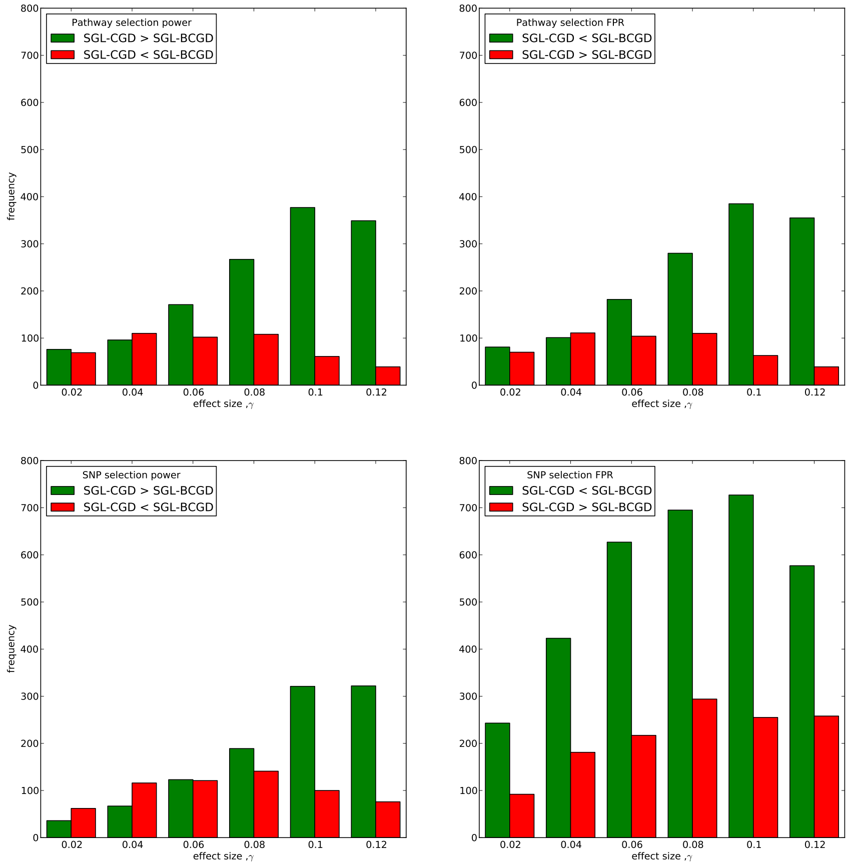


Figure 6: SGL-CGD vs SGL-BCGD performance, measured across 2000 MC simulations. *Top row:* Pathway selection performance. (Left) green bars indicate the number of MC simulations where SGL-CGD has greater pathway selection power than SGL. Red bars indicate where SGL-BCGD has greater power than SGL-CGD. (Right) green bars indicate the number of MC simulations where SGL-CGD has a lower FPR than SGL. Red bars indicate the opposite. *Bottom row:* As above, but for SNP selection performance.

We first consider pathway selection performance (top row of Figure 6). For both methods, the same number of pathways are selected on average, across all effect sizes (Table 1). At low effect sizes, there is no difference in performance between the two methods for the large majority of MC simulations, and where there is a difference, the two methods are evenly balanced. As with SGL Simulation Study 1, this is the region (with $\gamma \leq 0.04$) where pathway selection fairs no better than chance. With $\gamma > 0.04$, SGL-CGD consistently outperforms SGL, both in terms of pathway selection sensitivity and control of false positives (measured by FPR).

To understand why, we turn to SNP selection performance (bottom row of Figure 6). At small effect sizes ($\gamma \leq 0.04$), in the small minority of simulations where the correct pathways are identified, SGL-BCGD tends to demonstrate greater power than SGL-CGD (Figure 6 bottom left). However, this is at the expense of lower specificity (Figure 6 bottom right). These difference are due to the slightly larger number of SNPs selected by SGL-BCGD (see Table 1), which in turn is due to the 'screening out' of previously selected SNPs from the adjacent causal pathway during BCGD, as described previously. This results in the selection of a larger number of SNPs when any two overlapping pathways are selected by the model. In the case where two causal pathways are selected, SNP selection power is then likely to be higher, although at the expense of a greater number of false positives.

When pathway effects are just on the margin of detectability ($\gamma = 0.06$), SGL-CGD is more often able to select both causal pathways, although this doesn't translate into increased SNP selection power. This is most likely because at this effect size neither model can detect SNPs with low MAF, so that SGL-CGD is detecting the same (overlapping) SNPs in both causal pathways. Note that once again SGL-BCGD typically has a higher FPR than SGL-CGD, since more SNPs are selected from non-causal pathways.

As the effect size increases, the number of simulations in which SGL-CGD outperforms SGL-BCGD for SNP selection power grows, paralleling the former method's enhanced pathway selection power. This is again a demonstration of the screening effect with SGL-BCGD described previously. This means that SGL-CGD is more often able to select both causal pathways, and to select additional causal SNPs that are missed by SGL. These additional SNPs are likely to be those with lower MAF, for example, that are harder to detect with SGL, once the effect of overlapping SNPs are screened out during estimation using BCGD. Interestingly, as before SGL-CGD continues to exhibit lower false positive rates than SGL. This suggests that, with the simulated data considered here, the independence assumption offers better control of false positives by enabling the selection of causal SNPs in each and every pathway to which they are mapped. In contrast, where causal SNPs are successively screened out during the estimation using BCGD, too many SNPs with spurious effects are selected.

The relative advantage of SGL-CGD over SGL-BCGD on all performance measures starts to decrease around $\gamma = 0.1$, as SGL-BCGD becomes better able to detect all causal pathways and SNPs, irrespective of the screening effect.

## Pathway and SNP selection bias

One issue that must be addressed is the problem of selection bias, by which we mean the tendency of SGL to favour the selection of particular pathways or SNPs under the null, where no SNPs influence the phenotype. Possible biasing factors include variations in pathway size, that is the number of SNPs mapping to a pathway, or varying patterns of SNP-SNP correlations and gene sizes. Common strategies for bias reduction include the use of dimensionality reduction techniques and permutation methods [87, 35, 89, 16].

In earlier work we described an adaptive weight-tuning strategy, designed to reduce selection bias in a group lasso-based pathway selection method [66]. This works by tuning the pathway weight vector, $\mathbf{w} = (w_1, w_2, \ldots, w_L)$, so as to ensure that pathways are selected with equal probability under the null. This strategy can be readily extended to the case of dual-level sparsity with the SGL.

Our procedure rests on the observation that for pathway selection to be unbiased, each pathway must have an equal chance of being selected. For a given $\alpha$, and with $\lambda$ tuned to ensure that a single pathway is selected, pathway selection probabilities are then described by a uniform distribution, $\Pi_l = 1/L$, for $l = 1, \ldots, L$. We proceed by calculating an empirical pathway selection frequency distribution, $\Pi^*(\mathbf{w})$, by determining which pathway will first be selected by the model as $\lambda$ is reduced from its maximal value, $\lambda_{max}$, over multiple permutations of the response, $\mathbf{y}$. This process is described in detail in Supplementary Information D.

Our iterative weight tuning procedure then works by applying successive adjustments to the pathway weight vector, $\boldsymbol{w}$, so as to reduce the difference, $d_l = \Pi_l^*(\boldsymbol{w}) - \Pi_l$, between the unbiased and empirical (biased) distributions for each pathway. At iteration $\tau$, we compute the empirical pathway selection probability distribution $\Pi^*(\mathbf{w}^{(\tau)})$, determine $d_l$ for each pathway, and then apply the following weight adjustment

$$w_l^{(\tau+1)} = w_l^{(\tau)} \left[1 - \text{sign}(d_l)(\eta - 1)L^2 d_l^2\right] \qquad 0 < \eta < 1, \quad l = 1, \ldots, L.$$

The parameter $\eta$ controls the maximum amount by which each $w_l$ can be reduced in a single iteration, in the case that pathway $l$ is selected with zero frequency. The square in the weight adjustment factor ensures that large values of $|d_l|$ result in relatively large adjustments to $w_l$. Iterations continue until convergence, where $\sum_{l=1}^{L} |d_l| < \epsilon$.

Note that when multiple pathways are selected by the model, the expected pathway selection frequency distribution under the null will not be uniform. This is because pathways overlap, so that selection frequencies will reflect the complex distribution of overlapping genes, as indeed will unbiased empirical selection frequencies. We have shown previously in extensive simulations that this adaptive weight-tuning procedure gives rise to substantial gains in sensitivity and specificity with regard to pathway selection [66].

## Pathway, SNP and gene ranking

With most variable selection methods, a choice for the regularisation parameter, $\lambda$, must be made, since this determines the number of variables selected by the model. Common strategies include the use of cross validation to choose a $\lambda$ value that minimises the prediction error between training and test datasets [31]. One drawback of this approach is that it focuses on optimising the *size* of the set, $\hat{\mathcal{C}}$, of selected pathways (more generally, selected variables) that minimises the cross validated prediction error. Since the variables in $\hat{\mathcal{C}}$ will vary across each fold of the cross validation, this procedure is not in general a good means of establishing the importance of a unique set of variables, and can give rise to the selection of too many variables [85, 54]. For the lasso, alternative approaches, based on data subsampling or bootstrapping have been shown to improve model consistency, in the sense that the correct model is selected with a high probability [4, 54, 13]. These methods work by recording selected variables across multiple subsamples of the data, and forming the final set of selected variables either as the intersection of variables selected at each model fit, or by assessing variable selection frequencies. Examples of the use of such approaches can be found in a number of recent gene mapping studies involving model selection using either the lasso or elastic net [17, 21, 56, 85]. Motivated by these ideas, we adopt a resampling strategy in which we calculate pathway, gene and SNP selection frequencies by repeatedly fitting the model over $B$ subsamples of the data, at fixed values for $\alpha$ and $\lambda$. Each random subsample of size $N/2$ is drawn without replacement. Our motivation here is to exploit knowledge of finite sample variability obtained by subsampling, to achieve better estimates of a variable's importance. With this approach, which in some respects resembles the 'pointwise stability selection' strategy of Meinshausen and Bühlmann [54], selection frequencies provide a direct measure of confidence in the selected pathways in a finite sample. This resampling strategy also allows us to rank pathways, genes and SNPs in order of their strength of association with the phenotype, so that we expect the true set of causal variables to achieve a high ranking, whereas non-causal variables will be ranked low.

For pathway ranking, we denote the set of selected pathways at subsample $b$ by

$$\hat{\mathcal{C}}^{(b)} = \{l : \hat{\boldsymbol{\beta}}_l^{(b)} \neq \mathbf{0}\} \quad b = 1, \ldots, B,$$

where $\hat{\boldsymbol{\beta}}_l^{(b)}$ is the estimated SNP coefficient vector for pathway $l$ at subsample $b$. The selection probability for pathway $l$ measured across all $B$ subsamples is then

$$\pi_l^{path} = \frac{1}{B} \sum_{b=1}^{B} I_l^{(b)} \quad l = 1, \ldots, L$$

where the indicator function, $I_l^{(b)} = 1$ if $l \in \hat{\mathcal{C}}^{(b)}$, and 0 otherwise. Pathways are ranked in order of their selection probabilities, $\pi_{l_1}^{path} \geq, \ldots, \geq \pi_{l_L}^{path}$.

For SNP and gene ranking, we denote the set of SNPs selected at subsample $b$ (in the unexpanded variable space) by $\hat{\mathcal{S}}^{(b)}$, and further denote the set of selected genes to which the SNPs in $\hat{\mathcal{S}}^{(b)}$ are mapped by $\hat{\phi}^{(b)} \subset \Phi$, where $\Phi = \{1, \ldots, G\}$ is the set of gene indices corresponding to all $G$ mapped genes. Using the same strategy as for pathway ranking, we obtain an expression for the selection probability of SNP $j$ across $B$ subsamples as

$$\pi_j^{SNP} = \frac{1}{B} \sum_{b=1}^{B} J_j^{(b)}$$

where the indicator function, $J_j^{(b)} = 1$ if $j \in \hat{\mathcal{S}}^{(b)}$, and 0 otherwise. A similar expression for the selection probability for gene $g$ is

$$\pi_g^{gene} = \frac{1}{B} \sum_{b=1}^{B} K_g^{(b)}$$

where the indicator function, $K_g^{(b)} = 1$ if $g \in \hat{\phi}^{(b)}$, and 0 otherwise. SNPs and genes are then ranked in order of their respective selection frequencies.

# Results

## Subjects, genotypes and phenotypes

The analysis is carried out using data from two separate cohorts of Asian adults. These datasets have previously been used to search for novel variants associated with type 2 diabetes mellitus (T2D) in Asian populations. The first (discovery) cohort is from the Singapore Prospective Study Program, hereafter referred to as 'SP2', and the second (replication) dataset is from the Singapore Malay Eye Study or 'SiMES'. Detailed information on both datasets can be found in Sim et al. [68], but we briefly outline some salient features here.

Both datasets comprise whole genome data for T2D cases and controls, genotyped on the Illumina HumanHap 610 Quad array. For the present study we use controls only, since variation in lipid levels between cases and controls can be greater than the variation within controls alone. The use of both cases and controls in our analysis might then lead to a confounded analysis, where any associations could be linked to T2D status or some other spurious factor.

The SP2 dataset consists entirely of ethnic Chinese, and shows no evidence of population stratification. The SiMES dataset comprises ethnic Malays, and shows some evidence of cryptic relatedness between samples. For this reason, the first two principal components of a PCA for population structure are used as covariates in our analysis of this dataset. Again full details of the stratification analysis can be found in Sim et al. [68] and associated supplementary information.

A summary of information pertaining to genotypes for each dataset, both before and after imputation and pathway mapping, is given in Table 2, along with a list of phenotypes and covariates.

Table 2: Genotype and phenotype information corresponding to the SP2 and SiMES datasets used in the study.

| | SP2 | Simes |
|---|---|---|
| Sample size | $N = 1,040$ | $N = 1,099$ |
| | | |
| **Genotypes** | | |
| *Before imputation* | | |
| SNPs available for analysis[1] | $542,297$ | $557,824$ |
| SNPs with missing genotypes[2] | $152,372$ | $282,549$ |
| *Post imputation* | | |
| SNPs available for analysis[3] | $492,639$ | $515,503$ |
| | | |
| **Phenotypes/covariates** | | |
| quantitative trait (phenotype)[4] | HDLC | HDLC |
| covariates | gender, age, age$^2$, BMI[5] | gender, age, age$^2$, BMI, PC1, PC2[6] |

[1] after first round of quality control [68] and removal of monomorphic SNPs

[2] maximum 5% missing rate per SNP

[3] after imputation and removal of SNPs with MAF$<$ 0.01

[4] mg/dL

[5] body mass index (kg/m$^2$)

[6] principal components relating to cryptic relatedness

### Genotype imputation

After the initial round of quality control, genotypes for both datasets have a maximum SNP missingness of 5%. Since our method cannot handle missing values, we perform ‘missing holes’ SNP imputation, so that all missing SNP calls are estimated against a reference panel of known haplotypes.

SNP imputation proceeds in two stages. First, imputation requires accurate estimation of haplotypes from diploid genotypes (phasing). This is performed using SHAPEIT v1 (http://www.shapeit.fr). This uses a hidden Markov model to infer haplotypes from sample genotypes using a map of known recombination rates across the genome [18]. The recombination map must correspond to genotype coordinates in the dataset to be imputed, so we use recombination data from HapMap phase II, corresponding to genome build NCBI b36 (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2008-03_rel22_B36/).

Following the primary phasing stage, SNP imputation is performed using IMPUTE v2.2.2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html). IMPUTE uses a reference panel of known haplotypes to infer unobserved genotypes, given a set of observed sample haplotypes [37]. The latest version (IMPUTE 2) uses an updated, efficient algorithm, so that a custom reference panel can be used for each study haplotype, and for each region of the genome, enabling the full range of reference information provided by HapMap3 [77] to be used. Following IMPUTE 2 guidelines, we use HapMap3 reference data corresponding to NCBI b36 (http://mathgen.stats.ox.ac.uk/impute/data_download_hapmap3_r2.html) which includes haplotype data for 1,011 individuals from Africa, Asia, Europe and the Americas. SNPs are imputed in 5MB chunks, using an effective population size ($Ne$) of 15,000, and a buffer of 250kb to avoid edge effects, again as recommended for IMPUTE 2.

The phasing and imputation process is complex and computationally intensive. For this reason we implement a pipeline in Python, with phasing and imputation for each chromosome conducted in parallel across multiple nodes in a computing cluster. This enables full genome imputation that would otherwise take days, to be completed in a matter of hours.

### Pathway mapping

Pathways GWAS methods rely on prior information mapping SNPs to functional networks or pathways. Since pathways are typically defined as groups of interacting genes, SNP to pathway mapping is a two-part process, requiring the mapping of genes to pathways, and of SNPs to genes. A consistent strategy for this mapping process has however yet to be established, a situation compounded by a lack of agreement on what constitutes a pathway in the first place [10].

The number and size of databases devoted to classifying genes into pathways is growing rapidly, as is the range and diversity of gene interactions considered (see for example http://www.pathguide.org/). Databases such as those provided by KEGG (http://www.genome.jp/kegg/pathway.html), Reactome (http://www.reactome.org/) and Biocarta (http://www.biocarta.com/) classify pathways across a number of functional domains, for example apoptosis, cell adhesion or lipid metabolism; or crystallise current knowledge on specific disease-related molecular reaction networks. Strategies for pathways database assembly range from a fully-automated text-mining approach, to that of careful curation by experts. Inevitably therefore, there is considerable variation between databases, in terms of both gene coverage and consistency [71], so that the choice of database(s) will itself influence results in pathways GWAS.

The mapping of SNPs to genes adds a further layer of complexity, since although many SNPs may occur within gene boundaries, on a typical GWAS array the vast majority of SNPs will reside in inter-genic regions. In an attempt to include variants potentially residing in functionally significant regions lying outside gene boundaries, SNPs may be mapped to nearby genes using various distance thresholds. Various values for SNP to gene mapping distances, measured in thousands of nucleotide base pairs (kb), have been suggested in the literature, ranging from mapping SNPs to genes only if they fall within a specific gene, to the attempt to encompass upstream promoters and enhancers by extending the range to 10, 20 or even 500kb and beyond [87, 20, 10]. This process is illustrated schematically in Figure 7. Notable features of the SNP to pathway mapping process include the fact that genes (and therefore SNPs) may map to more than one pathway, and also that many SNPs and genes do not currently map to any known pathway [25].

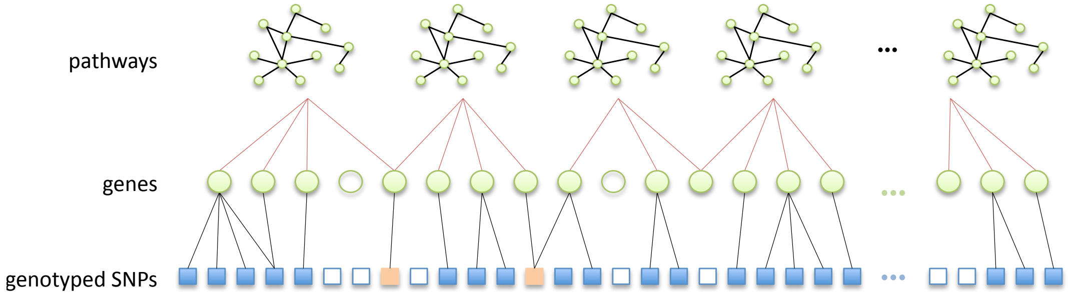


Figure 7: Schematic illustration of the SNP to pathway mapping process. (i) Genes (green circles) are mapped to pathways using information on gene-gene interactions (top row), obtained from a gene pathways database. Many genes do not map to any known pathway (unfilled circles). Also, some genes may map to more than one pathway. (ii) Genes that map to a pathway are in turn mapped to genotyped SNPs within a specified distance. Many SNPs cannot be mapped to a pathway since they do not map to a mapped gene (unfilled squares). Note SNPs may map to more than one gene. Some SNPs (orange squares) may map to more than one pathway, either because they map to multiple genes belonging to different pathways, or because they map to a single gene that belongs to multiple pathways.

Following imputation, SNPs for both datasets in the present study are mapped to KEGG canonical pathways from the MSigDB database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). We exclude the largest KEGG pathway (by number of mapped SNPs), 'Pathways in Cancer', since it is highly redundant in that it contains multiple other pathways as subsets. Details of the pathway mapping process are given in Figures 8 and 9.

Note that there is a difference in the number of SNPs available for the pathway mapping between the two datasets, and this results in a small discrepancy in the total number of mapped genes (SP2: 4,734 mapped genes; SiMES: 4,751). However, both datasets map to all 185 KEGG pathways, and a large majority of mapped genes and SNPs overlap both datasets. Detailed information on the pathway mapping process for the two datasets is presented in Table 3.

We perform pathways-driven SNP selection on both datasets, using the procedures described in Methods. We present results for each dataset separately below.

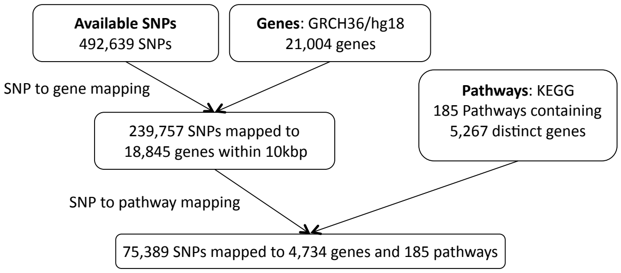


Figure 8: *SP2 dataset.* SNP to pathway mapping.

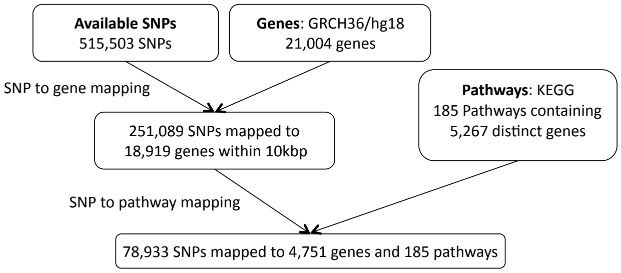


Figure 9: *SiMES dataset.* SNP to pathway mapping.

Table 3: Comparison of SNP and gene to pathway mappings for the SP2 and SiMES datasets.

| | SP2 | SiMES |
|---|---|---|
| Total SNPs mapping to pathways | 75,389 | 78,933 |
| Total SNPs mapping to pathways in both datasets (intersection) | 74,864 | |
| Total mapped genes | 4,734 | 4,751 |
| Total genes mapping to pathways in both datasets (intersection) | 4,726 | |
| Total mapped pathways | 185 | 185 |
| Minimum number of genes mapping to single pathway | 11 | 11 |
| Maximum number of genes mapping to single pathway | 63 | 63 |
| Minimum number of SNPs mapping to single pathway | 66 | 67 |
| Maximum number of SNPs mapping to single pathway | 5,759 | 6,058 |
| Minimum number of pathways mapping to a single SNP | 1 | 1 |
| Maximum number of pathways mapping to a single SNP | 45 | 45 |

## SP2 Analysis

For the SP2 dataset we consider two separate scenarios for the regularisation parameters $\lambda$ and $\alpha$. For the two scenarios we set the sparsity parameter, $\lambda = 0.95\lambda_{max}$, but consider two values for $\alpha$, namely $\alpha = 0.95, 0.85$. We test each scenario over 1000 $N/2$ subsamples. We also compare the resulting pathway and SNP selection frequency distributions with null distributions, again over 1000 $N/2$ subsamples, but with phenotype labels permuted, so that no SNPs can influence the phenotype.

The parameter $\alpha$ controls how the regularisation penalty is distributed between the $\ell_2$ (pathway) and $\ell_1$ (SNP) norms of the coefficient vector. Each scenario therefore entails different numbers of selected pathways and SNPs, and this information is presented in Table 4.

Table 4: Separate combinations of regularisation parameters, $\lambda$ and $\alpha$ used for analysis of the SP2 dataset. For each $\lambda$, $\alpha$ combination, the mean (±SD) number of selected pathways and SNPs across all 1000 subsamples is reported.

| | $\lambda = 0.95\lambda_{max}$ | |
|---|---|---|
| | $\alpha = 0.85$ | $\alpha = 0.95$ |
| *empirical* | | |
| selected pathways | $7.9 \pm 6.1$ | $4.8 \pm 4.1$ |
| selected SNPs | $1551 \pm 1294$ | $160 \pm 185$ |
| *null* | | |
| selected pathways | $9.1 \pm 7.2$ | $5.0 \pm 4.55$ |
| selected SNPs | $1656 \pm 1401$ | $155 \pm 194$ |

Comparisons of empirical and null pathway selection frequency distributions for each scenario are presented in Figure 10. The same comparisons for SNP selection frequencies are presented in Figure 11. In these plots, null distributions (coloured blue) are ordered along the $x$-axis according to their corresponding ranked empirical selection frequencies (marked in red). This is to help visualise any potential biases that may be influencing variable selection (see below).

To interpret these results, we begin by noting from Table 4 that many more SNPs are selected with $\alpha = 0.85$, resulting in higher SNP selection frequencies, compared to those obtained with $\alpha = 0.95$ (see Figure 11). This is as expected, since a lower value for $\alpha$ implies a reduced $\ell_1$ penalty on the SNP coefficient vector, resulting in more SNPs being selected. Perhaps surprisingly, given that the $\ell_2$ group penalty $(1 - \alpha)\lambda$ is increased, the number of selected pathways is also greater. This must reflect the reduced $\ell_1$ penalty, which allows a greater number of SNPs to contribute to a putative selected pathway's coefficient vector. This in turn increases the number of pathways that pass the threshold for selection.

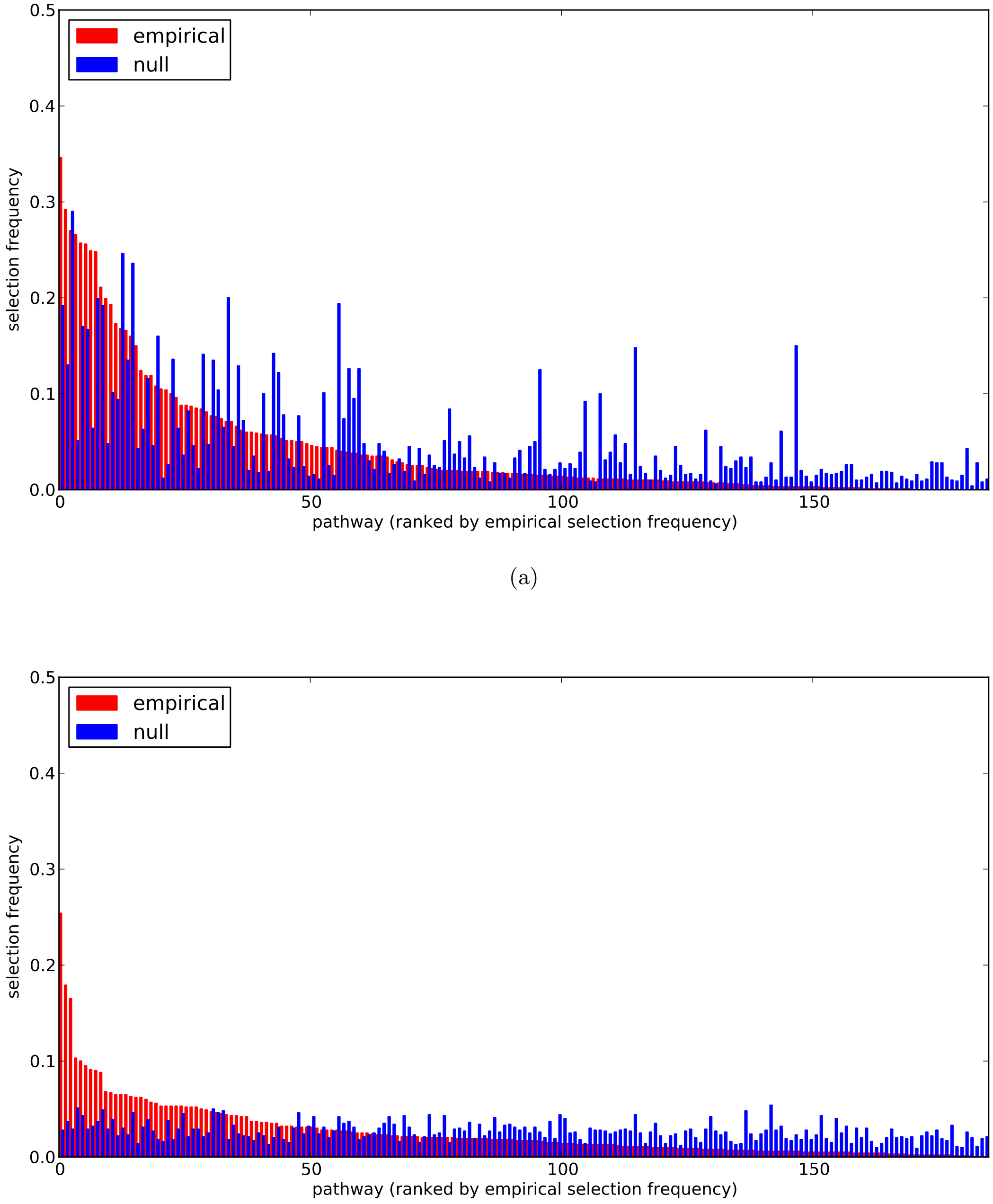
This raises the question of what might be considered to be an optimal choice for the regularisation-distributional parameter $\alpha$, since different assumptions about the number of SNPs potentially influencing the phenotype may affect the resulting pathway and SNP rankings. To answer this, we turn our attention to the pathway and SNP selection frequency distributions for each $\alpha$ value in Figures 10 and 11. At the lower value of $\alpha = 0.85$ (top plots in Figures 10 and 11), empirical pathway and SNP selection frequency distributions appear to be biased, in the sense that there is a suggestion that pathways and SNPs with the highest empirical selection frequencies also tend to be selected with a higher frequency under the null, where there is no association between genotype and phenotype. This relationship appears to be diminished with $\alpha = 0.95$, when fewer SNPs are selected by the model. We investigate this further by plotting empirical vs. null selection frequencies as a sequence of scatter plots in Figure 12, and we report Pearson correlation coefficients and p-values for these in Table 5.

Table 5: SP2 dataset: Pearson correlation coefficients ($r$)and p-values for the data plotted in Figure 12. $n$ denotes the number of predictors considered. For SNPs, coefficients describe correlations for all predictors selected with nonzero empirical selection frequencies only, since a large number of SNPs are not selected by the model at any subsample.

| | $\alpha = 0.85$ | | | $\alpha = 0.95$ | | |
|---|---|---|---|---|---|---|
| | $n$ | $r$ | p-value | $n$ | $r$ | p-value |
| pathways | 185 | 0.66 | $1.3 \times 10^{-24}$ | 185 | 0.26 | $2.9 \times 10^{-4}$ |
| SNPs | 62,965 | 0.37 | 0 | 30,027 | 0.11 | $1.2 \times 10^{-84}$ |

These provide further evidence of increased correlation between empirical and null selection frequency distributions at the lower $\alpha$ value for both pathways and SNPs, again suggesting increased bias in the empirical results, in the sense that certain pathways and SNPs tend to be selected with a higher frequency, irrespective of whether or not a true signal may be present. Further qualitative evidence of reduced bias with $\alpha = 0.95$ is suggested by the clearer separation of empirical and null distributions at the higher $\alpha$ value in Figures 10 and 11. For example, the maximum empirical pathway selection frequency is reduced by a factor of 0.29 (0.35 to 0.25) as $\alpha$ is increased from 0.85 to 0.95, whereas the maximum pathway selection frequency under the null is reduced by a factor of 0.81 (0.29 to 0.054). Similarly for SNPs, the maximum empirical SNP selection frequency is reduced by a factor of 0.37 (0.52 to 0.33), whereas the maximum SNP selection frequency under the null is reduced by a factor of 0.9 (0.11 to 0.011).

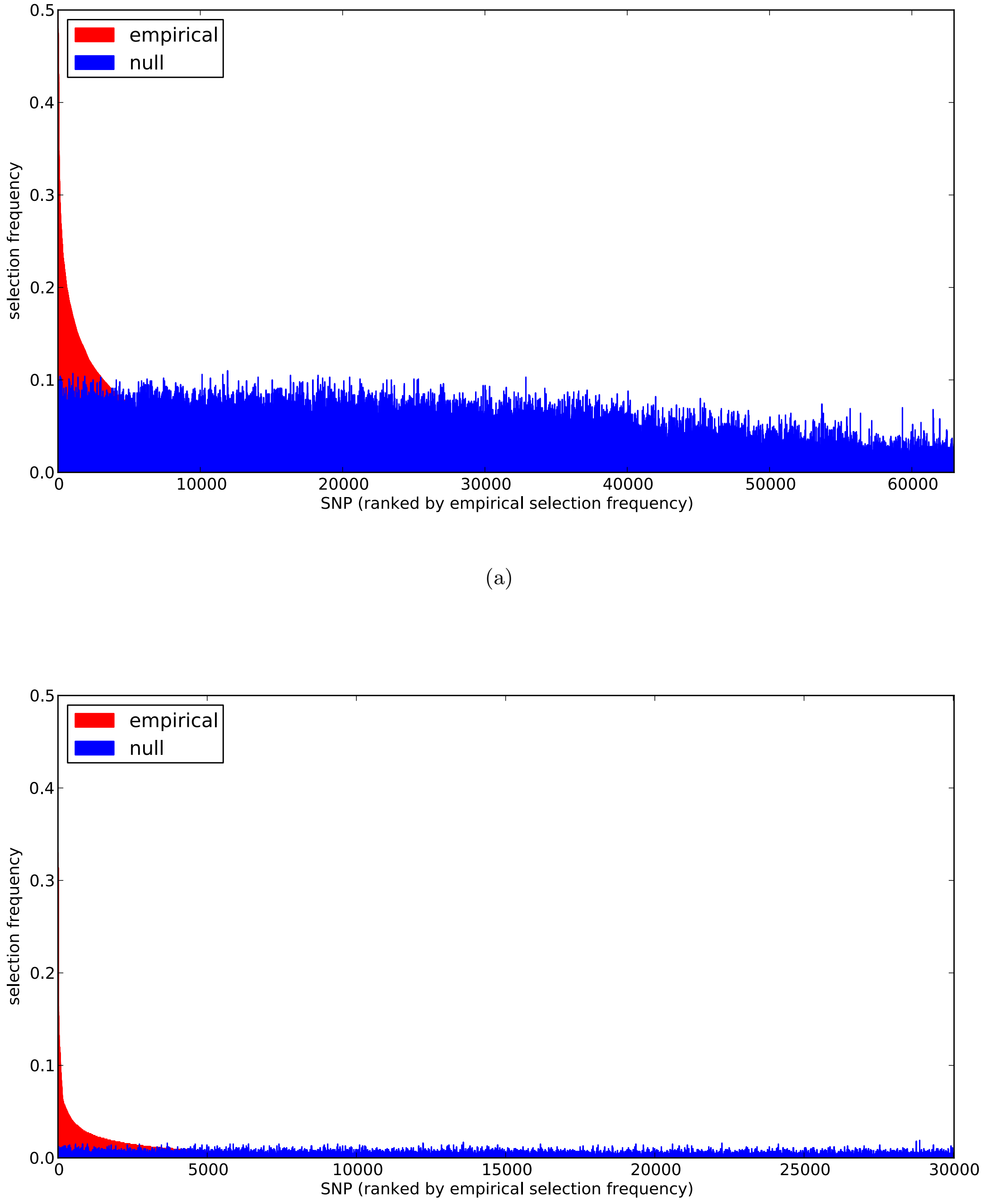
The increased bias with $\alpha = 0.85$ is most likely due to the selection of too many SNPs,



(a)

(b)

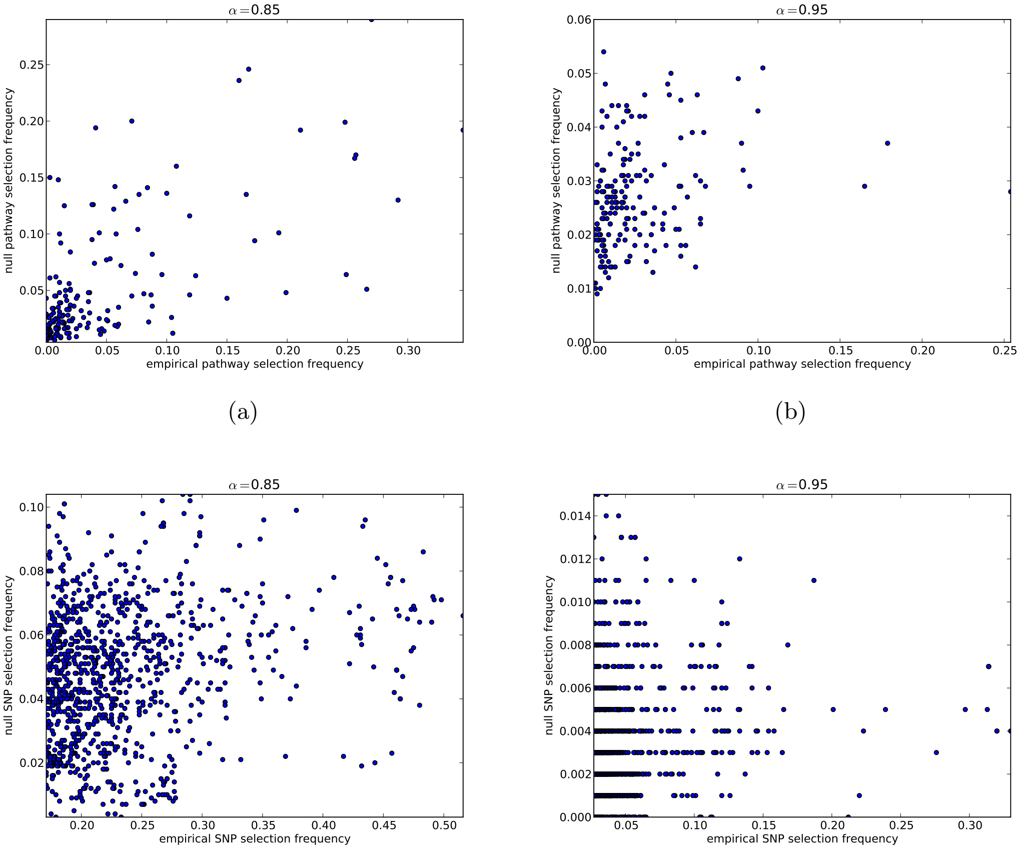
Figure 10: Empirical and null pathway selection frequency distributions for all 185 KEGG pathways with the SP2 dataset. For each scenario, pathways are ranked along the $x$-axis in order of their empirical pathway selection frequency, $\pi_{l_1}^{path} >, \ldots, > \pi_{l_L}^{path}$. (a) $\alpha = 0.85$. (b) $\alpha = 0.95$.



(a)

(b)

Figure 11: Empirical and null SNP selection frequency distributions with the SP2 dataset. For each scenario, SNPs are ranked along the $x$-axis in order of their empirical pathway selection frequency, $\pi_{j_1}^{SNP} > \pi_{j_2}^{SNP} > \ldots$. (a) $\alpha = 0.85$. (b) $\alpha = 0.95$. Note fewer SNPs are selected with nonzero empirical selection frequency with $\alpha = 0.95$, so that the $x$-axis range in (b) is reduced.



(a)

(b)

(c)

(d)

Figure 12: SP2 dataset: Scatter plots comparing empirical and null selection frequencies presented in Figures 10 and 11. *(a)* and *(b)*: Pathway selection frequencies with $\alpha = 0.85, 0.95$ respectively. *(c)* and *(d)*: SNP selection frequencies for the same $\alpha$ values. For clarity, SNP selection frequencies are plotted for the top 1000 SNPs (by empirical selection frequency) only. Corresponding correlation coefficients (for all ranked SNPs) are presented in Table 5. Note that pathway and SNP selection frequencies are much higher at the lower $\alpha$ value (left hand plots), since many more variables are selected (see Table 4.)

in the sense that many selected SNPs do not exhibit real phenotypic effects. These extra SNPs effectively add noise to the model, in the form of multiple weak, spurious signals. This in turn will add bias to the resulting selection frequency distributions, tending to favour, for example, SNPs that overlap multiple pathways, and the pathways that contain them. As $\alpha$ is increased, we would expect this biasing effect to be reduced, until a point where too few SNPs are selected, when there is then a risk that some of the true signal may be lost.

Note that the reduced but still significant correlations between empirical and null selection frequency distributions at $\alpha = 0.95$ in Table 5 are not unexpected. These may reflect the complex overlap structure between pathways, meaning that pathways (and associated SNPs) with a relatively high degree of overlap with other pathways, due for example to the presence of so called 'hub genes', are more likely to harbour true signals, as well as spurious ones [48, 11, 42].

Taking all the above into consideration, we choose to report results with $\alpha = 0.95$, where there is less evidence of bias due to the selection of too many SNPs. The top 30 pathways, ranked by selection frequency are presented in Table 6, and the top 30 ranked SNPs, together with corresponding genes to which they are mapped are presented in Table 7. Versions of these tables extending to lower ranks are provided as supplementary information.

Table 6: SP2 dataset: Top 30 pathways, ranked by pathway selection frequency, $\pi^{path}$. The final column lists genes in the pathway that are in the top 30 ranked genes selected in the study (see Table 7). Pathways falling in the consensus set, $\Psi_{25}^{path}$, obtained by comparing pathway ranking results from both SP2 and SiMES datasets (see Table 11), are marked with a $^*$.

| Rank | KEGG pathway name | $\pi^{path}$ | Size (# SNPs) | top 30 ranked genes in pathway |
|---|---|---|---|---|
| 1 | Toll Like Receptor Signaling Pathway | 0.254 | 766 | *TIRAP RAC1 IFNAR1 CD80 IL12B PIK3R1* |
| 2 | Jak Stat Signaling Pathway | 0.179 | 1447 | *PIAS2 IL5RA TPO IFNAR1 IL12B PIK3R1 IL2RA* |
| 3 | Ubiquitin Mediated Proteolysis | 0.165 | 1603 | *PIAS2 RFWD2 PARK2* |
| 4 | *Dilated Cardiomyopathy | 0.103 | 3054 | *ADCY2 TGFB3 PRKACB RYR2 ITGB8 ITGA1 CACNA2D3 LAMA2 CACNA1C* |
| 5 | Cytokine Cytokine Receptor Interaction | 0.100 | 2553 | *IL5RA IL12B TGFB3 EGFR TPO IFNAR1 IL2RA* |
| 6 | Ecm Receptor Interaction | 0.095 | 2271 | *ITGB8 ITGA1 LAMA2* |
| 7 | Arginine And Proline Metabolism | 0.091 | 432 | *NOS1* |
| 8 | Parkinson's Disease | 0.090 | 1320 | *PARK2* |
| 9 | * Hypertrophic Cardiomyopathy | 0.088 | 2819 | *TGFB3 RYR2 ITGB8 ITGA1 CACNA2D3 LAMA2 CACNA1C* |
| 10 | Small Cell Lung Cancer | 0.068 | 1808 | *PIAS2 PIK3R1 LAMA2* |
| 11 | Natural Killer Cell Mediated Cytotoxicity | 0.067 | 1781 | *KRAS RAC1 VAV3 VAV2 PRKCA IFNAR1 PRKCB PIK3R1* |
| 12 | * T Cell Receptor Signaling Pathway | 0.065 | 1541 | *KRAS VAV3 VAV2 PIK3R1* |
| 13 | Tgf Beta Signaling Pathway | 0.065 | 947 | *TGFB3* |
| 14 | Olfactory Transduction | 0.065 | 2497 | *PRKACB* |
| 15 | * Arrhythmogenic Right Ventricular Cardiomyopathy | 0.063 | 3726 | *RYR2 TCF7L1 ITGB8 ITGA1 CACNA2D3 LAMA2 CACNA1C* |
| 16 | * Ppar Signaling Pathway | 0.062 | 758 | |
| 17 | Taste Transduction | 0.062 | 941 | *PRKACB* |
| 18 | Type I Diabetes Mellitus | 0.060 | 776 | *CD80 IL12B* |
| 19 | * Ribosome | 0.057 | 261 | |
| 20 | * Terpenoid Backbone Biosynthesis | 0.056 | 147 | |
| 21 | Neuroactive Ligand Receptor Interaction | 0.053 | 5745 | *GRIN3A* |
| 22 | Regulation Of Actin Cytoskeleton | 0.053 | 3803 | *KRAS RAC1 EGFR ITGB8 VAV3 ITGA1 VAV2 PIK3R1* |
| 23 | Mismatch Repair | 0.053 | 222 | |
| 24 | Cell Adhesion Molecules Cams | 0.053 | 3977 | *ITGB8 CD80* |
| 25 | Maturity Onset Diabetes Of The Young | 0.053 | 239 | |
| 26 | Butanoate Metabolism | 0.052 | 383 | |
| 27 | Purine Metabolism | 0.052 | 3224 | *ADCY2* |
| 28 | P53 Signaling Pathway | 0.052 | 598 | *RFWD2* |
| 29 | Dorso Ventral Axis Formation | 0.050 | 581 | *KRAS EGFR* |
| 30 | Basal Cell Carcinoma | 0.049 | 589 | *TCF7L1* |

Table 7: SP2 dataset: Top 30 SNPs and genes, respectively ranked by SNP and gene selection frequency. Genes falling in the top 30 ranks of the consensus gene set, $\Psi^{gene}_{244}$, obtained by comparing gene ranking results from both SP2 and SiMES datasets (see Table 13), are marked with a $^*$.

| SNP RANKING | | | | GENE RANKING | | |
|---|---|---|---|---|---|---|
| Rank | SNP | $\pi^{SNP}$ | Mapped gene(s) | Gene | $\pi^{gene}$ | # mapped SNPs |
| 1 | rs2257167 | 0.33 | *IFNAR1* | *IFNAR1* | 0.33 | 11 |
| 2 | rs2254315 | 0.32 | *IFNAR1* | *IL12B* | 0.30 | 9 |
| 3 | rs1041868 | 0.32 | *IFNAR1* | *PIAS2* | 0.30 | 7 |
| 4 | rs7364085 | 0.31 | *IFNAR1* | *TIRAP* | 0.22 | 5 |
| 5 | rs2850021 | 0.31 | *IFNAR1* | *RAC1* | 0.21 | 10 |
| 6 | rs2253413 | 0.31 | *IFNAR1* | *LAMA2*$^*$ | 0.19 | 111 |
| 7 | rs2243590 | 0.31 | *IFNAR1* | *ADCY2*$^*$ | 0.19 | 94 |
| 8 | rs2834204 | 0.31 | *IFNAR1* | *PIK3R1* | 0.19 | 28 |
| 9 | rs3181224 | 0.30 | *IL12B* | *PARK2* | 0.19 | 460 |
| 10 | rs512868 | 0.28 | *PIAS2* | *IL2RA* | 0.19 | 55 |
| 11 | rs641366 | 0.24 | *PIAS2* | *PRKCA*$^*$ | 0.19 | 123 |
| 12 | rs2032215 | 0.22 | *PIAS2* | *ITGB8* | 0.18 | 27 |
| 13 | rs8177375 | 0.22 | *TIRAP* | *TCF7L1* | 0.18 | 55 |
| 14 | rs10893493 | 0.21 | *TIRAP* | *CD80*$^*$ | 0.18 | 21 |
| 15 | rs4890341 | 0.20 | *PIAS2* | *GRIN3A* | 0.18 | 60 |
| 16 | rs2303361 | 0.19 | *RAC1* | *PRKCB*$^*$ | 0.18 | 83 |
| 17 | rs7873495 | 0.17 | *GRIN3A PPP3R2* | *CACNA1C*$^*$ | 0.17 | 180 |
| 18 | rs1323653 | 0.17 | *IL2RA* | *TGFB3* | 0.16 | 7 |
| 19 | rs11762117 | 0.16 | *ITGB8* | *PRKACB* | 0.16 | 16 |
| 20 | rs3807955 | 0.16 | *ITGB8* | *KRAS*$^*$ | 0.16 | 21 |
| 21 | rs10462842 | 0.15 | *ADCY2* | *VAV3* | 0.16 | 97 |
| 22 | rs10215885 | 0.15 | *ITGB8* | *IL5RA* | 0.15 | 38 |
| 23 | rs3807936 | 0.15 | *ITGB8* | *ITGA1*$^*$ | 0.15 | 77 |
| 24 | rs530205 | 0.15 | *PIAS2* | *VAV2*$^*$ | 0.15 | 85 |
| 25 | rs3823974 | 0.15 | *ITGB8* | *EGFR*$^*$ | 0.14 | 61 |
| 26 | rs2074425 | 0.15 | *ITGB8* | *TPO* | 0.14 | 50 |
| 27 | rs6725799 | 0.15 | *TCF7L1* | *CACNA2D3*$^*$ | 0.14 | 283 |
| 28 | rs2301727 | 0.15 | *ITGB8* | *RYR2*$^*$ | 0.14 | 214 |
| 29 | rs10486391 | 0.14 | *ITGB8* | *NOS1* | 0.14 | 49 |
| 30 | rs3779505 | 0.14 | *ITGB8* | *RFWD2* | 0.13 | 31 |

## SiMES Analysis

For the replication SiMES dataset, we repeat the above analysis design, but consider only the 'low bias' scenario where $\lambda = 0.95\lambda_{max}$ and $\alpha = 0.95$. Once again we test each scenario over 1000 $N/2$ subsamples, and compare the resulting pathway and SNP selection frequency distributions with null distributions generated over 1000 $N/2$ subsamples with phenotype labels permuted. Pathway and SNP selection frequency distributions are presented in Figure 14. An investigation of pathway and SNP selection bias is presented in the form of scatter plots illustrating potential correlation between empirical and null selection frequencies in Figure 13, with corresponding Pearson correlation coefficients and p-values presented in Table 8. The top 30 ranked pathways, and SNPs and genes are presented in Tables 9 and 10 respectively.

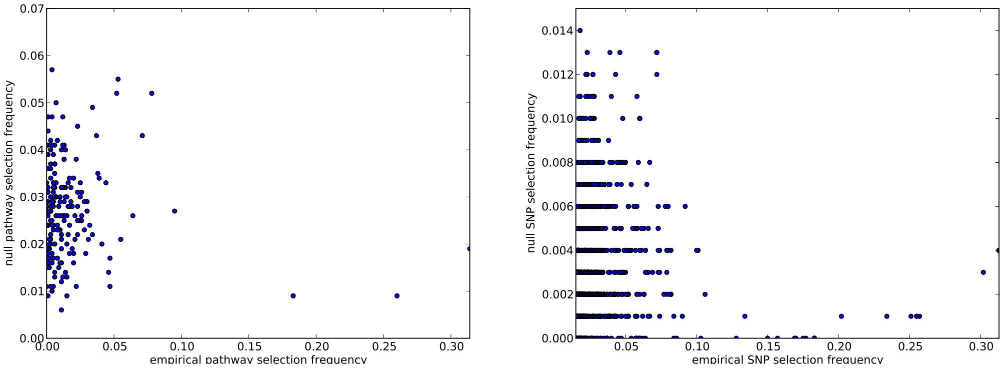


Figure 13: SiMES dataset: Scatter plots comparing empirical and null pathway (*left*) and SNP (*right*) selection frequencies presented in Figure 14. For clarity, SNP selection frequencies are plotted for the top 1000 SNPs (by empirical selection frequency) only.

Table 8: SiMES dataset: Pearson correlation coefficients ($r$)and p-values for the data plotted in Figure 13. $n$ denotes the number of predictors considered. For SNPs, coefficients describe correlations for all predictors selected with nonzero empirical selection frequencies only, since a large number of SNPs are not selected by the model at any subsample.

| | $n$ | $r$ | p-value |
|---|---|---|---|
| pathways | 185 | $-0.094$ | 0.20 |
| SNPs | $20,006$ | 0.058 | $2.63 \times 10^{-16}$ |

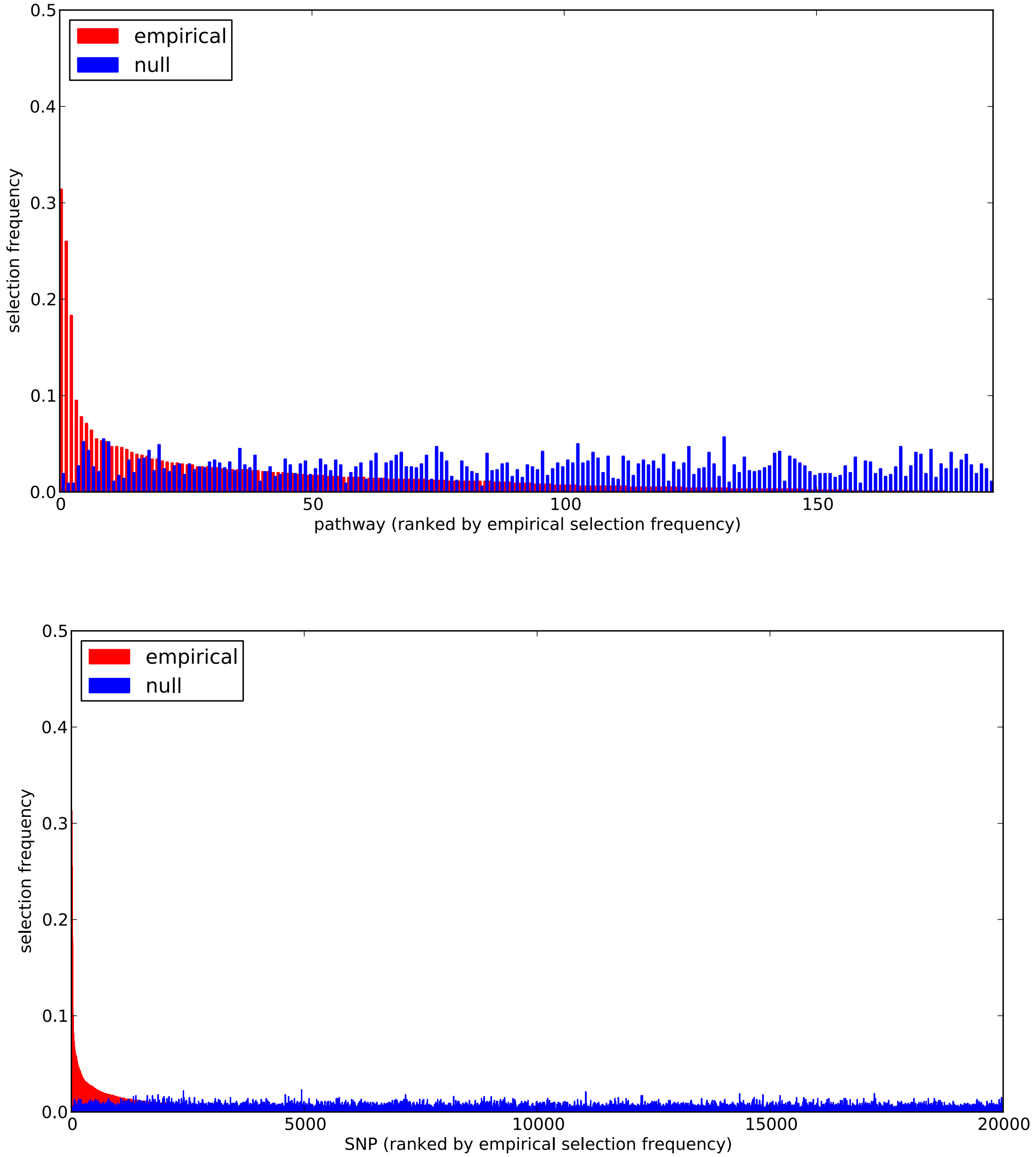


Figure 14: Empirical and null pathway (*top*) and SNP (*bottom*) selection frequency distributions for the SiMES dataset. $\alpha = 0.95$. For both empirical (red) and null (blue) distributions, variables (pathways and SNPs) are ranked along the $x$-axis in order of their empirical selection frequencies.

Table 9: SiMES dataset: Top 30 pathways, ranked by pathway selection frequency, $\pi^{path}$. The final column lists genes in the pathway that are in the top 30 ranked genes selected in the study (i.e. genes in the top 30 gene ranking in Table 10). Pathways falling in the consensus set, $\Psi_{25}^{path}$, obtained by comparing pathway ranking results from both SP2 and SiMES datasets (see Table 11), are marked with a $^*$.

| Rank | KEGG pathway name | $\pi^{path}$ | Size<br>(# SNPs) | top 30 ranked genes in pathway |
|---|---|---|---|---|
| 1 | Oxidative Phosphorylation | 0.314 | 871 | *PPA2 NDUFA4 SDHB SDHC ATP6V0A4* |
| 2 | * Terpenoid Backbone Biosynthesis | 0.260 | 158 | *PDSS2* |
| 3 | Regulation Of Autophagy | 0.183 | 215 | *GABARAPL1* |
| 4 | Glycerolipid Metabolism | 0.095 | 1074 | *ALDH7A1 DGKB DGKH ALDH2 LIPC* |
| 5 | *Dilated Cardiomyopathy | 0.078 | 3177 | *ADCY2 RYR2 ITGA11 ITGB1 SLC8A1 ITGA1 CACNA2D3 LAMA2 CACNA1C CACNA1D* |
| 6 | * Hypertrophic Cardiomyopathy | 0.071 | 2932 | *PRKAG2 RYR2 ITGA11 ITGB1 SLC8A1 ITGA1 CACNA2D3 LAMA2 CACNA1C CACNA1D* |
| 7 | * Ribosome | 0.064 | 270 | |
| 8 | Glutathione Metabolism | 0.055 | 389 | *MGST3* |
| 9 | * Arrhythmogenic Right Ventricular Cardiomyopathy | 0.053 | 3899 | *RYR2 ITGA11 ITGB1 SLC8A1 ITGA1 CACNA2D3 LAMA2 CACNA1C CACNA1D* |
| 10 | * T Cell Receptor Signaling Pathway | 0.052 | 1624 | *PAK7 FYN* |
| 11 | Cardiac Muscle Contraction | 0.047 | 1952 | *RYR2 SLC8A1 CACNA2D3 CACNA1C CACNA1D* |
| 12 | Biosynthesis Of Unsaturated Fatty Acids | 0.047 | 282 | |
| 13 | Lysosome | 0.046 | 1322 | *ATP6V0A4* |
| 14 | Apoptosis | 0.044 | 954 | *BCL2* |
| 15 | Pathogenic Escherichia Coli Infection | 0.041 | 538 | *ITGB1 FYN* |
| 16 | Metabolism Of Xenobiotics By Cytochrome P450 | 0.039 | 880 | *MGST3* |
| 17 | Drug Metabolism Cytochrome P450 | 0.038 | 910 | *MGST3* |
| 18 | Autoimmune Thyroid Disease | 0.037 | 686 | |
| 19 | Focal Adhesion | 0.034 | 4787 | *ITGA11 LAMA2 BCL2 FYN EGFR ITGB1 ITGA1 PAK7 PRKCB IGF1R* |
| 20 | Leishmania Infection | 0.034 | 718 | *PRKCB ITGB1* |
| 21 | * Ppar Signaling Pathway | 0.032 | 800 | |
| 22 | Rna Polymerase | 0.031 | 193 | |
| 23 | Lysine Degradation | 0.030 | 423 | *ALDH7A1 ALDH2* |
| 24 | Endocytosis | 0.030 | 3436 | *EGFR IGF1R* |
| 25 | Glycosaminoglycan Biosynthesis Chondroitin Sulfate | 0.029 | 727 | |
| 26 | Melanoma | 0.028 | 1189 | *EGFR IGF1R* |
| 27 | Nucleotide Excision Repair | 0.028 | 330 | |
| 28 | Prostate Cancer | 0.026 | 1419 | *EGFR IGF1R BCL2* |
| 29 | Renal Cell Carcinoma | 0.026 | 1004 | *PAK7* |
| 30 | Glycine Serine And Threonine Metabolism | 0.026 | 268 | |

Table 10: SiMES dataset: Top 30 SNPs and genes, respectively ranked by SNP and gene selection frequency. Genes falling in the top 30 ranks of the consensus gene set, $\Psi_{244}^{gene}$, obtained by comparing gene ranking results from both SP2 and SiMES datasets (see Table 13), are marked with a $^*$.

| SNP RANKING | | | | GENE RANKING | | |
|---|---|---|---|---|---|---|
| Rank | SNP | $\pi^{SNP}$ | Mapped gene(s) | Gene | $\pi^{gene}$ | # mapped SNPs |
| 1 | rs2636698 | 0.31 | *PPA2* | *PPA2* | 0.31 | 16 |
| 2 | rs2726503 | 0.31 | *PPA2* | *PDSS2* | 0.26 | 59 |
| 3 | rs2713829 | 0.31 | *PPA2* | *GABARAPL1* | 0.18 | 11 |
| 4 | rs2636726 | 0.31 | *PPA2* | *ATP6V0A4* | 0.15 | 35 |
| 5 | rs2713834 | 0.31 | *PPA2* | *ITGB1* | 0.13 | 14 |
| 6 | rs2726471 | 0.31 | *PPA2* | *CACNA1C*$^*$ | 0.11 | 186 |
| 7 | rs2636713 | 0.31 | *PPA2* | *PRKCB*$^*$ | 0.11 | 84 |
| 8 | rs2726516 | 0.31 | *PPA2* | *FYN* | 0.11 | 46 |
| 9 | rs2636739 | 0.31 | *PPA2* | *BCL2*$^*$ | 0.10 | 61 |
| 10 | rs2713861 | 0.31 | *PPA2* | *PAK7*$^*$ | 0.10 | 127 |
| 11 | rs2636751 | 0.31 | *PPA2* | *DGKB* | 0.10 | 233 |
| 12 | rs2298733 | 0.30 | *PPA2* | *LAMA2*$^*$ | 0.10 | 118 |
| 13 | rs6568474 | 0.26 | *PDSS2* | *NDUFA4* | 0.10 | 7 |
| 14 | rs9386622 | 0.26 | *PDSS2* | *DGKH* | 0.10 | 70 |
| 15 | rs6924886 | 0.26 | *PDSS2* | *ADCY2*$^*$ | 0.09 | 104 |
| 16 | rs759440 | 0.25 | *PDSS2* | *LIPC* | 0.09 | 69 |
| 17 | rs11759792 | 0.23 | *PDSS2* | *SLC8A1*$^*$ | 0.09 | 240 |
| 18 | rs10457161 | 0.20 | *PDSS2* | *EGFR*$^*$ | 0.09 | 74 |
| 19 | rs12821011 | 0.18 | *GABARAPL1* | *PRKAG2* | 0.09 | 118 |
| 20 | rs11053685 | 0.18 | *GABARAPL1* | *CACNA1D* | 0.09 | 83 |
| 21 | rs4764324 | 0.18 | *GABARAPL1* | *ITGA11*$^*$ | 0.09 | 63 |
| 22 | rs4764327 | 0.18 | *GABARAPL1* | *IGF1R*$^*$ | 0.09 | 100 |
| 23 | rs9373924 | 0.18 | *PDSS2* | *SDHC* | 0.09 | 9 |
| 24 | rs10845074 | 0.18 | *GABARAPL1* | *CACNA2D3*$^*$ | 0.08 | 294 |
| 25 | rs9320215 | 0.17 | *PDSS2* | *RYR2*$^*$ | 0.08 | 221 |
| 26 | rs10845073 | 0.17 | *GABARAPL1* | *ITGA1*$^*$ | 0.08 | 77 |
| 27 | rs4946826 | 0.16 | *PDSS2* | *ALDH7A1* | 0.08 | 23 |
| 28 | rs6938393 | 0.15 | *PDSS2* | *MGST3*$^*$ | 0.08 | 40 |
| 29 | rs13202332 | 0.13 | *PDSS2* | *ALDH2* | 0.08 | 12 |
| 30 | rs9480754 | 0.13 | *PDSS2* | *SDHB* | 0.08 | 13 |

### Comparison of ranked pathway lists

We now consider the problem of comparing rankings (pathways, genes and SNPs) obtained for each dataset. To do this we require some measure of distance between each pair of lists. Ideally this measure should place more emphasis on differences between highly-ranked variables, since we expect the association signal, and hence agreement between the ranked lists, to be strongest there. By the same reasoning, we expect there to be little or no agreement between variables at lower rankings, where selection frequencies are low. Indeed a consideration of empirical and null selection frequency distributions (Figures 10 and 14, top) suggests that only the very top ranked variables are likely to reflect any true signal, so that we would additionally like our distance metric to be able to accommodate consideration of the top-$k$ variables only, with $k < p$, where $p$ is the total number of variables ranked in either dataset. One complication with top-$k$ lists is that they are *partial*, in the sense that unlike complete ($k = p$) lists, a variable may occur in one list, but not the other.

In order to consider this problem, we introduce the following notation. We denote the complete set of ranked predictors by $\mathcal{L} = \{1, \ldots, p\}$, and begin by assuming that all variables are ranked in both datasets. We denote the rank of each variable in list 1 by $\tau(i), i = 1, \ldots, p$, so that $\tau(5) = 1$ if variable 5 is ranked first and so on. The corresponding ranks for list 2 are denoted by $\sigma(i), i = 1, \ldots, p$. A suitable metric describing the distance between two top-$k$ rankings is the *Canberra distance* [44],

$$Ca(k, \tau, \sigma) = \sum_{i=1}^{p} \frac{|\min\{\tau(i), k+1\} - \min\{\sigma(i), k+1\}|}{\min\{\tau(i), k+1\} + \min\{\sigma(i), k+1\}}. \tag{6}$$

This has the properties that we require, in that the denominator ensures more emphasis is placed on differences in the ranks of highly ranked variables in either dataset. Furthermore, this distance measure allows comparisons between partial, top-$k$ lists, since a variable occurring in one top-$k$ list but not the other is assigned a ranking of $k + 1$ in the list from which it is missing. Note also that a variable $i$ that is not in either of the top-$k$ ranks, that is $\tau(i), \sigma(i) > k$, makes no contribution to $Ca(k, \tau, \sigma)$.

In order to gauge the extent to which the distance measure (6) differs from that expected between two random lists, we require a value for the expected Canberra distance between two random lists, which we denote $E[Ca(k, p)]$. Jurman et al. [44] derive an expression for this quantity, and we use this to compute the normalised Canberra distance,

$$Ca^*(k, \tau, \sigma) = \frac{Ca(k, \tau, \sigma)}{E[Ca(k, p)]}. \tag{7}$$

Note that this has a lower bound of 0, corresponding to exact agreement between the lists. For two random lists, the upper bound will generally be close to 1, although it can exceed 1, particularly for small $k$, since the expected value for random lists is not necessarily the highest value.

We illustrate the variation of the normalised Canberra distance (7) between SP2 and SiMES pathway rankings in the left hand plot in Figure 15 (blue curve). We consider all possible top-$k$ lists, $k = 1, \ldots, 185$ since all 185 pathways are ranked in both datasets. In the same plot, we also show

$$Ca^*_\pi(k, \tau, \sigma) = \frac{1}{Z} \sum_{\pi=1}^{Z} \frac{Ca(k, \tau, \sigma^\pi)}{E[Ca(k, p)]} \quad k = 1, \ldots, 185 \tag{8}$$

obtained by comparing empirical SP2 rankings ($\tau$) against $Z = 10,000$ permutations of the SiMES pathway rankings, $\sigma^\pi, \pi = 1, \dots, 10,000$ (green curve). This latter curve confirms that the expected value, $E[Ca(k,p)]$, is indeed a good measure of $Ca$ in the random case where there is no agreement between rankings.

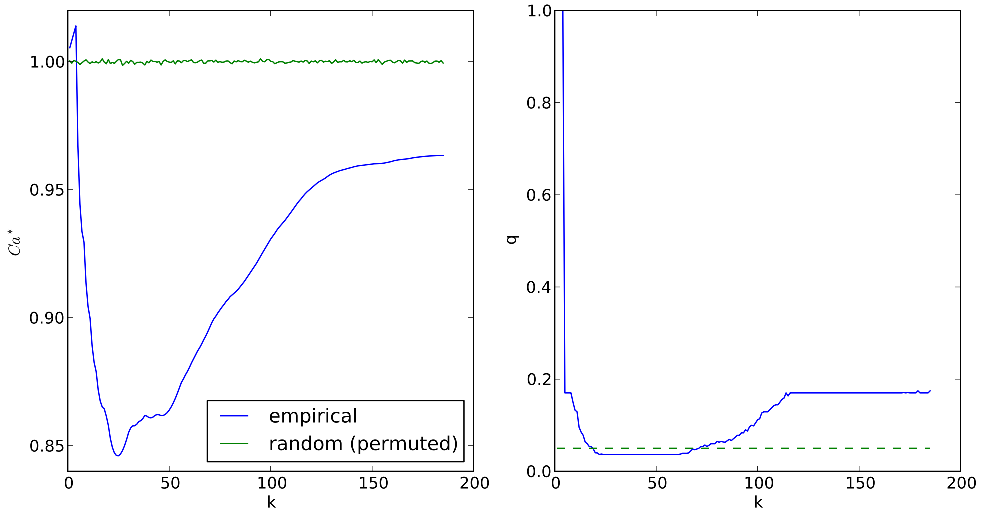


Figure 15: Comparison of top-$k$ SP2 and SiMES pathway rankings. *Left hand plot:* Variation of normalised Canberra distance, $Ca^*$ with $k$ (7) (blue curve). Corresponding mean values over $Z = 10,000$ permutations of SiMES rankings (8) (green curve). *Right hand plot:* FDR $q$-values (blue curve). Dotted green line shows the threshold for FDR control at the 5% level.

Using the same permuted rankings, $\sigma^\pi$, we next test the null hypothesis that the observed normalised Canberra distance, $Ca^*(k, \tau, \sigma)$, is not significantly different from that between $\tau$ and a random list $\sigma^\pi$, by computing a p-value as

$$p^*(k) = \frac{1}{Z} \sum_{\pi=1}^{Z} I_{Ca^*(k,\tau,\sigma) \leq Ca^*(k,\tau,\sigma^\pi)},$$

for $k = 1, \dots, 185$. We then obtain FDR q-values using the Benjamini-Hochberg procedure [5] and illustrate these for each $k$ in the right hand plot of Figure 15. FDR is controlled at a nominal 5% level for $19 \leq k \leq 71$, indicating that the distance between the top-$k$ pathway rankings for both datasets is significantly different from the random ranking case for a wide range of possible values of $k$. The distance $Ca^*$ between SP2 and SiMES pathway rankings however attains its minimum value when $k = 25$ with q(25) = 0.037, so that on this measure, the two pathway rankings are in closest agreement when we consider the top 25 pathways in each ranked list only. Some intuitive understanding of why this might be so can be gained by considering the empirical vs. null pathway selection frequency distributions for each dataset in Figures 10 (b) and 14 (top). Here we see that the separation between empirical and null selection frequencies is most clear for values of $k$ below around 30 for SP2, and around 15 for SiMES.

If we assume that the two pathway rankings are indeed in closest agreement when $k = 25$, then one means of obtaining a consensus set of important pathways is to consider their intersection,

$$\Psi_{25}^{path} = \{i : \tau^{-1}(i) \leq 25\} \cap \{j : \sigma^{-1}(j) \leq 25\},$$

from which we can obtain a set of average rankings as

$$\psi_{25}^{path} = \{\frac{\tau(z) + \sigma(z)}{2} : z \in \Psi_{25}^{path}\}.$$

Both the intersection set, $\Psi_{25}^{path}$, and ordered average rankings, $\psi_{25}^{path}$ for the two datasets under consideration are shown in Table 11. We additionally mark the consensus set $\Psi_{path}^{25}$ with asterisks in Tables 6 and 9.

Table 11: Consensus set of pathways, $\Psi_{25}^{path}$, for SP2 and SiMES datasets with $k = 25$. Consensus pathways are ordered by their average rankings in $\psi_{25}^{path}$.

| Pathway | Average rank ($\psi_{25}^{path}$) |
|---|---|
| Dilated Cardiomyopathy | 4.5 |
| Hypertrophic Cardiomyopathy | 7.5 |
| T Cell Receptor Signaling Pathway | 11.0 |
| Terpenoid Backbone Biosynthesis | 11.0 |
| Arrhythmogenic Right Ventricular Cardiomyopathy | 12.0 |
| Ribosome | 13.0 |
| Ppar Signaling Pathway | 18.5 |

## Comparison of ranked gene and SNP lists

A number of factors complicate the comparison of ranked gene and SNP lists across both datasets. Firstly, sets of mapped SNPs and genes differ slightly between the two datasets (see Table 2). Secondly, even if we consider only those variables *mapped* in both datasets, different, though overlapping sets of variables are *ranked* in each. Thirdly, ranked variables are not independent [44]. For example, genes may be grouped into pathways, so that a reordering of genes within a pathway might be considered less significant than a reordering of genes mapping to different pathways. Similarly a reordering of SNPs mapping to a single gene might be considered less significant than a reordering of SNPs mapping to different genes.

In order to compute a distance measure between pairs or ranked lists, we therefore make two simplifying assumptions. First, we consider only variables ranked in one or both datasets. This seems reasonable, since we can necessarily only compile a distance measure from variables that are ranked in one or both datasets. Second, we assume that variables are independent. This makes our distance measure conservative, in the sense that it will treat all reordering of SNPs or genes equally, irrespective of any potential functional relationship between them.

With these assumptions in mind, we begin by denoting the set of all $p^*$ variables (genes or SNPs) that are ranked in *either* dataset by $\mathcal{L} = \{1, \ldots, p^*\}$. We further denote the corresponding sets of ranked variables for SP2 and SiMES datasets by $\mathcal{L}_\tau$ and $\mathcal{L}_\sigma$ respectively. We then have the following set relations: $\mathcal{L}_\tau, \mathcal{L}_\sigma \subset \mathcal{L}$; $\mathcal{L}_\tau \neq \mathcal{L}_\sigma$; and $|\mathcal{L}_\tau| \neq |\mathcal{L}_\sigma|$.

We now extend the previous Canberra distance measure to encompass the above set relations. We begin, as before, by defining two ranked lists corresponding to the rankings of all the variables in $\mathcal{L}$ for each dataset, although this time we must account for the fact that not all variables in $\mathcal{L}$ are ranked in both. We denote SP2 rankings by $\tau(i), i = 1, \ldots, p^*$, where $\tau(i)$ is the rank of variable $i$ if $i \in \mathcal{L}_\tau$, and $\tau(i) = p^*$ otherwise. SiMES rankings are defined in the same way, and denoted by $\sigma(i), i = 1, \ldots, p^*$.

Applying this revised ranking scheme, we can then define a top-$k$ normalised Canberra distance (6) as

$$Ca^*(k, \tau, \sigma) = \frac{Ca(k, \tau, \sigma)}{E[Ca(k, p^*)]}. \tag{9}$$

for any $k \leq \min\{|\mathcal{L}_\tau|, |\mathcal{L}_\sigma|\}$. The restriction on $k$ follows from the fact that we cannot distinguish between top-$k$ rankings for all $k > \min\{|\mathcal{L}_\tau|, |\mathcal{L}_\sigma|\}$.

*Gene rankings*

Information summarising the relationship between the two ranked lists of genes is given in Table 12.

Table 12: Summary of genes analysed and ranked in SP2 and SiMES datasets.

| | SP2 | SiMES |
|---|---|---|
| number of genes mapped to pathways | 4,734 | 4,751 |
| number of genes mapping to both datasets | 4,726 | |
| number of ranked genes ($|\mathcal{L}_\tau|, |\mathcal{L}_\sigma|$) | 3,430 | 2,815 |
| number of genes ranked in either dataset ($p^*$) | 3,913 | |
| number of genes ranked in both datasets ($|\mathcal{L}_\tau| \cap |\mathcal{L}_\sigma|$) | 2,332 | |

We consider normalised Canberra distances, $Ca^*(k, \tau, \sigma)$, for $k = 1, \ldots, 500$ only, and plot these in Figure 16 (left, blue curve), along with $Ca^*_\pi(k, \tau, \sigma)$ (8) for $Z = 10,000$ permutations of the SiMES pathway rankings, $\sigma^\pi, \pi = 1, \ldots, 10,000$ (green curve). Once again this latter curve confirms that the expected value, $E[Ca(k, p^*)]$, is indeed a good measure of $Ca$ in the random case where there is no agreement between rankings. We also plot FDR $q$-values using the same procedure as described previously for pathways. FDR is controlled at a nominal 5% level for all $k > 13$ in the region tested ($1 \leq k \leq 500$). The distance $Ca^*$ between SP2 and SiMES pathway rankings attains its minimum value when $k = 244$, so that on this measure, the two gene rankings are in closest agreement when we consider the top 244 pathways in each ranked list only.

Following the same strategy as implemented for pathways, we then form the consensus set, $\Psi^{gene}_{244}$, and average rankings $\psi^{gene}_{244}$. The consensus set contains 84 genes, and we list the top 30 genes ordered by their average rank in the two datasets, in Table 13.

Table 13: Top 30 consensus genes ordered by their average rank, $\psi_{244}^{gene}$.

| Rank | Gene | Average rank ($\psi_{244}^{gene}$) |
|---|---|---|
| 1 | *LAMA2* | 9.0 |
| 2 | *ADCY2* | 11.0 |
| 3 | *CACNA1C* | 11.5 |
| 4 | *PRKCB* | 11.5 |
| 5 | *PRKCA* | 21.0 |
| 6 | *EGFR* | 21.5 |
| 7 | *ITGA1* | 24.5 |
| 8 | *CACNA2D3* | 25.5 |
| 9 | *RYR2* | 26.5 |
| 10 | *IGF1R* | 30.5 |
| 11 | *PAK7* | 36.5 |
| 12 | *ADCY8* | 37.5 |
| 13 | *VAV2* | 41.0 |
| 14 | *SLC8A1* | 41.5 |
| 15 | *CACNB2* | 42.5 |
| 16 | *CACNA2D1* | 43.0 |
| 17 | *ITGA9* | 44.0 |
| 18 | *KRAS* | 47.5 |
| 19 | *MAPK10* | 50.5 |
| 20 | *CACNA1S* | 51.0 |
| 21 | *VAV3* | 54.0 |
| 22 | *PLCG2* | 55.5 |
| 23 | *BCL2* | 57.0 |
| 24 | *CD80* | 60.0 |
| 25 | *ITGA11* | 60.5 |
| 26 | *CTNNA2* | 61.0 |
| 27 | *ALDH1B1* | 61.5 |
| 28 | *MGST3* | 63.0 |
| 29 | *NEDD4L* | 63.0 |
| 30 | *PRKAG2* | 66.0 |

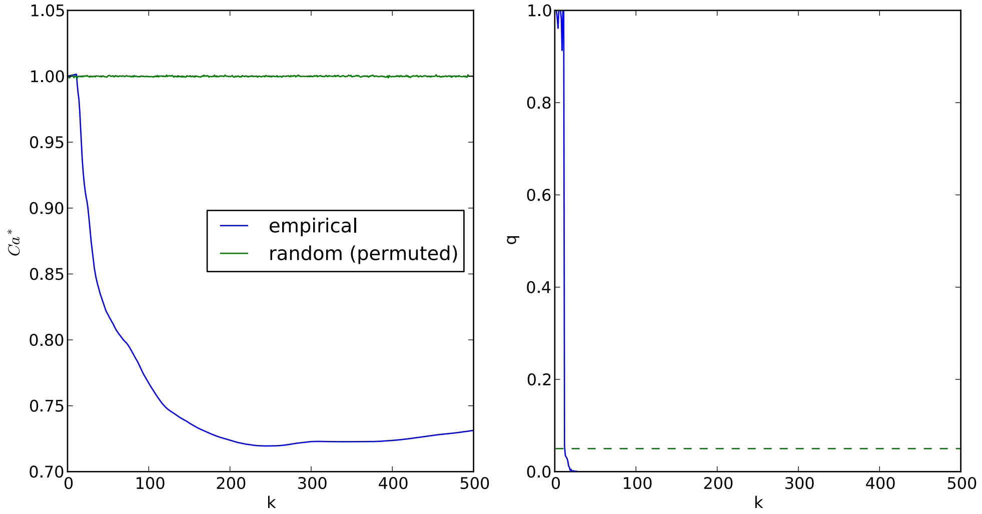


Figure 16: Comparison of top-$k$ SP2 and SiMES gene rankings, for $k = 1, \ldots, 500$. *Left hand plot:* Variation of normalised Canberra distance, $Ca^*$ with $k$ (9) (blue curve), and corresponding mean values over 10,000 permutations of SiMES rankings (8) (green curve). *Right hand plot:* FDR $q$-values (blue curve). Dotted green line shows the threshold for FDR control at the 5% level.

*SNP rankings*

Information summarising the relationship between the two ranked lists of SNPs is given in Table 14. In contrast to both pathway and gene rankings, it is apparent that relatively few ranked SNPs overlap both datasets – 8,151 out of 41,452 SNPs that are ranked in either dataset. This results in values for $Ca^*(k)$ that are close to 1, corresponding to the random list case, over a wide range of possible values for $k$ (data not shown).

For this reason, we compute a simple summary measure

$$\psi^{SNP} = \{\frac{\tau(j) + \sigma(j)}{2} : j \in \mathcal{L}_\tau \cap \mathcal{L}_\sigma\} \tag{10}$$

and report only the top ranking SNPs using this measure in Table 15.

Table 14: Summary of SNPs analysed and ranked in SP2 and SiMES datasets.

| | SP2 | SiMES |
|---|---|---|
| number of SNPs mapped to pathways | 75,389 | 78,933 |
| number of SNPs mapping to both datasets | 74,864 | |
| number of ranked SNPs ($\lvert\mathcal{L}_\tau\rvert, \lvert\mathcal{L}_\sigma\rvert$) | 30,027 | 20,006 |
| number of SNPs ranked in either dataset ($p^*$) | 41,452 | |
| number of SNPs ranked in both datasets ($\lvert\mathcal{L}_\tau\rvert \cap \lvert\mathcal{L}_\sigma\rvert$) | 8,581 | |

Table 15: Top 30 SNPs ranked in both SP2 and SiMES datasets, ranked in order of mean ranking,

| rank | SNP | $\psi^{SNP}$ | mapped gene(s) | | |
|---|---|---|---|---|---|
| 1 | rs897799 | 133.0 | COX6B2 | IL11 | |
| 2 | rs2126953 | 203.0 | ITGA1 | | |
| 3 | rs7714110 | 213.0 | ADCY2 | | |
| 4 | rs6924886 | 274.5 | PDSS2 | | |
| 5 | rs2447867 | 275.5 | ITGA1 | | |
| 6 | rs9386622 | 283.0 | PDSS2 | | |
| 7 | rs6568474 | 283.5 | PDSS2 | | |
| 8 | rs10446497 | 349.5 | PAK2 | | |
| 9 | rs6583177 | 385.5 | PAK2 | | |
| 10 | rs4765961 | 429.0 | CACNA1C | | |
| 11 | rs759440 | 457.0 | PDSS2 | | |
| 12 | rs10457161 | 465.0 | PDSS2 | | |
| 13 | rs10462842 | 479.5 | ADCY2 | | |
| 14 | rs9373932 | 529.0 | PDSS2 | | |
| 15 | rs12206487 | 532.0 | LAMA2 | | |
| 16 | rs743567 | 543.0 | MYH7 | | |
| 17 | rs12472674 | 543.0 | CTNNA2 | | |
| 18 | rs11759792 | 557.5 | PDSS2 | | |
| 19 | rs9373924 | 566.0 | PDSS2 | | |
| 20 | rs319070 | 623.0 | PDSS2 | | |
| 21 | rs751877 | 630.0 | ADCY4 | LTB4R | RIPK3 |
| 22 | rs2047698 | 714.0 | PDGFD | | |
| 23 | rs4804505 | 727.5 | PDE4A | KEAP1 | |
| 24 | rs12672417 | 764.0 | SMURF1 | | |
| 25 | rs7766689 | 800.5 | LAMA2 | | |
| 26 | rs157694 | 860.0 | MAP3K7 | | |
| 27 | rs554192 | 878.0 | NEDD4L | | |
| 28 | rs2746543 | 896.5 | SDHB | | |
| 29 | rs742257 | 942.0 | LAMB3 | | |
| 30 | rs1798619 | 944.5 | PAK2 | | |

# Discussion

We have outlined a method for the detection of pathways, SNPs and genes associated with a quantitative trait. Our method uses a sparse regression model, the sparse group lasso, that enforces sparsity at the pathway and SNP level. As well as identifying important pathways, this approach is designed to maximise the power to detect causal SNPs, possibly of low effect size, that might otherwise be missed if pathways information is ignored. In a simulation study we demonstrated that where causal SNPs are enriched within a single causal pathway, SGL does indeed have greater SNP selection power, compared to an alternative sparse regression model, the lasso, that disregards pathways information. These results mirror previous findings that support the intuition that a sparse selection penalty that promotes dual-level sparsity is better able to recover the true model in these circumstances [26, 69].

We then argued from a theoretical standpoint that where individual SNPs can map to multiple pathways, a modification (SGL-CGD) of the standard SGL-BCGD estimation algorithm that treats pathways as independent, may offer greater sensitivity for the detection of causal SNPs and pathways. A potential concern is that this gain in power may be accompanied by an inflated number of false positives. However, in a simulation study with overlapping pathways we found relative gains in both sensitivity and specificity, under the independence assumption. This gain in specificity was unexpected, and appears to arise directly from treating pathways as independent in the model estimation. As with the group lasso, the ability of SGL to recover the true model is likely to be affected by the complexity of the pathway overlap structure [60], although we expect that the gains in power and sensitivity achieved with the independence assumption will also be apparent with real data.

Our method combines the SGL model and SGL-CGD estimation algorithm with a weight-tuning algorithm to reduce selection bias, and a resampling technique designed to provide a robust measure of SNP, gene and pathway importance in a finite sample. As such, the latter is expected to confer advantages, in terms of the down ranking of unimportant predictors, previously observed for the lasso [54, 13]. Once again it would be interesting to explore this further using simulations derived from real pathway and genotype data.

We do not explore the issue of determining a selection frequency threshold for the control of false positives here. In principal such a threshold could be determined by comparing empirical selection frequency distributions with those obtained under the 'null', through permutations, although this is not a trivial exercise [83]. An alternative method for error control has been investigated in the context of lasso selection [54], but the direct application of this approach to the present case is not feasible, since overlapping pathways make clear distinctions between causal and noise variables problematic. We instead develop a heuristic measure of ranking performance in our application study identifying SNPs and pathways associated with serum high-density lipoprotein cholesterol levels (HDLC). Firstly, by comparing empirical and null pathway and SNP rankings for each dataset, we gain some confidence that pathway and SNP signals captured in the top rankings can be distinguished from those arising from noise or spurious associations. Secondly, we take advantage of the fact that we are able to compare results from two independent GWAS datasets. On the assumption that similar patterns of genetic variation are likely to impact HDLC levels in both cohorts, we set a ranking threshold based on computing distances between ranked lists from each dataset.

Interestingly, when a comparison between empirical and null rankings is made with a reduced value for the regularisation parameter $\alpha$, there is evidence of selection bias, in the sense that pathways and SNPs tend to be highly ranked both empirically and under the null (see Figures 10 and 11). Since a smaller $\alpha$ corresponds to a greater number of SNPs being selected at each subsample, this would seem to suggest that too many SNPs are being selected. In this case, pathway and SNP rankings may in part reflect spurious associations, with a bias towards SNPs overlapping multiple pathways.

There are other potentially interesting areas to explore with regard to the subsampling method used here. For example, standard approaches consider only the set of variables selected at each subsample, and ignore potentially relevant information captured in the coefficient estimates themselves. The use of this additional information would result in a set of ranked lists, one for each subsample, and the joint consideration of these lists has the potential to provide a more robust measure of variable importance, by taking account of the relative importance of each variable for each subsample [64, 47, 43].

Turning to the study results, we conduct two separate analyses on independent discovery and replication datasets. Since subjects from both datasets are genotyped on the same platform, the large majority of SNPs mapping to pathways in one dataset do so also in the other dataset. Thus 99.3% of SNPs mapping to pathways in the SP2 dataset are similarly mapped in the SiMES dataset. For the SiMES dataset, the corresponding figure is 94.8%. As expected, the concordance of gene coverage is even greater. Thus 99.8% of mapped genes in the SP2 dataset are also mapped in the SiMES dataset, and 99.5% of mapped genes in the SiMES dataset are also mapped in SP2. This large overlap in gene (and pathway) coverage between datasets is likely to occur even when datasets are genotyped on different SNP arrays. Indeed this is one advantage of methods such as the one described here that enable comparisons between pathway and gene rankings.

We obtain consensus pathway and gene rankings by considering only the top $k$ ranks in each dataset, with $k$ obtained as the value that minimises the distance between the two rankings. We additionally derive a significance measure for each top-$k$ distance by comparing empirical distances against a null distribution obtained by permuting ranks in one list. We note that this can only be an approximation of the true null, since in reality rankings for both datasets may be influenced by the extent to which genes and SNPs overlap multiple pathways. However, some support for the reasonableness of this approximation can be gained from our earlier analysis, showing that the correlation between empirical and null pathway and SNP rankings is low, so that rankings under the null are indeed approximately random.

Considering the consensus pathway rankings in Table 11, three out of the seven consensus pathways (ranked 1, 2 and 5), are related to cardiomyopathy. These three pathways are the only cardiomyopathy-related pathways amongst the 185 KEGG pathways used in our analysis, so it is noteworthy that all three fall within the consensus pathway rankings. The link between HDLC levels and cardiomyopathy is already well established [1, 29, 81, 24, 27]. Furthermore, numerous references in the literature also describe the links between lipid metabolism and T cell receptor (consensus pathway ranking 3) and PPAR signaling (rank 7) [40, 9, 73, 6].

Turning to a consideration of the top 30 consensus genes and SNPs presented in Tables 13 and 15 (and see also pathway ranking tables 6, 9 and 11, and extended results in supplementary information). We found that many are enriched in one of several gene families:

1. L-type calcium channel genes, including *CACNA1C, CACNA1S, CACNA2D1, CACNA2D3* and *CACNB2*

2. Adenylate cyclase genes, including *ADCY2*, *ADCY4* and *ADCY8*

3. Integrin and laminin genes, including *ITGA1*, *ITGA9*, *ITGA11*, *LAMA2*, and *LAMA3*

4. MAPK signaling pathway genes, including *MAPK10* and *MAP3K7*
5. Immunological pathway genes, including *PAK2*, *PAK7*, *PRKCA*, *PRKCB*, *VAV2* and *VAV3*

These genes are highly enriched in several high ranking pathways from both datasets. Notably, the focal adhesion pathway alone has 12 gene hits, as does the dilated cardiomyopathy pathway. Cardiomyopathy pathways as a whole have 30 genes hits (several of the genes overlap more than one cardiomyopathy pathway). 10 of these genes feature in the MAPK signaling pathway, while GnRH (8 genes), T and B cell receptor (8), calcium (7), ErbB (5), and Wnt signling (4) pathways also contain several genes in the list. To elucidate the biological relevance of these gene families and the connections between them, we investigated their known functional links with cardiovascular phenotypes (not restricted to HDLC) by referencing the KEGG and Genetic Association (http://geneticassociationdb.nih.gov) databases.

*Voltage dependent L-type calcium channel gene family*
The genes in this family encode the subunits of the human voltage dependent L-type calcium channel (CaV1). The $\alpha$-1 subunit (encoded by *CACNA1C*, *A1S*, *A2D1* and *A2D3* in our study) determines channel function in various tissues. CaV1 function has significant impact on the activity of heart cells and smooth muscles. For example, patients with malfunctioning CaV1 develop arrhythmias and shortened QT interval [72, 2, 75]. Furthermore, *CACNA1C* polymorphisms have been associated with variation in blood pressure in Caucasian and East Asian populations by pharmacogenetic analysis. In 120 Caucasians, 3 SNPs in this gene were significantly associated with the response to a widely applied antihypertensive CaV1 blocker [8]. Kamide et al. [45] also found that polymorphisms in *CACNA1C* were associated with sensitivity to an antihypertensive in 161 Japanese patients. The CaV1 $\beta$ subunit encoding *CACNB2* has also been associated with blood pressure [49].

This gene family was mapped to several pathways in our study, with the KEGG dilated cardiomyopathy pathway achieving highest rank both within individual datasets, and in the consensus pathway rankings. Dilated cardiomyopathy is the most common form of cardiomyopathy, and features enlarged and weakened heart muscles. Although high levels of serum HDLC lowers the risk of heart disease [12, 81], there is still no direct evidence that CaV1 is involved in HDLC metabolism.

*Adenylate cyclase gene family*
Three adenylate cyclases genes, *ADCY2*, *ADCY4* and *ADCY8* were highly ranked in our study. Currently, there are no reported associations of these genes with cardiovascular disease or lipid levels. Adenylate cylcase genes catalyse the formation of cyclic adenosine monophosphate (cAMP) from adenosine triphosphate (ATP), while cAMP servers as the second messenger in cell signal transduction. Note that *ADCY2* is insensitive to calcium concentration, suggesting that any association of this gene family with HDLC levels may not be due to any interactions with the CaV1 gene family.

Among high ranking pathways, *ADCY2* and *ADCY8* feature in the dilated cardiomyopathy pathway. Although ADCY4 was not in the top 30 consensus genes, rs751877 in this gene was among the top 30 consensus SNPs.

*Integrin and laminin gene families*

We found 3 genes encoding integrin subunits in our study. Integrins hook to the extracellular matrix (ECM) from the cell surface, and are also important signal transduction receptors which communicate aspects of the cell's physical and chemical environment [58]. Interestingly, laminins are the major component of the ECM, and are relevant to the shape and migration of almost every type of tissue. Both of these two families of genes are therefore highly relevant to the survival and shape of heart muscles. A recent GWAS conducted in a Japanese population confirmed a previous association between *ITGA9* and blood pressure in European populations [74].

Integrin family genes and *LAMA2* were selected primarily within high-ranking cardiomyopathy, focal adhesion and ECM receptor signaling pathways, with once again the dilated cardiomyopathy pathway achieving the highest ranks. However, evidence for *LAMA3* association is weaker, since it was not in the top 30 consensus genes, although a SNP from the *LAMB3* (laminin $\beta$-3) gene was ranked 29 in the consensus SNP list.

*MAPK signaling pathway*

TAK1 (*MAP3K7*) and JNK3 (*MAPK10*) are kinases which regulate cell cycling. They activate or depress downstream transcription factors which mediate cell proliferation, differentiation and inflammation.

JNK activity has been associated with obesity in a mouse model, where the absence of JNK1 (*MAPK8*), a protein in the same family as *MAPK10*, protects against the obesity-induced insulin resistance [33]. The negative correlation between HDLC level and obesity has been well accepted [36].

*Immunological pathways*

PAK (*PAK2* and *PAK7*) genes feature in the high ranking T cell signaling pathway in both SP2 and SiMES datasets. *PRKC* (including *PRKCA* and *PRKCB*), along with *VAV* (*VAV2* and *VAV3*) genes also feature in various high ranking immunological pathways including T cell signaling, Pathogenic Escherichia Coli Infection and Natural Killer Cell Mediated Cytotoxicity. Genes from all 3 of these families are frequently top ranked in these pathways.

*PAK* and *VAV* are activated by antigens, and regulate the T cell cytoskeleton, indicating a possible impact on T cell shape and mobility. In a candidate gene association analysis, *PRKCA* was reported to be associated with HDLC at a nominally significant level, but was not significant after adjusting for multiple testing [51].

In summary, genes enriched in the above gene clusters and pathways may be relevant to heart muscle cell signal transduction, shape and migration, and may thus have functional relevance to the onset of cardiovascular diseases. Many highly ranked genes in our study are also involved in neurological pathways. For example polymorphisms in *CACNA1C* have been associated with bipolar disorder, schizophrenia and major depression [22, 55, 30]. This points to an interesting hypothesis that serum HDLC levels might be regulated not only by metabolism but also by neurological pathways, although the elucidation of any putative biological mechanism underlying such an association obviously exceeds the scope of this study.

Besides the gene families and associated pathways discussed above, a notable feature of the top 30 consensus SNP ranking results presented in Table 15 is the inclusion of 9 SNPs mapping to the *PDSS2* gene. *PDSS2* achieves its high ranking in the consensus SNP list, through its inclusion in the highly ranked terpenoid backbone biosynthesis pathway in SiMES, and this gene is in fact the second highest consensus ranking gene with this dataset. In contrast, this gene is ranked low (216) in the SP2 dataset, which explains why it fails to make the top 30 consensus gene rankings. *PDSS2* encodes subunit 2 of prenyl diphosphate synthase, which determines the length of the isoprenoid chain of coenzyme Q10 (CoQ10) [61]. A deficiency in biosynthesis of CoQ10 has previously been associated with delayed motor development and abnormal renal function, with excess serum lipids [70].

Despite the well established links between lipid metabolism and PPAR signaling noted above, no genes in this high-ranked pathway fall in the top 30 gene rankings for either dataset (see Tables 6 and 15). This could be because the association signal in this pathway is more widely distributed, compared to other high ranking pathways, perhaps indicating heterogeneity in genetic causal factors within our sample, so that different genes and SNPs are highlighted in different subsamples. This would result in reduced gene selection frequencies. Also, genes that overlap multiple putative causal pathways are more likely to be selected in a given subsample, meaning that associated genes mapping to pathways with relatively few overlaps may have lower selection frequencies. This may be the case with genes in the PPAR signaling pathway, whose 63 genes map to an average $2.7 \pm 1.8$ pathways. As a comparison, the 84 genes in the top-ranked dilated cardiomyopathy pathway map to an average $7.2 \pm 3.8$ pathways.

Our study failed to highlight HDLC-associated SNPs identified in previous GWAS (see for example www.genome.gov/gwastudies for an up to date list). A primary reason for this is that the large majority of SNPs identified in previous studies do not map to pathways in our study, either because they fall in intergenic regions, or because they do not feature on the Illumina arrays used here. In addition our method is designed to highlight distributed, small SNP effects that accumulate across gene pathways, and so will likely fail to identify those SNPs with significant marginal effects targeted by GWAS. Furthermore, where there are common mechanisms affecting phenotypes in both cohorts, we would expect to observe the most concordance between the two studies at the pathway level, followed by genes, and lastly SNPs. Indeed this increased heterogeneity at the SNP, and to a lesser extent at the gene level is one motivation for adopting a pathways approach in the first place [35, 34, 10]. This reduced concordance at the SNP level may be due to increased heterogeneity of genetic risk factors between the two datasets. Another important source of variation in SNP selection frequencies is LD between SNPs. The within-pathway lasso penalty will tend to select one of a group of highly correlated SNPs at random, reducing SNP selection frequencies within LD blocks harbouring causal SNPs. An alternative approach would be to consider a different penalty within selected pathways, for example the elastic net [92], which selects groups of correlated variables jointly, although this comes at the cost of introducing a further regularisation parameter to be tuned.

Finally, as with all pathways analyses, a number of limitations with this general approach should be noted. Despite great efforts, pathway assembly is still in its infancy, and the relative sparsity of gene-pathway annotations reflects the fact that our understanding of how the majority of genes functionally interact is at an early stage. As a consequence, annotations from different pathways databases often vary [71], so that the choice of pathways database will impact results [19, 10]. Results are also subject to bias resulting from SNP to gene mapping strategies, so that for example SNP to gene mapping distances will affect the number of unmapped SNPs falling within gene 'deserts' [20]; SNPs may map to relatively large numbers of genes in gene rich areas of the genome; and the mapping of a SNP to its closest gene may obscure a true functional relationships with a more distant gene [87]. Indeed recent research from the ENCODE project indicates that functional elements may in fact be densely distributed throughout the genome [7, 62], and this information has the potential to radically alter future pathways analysis. These issues, together with the fact that PGAS methods are by construction designed to highlight distributed, moderate to small SNP effects, serve to further illustrate the point that pathways analysis should be seen as complementary to studies searching for single markers [86].

# References

[1] Benjamin J Ansell et al. "High-density lipoprotein function recent advances." *Journal of the American College of Cardiology* 46.10 (2005), pp. 1792–8.

[2] Charles Antzelevitch et al. "Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death." *Circulation* 115.4 (2007), pp. 442–9.

[3] Kristin L Ayers and Heather J Cordell. "SNP selection in genome-wide and candidate gene studies via penalized logistic regression." *Genetic epidemiology* 34.8 (2010), pp. 879–91.

[4] Francis R Bach. "Bolasso : Model Consistent Lasso Estimation through the Bootstrap". *Proceedings of the 25th International Conference on Machine Learning.* 2004. 2008.

[5] Yoav Benjamini and Y Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society, Series B* 57.1 (1995), pp. 289–300.

[6] Steven J Bensinger and Peter Tontonoz. "Integration of metabolism and inflammation by lipid-activated nuclear receptors." *Nature* 454.7203 (2008), pp. 470–7.

[7] Bradley E. Bernstein et al. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489.7414 (2012), pp. 57–74.

[8] Troy Bremer et al. "CACNA1C polymorphisms are associated with the efficacy of calcium channel blockers in the treatment of hypertension." *Pharmacogenomics* 7.3 (2006), pp. 271–9.

[9] Philip C. Calder and Parveen Yaqoob. "Lipid Rafts–Composition, Characterization, and Controversies". *J. Nutr.* 137.3 (2007), pp. 545–547.

[10] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application". *American Journal of Human Genetics* 86 (2010), pp. 6–22.

[11] Scott L Carter et al. "Gene co-expression network topology provides a framework for molecular characterization of cellular state." *Bioinformatics (Oxford, England)* 20.14 (2004), pp. 2242–50.

[12] W P Castelli. "Cholesterol and lipids in the risk of coronary artery disease–the Framingham Heart Study." *The Canadian journal of cardiology* 4 Suppl A (1988), 5A–10A.

[13] A. Chatterjee and S.N. Lahiri. "Bootstrapping Lasso Estimators". *Journal of the American Statistical Association* 106.494 (2011), pp. 608–625.

[14] Soumyadeep Chatterjee et al. “Sparse Group Lasso for Regression on Land Climate Variables”. *2011 IEEE 11th International Conference on Data Mining Workshops* (2011), pp. 1–8.

[15] Lin S Chen et al. “Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data”. *American Journal of Human Genetics* 86.6 (2010), pp. 860–871.

[16] Xi Chen and Han Liu. “An Efficient Optimization Algorithm for Structured Sparse CCA, with Applications to eQTL Mapping”. *Statistics in Biosciences* 4.1 (2011), pp. 3–26.

[17] Seoae Cho et al. “Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis.” *Annals of human genetics* 74.5 (2010), pp. 416–28.

[18] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. “A linear complexity phasing method for thousands of genomes.” *Nature methods* 9.2 (2012), pp. 179–81.

[19] Clara C Elbers et al. “Using genome-wide pathway analysis to unravel the etiology of complex diseases.” *Genetic epidemiology* 33.5 (2009), pp. 419–31.

[20] Hariklia Eleftherohorinou et al. “Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases.” *PloS one* 4.11 (2009), e8068.

[21] Hariklia Eleftherohorinou et al. “Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways”. *Human molecular genetics* (2011), pp. 1–13.

[22] Manuel A R Ferreira et al. “Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder.” *Nature genetics* 40.9 (2008), pp. 1056–8.

[23] Rina Foygel and Mathias Drton. “Exact block-wise optimization in group lasso and sparse group lasso for linear regression”. 2010.

[24] H.F.G. Freitas et al. “Association of HDL cholesterol and triglycerides with mortality in patients with heart failure”. *Brazilian Journal of Medical and Biological Research* 42.5 (2009), pp. 420–425.

[25] Brooke L Fridley and Joanna M Biernacka. “Gene set analysis of SNP data: benefits, challenges, and future directions.” *European journal of human genetics : EJHG* 19.8 (2011), pp. 837–843.

[26] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “A note on the group lasso and a sparse group lasso” (2010), pp. 1–8. arXiv:/arxiv.org/abs/1001.0736 [http:].

[27] Sainath Gaddam et al. “Serum lipoprotein levels in takotsubo cardiomyopathy vs. myocardial infarction.” *International archives of medicine* 4.1 (2011), p. 14.

[28] David B Goldstein. “Common genetic variation and human traits.” *The New England journal of medicine* 360.17 (2009), pp. 1696–8.

[29] D. J. Gordon et al. “High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies”. *Circulation* 79.1 (1989), pp. 8–15.

[30] E K Green et al. “The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia.” *Molecular psychiatry* 15.10 (2010), pp. 1016–22.

[31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* 2nd. Springer, New York, 2008.

[32] Lucia A Hindorff et al. “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.” *Proceedings of the National Academy of Sciences of the United States of America* 106.23 (2009), pp. 9362–7.

[33] Jiro Hirosumi et al. “A central role for JNK in obesity and insulin resistance.” *Nature* 420.6913 (2002), pp. 333–6.

[34] Joel N Hirschhorn. “Genomewide association studies–illuminating biologic pathways.” *The New England journal of medicine* 360.17 (2009), pp. 1699–701.

[35] Peter Holmans et al. “Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder.” *American journal of human genetics* 85.1 (2009), pp. 13–24.

[36] Barbara V Howard, Giacomo Ruotolo, and David C Robbins. “Obesity and dyslipidemia.” *Endocrinology and metabolism clinics of North America* 32.4 (2003), pp. 855–67.

[37] Bryan Howie, Jonathan Marchini, and Matthew Stephens. “Genotype Imputation with Thousands of Genomes”. *G3 (Bethesda)* 1.November (2011), pp. 457–469.

[38] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. “Learning with Structured Sparsity”. *Journal of Machine Learning Research* 12.12 (2011), pp. 3371–3412.

[39] Laurent Jacob, Guillaume Obozinski, and Jean-philippe Vert. “Group Lasso with Overlap and Graph Lasso”. *Proceedings of the 26th International Conference on Machine Learning.* 2009.

[40] P W Janes et al. “The role of lipid rafts in T cell antigen receptor (TCR) signalling.” *Seminars in immunology* 12.1 (2000), pp. 23–34.

[41] Rodolphe Jenatton and Francis Bach. “Structured Variable Selection with Sparsity-Inducing Norms”. *Journal of Machine Learning Research* 12 (2011), pp. 2777–2824.

[42] H Jeong et al. “Lethality and centrality in protein networks.” *Nature* 411.6833 (2001), pp. 41–2.

[43] Giuseppe Jurman et al. “Algebraic comparison of partial lists in bioinformatics.” *PloS one* 7.5 (2012), e36540.

[44] Giuseppe Jurman et al. “Algebraic stability indicators for ranked lists in molecular profiling.” *Bioinformatics (Oxford, England)* 24.2 (2008), pp. 258–64.

[45] Kei Kamide et al. “Genetic polymorphisms of L-type calcium channel alpha1C and alpha1D subunit genes are associated with sensitivity to the antihypertensive effects of L-type dihydropyridine calcium-channel blockers.” *Circulation journal : official journal of the Japanese Circulation Society* 73.4 (2009), pp. 732–40.

[46] Yoo-Ah Kim, Stefan Wuchty, and Teresa M Przytycka. “Identifying causal genes and dysregulated pathways in complex diseases.” *PLoS computational biology* 7.3 (2011), e1001095.

[47] Raivo Kolde et al. "Robust rank aggregation for gene list integration and meta-analysis." *Bioinformatics (Oxford, England)* 28.4 (2012), pp. 573–80.

[48] Ben Lehner et al. "Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways." *Nature genetics* 38.8 (2006), pp. 896–903.

[49] Daniel Levy et al. "Genome-wide association study of blood pressure and hypertension." *Nature genetics* 41.6 (2009), pp. 677–87.

[50] Jun Liu and Jieping Ye. "Fast Overlapping Group Lasso" (2010), pp. 1–14. arXiv:1009.0306.

[51] Yingchang Lu et al. "Multiple genetic variants along candidate pathways influence plasma high-density lipoprotein cholesterol concentrations." *Journal of lipid research* 49.12 (2008), pp. 2582–9.

[52] Teri a. Manolio et al. "Finding the missing heritability of complex diseases". *Nature* 461.7265 (2009), pp. 747–753.

[53] Mark I McCarthy et al. "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." *Nature Reviews Genetics* 9.5 (2008), pp. 356–69.

[54] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010), pp. 417–473.

[55] V Moskvina et al. "Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk." *Molecular psychiatry* 14.3 (2009), pp. 252–60.

[56] Allan J Motyer et al. "LASSO model selection with post-processing for a genome-wide association study data set." *BMC proceedings*. Vol. 5 Suppl 9. Suppl 9. BioMed Central Ltd, 2011, S24.

[57] K K Namboodiri et al. "The Collaborative Lipid Research Clinics Family Study: biological and cultural determinants of familial resemblance for plasma lipids and lipoproteins." *Genetic epidemiology* 2.3 (1985), pp. 227–54.

[58] M V Nermut et al. "Electron microscopy and structural model of human fibronectin receptor." *The EMBO journal* 7.13 (1988), pp. 4093–9.

[59] Jie Peng et al. "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer". *The Annals of Applied Statistics* 4.1 (2010), pp. 53–77.

[60] Daniel Percival. "Theoretical properties of the overlapping groups lasso". *Electronic Journal of Statistics* 6 (2012), pp. 269–288.

[61] Ryoichi Saiki et al. "Characterization of solanesyl and decaprenyl diphosphate synthases in mice and humans." *The FEBS journal* 272.21 (2005), pp. 5606–22.

[62] Amartya Sanyal et al. "The long-range interaction landscape of gene promoters". *Nature* 489.7414 (2012), pp. 109–113.

[63] Eric E Schadt. "Molecular networks as sensors and drivers of common human diseases." *Nature* 461.7261 (2009), pp. 218–23.

[64] D Sculley. "Rank Aggregation for Similar Items". *Proceedings of the 2007 SIAM International Conference on Data Mining* i (2007), pp. 587–592.

[65] Gang Shi et al. “Mining Gold Dust Under the Genome Wide Significance Level: A Two-Stage Approach to Analysis of GWAS”. *Genetic epidemiology* 35.2 (2011), pp. 117–118.

[66] Matt Silver and Giovanni Montana. “Fast Identification of Biological Pathways Associated with a Quantitative Trait Using Group Lasso with Overlaps”. *Statistical Applications in Genetics and Molecular Biology* 11.1 (2012).

[67] Matt Silver et al. “Identification of gene pathways implicated in Alzheimer’s disease using longitudinal imaging phenotypes with sparse regression”. *NeuroImage* 63.3 (2012), pp. 1681–1694.

[68] Xueling Sim et al. “Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia.” *PLoS genetics* 7.4 (2011), e1001363.

[69] Noah Simon et al. “A sparse-group lasso”. *Journal of Computational and Graphical Statistics* In press (2012), pp. 1–13.

[70] C Sobreira et al. “Mitochondrial encephalomyopathy with coenzyme Q10 deficiency.” *Neurology* 48.5 (1997), pp. 1238–43.

[71] Donny Soh et al. “Consistency, comprehensiveness, and compatibility of pathway databases”. *BMC Bioinformatics* 11.1 (2010), p. 449.

[72] Igor Splawski et al. “Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism.” *Cell* 119.1 (2004), pp. 19–31.

[73] B. Staels et al. “Mechanism of Action of Fibrates on Lipid and Lipoprotein Metabolism”. *Circulation* 98.19 (1998), pp. 2088–2093.

[74] Fumihiko Takeuchi et al. “Blood pressure and hypertension are associated with 7 loci in the Japanese population.” *Circulation* 121.21 (2010), pp. 2302–9.

[75] Christian Templin et al. “Identification of a novel loss-of-function calcium channel gene mutation in short QT syndrome (SQTS6).” *European heart journal* 32.9 (2011), pp. 1077–88.

[76] Tanya M Teslovich et al. “Biological, clinical and population relevance of 95 loci for blood lipids.” *Nature* 466.7307 (2010), pp. 707–13.

[77] The 1000 Genomes Project Consortium. “A map of human genome variation from population-scale sequencing”. *Nature* 467 (2011), pp. 1061–1073.

[78] Rob Tibshirani. “Regression shrinkage and selection via the Lasso”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1 (1996), pp. 267–288.

[79] Robert Tibshirani and Pei Wang. “Spatial smoothing and hot spot detection for CGH data using the fused lasso.” *Biostatistics (Oxford, England)* 9.1 (2008), pp. 18–29.

[80] Robert Tibshirani et al. “Sparsity and smoothness via the fused lasso”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1 (2005), pp. 91–108.

[81] Peter P Toth. “Cardiology patient page. The ”good cholesterol”: high-density lipoprotein.” *Circulation* 111.5 (2005), e89–91.

[82] Paul Tseng and Sangwoon Yun. “A coordinate gradient descent method for nonsmooth separable minimization”. *Mathematical Programming* 117.1 (2009), pp. 387–423.

[83] William Valdar et al. “Reprioritizing genetic associations in hit regions using LASSO-based resample model averaging.” *Genetic epidemiology* 36.5 (2012), pp. 451–62.

[84] Peter M Visscher et al. “Five years of GWAS discovery.” *American journal of human genetics* 90.1 (2012), pp. 7–24.

[85] Maria Vounou et al. “Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer’s disease.” *NeuroImage* 60 (2011), pp. 700–716.

[86] Kai Wang, Mingyao Li, and Hakon Hakonarson. “Analysing biological pathways in genome-wide association studies”. *Nature Reviews Genetics* 11.12 (2010), pp. 843–854.

[87] Kai Wang et al. “Diverse Genome-wide Association Studies Associate the IL12 / IL23 Pathway with Crohn Disease”. *American journal of human genetics* 84.3 (2009), pp. 399–405.

[88] Tong Tong Wu et al. “Genome-wide association analysis by lasso penalized logistic regression.” *Bioinformatics (Oxford, England)* 25.6 (2009), pp. 714–21.

[89] Jingyuan Zhao et al. “Pathway-based analysis using reduced gene subsets in genome-wide association studies.” *BMC bioinformatics* 12.1 (2011), p. 17.

[90] Peng Zhao, Guilherme Rocha, and Bin Yu. “The composite absolute penalties family for grouped and hierarchical variable selection”. *The Annals of Statistics* 37.6A (2009), pp. 3468–3497.

[91] Hua Zhou et al. “Association Screening of Common and Rare Genetic Variants by Penalized Regression.” *Bioinformatics (Oxford, England)* 26.19 (2010), pp. 2375–2382.

[92] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

# Supplementary Information

## A SGL estimation algorithm

For the estimation of $\boldsymbol{\beta}^{SGL}$ we proceed by noting that the optimisation (1) is convex, and (in the case of non-overlapping groups) that the penalty is block-separable, so that we can obtain a solution using block, or group-wise coordinate descent [82]. For a single group, $l$, the corresponding minimising function is given by

$$f(\boldsymbol{\beta}_l) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + (1-\alpha)\lambda w_l||\boldsymbol{\beta}_l||_2 + \alpha\lambda||\boldsymbol{\beta}_l||_1. \tag{11}$$

An optimal solution for SNP coefficient $\beta_j$ is then derived from the subgradient equations

$$-\mathbf{x}_j'(\hat{\mathbf{r}}_l - \sum_{k\neq j}\mathbf{x}_k\hat{\beta}_k - \mathbf{x}_j\beta_j) + (1-\alpha)\lambda w_l s_j + \alpha\lambda t_j = 0 \quad j = l_1, \ldots, l_{P_l}, \tag{12}$$

where $\hat{\beta}_k, k \neq j$ are the current estimates for other SNP coefficients in group $l$, and the group partial residual, $\hat{\mathbf{r}}_l = \mathbf{y} - \sum_{m\neq l}\mathbf{X}_m\hat{\boldsymbol{\beta}}_m$. Here $s_j$ and $t_j$ are the respective subgradients of

$||\boldsymbol{\beta}_l||_2$ and $|\beta_j|$, with

$$s_j = \begin{cases} \frac{\beta_j}{||\boldsymbol{\beta}_l||_2} & \text{if } ||\boldsymbol{\beta}_l||_2 \neq \mathbf{0} \\ \in [-1, 1] & \text{if } ||\boldsymbol{\beta}_l||_2 = \mathbf{0} \end{cases}$$
$$t_j = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \in [-1, 1] & \text{if } \beta_j = 0. \end{cases} \tag{13}$$

If $\boldsymbol{\beta}_l = \mathbf{0}$, that is group $l$ is not selected by the model, then from (12)

$$-\mathbf{x}_j'\hat{\mathbf{r}}_l + (1-\alpha)\lambda w_l s_j + \alpha\lambda t_j = 0, \quad j = l_1, \ldots, l_{P_l}. \tag{14}$$

Substituting $\mathbf{a} = \mathbf{X}_l'\hat{\mathbf{r}}_l$ gives

$$a_j = (1-\alpha)\lambda w_l s_j + \alpha\lambda t_j, \quad j = l_1, \ldots, l_{P_l}$$

so that

$$s_j^2 = \frac{1}{(1-\alpha)^2\lambda^2 w_l^2}(a_j - \alpha\lambda t_j)^2, \quad j = l_1, \ldots, l_{P_l},$$

and

$$\sum_j s_j^2 = \frac{1}{(1-\alpha)^2\lambda^2 w_l^2}\sum_j (a_j - \alpha\lambda t_j)^2.$$

From (13), when $\boldsymbol{\beta}_l = \mathbf{0}$, $||s||_2 = (\sum_j s_j^2)^{\frac{1}{2}} \leq 1$, so that

$$\sum_j (a_j - \alpha\lambda t_j)^2 \leq (1-\alpha)^2\lambda^2 w_l^2. \tag{15}$$

Also from (13), one further condition when $\boldsymbol{\beta}_l = \mathbf{0}$ is that $t_j \in [-1, 1]$. The values, $\hat{t}_j$ that minimise the left hand size of (15) are therefore given by

$$\hat{t}_j = \begin{cases} \frac{a_j}{\alpha\lambda} & \text{if } |\frac{a_j}{\alpha\lambda}| \leq 1 \\ \text{sign}(\frac{a_j}{\alpha\lambda}) & \text{if } |\frac{a_j}{\alpha\lambda}| > 1. \end{cases}$$

Substituting for $a_j$, we can then write the values for $a_j - \alpha\lambda t_j$ that minimise the left hand side of (15) as

$$\begin{aligned} a_j - \alpha\lambda t_j &= \begin{cases} 0 & \text{if } |\mathbf{x}_j'\hat{\mathbf{r}}_l| \leq \alpha\lambda \\ \text{sign}(\mathbf{x}_j'\hat{\mathbf{r}}_l)(|\mathbf{x}_j'\hat{\mathbf{r}}_l| - \alpha\lambda) & \text{if } |\mathbf{x}_j'\hat{\mathbf{r}}_l| > \alpha\lambda \end{cases} \\ &= S(\mathbf{x}_j'\hat{\mathbf{r}}_l, \alpha\lambda) \end{aligned}$$

for $j = l_1, \ldots, l_{P_l}$, where

$$S(\mathbf{x}_j'\hat{\mathbf{r}}_l, \alpha\lambda) = \text{sign}(\mathbf{x}_j'\hat{\mathbf{r}}_l)(|\mathbf{x}_j'\hat{\mathbf{r}}_l| - \alpha\lambda)_+ \tag{16}$$

is the lasso soft thresholding operator. Finally, we can now rewrite the condition for $\hat{\boldsymbol{\beta}}_l = \mathbf{0}$, (15) as

$$||S(\mathbf{X}_l'\hat{\mathbf{r}}_l, \alpha\lambda)||_2 \leq (1-\alpha)\lambda w_l, \tag{17}$$

Where the vector $S(\mathbf{X}_l'\hat{\mathbf{r}}_l, \alpha\lambda) = [S(\mathbf{x}_{l_1}'\hat{\mathbf{r}}_l, \alpha\lambda), \ldots, S(\mathbf{x}_{l_P}'\hat{\mathbf{r}}_l, \alpha\lambda)]$. Note that with $\alpha = 0$, this reduces to the group lasso group selection criterion.

In the case that $\boldsymbol{\beta}_l \neq \mathbf{0}$, that is group $l$ is selected by the model, from (12) and (13) we see that $\beta_j = 0$ when

$$-\mathbf{x}_j'(\hat{\mathbf{r}}_l - \sum_{k \neq j} \mathbf{x}_k \hat{\beta}_k) \leq |\alpha\lambda|. \tag{18}$$

For completeness, we rewrite the criterion for selecting pathway $l$ from (17) as

$$||S(\mathbf{X}_l'\hat{\mathbf{r}}_l, \alpha\lambda)||_2 > (1-\alpha)\lambda w_l \tag{19}$$

and the criterion for selecting SNP $j$ in selected pathway $l$ from (18) as

$$|\mathbf{x}_j'\hat{\mathbf{r}}_{l,j}| > \alpha\lambda \tag{20}$$

where $\hat{\mathbf{r}}_{l,j} = \hat{\mathbf{r}}_l - \sum_{k \neq j} \mathbf{x}_k \hat{\beta}_k$ is the SNP partial residual, obtained by regressing out the current estimated effects of all other predictors in the model, except for predictor $j$.

A number of methods for the estimation of $\boldsymbol{\beta}_l$ in the case that $||\boldsymbol{\beta}_l||_2 \neq \mathbf{0}$ have been proposed [26, 23, 50, 69]. A complicating factor is the discontinuities in the first (and second) derivatives of $s_j$ at $||\boldsymbol{\beta}_l||_2 = 0$, that is where $||\boldsymbol{\beta}_l||_2$ first moves away from zero, and of $t_j$ when $\beta_j = 0$. As with GL, Friedman, Hastie, and Tibshirani [26] describe a numerical method using coordinate descent, by combining a golden search over $\beta_j$ with parabolic interpolation. However we find this too computationally intensive for the large datasets we wish to analyse. Simon et al. [69] propose an accelerated, block gradient descent method in which $\boldsymbol{\beta}_l$ is iteratively updated in a single step along the line of steepest descent of the block objective function until convergence. We instead use a block, coordinate-wise gradient descent (BCGD) method that uses a Newton update, similar to that proposed by Zhou et al. [91], and we describe this below.

To update $\beta_j$ from its current estimate, $\hat{\beta}_j$, we note from (12) and (13) that if $\hat{\beta}_j \neq 0$, the subgradient equation for predictor $j$ is given by

$$\partial_j = -\mathbf{x}_j'(\hat{\mathbf{r}}_l - \mathbf{X}_l\hat{\boldsymbol{\beta}}_l) + (1-\alpha)\lambda w_l \frac{\hat{\beta}_j}{||\hat{\boldsymbol{\beta}}_l||_2} + \alpha\lambda \cdot \mathrm{sign}(\hat{\beta}_j). \tag{21}$$

We then descend along the gradient at $\hat{\beta}_j$ towards the minimum using Newton's method. The Newton update, $\beta_j^*$, is then given by

$$\begin{aligned} \hat{\beta}_j^* &= \hat{\beta}_j - \frac{\partial_j}{\partial_j'}, \\ \text{where} \quad \partial_j' &= 1 + \frac{(1-\alpha)\lambda w_l}{||\hat{\boldsymbol{\beta}}_l||_2}\Big(1 - \frac{\hat{\beta}_j^2}{||\hat{\boldsymbol{\beta}}_l||_2^2}\Big) \end{aligned} \tag{22}$$

is the derivative of (21) at $\hat{\beta}_j$. The update (22) is repeated until convergence.

We must also deal with the case where $\hat{\beta}_j = 0$. Here we adopt a slightly different strategy, since the partial derivative, $t_j$ of $\beta_j$ is not continuous. We avoid this discontinuity by testing the 'directional derivatives', $\partial_j^+$ and $\partial_j^-$, respectively representing the partial derivatives at $\beta_j = 0$ in the direction of increasing and decreasing $\beta_j$. Recalling that we are dealing with the case $||\boldsymbol{\beta}_l||_2 \neq \mathbf{0}$, at $\beta_j = 0$ the group penalty term in (21) disappears. That is, once a group is selected by model it becomes easier for each SNP coefficient to move away from zero. The two directional derivatives are then given by

$$\begin{aligned}\partial_j^+ &= -\mathbf{x}_j'(\hat{\mathbf{r}}_l - \mathbf{X}_l\hat{\boldsymbol{\beta}}_l) + \alpha\lambda \\ \partial_j^- &= -\mathbf{x}_j'(\hat{\mathbf{r}}_l - \mathbf{X}_l\hat{\boldsymbol{\beta}}_l) - \alpha\lambda.\end{aligned} \tag{23}$$

Since the minimising function (11) is convex, there are three possible outcomes, and we substitute for $\partial_j$ in (22) accordingly:

$$\partial_j \leftarrow \begin{cases} \partial_j^- & \text{if} \quad \partial_j^- > 0 \quad \text{and} \quad \partial_j^+ > 0 \\ \partial_j^+ & \text{if} \quad \partial_j^- < 0 \quad \text{and} \quad \partial_j^+ < 0 \\ 0 & \text{if} \quad \partial_j^- > 0 \quad \text{and} \quad \partial_j^+ < 0 \end{cases} \tag{24}$$

In the third case, $f(\boldsymbol{\beta}_l)$ is increasing either side of $\beta_j = 0$, so that $\hat{\beta}_j$ must remain at zero. We can then proceed with the standard Newton update (22).

Finally, since the Newton update may occasionally overstep the minimum (where $\partial_j = 0$), a simple remedy proposed by Zhou et al. [91] is to check that $f(\boldsymbol{\beta}_l)$ is decreasing at each iteration. If this is not the case, then the step size in (22) is halved. The complete algorithm for SGL estimation using BCGD is presented in Box 1.

One remaining practical issue is the obtaining of a value for $\lambda_{max}$, the smallest value of $\lambda$ at which no groups are selected by the model. Noting that $\hat{\mathbf{r}}_l = \mathbf{y}$ when no groups are selected, from (17) we obtain the smallest value, $\lambda_l^{min}$, for the minimum value of $\lambda$ at which group $l$ is not selected as

$$\lambda_l^{min} = \frac{||S(\mathbf{X}_l'\mathbf{y}, \alpha\lambda_l^{min})||_2}{(1-\alpha)w_l} \tag{25}$$

We can solve this in its quadratic form by first setting an upper bound for $\lambda$ at the point $\lambda_l^*$, where the soft thresholding function $S(\mathbf{X}_l'\mathbf{y}_l, \alpha\lambda) = \mathbf{0}$, that is when no SNPs are selected by the model. We then obtain the solution by solving

$$||S(\mathbf{X}_l'\mathbf{y}, \alpha\lambda_l^{min})||_2^2 - (1-\alpha)^2(\lambda_l^{min})^2 w_l^2 = 0 \quad 0 < \lambda_l^{min} < \lambda_l^* \tag{26}$$

for $\lambda_l^{min}$, where

$$\lambda_l^* = \max_j \frac{|\mathbf{x}_j'\mathbf{y}|}{\alpha}, \quad j = l_1, \ldots, l_{P_l}.$$

We do this using the 1d root-finding function, *brentq*, in Python's *scipy* library. Finally, we obtain a value for $\lambda_{max}$ as

$$\lambda_{max} = \max_l \lambda_l^{min}, \quad l = 1, \ldots, L. \tag{27}$$

# B SGL with overlaps

We assume that $\mathbf{X}$ and $\boldsymbol{\beta}$ have been expanded to account for overlaps, but we drop the $^*$ notation for clarity. We proceed as before by solving the block-separable optimisation (4) for each group or pathway in turn. However, for overlapping pathways, the assumption of pathway independence requires that each $\mathbf{X}_l, (l = 1, \ldots, L)$ is regressed against the full phenotype vector $\mathbf{y}$ rather than the partial residual, $\hat{\mathbf{r}}_l$. With this in mind, the revised subgradient equations for group $l$ (12) are given by

$$-\mathbf{x}_j'(\mathbf{y} - \sum_{k \neq j} \mathbf{x}_k \hat{\beta}_k - \mathbf{x}_j \beta_j) + (1-\alpha)\lambda w_l s_j + \alpha\lambda t_j = 0 \quad j = l_1, \ldots, l_{P_l}. \tag{28}$$

The estimation for group $l$ then proceeds as described in the previous section, but with the partial residual $\hat{\mathbf{r}}_l$ replaced by $\mathbf{y}$, so that the group sparsity condition (17) for $||\hat{\boldsymbol{\beta}}_l||_2 = \mathbf{0}$ becomes

$$||S(\mathbf{X}_l'\mathbf{y}, \alpha\lambda)||_2 \leq (1-\alpha)\lambda w_l. \tag{29}$$

As before, where group $l$ is selected by the model, the update for $\beta_j$, with current estimate $\hat{\beta}_j$, is derived from the partial derivative (21), which under the independence assumption is given by

$$\partial_j = -\mathbf{x}_j'(\mathbf{y} - \mathbf{X}_l\hat{\boldsymbol{\beta}}_l) + (1-\alpha)\lambda w_l \frac{\hat{\beta}_j}{||\hat{\boldsymbol{\beta}}_l||_2} + \alpha\lambda \cdot \text{sign}(\hat{\beta}_j), \tag{30}$$

for $j = l_1, \ldots, l_{P_l}$. The Newton update (22) remains the same. When $\hat{\beta}_j = 0$, the revised directional derivatives (23) are given by

$$\begin{aligned} \partial_j^+ &= -\mathbf{x}_j'(\mathbf{y} - \mathbf{X}_l\hat{\boldsymbol{\beta}}_l) + \alpha\lambda \\ \partial_j^- &= -\mathbf{x}_j'(\mathbf{y} - \mathbf{X}_l\hat{\boldsymbol{\beta}}_l) - \alpha\lambda. \end{aligned} \tag{31}$$

As before the conditions for SNP sparsity within a selected group are determined by (24).

The value of $\lambda_{max}$, the smallest $\lambda$ value at which no group is selected by the model, is determined in the same way as before, since this procedure (described in (25), (26) and (27)) does not depend on $\hat{\mathbf{r}}_l$.

Importantly, since each group is regressed independently against the phenotype vector $\mathbf{y}$, there is no block coordinate descent stage in the estimation, that is the revised algorithm utilises only coordinate gradient descent within each selected pathway. For this reason we use the acronym SGL-CGD for the revised algorithm. The new algorithm is described in Box 2. Note that since the block coordinate descent stage is avoided, the new algorithm has the added benefit of being much faster than would otherwise be the case.

# C Simulation study 1

A baseline phenotype, $y$ is sampled from $\mathcal{N}(10, 1)$. To generate SNP effects, we first select a single pathway, $\mathcal{G}_l$, at random. From this pathway we randomly select 5 SNPs to from the set $\mathcal{S} \subset \mathcal{G}_l$ of causal SNPs. At each MC simulation we generate a genetic effect and adjust $y$ so that

$$y^* = y + w$$

where

$$w = \delta \sum_{k \in \mathcal{S}} \zeta_k x_k.$$

Here $\delta$ controls the overall additive genetic effect on phenotype $y$ due to all casual SNPs in $\mathcal{S}$, and $\zeta_k$ determines the contribution from causal SNP $k$, with $\sum_{k \in \mathcal{S}} \zeta_k = 1$. In our simulations we maintain a constant overall genetic effect size,

$$\gamma = \mathrm{E}(w)/\mathrm{E}(y)$$

across all affected phenotypes, so that $\gamma$ represents the proportionate increase in the mean value of $y$ due to all genetic effects. We also set $\zeta_k = 1/5$, for $k \in \mathcal{S}$, so that the contribution from each causal SNP allele is equal. This enables us to determine $\delta$ for a given $\gamma$ as

$$\delta = \frac{5\gamma \mathrm{E}(y)}{2\sum_{k\in\mathcal{S}} m_k}.$$

Note that for constant $\gamma$, the proportionate effect on the mean value of $y$ due to SNP $k$ is MAF dependent, and is given by $2\delta m_k/\mathrm{E}(y)$.

# D Weight tuning for bias reduction

For fixed $\alpha$, and with $\lambda$ tuned to select a single pathway, we need to establish which pathway enters the model first, as $\lambda$ is reduced from its maximal value, $\lambda_{max}$. From (27), at phenotype permutation $r$, the pathway $\hat{\mathcal{C}}_r$ selected with permuted phenotype $\mathbf{y}_r$ is given by

$$\hat{\mathcal{C}}_r = \arg\max_l \lambda_l^{min}, \quad , l = 1, \dots, L.$$

$\lambda_l^{min}$ is obtained by solving

$$\lambda_l^{min} = \frac{||S(\mathbf{X}_l'\mathbf{y}_r, \alpha\lambda_l^{min})||_2}{(1-\alpha)w_l},$$

using the procedure described at the end of Section A. For $R$ permutations of the phenotype vector, $\mathbf{y}$, the empirical pathway selection frequency distribution is then given by

$$\Pi^*(\mathbf{w}) = \frac{1}{R}\sum_{r=1}^{R}\{\hat{\mathcal{C}}_r = l\}, \quad l = 1, \dots, L.$$