Bioinformatics, 34(16), 2018, 2788–2796
doi: 10.1093/bioinformatics/bty187
Advance Access Publication Date: 28 March 2018
Original Paper

OXFORD

Genetics and population analysis

LSMM: a statistical approach to integrating functional annotations with genome-wide association studies

Jingsi Ming[1], Mingwei Dai[2,3], Mingxuan Cai[1], Xiang Wan[4], Jin Liu[5,]* and Can Yang[3,]*

[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong, [2]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China, [3]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, [4]Shenzhen Research Institute of Big Data, Shenzhen, China and [5]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on October 23, 2017; revised on February 14, 2018; editorial decision on March 22, 2018; accepted on March 27, 2018

Abstract

Motivation: Thousands of risk variants underlying complex phenotypes (quantitative traits and diseases) have been identified in genome-wide association studies (GWAS). However, there are still two major challenges towards deepening our understanding of the genetic architectures of complex phenotypes. First, the majority of GWAS hits are in non-coding region and their biological interpretation is still unclear. Second, accumulating evidence from GWAS suggests the polygenicity of complex traits, i.e. a complex trait is often affected by many variants with small or moderate effects, whereas a large proportion of risk variants with small effects remain unknown.

Results: The availability of functional annotation data enables us to address the above challenges. In this study, we propose a latent sparse mixed model (LSMM) to integrate functional annotations with GWAS data. Not only does it increase the statistical power of identifying risk variants, but also offers more biological insights by detecting relevant functional annotations. To allow LSMM scalable to millions of variants and hundreds of functional annotations, we developed an efficient variational expectation-maximization algorithm for model parameter estimation and statistical inference. We first conducted comprehensive simulation studies to evaluate the performance of LSMM. Then we applied it to analyze 30 GWAS of complex phenotypes integrated with nine genic category annotations and 127 cell-type specific functional annotations from the Roadmap project. The results demonstrate that our method possesses more statistical power than conventional methods, and can help researchers achieve deeper understanding of genetic architecture of these complex phenotypes.

Availability and implementation: The LSMM software is available at https://github.com/mingjingsi/LSMM.

Contact: macyang@ust.hk or jin.liu@duke-nus.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Since the success of the first GWAS on age-related macular degeneration (Klein *et al.*, 2005), >40 000 single-nucleotide polymorphisms (SNPs) have been reported in about 3300 GWAS at the genome-wide significance level (see GWAS Catalog http://www.ebi.ac.uk/gwas/) (Welter *et al.*, 2014). Despite these fruitful discoveries, the emerging evidence from GWAS presents great challenges towards deeper understanding of the genetic architectures of complex phenotypes. First, >85% genome-wide significant hits are located in the non-coding region (Welter *et al.*, 2014) and most of their functional roles are still elusive. Second, complex phenotypes are often highly polygenic, i.e. they are affected by a vast number of risk variants with individually small effects. For example, it is widely accepted that 70–80% of the variation in human height can be attributed to genetics (Visscher *et al.*, 2008). However, Wood *et al.* (2014) collected >250 000 samples and identified 697 variants at genome-wide significance level, and all these variants together can only explain 20% of heritability. A recent estimate (Boyle *et al.*, 2017) suggests that about 100 000 variants may be associated with human height. Given current sample sizes, a large proportion of risk variants underlying complex phenotypes remain unknown yet.

Fortunately, an increasing number of reports suggest that the functional importance of SNPs may not be equal (Schork *et al.*, 2013), which provides a direction to address the above challenges. On the one hand, SNPs in or near genic regions can explain more heritability of complex phenotypes (Smith *et al.*, 2011; Yang *et al.*, 2011). For example, the partition of genic category annotations for SNPs has revealed that SNPs in 5′ UTR, exon and 3′ UTR are significantly enriched across diverse complex traits (Schork *et al.*, 2013). On the other hand, cell-type specific functional annotations can provide information that is complementary to genic category annotations, for dissecting genetic contribution to complex diseases in a cell-type specific manner. To name a few, genetic variants related to functions of immune cells are significantly enriched for immune diseases, such as rheumatoid arthritis, coeliac disease and type 1 diabetes; variants with liver functions are enriched for metabolic traits, such as LDL, HDL and total cholesterol; variants with pancreatic islet functions are enriched for fasting glucose (Kundaje *et al.*, 2015). Additionally, SNPs in genes that are preferentially expressed in the central nervous system are significantly enriched in psychiatric disorders (e.g. schizophrenia and bipolar disorder) (Chung *et al.*, 2014).

A large amount of functional annotation data has become publicly available and the volume is still expanding. The Encyclopedia of DNA Elements (ENCODE) project (The ENCODE Project Consortium, 2012) has conducted >1650 experiments on 147 cell lines to delineate functional elements across the human genome, such as DNase I hypersensitive sites and transcription factor binding. The NIH Roadmap Epigenomics Mapping Consortium (Kundaje *et al.*, 2015) is generating high-quality genome-wide human epigenomic maps of histone modifications, chromatin accessibility, DNA methylation and mRNA expression across more than one hundred of human cell types and tissues.

With the availability of rich functional annotations, we aim to (i) integrate genic category annotations and cell-type specific functional annotations with GWAS to increase the statistical power of identifying risk SNPs, and (ii) detect relevant cell-type specific functional annotations among a large amount of available annotation data to have a more biologically insightful interpretation of GWAS results. Statistical methods to incorporate functional annotations with GWAS can be roughly divided into two categories: methods based on individual-level genotype data, e.g. bfGWAS (Yang *et al.*, 2017) and FST (He *et al.*, 2017), and methods based on summary statistics. In practice, the access to the individual-level data of a large sample size is often very difficult. Therefore, methods in the second category play the major role and also have computational advantages. However, existing methods based on summary statistics, e.g. stratified FDR methods (Schork *et al.*, 2013), cmfdr (Zablocki *et al.*, 2014), GPA (Chung *et al.*, 2014), GenoWAP (Lu *et al.*, 2016) and EPS (Liu *et al.*, 2016), were designed to handle only a few number of functional annotations and can not be scalable to a large-scale integrative analysis.

In this study, we propose a latent sparse mixed model (LSMM) to integrate genic category annotations and cell-type specific functional annotations with GWAS data. The 'latent' statuses are used to connect the observed summary statistics from GWAS with functional annotations. 'Mixed' models are designed to simultaneously consider both genic category and cell-type specific annotations, where genic category annotations are put into the design matrix of fixed effects, and cell-type specific annotations are encoded in the design matrix of random effects. We further impose a 'sparse' structure on the random effects to adaptively select relevant cell-type specific annotations. We first conducted comprehensive simulations to investigate the properties of LSMM and then applied LSMM to real data. We integrated summary statistics from 30 GWAS with nine genic category annotations and 127 cell-type specific functional annotations from the Roadmap project. Compared with existing methods, our method is able to increase the statistical power in the identification of risk variants and detection of cell-type specific functional annotations, and thus provides a deeper understanding of genetic architecture of complex phenotypes.

2 Latent sparse mixed model

2.1 Model

Suppose we have the summary statistics (P-values) of M SNPs from GWAS. Consider the two-groups model (Efron, 2008), i.e. SNPs either belong to null or non-null group. Let γ_j be the latent variable indicating the membership of the j-th SNP, i.e. $\gamma_j = 0$ or $\gamma_j = 1$ indicates the j-th SNP from null or non-null group, respectively. The proportion of null and non-null group are denoted as π_0 and π_1, respectively. Then we model the observed P-values as (Chung *et al.*, 2014),

$$p_j \sim \begin{cases} U[0,1], & \gamma_j = 0, \\ Beta(\alpha, 1), & \gamma_j = 1, \end{cases} \tag{1}$$

where $U[0,1]$ denotes the uniform distribution on [0, 1] and $Beta(\alpha, 1)$ is the beta distribution with parameter $(\alpha, 1)$. We constrain $0 < \alpha < 1$ to model the fact that P-values from the non-null group tend to be closer to 0 rather than 1.

Suppose that we have collected not only the P-values of M SNPs from GWAS, but also functional annotations of these SNPs. To incorporate information from functional annotations for prioritization of risk variants and detection of cell-type specific functions for a complex phenotype, we consider the following latent sparse mixed model:

$$\log \frac{\Pr\left(\gamma_j = 1 | \mathbf{Z}_j, \mathbf{A}_j\right)}{\Pr\left(\gamma_j = 0 | \mathbf{Z}_j, \mathbf{A}_j\right)} = \mathbf{Z}_j \mathbf{b} + \mathbf{A}_j \boldsymbol{\beta}, \tag{2}$$

where $\mathbf{Z} \in \mathbb{R}^{M \times (L+1)}$ is the design matrix for fixed effects, comprised of an intercept and L covariates, $\mathbf{b} \in \mathbb{R}^{L+1}$ is the vector of fixed effects, $\mathbf{A} \in \mathbb{R}^{M \times K}$ is the design matrix for random effects, $\boldsymbol{\beta} \in \mathbb{R}^{K}$ is the vector of random effects, and K is the number of random effects. Both the j-th row of $\mathbf{Z}$ (i.e. $\mathbf{Z}_j$) and $\mathbf{A}$ (i.e. $\mathbf{A}_j$) corresponds to the j-th

SNP. Note that γ_j is a latent variable in model (2) but its corresponding p_j is observed. This makes our model different from the standard generalized linear mixed model.

Now we partition functional annotations into two categories: genic category annotations and cell-type specific annotations. According to (Schork *et al.*, 2013), genomic regions, such as exon, intron, 5′UTR and 3′UTR, are considered as genic category annotations. For cell-type specific annotations, we used epigenetic markers (H3k4me1, H3k4me3, H3k36me3, H3k27me3, H3k9me3, H3k27ac, H3k9ac and DNase I Hypersensitivity) of multiple tissues from the Roadmap project. As we are more interested in the detection of cell-type specific results, we put genic category annotation data into **Z** and cell-type specific annotation data into **A**, where each column of **Z** corresponds to a genic functional category and each column of **A** corresponds to a cell-type specific functional category. In the simplest case, the entries in **Z** and **A** are binary. For example, $Z_{jl} = 1$ means that the j-th SNP has a function in the l-th genic category and $Z_{jl} = 0$ otherwise. Our model also allows the entries in **Z** and **A** to be continuous variables, e.g. a score Z_{jl} between 0 and 1 can be used to indicate the degree that the j-the SNP has a function in the l-th category. The closer to 1, the more likely it has a functional role. The entries in **A** are defined in the same way as those of **Z**.

To adaptively select cell-type specific annotations, we assign a spike-slab prior on β_k:

$$\beta_k \sim \begin{cases} N\left(\beta_k|0,\sigma^2\right), & \eta_k = 1, \\ \delta_0(\beta_k), & \eta_k = 0, \end{cases} \tag{3}$$

where $N\left(\beta_k|0,\sigma^2\right)$ denotes the Gaussian distribution with mean 0 and variance σ^2, δ_0 denotes the Dirac delta function at zero, $\eta_k = 1$ or $\eta_k = 0$ means the k-th annotation is relevant or irrelevant to the given phenotype, respectively. Here, η_k is a Bernoulli variable with probability ω being 1:

$$\eta_k \sim \omega^{\eta_k}(1-\omega)^{1-\eta_k}, \tag{4}$$

where ω can be interpreted as the proportion of relevant annotations corresponding to this phenotype.

Let $\boldsymbol{\theta} = \left\{\alpha, \mathbf{b}, \sigma^2, \omega\right\}$ be the collection of model parameters. The logarithm of the marginal likelihood can be written as

$$\log \Pr(\mathbf{p}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\theta}) = \log \sum_{\gamma} \sum_{\boldsymbol{\eta}} \int \Pr(\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\theta}) d\boldsymbol{\beta}, \tag{5}$$

where

$$\Pr(\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\theta}) = \Pr(\mathbf{p}|\boldsymbol{\gamma}; \alpha) \Pr(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{A}, \boldsymbol{\beta}; \mathbf{b}) \Pr\left(\boldsymbol{\beta}|\boldsymbol{\eta}; \sigma^2\right) \Pr(\boldsymbol{\eta}|\omega). \tag{6}$$

Our goal is to maximize the marginal likelihood to obtain the estimation $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and compute the posterior

$$\Pr\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{p}, \mathbf{Z}, \mathbf{A}; \widehat{\boldsymbol{\theta}}\right) = \frac{\Pr\left(\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \widehat{\boldsymbol{\theta}}\right)}{\Pr\left(\mathbf{p}|\mathbf{Z}, \mathbf{A}; \widehat{\boldsymbol{\theta}}\right)}. \tag{7}$$

Then we can infer the risk SNPs and relevant cell-type specific functional annotations for this phenotype and calculate the false discovery rate.

2.2 Algorithm

Exact evaluation of posterior (7) is intractable. One difficulty is due to the sigmoid function resulting from the logistic model. The other comes from the spike-slab prior. To address this issue, we propose a variational expectation-maximization (EM) algorithm for parameter estimation and posterior approximation.

Before starting the derivation of our algorithm, we first reparametrize the spike-slab prior (3) by introducing a new Gaussian variable $\tilde{\beta}_k \sim N\left(0,\sigma^2\right)$, then the product $\eta_k\tilde{\beta}_k$ has the same distribution with β_k in model (3). So model (2) can be written as

$$\log \frac{\Pr\left(\gamma_j = 1|\mathbf{Z}_j, \mathbf{A}_j\right)}{\Pr\left(\gamma_j = 0|\mathbf{Z}_j, \mathbf{A}_j\right)} = \mathbf{Z}_j\mathbf{b} + \sum_{k=1}^{K} A_{jk}\beta_k = \mathbf{Z}_j\mathbf{b} + \sum_{k=1}^{K} A_{jk}\eta_k\tilde{\beta}_k. \tag{8}$$

Hence the complete-data likelihood $\Pr(\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\theta})$ can be rewritten as:

$$\Pr\left(\mathbf{p}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\theta}\right) = \Pr(\mathbf{p}|\boldsymbol{\gamma}; \alpha) \Pr\left(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{A}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}; \mathbf{b}\right) \Pr\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\sigma^2, \omega\right), \tag{9}$$

where

$$\Pr(\mathbf{p}|\boldsymbol{\gamma}; \boldsymbol{\alpha}) = \prod_{j=1}^{M} \Pr\left(p_j|\gamma_j; \alpha\right) = \prod_{j=1}^{M} \left(\alpha p_j^{\alpha-1}\right)^{\gamma_j}, \tag{10}$$

$$\begin{aligned} \Pr\left(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{A}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}; \mathbf{b}\right) &= \prod_{j=1}^{M} \Pr\left(\gamma_j|\mathbf{Z}_j, \mathbf{A}_j, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}; \mathbf{b}\right) \\ &= \prod_{j=1}^{M} e^{\gamma_j\left(\mathbf{Z}_j\mathbf{b} + \sum_k A_{jk}\eta_k\tilde{\beta}_k\right)} S\left(-\mathbf{Z}_j\mathbf{b} - \sum_{k=1}^{K} A_{jk}\eta_k\tilde{\beta}_k\right), \end{aligned} \tag{11}$$

$$\Pr\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\sigma^2, \omega\right) = \Pr\left(\tilde{\boldsymbol{\beta}}|\sigma^2\right) \Pr(\boldsymbol{\eta}|\omega) = \prod_{k=1}^{K} N\left(\tilde{\beta}_k|0, \sigma^2\right) \omega^{\eta_k}(1-\omega)^{1-\eta_k}, \tag{12}$$

where $S(\cdot)$ is the sigmoid function and $S(x) = (1+e^{-x})^{-1}$. With this reparameterization, we get rid of the Dirac delta function.

Due to the intractability caused by the sigmoid function inside integration (5), we consider the JJ bound (Jaakkola and Jordan, 2000):

$$S(x) \geq S(\xi) \exp\left\{(x-\xi)/2 - \lambda(\xi)\left(x^2 - \xi^2\right)\right\}, \tag{13}$$

where $\lambda(\xi) = \frac{1}{2\xi}\left[S(\xi) - \frac{1}{2}\right]$ and the right-hand-side of the inequality (13) is the JJ bound. Clearly, the JJ bound is in the exponential of a quadratic form. Applying this bound to (11), we can get a tractable lower bound of $\Pr\left(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{A}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}; \mathbf{b}\right)$, denoted as $h\left(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{A}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}; \mathbf{b}, \boldsymbol{\xi}\right)$, where $\boldsymbol{\xi} \in \mathbb{R}^M$ is variational parameter. Let $\boldsymbol{\Theta} = \left\{\alpha, \mathbf{b}, \boldsymbol{\xi}, \sigma^2, \omega\right\}$. The lower bound of the complete-data likelihood is defined as

$$f\left(\mathbf{p}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\Theta}\right) = \Pr(\mathbf{p}|\boldsymbol{\gamma}; \alpha) h\left(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{A}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}; \mathbf{b}, \boldsymbol{\xi}\right) \Pr\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\sigma^2, \omega\right). \tag{14}$$

Next, we derive the variational EM algorithm. Let $q\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}\right)$ be an approximation of the posterior $\Pr\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\mathbf{p}, \mathbf{Z}, \mathbf{A}; \boldsymbol{\theta}\right)$. We can obtain a lower bound of the logarithm of the marginal likelihood

$$\begin{aligned} &\log \Pr(\mathbf{p}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\theta}) \\ &= \log \textstyle\sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int \Pr\left(\mathbf{p}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\theta}\right) d\tilde{\boldsymbol{\beta}} \\ &\geq \log \textstyle\sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int f\left(\mathbf{p}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\Theta}\right) d\tilde{\boldsymbol{\beta}} \\ &\geq \textstyle\sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int q\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}\right) \log \dfrac{f\left(\mathbf{p}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\Theta}\right)}{q\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}\right)} d\tilde{\boldsymbol{\beta}} \\ &= \mathbf{E}_q\left[\log f\left(\mathbf{p}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}|\mathbf{Z}, \mathbf{A}; \boldsymbol{\Theta}\right) - \log q\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}\right)\right] \triangleq L(q), \end{aligned} \tag{15}$$

where $L(q)$ is the lower bound. The first inequality is based on the JJ bound. The second inequality follows Jensen's inequality. To make it feasible to evaluate the lower bound, we use the mean-field theory and assume that $q(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta})$ can be factorized as

$$q(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}) = \left(\prod_{k=1}^{K} q\left(\tilde{\beta}_k, \eta_k\right)\right)\left(\prod_{j=1}^{M} q\left(\gamma_j\right)\right), \tag{16}$$

where $q\left(\tilde{\beta}_k, \eta_k\right) = q\left(\tilde{\beta}_k | \eta_k\right) q(\eta_k)$. It turns out that $q(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta})$ can be obtained analytically and thus the lower bound $L(q)$ can be exactly evaluated. By setting the derivative of $L(q)$ with respect to the parameters in $\boldsymbol{\Theta}$ be zero, we can obtain the updating equations for parameter estimation. The detailed derivation of the algorithm can be found in Supplementary Section S1.

It is worth noting that LSMM covers two special cases: (i) Two-groups model only (denoted as TGM) when all the coefficients in $\mathbf{b}$ (except the intercept term) and $\boldsymbol{\beta}$ are zero; (ii) Two-groups model plus fixed effects model only (denoted as LFM for the abbreviation of latent fixed effect model) when all coefficients in $\boldsymbol{\beta}$ are zero. This motivates us developing a four-stage algorithm based on warm starts. More specifically, in the first stage, we run an EM algorithm to obtain the two parameters (α and the proportion of non-null group π_1) in the TGM. Then we use the estimated parameters as the starting point to run the second stage variational EM algorithm to fit the LFM and obtain the parameter α, $\mathbf{b}$ and the posterior probability of $\boldsymbol{\gamma}$. In the third stage, we treat the obtained posterior as the value of $\boldsymbol{\gamma}$ and fit the logistic sparse mixed model to obtain the required initial value for the parameters in the next stage. Finally, in the fourth stage, we run the above variational EM algorithm with the obtained parameters at the second and third stage until convergence. Since all the iterations are built upon the framework of EM algorithm, the lower bound is guaranteed to increase at each iteration. The details of the algorithm design are provided in Supplementary Section S2.

2.3 Identification of risk SNPs and detection of relevant cell-type specific functional annotations

After the convergence of the variational EM algorithm, the approximated posterior of latent variables $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ can be obtained. Using this information, we are able to prioritize risk SNPs and relevant cell-type specific functional annotations.

Risk SNPs are identified based on $q\left(\gamma_j = 1\right)$, an approximation of the posterior probability that the j-th SNP is associated with this phenotype. Accordingly, we can calculate the approximated local false discovery rate $fdr_j = 1 - q\left(\gamma_j = 1\right)$. To control the global false discovery rate (FDR), we sort SNPs by fdr from the smallest to the largest and regard the j-th re-ordered SNP as a risk SNP if

$$FDR_{(j)} = \frac{\sum_{i=1}^{j} fdr_{(i)}}{j} \leq \tau, \tag{17}$$

where $fdr_{(i)}$ is the i-th ordered fdr, $FDR_{(j)}$ is the corresponding global FDR, and τ is the threshold of global FDR. In simulations, we chose $\tau = 0.1$.

Relevant cell-type specific functional annotations are inferred from $q(\eta_k = 1)$, an approximation of the posterior probability that annotation k is relevant to this phenotype. Similarly, we can calculate the approximated local false discovery rate $fdr_k = 1 - q$ $(\eta_k = 1)$ and convert it into the global false discovery rate. We can either control the local false discovery rate (e.g. $fdr_k \leq 0.1$) or global false discovery rate with $\tau = 0.1$.

3 Results

3.1 Simulation

We conducted simulations to evaluate the performance of the proposed LSMM. The simulation data was generated as follows. The numbers of SNPs, fixed effects (genic category annotations) and random effects (cell-type specific functional annotations) were set to be $M = 100\,000$, $L = 10$ and $K = 500$ respectively. The entries in design matrices Z_{jl} and A_{jk} were generated from $Bernoulli(0.1)$, $j = 1, \ldots, M$, $l = 1, \ldots, L$ and $k = 1, \ldots, K$. Given the proportion of relevant cell-type specific functional annotations ω, η_k was drawn from $Bern oulli(\omega)$ and the corresponding nonzero entries of random effects $\boldsymbol{\beta}$ were simulated from $N(0, 1)$. The first entry of the coefficients of fixed effects $\mathbf{b}$, i.e. the intercept in the logistic model, was fixed at –2 and other entries were generated from $N(0, 1)$ and then kept fixed in multiple replications. After that, we simulated γ_j from Bernoulli distribution with probability $S(\mathbf{Z}_j\mathbf{b} + \mathbf{A}_j\boldsymbol{\beta})$, and then generated p_j from $U[0, 1]$ if $\gamma_j = 0$ and $Beta(\alpha, 1)$ otherwise.

We first evaluated the performance of LSMM in the identification of risk SNPs. We compared LSMM with two special cases, LFM (with fixed effects only) and TGM (without fixed effects and random effects). After prioritizing the risk SNPs using these methods, we made a comparison upon their empirical FDR, power, area under the receiver operating characteristic curve (AUC) and partial AUC. We varied the proportion of relevant random effects ω at $\{0, 0.01, 0.05, 0.1, 0.2\}$. Figure 1 shows the performance of these three models with $\alpha = 0.2$ and $K = 500$ (results for other scenarios are shown in Supplementary Figs S1–S8). As shown in Figure 1, the empirical FDRs are indeed controlled at the nominal level ($\tau = 0.1$) for all these models. For TGM and LFM, the powers increase as the proportion of relevant functional annotations ω increases. This is because a larger ω could result in an increasing proportion of non-null group for SNPs. However, the AUC and partial AUC of LFM slightly decrease because the estimates of fixed effects using LFM would become less accurate when the impact of functional annotations becomes larger. LSMM can adaptively select relevant functional annotations to improve its performance. As expected, it outperforms both TGM and LFM in terms of the power, AUC and partial AUC. One may wonder what if we do not do variable selection and simply treat the effects of all covariates as fixed effects. We evaluated this approach and found that, without variable selection, the FDR would be inflated when the GWAS signal is relatively weak (see Supplementary Fig. S9).

Next, we evaluated the performance of LSMM in the detection of relevant cell-type specific functional annotations in terms of the FDR, power, AUC and partial AUC. We varied the proportion of relevant cell-type specific functional annotations ω at $\{0.01, 0.05, 0.1, 0.2\}$. The results are given in Figure 2 with $\alpha = 0.2$ and $K = 500$ (results for other scenarios are shown in Supplementary Figs S10–S17). The empirical FDR is controlled at 0.1 with conservativeness. This is because the variational approach is adopted to approximate the posterior, e.g. the JJ bound and mean-field approximation. When the signal of the GWAS data is relatively strong, i.e. α is relatively small, LSMM has a very good performance of detecting relevant functional annotations, as indicated by power, AUC and partial AUC. When the number of SNPs becomes larger (e.g. $M = 500\,000$), for a fixed signal strength, the empirical FDR becomes less conservative and the performance becomes better (see Supplementary Figs S18–S20). Simulations (details are given in Supplementary Section S3.5) also show that when the annotations integrated are of greater importance, the performance of LSMM becomes better, especially when the number of

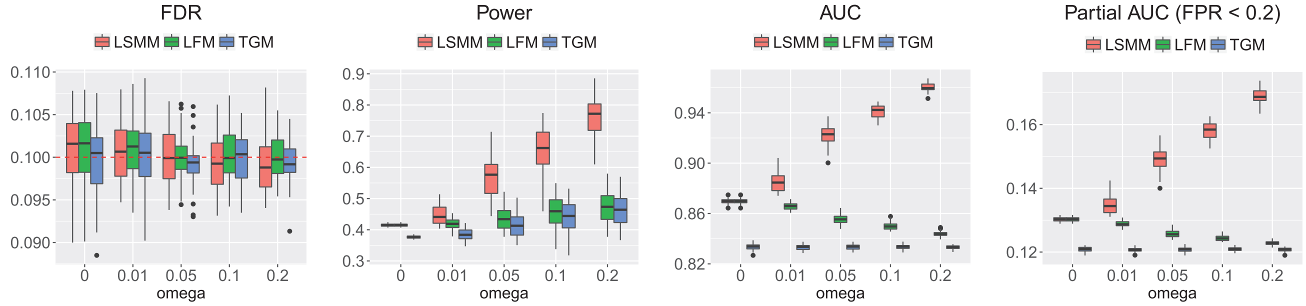

Fig. 1. FDR, power, AUC and partial AUC of LSMM, LFM and TGM for identification of risk SNPs with $\alpha = 0.2$ and $K = 500$. We controlled global FDR at 0.1 to evaluate empirical FDR and power. The results are summarized from 50 replications

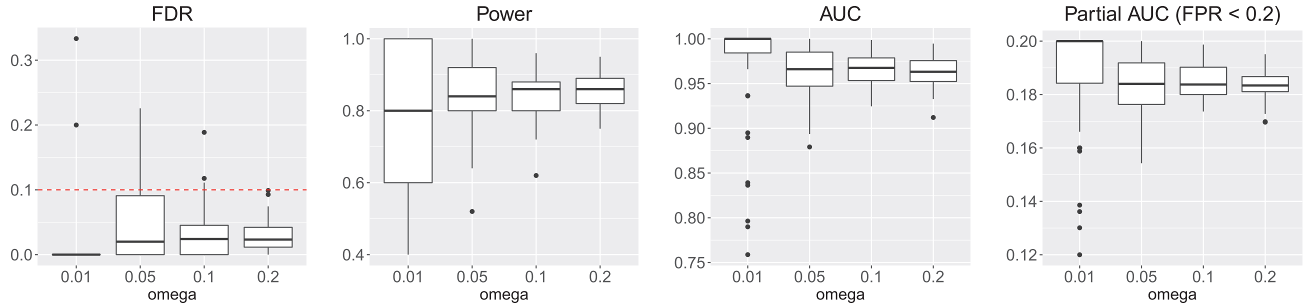

Fig. 2. FDR, power, AUC and partial AUC of LSMM for detection of relevant annotations with $\alpha = 0.2$ and $K = 500$. We controlled global FDR at 0.1 to evaluate empirical FDR and power. The results are summarized from 50 replications

SNPs is small (see Supplementary Figs S21–S22). Therefore, the performance of LSMM in the detection of relevant annotations is influenced by the signal strength of the GWAS data, the number of SNPs and the importance of annotations. We also conducted the following simulations to examine the role of adjusting covariates (i.e. genic category annotations) using fixed effects for detecting relevant cell-type specific annotations. We consider the case that genic category annotations and some cell-type specific annotations are correlated and **b**, the vector of coefficients corresponding to genic category annotations, is nonzero. Without adjusting genic category annotations, some irrelevant cell-type specific annotations will be falsely included in the model due to their correlation with genic category annotations. To verify this, we simulated a case that 10 genic category annotations and first 50 cell-type specific annotations are correlated with correlation coefficient varied at $\{0, 0.2, 0.4, 0.6, 0.8\}$ and the remaining annotations are generated independently. The simulation details are given in Supplementary Section S3.6. In the presence of correlation, as expected, a larger FDR of detecting relevant cell-type specific annotations is observed without adjusting genic category annotations (see Supplementary Fig. S23).

LSMM is not sensitive to initial parameter specification except when the true proportion of risk SNPs is extremely small and the signal of GWAS data is very weak (details of simulation and results are given in Supplementary Section S3.7 and Supplementary Fig. S24). Regarding parameter estimation, LSMM provides a satisfactory estimate of α, the parameter in Beta distribution (see Supplementary Figs S25–S27). When the signal strength of GWAS data is not very weak, the estimated fixed effects **b** (Supplementary Figs S28–S38) and the proportion of non-zero random effects ω (Supplementary Fig. S39) are relatively accurate. When we generate *P*-values for SNPs from individual-level data instead of the generative model (1), we note that the value of α is determined by both heritability and sample size of individual-level data (details of simulation and results are given in Supplementary Section S3.9 and Supplementary Fig. S40). Given fixed sample sizes and nonzero proportion, smaller α corresponds to larger heritability.

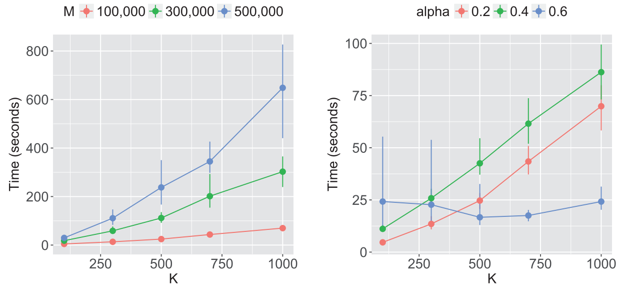

Fig. 3. Computational time of LSMM. Left: We varied the number of SNPs *M* and the number of random effects *K*, with $\alpha = 0.2$. Right: We varied the number of random effects *K* and the strength of GWAS signal α with $M = 100\ 000$. The results are summarized from 10 replications

The computational time of LSMM depends on the strength of GWAS signal, the number of SNPs and the number of random effects. Figure 3 (left) shows that the computational time is nearly linear with respect to *M* and *K* with $\alpha = 0.2$. In the right panel, we fixed $M = 100\ 000$ and varied *K* and α. When the GWAS signal is relatively weak, e.g. $\alpha = 0.6$, the timings of LSMM remain the same for different scales of random effects. This is because LSMM adopts a warm-start strategy and its last two stages start from the estimates at the second stage (i.e. fixed effects only) and converge in a few iterations when the GWAS signal is too weak to provide information for updating the random effects.

Then we conducted simulations to test the robustness of LSMM under the situation that some assumptions in our proposed model are violated.

- LSMM assumes that the P-values of non-null SNPs follow the Beta distribution. Here we simulated the underlying distribution of P-values in non-null group from other distributions. The experimental results (Supplementary Figs S41–S43) indicate that the FDR of LSMM is still well controlled at the nominal level.
- We assume independence among SNPs, which greatly facilitates the computation and inference of LSMM. We conducted simulation (details are given in Suplemenatry Section S3.11) to evaluate the impact of this assumption on LSMM. Because GWAS only aim to identify the local genomic region in LD with true risk genetic variants, it is reasonable to consider the identified SNPs not as false positives if they are in the flanking region of the true risk SNPs. In this sense, the results (Supplementary Fig. S44) suggest that LSMM can provide a satisfactory FDR control.
- We assume the proportion of risk variants is not very small due to the polygenic effect in the context of GWAS. In the simulation, we used the TGM to generate data such that we can evaluate whether the estimates converge to their true values. We varied the true value of the proportion of risk SNPs $\pi_1 \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2\}$. We also used Higher Criticism to estimate the proportion of non-null effects as a comparison. The results (Supplementary Fig. S45) show that when the true proportion of risk SNPs is extremely small (e.g. $\pi_1 \leq 0.001$ for $\alpha = 0.4$) and the signal of GWAS data is weak (e.g. $\pi_1 \leq 0.01$ for $\alpha = 0.6$), the estimation using LSMM is not very accurate. However, LSMM can still provide a valid FDR control. The performance of Higher Criticism is quite opposite. Although it can provide stable estimation when the true proportion of risk SNPs is small ($\pi_1 \leq 0.01$), its performance is not as well as LSMM when π_1 is relatively large, e.g. $\pi_1 \geq 0.05$.
- In LSMM, we use generative model (2) to integrate functional annotations. Here we conducted simulations based on probit model:

$$y_j = \mathbf{Z}_j\mathbf{b} + \mathbf{A}_j\boldsymbol{\beta} + e_j, \tag{18}$$

  where $e_j \sim N(0, \sigma_e^2)$. And we set $\gamma_j = 1$ if $y_j > 0$, $\gamma_j = 0$ if $y_j \leq 0$. The details of the simulation are given in Suplemenatry Section S3.13. The results are provided in Suplemenatry Figures S46–S51. Noting that FDRs are all well-controlled at the nominal level. LSMM shows the best performance in power, AUC and partial AUC in identification of risk SNPs, and the advantages of LSMM over LFM and TGM are more noticeable as the signal-noise ratio increases.

- The correlation among random effects is ignored in LSMM. However, simulation results on correlated random effects (Supplementary Figs S52–S53) indicate the robustness of LSMM.
- LSMM assumes a common variance, σ^2, for random effects. When this assumption is violated, e.g. the variance for each random effect is from $U[1, 10]$, the performance of LSMM (Supplementary Figs S54–S55) is still comparable to the ideal case shown in Figures 1 and 2.

In summary, the above simulation results suggest the robustness of LSMM and its potentially wide usage.

We compared LSMM with GPA in the identification of risk variants and detection of cell-type specific annotations. As LSMM can integrate both genic category and functional annotations, we compared GPA with LSMM without fixed effects (integrate functional annotations only) for a fair comparison. From the model setup, one main difference between GPA and LSMM is that GPA assumes conditional independence among annotations, whereas in LSMM we do not make this assumption. To check the influence of correlated functional annotations, we simulated a case that the first 10 functional annotations were correlated and all the others were independent. We varied the correlation among annotations at $\{0, 0.2, 0.4, 0.6, 0.8\}$. The results are shown in Supplementary Figures S56–S59. We observe that the empirical FDRs of LSMM and LSMM without fixed effects are indeed controlled at 0.1, but the FDR of GPA inflates very much when annotations are correlated. As the FDR of GPA is not controlled, the power of GPA is not comparable to the other two models. According to the AUC and partial AUC, the performance of GPA becomes worse as the correlation among annotations increase, while the performance of LSMM is still stable and outstanding. It implies that LSMM is able to identify true relevant annotations among correlated misleading ones.

We also conducted simulations to compare LSMM with cmfdr, a fully Bayesian approach to incorporate genic category annotations in GWAS using MCMC sampling algorithm. We find that cmfdr is not able to handle a large number of annotations and the MCMC sampling algorithm is very time-consuming. Besides the computational time, we observe the empirical FDR of cmfdr is slightly inflated and its performance for prioritization of risk variants is inferior to LSMM in terms of AUC and partial AUC (see Supplementary Fig. S60).

As a comparison, we also used GenoWAP, a GWAS signal prioritization method that integrates genomic functional annotation and GWAS test statistics, to prioritize SNPs in our simulation. As GenoWAP can only integrate one annotation at a time, in the simulation we set $L = 1$ and let $\mathbf{Z}$ be the functional annotation integrated using GenoWAP. The performance of LSMM, LFM and GenoWAP for identification of risk SNPs are shown in Suplemenatry Figure S61. We observe that for GenoWAP, the empirical FDRs are very conservative and its power, AUC and pAUC are all very low. That is because the SNPs that GenoWAP detects are disease-specific functional, i.e. only SNPs which are annotated to be functional have a chance be detected. In our simulation, 90% SNPs are not functional in the annotation category and thus are not identified by GenoWAP. To evaluate the influence of the functional proportion on GenoWAP, we conducted the following simulations with no tissue-specific functional annotations. We set $L = 1$ and generated data from LFM. The results for different functional proportions are shown in Supplementary Figure S62. The empirical FDRs of GenoWAP are still very small. Each of the power, AUC and pAUC of GenoWAP shows an increasing trend as the functional proportion increases, indicating that the performance of GenoWAP is influenced by the quality of annotation.

3.2 Real data analysis

We applied LSMM to analyze 30 GWAS of complex phenotypes. The source of the 30 GWAS is given in Supplemenatry Table S2. We used ANNOVAR (Wang *et al.*, 2010) to provide the genic category annotations: upstream, downstream, exonic, intergenic, intronic, ncRNA_exonic, ncRNA_intronic, UTR3 and UTR5, where ncRNA means variant overlaps a transcript without coding annotation in the gene definition. We obtained 127 cell-type specific functional annotations from GenoSkylinePlus (Lu *et al.*, 2017) (http://genocanyon.med.yale.edu/GenoSkyline). To avoid unusually large GWAS signals in the MHC region (Chromosome 6, 25–35 Mb), we excluded the SNPs in this region.

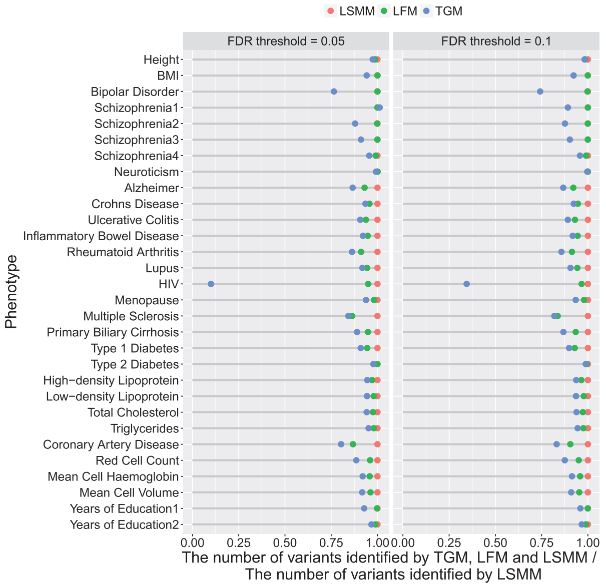

Fig. 4. The number of risk variants identified by TGM, LFM and LSMM for 30 GWAS, under the same level of global FDR control (0.05 and 0.1). For visualization purpose, these numbers are normalized by dividing the corresponding number of variants identified by LSMM

We compared the number of identified risk SNPs using TGM, LFM and LSMM for 30 GWAS. Using LSMM as a reference, we calculated the ratio of the number of risk SNPs each method identified to that from LSMM under FDR thresholds $\tau = 0.05$ and $\tau = 0.1$. The results are shown in Figure 4. For detecting the relevant cell-type specific functional annotations, we controlled the local fdr at 0.1. Figure 5 shows the approximated posterior probability for annotations and phenotypes, where the darkness of the red entry implies the level of relevance between the corresponding cell-type specific functional annotation and the phenotype, the darker the more relevant.

Figure 4 shows that LSMM can identify more risk variants than TGM and LFM, under the same level of FDR control. The differences between TGM and LFM are due to the impact of genic category annotations and the differences between LFM and LSMM can be attributed to cell-type specific functional annotations. For HIV (McLaren *et al.*, 2013) and bipolar disorder (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011), a clear improvement in the identification of risk SNPs can be found from TGM to LFM, reflecting a large enrichment of genic category annotations. The contribution of cell-type specific annotations can be clearly seen with the improvement from LFM to LSMM in several GWAS analyses, such as multiple sclerosis (Sawcer *et al.*, 2011) and coronary artery disease (CAD) (Schunkert *et al.*, 2011). For multiple sclerosis, genic category annotations do not show huge contributions, however, the contributions of cell-type specific annotations are substantial. As shown in Figure 5, its relevant cell-type specific annotations are related with immune system, GM12878 lymphoblastoid cells and primary B cells from peripheral blood. For CAD, both enrichment of genic category and cell-type specific annotations are estimated and its relevant cells are from a few different tissues, including blood, heart, lung and skin (see Fig. 5). As a cardiovascular disease, it is reasonable to discover the relevance of these cells to CAD, and Fernández-Ruiz (2016) has shown its relationship with immune system. The annotations in lung and skin we detected may provide some new insights about the disease.

Among the 30 GWAS, we analyzed four GWAS of schizophrenia with different sample sizes, Schizophrenia1 (9379 cases and 7736 controls) (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013), Schizophrenia2 (9394 cases and 12 462 controls) (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011), Schizophrenia3 (13 833 cases and 18 310 controls) (Ripke *et al.*, 2013) and Schizophrenia4 (36 989 cases and 113 075 controls) (Ripke *et al.*, 2014). The detailed results are summarized in Supplemenatry Table S3. The Manhattan plots using TGM and LSMM are provided in Supplementary Figure S63. Clearly, LSMM steadily improves over TGM and LFM in the analysis of schizophrenia, a highly polygenic trait, with different sample sizes. In particular, for Schizophrenia3, LSMM identified 1492 risk variants which could not be identified by TGM. Interestingly, the majority of them (872 variants) can be re-identified in Schizophrenia4 using TGM. This indicates that LSMM has a better power in prioritizing risk variants than TGM. For Schizophrenia4, four cell-type specific functional annotations are detected. In our analysis, both genetic variants related to functions of brain cells (brain angular gyrus) and blood cells (K562 leukemia cells) are detected to be relevant. This evidence not only connects Schizophrenia with brain, but also suggests the biological link between Schizophrenia and immune system (Ripke *et al.*, 2014). To make a comparison, we also used GenoWAP to analyze Schizophrenia3 and Schizophrenia4 by integrating each of the nine genic category annotations. The results are shown in Supplemenatry Table S4. With the nominal local FDR controlled at 0.1, even we collected the risk SNPs identified by integrating every annotation using GenoWAP, the total number is still much less than TGM, LFM and LSMM, suggesting GenoWAP is too conservative for real data analysis. We also analyzed two GWAS of years of education, Years of Education 1 (Rietveld *et al.*, 2013) and Years of Education 2 (Okbay *et al.*, 2016). Compared with Years of Education 1, the GWAS dataset for Years of Education 2 is based on a larger sample size, and thus it enables LSMM to detect relevant functional annotations in brain and immune system. Our results are consistent with Finucane *et al.* (2015).

More findings about the relevance between cell-type specific annotations and GWAS are shown in Figure 5. Some are concordant with previous GWAS analyses. For example, we detect the functional annotation in liver to be relevant to the lipid-related phenotypes, including low-density lipoprotein, high-density lipoprotein, triglycerides and total cholesterol (Global Lipids Genetics Consortium, 2013). Similar functional enrichment has been found by Finucane *et al.* (2015), Kundaje *et al.* (2015)and Lu *et al.* (2017). For height (Wood *et al.*, 2014), >40 cell-type specific functional annotations are detected to be relevant using LSMM, which reflects its highly polygenic genetic architecture. These relevant annotations include cells in bone, vascular and skeletal muscle which were also shown significant enrichments for height by Finucane *et al.* (2015). Recent research has linked some neurodegenerative diseases, which were believed to be more related to brain and neural system, to the immune system, such as Alzheimer's disease (Sims *et al.*, 2017) and Parkinson's disease (Sulzer *et al.*, 2017). For Alzheimer's disease (Lambert *et al.*, 2013), similar results have been found using LSMM. The relevant functional annotations are from blood cells, including monocytes-CD14+ and K562 leukemia cells. For autoimmune diseases including Crohn's disease (Jostins *et al.*, 2012), ulcerative colitis (Jostins *et al.*, 2012), inflammatory bowel disease (Jostins *et al.*, 2012),

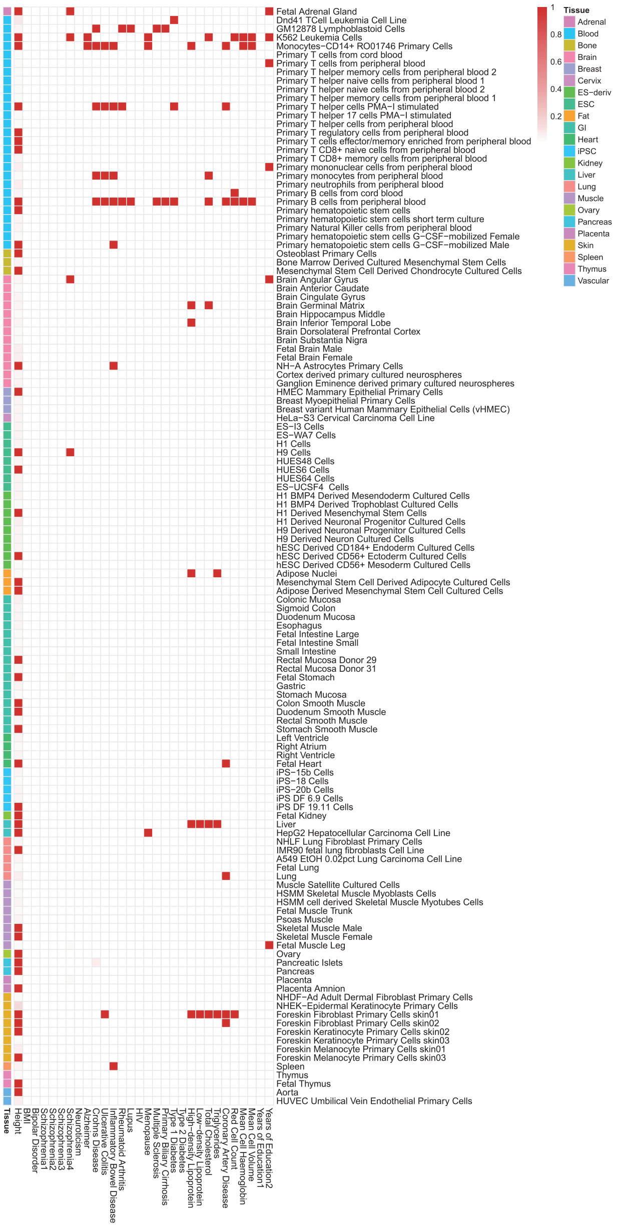

Fig. 5. Relevant cell-type specific functional annotations for 30 GWAS

rheumatoid arthritis (Okada *et al.*, 2014), lupus (Bentham *et al.*, 2015), menopause (Day *et al.*, 2015), multiple sclerosis (Sawcer *et al.*, 2011) and primary biliary cirrhosis (Cordell *et al.*, 2015), the detected relevant functional annotations are mainly from the immune system and have many overlaps. Our results also provide the genomic level supports to previous medical literature, such as the relevance between spleen and inflammatory bowel disease (Muller *et al.*, 1993), between liver and menopause (Mucci *et al.*, 2001). The result also provides several new insights. Lipid-related phenotypes including high-density lipoprotein and total cholesterol are also relevant to functional annotations in immune system and brain. Additionally, annotations in immune system are considered relevant to blood-related phenotypes including red cell count, mean cell haemoglobin and mean cell volume (Pickrell, 2014). The foreskin fibroblast primary cells in skin are relevant to ulcerative colitis, four lipid-related phenotypes and red cell count.

Regarding the computational time, LSMM takes <6 min to handle each of the 30 GWAS datasets. We also recorded timings of cmfdr as a comparison. As cmfdr is not scalable to a large number of covariates, we only integrated the nine genic category annotations in cmfdr. The MCMC algorithm was suggested (Zablocki *et al.*, 2014) to run with 5000 burn-in and 20 000 main iterations. According to our estimates, cmfdr takes >10 days for most phenotypes. The detailed timing results are shown in Supplementary Figure S64.

If we did not adjust the genic category annotation, more relevant cell-type specific functional annotations would be detected (results are shown in Supplementary Fig. S65). This indicates that LSMM could adjust covariates' effects and provide a more reliable identification of relevant functional annotations.

4 Conclusion

We have presented a statistical approach, LSMM, to integrate genic category annotations and a large amount of cell-type specific functional annotations with GWAS data. LSMM can not only improve the statistical power in the identification of risk SNPs, but also infer relevant cell-type specific functional annotations to the phenotype, offering new insights to explore the genetic architecture of complex traits or diseases. Through comprehensive simulations and real data analysis of 30 GWAS, LSMM is shown to be statistically efficient and computationally scalable. As more annotation data will become publicly available in the future, we believe LSMM is widely useful for integrative analysis of genomic data.

Funding

This work was supported in part by the National Science Funding of China [61501389]; the Hong Kong Research Grant Council [22302815, 12316116 and 12301417]; The Hong Kong University of Science and Technology [startup grant R9405]; Innovative Technology Funding of Hong Kong [ITF391/15FX (P0162)]; Duke-NUS Medical School WBS [R-913-200-098-263]; Ministry of Education, Singapore [MOE2016-T2-2-029]; Shenzhen Fundamental Research Fund [KQTD2015033114415450].

Conflict of Interest: none declared.

References

Bentham,J. *et al.* (2015) Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.*, **47**, 1457–1464.

Boyle,E.A. *et al.* (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**, 1177–1186.

Chung,D. *et al.* (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet.*, **10**, e1004787.

Cordell,H.J. *et al.* (2015) International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.*, **6**, 8019.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, **381**, 1371–1379.

Day,F.R. *et al.* (2015) Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.*, **47**, 1294–1303.

Efron,B. (2008) Microarrays, empirical bayes and the two-groups model. *Stat. Sci.*, **23**, 1–22.

Fernández-Ruiz,I. (2016) Immune system and cardiovascular disease. *Nat. Rev. Cardiol.*, **13**, 503.

Finucane,H.K. *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.

Global Lipids Genetics Consortium (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.

He,Z. *et al.* (2017) Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *Am. J. Hum. Genet.*, **101**, 340–352.

Jaakkola,T.S. and Jordan,M.I. (2000) Bayesian parameter estimation via variational methods. *Stat. Comput.*, **10**, 25–37.

Jostins,L. *et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119–124.

Klein,R.J. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Kundaje,A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Lambert,J.C. *et al.* (2013) Meta-analysis of 74, 046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, **45**, 1452–1458.

Liu,J. *et al.* (2016) EPS: an empirical Bayes approach to integrating pleiotropy and tissue-specific information for prioritizing risk genes. *Bioinformatics*, **32**, 1856–1864.

Lu,Q. *et al.* (2016) GenoWAP: gWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics*, **32**, 542–548.

Lu,Q. *et al.* (2017) Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genet.*, **13**, e1006933–e1006924.

McLaren,P.J. *et al.* (2013) Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathog.*, **9**, e1003515–e1003519.

Mucci,L.A. *et al.* (2001) Age at menarche and age at menopause in relation to hepatocellular carcinoma in women. *BJOG*, **108**, 291–294.

Muller,A. *et al.* (1993) Splenic function in inflammatory bowel disease: assessment by differential interference microscopy and splenic ultrasound. *Q J Med.*, **86**, 333–340.

Okada,Y. *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, 376–381.

Okbay,A. *et al.* (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, **533**, 539–542.

Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.*, **43**, 977–983.

Rietveld,C.A. *et al.* (2013) GWAS of 126, 559 individuals identifies genetic variants associated with educational attainment. *Science*, **340**, 1467–1471.

Ripke,S. *et al.* (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.*, **45**, 1150–1159.

Ripke,S. *et al.* (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Sawcer,S. *et al.* (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, **476**, 214–219.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.*, **43**, 969–976.

Schork,A.J. *et al.* (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.*, **9**, e1003449–e1003441.

Schunkert,H. *et al.* (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.*, **43**, 333–338.

Sims,R. *et al.* (2017) Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet.*, **49**, 1373–1384.

Smith,E.N. *et al.* (2011) Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. *PLoS Genet.*, **7**, e1002134.

Sulzer,D. *et al.* (2017) T cells from patients with Parkinson's disease recognize α-synuclein peptides. *Nature*, **546**, 656–661.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Visscher,P.M. *et al.* (2008) Heritability in the genomics era - concepts and misconceptions. *Nat. Rev. Genet.*, **9**, 255–266.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Wood,A.R. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173–1186.

Yang,J. *et al.* (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.*, **43**, 519–525.

Yang,J. *et al.* (2017) A scalable bayesian method for integrating functional information in genome-wide association studies. *Am. J. Hum. Genet.*, **101**, 404–416.

Zablocki,R.W. *et al.* (2014) Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*, **30**, 2098–2104.