Bioinformatics, 2019, 1–9
doi: 10.1093/bioinformatics/btz947
Advance Access Publication Date: 20 December 2019
Original Paper

Genetics and population analysis

LPM: a latent probit model to characterize the relationship among complex traits using summary statistics from multiple GWASs and functional annotations

Jingsi Ming[1], **Tao Wang**[2,3] **and Can Yang**[1,*]

[1]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR, China, [2]Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China and [3]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on July 21, 2019; revised on December 13, 2019; editorial decision on December 16, 2019; accepted on December 18, 2019

Abstract

Motivation: Much effort has been made toward understanding the genetic architecture of complex traits and diseases. In the past decade, fruitful GWAS findings have highlighted the important role of regulatory variants and pervasive pleiotropy. Because of the accumulation of GWAS data on a wide range of phenotypes and high-quality functional annotations in different cell types, it is timely to develop a statistical framework to explore the genetic architecture of human complex traits by integrating rich data resources.

Results: In this study, we propose a unified statistical approach, aiming to characterize relationship among complex traits, and prioritize risk variants by leveraging regulatory information collected in functional annotations. Specifically, we consider a latent probit model (LPM) to integrate summary-level GWAS data and functional annotations. The developed computational framework not only makes LPM scalable to hundreds of annotations and phenotypes but also ensures its statistically guaranteed accuracy. Through comprehensive simulation studies, we evaluated LPM's performance and compared it with related methods. Then, we applied it to analyze 44 GWASs with 9 genic category annotations and 127 cell-type specific functional annotations. The results demonstrate the benefits of LPM and gain insights of genetic architecture of complex traits.

Availability and implementation: The LPM package, all simulation codes and real datasets in this study are available at https://github.com/mingjingsi/LPM.

Contact: macyang@ust.hk

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

1 Introduction

In the past decade, genome-wide association studies (GWASs) have been conducted for hundreds of complex phenotypes, including complex diseases and quantitative traits, resulting in the identification of tens of thousands of single-nucleotide polymorphisms (SNPs) associated with one or more complex traits at the genome-wide significance level (Welter *et al.*, 2014). By exploring these fruitful findings, genetic variants that affect multiple seemingly irrelevant traits have been discovered. This phenomenon is known as 'pleiotropy' (Solovieff *et al.*, 2013). Recently, accumulating studies suggest the pervasiveness of pleiotropy. Pleiotropic effects can be characterized from both local and global perspectives (Yang *et al.*, 2015). On the one hand, localization of pleiotropic risk variants offers more insights on the genetic architecture of human complex traits. For example, a non-synonymous variant (rs13107325) in the zinc transporter SLC39A8 influences both schizophrenia (SCZ) and Parkinson disease (Pickrell *et al.*, 2016); and Ellinghaus *et al.* (2016) identified 187 independent multi-disease loci in an analysis of five chronic inflammatory diseases. On the other hand, genetic correlation between two complex traits has been widely explored in recent studies (Zheng *et al.*, 2017), providing a comprehensive view on disease classification (Wang *et al.*, 2017). Substantial genetic correlations have been revealed among psychiatric disorders, such as the high correlation between SCZ and bipolar disorder (BIP) and moderate correlation between SCZ and major depressive disorder (MDD) (Lee *et al.*, 2013). For autoimmune diseases, primary sclerosing cholangitis and ulcerative colitis (UC), as well as UC and Crohn's disease (CD), are suggested to have relatively high genome-wide genetic correlations (Ji *et al.*, 2017).

The evidence of pervasive pleiotropy not only deepens our understanding of genetic basis underlying complex traits, but also allows the improved statistical power of identification of risk variants by joint analysis of multiple traits. To name a few, joint analysis of SCZ and BIP could significantly improve association mapping power for each of the diseases (Chung *et al.*, 2014). The power to identify associated variants for systolic blood pressure was increased by considering GWASs of other phenotypes, such as low-density lipoprotein (LDL), body mass index and type 1 diabetes (T1D) (Andreassen *et al.*, 2014).

An increasing number of reports suggest that SNPs with important functional implications can explain more heritability of complex traits (Smith *et al.*, 2011; Yang *et al.*, 2011) and the pattern of enrichment in a specific genic annotation category is consistent across diverse phenotypes (Schork *et al.*, 2013). For example, SNPs in 5′UTR, exons and 3′UTR of genes are significantly enriched, SNPs in introns are moderately enriched and intergenic SNPs are not enriched in height, SCZ and tobacco smoking (Schork *et al.*, 2013). It is coincidence with the finding that pleiotropic SNPs are more often exonic and less often intergenic compared with non-pleiotropic SNPs (Sivakumaran *et al.*, 2011). Additionally, some cell-type specific functional annotations are shown to be relevant to complex traits. For example, functional annotations in liver are relevant to lipid-related traits, such as LDL, high-density lipoprotein (HDL) and total cholesterol (TC) (Kundaje *et al.*, 2015; Ming *et al.*, 2018); functional annotations in immune system are relevant to many autoimmune diseases, such as CD, UC and rheumatoid arthritis (RA) (Ming *et al.*, 2018). Large amounts of functional annotation data have been provided by the Encyclopedia of DNA Elements project (ENCODE Project Consortium, 2012) and the NIH Roadmap Epigenomics Mapping Consortium (Kundaje *et al.*, 2015).

With the availability of functional annotation data and summary statistics from GWASs on a wide spectrum of phenotypes, we aim to propose a unified framework which can (i) characterize relationship among complex traits, including identifying pleiotropic associations and estimating correlations among traits, (ii) increase the association mapping power for one or more traits and (iii) investigate the effect of functional annotations. Existing statistical methods based on summary statistics are not able to achieve these aims simultaneously. Methods, such as cross-trait linkage disequilibrium (LD) score regression (Bulik-Sullivan *et al.*, 2015) and GNOVA (Lu *et al.*, 2017a), provide genetic correlation estimation for pair of traits, but are not able to prioritize GWAS results. In contrast, RiVIERA (Li and Kellis, 2016) can prioritize disease-associated variants by joint analysis of summary statistics across multiple traits and epigenomic annotations, but does not measure pleiotropy. Other methods, such as GPA (Chung *et al.*, 2014) and graph-GPA (Chung *et al.*, 2017), can both infer the relationship among traits and identify causal variants. However, statistical and computational challenges arise as the number of traits increases. GPA assumes a four-group model for the case of two GWASs. The number of groups increases exponentially with the number of traits. Graph-GPA is not able to integrate functional annotations, and its implementation is based on a Markov chain Monte Carlo algorithm which is time-consuming. Additionally, the relationship among traits inferred by graph-GPA is hard to interpret in real data analysis, because graph-GPA suggests a graphical model based on a Markov random field which represents a conditional independent structure for genetic relationship among traits and the structure may change when adding or removing some traits.

Here, we propose a latent probit model (LPM) to characterize relationship among complex traits by integrating summary statistics from multiple GWASs and functional annotations. To make LPM scalable to millions of SNPs and hundreds of traits, instead of working with a brute-force algorithm to handle all the data simultaneously, we develop an efficient parameter-expanded expectation–maximization (PX-EM) algorithm for pair-wise analysis and implement a dynamic threading strategy to enhance its parallel property. This pair-wise strategy is guaranteed to give consistent results by our theoretical analysis from the perspective of the composite likelihood approach (Varin *et al.*, 2011). We conducted comprehensive simulations to evaluate the performance of LPM. Then, we analyzed 44 GWASs of complex traits with 9 genic category annotations and 127 cell-type specific functional annotations using LPM. The results demonstrate that our method can not only fulfill the three goals (characterizing relationship, prioritizing SNPs and integrating functional annotations) under a unified framework, but also achieve a better performance comparing with conventional methods.

2 Latent probit model

2.1 Model

Suppose we have the summary statistics (*P*-values) for M SNPs in K GWASs. In this article, we use $j = 1, \ldots, M$ to index SNPs and $k = 1, \ldots, K$ to index GWAS datasets. For each GWAS, we consider the *P*-values following a two-group model (Efron, 2008), i.e. a mixture of null and non-null distributions, and introduce a latent variable η_{jk} to indicate which group the j-th SNP belongs to for the k-th GWAS. Here $\eta_{jk} = 0$ and $\eta_{jk} = 1$ indicate the j-th SNP is un-associated (in the null group) and associated (in the non-null group) with the k-th trait, respectively. We assume the *P*-values in the k-th GWAS to be distributed as

$$P_{jk} \sim \begin{cases} U[0,1], & \eta_{jk} = 0, \\ Beta\ (\alpha_k, 1), & \eta_{jk} = 1, \end{cases}$$

where $U[0,1]$ is the uniform distribution on $[0,1]$ and $Beta(\alpha_k, 1)$ is the beta distribution with the constraint $0 < \alpha_k < 1$. This model is designed to capture the pattern that *P*-values from the non-null group have higher density near zero (Chung *et al.*, 2014).

To adjust the effect of functional annotations and model the relationship of traits, we consider the LPM:

$$\eta_{jk} = \begin{cases} 1, & \text{if } Z_{jk} > 0, \\ 0, & \text{if } Z_{jk} \leq 0, \end{cases} \tag{1}$$

$$\mathbf{Z}_j = \beta \mathbf{X}_j + \epsilon_j, \epsilon_j \sim N(0, \mathbf{R}), \tag{2}$$

where $\mathbf{Z} \in \mathbb{R}^{M \times K}$ is the matrix of latent variables, $\mathbf{X} \in \mathbb{R}^{M \times (D+1)}$ is the design matrix of functional annotations, comprised of an intercept and D annotations, $\beta \in \mathbb{R}^{K \times (D+1)}$ is the matrix of coefficients and ϵ is the part un-captured by functional annotations. For the j-th SNP, $\mathbf{Z}_j^T$, $\mathbf{X}_j^T$ and ϵ_j^T correspond to the j-th row of $\mathbf{Z}$, $\mathbf{X}$ and ϵ, respectively. The association status η is modulated by two parts: functional annotations and residual part ϵ, where $\mathbf{R} \in \mathbb{R}^{K \times K}$ measures correlation of ϵ among K traits after adjusting functional annotations. We assume that known functional categories may not be able to fully characterize the genetic effects on complex traits due to the polygenicity. Therefore, the residual part, whose correlation is modeled in $\mathbf{R}$, is used to capture the remaining polygenic effect. Based on this pair-wise correlation, conditional dependence for interested traits can also be inferred (see details for calculation in Supplementary Section 1).

The probit link is used for modeling the correlation among multiple traits easily. Furthermore, based on the composite likelihood approach, we can analyze the GWASs in a pair-wise manner instead of working with K traits simultaneously. We denote this model as bivariate LPM (bLPM):

$$\tilde{P}_{jk} \sim \begin{cases} U[0,1], & \tilde{\eta}_{jk} = 0, \\ Beta\ (\tilde{\alpha}_k, 1), & \tilde{\eta}_{jk} = 1, \end{cases}$$

$$\tilde{\eta}_{jk} = \begin{cases} 1, & \text{if } \tilde{Z}_{jk} > 0, \\ 0, & \text{if } \tilde{Z}_{jk} \leq 0, \end{cases}$$

$$\tilde{\mathbf{Z}}_j = \tilde{\beta}\mathbf{X}_j + \tilde{\epsilon}_j, \tilde{\epsilon}_j \sim N(0, \tilde{\mathbf{R}}), \tag{3}$$

where $k = 1, 2$, $\tilde{\beta} \in \mathbb{R}^{2\times(D+1)}$ and $\tilde{\mathbf{R}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

We let $\theta = \{\alpha, \beta, \mathbf{R}\}$ and $\tilde{\theta} = \{\tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{R}}\}$ be the collection of parameters in LPM and bLPM, respectively. The logarithm of the marginal likelihood for bLPM can be written as

$$\begin{aligned} \log \Pr(\tilde{\mathbf{P}}|\mathbf{X}; \tilde{\theta}) &= \log \textstyle\sum_{\tilde{\eta}} \int \Pr(\tilde{\mathbf{P}}, \tilde{\eta}, \tilde{\mathbf{Z}}|\mathbf{X}; \tilde{\theta}) d\tilde{\mathbf{Z}} \\ &= \log \textstyle\sum_{\tilde{\eta}} \int \Pr(\tilde{\mathbf{P}}|\tilde{\eta}; \tilde{\alpha}) \Pr(\tilde{\mathbf{Z}}|\mathbf{X}; \tilde{\beta}, \tilde{\mathbf{R}}) d\tilde{\mathbf{Z}}, \end{aligned} \tag{4}$$

where

$$\Pr(\tilde{\mathbf{P}}|\tilde{\eta}; \tilde{\alpha}) = \prod_{j=1}^{M} \prod_{k=1}^{2} (\tilde{\alpha}_k \tilde{P}_{jk}^{\tilde{\alpha}_k - 1})^{\tilde{\eta}_{jk}}, \tag{5}$$

$$\Pr(\tilde{\mathbf{Z}}|\mathbf{X}; \tilde{\beta}, \tilde{\mathbf{R}}) = \prod_{j=1}^{M} N(\tilde{\mathbf{Z}}_j; \tilde{\beta}\mathbf{X}_j, \tilde{\mathbf{R}}). \tag{6}$$

In Equations (5) and (6), we assume conditional independence among M SNPs given annotation matrix $\mathbf{X}$. Our goal is to find $\tilde{\theta}$ which maximizes the marginal likelihood in Equation (4) for each pair of GWASs and then obtain an estimate of θ in LPM which is denoted as $\hat{\theta}$. We can make statistical inferences on the association of SNPs, relationship among traits and annotation enrichment.

2.2 Algorithm

Instead of using the standard EM algorithm, we propose a PX-EM algorithm, which converges much faster (Liu *et al.*, 1998), for parameter estimation and posterior calculation in bLPM.

We expand the parameter in bLPM to $\mathbf{\Theta} = \{\tilde{\alpha}, \gamma, \mathbf{\Sigma}\}$. Accordingly, model (3) is expanded to

$$\tilde{\mathbf{Z}}_j = \gamma\mathbf{X}_j + \tilde{\epsilon}_j, \ \ \tilde{\epsilon}_j \sim N(0, \Sigma),$$

where $\gamma = \mathbf{D}\tilde{\beta}$, $\mathbf{\Sigma} = \mathbf{D}\tilde{\mathbf{R}}\mathbf{D} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ and $\mathbf{D} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$ is the auxiliary parameter whose value is fixed at the identity matrix in the original model. For the expanded model, the complete-data log-likelihood can be written as

$$\log \Pr(\tilde{\mathbf{P}}, \tilde{\eta}, \tilde{\mathbf{Z}}|\mathbf{X}; \Theta) = \log \Pr(\tilde{\mathbf{P}}|\tilde{\eta}; \tilde{\alpha}) + \log \Pr(\tilde{\mathbf{Z}}|\mathbf{X}; \gamma, \Sigma),$$

where $\Pr(\tilde{\mathbf{Z}}|\mathbf{X}; \gamma, \mathbf{\Sigma}) = \prod_{j=1}^{M} \mathrm{N}(\tilde{\mathrm{Z}}_{\mathrm{j}}; \gamma\mathbf{X}_{\mathrm{j}}, \mathbf{\Sigma})$.

In the PX-E step, the Q function is evaluated as $Q = E_{\tilde{\eta},\tilde{\mathbf{Z}}} \log \Pr(\tilde{\mathbf{P}}, \tilde{\eta}, \tilde{\mathbf{Z}}|\mathbf{X}; \mathbf{\Theta})$, where the expectation is calculated based on the current $\mathbf{\Theta}$ in the original model. In the PX-M step, we maximize the Q function with respect to $\mathbf{\Theta}$ and obtain the updating equations of $\mathbf{\Theta}^{(t)} = \{\tilde{\alpha}^{(t)}, \gamma^{(t)}, \mathbf{\Sigma}^{(t)}\}$ for each iteration, where the superscript (t) denotes the t-th iteration. In the reduction step, we obtain the estimates for the original parameters $\tilde{\theta}^{(t)} = \{\tilde{\alpha}^{(t)}, \tilde{\beta}^{(t)}, \tilde{\mathbf{R}}^{(t)}\}$. Derivation can be found in Supplementary Section 2. If the correlation coefficient ρ is zero, we can analyze the traits independently, which provides warm starts for generating our three-stage algorithm for bLPM (see Supplementary Section 3).

For K GWASs, we analyze them pairwisely using the above algorithm and obtain the corresponding estimates $\hat{\tilde{\theta}}$. We can implement this procedure parallelly. To obtain the estimates $\hat{\alpha}_k$ and $\hat{\beta}_k$ for LPM, we use the average over the pairs containing the k-th GWAS: $\hat{\alpha}_k = \sum_{l\neq k}^{K} \hat{\tilde{\alpha}}_{lk(k)}$, $\hat{\beta}_k = \sum_{l\neq k}^{K} \hat{\tilde{\beta}}_{lk(k)}$, where $\hat{\tilde{\alpha}}_{lk}$ and $\hat{\tilde{\beta}}_{lk}$ are the estimate of $\tilde{\alpha}$ and $\tilde{\beta}$, respectively in bLPM for the l-th trait and the k-th trait, and (k) means the entry for the k-th trait. We can also form a matrix $\hat{\mathbf{R}}_{pair}$ using the corresponding estimation $\hat{\rho}$ in pair-wise analysis. In real data analysis, the number of SNPs M is often different in each GWAS due to different genotyping platform and quality control. To avoid losing much information, we allow different M in each bLPM. However, since the pair-wise analysis is not based on the same data, $\hat{\mathbf{R}}_{pair}$ may not be positive semidefinite. As such, we solve the following optimization problem to obtain the nearest correlation matrix $\hat{\mathbf{R}}$: $\min \frac{1}{2} ||\mathbf{R} - \hat{\mathbf{R}}_{pair}||^2$, s.t. $\mathbf{R} \in \mathcal{S}_+^K, R_{kk} = 1, ; k = 1, \ldots, K$, where $\mathcal{S}_+^K$ is the cone of positive semidefinite matrices in the space of $K \times K$ symmetric matrices, and $||\cdot||$ is the Frobenius norm. This problem can be efficiently solved by a Newton-type method (Qi and Sun, 2006).

2.3 Inferences based on LPM

2.3.1 Identification of risk SNPs

After we obtain the estimates of parameters in LPM, we are able to prioritize risk SNPs based on the posterior of η, which indicates the strength of association of the SNPs with the traits.

If we consider the traits separately, the association mapping of the j-th SNP on the k-th trait can be inferred from $\Pr(\eta_{jk} = 1|P_{jk}, \mathbf{X})$. In this case, the relationship among traits is ignored and only the current GWAS data are used.

If two traits are considered, risk SNPs for both the k-th trait and the k'-th trait can be inferred from $\Pr(\eta_{jk} = 1, \eta_{jk'} = 1|P_{jk}, P_{jk'}, \mathbf{X})$. In addition, we can infer the risk SNPs for the k-th trait by calculating the marginal posterior $\Pr(\eta_{jk} = 1|P_{jk}, P_{jk'}, \mathbf{X})$. Similarly, we can consider more than two traits.

Moreover, we can calculate the local false discovery rate (FDR) and use the direct posterior probability approach (Newton *et al.*, 2004) to control the global FDR. The details are shown in Supplementary Section 4.

2.3.2 Relationship test among traits

We test the relationship between two traits in the pair-wise analysis by the hypothesis:

$$H_0 : \rho = 0 \text{ v.s. } H_1 : \rho \neq 0.$$

We use the likelihood ratio test. The test statistic is

$$\lambda = 2[\log \Pr(\tilde{\mathbf{P}}|\mathbf{X}; \tilde{\theta}) - \log \Pr(\tilde{\mathbf{P}}|\mathbf{X}; \tilde{\theta}_0)],$$

where $\tilde{\theta}_0$ is the parameter estimates under H_0, i.e. the estimates we directly obtain in the second stage of the algorithm. The probability distribution of λ is asymptotically a χ^2 distribution with $df = 1$ under the null.

2.3.3 Hypothesis testing of annotation enrichment

When we integrate functional annotation data, we are interested in the enrichment of annotation for a specific trait. We consider the following test:

$$H_0 : \beta_{kd} = 0 \text{ v.s. } H_1 : \beta_{kd} \neq 0.$$

As the inverse of observed information matrix provides an estimator of the asymptotic covariance matrix, the Wald test statistic is

$$W = \frac{\hat{\beta}_{kd}^2}{[\mathcal{I}(\hat{\theta}_k)]_{d+1,d+1}^{-1}},$$

where $\mathcal{I}(\hat{\theta}_k)$ is the observed information matrix at $\hat{\theta}_k = (\hat{\alpha}_k, \hat{\beta}_k^T)^T$ when only consider the k-th trait. The details for calculation are shown in Supplementary Section 5. The probability distribution of W is approximately a χ^2 distribution with $df = 1$ under the null.

3 Results

3.1 Simulation

3.1.1 Performance in characterizing the correlations among traits

We simulated summary statistics of eight GWASs and design matrix of five functional annotations as follows. The eight traits were

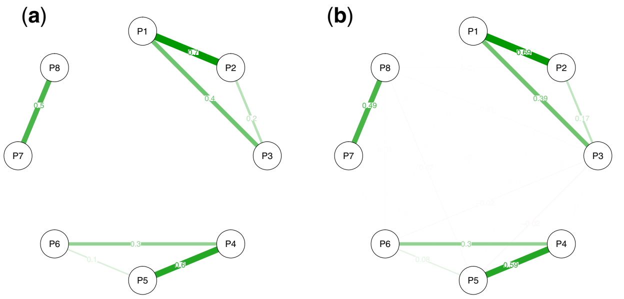

Fig. 1. Correlation graphs. (**a**) True graph. (**b**) Average estimated correlation graph using LPM. The numbers on the edges and the widths of the edges indicate the correlation between the connected traits. The results are summarized from 50 replications

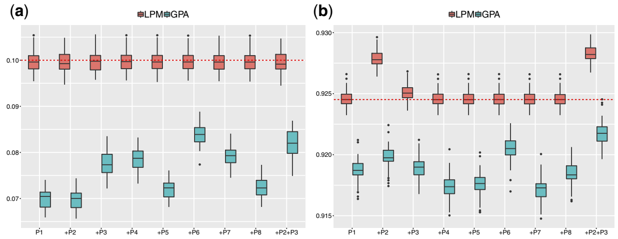

Fig. 2. (**a**) FDR and (**b**) AUC of LPM and GPA for identification of risk SNPs for P1. The x-axis indicates which traits are used in analysis, e.g. P1 indicates separate analysis using only P1, indicates joint analysis of P1 and P2. We controlled global FDR at 0.1 to evaluate empirical FDR. The red horizontal line in (b) was set at the median AUC in separate analysis using LPM as a reference line. The results are summarized from 50 replications. (Color version of this figure is available at *Bioinformatics* online.)

divided into three groups: (i) P1, P2, P3; (ii) P4, P5, P6 and (iii) P7, P8. Correlation existed only within the groups. Specifically, we set the correlation matrix $\mathbf{R}$ with corresponding entries $\rho_{12} = 0.7$, $\rho_{13} = 0.4$, $\rho_{23} = 0.2$, $\rho_{45} = 0.6$, $\rho_{46} = 0.3$, $\rho_{56} = 0.1$ and $\rho_{78} = 0.5$ (all the other entries were set to zeros). The relationship among the traits is depicted in Figure 1a. The numbers of SNPs and functional annotations were set to be $M = 100\,000$ and $D = 5$, respectively. First, we generated design matrix $\mathbf{X}$ and coefficients β of functional annotations. The entries in $\mathbf{X}$ excluded the intercept were generated from $Bernoulli(0.2)$. The entries in first column of β were set to be -1 and the other entries were first generated from $N(0, 1)$ and then transformed to control the relative signal strength between annotated part and un-annotated part $r = \frac{[Var(\mathbf{X}\beta^T)]_{kk}}{[Var(\epsilon)]_{kk}}$ for $k = 1, \ldots, 8$ to be 1. Both $\mathbf{X}$ and β were kept fixed in multiple replications. Then, we simulated η_{jk} according to multivariate probit models (1) and (2). Finally, we generated P_{jk} from $U[0, 1]$ if $\eta_{jk} = 0$ and $Beta(\alpha_k, 1)$ if $\eta_{jk} = 1$ with $\alpha_1 = 0.2$, $\alpha_2 = 0.35$, $\alpha_3 = 0.5$, $\alpha_4 = 0.3$, $\alpha_5 = 0.45$, $\alpha_6 = 0.55$, $\alpha_7 = 0.25$ and $\alpha_8 = 0.4$.

Figure 1b shows that the correlation graph is accurately estimated using LPM. We also evaluated the type I error rate and power of LPM for the relationship test among the traits and compared it with GPA and graph-GPA. For LPM, as shown in Supplementary Figure S1, the type I error rates are almost 0 for all the pairs with no correlation and the powers are almost 1 except for two pairs (P2 and P3, P5 and P6) in which cases the correlations are relatively small and the signal strength is relatively weak, i.e. the corresponding ρs are relatively small and αs are relatively large. As the relationship both GPA and graph-GPA measured does not adjust the effect of functional annotations, more significant relationships are detected. If all the functional annotations have no role, i.e. $\mathbf{X}$ only had the intercept term, the relationship test graphs of LPM and GPA are similar and some connections are not detected by graph-GPA as it represents a conditional independent structure (see Supplementary Fig. S2). For a more dense correlation graph, similar results are given in Supplementary Figures S3–S6.

To provide a better illustration for the performance of LPM, we conducted simulations which considered only two traits to obtain the type I error rate and power of LPM for the relationship test. As shown in Supplementary Figure S7, the type I error rates are well controlled in all cases and the power increases as signal strength α decreases and as ρ increases. However, we noted that a large relative signal strength r could lead to a small power. This is because the correlation resulting from annotations increases as r increases and the correlation we aim to estimate becomes relatively smaller.

3.1.2 Performance in the identification of risk SNPs for one or multiple traits

To identify risk SNPs for one specific trait, we consider three cases (i) separate analysis of the target trait, (ii) joint analysis of the target trait with another trait and (iii) joint analysis of the target trait with other two traits, using LPM. If the integrated traits are correlated with the target trait, the power to identify risk SNPs is expected to increase in joint analysis.

We compared LPM with GPA under these three cases in terms of their empirical FDR and AUC. The results for P1 are shown in Figure 2 (results for other traits are given in Supplementary Figs S8–S14). The empirical FDRs of LPM are indeed controlled at the nominal level. However, the FDRs of GPA are inflated in some cases when the GWAS signal is relatively weak and are conservative when the GWAS signal is relatively strong. Moreover, LPM outperformed GPA for all the cases in terms of AUC. As expected, the AUC of LPM increases when correlated traits are integrated. For example, as shown in Figure 2b, the power to identify risk SNPs for P1 increases when the correlated traits (P2 and P3) are jointly analyzed. Specifically, integrating traits with high correlation with the target trait could result in a better improvement of AUC (simulations are described in Supplementary Section 6.6). We also compared LPM with RiVIERA and found LPM achieved a better performance (see details of simulation and results in Supplementary Section 6.7).

For the identification of SNPs associated with two and three traits, the comparison performance of LPM and GPA is shown in Supplementary Figures S18 and S19. LPM performed better in terms of FDR control and AUC. In the identification of risk SNPs for both P1 and P4, a larger AUC can be achieved by integrating traits correlated with either P1 or P4, i.e. integrating P2, P3, P5 or P6.

When functional annotations do not play a role, we have shown the relationship test graphs are similar for LPM, GPA and graph-GPA. In this case, we also compared their performance in identification of risk SNPs for one specific trait. The results are shown in Supplementary Figures S20–S27. The performance of LPM and GPA is very close in terms of FDR and AUC. However, for graph-GPA, the empirical FDRs are conservative and AUCs are relatively small.

3.1.3 Type I error rate and power for the hypothesis testing of annotation enrichment

We conducted simulations which considered only two traits with number of annotations $D = 1$, correlation coefficient $\rho = 0$ and same signal strength for the traits, i.e. $\alpha_1 = \alpha_2 = \alpha$. We varied α in $\{0.2, 0.4, 0.6\}$ to obtain the type I error rate, and varied the coefficient of annotation β in $\{-0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4\}$ to obtain the power of LPM for the enrichment test of the annotation. The results are shown in Supplementary Figure S28. We observed that the type I error rate was indeed controlled at the nominal level and the power was close to one when the signal strength was relatively strong (i.e. $\alpha = 0.2$ or 0.4), and the coefficient was not very small (i.e. $|\beta| \geq 0.2$). When functional annotations only correspond to a small proportion of genome, the power decreases with lower functional proportion especially when the absolute value of effect size is small (see Supplementary Fig. S29). However, LPM can still provide a valid type I error rate control, as shown in Supplementary Figure S30.

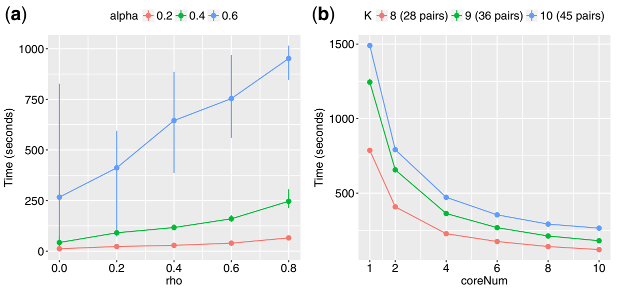

Fig. 3. Computational time of LPM. (**a**) For one pair of traits, we varied the signal strength α and the correlation ρ. (**b**) For different numbers of traits, we varied the number of cores we used. The results are summarized from 10 replications

3.1.4 Computational time

Figure 3 shows the computational time of LPM with $M = 100\ 000$ and $D = 5$. For one pair of traits, the computational time depends on the signal strength of GWAS data and their correlation. When the number of traits increases, the time can be largely shortened by using more cores in parallel computation. However, the time is not linear in the number of cores. One reason is that the time we recorded not only included the time to fit bLPM for pairs of GWASs which is implemented parallelly, but also included the time for data preparation. Another reason is that we used Armadillo (Sanderson and Curtin, 2016) in our R package which has already executed many functions (e.g. matrix multiplication) in parallel.

3.2 Real data analysis

We applied LPM to analyze 44 GWASs with 9 genic category annotations and 127 cell-type specific functional annotations. The source of the summary statistics and sample sizes of 44 GWASs are given in Supplementary Tables S2 and S3, respectively. The genic category annotations were provided by ANNOVAR (Wang *et al.*, 2010). In this article, we used nine genic category annotations which include upstream, downstream, exonic, intergenic, intronic, ncRNA_exonic, ncRNA_intronic, UTR′3 and UTR′5, where ncRNA referred to RNA without coding annotation. The cell-type specific functional annotations were generated using epigenetic markers (H3k4me1, H3k4me3, H3k36me3, H3k27me3, H3k9me3, H3k27ac, H3k9ac and DNase I Hypersensitivity) in 127 tissues and cell types from the Epigenomics Roadmap Project. We collected 127 cell-type specific functional annotations from GenoSkylinePlus (Lu *et al.*, 2017b). We excluded the SNPs in the MHC region (Chromosome 6, 25–35 Mb) to avoid unusually large GWAS signals.

3.2.1 Correlations among 44 GWASs

The estimated correlations among 44 GWASs are shown in Figure 4. We have observed that the correlations among traits were quite dense (see Supplementary Table S4) and most of the correlations are positive. This is because the correlation given by LPM measures the correlation of association status η in the probit scale. The positive values indicate that the association status of two traits is largely overlapped, implying the ubiquity of pleiotropy, i.e. many SNPs can affect both traits under consideration.

According to Figure 4, traits can be divided into several groups with relatively strong correlations and these are consistent with the categories of traits. Main groups are psychiatric disorders (average of the estimated correlations in this group $\bar{\rho} = 0.73$), which include BIP (Sklar *et al.*, 2011), SCZ (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Ripke *et al.*, 2011, 2013, 2014), neuroticism, depressive symptoms (Okbay *et al.*, 2016a), MDD (Wray *et al.*, 2018), attention deficit hyperactivity disorder (Demontis *et al.*, 2019), autism spectrum disorder (Grove *et al.*, 2019) and anorexia nervosa (Duncan *et al.*, 2017); hematopoietic traits ($\bar{\rho} = 0.85$), which include mean cell hemoglobin concentration, mean cell hemoglobin, mean cell volume, red blood cell count, hemoglobin and packed cell volume (Pickrell, 2014); autoimmune diseases ($\bar{\rho} = 0.73$), which include systemic lupus erythematosus (Bentham *et al.*, 2015), atopic dermatitis (Paternoster *et al.*, 2015), primary biliary cirrhosis (Cordell *et al.*, 2015), CD, UC, inflammatory bowel disease (Jostins *et al.*, 2012), celiac disease (Dubois *et al.*, 2010), RA (Okada *et al.*, 2014), multiple sclerosis (Sawcer *et al.*, 2011) and T1D (Bradfield *et al.*, 2011) and lipid-related traits ($\bar{\rho} = 0.84$), which include HDL, triglycerides, LDL and TC (Willer *et al.*, 2013).

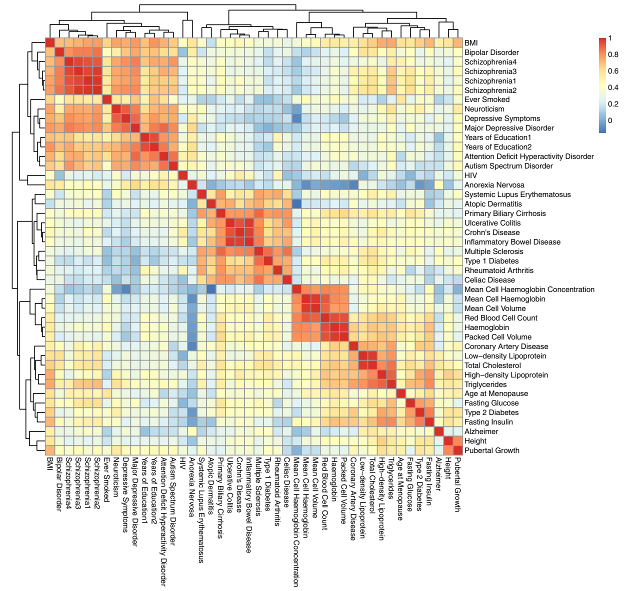

Fig. 4. The estimated $\hat{\mathbf{R}}$ for 44 GWAS with nine genic category annotations and 127 cell-type specific functional annotations integrated

We also found some significant relationships between complex diseases and metabolic traits. For example, relatively high correlations were observed between coronary artery disease (Schunkert *et al.*, 2011) and lipid-related traits ($\bar{\rho} = 0.79$), among type 2 diabetes (Morris *et al.*, 2012), fasting glucose and fasting insulin (Manning *et al.*, 2012) ($\bar{\rho} = 0.77$). We observed that psychiatric disorders were correlated with many other traits, such as body mass index (Speliotes *et al.*, 2010) ($\bar{\rho} = 0.72$), years of education (Okbay *et al.*, 2016b; Rietveld *et al.*, 2013) ($\bar{\rho} = 0.72$), HIV (McLaren *et al.*, 2013) ($\bar{\rho} = 0.65$) and ever smoked (Furberg *et al.*, 2010) ($\bar{\rho} = 0.69$). Similar evidence of the relationship has been found by Hartwig *et al.* (2016), Breslau *et al.* (2008), Chandra *et al.* (2005) and Lawrence *et al.* (2009). We also discovered connections between height (Wood *et al.*, 2014) and pubertal growth (Cousminer *et al.*, 2013) ($\hat{\rho} = 0.81$), between age at menopause (Day *et al.*, 2015) and fasting insulin ($\hat{\rho} = 0.62$), between Alzheimer (Lambert *et al.*, 2013) and lipid-related traits ($\bar{\rho} = 0.71$).

As a comparison, we used cross-trait LD score regression to estimate the genic correlations among several traits. The results are shown in Supplementary Figure S50. Although the definitions of correlation in cross-trait LD score regression and LPM are different, the trends of relationship among traits are quite similar because the two versions of correlation essentially capture the dependence among traits. We also applied graph-GPA to infer the relationship among traits. As graph-GPA was not scalable to a large number of traits, we can only analyze a subset of the traits. The relationship estimated by graph-GPA changed a lot as the number of traits changed (see Supplementary Fig. S51). This was because graph-GPA represented the relationship from a conditional independent structure, and the conditional independent structure might change when adding more traits or removing some included traits.

We compared the pair-wise correlations for 44 GWASs with and without considering functional annotations. The results are shown in Supplementary Figure S52. After incorporating the functional annotations, the correlations for 641 out of 946 pairs of traits decrease. However, the differences are not significant indicating that the correlation between two traits should be mainly attributed to the remaining polygenic effects rather than the known functional annotations. For example, the correlation between RA and fasting insulin is estimated to be $\hat{\rho} = 0.2263$ (se = 0.0641) without considering functional annotations. The correlation decreases after adjusting functional annotations ($\hat{\rho} = 0.0913$, se = 0.0639). For CD and UC, their correlation only slightly decreases from $\hat{\rho} = 0.9926$ (se = 0.0092) to $\hat{\rho} = 0.9920$ (se = 0.0095) after adjusting functional annotations.

3.2.2 Association mapping

We compared the number of SNPs identified to be associated with each of the 44 traits using LPM by five different analysis approaches: (i) separate analysis without annotation, (ii) separate analysis with genic category annotations, (iii) separate analysis with all annotations (genic category annotations and cell-type specific annotations), (iv) joint analysis of the top 1 correlated trait with all annotations and (v) joint analysis of the top 2 correlated traits with all annotations. The details of the top correlated traits are given in Supplementary Table S5. Using the fifth approach as a reference, we calculated the ratio of the number of risk SNPs identified using each approach. The results are given in Supplementary Figure S53.

The results show that more risk SNPs can be identified by integrating functional annotations and correlated traits. For HIV, age at menopause and Alzheimer, a clear improvement was observed between the first two approaches, reflecting a significant enrichment of genic category annotations. Comparing the second and the third approaches, the contributions of cell-type specific functional annotations were observed to be significant for many traits, such as atopic dermatitis, fasting insulin, multiple sclerosis, primary biliary cirrhosis, RA and T1D. The differences between the fourth and fifth approaches were due to pleiotropy. For example, LDL and TC were observed to be highly correlated ($\hat{\rho} = 0.97$). As a result, joint analysis of these two traits led to an improvement in identifying risk SNPs. Specifically, for the identification of risk SNPs for LDL with all annotations integrated, 3758 SNPs were identified in separate analysis of LDL, whereas in joint analysis of LDL and TC 7845 SNPs were identified, when the global FDR was controlled at 0.1. The Manhattan plots are provided in Supplementary Figure S54.

3.2.3 Enrichment of functional annotations

The results for the enrichment test of 9 genic category annotations and 127 cell-type specific functional annotations are shown in Figure 5. The detailed results of enrichment test are given in Supplementary Figures S55–S61 and the estimated coefficients of genic category annotations and cell-type specific functional annotation are given in Supplementary Figures S62–S66 and Figures S67–S73, respectively.

We noted that the estimated coefficients for genic category annotations except for intergenic were positive for most traits. For intergenic annotation, the estimated coefficients were negative for many traits, e.g. pubertal growth ($\hat{\beta} = -1.66$, se = 1.00) and ever smoked ($\hat{\beta} = -0.60$, se = 0.78). Comparing the number of traits with significant coefficients for each genic category annotation, we found that exonic and UTR′3 were enriched most. Our results are consistent with findings in the previous study (Schork *et al.*, 2013). Note that standard errors of the estimated coefficients are influenced by sample size and genetic architecture. The overall GWAS signal strength can be seen from estimated α, as given in Supplementary Table S3. Phenotypes with weak signal strength, i.e. large α, are likely to have large standard errors (e.g. HIV: $\hat{\alpha} = 0.75$, ever smoked: $\hat{\alpha} = 0.74$ and pubertal growth: $\hat{\alpha} = 0.67$).

For cell-type specific functional annotations, we detected enrichment of functional annotations for lipid-related traits in liver (liver and HepG2 cells) and fat (adipose nuclei). Specifically, the enrichment in liver was significant for all the four lipid-related traits (HDL: $\hat{\beta} = 0.30$, se = 0.03; LDL: $\hat{\beta} = 0.36$, se = 0.03; TC: $\hat{\beta} = 0.38$, se = 0.02 and triglycerides: $\hat{\beta} = 0.25$, se = 0.03), which was consistent with findings in previous studies (Kundaje *et al.*, 2015; Lu *et al.*, 2017b; Ming *et al.*, 2018). SNPs annotated in cells of immune system were observed to be enriched for many traits, including autoimmune diseases, lipid-related traits, hematopoietic traits, some psychiatric disorders, such as SCZ and BIP. For height, significant functional annotations included cells in immune system, bone, liver, muscle and skin. The foreskin fibroblast primary cells were shown to be enriched for some autoimmune diseases and lipid-related traits (Ming *et al.*, 2018).

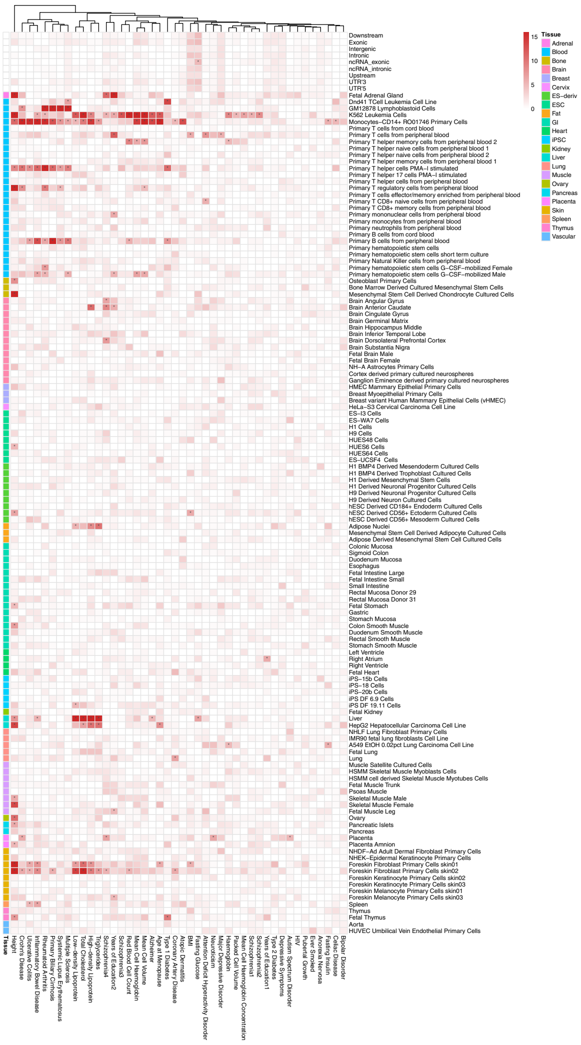

Fig. 5. The $-\log_{10}(P - \text{value})$ for enrichment test of nine genic category annotations and 127 cell-type specific functional annotations. The *P*-values which are smaller than 10^{-15} are set to be 10^{-15}. The symbol '*' means the *P*-value is significant after Bonferroni correction at level 0.05

3.2.4 Replications

Among the 44 GWASs, we analyzed four different GWASs of SCZ (SCZ1–4). We found that the correlations among them were very high ($\bar{\rho} = 0.95$) and their enrichment of genic category annotations (see Supplementary Figs S62–S66) and cell-type specific functional annotations (see Supplementary Figs S74 and S75) were highly consistent, indicating the replicability of the findings of LPM. As the sample size of GWASs from SCZ1 to SCZ4 became larger (see Supplementary Table S6), the standard error of the estimated coefficients became smaller and the *P*-values of the enrichment test for annotations became more significant. For SCZ1–3, only SNPs annotated in K562 Leukemia cells were enriched (SCZ1: $\hat{\beta} = 0.18$, se = 0.04; SCZ2: $\hat{\beta} = 0.18$, se = 0.03; SCZ3: $\hat{\beta} = 0.19$, se = 0.03). For SCZ4, besides K562 leukemia cells ($\hat{\beta} = 0.08$, se = 0.02), more enrichment in functional annotations was detected, such as brain anterior caudate ($\hat{\beta} = 0.11$, se = 0.02) and brain angular gyrus ($\hat{\beta} = 0.12$, se = 0.02). As shown in

Supplementary Figure S53, more risk SNPs were identified to be associated with SCZ by jointly analyzing these GWASs. Similar results can be found (see Supplementary Figs S76 and S77) for GWASs of educational traits, where Years of Education2 has a larger sample size than Years of Education1.

We also compared the results of UC, CD with IBD and depressive symptoms with MDD. As IBD is comprised of two major disorders: UC and CD, depressive symptoms includes MDD, the consistent results from analyzing these GWASs also indicate the replicability of our method (see Supplementary Figs S78–S81).

4 Discussion

From the perspective of statistical modeling, a remarkable benefit of our model is its scalability to a large number of GWASs. The LPM model only accounts for the first-order ($p(\eta_{jk})$) and second-order ($p(\eta_{jk}, \eta_{jk'})$) terms so that the number of parameters could increase quadratically with the number of traits. While some other methods, such as GPA, which considered all the high-order terms, the number of parameters will increase exponentially. Here, we made the assumption that the majority of information could be captured by the first-order and second-order terms. Therefore, the computational efficiency could be largely increased without losing too much accuracy. In particular, the design of LPM naturally allows a parallel algorithm such that the model fitting can be computationally efficient. The feasibility of the approach is demonstrated by our simulations in which the accuracy of parameter estimation using LPM is evaluated. Supplementary Figure S31 shows that LPM provides satisfactory estimate of α, β and $\mathbf{R}$. Specifically, the estimates of $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\mathbf{R}}$ using bLPM for different pairs are stable which shows the consistency and reliability of our algorithm (Supplementary Figs S32–S41). It also has theoretical supports which are based on the composite likelihood approach. The details of related theorem are provided in Supplementary Section 9. Because of the pair-wise analysis, the number of SNPs for each pair of GWASs can be different. There is no need to remove SNPs with missing values in any one of the GWASs, avoiding the huge information loss especially for large amounts of traits.

Besides the perspective of computational efficiency, another difference between LPM and GPA is that GPA assumes the conditional independence among multiple annotations given the latent association status indicator while LPM not. In the presence of multiple highly correlated annotations, LPM can provide more stable estimation and hypothesis testing results. For correlated annotations, the functional enrichment estimates may be different when testing each annotation at a time or testing all annotations simultaneously. This is quite similar to standard linear regression: marginal analysis and joint analysis. The choice of these two methods depends on our goal, the marginal effect for each annotation or the effect with all annotations taken into consideration. We provided a demonstration to show the difference in Supplementary Figure S82.

LPM assumes that the P-values of SNPs in each GWAS are from a mixture of uniform and beta distributions. This mixture model adopted in Pounds and Morris (2003), GPA and LSMM have been shown to be an excellent approximation to observed distribution of P-values arising from either microarray study or GWAS. In the two-group model, effect sizes and standard errors are not modeled directly [e.g. unlike Zhu and Stephens (2017)] because we are aiming to infer the association status collected in η instead of to estimate the true effect sizes. However, the information of sample size and heritability in the k-th study can be captured by parameter α_k in the beta distribution (see Supplementary Fig. S42). To check whether LPM is robust to violation of this assumption, we considered the situations when P-values are obtained from individual-level data and when P-values in non-null group are from distributions other than beta distribution (details are given in Supplementary Sections 6.14 and 6.15, respectively). The results show that the type I error rate of LPM for the relationship test and the empirical FDR for identification of risk SNPs are well controlled at the nominal level.

We assume the proportion of risk variants should not be extremely small to ensure the accuracy of parameter estimation. We conducted simulations for two traits with same signal strength ($\alpha_1 = \alpha_2 = \alpha$) and no annotations. In this simulation, we varied the proportion of risk variants π_1 in $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2\}$ and evaluated the accuracy of the estimation of π_1 and ρ using LPM. The results are shown in Supplementary Figures S47 and S48, respectively. When the true proportion of risk SNPs is extremely small and the signal of GWAS data are weak (e.g. $\pi_1 \le 0.001$ for $\alpha = 0.4$ and $\pi_1 \le 0.01$ for $\alpha = 0.6$), the estimation of π_1 using LPM can be inaccurate and ρ is underestimated under similar circumstances ($\pi_1 \le 0.01$ for $\alpha \le 0.4$ and $\pi_1 \le 0.05$ for $\alpha = 0.6$). However, we believe LPM is widely applicable to complex traits because the proportion of risk variants is not very small due to the polygenic.

Here, we briefly discuss the relationship between LPM and related methods for prioritizing risk variants and characterizing functional enrichment.

- As causal variants may not be directly genotyped in GWAS, LPM is designed to prioritize a local genome region which contains causal variants rather than directly identify causal variants. To achieve this goal, SNPs are assumed to be conditionally independent given the functional annotations. This assumption greatly facilitates computation and inference of LPM when it is applied to genome-wide SNP data generated by genotyping arrays. Differently from LPM, fine-mapping methods, such as PAINTOR (Kichaev *et al.*, 2014) and RiVIERA-MT (Li *et al.*, 2017), take LD into account, aiming to directly identify causal variants. An inexplicit assumption of these fine-mapping methods is that the causal variants have been genotyped or accurately imputed in the given data. When the assumption holds, these methods may have better performance over LPM for prioritizing causal variants. However, accounting for LD also brings much more computational cost. To reduce the computational burden, PAINTOR makes restriction on the number of causal variants in a given LD block (two causal variants as the default setting) and RiVIERA-MT approximates posterior of causal configurations through stochastic search. Generally speaking, we should have a trade-off between statistical accuracy and computational cost in different applications. Fine-mapping methods could be preferred to handle GWAS data with tens of millions of imputed SNPs if computational cost is affordable. LPM can be applied to GWAS data with about a million of SNPs where LD effect is not a major issue for localizing risk variants in a small region of the genome. To verify this, we conducted simulations using real genotype and real annotations to evaluate the impact of LD effects on LPM. Details of simulations are given in Supplementary Section 6.17. The results (Supplementary Fig. S49) indicate that LPM can provide a satisfactory FDR control in terms of identifying a local genomic region of the risk SNPs.
- For functional enrichments, sLDSC (Finucane *et al.*, 2015) is a popular way addressing for LD. It is based on the proportion of variance explained by functional annotation. The analytic tool to characterize enrichment in sLDSC is the variance component model. However, the definition of functional enrichment in LPM is different, which is based on the proportion of association status in the annotation. A probit model is used in LPM and logistic model was adopted in PAINTOR to estimate the functional enrichment. As a result, the enrichment results from LPM (or PAINTOR) may not be directly comparable to sLDSC.

In summary, we have presented a statistical approach, LPM, to integrate summary statistics from multiple GWASs and functional annotations. This unified framework can characterize relationship among complex traits, increase the statistical power for association mapping, integrate and investigate the effect of functional annotations simultaneously. With extensive simulations and real data analysis of 44 GWASs, we have demonstrated the statistical efficiency and computational scalability of LPM.

Funding

This work was supported in part by the National Natural Science Foundation of China [61501389, 11601326, 11971017]; the Hong Kong Research Grant Council [22302815, 12316116, 12301417, 16307818]; the National Key R&D Program of China [2018YFC0910500]; the Shanghai Municipal Science and Technology Major Project [2017SHZDZX01]; the University Grants Committee [initiation grant NO. IGN17SC02]; Neil Shen's SJTU Medical Research Fund; and The Hong Kong University of Science and Technology [startup grant R9405, Z0428 from the Big Data Institute].

Conflict of Interest: none declared.

References

Andreassen,O.A. *et al.* (2014) Identifying common genetic variants in blood pressure due to polygenic pleiotropy with associated phenotypes novelty and significance. *Hypertension*, **63**, 819–826.

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.

Bradfield,J.P. *et al.* (2011) A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.*, **7**, e1002293.

Breslau,J. *et al.* (2008) Mental disorders and subsequent educational attainment in a US national sample. *J. Psychiatr. Res.*, **42**, 708–716.

Bulik-Sullivan,B. *et al.*; ReproGen Consortium. (2015) An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**, 1236–1241.

Chandra,P.S. *et al.* (2005) HIV & psychiatric disorders. *Indian J. Med. Res.*, **121**, 451–467.

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

Chung,D. *et al.* (2017) graph-GPA: a graphical model for prioritizing GWAS results and investigating pleiotropic architecture. *PLoS Comput. Biol.*, **13**, e1005388.

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

Cousminer,D.L. *et al.*; The ReproGen Consortium. (2013) Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth, pubertal timing and childhood adiposity. *Hum. Mol. Genet.*, **22**, 2735–2747.

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.*; The PRACTICAL Consortium. (2015) Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.*, **47**, 1294–1303.

Demontis,D. *et al.*; ADHD Working Group of the Psychiatric Genomics Consortium (PGC). (2019) Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.*, **51**, 63–75.

Dubois,P.C.A. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295–302.

Duncan,L. *et al.*; Eating Disorders Working Group of the Psychiatric Genomics Consortium. (2017) Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *Am. J. Psychiatry*, **174**, 850–858.

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

Ellinghaus,D. *et al.*; The International IBD Genetics Consortium (IIBDGC). (2016) Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.*, **48**, 510–518.

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

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

Furberg,H. *et al.* (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.*, **42**, 441.

Grove,J. *et al.*; Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium. (2019) Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.*, **51**, 431–444.

Hartwig,F.P. *et al.* (2016) Body mass index and psychiatric disorders: a Mendelian randomization study. *Sci. Rep.*, **6**, 32730.

Ji,S.-G. *et al.*; The UK-PSC Consortium. (2017) Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat. Genet.*, **49**, 269–273.

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

Kichaev,G. *et al.* (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.*, **10**, e1004722.

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

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

Lawrence,D. *et al.* (2009) Smoking and mental illness: results from population surveys in Australia and the United States. *BMC Public Health*, **9**, 285.

Lee,S.H. *et al.*; International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.*, **45**, 984–994.

Li,Y. and Kellis,M. (2016) Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res.*, **44**, e144.

Li,Y. *et al.* (2017) A probabilistic framework to dissect functional cell-type-specific regulatory elements and risk loci underlying the genetics of complex traits. *bioRxiv*.

Liu,C. *et al.* (1998) Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**, 755–770.

Lu,Q. *et al.* (2017a) A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *Am. J. Hum. Genet.*, **101**, 939–964.

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

Manning,A.K. *et al.*; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012) A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.*, **44**, 659–669.

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.

Ming,J. *et al.* (2018) LSMM: a statistical approach to integrating functional annotations with genome-wide association studies. *Bioinformatics*, **34**, 2788–2796.

Morris,A.P. *et al.*; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.*, **44**, 981–990.

Newton,M.A. *et al.* (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

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

Okbay,A. *et al.*; LifeLines Cohort Study. (2016a) Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.*, **48**, 624–633.

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

Paternoster,L. *et al.*; Australian Asthma Genetics Consortium (AAGC). (2015) Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.*, **47**, 1449–1456.

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.

Pickrell,J.K. *et al.* (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.*, **48**, 709–717.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Qi,H. and Sun,D. (2006) A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM J. Matrix Anal. Appl.*, **28**, 360–385.

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

Ripke,S. *et al.* (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.*, **43**, 969–976.

Ripke,S. *et al.*; Multicenter Genetic Studies of Schizophrenia Consortium. (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.

Sanderson,C. and Curtin,R. (2016) Armadillo: a template-based C library for linear algebra. *J. Open Source Softw.*, **1**, 26.

Sawcer,S. *et al.*; Wellcome Trust Case Control Consortium 2. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, **476**, 214–219.

Schork,A.J. *et al.*; The Tobacco and Genetics Consortium. (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.

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

Sivakumaran,S. *et al.* (2011) Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.*, **89**, 607–618.

Sklar,P. *et al.* (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.*, **43**, 977–983.

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.

Solovieff,N. *et al.* (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, **14**, 483–495.

Speliotes,E.K. *et al.* (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, **42**, 937–948.

Varin,C. *et al.* (2011) An overview of composite likelihood methods. *Stat. Sin.*, **21**, 5–42.

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

Wang,K. *et al.* (2017) Classification of common human diseases derived from shared genetic and environmental determinants. *Nat. Genet.*, **49**, 1319–1325.

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

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

Wood,A.R. *et al.*; The Electronic Medical Records and Genomics (eMERGE) Consortium. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173–1186.

Wray,N.R. *et al.* (2018) Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.*, **50**, 668–681.

Yang,C. *et al.* (2015) Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. *Front. Genet.*, **6**, 229.

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

Zheng,J. *et al.*; Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium. (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, **33**, 272–279.

Zhu,X. and Stephens,M. (2017) Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.*, **11**, 1561–1592.