*Bioinformatics*, 33(24), 2017, 3878–3886
doi: 10.1093/bioinformatics/btx512
Advance Access Publication Date: 14 August 2017
Original Paper

Genome analysis

# LLR: a latent low-rank approach to colocalizing genetic risk variants in multiple GWAS

**Jin Liu[1,†], Xiang Wan[2,†], Chaolong Wang[3], Chao Yang[4], Xiaowei Zhou[5] and Can Yang[6,7,*]**

[1]Center for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore, [2]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China, [3]Genome Institute of Singapore, A*STAR, Singapore, Singapore, [4]Baidu Inc, Shanghai, China, [5]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA, [6]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China and [7]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on August 29, 2016; revised on July 9, 2017; editorial decision on August 8, 2017; accepted on August 9, 2017

## Abstract

**Motivation**: Genome-wide association studies (GWAS), which genotype millions of single nucleotide polymorphisms (SNPs) in thousands of individuals, are widely used to identify the risk SNPs underlying complex human phenotypes (quantitative traits or diseases). Most conventional statistical methods in GWAS only investigate one phenotype at a time. However, an increasing number of reports suggest the ubiquity of pleiotropy, i.e. many complex phenotypes sharing common genetic bases. This motivated us to leverage pleiotropy to develop new statistical approaches to joint analysis of multiple GWAS.

**Results**: In this study, we propose a latent low-rank (LLR) approach to colocalizing genetic risk variants using summary statistics. In the presence of pleiotropy, there exist risk loci that affect multiple phenotypes. To leverage pleiotropy, we introduce a low-rank structure to modulate the probabilities of the latent association statuses between loci and phenotypes. Regarding the computational efficiency of LLR, a novel expectation-maximization-path (EM-path) algorithm has been developed to greatly reduce the computational cost and facilitate model selection and inference. We demonstrate the advantages of LLR over competing approaches through simulation studies and joint analysis of 18 GWAS datasets.

**Availability and implementation**: The LLR software is available on https://sites.google.com/site/liujin810822.

**Contact**: macyang@ust.hk.edu

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

## 1 Introduction

Thousands of genome-wide association studies (GWAS) have been conducted over the past decade to identify the genetic risk variants [i.e. single-nucleotide polymorphisms (SNPs)] underlying such complex phenotypes as human height, diabetes and psychiatric disorders [see GWAS catalog (Welter *et al.*, 2014) http://www.genome.gov/gwastudies/]. The progress achieved by GWAS suggests that complex phenotypes are often affected by many variants with weak individual effects rather than just a few variants with large effects (Visscher *et al.*, 2012; Yang *et al.*, 2010). In conventional GWAS data analysis, association mapping is performed on one phenotype at a time (Stephens and Balding, 2009). Although many methods have been proposed to improve the power of association mapping (Cantor *et al.*, 2010), those improvements are often limited due to polygenicity (Gamazon *et al.*, 2015).

Recently, there is accumulating evidence to suggest the ubiquity of pleiotropy, i.e. many complex phenotypes sharing common genetic bases (Cotsapas *et al.*, 2011; Solovieff *et al.*, 2013; Visscher and Yang, 2016; Wang *et al.*, 2015; Yang *et al.*, 2015). Examples include the *PTPN22* gene associated with multiple auto-immune disorders, such as rheumatoid arthritis, Crohn's disease and type I diabetes (T1D) (Cotsapas *et al.*, 2011), and the *ABO* gene that is associated with both coronary artery disease (CAD) and tonsillitis (Pickrell *et al.*, 2016). The Psychiatric Genomics Consortium (PGC) investigated the shared genetic etiology of five psychiatric disorders by analyzing GWAS data on 33 332 cases and 27 888 controls (Psychiatric Genomics Consortium, 2013), and identified four loci, including *CACNA1C* and *CACNB2*. Further analysis revealed a significant genetic correlation among the psychiatric disorders considered. For example, the degree of genetic correlation between schizophrenia (SCZ) and bipolar disorder (BPD) was estimated to be around 0.68 (Cross Disorder Group of the Psychiatric Genomics Consortium, 2013). Therefore, leveraging pleiotropy in the joint analysis of multiple GWAS appears to be a promising strategy for association mapping (Segura *et al.*, 2012) and risk prediction (Li *et al.*, 2014).

Several studies have reported encouraging results on power improvement through the joint analysis of multiple GWAS (Segura *et al.*, 2012; Zhou and Stephens, 2014). However, these methods require individual-level genotype data as their input. This requirement is likely to be a major obstacle in the joint analysis of multiple GWAS because individual-level data from multiple GWAS are often unavailable to research groups. Data sharing agreements among multiple research groups and privacy protection regulations require considerable efforts in practice. Instead, the summary statistics (such as *z*-values and *P*-values) of many GWAS are publicly available. To make use of such a rich data resource, serval statistical approaches have been proposed, including GPA (Chung *et al.*, 2014), fgwas (Pickrell, 2014), PAINTOR (Kichaev *et al.*, 2014), trans-ethnic PAINTOR (Kichaev and Pasaniuc, 2015), CAPSSOC (Zhu *et al.*, 2015), CAVIAR (Hormozdiari *et al.*, 2014) and MGAS (Van der Sluis *et al.*, 2015) but most are limited to the analysis of one or two GWAS. Effective methods for harnessing the summary statistics from multiple GWAS to colocalize risk variants remain lacking.

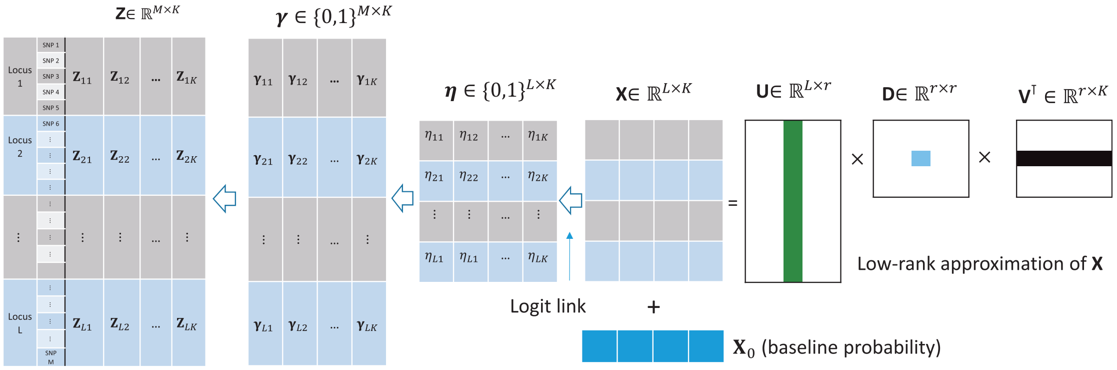
In this article, we propose a latent low-rank (LLR) approach to colocalizing genetic risk variants through the analysis of summary statistics, specifically *Z*-scores. In the presence of pleiotropy, a locus can be associated with multiple phenotype. This offers us an opportunity to improve the power of identifying risk locus by borrowing information across multiple studies. To do so, we introduce a low-rank structure to modulate the association probabilities between loci and phenotypes. Because the association status between the locus and the phenotype is not directly observable, the proposed low-rank structure is a latent variable model. Although the standard expectation-maximization (EM) algorithm is applicable to our model, it is too computationally expensive to handle genome-wide summary statistics from multiple GWAS. To address this issue, we have developed a novel EM-path algorithm that greatly reduces the computational costs and facilitates parameter tuning. We show through simulations that LLR consistently outperforms competing approaches, and also illustrate its benefits in the joint analysis of 18 phenotypes.

## 2 LLR: model, algorithm and inference

Before introducing LLR in details, we first outline its model structure, which is illustrated in Figure 1. LLR only requires the *Z*-scores of multiple GWAS as its input. For modeling convenience, we assume that the entire genome partition has been partitioned into nearly independent loci (Berisa and Pickrell, 2016). Given the association status of SNPs and loci, the probabilistic model of *Z*-scores can be derived, as discussed in Section 2.1. In Section 2.2, we propose a low-rank structure to modulate the prior probability of a given association status between loci and phenotypes, where the correlation induced by pleiotropy is taken into account. We then introduce a novel EM-path algorithm to render LLR applicable to large-scale genomic data analysis. Finally, we discuss how to use the false discovery rate obtained by LLR to prioritize risk variants.

### 2.1 Basic model for a single GWAS

Suppose that we have collected the *Z*-scores of $M$ SNPs from $K$ GWAS in matrix $\mathbf{Z} = [Z_{jk}] \in R^{M\times K}$, where $Z_{jk}$ corresponds to the *Z*-score of the *j*th SNP in the *k*th GWAS. Note that the *Z*-scores are obtained by testing one SNP at a time in a single GWAS analysis. Suppose that the $M$ SNPs can be partitioned into $L$ nearly



$\mathbf{Z} \in \mathbb{R}^{M\times K}$
$\boldsymbol{\gamma} \in \{0,1\}^{M\times K}$
$\boldsymbol{\eta} \in \{0,1\}^{L\times K}$
$\mathbf{X} \in \mathbb{R}^{L\times K}$
$\mathbf{U} \in \mathbb{R}^{L\times r}$
$\mathbf{D} \in \mathbb{R}^{r\times r}$
$\mathbf{V}^{\mathrm{T}} \in \mathbb{R}^{r\times K}$
Locus 1
Locus 2
Locus L
SNP 1
SNP 2
SNP 3
SNP 4
SNP 5
SNP 6
SNP M
$Z_{11}$ $Z_{12}$ ... $Z_{1K}$
$Z_{21}$ $Z_{22}$ ... $Z_{2K}$
$Z_{L1}$ $Z_{L2}$ ... $Z_{LK}$
$\gamma_{11}$ $\gamma_{12}$ ... $\gamma_{1K}$
$\gamma_{21}$ $\gamma_{22}$ ... $\gamma_{2K}$
$\gamma_{L1}$ $\gamma_{L2}$ ... $\gamma_{LK}$
$\eta_{11}$ $\eta_{12}$ ... $\eta_{1K}$
$\eta_{21}$ $\eta_{22}$ ... $\eta_{2K}$
$\eta_{L1}$ $\eta_{L1}$ ... $\eta_{LK}$
=
×
×
Low-rank approximation of **X**
Logit link
+
$\mathbf{X}_0$ (baseline probability)

**Fig. 1.** Model structure of LLR. The LLR input is the *Z*-scores from *K* studies, denoted as $\mathrm{Z} \in R^{M\times K}$, where *M* is the number of SNPs. By partitioning the genome into *L* loci, matrix Z is partitioned accordingly. Note that $\mathrm{Z}_{lk} \in R^{M_l\times 1}$ is the collection of *Z*-scores corresponding to locus *l* and phenotype *k*. The distribution of *Z*-scores depends on the association status of the SNPs and loci, denoted as $\gamma$ and $\boldsymbol{\eta}$, as given in Section 2.1. A low-rank structure $\mathrm{X} = \mathrm{UDV}^T$ is introduced to incorporate pleiotropy, and it modulates the prior probability of $\boldsymbol{\eta}$ via the logit link, as demonstrated in Section 2.2

independent loci. For locus $l$, we have $M_l$ SNPs, where $M = \sum_{l=1}^{L} M_l$. Given the SNPs at locus $l$ for phenotype $k$, we denote their $Z$-scores as $\mathbf{Z}_{lk}$ and association status matrix as $\gamma_{lk} \in \{0,1\}^{M_l \times 1}$, respectively. If all of the $M_l$ SNPs are independent, then the probability model for $\mathbf{Z}_{lk}$ is

$$\Pr(\mathbf{Z}_{lk}|\boldsymbol{\lambda}_{lk}, \gamma_{lk}) = \mathcal{N}(\boldsymbol{\lambda}_{lk} \circ \gamma_{lk}, \mathbf{I}), \tag{1}$$

where $\boldsymbol{\lambda}_{lk} \in R^{M_l \times 1}$ is the non-central parameter (NCP) and $\circ$ denotes the element-wise product. In the presence of the linkage disequilibrium (LD), the $Z$-scores are no longer independent but correlated. The above model can be modified as follows to adjust the LD effects (Hormozdiari *et al.*, 2014; Kichaev *et al.*, 2014):

$$\Pr(\mathbf{Z}_{lk}|\gamma_{lk}, \boldsymbol{\lambda}_{lk}, \boldsymbol{\Sigma}_l) = \mathcal{N}(\boldsymbol{\Sigma}_l(\boldsymbol{\lambda}_{lk} \circ \gamma_{lk}), \boldsymbol{\Sigma}_l), \tag{2}$$

where $\boldsymbol{\Sigma}_l \in R^{M_l \times M_l}$ is the correlation among the SNPs at locus $l$, which can be accurately estimated from 1000 genome reference panel data (1000 Genomes Project Consortium, 2010). Hence, it can be treated as known. For NCP $\boldsymbol{\lambda}_{lk}$, we follow the same strategy as that of PAINTOR (Kichaev *et al.*, 2014) to treat them as fixed, i.e. the NCP is set to be the observed $Z$-score if the absolute value of the $Z$-score is larger than 5.3, or the sign of the observed $Z$-score multiplied by 5.3 otherwise. Because $\boldsymbol{\lambda}_{lk}$ and $\boldsymbol{\Sigma}_l$ are treated as fixed, we denote $\Pr(\mathbf{Z}_{lk}|\gamma_{lk}, \boldsymbol{\lambda}_{lk}, \boldsymbol{\Sigma}_l)$ as $\Pr(\mathbf{Z}_{lk}|\gamma_{lk})$.

We now specify the joint probabilistic model for $\mathbf{Z}_k = [\mathbf{Z}_{1k}, \ldots, \mathbf{Z}_{Lk}]^T \in R^M$ and $\gamma_k = [\gamma_{1k}, \gamma_{1k}, \ldots, \gamma_{Lk}]^T \in \{0,1\}^M$

$$\Pr(\mathbf{Z}_k, \gamma_k) \quad = \prod_{l=1}^{L} \Pr(\mathbf{Z}_{lk}, \gamma_{lk}) = \prod_{l=1}^{L} \Pr(\mathbf{Z}_{lk}|\gamma_{lk}) \Pr(\gamma_{lk}), \tag{3}$$

where the equality holds due to the independence assumption among the loci. For locus $l$, what remains unknown is the probability of the $2^{M_l}$ configurations in $\gamma_{lk}$. To obtain that probability, we could in principle use the EM algorithm. However, without any constraints, there would be too many parameters to estimate, leading to expensive computation and inefficient statistical inferences. As a result, PAINTOR restricts the search for the maximum number of all possible causal variants to a default value of 2. In practice, Pickrell (2014) and Pickrell *et al.* (2016) set the number of causal variants within a locus to 1.

## 2.2 Latent low-rank model

Due to the polygenicity of complex phenotypes, the effect sizes of individual SNPs are very weak. However, the joint effects of multiple SNPs are still detectable. To leverage the strength of multiple studies, rather than relying solely on the association signals at the SNP level, we introduce a locus-level association status matrix $\boldsymbol{\eta} = [\eta_{lk}] \in \{0,1\}^{L \times K}$, where $\eta_{lk} = 1$ if the $l$th locus is associated with the $k$th phenotype, and $\eta_{lk} = 0$ otherwise. To avoid the combinatorial search of possible configurations in $\gamma_{lk}$ when $\eta_{lk} = 1$, we follow the same strategy as Pickrell (2014), which is to assume that there is only one risk SNP at locus $l$ for phenotype $k$ and that all $M_l$ SNPs have the same prior probability. Therefore, we have the following conditional probability.

$$\begin{aligned} &\Pr(\gamma_{lk} = 0|\eta_{lk} = 0) = 1, \\ &\Pr(\gamma_{j,lk} = 1, \gamma_{-j,lk} = 0|\eta_{lk} = 1) = 1/M_l, \end{aligned} \tag{4}$$

where $\gamma_{lk} = 0$ means that none of the SNPs at locus $l$ is associated with phenotype $k$, and $(\gamma_{j,lk} = 1, \gamma_{-j,lk} = 0)$ indicates that only SNP $j$ at locus $l$ is associated with phenotype $k$. To keep our notation simple, we use $\gamma_{lk}(j=1)$ to denote $\left(\gamma_{j,lk} = 1, \gamma_{-j,lk} = 0\right)$. Based on (3) and (4), we have the following joint probabilistic model.

$$\begin{aligned} &\Pr(\mathbf{Z}, \gamma, \boldsymbol{\eta}) \\ = \; &\Pr(\mathbf{Z}|\gamma)\Pr(\gamma|\boldsymbol{\eta})\Pr(\boldsymbol{\eta}) \\ = \; &\prod_{k=1}^{K}\prod_{l=1}^{L}\left\{\left(\Pr(\eta_{lk} = 0)\Pr(\mathbf{Z}_{lk}|\gamma_{lk} = 0)\right)^{1-\eta_{lk}} \left[\Pr(\eta_{lk} = 1)\right.\right. \\ &\left.\left.\prod_{j=1}^{M_l}\left(\frac{1}{M_l}\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1))\right)^{\mathbb{I}(\gamma_{lk}(j=1))}\right]^{\eta_{lk}}\right\}, \end{aligned} \tag{5}$$

where $\Pr(\mathbf{Z}_{lk}|\gamma_{lk} = 0) = \mathcal{N}(0, \boldsymbol{\Sigma}_l)$ and $\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1)) = \mathcal{N}(\boldsymbol{\Sigma}_l[\boldsymbol{\lambda}_{lk} \circ \gamma_{lk}(j=1)], \boldsymbol{\Sigma}_l)$. Integrating out latent variables $\gamma$ and $\boldsymbol{\eta}$ (see details of derivation in the Supplementary document), the incomplete-data likelihood becomes

$$\begin{aligned} &\Pr(\mathbf{Z}|\pi_1, \pi_2, \ldots, \pi_K) \\ = \; &\prod_{k=1}^{K}\prod_{l=1}^{L}\left\{(1-\pi_k)\mathcal{N}(0, \boldsymbol{\Sigma}_l)\right. \\ &\left.+\pi_k\left[\sum_{j=1}^{M_l}\frac{1}{M_l}\mathcal{N}\left(\boldsymbol{\Sigma}_l[\boldsymbol{\lambda}_{lk} \circ \gamma_{lk}(j=1)], \boldsymbol{\Sigma}_l\right)\right]\right\}, \end{aligned} \tag{6}$$

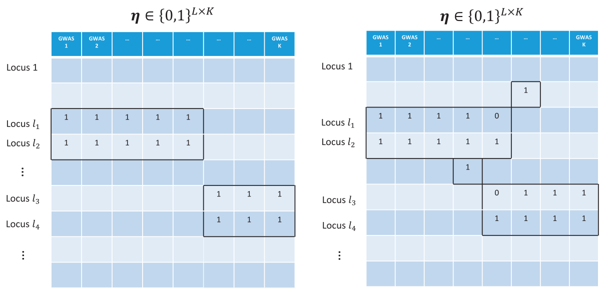
where $\pi_k = \Pr(\eta_{lk} = 1)$ denotes the prior probability that a locus is associated with phenotype $k$.

We now consider incorporating pleiotropic information into our model by modulating prior probability $\Pr(\eta_{lk})$ because the columns of $\boldsymbol{\eta}$ are correlated in the presence of loci that can affect multiple phenotypes. Such pleiotropy-induced correlation allows us to impose a low-rank structure on latent status matrix $\boldsymbol{\eta}$.

Let us consider the association pattern in $\boldsymbol{\eta}$. Ideally, if the pleiotropic loci affect the same group of phenotypes, as illustrated in the left panel of Figure 2, then $\boldsymbol{\eta}$ is an exact low-rank matrix. However, such an ideal case is extremely unlikely in practice. As we can see in the figure, the rank of $\boldsymbol{\eta}$ increases dramatically with even a small perturbation, which implies that the hard constraint that $\boldsymbol{\eta}$ has a low-rank structure in the presence of pleiotropy leads to the unstable estimation of $\boldsymbol{\eta}$. To overcome this difficulty, we propose a soft constraint on latent status matrix $\boldsymbol{\eta}$, i.e. we assume that there exists a low-rank matrix $\mathbf{X}$ that can modulate the probability of the latent status in $\boldsymbol{\eta}$ through a logit link:

$$\log\frac{\Pr(\eta_{lk} = 1|\mathbf{X}, \mathbf{x}_0)}{\Pr(\eta_{lk} = 0|\mathbf{X}, \mathbf{x}_0)} = X_{lk} + x_{0k}, \tag{7}$$

where $\mathbf{X} = [X_{lk}] \in R^{L \times K}$ is assumed to be a low-rank matrix and $x_{0k}$ is the intercept for GWAS $k$. In the absence of pleiotropy, each



$\boldsymbol{\eta} \in \{0,1\}^{L \times K}$

| | GWAS 1 | GWAS 2 | ... | ... | ... | ... | ... | GWAS K |
|---|---|---|---|---|---|---|---|---|
| Locus 1 | | | | | | | | |
| | | | | | | | | |
| Locus $l_1$ | 1 | 1 | 1 | 1 | 1 | | | |
| Locus $l_2$ | 1 | 1 | 1 | 1 | 1 | | | |
| ⋮ | | | | | | | | |
| Locus $l_3$ | | | | | | 1 | 1 | 1 |
| Locus $l_4$ | | | | | | 1 | 1 | 1 |
| ⋮ | | | | | | | | |
| | | | | | | | | |

$\boldsymbol{\eta} \in \{0,1\}^{L \times K}$

| | GWAS 1 | GWAS 2 | ... | ... | ... | ... | ... | GWAS K |
|---|---|---|---|---|---|---|---|---|
| Locus 1 | | | | | | | | |
| | | | | | | 1 | | |
| Locus $l_1$ | 1 | 1 | 1 | 1 | 0 | | | |
| Locus $l_2$ | 1 | 1 | 1 | 1 | 1 | | | |
| | | | | 1 | | | | |
| Locus $l_3$ | | | | | 0 | 1 | 1 | 1 |
| Locus $l_4$ | | | | | 1 | 1 | 1 | 1 |
| ⋮ | | | | | | | | |
| | | | | | | | | |

**Fig. 2.** Left panel: illustration of the low-rank structure of $\boldsymbol{\eta}$ in an ideal case. Right panel: the association pattern after a small perturbation in the ideal case. Clearly, the rank of $\boldsymbol{\eta}$ increases dramatically, which motivates us to impose a low-rank structure on $\boldsymbol{\eta}$ via the logit link, as shown in Figure 1

GWAS should be analyzed separately. Our model (7) includes this as a special case with $\Pr(\eta_{lk}=1|\mathbf{X},\mathbf{x}_0)=\frac{1}{1+\exp(-x_{0k})}=\pi_k$ by setting $\mathbf{X}=0$. In this case, the prior of locus $l$ associated with phenotype $k$ is constant with respect to locus $l$ and depends only on $k$. As the pattern of pleiotropy becomes increasingly complex, the rank of $\mathbf{X}$ is allowed to increase to capture the induced correlation structure. The prior of locus $l$ associated with phenotype $k$ becomes $\pi_{lk}=\frac{1}{1+\exp(-X_{lk}-x_{0k})}$, which depends on both $l$ and $k$, indicating that its prior is locus-specific. As we shall see below, the estimation of $X_{lk}$ indeed borrows information from all $K$ GWAS, which is how we incorporate pleiotropy into our model.

More specifically, we consider the most commonly used norm, i.e. the nuclear norm of $\mathbf{X}$, to regularize its rank (Candès and Recht, 2009; Yang *et al.*, 2013; Zhou *et al.*, 2015). We denote it as $||\mathbf{X}||_*=\sum_{i=1}^{r}\sigma_i$, where $r$ is the rank of matrix $\mathbf{X}$ and $\sigma_i$ is its $i$th singular value. Let $\mathbf{\Theta}=\{\mathbf{X},\mathbf{x}_0\}$ be the collection of model parameters. The regularized incomplete-data log-likelihood of model (6) can then be written as

$$\begin{aligned}\ell^{\mathrm{p}}(\mathbf{\Theta}) = \sum_{k=1}^{K}\sum_{l=1}^{L}\log\Bigg\{&(1-\pi_{lk})\mathcal{N}(0,\mathbf{\Sigma}_l)\\ &+\pi_{lk}\left[\sum_{j=1}^{M_l}\frac{1}{M_l}\mathcal{N}\big(\mathbf{\Sigma}_l[\boldsymbol{\lambda}_{lk}\circ\gamma_{lk}(j=1)],\mathbf{\Sigma}_l\big)\right]\Bigg\}-\kappa||\mathbf{X}||_*,\end{aligned}\tag{8}$$

where $\pi_{lk}=\Pr(\eta_{lk}=1|\mathbf{X},\mathbf{x}_0)=\frac{1}{1+\exp(-X_{lk}-x_{0k})}$, and $\kappa$ is the regularization parameter. Clearly, when $\kappa\to\infty$, the maximizer of log-likelihood (8), $\widehat{\mathbf{\Theta}}(\infty)$, gives $\mathbf{X}=0$ that makes LLR equivalent with the separate analysis, i.e. the separate analysis is a special case of LLR. As $\kappa$ decreases, $\widehat{\mathbf{\Theta}}(\kappa)$ produces a low-rank structure of $\mathbf{X}$ that naturally incorporates pleiotropy information. The tuning parameter $\kappa$ allows our model to adapt to pleiotropy patterns with different degrees of complexity.

## 2.3 Algorithm

The direct maximization of log-likelihood (8) is not an easy task, and we can instead consider using the standard EM algorithm to estimate model parameters $\mathbf{\Theta}=\{\mathbf{X},\mathbf{x}_0\}$. As we shall see, however, the standard EM algorithm is computationally expensive when regularization parameter $\kappa$ needs to be tuned for model selection. To overcome this challenge, we have developed a novel EM-path algorithm that greatly reduces the computational cost and facilitates model selection.

### *2.3.1* Standard EM algorithm

Let $\mathbf{\Theta}^{(t)}=\{\mathbf{X}^{(t)},\mathbf{x}_0^{(t)}\}$ denote the estimated parameters at the $t$th EM iteration.

**E-step:** Consider the complete-data log-likelihood of model (5)

$$\begin{aligned}&\ell_c(\mathbf{Z},\gamma,\boldsymbol{\eta};\mathbf{\Theta})\\ =&\sum_{k=1}^{K}\sum_{l=1}^{L}\Bigg\{(1-\eta_{lk})\big[\log(1-\pi_{lk})+\log\Pr(\mathbf{Z}_{lk}|\gamma_{lk}=0)\big]\\ &+\eta_{lk}\left[\log\pi_{lk}+\sum_{j=1}^{M_l}\mathbb{I}(\gamma_{lk}(j=1))\left(\log\frac{1}{M_l}\right.\right.\\ &\left.\left.+\log\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1))\right)\right].\end{aligned}\tag{9}$$

We can calculate the $Q$ function as

$$Q\Big(\mathbf{\Theta};\mathbf{\Theta}^{(t)}\Big)=\mathbb{E}_{\mathbf{\Theta}^{(t)}}\{\ell_c(\mathbf{Z},\gamma,\boldsymbol{\eta};\mathbf{\Theta})|\mathbf{Z}\}-\kappa||\mathbf{X}||_*,$$

where the expectation is taken w.r.t. $\boldsymbol{\eta}$ and $\gamma$ given current estimated parameter $\mathbf{\Theta}$ and data $\mathbf{Z}$. Thus, the $Q$ function can be further written as

$$\begin{aligned}&Q\Big(\mathbf{\Theta};\mathbf{\Theta}^{(t)}\Big)\\ =&\sum_{k=1}^{K}\sum_{l=1}^{L}\Bigg\{\mathbb{E}_{\mathbf{\Theta}^{(t)}}[1-\eta_{lk}|\mathbf{Z}][\log(1-\pi_{lk})+\log\Pr(\mathbf{Z}_{lk}|\gamma_{lk}=0)]\\ &+\mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}|\mathbf{Z}]\log\pi_{lk}+\left[\sum_{j=1}^{M_l}\mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}\mathbb{I}(\gamma_{lk}(j=1))|\mathbf{Z}]\left(\log\frac{1}{M_l}\right.\right.\\ &\left.\left.+\log\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1))\right)\right]-\kappa||\mathbf{X}||_*,\end{aligned}\tag{10}$$

where

$$\begin{aligned}&\mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}|\mathbf{Z}]\\ =&\Pr(\eta_{lk}=1|\mathbf{Z};\mathbf{\Theta}^{(t)})\\ =&\frac{\pi_{lk}^{(t)}\sum_{j=1}^{M_l}\left[\frac{1}{M_l}\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1)\right]}{\pi_{lk}^{(t)}\sum_{j=1}^{M_l}\left[\frac{1}{M_l}\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1)\right]+(1-\pi_{jk}^{(t)})\Pr(\mathbf{Z}_{lk}|\gamma_{lk}=0)},\\ &\mathbb{E}_{\mathbf{\Theta}^{(t)}}\left[1-\eta_{lk}^{(t)}|\mathbf{Z}\right]=1-\mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}|\mathbf{Z}],\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_{\mathbf{\Theta}^{(t)}}(\eta_{lk}\mathbb{I}[\gamma_{lk}(j=1)|\mathbf{Z}])\\ =&\Pr\Big(\eta_{lk}=1,\mathbb{I}[\gamma_{lk}(j=1)]|\mathbf{Z};\mathbf{\Theta}^{(t)}\Big)\\ =&\Pr\Big(\eta_{lk}=1|\mathbf{Z};\mathbf{\Theta}^{(t)}\Big)\Pr\Big(\gamma_{lk}(j=1)|\eta_{lk}=1,\mathbf{Z};\mathbf{\Theta}^{(t)}\Big)\\ =&\Pr\Big(\eta_{lk}=1|\mathbf{Z};\mathbf{\Theta}^{(t)}\Big)\frac{\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1))\frac{1}{M_l}}{\sum_{j=1}^{M_l}\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1))\frac{1}{M_l}}.\\ =&\Pr\Big(\eta_{lk}=1|\mathbf{Z};\mathbf{\Theta}^{(t)}\Big)\frac{\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1))}{\sum_{j=1}^{M_l}\Pr(\mathbf{Z}_{lk}|\gamma_{lk}(j=1))}.\end{aligned}$$

**M-step:** We need to consider only the terms involving model parameters $\mathbf{\Theta}=\{\mathbf{X},\mathbf{x}_0\}$ in the $Q$ function. Therefore, the optimization problem can be written as

$$\max_{\mathbf{X},\mathbf{x}_0}\sum_{k=1}^{K}\sum_{l=1}^{L}\Big\{\mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}|\mathbf{Z}]\log\pi_{lk}+\mathbb{E}_{\mathbf{\Theta}^{(t)}}[1-\eta_{lk}|\mathbf{Z}]\log(1-\pi_{lk})\Big\}-\kappa||\mathbf{X}||_*,\tag{11}$$

where $\pi_{lk}$ is a function of $X_{lk}$,$x_{0k}$, as given by Equation (7). It turns out that the optimization problem (11) is actually the log-likelihood of a logistic regression problem with nuclear norm regularization, and fast algorithms are available for such convex optimization (Zhou *et al.*, 2015). Therefore, for a given regularization parameter $\kappa$, the standard EM algorithm repeats the foregoing E-step and M-step until convergence to obtain $\widehat{\mathbf{\Theta}}(\kappa)$. However, the computational cost is very high because the M-step involves solving a regularized logistic regression problem, which requires singular value decomposition (SVD) to be performed numerous times. Because the EM algorithm often requires hundreds of iterations, we would need to solve the large-scale logistic regression problem hundreds of times. Furthermore, to select an optimal value for $\kappa$, the entire EM would need to be invoked multiple times, e.g. the $\kappa$ sequence $\{\kappa_1,\kappa_2,\ldots,\kappa_{100}\}$ would need to be solved.

#### *2.3.2* Efficient EM-path algorithm

We propose the integration of the EM and boosting algorithms (Friedman *et al.*, 2000; Friedman, 2001) to address the computational challenges discussed above. Our new algorithm is motivated by integration of the following facts of EM and boosting algorithms: (i) the convergence of EM is guaranteed as long as the ascent condition of the $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)})$ function holds, i.e. $Q(\mathbf{\Theta}^{(t+1)}|\mathbf{\Theta}^{(t)}) \geq Q(\mathbf{\Theta}^{(t)}|\mathbf{\Theta}^{(t)})$; (ii) boosting algorithms can be viewed as a type of gradient method (Friedman, 2001, 2012; Hastie *et al.*, 2007), e.g. the steepest descent method (Tibshirani, 2015); and (iii) the regularized solution path (e.g. the $L_1$ norm and nuclear norm) can be closely approximated by the boosting path whose statistical properties are also guaranteed (Hastie *et al.*, 2009; Tibshirani, 2015). In other words, the gradient view of boosting ensures that the ascent condition holds during the EM iterations, and the boosting updates generate the similar regularized path without tuning regularization parameter $\kappa$. Therefore, we only need to run EM once to generate all of the solution paths. We thus refer to our proposed algorithm as the EM-path algorithm.

In more details, the EM-path algorithm optimizing log-likelihood (8) is given as follows. We initialize $\mathbf{x}_0^{(0)}$ using separate estimates on each phenotype $k$ and $\mathbf{X}^{(0)} = 0$. The algorithm then simply repeats the following $\mathbb{E}$-step and $\mathbb{M}$-step for $t = 1, 2, \ldots$ until convergence.

**E-step:** the same as the E-step in the standard EM algorithm given in Section 2.3.1.

**M-step:** Let $f(\mathbf{X}, \mathbf{x}_0)$ be part of the logistic log-likelihood function (without the regularization term) in the $Q$ function:

$$f(\mathbf{X}, \mathbf{x}_0) = \sum_{k=1}^{K}\sum_{l=1}^{L}\Big\{\mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}|\mathbf{Z}]\log\pi_{lk} \\ +\mathbb{E}_{\mathbf{\Theta}^{(t)}}[1-\eta_{lk}|\mathbf{Z}]\log(1-\pi_{lk})\Big\}.$$

We can obtain the partial derivatives w.r.t. $\mathbf{X}$ and $\mathbf{x}_0$ as

$$\mathbf{G} = \frac{\partial f(\mathbf{X}, \mathbf{x}_0)}{\partial \mathbf{X}} = [G_{lk}] \in R^{L\times K},$$
$$\mathbf{g} = \frac{\partial f(\mathbf{X}, \mathbf{x}_0)}{\partial \mathbf{x}_0} = [g_k] \in R^{K\times 1},$$

where

$$\begin{aligned} G_{lk} = &\quad \mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}|\mathbf{Z}] - \pi_{lk}, \\ g_k = &\quad \sum_{l=1}^{L}\mathbb{E}_{\mathbf{\Theta}^{(t)}}[\eta_{lk}|\mathbf{Z}] - \pi_{lk}, \end{aligned}$$

and then update $\mathbf{x}_0$ and $\mathbf{X}$ as

$$\mathbf{x}_0^{(t+1)} = \mathbf{x}_0^{(t)} + \epsilon\cdot\mathrm{sign}\left(\mathbf{g}^{(t)}\right), \mathbf{X}^{(t+1)} = \mathbf{X}^{(t)} + \epsilon\cdot\mathbf{u}\mathbf{v}^{\top}, \qquad (12)$$

where $\mathbf{u}$ and $\mathbf{v}$ are the leading singular vectors of $\mathbf{G}^{(t)}$, which can be efficiently obtained by the power method without SVD, and $\epsilon$ is a fixed small step size (e.g. 0.01). In fact, this update can be characterized as a steepest ascent w.r.t. the nuclear norm. A more detailed derivation is provided in the Supplementary document. As can be seen, the proposed EM-path algorithm needs to run only once to generate the solution path. In addition, there is no need to explicitly tune the regularization parameter $\kappa$, and the update in the M-step is much cheaper in computational terms than that in the standard EM algorithm.

### 2.4 Model selection and inference

#### *2.4.1* Model selection

We search for optimal number of EM steps using $V$-fold cross-validation [$V = 5$ in the numerical study (Hastie *et al.*, 2009)]. Briefly, we randomly partition $L\times K$ entries in $\boldsymbol{\eta}$ into five groups with roughly equal sizes, $\Omega_1,\ldots,\Omega_5$, such that $\Omega_1\cup\ldots\cup\Omega_5 = \Omega$ and $\Omega_1\cap\ldots\cap\Omega_5 = \varnothing$. We choose four of them as the training set, and the remaining one as the testing set. We then evaluate incomplete-data likelihood (6) of each iteration on the testing set. The optimal number of iterations is chosen to maximize the testing likelihood averaged in cross-validation. More details on cross-validation are given in the Supplementary document.

#### *2.4.2* Statistical inference

After the parameters in the LLR model are estimated, SNPs can be prioritized on the basis of their local false discovery rates (FDRs) (Efron, 2010), i.e. the lower the local FDR, the higher the priority. The estimated FDR of the $l$th locus for phenotype $k$ is given as

$$\widehat{\mathrm{fdr}}_{lk}^{\text{locus}} = 1 - \Pr(\eta_{lk} = 1|\mathbf{Z};\widehat{\mathbf{\Theta}}). \qquad (13)$$

Similarly, we can evaluate the local FDR at the SNP level by

$$\widehat{\mathrm{fdr}}_{jlk}^{\text{SNP}} = 1 - \Pr(\eta_{lk} = 1, \gamma_{jlk} = 1|\mathbf{Z};\widehat{\mathbf{\Theta}}). \qquad (14)$$

Clearly, these probabilities are naturally provided at the E-step after the convergence of the EM algorithm.

## 3 Results

### 3.1 Simulation

We designed our simulation based on the following thinking: Our model presented in Section 2 is design for the analysis of the summary statistics from multiple studies when the individual-level genotype data is not available for sharing. In the first scenario, we directly simulated summary statistics from the generative model to evaluate the performance of LLR, which we refer to as 'summary-statistic-level simulation'. Basically, the first scenario is used to validate our designed algorithm, model selection and inference. However, in real data analysis, the summary statistics are often obtained from individual level phenotype data with the corresponding phenotype. Therefore, in the second scenario, we mimics the real data analysis by first simulating the genotype data and then computing the phenotype of each sample using Eq. (16) in the main manuscript. Next, we compute the summary statistics from the simulated genotype and phenotype. We refer this simulation setting as 'individual-level simulation'. In each of the two scenarios, we varied some important parameters, such as heritability $h^2$, within-loci correlation $\rho$ and the number of loci $L$, to obtain compare the LLR's performance with that of three alternative methods, namely, GPA, GPA-Joint (which refers to the joint analysis of two GWASs using GPA) and PAINTOR.

### *3.1.1* Simulation settings

In both scenarios, we considered $K = 20$ studies with loci $L = 500$ or $L = 2000$. We used the autoregressive correlation structure $\mathbf{\Sigma}(\rho)$ to simulate the LD effects within a locus and varied $\rho$ from small to large (i.e. $\rho = 0.2$, 0.5 and 0.8) to evaluate the influence of LD. For a given locus, the number of SNPs was fixed at 20. The proportion of non-null loci for each study was fixed at 0.2, meaning that $\pi_{lk} = 0.2$ for all loci.

In the first scenario, the summary statistics were generated as follows. We first randomly generated the hidden status for each SNP and each locus. More specifically, we generated an $L \times r$ matrix $\mathbf{R}$ and an $r \times K$ matrix $\mathbf{B}$ to form the low-rank matrix $\mathbf{X} = \mathbf{RB}$, where we set $r = 2$. The entries in $\mathbf{R}$ were independently drawn from the standard normal distribution, and $\mathbf{B}$ was designed to partition the $K$ studies into two groups, i.e.

$$\mathbf{B} = \left( \begin{array}{cc} 1 \quad \dots \quad 1 & 1 \quad \dots \quad 1 \\ \underbrace{-1 \quad \dots \quad -1}_{10 \ \text{columns}} & \underbrace{1 \quad \dots \quad 1}_{10 \ \text{columns}} \end{array} \right).$$

Using such a simulation setting can generate moderate within-group pleiotropic effects. The intercept term $\mathbf{x}_0$ was set to control the proportion of non-null loci to around 20%. Then, latent variable $\boldsymbol{\eta}$ was generated by $\Pr(\eta_{lk} = 1) = \frac{\exp(x_{0k} + X_{lk})}{1 + \exp(x_{0k} + X_{lk})}$. Next, we randomly chose a causal SNP within locus $l$ in the $k$th trait if $\eta_{lk} = 1$ and assigned its effect size as follows.

$$\beta_{jlk} = \begin{cases} \mathcal{N}(0, \sigma_\beta^2), & \text{if } \gamma_{ljk} = 1 \\ 0, & \text{if } \gamma_{jlk} = 0. \end{cases} \tag{15}$$

Next, we generated $Z$-scores according to distribution (2), with NCP $\lambda_{jlk} = \frac{\beta_{jlk}\sqrt{n_k}}{\sigma_{e_k}}$. Here, we used the same $\sigma_e^2$ for all $\sigma_{e_k}^2$s. Heritability $h^2 = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_e^2}$ was controlled at different values by adjusting ratio $\sigma_e^2/\sigma_\beta^2$, e.g. 0.3, 0.4 and 0.5. In this scenario, the effective sample size $n_k$ was set to 5000. To mimic the situation that the true LD structure $\mathbf{\Sigma}$ is unknown in practice, we simulated an additional dataset with sample size $n_{\text{ref}} = 400$ from the multivariate normal distribution with true covariance $\mathbf{\Sigma}(\rho)$. We used this dataset as a reference panel to estimate the LD structure, and then plugged it into our LLR model.

In the second simulation scenario, we gauged LLR's performance relative to the three alternative approaches using simulated raw genotype data. We first generated the minor allele frequencies for all of the SNPs from a uniform distribution $\mathcal{U}(0.05, 0.5)$. We then sampled data matrix $\mathbf{W}$ from the multivariate normal distribution with $\mathbf{\Sigma}(\rho)$, and categorized $\mathbf{W}$ into genotype data 0, 1, 2 according to the Hardy-Weinberg principle, denoted as $\mathbf{G}$. After generating the raw genotype data, we used the same strategy to simulate effect sizes $\beta_{jlk}$ and obtained quantitative phenotypes as

$$\mathbf{y}_k = \mathbf{G}\boldsymbol{\beta}_\text{k} + \epsilon_k, \tag{16}$$

where $\mathbf{G} \in 0, 1, 2^{n \times M}$, $\boldsymbol{\beta}_\text{k} = [\beta_{jlk}] \in R^{M \times 1}$, and $\epsilon_k \sim \mathcal{N}(0, \sigma_{e_k}^2 \mathbf{I})$. By adjusting $\sigma_{e_k}^2$, we controlled heritability at 0.3, 0.4 and 0.5.

### *3.1.2* Results

With the simulated datasets, GPA and PAINTOR were used to perform analysis of a single GWAS. The performance of these two approaches served as the baseline, and the difference between them elucidated the role of accounting for LD effects. We did not use trans-PAINTOR (Kichaev and Pasaniuc, 2015) to simultaneously analyze all of the studies because it assumes all causal variants to be identical across studies. That assumption is reasonable when analyzing the same phenotype across different populations, but is not appropriate in our setting, i.e. analysis of different phenotypes in the same population. GPA-Joint was applied to analyze two studies within the same group. The comparison between GPA and GPA-Joint provided evidence of the role pleiotropy plays. LLR was applicable to all 20 studies considered in the simulation. The comparison between LLR and GPA-Joint allowed us to evaluate the gain in power achieved by the joint analysis of more than two studies.

Figure 3 shows performance comparison of all four methods for SNP prioritization in summary-statistic-level simulation (i.e. the first scenario) with the number of loci $L = 500$ (left panel) and $L = 2000$ (right panel). In terms of risk variant ranking, LLR consistently outperformed PAINTOR because of its ability to simultaneously integrate information from multiple studies. As correlation $\rho$ increases, GPA and GPA-Joint performed worse because GPA assumed independence among SNPs. Supplementary Figures S1 and S2 in the Supplementary document report AUC and FDR of all four methods for SNP prioritization in summary-statistic-level simulation with the number of loci $L = 500$ and $L = 2000$. It can be seen that LLR still outperforms the other methods in terms of the AUC measure. Note that the performances of all the methods degraded with $L$ increasing from 500 to 2000. This is because non-null proportion of loci

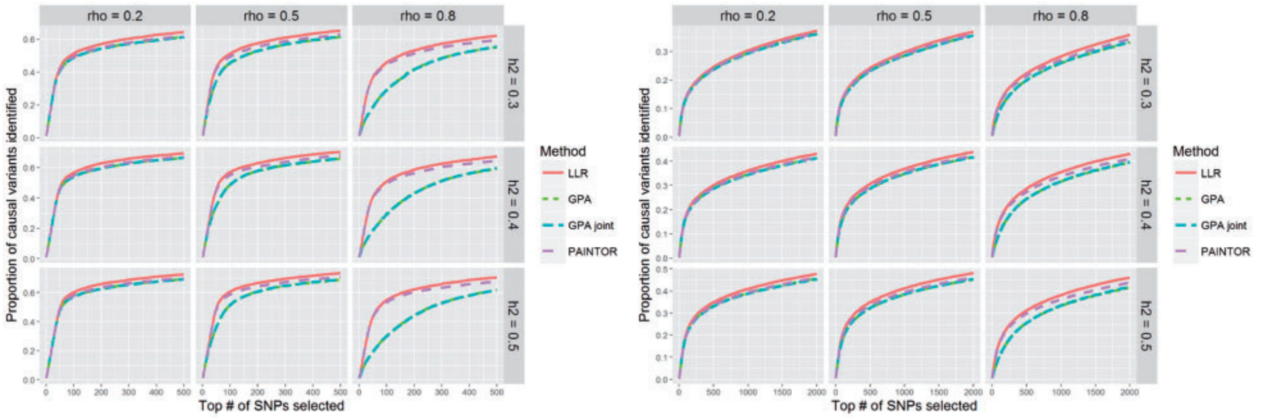


**Fig. 3.** Performance comparison of LLR, GPA, GPA-Joint and PAINTOR (summary-statistic-level simulation) with the number of loci $L = 500$ (left panel) and $L = 2000$ (right panel). In each panel, the four methods are tested with heritability $h^2 = 0.3, 0.4$, and 0.5, and within-locus correlation $\rho = 0.2, 0.5$, and 0.8. The results in each setting are summarized from 50 replications

remains fixed in each setting and the average signal strength of a locus becomes weaker. The results from individual-level simulation shown in Supplementary Figures S3, S4 and S5 have similar patterns, indicating that LLR's performance remains stable in both simulation scenarios with various configurations of correlation $\rho$ and heritability $h^2$.

The foregoing analysis demonstrates that LLR can make effective use of pleiotropic information. Next, we also evaluated its performance in the absence of pleiotropy using the given $K$ studies. We simulated the summary statistics using the aforementioned procedure with $\mathbf{X} = 0$, leading to independence among the $K$ studies. We then ran LLR on the simulated data and the results are presented in Supplementary Figures S6 and S7 in the Supplementary document. The two figures show that LLR performs essentially the same in the separate analyses, which is a desirable property. LLR works well in this setting because it includes separate analysis as a special case. In the absence of pleiotropy, there is no signal driving LLR from its origin (i.e. $\mathbf{X} = 0$ and $\mathbf{x}_0$ is initialized by the separate analysis) because the adaptive model selection strategy described in Section 2.4.1 prefers remaining at origin. In the presence of pleiotropy, LLR naturally generalizes the separate analyses, as guided by the EM-path algorithm. The model selection strategy applied in LLR basically prevents it from overfitting.

Although we assumed 'one causal SNP per locus', LLR still could identify more risk loci than risk SNPs at the same FDR cutoff since $\Pr(\eta_{lk} = 1|\mathbf{Z}; \widehat{\boldsymbol{\Theta}}) = \sum_{\gamma_{jlk}\in\{0,1\}} \Pr(\eta_{lk} = 1, \gamma_{jlk}|\mathbf{Z}; \widehat{\boldsymbol{\Theta}})$. As identification of risk SNPs often remains uncertain due to the polygenicity, LLR effectively takes the uncertainty at the SNP level into account at the locus level by the marginalization over $\gamma_{jlk}$, and further improves its power by using a low-rank structure to borrow information across loci in multiple studies. To verify this advantage, we also evaluated LLR's performance at the locus level. The results are given in Supplementary Figures S8 and S9 in the Supplementary document.

It should be noted that both LLR and PAINTOR involve the NCP threshold parameter, which is a critical parameter for FDR control. Our experimental results indicated that the NCP threshold at 3.7 adopted by PAINTOR may lead to an inflated FDR. To make the FDR of LLR controlled at the nominal level, we increased this threshold from 3.7 to 5.3 (Supplementary Figs S14). Empirical evidence from extensive simulation studies (Supplementary Figs S1–S9 and S15–S16) suggests that the NCP threshold at 5.3 can offer a satisfactory FDR control at nominal level 0.1. To see the importance of the NCP threshold on FDR control, we evaluated LLR's FDR with the true NCP. The results are reported in Supplementary Figures S10–S13 in the Supplementary document, in which the results of the separate analyses are also presented as a reference. We can observe from these two figures that the FDR is well controlled in LLR when the true NCP is used. Clearly, these experimental results imply that the method of handling NCP in both LLR and PAINTOR constitutes their major limitation. The issue thus deserves careful investigation in future work. A possible improvement would be to model the estimated effect sizes $\hat{\beta}_{jlk}$ and their standard errors $\mathrm{se}(\hat{\beta}_{jlk})$ rather than relying on the $Z$-scores (Stephens, 2017).

Regarding LLR's computing efficiency, we compared the EM-path algorithm with the standard EM algorithm. Supplementary Table S1 in the Supplementary document presents the results of this comparison under different model settings of $L$ and $\rho$. The solution paths of the two algorithms are depicted in Supplementary Figure S17 in the Supplementary document. Although their solution paths are very similar, the EM-path algorithm runs about four time faster than the standard EM regularized algorithm. This computing improvement greatly facilitates LLR's use in real large-scale genomic data applications.

## 3.2 Data analysis of 18 GWAS

We analyzed the data of 18 GWAS on multiple sclerosis (MS) (International Multiple Sclerosis Genetics Consortium, 2013), psychiatric diseases [i.e. bipolar disorder (BPD), major depression disorder (MDD), schizophrenia (SCZ) (Cross Disorder Group of the Psychiatric Genomics Consortium, 2013), attention-deficit/hyperacitivity disorder (ADHD) (Neale *et al.*, 2010)], neurodegenerative disorder [i.e. Alzheimer's disease (IGAP) (Lambert *et al.*, 2013), anorexia nervosa (GCAN) (Boraska *et al.*, 2014)], type 1 diabetes (T1D) (Cooper *et al.*, 2008), anthropometric traits [body mass index (BMI) for men and for women (Randall *et al.*, 2013)], cardiovascular disease [coronary artery disease (cardiogram) (Deloukas *et al.*, 2012)], glycemic traits [fasting glucose level (Manning *et al.*, 2012)], metabolic traits [docosahexaenoic acid (DHA) (Lemaitre *et al.*, 2011)], packed cell volume (PCV) (van der Harst *et al.*, 2012), smoking behavior [cigarettes per day (TAG cpd) (Tobacco and Genetics Consortium, 2010)] and human subcortical brain structures [caudate nucleus and putamen (Hibar *et al.*, 2015)]. We collected the publicly available summary statistics from these 18 GWASa from either dbGaP or consortium websites. Details (including download links) of the datasets are provided in Supplementary Table S2 in the Supplementary document. The summary statistics of some but not all of the GWAS were imputed. We matched the SNPs in all 18 GWAS datasets, for a total of 284 551 SNPs. We used the results from LDETECT (Berisa and Pickrell, 2016) to partition the entire genome into nearly independent 1703 loci, where 379 samples from European ancestry in the 1000 Genome Project were used as reference panel to estimate LD struture of those SNPs. Then we performed analysis using LLR on a desktop PC with 2.40 GHz CPU and 4GB RAM. The running time was around 5 min.

The Manhattan plots of the LLR analysis results are shown in Supplementary Figure S18 in the Supplementary document. We also conducted separate analyses on all 18 datasets, with the Manhattan plots reported in Supplementary Figure S19 of that document. In comparing the two figures, it can be seen that LLR identified more risk variants than the separate analyses. As explained in our discussion of the simulation study, LLR is a generalization of separate analyses by incorporating pleiotropy information, and thus all variants identified in the former will also be discovered by LLR. Therefore, we compared LLR with separate analysis to evaluate the gain of power. Supplementary Tables S3 and S4 in the Supplementary document summarize these detailed results on the locus level and SNP level, respectively.

There are a number of loci that are significantly associated with several studied phenotypes. For example, three loci are shared by SCZ, MS, PCV and T1D. All these three loci reside in the major histocompatibility complex (MHC) region of Chromosome 6, which harbours many genes whose primary function in regulating immune responsiveness to infection is to present foreign antigens to cytotoxic T lymphocytes (CTLs) and T helper cells. The first locus starting from 25 626 177 and ending at 26 735 343 contains 3 genes in SLC17 family, 19 histone H1 genes and 6 Butyrophilin (BTN) genes. A recent study indicates that common polymorphisms within the SLC17 family are associated with schizophrenia (Shi *et al.*, 2009). The SLC17 gene family consists of the three vesicular glutamate transporters and glutamate has been identified as an important risk factor of disease progression in multiple sclerosis (MS) in many studies (Frigo *et al.*, 2012; Groom *et al.*, 2003; Stojanovic *et al.*, 2014).

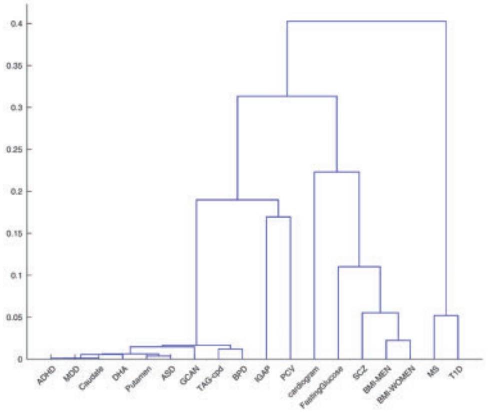


**Fig. 4.** The relationships of all 18 datasets using the coordinates derived from the first two right singular vectors of the estimated low-rank matrix **X**

The second locus starting from 26 895 127 and ending at 27 986 819 contains the gene ZNF184, which is closely related with schizophrenia (Shi *et al.*, 2009). The third locus starting from 27 991 166 and ending at 28 439 211 contains the gene PGBD1, which is linked with Alzheimer disease in recent studies (Feulner *et al.*, 2010; Guerreiro *et al.*, 2012). We believe that our proposed method can be an effective tool for the analysis of pleiotropy.

We took the first two right singular vectors of the estimated low-rank matrix **X** and used them as the coordinates of each study to generate a dendrogram plot of the hierarchical binary cluster tree in Figure 4. The cluster tree of the 18 datasets in Figure 4 conforms to many previously reported results. The first cluster involves seven datasets (TPG-cpd, DHA, ASD, ADHD, Putamen, Caudate and MDD), all of which are related to brain function. The connection between the three psychiatric disorders, ASD, ADHD and MDD, and brain function is straightforward. The other two datasets, Caudate and Putamen, come from a study investigating how genetic variants influence human subcortical brain structures, which primarily concern putamen and caudate nucleus volumes (Hibar *et al.*, 2015). The TPG-cpd dataset is from a study on the relationship between genetic factors and smoking behavior (Tobacco and Genetics Consortium, 2010). Many brain research studies have provided evidence on the effects of nicotine and its derivatives on brain function (Benwell *et al.*, 1988; Gallinat *et al.*, 2006; Janes *et al.*, 2010; Pentel *et al.*, 2000). The DHA dataset comes from a study on how common variants influence the plasma phospholipid level of n-3 fatty acids. It is well known that n-3 fatty acids provide DHA for the growth and function of nervous tissue. Reduced DHA is associated with impairments in cognitive and behavioral performance, the effects of which are particularly important during brain development (Innis, 2005, 2007). There is also considerable supporting evidence with regard to the other clusters. For example, both T1D and MS contribute substantially to the autoimmune disease burden in young adults, and the individual and familial co-occurrence of the two diseases is widely reported (Henderson *et al.*, 2000; Winer *et al.*, 2001). Finally, with regard to the relationship between BMI and SCZ, many studies have shown that individuals with SCZ to be more obese than those without, and certain genetic factors, such as the $5-HT_{2A}$ and $5-HT_{2C}$ receptors, have been reported to induce additive genetic effects on weight gain in SCZ patients (Allison *et al.*, 1999; Coodin, 2001; Ujike *et al.*, 2008).

## 4 Conclusion

Polygenicity renders the identification of risk variants in GWAS a challenging task. However, there is accumulating evidence to suggest that complex phenotypes can share common genetic bases, offering a new paradigm for exploring existing GWAS data resources. In this article, we propose LLR as a new statistical approach to prioritizing risk variants using the pleiotropy across multiple related studies. Compared with such existing approaches as PAINTOR and GPA, LLR demonstrates consistently reliable performance. The development of the EM-path algorithm allows LLR to efficiently handle the analysis of large-scale genomic data. These merits make LLR an attractive and effective tool for the integrative analysis of multiple GWAS data.

## Acknowledgement

We thank the editor and two anonymous reviewers for their constructive comments which have helped us to improve our manuscript.

## Funding

This work was supported in part by grant No. 61501389 from National Science Funding of China, grants No. 22302815, No. 12316116 and No. 12202114 from the Hong Kong Research Grant Council, and grants FRG2/14-15/069, FRG2/14-15/077 and FRG2/15-16/011 from Hong Kong Baptist University, and Duke-NUS Graduate Medical School WBS: R-913-200-098-263 and MOE2016-T2-2-029 from Ministry of Education, Singapore.

*Conflict of Interes't*: none declared.

## References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Allison,D.B. *et al.* (1999) The distribution of body mass index among individuals with and without schizophrenia. *J. Clin. Psychiatry*, **60**, 215–220.

Benwell,M.E. *et al.* (1988) Evidence that tobacco smoking increases the density of (-)-[3h] nicotine binding sites in human brain. *J. Neurochem.*, **50**, 1243–1247.

Berisa,T. and Pickrell,J.K. (2016) Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics (Oxford, England)*, **32**, 283.

Boraska,V. *et al.* (2014) A genome-wide association study of anorexia nervosa. *Mol. Psychiatry*, **19**, 1085–1094.

Candès,E.J. and Recht,B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.

Cantor,R.M. *et al.* (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, **86**, 6–22.

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

Coodin,S. (2001) Body mass index in persons with schizophrenia. *Can. J. Psychiatry*, **46**, 549–555.

Cooper,J.D. *et al.* (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.*, **40**, 1399–1401.

Cotsapas,C. *et al.* (2011) Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.*, **7**, e1002254.

Cross Disorder Group of the Psychiatric Genomics Consortium. (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nat. Genet.*, **45**, 984–994.

Deloukas,P. *et al.* (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.*, **45**, 25–33.

Efron,B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge.

Feulner,T. *et al.* (2010) Examination of the current top candidate genes for ad in a genome-wide association study. *Mol. Psychiatry*, **15**, 756–766.

Friedman,J. *et al.* (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.*, **28**, 337–407.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.

Friedman,J.H. (2012) Fast sparse regression and classification. *Int. J. Forecast.*, **28**, 722–738.

Frigo,M., G., Cogo,M., L., Fusco,M. *et al.* (2012) Glutamate and multiple sclerosis. *Curr. Med. Chem.*, **19**, 1295–1299.

Gallinat,J. *et al.* (2006) Smoking and structural brain deficits: a volumetric MR investigation. *Eur. J. Neurosci.*, **24**, 1744–1750.

Gamazon,E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.

Groom,A.J. *et al.* (2003) Multiple sclerosis and glutamate. *Ann. N. Y. Acad. Sci.*, **993**, 229–275.

Guerreiro,R.J. *et al.* (2012) The genetic architecture of alzheimer‘s disease: beyond app, psens and apoe. *Neurobiol. Aging*, **33**, 437–456.

Hastie,T. *et al.* (2007) Forward stagewise regression and the monotone lasso. *Electron. J. Stat.*, **1**, 1–29.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning*, 2nd edn. Springer, New York.

Henderson,R.D. *et al.* (2000) The occurrence of autoimmune diseases in patients with multiple sclerosis and their families. *J. Clin. Neurosci.*, **7**, 434–437.

Hibar,D.P. *et al.* (2015) Common genetic variants influence human subcortical brain structures. *Nature*, **520**, 224–229.

Hormozdiari,F. *et al.* (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.

Innis,S. (2005) Essential fatty acid transfer and fetal development. *Placenta*, **26**, S70–S75.

Innis,S.M. (2007) Dietary (n-3) fatty acids and brain development. *J. Nutrit.*, **137**, 855–859.

International Multiple Sclerosis Genetics Consortium (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.*, **45**, 1353–1360.

Janes,A.C. *et al.* (2010) Brain reactivity to smoking cues prior to smoking cessation predicts ability to maintain tobacco abstinence. *Biol. Psychiatry*, **67**, 722–729.

Kichaev,G. and Pasaniuc,B. (2015) Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.*, **97**, 260–271.

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

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.

Lemaitre,R.N. *et al.* (2011) Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the charge consortium. *PLoS Genet.*, **7**, e1002193–e1002193.

Li,C. *et al.* (2014) Improving genetic risk prediction by leveraging pleiotropy. *Hum. Genet.*, **133**, 639–650.

Manning,A.K. *et al.* (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.

Neale,B.M. *et al.* (2010) Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J. Am. Acad. Child Adolesc. Psychiatry*, **49**, 884–897.

Pentel,P.R. *et al.* (2000) A nicotine conjugate vaccine reduces nicotine distribution to brain and attenuates its behavioral and cardiovascular effects in rats. *Pharmacol. Biochem. Behav.*, **65**, 191–198.

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.

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

Randall,J.C. *et al.* (2013) Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet.*, **9**, e1003500.

Segura,V. *et al.* (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.*, **44**, 825–830.

Shi,J. *et al.* (2009) Common variants on chromosome 6p22. 1 are associated with schizophrenia. *Nature*, **460**, 753–757.

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

Stephens,M. (2017) False discovery rates: a new deal. *Biostatistics*, **18**, 275–294.

Stephens,M. and Balding,D.J. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.*, **10**, 681–690.

Stojanovic,I.R. *et al.* (2014) The role of glutamate and its receptors in multiple sclerosis. *J. Neural Trans.*, **121**, 945–955.

Tibshirani,R.J. (2015) A general framework for fast stagewise algorithms. *J. Mach. Learn. Res.*, **16**, 2543–2588.

Tobacco and Genetics Consortium. (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.*, **42**, 441–447.

Ujike,H. *et al.* (2008) Multiple genetic factors in olanzapine-induced weight gain in schizophrenia patients. *J. Clin. Psychiatry*, **69**, 1416–1422.

van der Harst,P. *et al.* (2012) Seventy-five genetic loci influencing the human red blood cell. *Nature*, **492**, 369–375.

Van der Sluis,S. *et al.* (2015) MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics*, **31**, 1007–1015.

Visscher,P.M. and Yang,J. (2016) A plethora of pleiotropy across complex traits. *Nat. Genet.*, **48**, 707.

Visscher,P.M. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.

Wang,Q. *et al.* (2015) Pervasive pleiotropy between psychiatric disorders and immune disorders revealed by integrative analysis of multiple GWAS. *Hum. Genet.*, **134**, 1–15.

Welter,D. *et al.* (2014) The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Winer,S. *et al.* (2001) Type i diabetes and multiple sclerosis patients target islet plus central nervous system autoantigens; nonimmunized nonobese diabetic mice can develop autoimmune encephalitis. *J. Immunol.*, **166**, 2831–2841.

Yang,C. *et al.* (2013) Accounting for non-genetic factors by low-rank representation and sparse regression for eQTL mapping. *Bioinformatics*, **29**, 1026–1034.

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.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.

‘Zhou,X. and Stephens,M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407–409.

Zhou,X. *et al.* (2015) Low-rank modeling and its applications in image analysis. *ACM Computing Surveys (CSUR)*, **47**, 36.

Zhu,X. *et al.* (2015) Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.*, **96**, 21–36.