Bioinformatics, 35(19), 2019, 3693–3700
doi: 10.1093/bioinformatics/btz167
Advance Access Publication Date: 9 March 2019
Original Paper

OXFORD

Genetics and population analysis

VIMCO: variational inference for multiple correlated outcomes in genome-wide association studies

Xingjie Shi[1,2], Yuling Jiao[3], Yi Yang[4], Ching-Yu Cheng[2], Can Yang[5], Xinyi Lin[2,*] and Jin Liu[2,*]

[1]Department of Statistics, Nanjing University of Finance and Economics, Nanjing, China, [2]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, [3]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China, [4]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China and [5]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

*To whom correspondence should be addressed.
Associate Editor: Russell Schwartz

Received on July 26, 2018; revised on December 22, 2018; editorial decision on March 4, 2019; accepted on March 8, 2019

Abstract

Motivation: In genome-wide association studies (GWASs) where multiple correlated traits have been measured on participants, a joint analysis strategy, whereby the traits are analyzed jointly, can improve statistical power over a single-trait analysis strategy. There are two questions of interest to be addressed when conducting a joint GWAS analysis with multiple traits. The first question examines whether a genetic loci is significantly associated with any of the traits being tested. The second question focuses on identifying the specific trait(s) that is associated with the genetic loci. Since existing methods primarily focus on the first question, this article seeks to provide a complementary method that addresses the second question.

Results: We propose a novel method, Variational Inference for Multiple Correlated Outcomes (VIMCO) that focuses on identifying the specific trait that is associated with the genetic loci, when performing a joint GWAS analysis of multiple traits, while accounting for correlation among the multiple traits. We performed extensive numerical studies and also applied VIMCO to analyze two datasets. The numerical studies and real data analysis demonstrate that VIMCO improves statistical power over single-trait analysis strategies when the multiple traits are correlated and has comparable performance when the traits are not correlated.

Availability and implementation: The VIMCO software can be downloaded from: https://github.com/XingjieShi/VIMCO.

Contact: xinyi.cindy.lin@duke-nus.edu.sg or jin.liu@duke-nus.edu.sg

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

1 Introduction

Genome-wide association studies (GWASs) conducted in the last decade have provided valuable insights into the genetic architecture underlying complex traits. GWAS are generally performed by analyzing individual traits, although multiple related traits are often collected, and in some cases these traits may reflect a common condition (Kim *et al.*, 2009). In a GWAS where multiple traits have been measured on the study individuals, a joint analysis strategy whereby the multiple traits are analyzed jointly, offers improved

statistical power when compared with a single-trait analysis strategy, when a genetic variant is associated with one or more correlated phenotypes (Korte *et al.*, 2012; Solovieff *et al.*, 2013). A joint analysis strategy may also be useful when the different phenotypes characterize the same underlying trait.

There are typically two questions of interest to be addressed in a joint GWAS analysis of multiple traits. The first question addresses whether a genetic loci is significantly associated with any of the (correlated) phenotypes being tested. The second question focuses on the identification of the phenotype(s) that is/are associated with the genetic loci. When multiple continuous traits and genotype data are available from the same study individuals, a joint GWAS analysis that addresses the first question can be performed using linear mixed models based methods (Casale, 2016). Examples of linear mixed models based methods include the multi-trait mixed models proposed by Korte *et al.* (2012) and the multivariate linear mixed models (mvLMM) implemented in the Genome-wide Efficient Mixed Model Association (GEMMA) software (Zhou and Stephens, 2014). A central limitation of these methods is that they cannot address the second question which seeks to identify the tested trait(s) that is/are associated with the genetic loci. Multi-trait variable selection methods based on penalization have also been proposed in this context (Liu *et al.*, 2016; Rothman *et al.*, 2010). However, they cannot be applied to assess associations while controlling the error rate (Carbonetto *et al.*, 2012).

In this article, we propose a novel method, Variational Inference for Multiple Correlated Outcomes (VIMCO), for joint analysis of multiple traits in GWAS that addresses the second question. VIMCO is applicable when individual-level data on multiple traits and genotype are available on the same study individuals. Our proposed method can be viewed as a complementary method to the widely used mvLMM implemented in the GEMMA package, in that it addresses a different but related scientific question and allows one to identify the specific trait that is associated with the genetic loci when performing a joint analysis of multiple traits. A variational Bayesian expectation-maximization (VBEM) algorithm is used to ensure computational efficiency. Through extensive numerical studies and real data analyses, we demonstrate that our proposed approach offers improved statistical power when compared with existing single-trait analysis strategies. The remainder of this article is organized as follows. In Section 2, we describe the model and algorithm that VIMCO uses to perform joint analysis of multiple traits. We then illustrate the performance of VIMCO using numerical simulations and real data analyses in Section 3 and conclude with a discussion in Section 4.

2 Variational inference for multiple correlated outcome

2.1 Model

In this section, we describe the notation and model used for joint modeling of multiple traits in a GWAS using VIMCO. Consider K continuous phenotypes/traits, $Y_1, \ldots, Y_K$, that are measured on N individuals, where Y_k is a $N \times 1$ vector for $k = 1, \ldots, K$. Assume that the genome-wide genotype data consists of p SNPs given by, $X_1, \ldots, X_p$where X_j is a $N \times 1$ vector for $j = 1, \ldots, p$. Denote $\mathbf{X} = [X_1, \ldots, X_p] \in \mathbb{R}^{N \times p}$and $\mathbf{Y} = [Y_1, \ldots, Y_K] \in \mathbb{R}^{N \times K}$. Without loss of generality, we assume that both the phenotypes and genotypes have been centered. We consider the following multivariate linear model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{1}$$

where $\mathbf{B} = [\beta_1, \ldots, \beta_p]^\top \in \mathbb{R}^{p \times K}$, $\beta_j = (\beta_{j1}, \ldots, \beta_{jK})^\top$, $j = 1, \ldots, p$, and $\mathbf{E} = [e_1^\top, \ldots, e_N^\top]^\top \in \mathbb{R}^{N \times K}$. We assume that $e_n \sim \mathcal{N}(0, \Theta^{-1})$, where Θ is the precision matrix of $\mathbf{E}$ with dimensionality $K \times K$ for individuals $n = 1, \ldots, N$. The entries in Θ are denoted by $\theta_{st}, s, t = 1, \ldots, K$. Under this model, the correlation in the traits (conditional on the genotypes) is modeled by the off-diagonal terms in Θ^{-1}.

We are interested in identifying genetic variants that are associated with one or more traits, which corresponds to the identification of nonzero entries in the matrix $\mathbf{B}$. We consider a spike-slab prior for $\mathbf{B}$and parameterize $\mathbf{B}$as a product of latent variables. Specifically, we assume that $\beta_j = \gamma_j \circ \tilde{\beta}_j = (\gamma_{j1}\tilde{\beta}_{j1}, \ldots, \gamma_{jK}\tilde{\beta}_{jK})^\top$, where $\circ$denotes the element-wise product and

$$\tilde{\beta}_{jk} \sim \mathcal{N}(0, \sigma^2_{\beta_k}), \;\; \gamma_{jk} \sim a_k^{\gamma_{jk}}(1 - a_k)^{1-\gamma_{jk}}.$$

Let $\Phi = \{a_1, \ldots, a_K, \sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_K}, \Theta\}$be the collection of (unknown) model parameters. Accordingly, our probabilistic model can be reparameterized as:

$$\Pr(\mathbf{Y}, \tilde{\beta}, \gamma | \mathbf{X}; \Phi) = \Pr(\mathbf{Y} | \mathbf{X}, \tilde{\beta}, \gamma; \Phi) \Pr(\tilde{\beta}, \gamma | \mathbf{X}; \Phi)$$

$$= \prod_{n=1}^{N} \mathcal{N}\left(\sum_{j=1}^{p} X_{nj}(\gamma_j \circ \tilde{\beta}_j), \Theta^{-1} \right) \prod_{j=1}^{p} \prod_{k=1}^{K} \left[a_k^{\gamma_{jk}} (1 - a_k)^{1-\gamma_{jk}} \mathcal{N}(0, \sigma^2_{\beta_k}) \right].$$

With this reparameterization, to identify genetic variants that are associated with one or more traits, we need to compute the posterior distribution of the latent variables $(\tilde{\beta}, \gamma)$:

$$\Pr(\tilde{\beta}, \gamma | \mathbf{Y}, \mathbf{X}; \Phi) = \frac{\Pr(\mathbf{Y}, \tilde{\beta}, \gamma | \mathbf{X}; \Phi)}{\Pr(\mathbf{Y} | \mathbf{X}; \Phi)}$$

$$= \frac{\Pr(\mathbf{Y}, \tilde{\beta}, \gamma | \mathbf{X}; \Phi)}{\sum_{\gamma} \int_{\tilde{\beta}} \Pr(\mathbf{Y}, \tilde{\beta}, \gamma | \mathbf{X}; \Phi) d\tilde{\beta}}. \tag{2}$$

We note that VIMCO is an empirical Bayesian approach where the parameters in the Spike-Slab and Normal priors are estimated from the data, i.e. it does not require specification of hyperparameters.

2.2 VBEM algorithm

In this section, we describe the algorithm used for computation of the model parameters and the posterior distribution of the latent variables $(\tilde{\beta}, \gamma)$in VIMCO. Exact computation of the posterior distribution (2) is computationally intensive due to the denominator which requires marginalizing over the latent variables $(\tilde{\beta}, \gamma)$. To overcome this computational intractability, we derive a computationally efficient VBEM algorithm (Bishop, 2006) to obtain an approximation for the posterior distribution (2). The key idea of the VBEM algorithm is to approximate our computationally intractable posterior distribution (2), with an approximating distribution, which is computationally tractable. The VBEM algorithm proceeds by specifying a family of (variational) distributions, which are parameterized by variational parameters, and choosing the optimal approximating distribution from this family by minimizing the KL divergence between the approximating distributions and our true posterior distribution (2) (Blei *et al.*, 2017). The KL divergence can be viewed as a measure of how different the approximating distribution is from our true posterior distribution (2). The optimal variational distribution is then used as an approximation for the true posterior distribution (2) (Bishop, 2006).

Let $q(\tilde{\beta},\gamma)$be a candidate approximating distribution of our true posterior distribution (2), and let $\mathrm{E}_q$denote the expectation taken with respect to $q(\tilde{\beta},\gamma)$. We can decompose the logarithm of the marginal likelihood as

$$\log p(\mathbf{Y}|\mathbf{X};\Phi) = \mathcal{L}_q + \mathrm{KL}\Big(q||\Pr(\tilde{\beta},\gamma|\mathbf{Y},\mathbf{X};\Phi)\Big), \tag{3}$$

where

$$\mathcal{L}_q = \mathrm{E}_q \log\left[\frac{\Pr(\mathbf{Y},\tilde{\beta},\gamma|\mathbf{X};\Phi)}{q(\tilde{\beta},\gamma)}\right]$$
$$\mathrm{KL}\Big(q||\Pr(\tilde{\beta},\gamma|\mathbf{Y},\mathbf{X};\Phi)\Big) = \mathrm{E}_q \log\left[\frac{q(\tilde{\beta},\gamma)}{\Pr(\tilde{\beta},\gamma|\mathbf{Y},\mathbf{X};\Phi)}\right].$$

Minimizing the KL divergence with respect to the approximating distribution $q(\tilde{\beta},\gamma)$is equivalent to maximizing the evidence lower bound (ELBO) $\mathcal{L}_q$. If we allow any possible choice for $q(\tilde{\beta},\gamma)$, then the Kullback-Leibler divergence $\mathrm{KL}\Big(q||\Pr(\tilde{\beta},\gamma|\mathbf{Y},\mathbf{X};\Phi)\Big)$is zero if and only if $q(\tilde{\beta},\gamma)$ is identical to our true posterior distribution (2) almost surely.

In order for the algorithm to be applicable to GWAS data, we require a family of distributions that is sufficiently flexible to accurately approximate the true posterior distribution, while being computationally tractable. We propose using a mean field variational family (Logsdon *et al.*, 2010), which contains distributions $q(\tilde{\beta},\gamma)$of the form

$$q(\tilde{\beta},\gamma) = \prod_j \prod_k [q(\tilde{\beta}_{jk},\gamma_{jk})].$$

The factorization used in the approximating mean field variational family of distributions makes the VBEM algorithm computationally efficient. The approximation is expected to perform best when the SNPs are independent and in the absence of pleiotropy. If more accurate inferences are needed, more computationally intensive methods may be required.

Given this variational family of distributions, the optimal variational distribution $q^*(\tilde{\beta}_{jk},\gamma_{jk})$that maximizes the ELBO $\mathcal{L}_q$has the form (Bishop, 2006)

$$\log q^*(\tilde{\beta}_{jk},\gamma_{jk}) = \mathrm{E}_{(j',k')\neq(j,k)}[\log \Pr(\mathbf{Y},\tilde{\beta},\gamma|\mathbf{X};\Phi)] + \text{constant}, \tag{4}$$

where the expectation is taken with respect to all other factors $q(\tilde{\beta}_{j'k'},\gamma_{j'k'})$ for $(j',k') \neq (j,k)$. After some derivations (details are provided in Supplementary Material SA), the optimal variational posterior distribution is given by:

$$\prod_j \prod_k [\alpha_{jk}^{\gamma_{jk}}(1-\alpha_{jk})^{1-\gamma_{jk}}\mathcal{N}(\mu_{jk},s_{jk}^2)^{\gamma_{jk}}\mathcal{N}(0,\sigma_{\beta_k}^2)^{1-\gamma_{jk}}],$$

where the variational parameters (μ_{jk}, s_{jk}^2, α_{jk}) are given by:

$$\mu_{jk} = \frac{\sum_t \theta_{kt} X_j^\top [Y_t - \sum_{j'\neq j}\alpha_{j't}\mu_{j't}X_{j'}] - \sum_{t\neq k}\theta_{kt}\alpha_{jt}\mu_{jt}\|X_j\|^2}{\theta_{kk}\|X_j\|^2 + \frac{1}{\sigma_{\beta_k}^2}},$$
$$s_{jk}^2 = \frac{1}{\theta_{kk}\|X_j\|^2 + \frac{1}{\sigma_{\beta_k}^2}},$$
$$\alpha_{jk} \equiv q(\gamma_{jk}=1) = \frac{1}{1+\exp\left(-\log\frac{a_k}{1-a_k} + \frac{1}{2}\left(\frac{\mu_{jk}^2}{s_{jk}^2} + \log\frac{s_{jk}^2}{\sigma_{\beta_k}^2}\right)\right)}. \tag{5}$$

To solve for the variational parameters (μ_{jk}, s_{jk}^2, and α_{jk}) and model parameters (Φ), the VBEM algorithm iterates between two optimization (expectation and maximization) steps until convergence. In the expectation step, we optimize the ELBO $\mathcal{L}_q$with respect to the variational parameters (μ_{jk}, $s_{jk}^2$and α_{jk}), while holding the model parameters fixed, i.e. compute variational parameters using Equation (5). In the maximization step, we optimize the ELBO $\mathcal{L}_q$with respect to the model parameters Φwhile holding the variational parameters fixed. With the optimal variational distribution, the ELBO $\mathcal{L}_q$can be evaluated in a closed form (details in Supplementary Material SA):

$$\begin{aligned}\mathcal{L}_q = &-\frac{1}{2}\sum_s\sum_t \theta_{st}(Y_s - \sum_j X_j\alpha_{js}\mu_{js})^\top(Y_t - \sum_j X_j\alpha_{jt}\mu_{jt})\\ &-\frac{1}{2}\sum_s \theta_{ss}\sum_j X_j^\top X_j[\alpha_{js}(\mu_{js}^2+s_{js}^2) - \alpha_{js}^2\mu_{js}^2]\\ &-\sum_j\sum_k\left[\alpha_{jk}\log\frac{\alpha_{jk}}{a_k} + (1-\alpha_{jk})\log\frac{1-\alpha_{jk}}{1-a_k}\right] + \frac{N}{2}\log|\Theta|\\ &+\frac{1}{2}\sum_j\sum_k \alpha_{jk}\left(1+\log\frac{s_{jk}^2}{\sigma_{\beta_k}^2} - \frac{\mu_{jk}^2+s_{jk}^2}{\sigma_{\beta_k}^2}\right) + \text{const.}\end{aligned} \tag{6}$$

By taking partial derivatives of the ELBO $\mathcal{L}_q$with respect to the model parameters and setting them to zero, we can solve for the model parameters and obtain the update equations for the maximization step:

$$\begin{aligned}a_k &= \frac{\sum_j \alpha_{jk}}{p},\\ \sigma_{\beta_k}^2 &= \frac{\sum_j \alpha_{jk}(\mu_{jk}^2+s_{jk}^2)}{\sum_j \alpha_{jk}},\\ (\Theta^{-1})_{kk} &= \frac{\|Y_k - \sum_j X_j\alpha_{jk}\mu_{jk}\|^2}{N}\\ &\quad + \frac{\sum_j \|X_j\|^2[\alpha_{jk}(\mu_{jk}^2+s_{jk}^2) - \alpha_{jk}^2\mu_{jk}^2]}{N},\\ (\Theta^{-1})_{kt} &= \frac{(Y_k - \sum_j X_j\alpha_{jk}\mu_{jk})^\top(Y_t - \sum_j X_j\alpha_{jt}\mu_{jt})}{N}.\end{aligned} \tag{7}$$

The VBEM algorithm iterates between the expectation (Equation 5) and maximization (Equation 7) steps until convergence. Further details on the VBEM algorithm are provided in Supplementary Material SA.

2.3 Inference

With the estimated variational parameters (μ_{jk}, $s_{jk}^2$and α_{jk}) and model parameters (Φ), the posterior probability $\Pr(\gamma_{jk}=1|\mathbf{Y},\mathbf{X};\Phi)$of whether genetic variant j is associated with trait k can be estimated by $\hat{\alpha}_{jk}$, and the local false discovery rate (lfdr) can be estimated by $1-\hat{\alpha}_{jk}$. Statistical inference can be conducted by identifying SNP-trait associations while controlling the global FDR at a fixed value. Specifically, given a cutoff for the global FDR, the cutoff ξ for the lfdr can be computed from global FDR $= \frac{\sum_j\sum_k \mathrm{lfdr}_{jk}\mathbb{I}(\mathrm{lfdr}_{jk}\leq\xi)}{\sum_j\sum_k \mathbb{I}(\mathrm{lfdr}_{jk}\leq\xi)}$(Newton *et al.*, 2004).

3 Results

3.1 Simulations

We conducted numerical simulations to evaluate the performance of VIMCO. We considered the scenario where we have $K=4$ continuous traits and $P=10\,000$ SNPs that were measured on $N=5000$ individuals. For each individual, the p genotypes were simulated by first generating a $p\times 1$multivariate normal distribution assuming

auto-regressive (AR) correlation with parameter ρ_x. We then discretized each variable to a trinary variable (0, 1, 2) by assuming Hardy-Weinberg equilibrium and with a minor allele frequency randomly selected from a uniform [0.05, 0.5] distribution. The genotype correlation was varied at $\rho_x = 0.2, 0.5, 0.8$. To generate the coefficient matrix **B**, for each trait, we selected 1% of the SNPs to be associated with the trait, where the effect sizes were generated from a standard normal distribution. To allow for pleiotropy where a SNP can be associated with more than one trait, we varied the proportion of causal SNPs that was associated with more than one trait. Let $g = \frac{\sum_j \mathbb{I}(\sum_k \gamma_{jk} \geq 2)}{\sum_j \sum_k \gamma_{jk}}$. The expectation of g was varied at 0, 0.15, 0.3, where increasing g reflects increasing pleiotropy. The error matrix **E** was generated with rows drawn independently from a multivariate normal distribution, with AR correlation parameter $\rho_e = 0.2, 0.5, 0.8$. A larger value of ρ_e implies a higher correlation between the traits. Each error variance was adjusted according to the pre-specified heritability of $h^2 = 0.3$.

To benchmark the performance of VIMCO, we considered both Bayesian [variational inference-based Bayesian variable selection regression (BVSR); Carbonetto *et al.*, 2012] and frequentist [single-trait linear mixed model (sLMM); Zhou and Stephens, 2012] single-trait analysis approaches. Similar to VIMCO, BVSR utilizes a VBEM algorithm, but can only be applied to single traits. sLMM was implemented using the GEMMA software package. For VIMCO and BVSR, we report the statistical power with a global FDR controlled at 0.1. The global FDR for VIMCO and BVSR controls for multiple testing across the multiple traits and SNPs. For sLMM, we report power while controlling for the family wise Type 1 error rate at 0.05, by applying a Bonferroni correction for both the number of SNPs and number of traits tested. We note that this comparison between of VIMCO/BVSR with sLMM may not be on the same footing as they are based on different metrics (FDR versus family wise error rate). The comparison, however, mimics how the methods are commonly used in practice, i.e. by controlling the global FDR for VIMCO and BVSR and by controlling the family wise error rate for sLMM. We compared the power of VIMCO, BVSR and sLMM by considering trait-SNP associations for all traits and SNPs. The power for VIMCO, BVSR and sLMM, for the scenario where the genotypes were strongly correlated ($\rho_x = 0.8$) and pleiotropy $g = 0$, is shown in the left panel of Figure 1. When the traits showed moderate ($\rho_e = 0.5$) or strong correlation ($\rho_e = 0.8$), VIMCO had higher statistical power than BVSR. When the traits were weakly correlated ($\rho_e = 0.2$), VIMCO had similar power as BVSR. In all cases, BVSR had higher power than sLMM. We also evaluated the global FDR control for the three approaches. Similar to earlier papers (Brzyski *et al.*, 2017), when evaluating the control of global FDR, the SNPs were evaluated as a cluster of SNPs. i.e. rejections within the same linkage disequilibrium block were grouped and counted as a single rejection. As shown in the middle panel in Figure 1, both VIMCO and BVSR had empirical FDRs that were close to the nominal 0.1 level. sLMM based on controlling for the family wise error rate had well-controlled FDR in this scenario, but the empirical FDR was conservative at lower genotype correlations $\rho_x = 0.2$ and 0.5 (Supplementary Figs SB1 and SB2), which is not surprising since it controls for the family wise error rate and not the FDR. We also compared the area under the curve (AUC) of VIMCO, BVSR and sLMM (right panel of Fig. 1). The AUC was evaluated by considering trait-SNP associations for all traits and SNPs. The AUC of VIMCO was higher than that of the single-trait approaches (BVSR and sLMM) when the traits showed moderate ($\rho_e = 0.5$) or strong correlation ($\rho_e = 0.8$). Simulations with lower genotype correlation $\rho_x = 0.2, 0.5$ and different levels of pleiotropy g are shown in Supplementary Figures SB1–SB3, and give similar conclusions.

As noted earlier, mvLMM assesses a different but related null hypothesis of whether any of the traits are associated with the genetic variants. Specifically, for the jth SNP, mvLMM evaluates the null hypothesis $H_{0b} : \beta_{j1} = \cdots = \beta_{jK} = 0$, whereas VIMCO evaluates the null hypothesis $H_{0a} : \beta_{jk} = 0$ for $k = 1, \ldots, K$ traits separately. We evaluated the performance of VIMCO for assessing the null hypothesis H_{0b}. For evaluating the performance in assessing the null hypothesis H_{0b}, we performed an *ad hoc* modification of VIMCO, BVSR and sLMM. In this *ad hoc* adaptation of VIMCO, BVSR and sLMM, we rejected the null hypothesis H_{0b} if H_{0a} is rejected for any of the traits. We examined the power of VIMCO and BVSR while controlling the global FDR at 0.1 For sLMM, we report power while controlling for the family wise Type 1 error rate at 0.05. We also compared the AUC of VIMCO, BVSR, sLMM and mvLMM. Results for the scenario where the genotypes were strongly correlated ($\rho_x = 0.8$) and the level of pleiotropy $g = 0$ are shown in Figure 2. Similar to results in evaluating H_{0a}, VIMCO improves statistical power and has higher AUC when the traits showed moderate or strong correlation. For this *ad hoc* adaptation of VIMCO and BVSR, the empirical FDR were well-controlled for the settings we considered (middle panel of Fig. 2). We note that evaluating the null hypothesis H_{0b} is not an intended use of VIMCO, and this *ad hoc* adaptation of VIMCO was performed in order to provide a

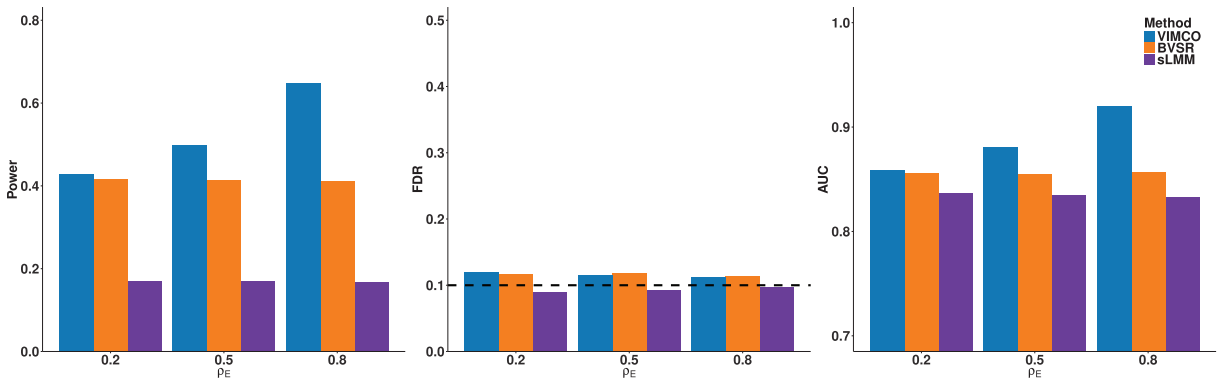

Fig. 1. Simulation results for evaluating null hypothesis H_{0a}, for different ρ_e (increasing levels of ρ_e imply increasing correlation between the traits) and genotype correlation parameter $\rho_x = 0.8$ Left panel: power of VIMCO, BVSR and sLMM; middle panel: empirical global FDR of VIMCO, BVSR and sLMM; right panel: AUC of VIMCO, BVSR and sLMM

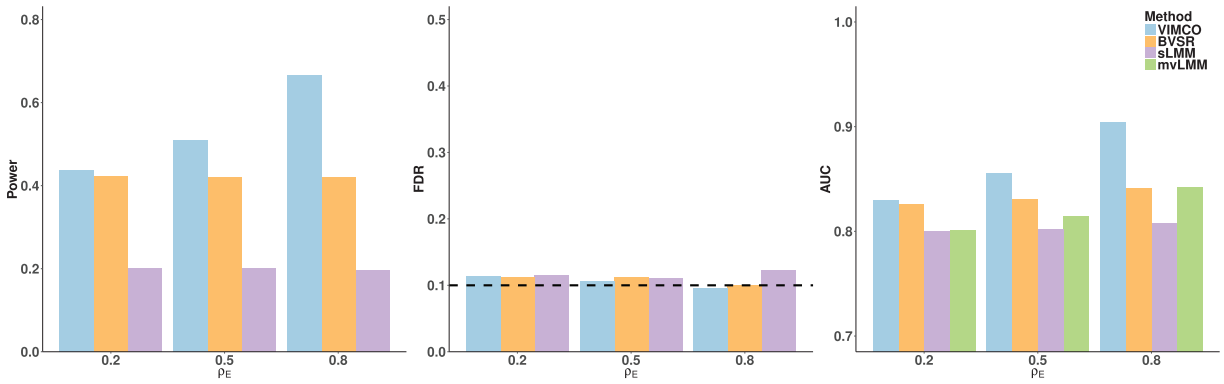

Fig. 2. Simulation results for evaluating null hypothesis H_{0b}, for different ρ_e (increasing levels of ρ_e imply increasing correlation between the traits) and genotype correlation parameter $\rho_x = 0.8$ Left panel: Power of VIMCO, BVSR and sLMM; middle panel: empirical global FDR of VIMCO, BVSR and sLMM; right panel: AUC of VIMCO, BVSR, sLMM and mvLMM

comparison with mvLMM. Simulations with different genotype correlation and levels of pleiotropy gave similar conclusions (Supplementary Figs SB4–SB6). To more closely mimic the linkage disequilibrium structure present in real genotype data, we also conducted simulations where we sampled genotypes from chromosome 1 in the NFBC1966 dataset. The linkage disequilibrium structure in real genotype data is likely to be more varied and complex instead of following a simple AR(ρ) structure that we previously examined. We observed slight inflation of the FDR for these simulations (Supplementary Figs SB7 and SB8). A possible reason is that under more complex linkage disequilibrium structures, SNPs correlated to causal SNP(s) may be selected even if they are not in the same linkage disequilibrium block(s), and determining if a selected SNP is representative of a causal SNP becomes more complicated. In cases where tighter control of FDR is required, more computationally accurate methods may be needed.

3.2 Real data analysis

To illustrate the performance of our proposed method VIMCO, we analyzed two datasets. The first dataset is a GWAS of four moderately correlated lipid traits from the Northern Finland Birth Cohort 1966 (NFBC1966). The second dataset is a GWAS of three weakly correlated eye measurements from the Singapore Indian Eye (SINDI) study (Cheng *et al.*, 2013).

3.2.1 NFBC1966

The NFBC1966 dataset consists of 10 metabolic traits and 364 590 SNPs from 5402 individuals (Sabatti *et al.*, 2009). The 10 metabolic traits include fasting lipid levels [total cholesterol (TC), high-density lipoprotein (HDL), low-density lipoprotein (LDL) and triglycerides (TG)], inflammatory marker C-reactive protein, markers of glucose homeostasis (glucose and insulin), body mass index and blood pressure (BP) measurements (systolic and diastolic BP). Quality control of the data was performed using PLINK (Purcell *et al.*, 2007) and GCTA (Yang *et al.*, 2011). Individuals with missing-ness in any of the traits and with genotype missing call-rates $> 5\%$ were excluded. We excluded SNPs with minor allele frequency $< 1\%$, missing call-rates $> 1\%$, or failed Hardy-Weinberg equilibrium. After quality control filtering and SNP pruning, 172 412 SNPs from 5123 individuals were available for analysis. We quantile-transformed each trait to a standard normal distribution, obtained the residuals after regressing out the effects of sex, oral contraceptives and pregnancy status, and quantile-transformed the residuals to a standard normal distribution.

We performed analysis on the four lipid traits (TC, LDL, HDL and TG). The pairwise Pearson correlation for the four lipid traits are given in Supplementary Figure SC1. Among the four traits, TC and LDL showed the strongest correlation (corr $= 0.88$). Modest correlation was also observed among the following pairs of traits: TG and TC (corr $= 0.41$), TG and LDL (corr $= 0.33$), TG and HDL (corr $= -0.40$). We applied VIMCO to perform joint analysis of the four traits. For comparison with the results from VIMCO, we also performed single-trait analyses whereby each of the four traits were analyzed separately, using both Bayesian (variational inference-based BVSR; Carbonetto *et al.*, 2012) and frequentist approaches sLMM (Zhou and Stephens, 2012).

For VIMCO and BVSR, we report significant SNP-trait associations at a global FDR of 0.1. The global FDR for VIMCO and BVSR controls for multiple testing across the multiple traits and SNPs. For sLMM, to control for multiple testing across both traits and SNPs, we report P-values for SNP-trait associations with P-value $< 1.25 \times 10^{-8}$ (we applied a Bonferroni adjustment for the four traits to the commonly used genome-wide significance threshold 5×10^{-8}).

Genomic locations of SNPs identified by VIMCO, BVSR and sLMM are shown in Figure 3. To control the global FDR at 0.1, a SNP has to have a lfdr < 0.73 and < 0.36 for VIMCO and BVSR, respectively (indicated by the horizontal red lines in the plots). VIMCO identified a total of 39 SNP-trait associations, whereas the single-trait analysis strategies BVSR and sLMM identified 34 and 10 SNP-trait associations, respectively. In terms of the total number of unique SNPs identified, VIMCO identified a smaller number of SNPs than BVSR; VIMCO identified 23 SNPs to be associated with at least one trait, while BVSR and sLMM identified 30 and 9 SNPs, respectively. A possible explanation for the lower number of unique SNPs identified by VIMCO is because for traits that showed strong correlation with each other (e.g. TC and LDL), VIMCO identified more SNPs than BVSR and sLMM; while for the remainder traits BVSR and sLMM identified fewer SNPs than VIMCO.

The lfdrs (VIMCO and BVSR) and P-values (sLMM) of identified SNPs are given in Supplementary Table SC1. For SNPs that were identified by either VIMCO, BVSR or sLMM, we also report their P-values from a mvLMM fitted using the GEMMA package (Zhou and Stephens, 2014). As noted earlier, mvLMM assesses a different but related hypothesis of whether any of the four traits are associated with the genetic variants.

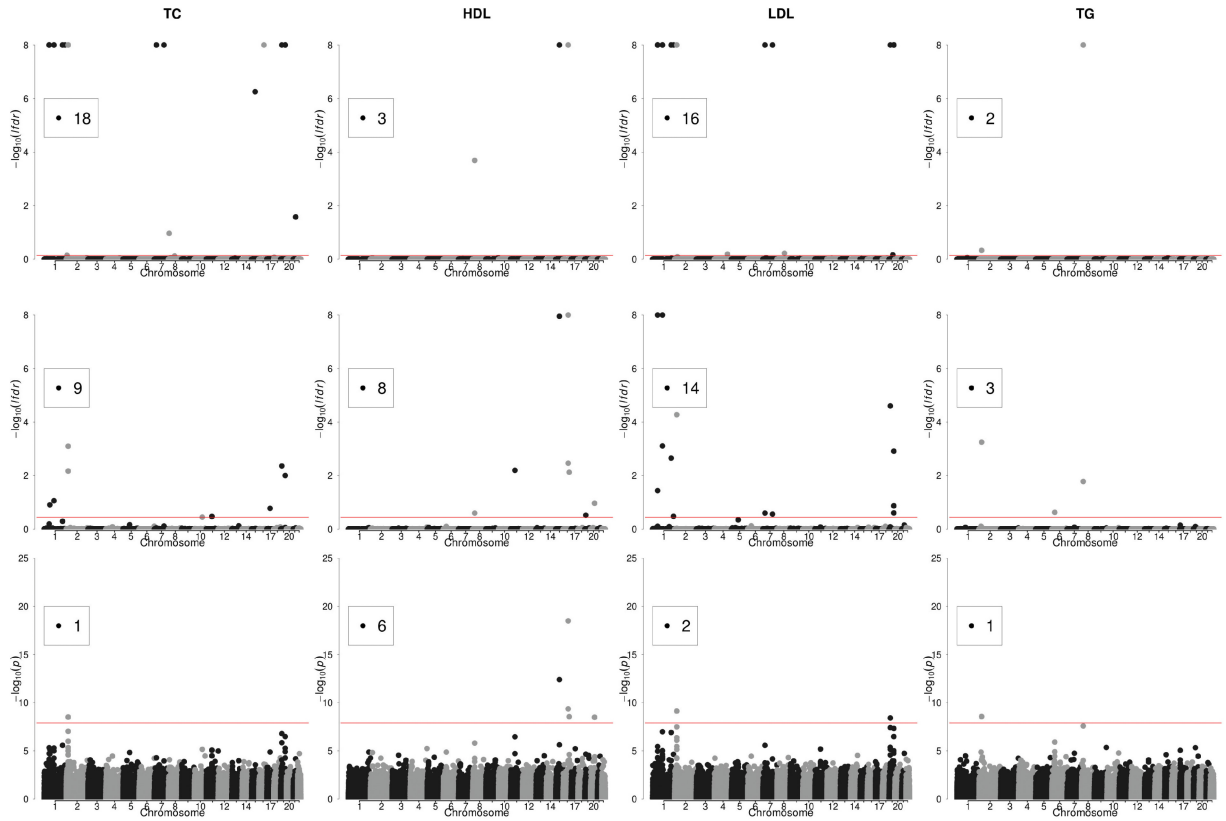

Fig. 3. Manhattan plots for the analysis results of VIMCO (first row), BVSR (second row) and sLMM (last row), in the NFBC1966 study. In the first and second row, the horizontal red lines correspond to a global FDR of 0.1. In the last row, the horizontal red line corresponds to a *P*-value cutoff of 1.25×10^{-8}. The number in the box indicates the number of SNP associations identified for each trait

Analysis of the NFBC1966 dataset was conducted on a machine with 3.0 GHz Intel Xeon CPU and 32 G memory. With the BVSR estimates as initial values (1.2 h), it took VIMCO an additional 2.5 h to complete the full variational inference procedure.

3.2.2 Singapore Indian Eye

The SINDI dataset contains three eye measurements: the ratio of BP to intraocular pressure (BP/IOP), central corneal thickness (CCT) and cup-to-disc ratio. These three traits are risk factors for glaucoma (Lavanya *et al.*, 2009). The pairwise correlation for these three traits (Supplementary Fig. SC2) was much weaker than those observed for the lipid traits in the NFBC1966 dataset, with the strongest correlation of 0.16 observed between BP/IOP and CCT. After quality control following previous studies (Cheng *et al.*, 2013) and SNP pruning, 2219 individuals and 257 736 SNPs were available for analysis. Locations of SNPs identified by VIMCO, BVSR and sLMM in the genome are shown in Figure 4. In this dataset where the traits were weakly correlated, the performance for VIMCO was similar as the single-trait approaches. VIMCO and BVSR identified the same two SNPs to be associated with CCT. Among the two associations, rs12447690 was also identified by sLMM (the *P*-value is 5.5×10^{-9}). This SNP was located in the Zinc-Finger protein (ZNF469) gene, and was previously reported to be associated with CCT (Gao *et al.*, 2016). The lfdrs and *P*-values of identified SNPs are given in Supplementary Table SC2.

Analysis of the SINDI dataset used BVSR estimates as initial values (6.8 h), and used an additional 1.2 h to complete the full variational inference procedure.

4 Discussion

In this article, we have proposed a novel method VIMCO that allows an investigator to identify the specific trait that is associated with the genetic loci when performing a joint GWAS analysis of multiple traits. Results from simulations and real data analyses demonstrate that VIMCO improved statistical power when the traits were correlated and had comparable performance when the traits were not correlated, when compared with single-trait analysis strategies. Furthermore, VIMCO utilizes a computationally efficient VBEM algorithm which allows it to handle genome-wide genotype data efficiently. The computational complexity of VIMCO is $O(K^2np + K^3)$. For a small number of traits (K) (around 10), the overall computational cost is $O(K^2np)$, i.e. the computational time is linear with respect to the sample size n and the number of SNPs p. To demonstrate the run times of VIMCO, Supplementary Figure SB9 shows the average run times for 10 iterations for: (i) $K = 4, n \in \{5000, 10\,000, 15\,000, 20\,000\}$ and $p \in \{50\,000, 100\,000, 500\,000, 1\,000\,000\}$; (ii) $p = 500\,000$, $n \in \{5000, 10\,000, 20\,000, 40\,000\}$and $K = 1, \ldots, 5$. Computations were conducted on a machine with 3.0 GHz Intel Xeon CPU and 32G memory. For a specific example, for $(K, n, p) = (4, 20\,000, 1\,000\,000)$, the total run time was 94 h. For larger sample sizes, VIMCO can be used if there is sufficient computational memory. For denser genotype data, to ensure computational efficiency, analysis can be performed for each chromosome separately.

VIMCO is, however, not without limitations. First, VIMCO is not applicable when the number of traits analyzed exceeds the number of samples. With increasing interest in performing phenome-wide

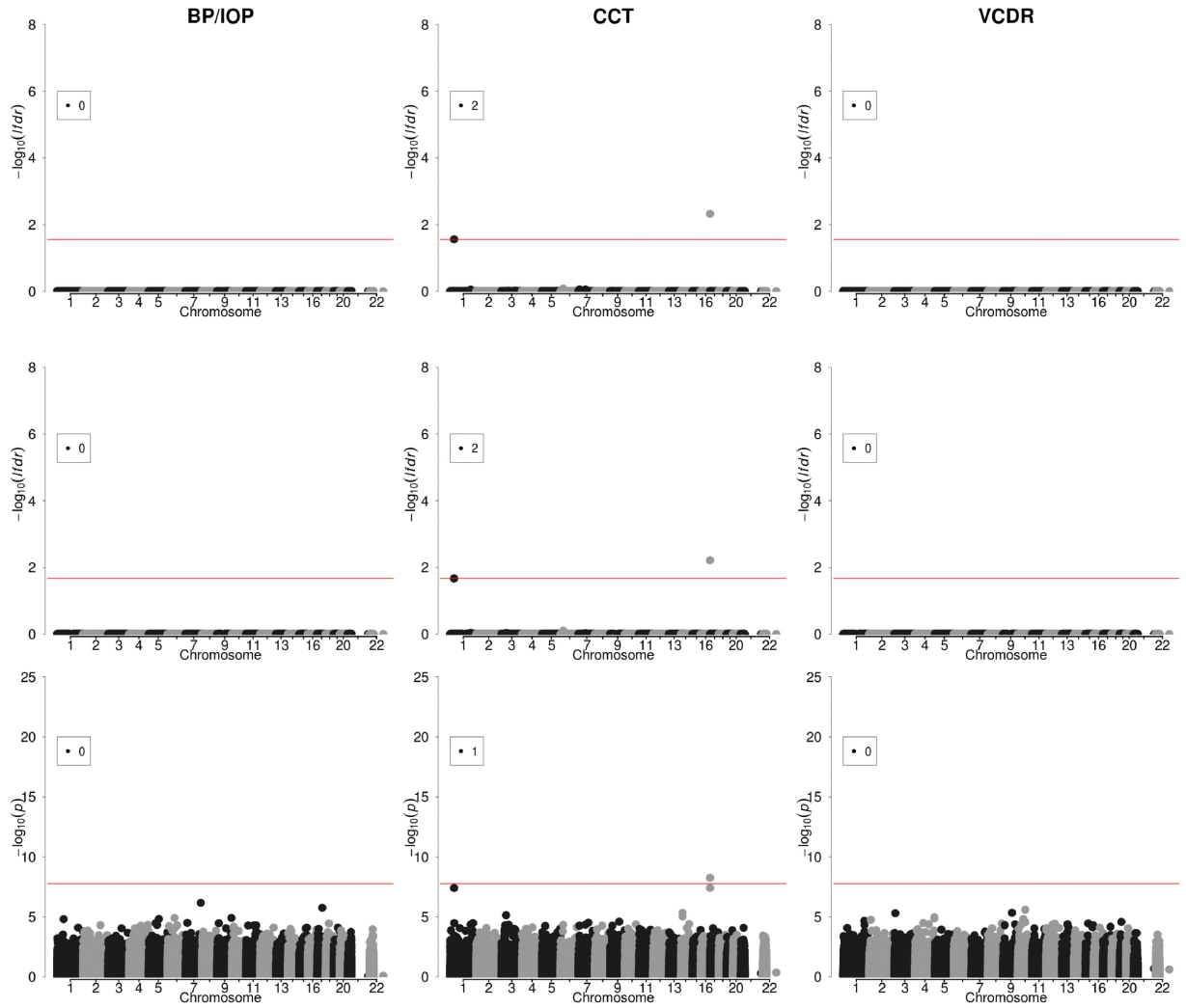

Fig. 4. Manhattan plots for the analysis results of VIMCO (first row), BVSR (second row) and sLMM (last row), in the SINDI study. In the first and second row, the horizontal red lines correspond to a global FDR of 0.1. In the last row, the horizontal red line corresponds to a *P*-value cutoff of 1.67×10^{-8}. The number in the box indicates the number of SNP associations identified for each trait

association studies whereby the number of phenotypes can be larger than the sample size, extending VIMCO to handle larger number of phenotypes is an avenue for future work. Second, VIMCO requires complete genotype and phenotype data for each individual. In practice, missingness can be handled using imputation techniques. Last, VIMCO requires individual-level trait and genotype data to be collected from the same individuals. The development of a method where summary statistics from different individuals can be used for analysis instead of individual-level data is another avenue for future research.

Acknowledgements

The authors are grateful to Minwei Dai for helpful comments and dicussion. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

Funding

This work was supported in part by [grant numbers 71501089, 11501579 and 71472023] from National Natural Science Foundation of China [grant numbers 22302815, 12316116 and 12301417] from the Hong Kong Research Grant Council [grant R-913-200-098-263 and R-913-200-127-263] from the Duke-NUS and AcRF Tier 2 [MOE2016-T2-2-029, MOE2018-T2-1-046 and MOE2018-T2-2-006] from the Ministry of Education, Singapore.

Conflict of Interest: none declared.

References

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, New York.

Blei,D.M. *et al.* (2017) Variational inference: a review for statisticians. *J. Am. Stat. Assoc.*, **112**, 859–877.

Brzyski,D. *et al.* (2017) Controlling the rate of gwas false discoveries. *Genetics*, **205**, 61–75.

Carbonetto,P. *et al.* (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.*, **7**, 73–108.

Casale,F.P. (2016) *Multivariate Linear Mixed Models for Statistical Genetics*. PhD Thesis, University of Cambridge.

Cheng,C.-Y. *et al.* (2013) Nine loci for ocular axial length identified through genome-wide association studies, including shared loci with refractive error. *Am. J. Hum. Genet.*, **93**, 264–277.

Gao,X. *et al.* (2016) Genome-wide association study identifies wnt7b as a novel locus for central corneal thickness in Latinos. *Hum. Mol. Genet.*, **25**, 5035–5045.

Kim,S. *et al.* (2009) A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, **25**, i204–i212.

Korte,A. *et al.* (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.*, **44**, 1066.

Lavanya,R. *et al.* (2009) Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. *Ophthalmic Epidemiol.*, **16**, 325–336.

Liu,J. *et al.* (2016) Analyzing association mapping in pedigree-based gwas using a penalized multitrait mixed model. *Genet. Epidemiol.*, **40**, 382–393.

Logsdon,B.A. *et al.* (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, **11**, 58.

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

Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rothman,A.J. *et al.* (2010) Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Stat.*, **19**, 947–962.

Sabatti,C. *et al.* (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.*, **41**, 35.

Solovieff,N. *et al.* (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, **14**, nrg3461.

Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.

Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821.

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