Bioinformatics, 32(12), 2016, 1856–1864
doi: 10.1093/bioinformatics/btw081
Advance Access Publication Date: 15 February 2016
Original Paper

OXFORD

Genetics and population analysis

EPS: an empirical Bayes approach to integrating pleiotropy and tissue-specific information for prioritizing risk genes

Jin Liu[1], Xiang Wan[2], Shuangge Ma[3] and Can Yang[4,*]

[1]Center of Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore, [2]Department of Computer Science, Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Kowloon, Hong Kong, [3]Department of Biostatistics, Yale University, New Heaven, CT, USA and [4]Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on May 15, 2015; revised on January 9, 2016; accepted on February 5, 2016

Abstract

Motivation: Researchers worldwide have generated a huge volume of genomic data, including thousands of genome-wide association studies (GWAS) and massive amounts of gene expression data from different tissues. How to perform a joint analysis of these data to gain new biological insights has become a critical step in understanding the etiology of complex diseases. Due to the polygenic architecture of complex diseases, the identification of risk genes remains challenging. Motivated by the shared risk genes found in complex diseases and tissue-specific gene expression patterns, we propose as an Empirical Bayes approach to integrating Pleiotropy and Tissue-Specific information (EPS) for prioritizing risk genes.

Results: As demonstrated by extensive simulation studies, EPS greatly improves the power of identification for disease-risk genes. EPS enables rigorous hypothesis testing of pleiotropy and tissue-specific risk gene expression patterns. All of the model parameters can be adaptively estimated from the developed expectation–maximization (EM) algorithm. We applied EPS to the bipolar disorder and schizophrenia GWAS from the Psychiatric Genomics Consortium, along with the gene expression data for multiple tissues from the Genotype-Tissue Expression project. The results of the real data analysis demonstrate many advantages of EPS.

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

Contact: eeyang@hkbu.edu.hk

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

1 Introduction

As of April 2015, genome-wide association studies (GWAS) have identified more than 15 000 disease-associated single nucleotide polymorphisms (SNPs) at a genome-wide significance level (i.e. *P*-value $< 5 \times 10^{-8}$). Despite these discoveries, these SNPs can only explain a small fraction of the genetic contribution to diseases (Welter *et al.*, 2014). It is widely agreed that this phenomenon is due to polygenicity of complex phenotypes and the large number of risk variants with weak effects that remain undiscovered (Visscher *et al.*, 2012).

Growing evidence suggests that there are many genetic variants that may affect the multiple, seemingly different phenotypes (traits/diseases) (Solovieff *et al.*, 2013). In fact, a term named 'pleiotropy' was introduced to describe such a phenomenon more than 100 years ago (Stearns, 2010). The original definition of 'pleiotropy' only

© The Author 2016. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

concerns the nonzero effect of a locus on different phenotypes and allows the effect direction to be either the same (i.e. the risk allele increases the risk of two diseases) or different (i.e. the risk allele increases the risk of one disease while decreases the risk of another disease). For example, the G allele of rs6983267 increases the risk for prostate cancer and colorectal cancer (Thomas *et al.*, 2008; Tomlinson *et al.*, 2007) but the G allele of rs12720356 increases the risk for Crohn's disease and decreases the risk for psoriasis (Franke *et al.*, 2010, Genetic Analysis of Psoriasis Consortium and the Wellcome Trust Case Control Consortium 2 and others, 2010). A closely related but different concept is 'genetic correlation' (also known as 'co-heritability'—Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013a). This refers to the correlation between the genetic effect sizes of two phenotypes and thus the direction of these effects plays a role. To be clear, when referring to 'pleiotropy' in this article, we adopt its original definition. We are interested in the substantial overlap of genetic factors that underlie the multiple, seemingly different phenotypes. Fortunately, more GWAS results are becoming publicly available from GWAS repositories, such as the GWAS–GRASP (Eicher *et al.*, 2015). This offers us an unprecedented opportunity to explore pleiotropic effects.

Besides an enormous amount of GWAS data, diverse data resources for biological processes at different layers are also becoming available (Civelek and Lusis, 2014). For example, the ongoing Genotype-Tissue Expression project (GTEx) is generating a comprehensive atlas of gene expression and regulation across multiple human tissues (Lonsdale *et al.*, 2013). This provides abundant information to quantify biological processes that occur at the cellular level and thereby helps us to dissect the etiology of complex diseases. Moreover, the comprehensive gene expression data in multiple tissues may avoid the potential bias of the 'candidate tissue approach'. For example, in the study of obesity, the affected tissue is fat, but the causal tissue is, in some cases, the hypothalamus (Loos *et al.*, 2008). This suggests that the integration of GWAS data and gene expressions from multiple tissues may advance our understanding of complex diseases.

At present, the accumulating evidence on pleiotropy (Wang *et al.*, 2015; Yang *et al.*, 2015) and vast, publicly available data on tissue-specific gene expression [e.g. TiGER (Liu *et al.*, 2008) and GTEx (Moore, 2013)] require novel statistical approaches to gain a deeper understanding of complex diseases (Ritchie *et al.*, 2015). Statistical approaches that take pleiotropy into consideration have recently become very active (Solovieff *et al.*, 2013), such as the conditional false discovery rate (FDR) approach (Andreassen *et al.*, 2013), and linear mixed model-based approaches (Shriner, 2012; Zhou and Stephens, 2014). Various statistical methods have also been proposed to analyze tissue-specific gene expression data (Gutierrez-Arcelus *et al.*, 2015), such as joint eQTL analysis in multiple tissues (Flutre *et al.*, 2013) and covariate-modulated FDR (CMFDR) (Zablocki *et al.*, 2014). More recently, a statistical approach named 'GPA' (Chung *et al.*, 2014) was proposed to simultaneously integrate both pleiotropic information and functional annotation. Although GPA has demonstrated a very promising direction for the integration of multiple sources of genomic information, it does not fully account for the effects of linkage disequilibrium (LD) and only allows binary annotation as its input. These disadvantages may limit its use in the integrative analysis of large-scale genomic data.

In this article, we propose an Empirical Bayes approach to integrating Pleiotropy and Tissue-Specific (EPS) information for prioritizing risk genes. Compared with some existing approaches (Lee *et al.*, 2015; Torres *et al.*, 2014), EPS has several merits. First, EPS only requires the summary statistics at the gene level, rather than the genotype data at the individual level. The LD effects are carefully accounted for when grouping the SNP-level summary statistics into the gene level summary statistics by using reference panel data (e.g. 1000 genome data). Second, EPS is able to integrate both pleiotropy information and gene expression data (e.g. GTEx data), which greatly improves the power of risk gene prioritization. Third, EPS provides rigorous hypothesis testing to evaluate the significance of pleiotropy and tissue-specific patterns. These merits are demonstrated via extensive simulation studies and real data analysis.

2 EPS: basic model, algorithm and inference

2.1 Basic model

Suppose we have the *P*-values of all SNPs from one GWAS. Instead of working at the SNP level, we group the *P*-values at the SNP level into the *P*-values at the gene level using VEGAS (Liu *et al.*, 2010). There are several advantages to working at the gene level. First, the signals at the gene level are often more visible than those at the SNP level (Li *et al.*, 2011). Second, the gene itself is highly consistent across populations, which leads to more consistent results across a population and thus increases the replication rate (Neale and Sham, 2004). Third, it is more convenient to integrate expression data from multiple tissues with GWAS data at the gene level. We demonstrate these advantages in the real data analysis below.

Let us start with the simplest case. Consider *P*-values at the gene level: $P_1,\ldots,P_g,\ldots,P_G$, where G is the number of genes. We use the 'two-group model' (Efron *et al.*, 2008) and assume that the *P*-values come from the mixture of a null and a non-null distribution, with a probability of π_0 and $\pi_1 = 1 - \pi_0$, respectively. Let $\mathbf{Z}_g = (Z_{g0}, Z_{g1})$ be the hidden variable indicating whether the *P*-value for the gth gene is from the null or non-null group, i.e. $Z_{g0} \in \{0,1\}$, $Z_{g1} \in \{0,1\}$, and $Z_{g0} + Z_{g1} = 1$ (a gene can be either null or non-null). Here, $Z_{g0} = 1$ means that the gth gene is not associated with the trait (null) and $Z_{g1} = 1$ means that the gth gene is associated (non-null) with the trait. Thus, we have $\pi_0 = \Pr(Z_{g0} = 1)$ and $\pi_1 = \Pr(Z_{g1} = 1)$. Next, we model the conditional distribution of *P*-values given Z as: $P_g|Z_{g0} = 1 \sim \mathcal{U}(0,1)$ and $P_g|Z_{g1} = 1 \sim \mathcal{B}(\alpha,1)$, where $\mathcal{U}(0,1)$ is the uniform distribution on [0, 1], and $\mathcal{B}(\alpha,1)$ is the Beta distribution with parameters α and 1, where $0 < \alpha < 1$.

For the ease of presentation, we demonstrate the model for the case with two GWAS datasets but the generalization to more than two GWAS datasets is straightforward. Suppose, we have *P*-values from two GWAS datasets at the gene level and denote them by $\mathbf{P} = (P_{gk}) \in \mathbb{R}^{G\times 2}$, where $g \in \{1,\ldots,G\}$ is the index for the gene and $k \in \{1,2\}$ is the index for the GWAS dataset. Similar to the aforementioned basic model, we introduce the hidden variable $\mathbf{Z}_g = (Z_{g00}, Z_{g10}, Z_{g01}, Z_{g11})$ indicating the association between the gth gene and the two traits: $Z_{g00} = 1$ means the gth gene is associated with neither of them, $Z_{g10} = 1$ means it is only associated with the first trait, $Z_{g01} = 1$ means it is only associated with the second trait, and $Z_{g11} = 1$ means it is associated with both traits. Then, we extend the two-group model to the following 'four-group model'.

$$\pi_{00} = \Pr(Z_{g00} = 1) : P_{g1} \sim \mathcal{U}(0,1), \quad P_{g2} \sim \mathcal{U}(0,1),$$

$$\pi_{10} = \Pr(Z_{g10} = 1) : P_{g1} \sim \mathcal{B}(\alpha_1,1), \; P_{g2} \sim \mathcal{U}(0,1),$$

$$\pi_{01} = \Pr(Z_{g01} = 1) : P_{g1} \sim \mathcal{U}(0,1), \quad P_{g2} \sim \mathcal{B}(\alpha_2,1),$$

$$\pi_{11} = \Pr(Z_{g11} = 1) : P_{g1} \sim \mathcal{B}(\alpha_1,1), \; P_{g2} \sim \mathcal{B}(\alpha_2,1),$$

where $0 < \alpha_k < 1, k = 1,2$. Since the LD among SNPs has been taken into account by using simulation based on the LD structure of

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

a set of reference individuals from 1000 genome in VEGAS2 (Liu *et al.*, 2010; Mishra and Macgregor, 2015), it is reasonable to assume the independence of gene markers. Therefore, the joint distribution can be written as

$$\Pr(\mathbf{P}) = \prod_{g=1}^{G}\left(\sum_{l\in L}\Pr(\mathbf{P}_g|Z_{gl})\Pr(Z_{gl})\right) = \prod_{g=1}^{G}\left(\sum_{l\in L}\pi_l \Pr(\mathbf{P}_g|Z_{gl}=1)\right), \tag{1}$$

where $\mathbf{P}_g$ is the gth row of $\mathbf{P}$ and $L = \{00, 10, 01, 11\}$.

To incorporate the gene expression from multiple tissues, we extend model (1) as follows. Suppose we obtain expressions of G genes from T tissues and denote it by $\mathbf{E} \in \mathrm{R}^{G\times T}$. Without loss of generality, we assume that each of G gene expressions is normalized tissue by tissue (in column of $\mathbf{E}$) with mean equal to 0 and variance equal to 1. Similar to the distribution of P-values, given hidden variable $\mathbf{Z}_g$, we have

$$\mathbf{E}_g|Z_{g0} = 1 \sim N(\boldsymbol{\mu}_0, \Sigma) \text{ and } \mathbf{E}_g|Z_{g1} = 1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \tag{2}$$

where $\mathbf{E}_g$ is the gth row of $\mathbf{E}$ representing the gth gene expression across multiple tissues, $\boldsymbol{\mu}_l$ is a length-T vector representing the mean of gene expression across T tissues in the lth group ($l \in \{0, 1\}$ in the two-group model), and $\boldsymbol{\Sigma}$ is a $T \times T$ covariance matrix. Note that model (2) is very similar to linear discriminant analysis (LDA): conditioning on the indicating variable $\mathbf{Z}_g$, $\mathbf{E}_g$ is normally distributed with different mean but the same covariance. However, a fundamental difference between model (2) and LDA is that $\mathbf{Z}_g$ is not directly observed here, which makes our problem much more challenging. Regarding the Gaussian distribution on gene expression data, it is a reasonable assumption as long as gene expression data is appropriately normalized (Efron, 2010).

Like the way we model G gene expressions in two-group model, we have $\mathbf{E}_g|Z_{gl} = 1 \sim N(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}), \mathrm{l} \in \mathrm{L} = \{00, 10, 01, 11\}$ in the four-group model. Then, the joint distribution can be written as

$$\begin{aligned}\Pr(\mathbf{P}, \mathbf{E}) &= \prod_{g=1}^{G}\left(\sum_{l\in L}\pi_l \Pr(Z_{gl}=1)\Pr(\mathbf{P}_g, \mathbf{E}_g|Z_{gl}=1)\right)\\ &= \prod_{g=1}^{G}\left(\sum_{l\in L}\pi_l \Pr(\mathbf{P}_g|Z_{gl}=1)\Pr(\mathbf{E}_g|Z_{gl}=1)\right),\end{aligned} \tag{3}$$

where $\mathbf{P}_g$ and $\mathbf{E}_g$ are the gth row of $\mathbf{P}$ and $\mathbf{E}$. The first equality holds by assuming the independence among genes. The second equality holds by assuming the independence between $\mathbf{P}_g$ and $\mathbf{E}_g$, conditional on Z_{gl}. The independence among genes could be a strong assumption. To evaluate the impact of this assumption, we set up simulation studies using the real datasets from GTEx. The details of simulation studies are given in Section 4.1 and the Supplementary materials.

2.2 EM algorithm

In this section, we describe an EM algorithm to estimate parameter based on the joint model (3). Let $\boldsymbol{\Theta}$ collect all model parameters as $\boldsymbol{\Theta} = (\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}, \alpha_1, \alpha_2, \boldsymbol{\mu}_{00}, \boldsymbol{\mu}_{10}, \boldsymbol{\mu}_{01}, \boldsymbol{\mu}_{11}, \boldsymbol{\Sigma})$ and $\mathbf{Z}_g, g = 1, \ldots, G$, be the latent variable. Then the complete likelihood can be written as

$$L_c(\boldsymbol{\Theta}; \mathbf{P}, \mathbf{E}, \mathbf{Z}) = \prod_{g=1}^{\mathrm{G}}\prod_{\mathrm{l}\in\mathrm{L}}\left(\pi_l \Pr(\mathbf{P}_g|Z_{gl}=1)\Pr(\mathbf{E}_g|Z_{gl}=1)\right)^{Z_{gl}}. \tag{4}$$

Based on the complete likelihood, the $\mathbb{E}$-step and $\mathbb{M}$-step in the sth iteration of the EM algorithm involves the following calculations.

$\mathbb{E}$-step: For $l \in L$, posterior probabilities for association of the gth gene are obtained as

$$\begin{aligned}z_{gl}^{(s)} &= \Pr(Z_{gl}=1|\mathbf{P}, \mathbf{E}, \Theta^{(s)})\\ &= \frac{\pi_l^{(s)}\Pr(\mathbf{P}_g|Z_{gl}=1; \boldsymbol{\Theta}^{(s)})\Pr(\mathbf{E}_g|Z_{gl}=1; \boldsymbol{\Theta}^{(s)})}{\sum_{l'\in L}\pi_{l'}^{(s)}\Pr(\mathbf{P}_g|Z_{gl'}=1; \boldsymbol{\Theta}^{(s)})\Pr(\mathbf{E}_g|Z_{gl'}=1; \boldsymbol{\Theta}^{(s)})}.\end{aligned}$$

$\mathbb{M}$-step: Parameters for proportion of genes in each association status category are estimated as:

$$\pi_l^{(s+1)} = \frac{1}{G}\sum_{g=1}^{G} z_{gl}^{(s)}, l \in L.$$

The parameter α in the Beta distribution can be estimated as:

$$\alpha_1^{(s+1)} = \frac{\sum_{g=1}^{G} z_{g1*}^{(s)}}{\sum_{g=1}^{G} z_{g1*}^{(s)}(-\log P_{g1})}, \quad \alpha_2^{(s+1)} = \frac{\sum_{g=1}^{G} z_{g*1}^{(s)}}{\sum_{g=1}^{G} z_{g*1}^{(s)}(-\log P_{g2})},$$

where $z_{g1*}^{(s)} = z_{g10}^{(s)} + z_{g11}^{(s)}$ and $z_{g*1}^{(s)} = z_{g01}^{(s)} + z_{g11}^{(s)}$. Parameters in the Gaussian distribution are estimated as:

$$\boldsymbol{\mu}_{\mathrm{l}}^{(\mathrm{s}+1)} = \frac{\sum_{\mathrm{g}=11}^{\mathrm{G}} \mathrm{z}_{\mathrm{gl}}^{(\mathrm{s})}\mathbf{E}_{\mathrm{g}}}{\mathrm{n}_{\mathrm{l}}},$$

$$\boldsymbol{\Sigma}^{(\mathrm{s}+1)} = \frac{1}{n}\sum_{l\in L}\sum_{g=1}^{G}\left(z_{gl}^{(s)}(\mathbf{E}_g - \boldsymbol{\mu}_{\mathrm{l}}^{(\mathrm{s}+1)})(\mathbf{E}_g - \boldsymbol{\mu}_{\mathrm{l}}^{(\mathrm{s}+1)})^{\top}\right),$$

where $n_l = \sum_{g=1}^{G} z_{gl}^{(s)}$, $n = \sum_{l\in L} n_l$ and $l \in L$.

2.3 Statistical inference

2.3.1 False discovery rate

After the parameters in the EPS model are estimated, genes can be prioritized based on their local FDRs (Efron, 2010). For single GWAS without incorporating gene expression, the estimated FDR of the gth gene is the probability that the gth gene belongs to the null group given its P-value, i.e.

$$\widehat{fdr}_g = \Pr(Z_{g0}=1|P_g; \hat{\boldsymbol{\Theta}}) = \frac{\hat{\pi}_0\Pr(P_g|Z_{g0}=1; \hat{\alpha})}{\sum_{l\in\{0,1\}}\hat{\pi}_l\Pr(P_g|Z_{gl}=1; \hat{\alpha})}. \tag{5}$$

For the joint analysis of two GWAS, we are interested in the three estimated local FDR of the gth gene if it is claimed to be associated with the first GWAS, the second GWAS and both GWAS, respectively.

$$\widehat{\mathrm{fdr}}_{g1} = \Pr(Z_{g0*}=1|\mathbf{P}_g; \hat{\boldsymbol{\Theta}}),$$

$$\widehat{\mathrm{fdr}}_{g2} = \Pr(Z_{g*0}=1|\mathbf{P}_g; \hat{\boldsymbol{\Theta}}),$$

$$\widehat{\mathrm{fdr}}_{g,share} = 1 - \Pr(Z_{g11}=1|\mathbf{P}_g; \hat{\boldsymbol{\Theta}}), \tag{6}$$

where $Z_{g0*} = Z_{g00} + Z_{g01}$, $Z_{g*0} = Z_{g00} + Z_{g10}$, and $\mathbf{P}_g$ is the gth row of the P-value matrix $\mathbf{P}$. To incorporate gene expression from multiple tissues, the estimated local FDRs are given as

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

$$\widehat{\text{fdr}}_{g1} = \Pr(Z_{g0*} = 1|\mathbf{P}_g, \mathbf{E}_g; \hat{\mathbf{\Theta}}),$$

$$\widehat{\text{fdr}}_{g2} = \Pr(Z_{g*0} = 1|\mathbf{P}_g, \mathbf{E}_g; \hat{\mathbf{\Theta}}),$$

$$\widehat{\text{fdr}}_{g,\text{share}} = 1 - \Pr(Z_{g11} = 1|\mathbf{P}_g, \mathbf{E}_g; \hat{\mathbf{\Theta}}), \quad (7)$$

where $\mathbf{E}_g$ is the gth row of expression matrix $\mathbf{E}$. Based on the known relationship between local FDR and global FDR (Efron, 2010), we can easily convert the estimated local FDR to global FDR using

$$\widehat{\text{FDR}}(\tau) = \frac{\sum_{g=1}^{G} \widehat{fdr}_g \mathbb{I}\left[\widehat{fdr}_g \leq \kappa\right]}{\sum_{g=1}^{G} \mathbb{I}\left[\widehat{fdr}_g \leq \kappa\right]} \leq \tau, \quad (8)$$

where κ is a pre-specified threshold for local FDR, function $\mathbb{I}(\cdot)$ is the indicator function which returns 1 if the argument is true, 0 otherwise, and the resulting global FDR is smaller than τ. By doing so, it is convenient for users to control FDR either in terms of global FDR or local FDR.

2.3.2 Hypothesis testing of risk genes differentially expressed in a tissue-specific manner

It is also very important to test whether the risk genes are differentially expressed in a tissue-specific manner, which offers us more biological insights on etiology of complex diseases. First, we consider evaluating whether the risk genes from a single GWAS are differentially expressed in the tth tissue. Let $\mu_{0,t}$ and $\mu_{1,t}$ be the mean expression values of genes from the null and non-null groups, respectively. Whether $\mu_{0,t}$ and $\mu_{1,t}$ are significantly different from each other can be evaluated via the following hypothesis testing:

$$\mathcal{H}_0^{(t)} : \mu_{0,t} = \mu_{1,t} \text{ versus } \mathcal{H}_1^{(t)} : \mu_{0,t} \neq \mu_{1,t}. \quad (9)$$

The likelihood ratio test (LRT) statistic is given by

$$\Lambda^{(t)} = 2\Big(\log\Pr(\mathbf{P}, \mathbf{E}_t; \hat{\mathbf{\Theta}}) - \log\Pr(\mathbf{P}, \mathbf{E}_t; \hat{\mathbf{\Theta}}_0^{(t)})\Big),$$

where $\mathbf{P}$ is the P-value vector of one GWAS, $\mathbf{E}_t$ is the standardized gene expression from the tth tissue, $\mathbf{\Theta}_0^{(t)}$ is the parameter estimates obtained under the $\mathcal{H}_0$ and its superscript t indicates the tth tissue. Based on the asymptotical theory (Van der Vaart, 2000), the test statistic $\Lambda^{(t)}$ asymptotically follows the $\chi^2_{\text{df}=1}$ under the null. This hypothesis testing allows us to evaluate whether risk genes are differentially expressed in a specific tissue. In such a way, we can marginally scan all tissues one at a time.

When there are two GWAS, the null hypothesis of testing with respect to the tth tissue can be naturally extended as follows:

$$\mathcal{H}_0^{(t)} : \mu_{00,t} = \mu_{10,t} = \mu_{01,t} = \mu_{11,t}. \quad (10)$$

The LRT test statistic asymptotically follows the $\chi^2_{\text{df}=3}$ under the null.

2.3.3 Hypothesis testing of pleiotropy

It is of great interest to test pleiotropy between two different traits/diseases. Based on the definition of pleiotropy, we are interested in whether there exists a significant overlapped genetic factor that underlies the two phenotypes. Under the null, there is no pleiotropy, i.e. the proportion of shared genes in two GWAS is independent of each other. Statistically, this can be evaluated by testing whether the joint probability of $\{Z_{g00}, Z_{g10}, Z_{g01}, Z_{g11}\}$ equals to the product of their marginal probability:

$$\mathcal{H}_0 : \pi_{11} = \pi_{1*}\pi_{*1} \text{ v.s. } \mathcal{H}_1 : \pi_{11} \neq \pi_{1*}\pi_{*1}, \quad (11)$$

where $\pi_{1*} = \pi_{10} + \pi_{11}$ and $\pi_{*1} = \pi_{01} + \pi_{11}$ are the marginal probability. Clearly, this hypothesis test can be done via the LRT, and the test statistic is given as

$$\Lambda^{(P)} = 2\Big(\log\Pr(\mathbf{P}; \hat{\mathbf{\Theta}}) - \log\Pr(\mathbf{P}; \hat{\mathbf{\Theta}}_0^{(P)})\Big), \quad (12)$$

where the superscript (P) indicates that the test aims at evaluating pleiotropy and $\hat{\mathbf{\Theta}}_0^{(P)}$ is the parameter estimates obtained under $\mathcal{H}_0$. Under the null, the test statistic $\Lambda^{(P)}$ asymptotically follows χ^2 distribution with df = 1, denoted as $\chi^2_{\text{df}=1}$.

2.3.4 Empirical null for inflated *P*-values

In the case that the P-values are inflated, it is inappropriate to assume that the P-values from null group follow Uniform distribution on [0, 1] (Schwartzman, 2008). Alternatively, we may use $P_g|Z_{g0} = 1 \sim \mathcal{B}(\alpha_{\text{null}}, 1)$ to accommodate the inflation. The parameters α_{null} in null distribution can be estimated using P-values close to 1, e.g. P-value ≥ 0.5. Then we use $\mathcal{B}(\alpha_{\text{null}}, 1)$ to replace $\mathcal{U}[0, 1]$ as the null distribution and keep α_{null} unchanged during the EM updates. The strategy adopted here is similar to the one in the previous models (1) and (3) using the empirical null.

3 Extended model

3.1 Integrating a large number of tissues

If a large number of tissues are involved for gene expressions, the proposed model may not be appropriate because the number of parameters in $\mathbf{\Sigma}$ increases quadratically with respect to the number of tissues T (see Equation (2)). To overcome this difficulty, we propose a stage-wise approach when T is large, where penalized LDA is used to regularize our model.

In this section, we describe this stage-wise approach. Let us begin with a single-GWAS analysis, with the extension to a joint analysis straightforward.

We fit the model only with the P-value vector $\mathbf{P}$ from one GWAS and obtain the first-stage posterior probability for all G genes: $\mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2; \ldots; \mathbf{z}_G] \in \mathbb{R}^{G\times 2}$, where $\mathbf{z}_g = (z_{g0}, z_{g1}), g = 1, \ldots, G$.

Treating the first-stage posterior probability $\mathbf{z}$ as a response matrix, we use the penalized LDA (Witten and Tibshirani, 2011) to find a sparse linear discriminant vector, denoted as $\hat{\mathbf{B}} \in \mathrm{R}^T$, to separate the null ($z_{g,0}$) and non-null ($z_{g,1}$) group. The purpose here is to use a sparse discriminant vector to select the most useful tissue types and reduce the effects of irrelevant tissues. Using the discriminant vector $\hat{\mathbf{B}}$, we can update the gene expression based on the sparse projection as $\tilde{\mathbf{E}} = \mathbf{E}\hat{\mathbf{B}} \in \mathbb{R}^G$. The details of our modified LDA are given in the following section.

We fit model (4) using $\tilde{\mathbf{E}}$, a sparse compression of gene expression, to replace the original high-dimensional gene expression $\mathbf{E} \in \mathrm{R}^{G\times T}$ (T is large), and then compute the second-stage posterior probability with $\mathbf{P}_g, \tilde{\mathbf{E}}_g$ as $\Pr(Z_{gl} = 1|\mathbf{P}_g, \tilde{\mathbf{E}}_g), l \in \{0, 1\}$.

To extend the model to the joint analysis of two-GWAS, we need to find three sparse projection vectors $\hat{\mathbf{B}} \in \mathbb{R}^{T\times 3}$ because there are four groups $L \in \{00, 10, 01, 11\}$. As a result, $\tilde{\mathbf{E}} = \mathbf{E}\hat{\mathbf{B}}$ is a $G \times 3$ matrix. In this case, it is still applied to fit the model (4) using $\tilde{\mathbf{E}}$ to replace $\mathbf{E}$.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

3.2 Penalized LDA

We need to modify penalized LDA proposed by Witten and Tibshirani (2011) because here we only have the first-stage posterior probability ($\mathbf{z}_g$) of each gene from the null and non-null groups with no knowledge of its exact group membership.

Taking this uncertainty into account, the within-class covariance matrix $\mathbf{\Sigma}_w$ is estimated using the first-stage posterior probability:

$$\hat{\mathbf{\Sigma}}_w = \frac{1}{G}\sum_{g=1}^{G}\sum_{l\in L} z_{gl}(\mathbf{E}_g - \hat{\boldsymbol{\mu}}_l)(\mathbf{E}_g - \hat{\boldsymbol{\mu}}_l)^\top,$$

where $\hat{\mu}_l$ is the sample mean in the lth group:

$$\hat{\boldsymbol{\mu}}_l = \frac{1}{G}\sum_{g} z_{gl}\mathbf{E}_g.$$

Similarly, the between-class covariance matrix can be estimated as

$$\hat{\mathbf{\Sigma}}_b = \frac{1}{n}\mathbf{E}^\top\mathbf{z}(\mathbf{z}^\top\mathbf{z})^{-1}\mathbf{z}^\top\mathbf{E}.$$

The remaining part is the same as the original paper (Witten and Tibshirani, 2011). For completeness, we describe more technical details of penalized LDA in the Supplementary materials.

4 Results

4.1 Simulation

In this section, we report the simulation results to evaluate the performance of the model. We generated a gene expression matrix $\mathbf{E} \in \mathbb{R}^{G\times T}$ for G genes from T tissues, where each entry of $\mathbf{E}$ was from the standard normal distribution. Let $y_g = \mathbf{E}_g\mathbf{b} + \epsilon_g$ be the latent variable controlling the membership of the gth gene, where $\mathbf{b} \in \mathbb{R}^{T\times 1}$ is the effect sizes of genes from T tissues, $\mathbf{E}_g$ is the gth row of $\mathbf{E}$, and ϵ_g is the random noise, $g = 1, \ldots, G$. After generating y_g, we dichotomized the G genes into the null and non-null groups using a probit model. Denote the proportions of the null and non-null genes of both GWAS as π_0 and π_1, respectively. The P-values for the first GWAS were simulated as:

$$\begin{aligned} P_{g,1}|y_g &\sim \mathcal{B}(\alpha_1, 1) \quad \text{if } y_g \geq q_{\pi_0}, \\ P_{g,1}|y_g &\sim \mathcal{U}(0,1), \quad \text{if } y_g < q_{\pi_0}, \end{aligned} \tag{13}$$

where q_{π_0} ($\pi_0 = 0.8$ and 0.9) is the quantile of interest. To simulate the P-values of the second GWAS, we took pleiotropy between two GWAS into account. Specifically, we used $0 \leq \gamma \leq 1$ to control the number of genes with pleiotropic effects as $G\pi_1(\pi_1 + \gamma\pi_0)$. When $\gamma = 0$, the number of genes shared by two GWAS equals $G\pi_1^2$ which is exactly the shared proportion when the two GWAS are independent of each other. When $\gamma = 1$, the number of risk genes shared by two GWAS equals $G\pi_1$, i.e. all of the risk genes are shared by the two GWAS. The risk genes of the second GWAS were randomly selected such that the total number of genes with pleiotropic effects was $G\pi_1(\pi_1 + \gamma\pi_0)$ (note that the total proportion of risk genes in the second GWAS was kept to π_1 for simplicity). After the group membership was generated for each gene, the P-values of the second GWAS were simulated from $\mathcal{B}(\alpha_2, 1)$ and $\mathcal{U}(0,1)$ for risk genes and null genes, respectively.

First, we evaluated the type I errors of testing tissue-specific differential expression in a separate-GWAS analysis (9) and a joint-GWAS analysis (10), and the type I errors of testing pleiotropy (11). In this simulation, we set the number of genes $G = 20\,000$, the number of tissues $T = 5$, the effect size $\mathbf{b} = \mathbf{0}$, and the pleiotropy controlling parameter $\gamma = 0$. For simplicity, we chose $\alpha_1 = \alpha_2 = \alpha \in \{0.3, 0.4, 0.5, 0.6\}$ (the smaller α, the stronger GWAS signals) and $\pi_0 \in \{0.8, 0.9\}$. For each parameter setting, we ran the EPS 500 times. The experimental results are shown in Figure 1. Clearly, all type I errors were well controlled at the nominal level ($P = 0.05$). We noticed that the test of pleiotropy was underpower when the GWAS signal was weak (e.g. $\alpha \geq 0.5$) and the proportion of risk genes was small (e.g. $\pi_1 \geq 0.9$). More detailed information on these three tests is summarized in the QQ-plots in Supplementary Figures S1 and S2.

Next, we conducted a simulation to evaluate the power of the proposed method. We started with a small number of tissues, i.e. $T = 5$. We simulated $y_g = \mathbf{E}_g\mathbf{b} + \epsilon_g$, where the variance of ϵ_g was specified such that the signal-to-noise ratio (SNR) was controlled at around 2 (note that the SNR is defined as $\sqrt{\text{Var}(\mathbf{E}_g\mathbf{b})/\text{Var}(\epsilon_g)}$). We varied γ to check the influence of pleiotropy on the performance of gene prioritization. To quantitatively evaluate the performance of EPS, we calculated the area under the curve (AUC) and the power with the global FDR controlled at 0.2. The results are shown in Figure 2. Comparing the red boxplots (separate-GWAS analysis with gene expression) with the blue ones (separate-GWAS analysis without gene expression), we observed a clear improvement after incorporating informative gene expression data from T tissues. The benefit of integrating pleiotropy was also confirmed by comparing the green boxplots (joint-GWAS analysis without gene expression) with the blue ones (separate-GWAS analysis without gene expression). Integrating both pleiotropy and gene expression from multiple tissues further improved the performance of EPS, as indicated by the yellow boxplots (joint-GWAS analysis with gene expression). However, there are two remaining concerns: how are the results of the joint analysis affected when there is no pleiotropy, and how is the analysis affected when all of the incorporated tissues are non-informative? For the first concern, as shown in Figure 2, the joint analysis in EPS showed comparable performance with a separate

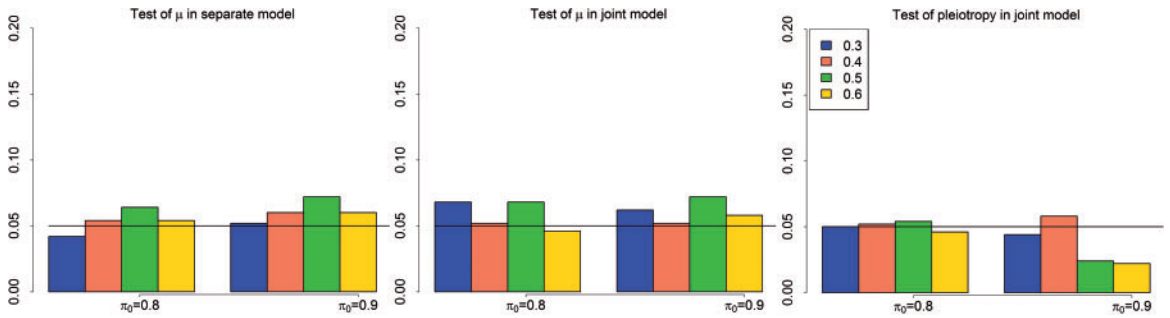

Fig. 1. Evaluation of type I errors of testing pleiotropy (11) and testing enrichment in single-GWAS (9) and two-GWAS (10) with parameter setting: $G = 20\,000$, $T = 5$, $\pi_0 \in \{0.8, 0.9\}$, $\alpha \in \{0.3, 0.4, 0.5, 0.6\}$ (Color version of this figure is available at *Bioinformatics* online.)

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

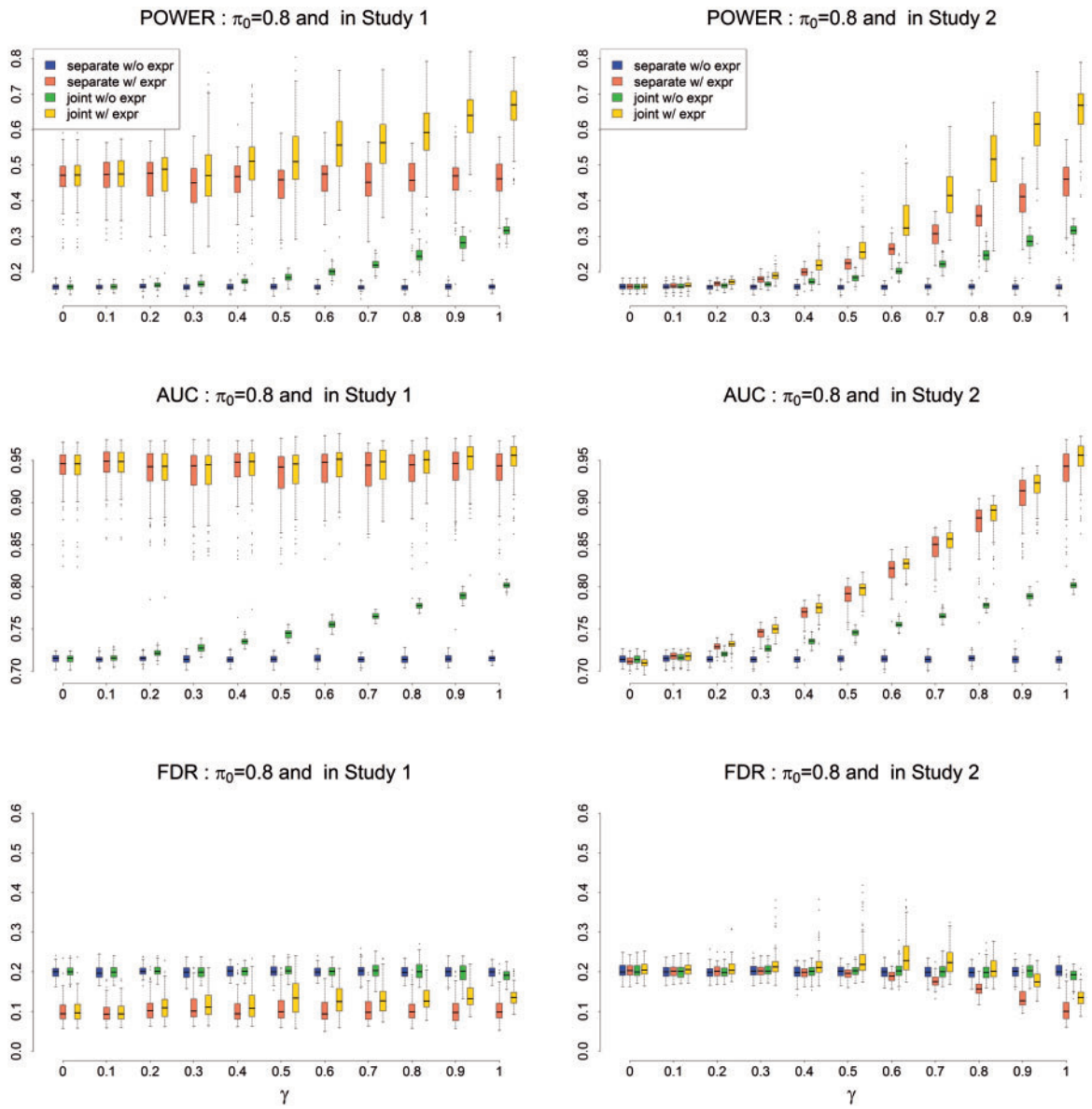

Fig. 2. Power (upper panel), AUC (middle panel) and FDR (lower panel) of EPS for gene prioritization with parameter setting: $G = 20\,000$, $T = 5$, $\pi_0 = 0.8$ and $\alpha_1 = \alpha_2 = 0.4$. The results are based on 500 simulations (Color version of this figure is available at *Bioinformatics* online.)

analysis in the absence of pleiotropy ($\gamma = 0$), suggesting that EPS can adapt to available pleiotropic information and avoid overfitting. To address the second concern, we performed a simulation study in which all of the tissues were non-information. As shown in Supplementary Figure S3, EPS performed robustly when a small number of non-informative gene expressions were incorporated into the analysis.

As observed in Fan and Fan (2008), inclusion of too many non-informative variables will generate too much noise, and thus degrade the performance. In such a case, we advocate using the extended model (Section 3), which uses penalized LDA to adaptively remove irrelevant tissues. To evaluate the performance of the extended model, we simulated gene expression from $T = 100$ tissues, only five of which were informative. The simulation results (Supplementary Fig. S4) suggest that the extended model performed very well for gene prioritization, even when most of the issues were non-informative. In addition to performance comparison, we investigated the accuracy of parameter estimation. These results are shown in Supplementary Figures S5 and S6. We compared EPS with CMFDR, which was designed to integrate one GWAS with functional annotation. The results shown in the left panel of Supplementary Figure S7 demonstrate that EPS performs better than CMFDR in terms of the AUC (top panel of Supplementary Fig. S7). CMFDR seems to have a greater power (middle panel of Supplementary Fig. S7), but it suffers from an uncontrolled FDR (lower panel of Supplementary Fig. S7).

Another issue with integrating multiple GWAS is the overlap of control samples. Although we did not take this issue into account in our analysis, we investigated its potential effects on EPS using simulation studies. The empirical results (shown in Supplementary Fig. S8) indicate that the FDR of EPS is indeed inflated in some extreme cases, but the inflation is small or moderate for the majority of cases (the true FDR is 0.3 where nominal FDR is 0.2). We plan to address this limitation in our future work.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

4.2 Analysis of schizophrenia (SCZ) and bipolar disorder (BPD) with the GTEx data

In this section, we applied EPS to the analysis of BPD and SCZ. Previous studies have shown that BPD and SCZ share some common polygenic variation and susceptibility genes (Maier *et al.*, 2006; Purcell *et al.*, 2009). Therefore, it is ideal to jointly analyze BPD and SCZ by considering the pleiotropy between them. Detailed information about these GWAS is provided in Cross-Disorder Group of the Psychiatric Genomics Consortium (2013b). We downloaded the summary statistics for BPD and SCZ from the Psychiatric Genomics Consortium (PGC) website. There are 1 233 533 and 1 237 959 SNPs reported in the PGC GWAS for BPD and SCZ, respectively. We took the intersection of these available SNPs and obtained 1 219 805 overlapping SNPs and their *P*-values. As all of the samples are of European ancestry, we used VEGAS2 (Liu *et al.*, 2010; Mishra and Macgregor, 2015) to combine *P*-values at the SNP level and obtained 17 763 *P*-values at the gene level, where 379 European ancestry samples in 1000 genome data served as the reference panel in the VEGAS method.

The GTEx (Lonsdale *et al.*, 2013) provides an up-to-date resource for the gene expression data in multiple tissues (http://www.gtexportal.org/home/). As of April 2015, the gene expression data for 53 tissues from 2921 samples were available on GTEx. As the sample sizes of some tissues may have been too small to provide reliable analysis results, we only considered the tissues that had at least 16 samples. We collected gene expression data from 46 tissues for analysis. For each tissue, there were 25 208 gene Ensembl IDs. We matched these genes with the list of genes from the output of VEGAS, and ended up with gene expression data for 13 815 genes from 46 tissues and their *P*-values from SCZ and BPD at the gene level. We further performed quantile normalization of the gene expression data. In practice, the expression of genes may not be independent, i.e. the rows of matrix **E** could be correlated. Because independence among genes is assumed in EPS, we performed additional simulation studies to evaluate the type I errors using this real data. The results shown in Supplementary Figures S7 and S8 indicate that EPS performed robustly in the presence of a correlation.

Next, we checked the distribution of the *P*-values from VEGAS and found that these gene-level *P*-values were inflated for both SCZ and BPD (the QQ-plots of these *P*-values are shown in Supplementary Fig. S9). Therefore, we chose to use the empirical null for the null distribution as described in Section 2.3.4. The parameters α_{null}s were estimated to be 0.72 and 0.80 for SCZ and BPD, respectively.

To identify tissue-specific-enrichment patterns of a complex trait, we applied EPS to integrate multiple tissues with their *P*-values from GWAS and then evaluated whether risk genes were differentially expressed in a tissue-specific manner. We performed this analysis for all 46 tissues using hypothesis testing (9) (a complete list of the test results is given in Supplementary Table S1). The test results suggested that the risk genes for BPD are differentially expressed in brain–cerebellar hemisphere and brain–cerebellum. Although a number of studies have suggested that the cerebellum plays a critical role in emotion processing (Hoppenbrouwers *et al.*, 2008), its importance in BPD is still uncertain (DelBello *et al.*, 1999; Mills *et al.*, 2005). A recent study based on quantitative T1ρ mapping (Johnson *et al.*, 2015) investigated brain abnormalities in BPD and highlighted the critical role of the cerebellum in BPD, which is consistent with the result of our analysis. The risk genes for SCZ also showed differential expression in adrenal-gland, pituitary, EBV-ETL cells, spleen, testis and whole-blood, in addition to the brain cerebellum regions. The significance in both the adrenal-gland and pituitary supports the important role of the hypothalamic–pituitary–adrenal axis function in SCZ (Walker *et al.*, 2008). The evidence that differential expression of SCZ risk genes in 'Cell–EBV–ETL' supports the recently discovered link between SCZ and the immune system (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Although the biological explanation of the differentially expressed genes in the spleen, testis and whole blood remains elusive, our method suggests that they warrant significant consideration. We briefly explain the reason for this in the next paragraph where we discuss the gene prioritization results of our method.

The gene prioritization results with or without the inclusion of gene expression data are summarized using Manhattan plots in Figure 3. To control false positive findings, we use the local FDR $\leq$ 0.1 as the criterion to report risk genes in the remainder of this section [the threshold for the local FDR, set at 0.1, is more stringent than the default setting of 0.2 in the classical local-FDR approach (Efron, 2010) (also see the 'locfdr' R package). In fact, setting the threshold for the local FDR at 0.1 often leads to a global FDR of around 0.05 (Efron, 2010)]. Without gene expression, EPS identified the same nine risk genes for both SCZ and BPD (see Supplementary Table S2 for details). For example, EPS identified two genes at 6p21-p22.1: *NKAPL* and *PGBD1*. These two genes have been identified as SCZ-risk genes in the Han Chinese population (Yue *et al.*, 2011). In this Han Chinese population, rs2142731 is the most significant variant in *PGBD1*, while its *P*-value is 0.7387 based on the PGC samples from European ancestry. Similarly, the *P*-value of rs1635 in an exon of *NKAPL* is 6.91×10^{-12} in the Han Chinese population, while its *P*-value is 0.7866 in the PGC samples. As genetic variants may have different allele frequencies and LD structures across different populations, it is difficult to replicate GWAS findings at the SNP level. However, genes can be highly consistent across ancestry (Neale and Sham, 2004). Here, EPS made use of the *P*-value at the gene level and successfully

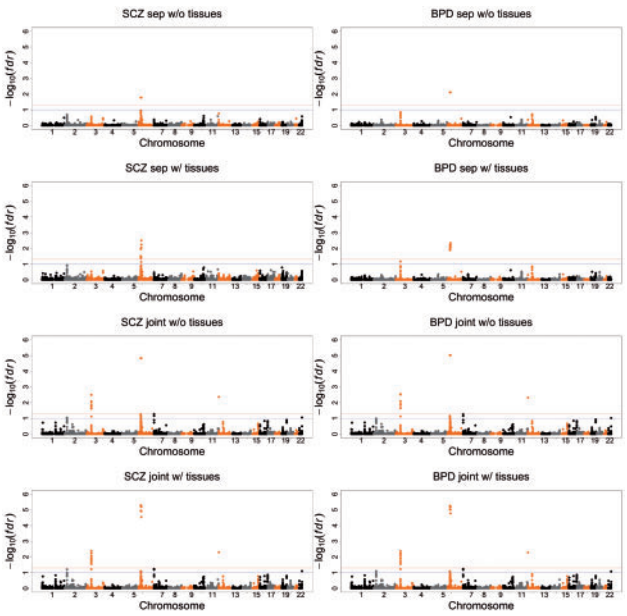

Fig. 3. Manhattan plots of BPD and SCZ. From top to bottom: separate analysis of BPD and SCZ without the GTEx gene expression data; separate analysis of BPD and SCZ with the GTEx gene expression data. Joint analysis of BPD and SCZ without the GTEx gene expression data; joint analysis of BPD and SCZ with the GTEx gene expression data. The red and blue lines indicate local FDR = 0.05 and 0.1, respectively. See detailed explanation in the main text

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

replicated these findings. According to the GTEx data, *PGBD1* had a relatively high expression in the brain–cerebellum and brain–cortex, suggesting that *PGBD1* may be more active in the brain than in other tissues. Another risk gene *NKAPL* had a relatively high expression in the testis, which explained why EPS detected some differentially expressed genes in the testis. These tissue-specific expression patterns of *PGBD1* and *NKAPL* are shown in Supplementary Figures S10 and S11.

When incorporating the gene expression data into our analysis, we did not use only the gene expression data from the significant tissues because using data twice may lead to a selection bias (Tibshirani and Efron, 2002). Instead, we incorporated the gene expression data from all tissues, and performed the analysis using penalized LDA as discussed in Section 3.2. The EPS results suggest that the gene *FLOT1* may be a risk gene for SCZ, but not for BPD, which is consistent with a very recent large sample study (Andreassen *et al.*, 2015). In fact, the GTEx data suggest that *FLOT1* is highly expressed not only in the blood, but also in adrenal-gland, indicating that *FLOT1* may be related to the functioning of the hypothalamic–pituitary–adrenal axis.

To demonstrate the possible benefits of leveraging the pleiotropic effects among different traits, we compared the separate GWAS analysis and the joint SCZ–BPD analysis without incorporating gene expression. In the joint analysis, EPS detected strong pleiotropy between SCZ and BPD, as indicated by the *P*-value $< 1 \times 10^{-30}$ [the estimated proportions are $\hat{\pi}_{00} = 0.972$, $\hat{\pi}_{10} = 0.001$, $\hat{\pi}_{01} = 0.001$ and $\hat{\pi}_{11} = 0.026$, then the test statistic (12) equals 209.35]. Leveraging the pleiotropic effects enabled EPS to identify many more risk genes than the separate analyses of SCZ and BPD, as shown in Supplementary Table S2. For example, the identification of the gene *CACNA1C* in the joint analysis is consistent with a recent GWAS with a larger sample size (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) and a study using the conditional FDR (Andreassen *et al.*, 2013). The gene *CACNA1C*, which encodes an alpha-1 subunit of a voltage-dependent calcium channel, has played an important role in the development of calcium antagonist drugs. As reported in Solovieff *et al.* (2013), this finding has been highlighted in clinical trials of calcium antagonist drugs, which are potentially effective in the treatment of psychiatric disorders.

Finally, we performed an integrative analysis of BPD–SCZ with the gene expression data from all tissues. As shown in Supplementary Table S2, the incorporation of the gene expression data enhanced the association signals of some genes. For example, *FNDC4* and *NT5DC2* were observed to be highly expressed in adrenal-gland, based on the GTEx data (Supplementary Figs S10 and S11). EPS also indicated that SCZ and BPD risk genes tend to be differentially expressed in the adrenal-gland. Combining all of these pieces of evidence, the strength of association between *FNDC4* and *NT5DC2* was adaptively enhanced by EPS. The association signals may also be weakened after incorporating gene expression data into the analysis. For example, *HIST1H2BN* is non-significant with a local FDR ≤ 0.1 because its expression levels are very low in all of the tissues from the GTEx data. To the best of our knowledge, there is no strong evidence that *HIST1H2BN* is a risk gene for SCZ or BPD. By incorporating gene expression data, EPS was able to reduce the association strength and possibly avoid a false positive finding.

5 Conclusion

We have presented a statistical approach, named EPS, that can integrate pleiotropy information from GWAS data and tissue-specific gene expression data. Compared with some existing approaches, such as linear mixed models, which require genotype data at the individual level, EPS only requires summary statistics for analysis. More importantly, EPS provides statistically rigorous evaluations of tissue-specific gene expression patterns and pleiotropic effects via hypothesis testing. These merits make EPS an attractive and effective tool for the integrative analysis of GWAS data with gene expression data from multiple tissues. Despite the promising statistical improvements we have made, the biological implications need to be independently replicated. One limitation of EPS is that it does not take into account overlapping samples. Addressing this issue in integrating multiple GWAS is an important area for future work.

Funding

This work was supported in part by grant NO. 61501389 from National Natural Science Foundation of China (NSFC), grants HKBU_22302815 and HKBU_12202114 from Hong Kong Research Grant Council, and grants FRG2/14-15/069 and FRG2/14-15/077 from Hong Kong Baptist University, and Duke-NUS Medical School WBS: R-913-200-098-263.

Conflict of Interest: none declared.

References

Andreassen,O. *et al.* (2015) Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. *Mol. Psychiatry*, **20**, 207–214.

Andreassen,O.A. *et al.* (2013) Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.*, **92**, 197–209.

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

Civelek,M. and Lusis,A.J. (2014) Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.*, **15**, 34–48.

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

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

DelBello,M.P. *et al.* (1999) MRI analysis of the cerebellum in bipolar disorder: a pilot study. *Neuropsychopharmacology*, **21**, 63–68.

Efron,B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume **1**. Cambridge University Press, Cambridge, UK.

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

Eicher,J.D. *et al.* (2015) GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res.*, **43**, D799–D804.

Fan,J. and Fan,Y. (2008) High dimensional classification using features annealed independence rules. *Ann. Stat.*, **36**, 2605.

Flutre,T. *et al.* (2013) A statistical framework for joint EQTL analysis in multiple tissues. *PLoS Genet.*, **9**, e1003486.

Franke,A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

Genetic Analysis of Psoriasis Consortium and the Wellcome Trust Case Control Consortium (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.*, **42**, 985–990.

Gutierrez-Arcelus,M. *et al.* (2015) Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*, **11**, e1004958.

Hoppenbrouwers,S.S. *et al.* (2008) The role of the cerebellum in the pathophysiology and treatment of neuropsychiatric disorders: a review. *Brain Res. Rev.*, **59**, 185–200.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

Johnson,C. *et al*. (2015) Brain abnormalities in bipolar disorder detected by quantitative t1ρ mapping. *Mol. Psychiatry*, **20**, 201–206.
Lee,D. *et al*. (2015) JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*, **31**, 1176–1182.
Li,M.X. *et al*. (2011) GATES: a rapid and powerful gene-based association test using extended SIMES procedure. *Am. J. Hum. Genet*., **88**, 283–293.
Liu,J.Z. *et al*. (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet*., **87**, 139–145.
Liu,X. *et al*. (2008) Tiger: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.
Lonsdale,J. *et al*. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet*., **45**, 580–585.
Loos,R.J. *et al*. (2008) Common variants near mc4r are associated with fat mass, weight and risk of obesity. *Nat. Genet*., **40**, 768–775.
Maier,W. *et al*. (2006) Schizophrenia and bipolar disorder: differences and overlaps. *Curr. Opin. Psychiatry*, **19**, 165–170.
Mills,N.P. *et al*. (2005) MRI analysis of cerebellar vermal abnormalities in bipolar disorder. *Am. J. Psychiatry*, **162**, 1530–1532.
Mishra,A. and Macgregor,S. (2015) Vegas2: software for more flexible gene-based testing. *Twin Res. Hum. Genet*., **18**, 86–91.
Moore,H.M. (2013) Acquisition of normal tissues for the GTEx program. *Biopreserv. Biobank*., **11**, 75–76.
Neale,B.M. and Sham,P.C. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet*., **75**, 353–362.
Purcell,S.M. *et al*. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.
Ritchie,M.D. *et al*. (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet*., **16**, 85–97.
Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.
Schwartzman,A. (2008) Empirical null and false discovery rate inference for exponential families. *Ann. Appl. Stat*., **2**, 13321359.
Shriner,D. (2012) Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Front. Genet*., **3**, 1.
Solovieff,N. *et al*. (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet*., **14**, 483–495.
Stearns,F.W. (2010) One hundred years of pleiotropy: a retrospective. *Genetics*, **186**, 767–773.
Thomas,G. *et al*. (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet*., **40**, 310–315.
Tibshirani,R.J. and Efron,B. (2002) Pre-validation and inference in microarrays. *Stat. Appl. Genet. Mol. Biol*., **1**, Article 1.
Tomlinson,I. *et al*. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet*., **39**, 984–988.
Torres,J.M. *et al*. (2014) Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *Am. J. Hum. Genet*., **95**, 521–534.
Van der Vaart,A.W. (2000). *Asymptotic Statistics*, volume **3**. Cambridge University Press, Cambridge, UK.
Visscher,P.M. *et al*. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet*., **90**, 7–24.
Walker,E. *et al*. (2008) Stress and the hypothalamic pituitary adrenal axis in the developmental course of schizophrenia. *Annu. Rev. Clin. Psychol*., **4**, 189–216.
Wang,Q. *et al*. (2015) Pervasive pleiotropy between psychiatric disorders and immune disorders revealed by integrative analysis of multiple GWAS. *Hum. Genet*., **134**, 1195–1209.
Welter,D. *et al*. (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., **42**, D1001–D1006.
Witten,D.M. and Tibshirani,R. (2011) Penalized classification using fisher's linear discriminant. *J. R. Stat. Soc. B*, **73**, 753–772.
Yang,C. *et al*. (2015) Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. *Front. Genet*., **6**, 229.
Yue,W.H. *et al*. (2011) Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nat. Genet*., **43**, 1228–1231.
Zablocki,R.W. *et al*. (2014) Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*, **30**, 2098–2104.
Zhou,X. and Stephens,M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407–409.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016