BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 9 2011, pages 1201–1206
doi:10.1093/bioinformatics/btr116

Genome analysis

Advance Access publication March 12, 2011

PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies

Jie Huang[1,2], Andrew D. Johnson[1,2] and Christopher J. O'Donnell[1,2,3,*]

[1]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA 01702, [2]Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, MD 20824 and [3]Cardiology Division, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The concept of pleiotropy was proposed a century ago, though up to now there have been insufficient efforts to design robust statistics and software aimed at visualizing and evaluating pleiotropy at a regional level. The Pleiotropic Region Identification Method (PRIMe) was developed to evaluate potentially pleiotropic loci based upon data from multiple genome-wide association studies (GWAS).

Methods: We first provide a software tool to systematically identify and characterize genomic regions where low association *P*-values are observed with multiple traits. We use the term Pleiotropy Index to denote the number of traits with low association *P*-values at a particular genomic region. For GWAS assumed to be uncorrelated, we adopted the binomial distribution to approximate the statistical significance of the Pleiotropy Index. For GWAS conducted on traits with known correlation coefficients, simulations are performed to derive the statistical distribution of the Pleiotropy Index under the null hypothesis of no genotype–phenotype association. For six hematologic and three blood pressure traits where full GWAS results were available from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, we estimated the trait correlations and applied the simulation approach to examine genomic regions with statistical evidence of pleiotropy. We then applied the approximation approach to explore GWAS summarized in the National Human Genome Research Institute (NHGRI) GWAS Catalog.

Results: By simulation, we identified pleiotropic regions including *SH2B3* and *BRAP* (12q24.12) for hematologic and blood pressure traits. By approximation, we confirmed the genome-wide significant pleiotropy of these two regions based on the GWAS Catalog data, together with an exploration on other regions which highlights the *FTO*, *GCKR* and *ABO* regions.

Availability and Implementation: The Perl and R scripts are available at http://www.framinghamheartstudy.org/research/gwas_pleiotropictool.html.

Contact: odonnellc@nhlbi.nih.gov

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

Received on October 27, 2010; revised on February 8, 2011; accepted on February 24, 2011

*To whom correspondence should be addressed.

1 INTRODUCTION

Pleiotropy describes the effect of a single genetic region on multiple phenotypic traits. The concept of pleiotropy was first defined a century ago as one mutation resulting in multiple distinct phenotypes (Stearns, 2010). Among the proposed mechanisms, the genetic region may encode a product that is used by multiple cell types, or it may have a signaling function affecting multiple targets, or the traits under study may themselves be highly interrelated at a physiological level (Hodgkin, 1998; Pyeritz, 1989). Pleiotropic genes have been shown in other species to affect environmental adaptation and tend to reside in central node positions in protein–protein interaction networks (Foster *et al*., 2004; Zou *et al*., 2008). The identification and characterization of pleiotropic genes and regions offers a unique window into the complexities of biological molecular interaction networks, and may potentially indicate evidence for epistasis (Tyler *et al*., 2009).

The presence and impact of pleiotropy in genome data for normal human characteristics and human disease traits merits further investigation, but thus far the efforts at statistical development are insufficient. Methods and software for analyzing multivariate phenotypes have been proposed (Ferreira and Purcell, 2009; Lange *et al*., 2003; Liu *et al*., 2009; Yang *et al*., 2010). However, the available methods, such as those by Ferreira and Purcell (2009), Lange *et al*. (2003) and Liu *et al*. (2009), all require use of individual-level phenotype data and thus cannot be used to study pleiotropy using only existing summarized genome-wide association studies (GWAS) results. The method of Yang *et al.* (2010) can be used on existing summarized GWAS results; however this method only considers single nucleotide polymorphism (SNP) level but not region-level pleiotropy.

A novel analytic approach was recently demonstrated to examine pleiotropic genes in psychiatric phenotypes (Huang *et al*., 2010). The current abundance of GWAS results provides an unprecedented opportunity to fully examine this phenomenon in a systematic manner. We previously analyzed results across 118 GWAS articles published from 2005 through 2008, creating a comprehensive database of 56 411 SNP-phenotype associations at a significance level of ≤ 0.001 (Johnson and O'Donnell, 2009). This study highlighted potential pleiotropic regions, presenting the 61 densest regions of associations from 118 GWAS. Polymorphisms associated with multiple traits in genes including *APOE* and the *MHC* region were identified, as well as novel candidates (*PIGU*, *RAPGEF1*, *COL4A1/2* and *OAS1*) that were subsequently replicated in other

studies (Johnson and O'Donnell, 2009). Another resource for the collection of significant GWAS associations is the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies (GWAS Catalog), which is regularly updated. A summary of results from 151 (of 237) published GWAS through December 2008 described 531 replicated SNP-trait associations and highlighted 18 regions of association with two or more traits (Hindorff *et al*., 2009). Since that time GWAS have continued to be published at a rapid pace, as of October 16, 2010, including 584 GWAS and 4054 SNP-trait associations reported in the GWAS Catalog. Other repositories of GWAS-related data exist such as the database of Genotypes and Phenotypes (dbGAP) but thus far these have not been widely applied to address hypotheses relating to pleiotropy (Mailman *et al*., 2007).

We developed the Pleiotropic Region Identification Method (PRIMe) and applied it to cardiovascular-related traits from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (Psaty *et al*., 2009). We provide an initial example involving six hematologic traits (hemoglobin concentration, hematocrit, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume and red blood cell count) and three blood pressure traits (systolic blood pressure, diastolic blood pressure and hypertension) used in two independent GWAS meta-analysis reports from the CHARGE Consortium (Ganesh *et al*., 2009; Levy *et al*., 2009). We further evaluated an approximation approach to estimate the statistical distribution of the Pleiotropy Index for uncorrelated traits, using data reported in the publicly accessible GWAS Catalog.

2 METHODS

2.1 PRIMe visualization

We implemented the PRIMe tool to systematically identify and characterize regions across multiple GWAS/traits. Here we use the terms 'GWAS' and 'trait' interchangeably when one GWAS reports association results for one phenotypic trait. However, the CHARGE datasets we used for pleiotropic analysis include two overall GWAS studies with a total of nine traits. The PRIMe visualization method defines a genomic region of interest out of the whole genome as follows: let P_S denote the threshold for association significance of SNPs, which can be user defined. Let r denote the correlation coefficient between an SNP pair, measured as the square root of the linkage disequilibrium (LD) measure of r^2. PRIMe iteratively finds SNPs with the lowest association P-value among all traits as the *driver*, and SNPs whose r^2 with the *driver* is above the user-specified threshold ($\geq$0.8 by default) as *passengers*. In order to define distinct regions out of a genome with extensive LD patterns, once a SNP is designated as a *passenger*, it will not be considered again as a new *driver* or *passenger*. After completion of this iterative process, regions are defined by one *driver* and zero or more *passengers*.

2.2 Statistical evaluation

To follow-up regions identified by the PRIMe visualization tool, we conducted statistical evaluation for scenarios ranging from uncorrelated to highly correlated traits. We use the term Pleiotropy Index to denote the number of traits with low association P-values at a particular genomic region. We estimated the statistical distribution of the Pleiotropy Index under the null hypothesis of no genotype–phenotype association for any of the traits. Let P_T denote the probability of the reaching a certain Pleiotropy Index value in a genomic region. Since the calculation of P_T depends on the particular LD pattern of a genomic region, which varies substantially across the genome, the same Pleiotropy Index value at different genomic regions would yield different P_T, and thus are not comparable. Therefore, we do not attempt to address the probability of attaining a Pleiotropy Index for any region across the genome *per se*. Instead, we assume that a region of interest has been identified through basic PRIMe queries or other means, and the main objective is to evaluate statistical significance for that particular region. We evaluated the simulated or approximated P_T value against the genome-wide significance threshold of 5×10^{-8} to assess pleiotropy (Pe'er *et al*., 2008).

Table 1. Correlation of six hematologic and three blood pressure traits

P	Hb	HCT	MCH	MCHC	MCV	RBC	DBP	SBP	HTN
Hb	1.00	0.88	0.17	0.18	0.12	0.54	0.11	0.09	0.07
HCT		1.00	0.06	−0.03	0.12	0.55	0.1	0.08	0.06
MCH			1.00	0.37	0.64	−0.23	0.02	0.00	0.00
MCHC				1.00	−0.01	0.07	0.04	0.03	0.02
MCV					1.00	−0.2	−0.02	−0.02	−0.03
RBC						1.00	0.07	0.05	0.04
DBP							1.00	0.72	0.56
SBP								1.00	0.64
HTN									1.00

Hb, hemoglobin concentration; HCT, hematocrit; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; DBP, diastolic blood pressure; SBP, systolic blood pressure; HTN, hypertension.

2.2.1 For uncorrelated traits: the binomial approximation When GWAS are conducted across different cohorts with non-overlapping samples, they can be assumed to be uncorrelated, under the assumption that there is no genotype–phenotype association for each individual GWAS. The number of independent SNPs within a genomic region could also be derived using the pairwise tagging approach (by default, r^2 threshold of 0.8) (de Bakker *et al*., 2005). For K uncorrelated traits and M independent SNPs, the statistical distribution of the Pleiotropy Index can be mathematically derived as follows. In a region with M independent SNPs, the number of SNPs with association P-values below P_S follows a binomial distribution $B(n,p)$, with n equal to M and p equal to P_S. Therefore, the probability that at least one out of M SNPs with association P-value below P_S is equal to $1-(1-P_S)^M$ because $(1-P_S)^M$ is the probability that none of the SNPs have a P-value below the threshold. For K uncorrelated traits each with M SNPs in the region, the number of traits with at least one SNP below P_S also follows a binomial distribution $B(n,p)$, this time with n equal to K and p equal to $1-(1-P_S)^M$. Therefore, for uncorrelated traits and independent SNPs, P_T can be approximated from a simple binomial approach.

2.2.2 For correlated traits: the multivariate simulation In the backdrop of deep phenotyping and large consortia, GWAS scans could be correlated due to both overlapping samples and similar phenotypic measurements. We estimated the overall trait correlation by calculating the correlation of z-statistics (β/SE) for all common SNPs (e.g. ~2.5 million SNPs based on HapMap2 imputation) in a number of GWAS from consortia (Table 1, Supplementary Tables S1 and S2). For K correlated traits in a region of M SNPs with known LD, the test statistics (z) of each SNP for each trait follows approximately a multivariate normal distribution (Conneely and Boehnke, 2007).

The probability that a trait indexed by k has at least one SNP with a P-value below the threshold of P_S is equivalent to the probability that the maximum of the absolute value of $z_{k1}, z_{k2}, \ldots z_{kM}$ statistics is greater than z_s (the z-value corresponding to the $P_S/2$ two-sided threshold). Theoretically, this probability can be derived from a multiple dimension integral of the multivariate normal density function of the z-statistics. However, this multiple dimension integral is challenging to evaluate numerically when the number of SNPs is large. Moreover, using the multivariate normal

Let $Cor(y_i, y_j) = \rho$ denote trait correlation

Let $Cor(g_{i1}, g_j) = r$ denote genotype correlation

$$\beta_{ij} = \frac{\sum (y_{ik} - \overline{y_i})(g_{ij} - \overline{g_j})}{\sum (g_{jk} - \overline{g_j})^2} \qquad SE(\beta_{ij}) = \sqrt{\frac{\sigma_e^2}{\sum (g_{jk} - \overline{g_j})^2}}$$

$$Z_{ij} = \frac{\sum (y_{ik} - \overline{y_i})(g_{jk} - \overline{g_j})}{\sqrt{\sigma_e^2 \sum (g_{jk} - \overline{g_j})^2}} \qquad E(Z_{ij}) = 0$$

$$E(Z_{ij}Z_{xy}) = \frac{E\left[\sum_k (y_{ik} - \overline{y_i})(g_{jk} - \overline{g_j}) \sum_l (y_{xl} - \overline{y_x})(g_{yl} - \overline{g_y})\right]}{N\sigma_e^2 \sigma_{gj} \sigma_{gk}}$$

$$= \frac{\sum_{kl} \left\{E\left[(y_{ik} - \overline{y_i})(y_{xl} - \overline{y_x})(g_{jk} - \overline{g_j})(g_{yl} - \overline{g_y})\right]\right\}}{N\sigma_e^2 \sigma_{gj} \sigma_{gk}}$$

$$= \frac{1}{N} \sum_k \frac{E\left[(y_{ik} - \overline{y_i})(y_{xk} - \overline{y_x})\right]}{\sigma_e^2} \frac{E\left[(g_{jk} - \overline{g_j})(g_{yk} - \overline{g_y})\right]}{\sigma_{gj} \sigma_{gk}}$$

$$= \frac{1}{N} N\rho_{ix} r_{jy} = \rho_{ix} r_{jy}$$

Fig. 1. Mathematical derivation of the test statistics correlation for K traits and M SNPs. This is based on the assumption that there is no genotype–phenotype association for any of the traits.

approximation from one GWAS to derive the null distribution of the Pleiotropy Index would require taking the correlation among the GWAS into consideration. Therefore, we take a simulation approach to evaluate the distribution of the Pleiotropy Index. We simulate the test statistics for all traits and all SNPs simultaneously, and to do so we use the fact that the correlation matrix between all $K \times M$ test statistics is the (kronecker) product of the M dimensional correlation matrix between SNPs (r) and the K dimensional correlation matrix of the traits analyzed (ρ) (Fig. 1).

3 RESULTS

3.1 PRIMe visualization

Figure 2 shows adjacent regions within the 12q24 region, which were highlighted by PRIMe for association with blood pressure and hematologic traits in the CHARGE Consortium. In Figure 2, the SNP in blue is designated as the *driver* SNP due to its lowest association *P*-value among all traits. The *driver* SNP can also be associated with other traits (shown in blue with less significant *P*-values). The SNPs shown in bright red are designated as *passengers*. The SNPs shown in weaker red are those whose r^2 with the *driver* SNP do not reach the user-specified threshold (0.8 by default).

3.2 For correlated traits: the multivariate simulation

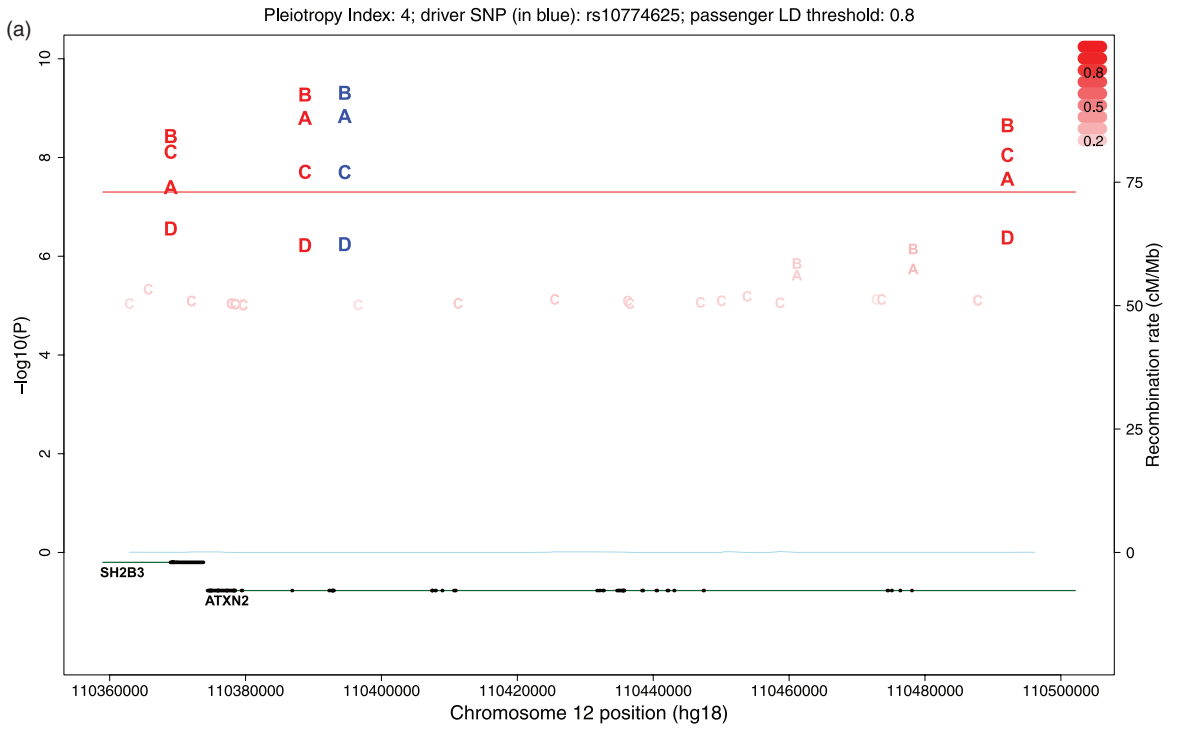
Table 1 shows the trait correlation matrix for nine hematologic and blood pressure traits from the CHARGE Consortium data. Correlation coefficients for these nine traits range widely from 0 to 0.88. To further illustrate the cause and extent of trait correlation, we calculated the pairwise correlation for two other publically available sets of data: the Wellcome Trust Case Control Consortium (WTCCC) and Global Lipids, as shown in Supplementary Tables S1 and S2 respectively (WTCCC 2007; Teslovich *et al.*, 2010). The latter two datasets are used only as references on trait correlation and not for pleiotropy analysis. We recommend that users of PRIMe consider any pair of traits with correlations <0.1 as largely uncorrelated.

For the nine hematologic and blood pressure traits, the two regions shown in Figure 2 (referred to as the *SH2B3* region and *BRAP* region hereafter) have the highest Pleiotropy Index of 4 among all identified regions with a Pleiotropy Index ≥ 2. For these nine traits in the *SH2B3* region, we performed 200 million simulations, and found only one instance where four of the nine traits all had at least one SNP with simulated association *P*-value (derived from the *z*-statistic) less than the pre-specified threshold of 1×10^{-5}. Therefore, we conclude that for these nine traits and the *SH2B3* region, based on a P_S of 1×10^{-5}, the simulated P_T value for the Pleiotropy Index of 4 is equal to $1/(2 \times 10^8) = 5.0 \times 10^{-9}$ (Supplementary Figure S1, left panel). Similarly, for these nine traits and using GWAS results from the *BRAP* region to estimate the SNP correlation, we performed 2×10^8 simulations and found six instances where four of the nine traits all had at least one SNP with simulated association *P*-value below the threshold. Therefore, the simulated P_T value for the Pleiotropy Index of 4 is equal to $6/(2 \times 10^8) = 3.0 \times 10^{-8}$ for the *BRAP* region (Supplementary Figure S1, right panel). Although the Pleiotropy Index is the same (i.e. 4), for both regions (out of a total of nine traits analyzed), the simulated P_T for the *BRAP* region is less significant than that for the *SH2B3* region, because the former resides in a longer haplotype block and, therefore, has a higher probability of observing more traits associated with SNPs in the region by chance. We conclude there is evidence that the proposed pleiotropy for these two regions is significant, because P_T is estimated to be $< 5.0 \times 10^{-8}$.

The *SH2B3* region was recently associated in a single report with platelet count, coronary artery disease, type 1 diabetes and celiac disease validating its likely pleiotropic effects (Soranzo *et al.*, 2009). In this region, the SNP *rs3185404* is non-synonymous and predicted to disrupt protein function. Furthermore, *SH2B3* is an important signaling protein and thus it is likely to affect cellular functions in a variety of human tissues (Barrett *et al.*, 2009; Hunt *et al.*, 2008).

3.3 For uncorrelated traits: the binomial approximation

We applied the PRIMe tool to the largest GWAS repository publicly available, the GWAS Catalog. For the same *SH2B3* and *BRAP* regions, Pleiotropy Indices of nine were observed for both regions (Supplementary Figures S2 and S3, respectively). There are no full-scale genome-wide *z* statistics to calculate trait correlations. Instead, we assume traits are not correlated for the 584 GWAS reported as of October 16, 2010. Given the independent number of SNPs (4 for the *SH2B3* region and 11 for the *BRAP* region), the calculated P_T is equal to 5.24×10^{-21} and 4.53×10^{-17}, respectively, based on the binomial approximation approach described above. Therefore, both regions are also deemed to demonstrate significant pleiotropy based on the GWAS Catalog data. A systematic analysis of data in the entire GWAS Catalog found that a total of 57 regions have a Pleiotropy Index of ≥ 5. The top regions include the *FTO* gene and



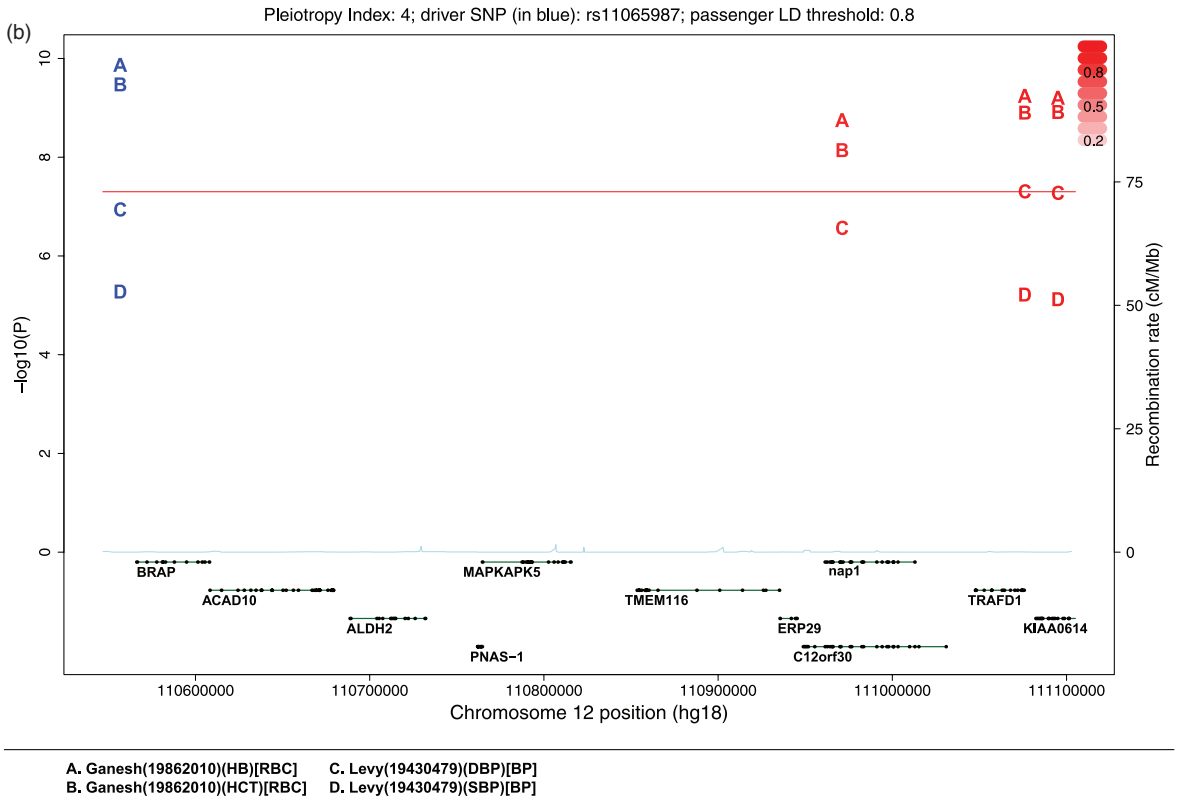


Fig. 2. PRIMe plot of the *SH2B3* region and the *BRAP* region for GWAS of six hematologic traits and three blood pressure traits. The *SH2B3* region (**a**) and the *BRAP* region (**b**). The driver SNPs (shown in blue) are *rs10774625* and *rs11065987*, respectively. A horizontal red line indicates the log of genome-wide significance *P*-value of 5×10^{-8}. The longest isoforms of RefSeq genes in the region are annotated including introns (green) and exons (black). The legend lists traits in the format of first_author(PubMed ID)(trait_name)[group_name].

the *GCKR* region (Supplementary Figures S4 and S5), both with a Pleiotropy Index of 18, while the well-known pleiotropic *ABO* region has a Pleiotropy Index of 9 (Supplementary Figure S6).

4 DISCUSSION

PRIMe does not rely on prespecified methods for genomic region grouping such as gene boundaries or physical position bins. Instead, association *P*-values for individual SNPs and LD between SNPs are used to define genomic regions where multiple traits show significant association with SNPs. Since pairwise LD (r) among SNPs is a key determinant of the width of a region; we used PLINK to precalculate LD based on the HapMap II + III samples from the CEU (Utah residents with ancestry from northern and western Europe) and TSI (Toscans in Italy) populations (Frazer *et al*., 2007; Purcell *et al*., 2007). We did not use the LD data downloadable from HapMap website, because it calculates LD only for SNPs up to 250 kb apart (Thorisson *et al*., 2005). To capture long-range haplotype blocks, we include LD for SNPs up to 1500 kb apart. Of note, the LD threshold chosen by investigators will not only affect the physical boundaries of a putative pleiotropic region, but also the statistical significance of the estimated P_T.

For the *SH2B3* and *BRAP* regions, if a LD threshold of $r^2 \geq 0.5$ rather than 0.8 is used, only one large region instead of two separate ones will be identified. This change of LD threshold will further complicate the choice of 5×10^{-8} as the threshold for evaluating genome-wide significance, since the total number of independent regions across the genome would be different when 'high LD' is defined as $r^2 \geq 0.5$ instead of 0.8. Furthermore, with a P_S threshold of 1×10^{-3} instead of 1×10^{-5} prespecified, the same *SH2B3* region is noted with a Pleiotropy Index of 6 but the simulated P_T would not have reached a genome-wide significant threshold. From a statistical testing perspective, higher values of Pleiotropy Index under less stringent P_S thresholds tend to result in a less significant P_T. On the other hand, a highly stringent P_S threshold will pose technical challenges for simulation. We performed 200 million simulations (i.e. $10/(5 \times 10^{-8})$) in an attempt to simulate the distribution for P_S up to 5×10^{-8}. However, as Supplementary Figure S1 shows, despite the high number of simulations, we still did not observe a Pleiotropy Index > 2 when P_S thresholds of 1×10^{-7} and 5×10^{-8} were used. Therefore, an exact P_T could not be specified other than $P_T < 1/(2 \times 10^8) = 5 \times 10^{-9}$. We suggest a P_S of 1×10^{-5}, which corresponds to the significance threshold currently used by the GWAS Catalog. Currently the direction of association effect is not considered for computing the Pleiotropy Index. Studies have suggested there may be biologically plausible mechanisms for different effect directions even among closely related phenotypes (Sirota *et al*., 2009).

For the approximation approach, we make the assumption that traits are not correlated. When traits studied in separate GWAS scans are not truly uncorrelated, the distribution of the Pleiotropy Index deviates from the binomial distribution, with higher proportions of correlated traits leading to greater deviations. As the correlation coefficient increases, the expected Pleiotropy Index increases and therefore is more likely (at any given significant P_T) to achieve larger values of the Pleiotropy Index (see Supplementary Figure S7). For the two CHARGE GWAS, five common cohorts contributed to a total of six cohorts for both GWAS, with the degree of sample overlap for these two GWAS being ~80%. However, a high degree of sample overlap alone does not imply that there will be high correlation for quantitative traits [e.g. 0.00 between mean corpuscular hemoglobin and systolic blood pressure for the CHARGE traits (Table 1), and −0.08 between HDL and LDL for the Global Lipids traits (Supplementary Table S2)]. For dichotomous traits illustrated by the WTCCC data, the use of a common control dataset generates an explicit overlap of phenotypic measurement, therefore leading to some degree of correlation among GWAS results for the seven traits (from 0.29 to 0.43, Supplementary Table S1). The highest correlation of 0.43 between type 1 diabetes and rheumatoid arthritis could also indicate common pathways for auto-immune diseases (Supplementary Table S1). For both quantitative and qualitative traits, under the null hypothesis of no genotype–phenotype association for each individual GWAS, even identically ascertained phenotypes will not yield trait correlation given no sample overlap.

The finding of the *FTO* region highlights how care must be taken in interpretation of conclusions on pleiotropy, since the commonly associated traits for the *FTO* region include many weight-related measures or for type 2 diabetes. Although a few of the association signals in the *GCKR* region are for similar serum lipids and glucose measures as well, there are distinct traits such as hematologic traits, serum uric acid and C-reactive protein. The significant test statistic in the case of *FTO* may be flagging either pleiotropy or further independent replication for adiposity traits, while the case of *GCKR* may reflect classic pleiotropy. For similar but not identical traits, suspected pleiotropy could be due to independent effects or through a chain of mediation. The identification of mediation can still provide useful biological information. A recent study on the genetics of *FTO* provides the first direct evidence that increased *FTO* expression causes obesity in mice. Mice with increased *FTO* expression on a high-fat diet develop glucose intolerance. The causal mechanism of *FTO* on human phenotypes warrants further gene expression profiling studies (Church *et al*., 2010). The null hypothesis of no genotype–phenotype association for each individual GWAS has been used as a basis for our simulation and approximation approaches, which could be invalid when traits with strong association with particular genomic regions have been well established and extensively replicated. Deriving the distribution under the alternative will depend on many variables (number of associated loci, strength of association with each trait, etc.) and could require a markedly different approach. We compared the power between our method and that of Ferreira and Purcell (2009), and found that their method is slightly more powerful under scenarios with various association *P*-value thresholds and type I error. However, our tool and method can be rapidly applied to summary GWAS results, together with other features including visualization.

In summary, we have developed a computational tool and statistical method to systematically identify, characterize and evaluate pleiotropic regions. PRIMe can be downloaded and run on local machines without uploading potentially sensitive GWAS data to public servers. The scripts are designed in a user friendly manner so that all customizable parameters can be specified in an interface script, which calls subsequent scripts that require no user manipulation. The availability of PRIMe may encourage sharing of more detailed GWAS data in a secure manner within collaborating groups in order to conduct discovery of potential pleiotropic regions. We anticipate that fine-mapping and re-sequencing will enable a

higher resolution view of pleiotropy, and further refine whether the same or distinct alleles in targeted regions contribute to multiple traits (Kitsios and Zintzaras, 2009). We have recently used PRIMe to help prioritize the selection of regions for sequencing in following up GWAS, gaining more potential return for smaller resource investments by focusing on regions with evidence for multiple traits. The value of this approach would be greatly improved by broader sharing of complete GWAS results among members of the scientific community with appropriate safeguards (Lumley and Rice, 2010), and the improvements in resources with more dense maps of SNPs provided by the 1000 Genomes and other sequencing projects that are underway. This new approach, together with efforts that pull together GWAS data into publicly accessible centralized repositories, is well suited to propel the study of the genetic architecture of complex diseases beyond individual investigators or individual datasets (Hindorff *et al*., 2009; Johnson and O'Donnell, 2009).

ACKNOWLEDGEMENTS

We acknowledge the invaluable expertise on the statistical simulation and power evaluation from Dr Josée Dupuis and Dr Qiong Yang at the Boston University Department of Biostatistics. This research was conducted in part using data and resources from the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. The analyses reflect intellectual input and resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. The authors acknowledge the essential role of the Cohorts for Heart and Aging Research in Genome Epidemiology (CHARGE) Consortium in development and support of this manuscript, especially acknowledging members from the CHARGE Consortium Blood Pressure and Hematological Working Groups for making their GWAS data fully available (Acknowledgement Section in Supplementary Material).

Funding: This project is supported by the National Heart, Lung and Blood Institute, Division of Intramural Research.

Conflict of Interest: none declared.

REFERENCES

Barrett,J.C. *et al*. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet.*, **41**, 703–707.

Church,C. *et al*. (2010) Overexpression of Fto leads to increased food intake and results in obesity. *Nat. Genet.*, **42**, 1086–1092.

Conneely,K.N. and Boehnke,M. (2007) So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests. *Am. J. Hum. Genet.*, **81**, 1158–1168.

de Bakker,P.I. *et al*. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.

Ferreira,M.A. and Purcell,S. M. (2009) A multivariate test of association. *Bioinformatics*, **25**, 132–133.

Foster,K.R. *et al*. (2004) Pleiotropy as a mechanism to stabilize cooperation. *Nature*, **431**, 693–696.

Frazer,K.A. *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Ganesh,S.K. *et al*. (2009) Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.*, **41**, 1191–1198.

Hindorff,L.A. *et al*. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Hodgkin,J. (1998) Seven types of pleiotropy. *Int. J. Dev. Biol.*, **42**, 501–505.

Huang,J. *et al*. (2010) Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression. *Am. J. Psychiatry*, **167**, 1254–1263.

Hunt,K.A. *et al*. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.*, **40**, 395–402.

Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.

Kitsios,G.D. and Zintzaras,E. (2009) Genomic convergence of genome-wide investigations for complex traits. *Ann. Hum. Genet.*, **73** (Pt 5), 514–519.

Lange,C. *et al*. (2003) A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*, **4**, 195–206.

Levy,D. *et al*. (2009) Genome-wide association study of blood pressure and hypertension. *Nat. Genet.*, **41**, 677–687.

Liu,J. *et al*. (2009) Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet. Epidemiol.*, **33**, 217–227.

Lumley,T. and Rice,K. (2010) Potential for revealing individual-level information in genome-wide association studies. *JAMA*, **303**, 659–660.

Mailman,M.D. *et al*. (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.

Pe'er,I. *et al*. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.*, **32**, 381–385.

Psaty,B.M. *et al*. (2009) Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ. Cardiovasc. Genet.*, **2**, 73–80.

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.

Pyeritz,R.E. (1989) Pleiotropy revisited: molecular explanations of a classic concept. *Am. J. Med. Genet.*, **34**, 124–134.

Sirota,M. *et al*. (2009) Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet.*, **5**, e1000792.

Soranzo,N. *et al*. (2009) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.*, **41**, 1182–1190.

Stearns,F.W. (2010) One hundred years of pleiotropy: a retrospective. *Genetics*, **186**, 767–773.

Teslovich,T.M. *et al*. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

Thorisson,G.A. *et al*. (2005) The International HapMap Project Web site. *Genome Res.*, **15**, 1592–1593.

Tyler,A.L. *et al*. (2009) Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*, **31**, 220–227.

WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Yang,Q. *et al*. (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet Epidemiol.*, **34**, 444–454.

Zou,L. *et al*. (2008) Systematic analysis of pleiotropy in C. elegans early embryogenesis. *PLoS Comput. Biol.*, **4**, e1000003.