European Journal of Human Genetics (2011) 19, 893–900
© 2011 Macmillan Publishers Limited All rights reserved 1018-4813/11
www.nature.com/ejhg

ARTICLE

Variable set enrichment analysis in genome-wide association studies

Wei Yang[1], Lisa de las Fuentes[2], Victor G Dávila-Román[2] and C Charles Gu*,[1,3]

Complex diseases such as hypertension are inherently multifactorial and involve many factors of mild-to-minute effect sizes. A genome-wide association study (GWAS) typically tests hundreds of thousands of single-nucleotide polymorphisms (SNPs), and offers opportunity to evaluate aggregated effects of many genetic variants with effects that are too small to detect individually. The gene-set-enrichment analysis (GSEA) is a pathway-based approach that tests for such aggregated effects of genes that are linked by biological functions. A key step in GSEA is the summary statistic (gene score) used to measure the overall relevance of a gene based on all SNPs tested in the gene. Existing GSEA methods use maximum statistics sensitive to gene size and linkage equilibrium. We propose the approach of variable set enrichment analysis (VSEA) and study new gene score methods that are less dependent on gene size. The new method treats groups of variables (SNPs or other variants) as base units for summarizing gene scores and relies less on gene definition itself. The power of VSEA is analyzed by simulation studies modeling various scenarios of complex multiloci interactions. Results show that the new gene scores generally performed better, some substantially so, than existing GSEA extension to GWAS. The new methods are implemented in an R package and when applied to a real GWAS data set demonstrated its practical utility in a GWAS setting.

European Journal of Human Genetics (2011) **19,** 893–900; doi:10.1038/ejhg.2011.46; published online 23 March 2011

Keywords: gene set enrichment; pathway-based analysis; SNP; genome-wide association

INTRODUCTION

Complex diseases often involve the interplay of many genes and environmental factors. The approach of genome-wide association study (GWAS) has shown great potential in dealing with such complexity, as demonstrated by the many exciting recent findings.[1] By scanning hundreds of thousands of single-nucleotide polymorphisms (SNPs), GWAS is less biased than candidate gene approach and more powerful than traditional linkage analysis in detecting genetic variants that confer modest disease risk. It provides new opportunities to evaluate a group of genes as a whole for potentially important synergetic effects. However, most published GWAS studies focus on marginal effects of individual SNPs. Seldom has the interplay of multiple genes been carefully evaluated, largely because of prohibitive computational burden and a lack of appropriate tools. In this study, we present a gene-set-based approach for evaluating the collective action of multiple SNPs in many genes with companion software that aims at reduced computational burden by taking advantage of known genetic pathways.

In gene expression analysis, the method of gene-set enrichment analysis (GSEA),[2] was proposed to overcome similar problems of prohibitive computation and lack of modeling for interactions. To measure the degree of 'enrichment' of association signals in a gene set, GSEA uses 'enrichment score' to summarize the association test scores of every gene in the gene set. If a gene set is over-represented by functional genes that are relevant to the disease of interest, GSEA will have enhanced power by combining strength of multiple genes. Indeed, the gene set identified by early GSEA application[3] led to subsequent functional validation of connections between impaired mitochondrial activity and insulin resistance.[4,5] The approach may be extended to GWAS studies. However, a key modification is needed for a scoring method that can combine association signals of multiple SNPs to obtain a single value representing the importance of that gene to the disease trait. Once such a 'gene score' is obtained, enrichment of association in gene sets can be tested as before.

In GWASs, the number of SNPs in a gene varies from a handful to hundreds, with only one or a few responsible for functional divergence, or in linkage disequilibrium (LD) with causal variants. Ideally, only such SNPs should be used to represent the gene. Therefore, a gene score may be defined by the maximum test statistics among all SNPs in a gene.[6] But genes with more genotyped SNPs or weak LD among them may score higher this way just by chance. A partial remedy for such a bias is to normalize gene-set enrichment scores. For example, the original GSEA method did so by scaling using the mean score estimated from permutation tests and by using the method described by Wang *et al*[6] by standardization.

An alternative approach is to perform normalization at the gene-score level, so that biases due to varying gene size and LD structure are minimized before deriving gene-set enrichment scores. We present several methods for normalizing gene scores and their utility in extending GSEA to GWAS. The new approach can evaluate enrichment of associated variables in predefined sets of genes, SNPs, and is termed 'variable-set enrichment analysis' (VSEA). An R software package (also named VSEA) is developed to implement the new methods, with a collection of gene sets constructed from known

[1]Division of Biostatistics, Washington University School of Medicine, St Louis, MO, USA; [2]Department of Medicine, Washington University School of Medicine, St Louis, MO, USA; [3]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA
*Correspondence: Dr C Charles Gu, Division of Biostatistics, Washington University School of Medicine, Campus Box 8067, 660 South Euclid Avenue, St Louis, MO 63110, USA. Tel: +314 362 3642; Fax: +314 362 2693; E-mail: gc@wubios.wustl.edu

Received 25 May 2010; revised 14 January 2011; accepted 9 February 2011; published online 23 March 2011

biological pathways to facilitate real GWAS analysis. We present evaluations of the new methods by simulation and analysis of real GWAS data, and discuss how VSEA is related to other existing gene-set -based methods and their suitable applications in different scenarios.

METHODS

The GSEA Approach

Subramanian's GSEA method. This method was proposed in gene expression analysis to detect significant 'enrichment' of disease association in a gene set. Let $L_0 = \{G_1, G_2, ..., G_N\}$ be the list of all genes measured on a gene expression microarray. Sorting their association test statistic values (for example, *t*-test score) from largest to smallest, we get $r_{(1)}, r_{(2)}, ..., r_{(N)}$ and a ranked gene list $L = \{G_{(1)}, G_{(2)}, ..., G_{(N)}\}$. Let $S \subset L_0$ be the gene-set of interest. The enrichment in *S* is evaluated in three steps:

Calculate the enrichment score (ES): A Kolmogorov–Smirnov statistic is used to measure over-representation of *S* at the top of the ranked list *L*. It is calculated by examining the ranked gene list *L*, increasing a running sum statistic when encountering a gene in *S* and by decreasing it when not. The enrichment score is defined as the maximum of the running sum statistics:

$$\mathrm{ES}(S, D) = \max_{1 \le j \le N} \left\{ \sum_{G_{(j*)} \in S, j* \le j} \frac{|r_{(j*)}|^P}{N_R} - \sum_{G_{(j*)} \notin S, j* \le j} \frac{1}{N - |S|} \right\} \quad (1)$$

where *D* denotes observed data, $|S|$ the number of genes in *S*, $N_R = \sum_{G_{(j*)} \in S} |r_{(j*)}|^P$, and *P* a tuning parameter giving higher weights to highly associated genes (P=1 was recommended in gene expression studies[2]).

Evaluate the significance of the enrichment score by permutation: ES is repeatedly calculated after shuffling disease status of samples, which generates an empirical null distribution of ES. An empirical *P*-value is estimated by the proportion of permutations that result in larger ES than originally observed.

Correct for multiple testing: When a large number of gene sets are tested simultaneously, false discovery rate (FDR) or family-wise error rate (FWER) may be used to correct for multiple testing. The enrichment scores are normalized first to minimize undesirable effects of varying gene-set size and within-gene-set correlations. Finally, the normalized ES (NES) is defined as:

$$\mathrm{NES}_{\mathrm{Subramania}}(S, D) = \mathrm{ES}(S, D)/\mathrm{mean}(\mathrm{ES}(S, \pi)) \quad (2)$$

where $\mathrm{ES}(S, \pi)$ is the enrichment score for permutation π.

Extension of GSEA to GWAS. A key step in extending GSEA to GWAS studies is to derive a summary score that combines signals from individual SNPs in each gene.[6] Denote the SNPs of gene G_k as $V_{k1}, V_{k2}, ..., V_{km}$, and their association test statistic as $t_{k1}, t_{k2}, ..., t_{km}$. In Wang's extension of GSEA to GWAS,[6] gene score for G_k is assigned the highest test statistic value among all the SNPs, $\max(t_{kj})$. Following this, the enrichment scores are calculated as in the original GSEA. To adjust for multiple testing, normalized ES is calculated as follows:

$$\mathrm{NES}_{\mathrm{Wang}}(S, D) = \frac{\mathrm{ES}(S, D) - \mathrm{mean}(\mathrm{ES}(S, \pi))}{\mathrm{SD}(\mathrm{ES}(S, \pi))} \quad (3)$$

We note that a summary 'gene score' simply defined by the maximum of per-SNP statistic favors larger genes (more genotyped SNPs with weak LD) because they may score higher by chance as a result of the greater number of independent tests. Therefore, new methods are needed to correct such biases for improved performance of gene-set enrichment analysis in GWAS.

New methods using normalized gene scores

We propose to perform normalization of gene scores before calculating gene-set enrichment scores, making the gene scores comparable for genes of different sizes. Several alternatives are introduced below in addition to the max statistic used by Wang *et al.*[6]

Gene scores based on maximum SNP statistics. The first class of gene scores aims to normalize the maximum SNP statistics by their empirical distribution derived from permutation analysis. For gene G_k, denote the observed maximum statistic as $m_k(D)$ and the permuted $m_k(\pi)$. We will consider new gene scores defined as follows based on maximum SNP statistics:

$$r_{WANG\_k} = m_k(D) \quad (4)$$

$$r_{CHI2MEAN\_k} = m_k(D)/\mathrm{mean}(m_k(\pi)) \quad (5)$$

$$r_{CHIMEAN\_k} = \sqrt{m_k(D)}/\mathrm{mean}(\sqrt{m_k(\pi)}) \quad (6)$$

$$r_{CHI2\_k} = [m_k(D) - \mathrm{mean}(m_k(\pi))]/\mathrm{SD}(m_k(\pi)) \quad (7)$$

$$r_{CHI\_k} = [\sqrt{m_k(D)} - \mathrm{mean}(\sqrt{m_k(\pi)})]/\mathrm{SD}(\sqrt{m_k(\pi)}) \quad (8)$$

$$r_{ABSZ\_k} = \Phi^{-1}(1 - P_k/2) \quad (9)$$

where $P_k = \hat{P}r(m_k(D)) = \Pr(m_k(\pi) > m_k(D))$ Φ is the standard normal cumulative distribution function. CHI2MEAN and CHIMEAN adjust for gene sizes by scaling and CHI2 and CHI by standardization, all based on the distribution of permuted per-SNP test scores. The square root function used in CHIMEAN and CHI was motivated by experiments using various transformations to obtain more comparable gene-score distributions.

Gene scores based on other multilocus test statistics. We consider two multilocus statistics. One is Hotelling's T^2-test, as implemented in the PLINK software.[7] The other is proposed by Zhou *et al*[8] to summarize SNP associations within a gene. We name the two gene scores as 'T^2' and 'LCMT'. For both methods, the *P*-value for a gene is calculated first and then mapped to a χ^2 distribution with one degree of freedom or the square root of the distribution. The mapped statistic is used as the gene score in the VSEA test.

Gene scores based on tagging SNPs. In this class of methods, we bypass the use of genes as analysis units and instead use a set of representative SNPs selected from these genes. Success of such tests depends on how to properly choose the SNPs. Wang's gene score is a special case that selects single SNPs with maximum statistics. At the other extreme, we may include all SNPs from all genes in the gene set (termed as 'ALL_SNP'). Between these extremes, we considered three other SNP-set-based methods.

TAG: Tag SNPs are selected as representatives using the method by Meng *et al*[9] and Lin and Altman.[10]

PCA1: For each gene, principal component analysis (PCA) is first performed on allele count correlation matrix. The eigenvectors for the first few components are used to obtain linear combinations of the SNP association test statistics and form the 'pseudo-SNP' statistic, which is further used in 'SNP-set' enrichment analysis:

$$t_{PCA1\_}k = |t^T \times e_k| / \sqrt{\lambda_k} \quad (10)$$

Where λ_k is the *k*th largest eigenvalue and e_k is the corresponding eigenvector.

PCA2: Similar to 'PCA1', this method uses another definition for 'pseudo-SNP' statistic:

$$t_{PCA2\_}k = t^T \times (e_k \circ e_k) \quad (11)$$

where operator o stands for the element-wise product of two vector/matrix of the same size. When there are SNPs in complete LD within the gene, both PCA methods will remove extra copies of repeated SNP statistics.

Implementation of the new algorithm

The gene scores presented above are implemented in the following procedure to perform VSEA in GWAS.

(1) Perform GWAS by calculating single SNP and/or multilocus association test scores. PLINK[7] is called to generate the test statistics for both original and permuted data sets.
(2) Calculate normalized gene scores; for SNP-set-based scoring algorithms, this step is replaced by selecting representative SNPs.

(3) Calculate enrichment score following the original GSEA algorithm for each gene set.
(4) Evaluate significance of the enrichment scores by contrasting with that in permuted data.
(5) Normalize enrichment scores by scaling (formula 2) or standardization (formula 3) when multiple gene sets are tested, correcting for multiple testing by FDR/FWER.

The procedure is implemented in an R package called VSEA with utility tools for data formatting and annotation of results. It uses PLINK for computational intensive calculations, in which more flexible modeling is available (for example, tests with environmental covariates included).

Simulation study

Simulation study is used to evaluate the performance of the proposed new gene scores and to compare two approaches with normalizing enrichment scores.

We simulated 10 disease models, all with the same prevalence of 25%, and six interacting disease loci. Disease SNPs contribute to the binary trait through interactions, which were specified through penetrances of joint genotypes at the six disease loci. As shown in Table 1, the minor allele frequencies (MAFs) of the six loci are 5%, 10%, 20%, 30%, 40% and 50%, respectively. The 10 scenarios are divided into two groups for lower (~12%) and higher (~20%) values of heritability. Each of the two groups is comprised of five scenarios: (S1) no marginal effect is observed for any of the six loci; (S2-1) two loci with low MAF have moderate marginal effects (OR=1.2 in recessive model) and the other four have no marginal effects. (S2-2) Similar to 2 but the two loci with moderate marginal effects have high MAFs; (S3-1) and (S3-2) are similar to (S2-1) and (S2-2), but two loci have strong marginal effects (OR=2 in recessive model). For each of the 10 scenarios, we considered 10 joint-penetrance tables randomly generated to satisfy the marginal constraints. Case–control data sets were then simulated using the SNAP software.[11]

Every simulated data set has 300 SNPs consisting of 30 blocks, each with 10 SNPs. SNPs are in high LD within blocks but LD-free across blocks. The six disease loci locate in the first six blocks, respectively. Their positions in corresponding blocks are totally random. We used 750 cases and 750 controls for all 10 scenarios. As we will see later, from the results using 750 cases and 750 controls, scenarios S2-1C1, S2-1C2, S2-2C2 and S2-2C2 tend to have lower power, thus additional sample sizes were used for these scenarios (1500:1500 and 3000:3000) to see how power would be improved; likewise, smaller samples were also used for scenarios that have higher power (250:250 for S3-1C1, S3-1C2, S3-2C2 and S3-2C2).

Table 1 Ratio of marginal genotypic effects at each individual locus in the 10 scenarios of simulation study

	L1	*L2*	*L3*	*L4*	*L5*	*L6*	
	Minor allele frequency						
Scenario	*0.05*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*Total heritability*
S1C1	1:1:1	1:1:1	1:1:1	1:1:1	1:1:1	1:1:1	
S2-1C1	**1:1:1.2**	**1:1:1.2**	1:1:1	1:1:1	1:1:1	1:1:1	
S2-2C1	1:1:1	1:1:1	1:1:1	1:1:1	**1:1:1.2**	**1:1:1.2**	~0.12
S3-1C1	**1:1:2**	**1:1:2**	1:1:1	1:1:1	1:1:1	1:1:1	
S3-2C1	1:1:1	1:1:1	1:1:1	1:1:1	**1:1:2**	**1:1:2**	
S1C2	1:1:1	1:1:1	1:1:1	1:1:1	1:1:1	1:1:1	
S2-1C2	**1:1:1.2**	**1:1:1.2**	1:1:1	1:1:1	1:1:1	1:1:1	
S2-2C2	1:1:1	1:1:1	1:1:1	1:1:1	**1:1:1.2**	**1:1:1.2**	~0.2
S3-1C2	**1:1:2**	**1:1:2**	1:1:1	1:1:1	1:1:1	1:1:1	
S3-2C2	1:1:1	1:1:1	1:1:1	1:1:1	**1:1:2**	**1:1:2**	

A population prevalence of K=25% is used in all scenarios.
All disease models in all scenarios are assumed to have six disease single-nucleotide polymorphisms with their minor allele frequencies listed under the corresponding marker names.
Entries in the table are relative genotypic risks of (minor homozygote) heterozygote (major homozygote).
Bold entries indicate loci with some marginal effects.

To define a gene in the simulated data, we selected a random streak of SNPs around each disease locus from the corresponding block. The six disease genes form a gene set in the VSEA test. Further, we defined a reference gene set consisting of irrelevant SNPs similarly, by randomly selecting six blocks that do not have disease loci, and took one locus from each of them to locate a 'gene' around it. Power and false-positive rates were calculated based on 100 replicates for each penetrance tables. Further, results were averaged over the 10-penetrance table configurations used in each scenario.

Before calculating gene scores, single SNP tests were performed using the 1 d.f. allelic test implemented by PLINK. In general, the power of VSEA is also dependent on the choice of appropriate single SNP test methods despite the gene score method that is used.

Real GWAS data analysis

To evaluate its practical use, we also applied VSEA to a pilot GWAS data set that characterizes the cardiovascular structural and functional manifestations of hypertension, collectively termed hypertensive heart disease (HHD).

Curation of pathway-based gene sets. To facilitate real data analysis, we constructed biologically interesting pathways from publicly available databases, including KEGG,[12] GO[13] and BioCarta. We retrieved 179 pathways from KEGG (release 44.0) and 313 pathways from BioCarta (November 2007). For pathways in GO, we constructed gene sets on the basis of GO level 4 annotations for biological process and molecular function. Some nodes in GO level 4 were excluded if they occured in levels 2 and 3 as well. For nodes in level 5 and onward, their genes were assigned to their ancestral GO annotations in level 4. Construction of these GO gene sets was carried out using the goTools package and the GO database from BioConductor 2.1, which used the 200708 release of GO. We obtained 2150 gene sets from GO database. The exact number of gene sets in analysis will differ depending on the genotyping platform for each GWAS data set. Detailed breakdown of gene sets constructed from KEGG,

Table 2 Numbers of gene sets constructed from KEGG, BioCarta and GO for various genotyping platforms

	Gene set	*Gene set size*				
Platform	*database*	*1–2*	*3–20*	*21–200*	*≥201*	*Total*
All curated	KEGG	7	56	112	4	179
	BioCarta	4	250	60	0	314
	GO	425	876	636	213	2150
	Total	436	1182	808	217	2643
Affymetrix 100k	KEGG	9	84	81	1	175
	BioCarta	23	275	13	0	311
	GO	526	611	212	31	1380
	Total	558	970	306	32	1866
Affymetrix 500k	KEGG	7	63	106	3	179
	BioCarta	7	264	43	0	314
	GO	563	663	263	48	1537
	Total	577	990	412	51	2030
Affymetrix 5.0	KEGG	7	64	105	3	179
	BioCarta	7	269	38	0	314
	GO	564	665	256	48	1533
	Total	578	998	399	51	2026
Affymetrix 6.0	KEGG	7	59	109	4	179
	BioCarta	6	261	47	0	314
	GO	570	659	273	49	1551
	Total	583	979	429	53	2044
Hypertensive heart disease pilot data	KEGG	7	64	105	3	179
	BioCarta	7	270	36	0	313
	GO	556	667	254	48	1525
	Total	570	1001	395	51	2017

GO and BioCarta for some Affymetrix SNP (Affymetrix Inc., Santa Clara, CA, USA) array platforms and our HHD pilot GWAS data are presented in Table 2. In the real data analysis, we only tested gene sets that contain at least three but no more than 200 genes represented by markers in a given GWA data set, to alleviate multiple testing and to avoid testing overly narrow or broad functional categories.

Example GWAS data set. To illustrate the utility of VSEA, we applied the method to a pilot data set of 74 cases and 70 controls from a GWAS study of HHD. Hypertension affects millions of people and HHD is associated with elevated cardiovascular morbidity and mortality.[14] Several ongoing GWAS studies attempt to characterize the genetic components of hypertension and related diseases.[15] In the pilot HHD sample, SNPs were genotyped using the Affymetrix Mapping 500K Array Set. The data of the SNPs underwent quality control (QC) using commonly accepted criteria on array quality (missing rate $\leq$0.05, mean heterozygosity between 0.25 and 0.3) and on marker quality (call rate $\geq$0.99 for SNPs with MAF $\leq$0.05, call rate $\geq$0.95 for all other SNPs and Hardy–Weinberg test *P*-value $>10^{-6}$). A total of 389 344 SNPs passed QC, which were mapped to genes based on annotation provided by Affymetrix. There were 15 863 genes associated with a total of 320 747 distinct SNPs available for VSEA analysis. We then applied VSEA to this data set using different gene score methods and 1000 permutations, and analyzed empirical distributions of statistics and compared results between gene score methods.

RESULTS

Simulation study

We first checked empirical false-positive rates of the proposed VSEA tests. As shown is Figure 1, the mean false-positive rate fluctuates around the nominal *P*-values for all scenarios considered by the simulation study. This indicates that the algorithms proposed for calculating a summary gene score, all performed reasonably well and their corresponding VSEA tests are comparable with each other in terms of type-I error rates.

However, the power of the VSEA test can differ substantially depending on the gene score method used. Among the algorithms based on maximum SNP statistics, most of our proposed methods (except for 'ABSZ' and 'ABSZ2'), including 'CHI', 'CHI2', 'CHIMEAN' and 'CHI2MEAN', performed better or similar to 'WANG' under all 10 scenarios considered with various sample sizes. The all-time-best performer is 'CHI2', which was consistently more powerful than others. As shown in Figure 2a, at a significance level of α=0.05 (nominal *P*-value), average improvement in power ranged between 14% to as much as 40% depending on the underlying model and the sample size.

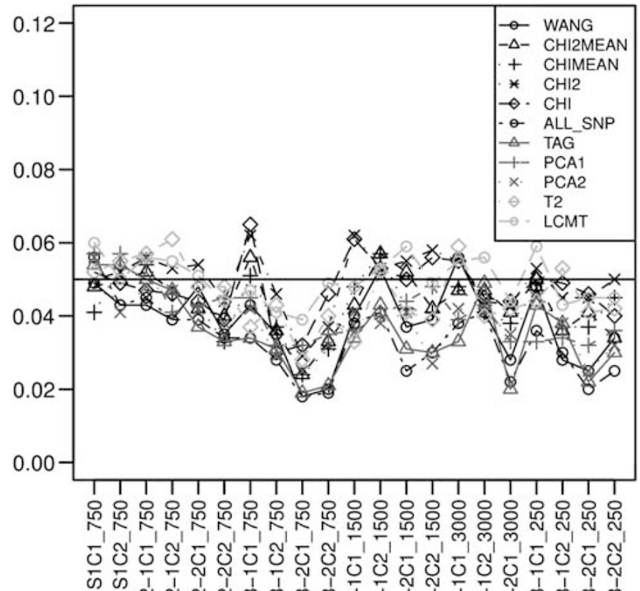

Figure 1 Mean false-positive rate of VSEA, estimated by averaging over 100 replicates of simulated data sets, under each of the scenarios considered by the simulation study. Sample sizes are shown at the end of each scenario name.

Using multilocus gene scores does improve VSEA test power in some scenarios (Figure 2b). LCMT usually has comparable or greater power compared with Wang's maximum statistics, except in S3-2C1 and S3-2C2 using 750:750 samples, in which there are extremely large single SNP effects. The other multilocus method, Hotelling's T^2, has slightly better power than Wang's maximum statistics in most cases, except when some risk SNPs have relatively large marginal effects (S3-1, S3-2).

VSEA tests using gene scores based on most of the tagging SNP-set-based methods performed at least as good as Wang's GSEA (see Figure 2c). The 'ALL_SNP' method seemed to be one of the best performers. However, even the best of these methods was still not as powerful as the 'CHI2' method.

When there was no marginal effect in any of the six disease SNPs (S1C1 and S1C2), all VSEA methods seemed to have no power at all (power close to nominal *P*-value). When there were even weak marginal effects in two of the six disease SNPs (S2-1C1, S2-1C2, S2-2C1 and S2-2C2), VSEA methods started to gain power and became useful with studies of larger sample sizes. For example, at α=0.05, VSEA test using the 'WANG' method had average power from 0.128 to 0.154 for the four configurations with a sample size of 750:750; the power increased to 0.205–0.264 for a sample size of 1500:1500 and to 0.373–0.566 for a sample size of 3000:3000. VSEA test using the more powerful 'CHI2' method achieved a power of 0.172–0.286 with a sample size of 750:750, to 0.345–0.555 for a sample size of 1500:1500 and to 0.694–0.919 for a sample size of 3000:3000. When there were strong marginal effects in two of the six disease SNPs, VSEA tests using all the proposed gene-score-calculating methods yielded good results. 'CHI2', 'CHI', 'CHI2-MEAN' and sometimes 'WANG' had powers close to 1 with 750 cases and 750 controls. In these scenarios, power of VSEA remained high even when we drop the sample sizes to 250 cases and 250 controls. In particular, 'CHI2' had an average power of ~0.99 for configurations S3-2C1 and S3-2C2, and around 0.68 for S3-1C1 and S3-1C2.

Real GWAS data

As the gene score method 'CHI2' outperformed others in the simulation study, we only report below results using 'CHI2' and those using 'WANG' for comparison. In the real data analysis, the PLINK permutation test (step-1 of the VSEA algorithm) took ~37 min on a single thread on a 2.4 GHz Dual-Core AMD Opteron Processor 2216 (PerformanceWare 1475, Pogo Linux, Inc., Redmond, WA, USA). After that, steps 2–4 of the VSEA analysis using 'WANG' or 'CHI2' gene score method each took ~14 h.

Nominal *P*-values from calculation using 'CHI2' and 'WANG' gene scores have a rank correlation of 0.65. In Table 3, we list the top 10 gene sets with smallest *P*-values by VSEA analyses using the 'CHI2' gene score method. For comparison, we also included ranks of the gene sets by the 'WANG' method. A similar table based on WANG is in Supplementary Table S1. Although there was some overlap between the two methods, a majority of the top 10 gene sets selected by the two methods were different. This could be related to the small sample size of the example data set (no gene set reached a significance level of 0.05 after adjustment for multiple testing). Therefore, biological interpretation of these results may be limited by the small sample size of this example data set.

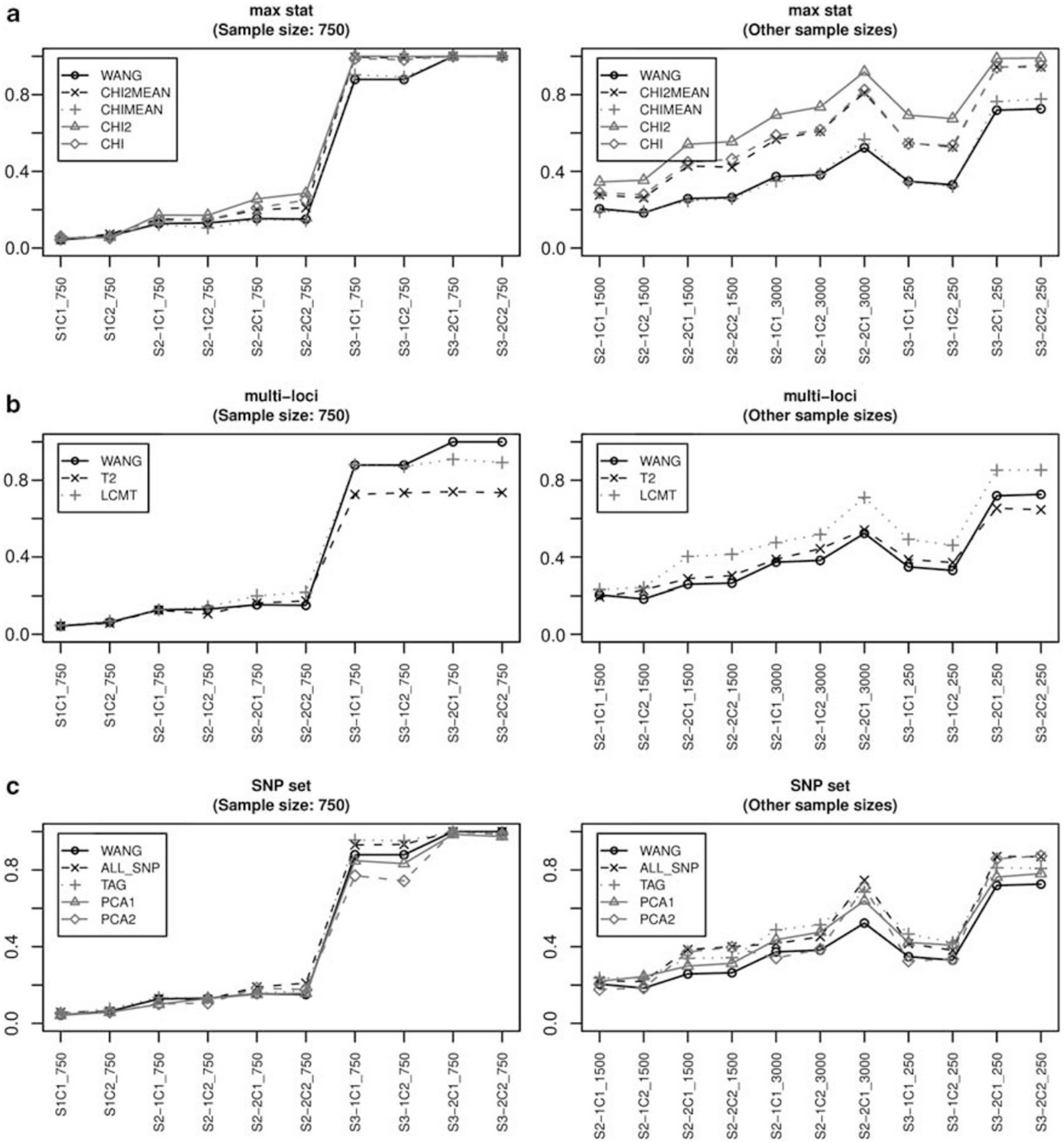

Figure 2 Estimated power of VSEA test by simulation study, using gene scores based on (**a**) maximum SNP statistics; (**b**) multilocus analysis; and (**c**) SNP-set enrichment. Sample sizes are shown at the end of each scenario name. The left panels use 750 cases and 750 controls, whereas the right panel uses other sample sizes (1500:1500, 3000:3000 and 250:250, identified by labels of the *x*-axis).

Table 3 Top gene sets identified by VSEA test (using CHI2 gene score) in the HHD pilot data

Gene set ID	*Pathway description*	*Gene set size*	*Rank by Wang et al*[6]	P*-value*	*FWER*_P
GO:0015101	Organic cation transporter activity	10	6	0.001	0.917
GO:0045785	Positive regulation of cell adhesion	8	50	0.004	0.996
h_IL12 Pathway	IL12- and Stat4-dependent signaling pathway in th1 development	20	316	0.006	1
GO:0046487	Glyoxylate metabolic process	3	372	0.006	1
h_RNApol3 Pathway	RNA polymerase III transcription	8	153	0.007	1
h_npp1 Pathway	Regulators of bone mineralization	5	54	0.007	1
h_pepiPathway	Proepithelin conversion to epithelin and wound repair control	4	4	0.008	1
GO:0006904	Vesicle docking during exocytosis	20	301	0.009	1
h_p53hypoxiaPathway	Hypoxia and p53 in the cardiovascular system	20	62	0.01	1
h_plcdPathway	Phospholipase c d1 in phospholipid associated-cell signaling	4	76	0.012	1

Abbreviations: FWER, family-wise error rate; HHD, hypertensive heart disease; VSEA, variable set enrichment analysis.

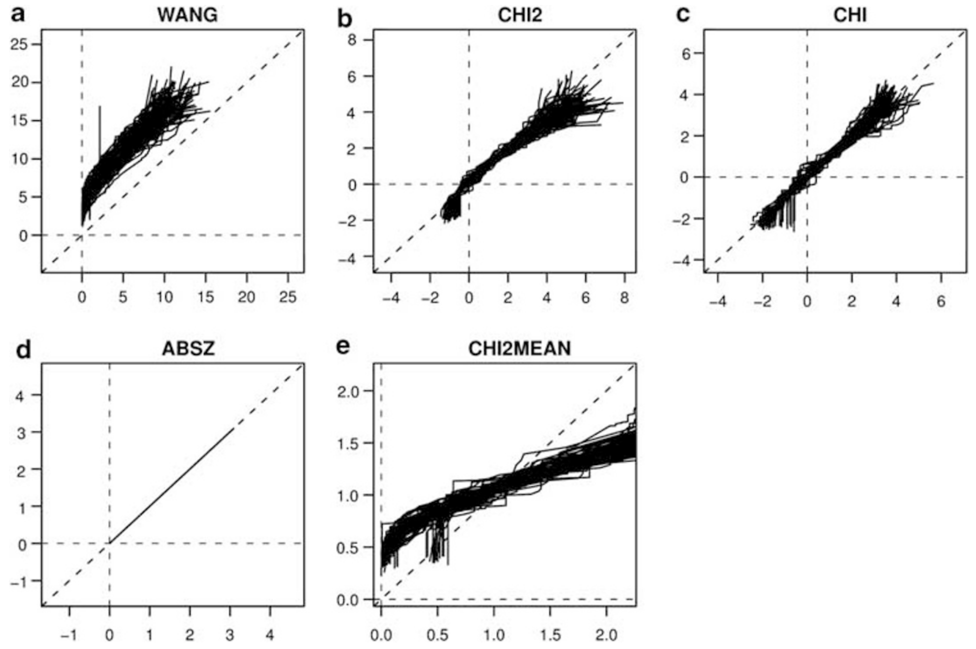

Figure 3 Comparison of gene-score distributions between large and small genes show effects of different gene-score normalization methods. Each panel presents overlaid QQ plots of gene-score distributions for 100 randomly selected pairs of genes in the example GWAS data set. Each pair comprises of a small gene with just a few (from 3 to 5) SNPs and a large one with at least 100 SNPs. The empirical distribution of gene scores was generated by calculating in the permuted data sets. The quantiles of the two distributions corresponding to each gene of the pair were then plotted against each other. In the plots, *x*-axis represents the smaller gene in the pair and *y*-axis represents the larger gene. Each panel is for a specific gene-score method: (**a**) WANG, (**b**) CHI2, (**c**) CHI, (**d**) ABSZ, (**e**) CHI2MEAN (see text for details of each method).

Nevertheless, the real data on genome-wide SNPs in a large number of genes and gene sets allowed us to examine the empirical null distributions of the proposed gene score and enrichment score statistics. This helped us to confirm that the objective of using the proposed statistics was accomplished; that is, normalization of gene scores had made the resulting statistics more comparable in spite of different gene sizes. Some results are shown in Figures 3, in which QQ plots of gene scores by various methods for 100 randomly selected pairs of genes are displayed. Each pair comprises of a gene with just a few (from 3 to 5) SNPs and another with many SNPs (≥ 100). In the QQ plots, gene scores were calculated in the permuted data sets based on the HHD pilot GWAS data set, and the quantiles of the gene scores in permutation tests for one gene were plotted against that for the other gene in the pair. They reflect the (dis)agreement between null distributions between large (*y*-axis) and small (*x*-axis) genes. Therefore, methods that produce the plots tightly centered around the diagonals are more superior in terms of generating comparable gene scores.

As seen in Figure 3, using the simple maximum statistic gene score without normalization, large genes tend to have much greater gene scores than smaller ones. On the other hand, using ‘CHI2’, ‘CHI’ or ‘ABSZ’ methods made distributions much more comparable between large and small genes, although ‘CHI2MEAN’ seemed to bias against large genes.

Similarly, comparisons were made between two normalized enrichment scores (formulae 2 and 3) to examine their effect in deriving comparable enrichment scores even when the sizes (number of genes) of gene sets may vary. In Figure 4, QQ plots are displayed to compare the empirical distributions of normalized enrichment scores for 100 randomly selected pairs of gene sets, using the ‘WANG’ (top row) and ‘CHI2’ gene scores (bottom row), respectively. For each gene-score method, three strategies of treating gene-set enrichment scores were compared: unnormalized (left), normalized by scaling (middle) and normalized by standardization (right). The plots reflect the (dis)-agreement between null distributions for different gene sets. Methods that generated the plots that were centered more tightly around the diagonals resulted in normalized scores that are less sensitive to gene-set sizes, and therefore are more suitable for multiple testing corrections. It is clear in Figure 4 that regardless of the gene score method used, normalization helped in deriving more comparable enrichment scores, and the standardization approach used by Wang is better than the simple scaling approach used in the original GSEA.

DISCUSSION

We presented new statistics for summarizing single SNP association test results in genes and gene sets for VSEA to evaluate enrichment of disease association in predefined gene sets. Similar to the original GSEA, VSEA tests whether the members of a gene set tend to co-occur near the top of the gene list ranked by single SNP analysis. We make several observations based on our evaluation of VSEA, using simulation and real GWAS data sets.

Gene sets as the basic unit for testing genotype–phenotype association

Traditional candidate gene studies consider single variants or haplotypes as basic units of analysis and generally had low replication of findings.[16–18] Because of differences in LD structure across populations, tagging SNPs or haplotypes of the risk variants could be quite different. This is particularly relevant under the common-disease–common-variant hypothesis, in which differential recombination histories across populations could have occurred during the long time before a mutation became prevalent. On the other hand, if there

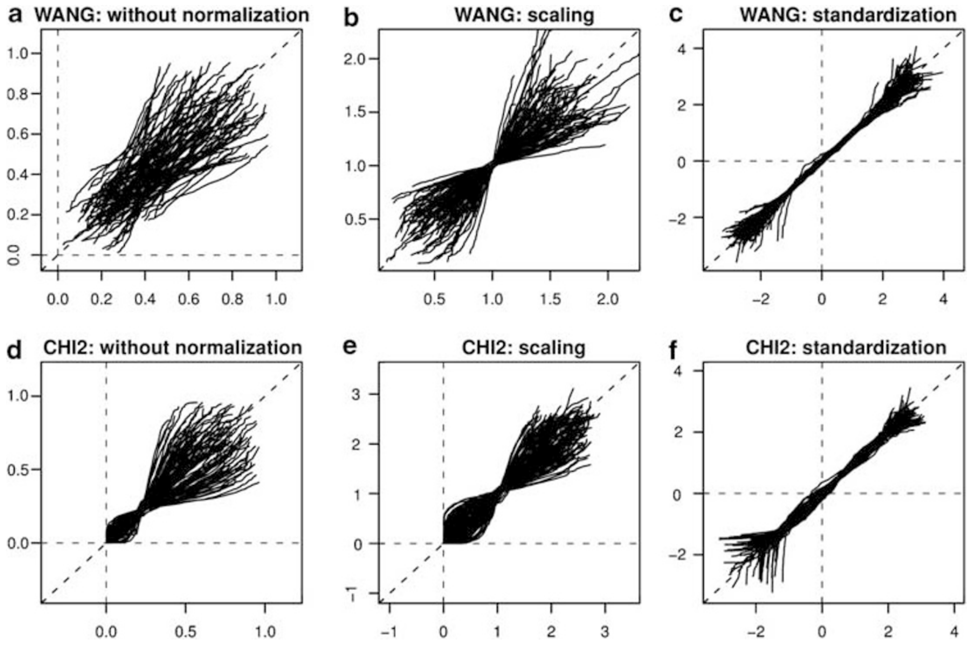

Figure 4 Comparison of enrichment scores when different enrichment score normalization methods were used. Each panel presents overlaid QQ plots of distributions of enrichment scores calculated in the example GWAS data, using 100 randomly selected pairs of gene sets. The empirical distribution of enrichment scores was estimated using 1000 permuted data sets. For each pair of gene sets, the quantiles of the two distributions corresponding to each set of the pair were plotted against each other. Columns of the figure correspond to three approaches to normalization and the rows to methods used for calculating gene scores: top row (**a**–**c**) using the 'WANG' method and bottom row (**d**–**f**) using the 'CHI2' method.

are multiple rare variants that are fairly recent in origin, allelic/locus heterogeneity is more likely across populations. In both cases, using single variants as basic units presents a challenge to successful replication studies.

In contrast, the positions, sequences and functionality of genes are highly consistent across diverse human populations. Furthermore, in complex diseases, groups of genes tend to work together and it seems reasonable to take gene sets as the unit for association analysis. This also has the added value of known genetic pathways and biological processes related to the disease of interest. This idea was successfully demonstrated in gene expression studies.[3–5] As gene expression (transcribed mRNA) levels connect genetic variations to clinically observed disease phenotypes, the value of GESA-type test in GWAS could be inferred. In fact, studies of so-called expression quantitative trait loci[19–21] convincingly demonstrated that genetic variations, together with environmental stimuli, may influence the location, timing and/or level of gene transcription. The findings support the abundance of *cis*-regulatory variations key to phenotypic variations in humans,[22] and motivate the continuing development of gene-set-based methods for GWAS studies.

Other gene-set-based methods available for GWAS

Several newly published methods also use the gene-set approach.[23–26] For example, Chasman[23] studied a GSEA-type method by pooling all SNPs in genes of a gene set to form a new set of SNPs with a hypergeometric test for enrichment. It is a special case of the 'tag-SNPs-set-based' VSEA, without accounting for the effect of varying gene sizes and gene-set sizes. They showed that the gene-set-based method was generally more efficient than conventional single-variant-based method when there are many variants with small effects. Peng *et al*[25] proposed a procedure similar to that of ours by testing for single SNP associations first, followed by gene-wise and then pathway-based analysis. However, the method was based solely on *P*-values with an underlying assumption about the independence of single SNP tests. In contrast, VSEA is based on a permutation procedure that properly accommodates the dependence of tests among single SNPs and those among genes. Therefore, Peng's method provides a quick way of reusing published *P*-values without genotype data, and VSEA is suitable for in-depth GWAS analysis when genotype data are available.

In general, enrichment test of gene-sets can be constructed using a broad range of statistics for gene scores, followed by restandardization of the gene scores, using permutation of phenotypes or genes or both.[27] The permutation procedure in GSEA-type method is known as 'phenotype shuffling'. A different kind of permutation is 'gene shuffling', in which gene scores are only calculated once, and then the enrichment score for a given gene set is calculated by comparison with randomly selected gene sets of the same size. Thus, gene shuffling can be much faster. However, the two permutation strategies have different underlying null hypotheses.[28,29] Moreover, whereas phenotype shuffling preserves important correlation structures between genes and among SNPs and is biologically more meaningful, gene shuffling risks destroying them. Therefore, the latter is not recommended unless the two strategies are combined to form a possibly more powerful test.[27] It remains to be seen whether this can materialize in GWAS because shuffling both phenotypes and genes simultaneously require significantly more computation.

The VSEA method is aimed at improving the 'accuracy' of gene scores so that they are less dependent on the size of genes and gene sets in consideration. Our study is not without its own limitations. For example, the simulation involved only 300 SNPs and a limited number of genes because of the lack of a realistic GWAS simulator that can handle complex interaction models. The method also did not explicitly model or directly test for interaction effects. Instead, it relies on predefined gene sets to capture a large collection of weak marginal effects, a common drawback of existing gene-set-based methods. Nevertheless, novel and improved GSEA methods continue being developed[30] in gene expression analysis. For GWAS studies, the persisting problem of 'missing heritability'[31] demands also for similar methods to detect collective actions of many risk factors. Therefore, further investigation is warranted for robust gene-set- and pathway-based methods that can more effectively incorporate biological information and be capable of providing functional understanding of new findings about the disease of interest.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research is supported in part by an NIH grants HL091028 and HL071782, and an AHA grant 0855626G.

LINKS

KEGG: Kyoto Encyclopedia of Genes and Genomes:
http://www.genome.jp/kegg/
BioCarta: http://www.biocarta.com
Gene Ontology: http://www.geneontology.org
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/

1 Manolio TA, Brooks LD, Collins FS: A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008; **118**: 1590–1605.
2 Subramanian A, Tamayo P, Mootha VK *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**: 15545–15550.
3 Mootha VK, Lindgren CM, Eriksson KF *et al*: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003; **34**: 267–273.
4 Patti ME, Butte AJ, Crunkhorn S *et al*: Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: potential role of PGC1 and NRF1. *Proc Natl Acad Sci USA* 2003; **100**: 8466–8471.
5 Petersen KF, Dufour S, Befroy D, Garcia R, Shulman GI: Impaired mitochondrial activity in the insulin-resistant offspring of patients with type 2 diabetes. *N Engl J Med* 2004; **350**: 664–671.
6 Wang K, Li M, Bucan M: Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007; **81**: 1278–1283.
7 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
8 Zhou H, Wei LJ, Xu X: Combining association tests across multiple genetic markers in case-control studies. *Hum Hered* 2008; **65**: 166–174.
9 Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG: Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* 2003; **73**: 115–130.
10 Lin Z, Altman RB: Finding haplotype tagging SNPs by use of principal components analysis. *Am J Hum Genet* 2004; **75**: 850–861.
11 Nothnagel M, Furst R, Rohde K: Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered* 2002; **54**: 186–198.
12 Kanehisa M, Goto S, Hattori M *et al*: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006; **34**: D354–D357.
13 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
14 Fields LE, Burt VL, Cutler JA, Hughes J, Roccella EJ, Sorlie P: The burden of adult hypertension in the United States 1999 to 2000: a rising tide. *Hypertension* 2004; **44**: 398–404.
15 Arnett DK, Baird AE, Barkley RA *et al*: Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: a scientific statement from the American Heart Association Council on Epidemiology and Prevention, the Stroke Council, and the Functional Genomics and Translational Biology Interdisciplinary Working Group. *Circulation* 2007; **115**: 2878–2901.
16 Ioannidis JP: Genetic associations: false or true? *Trends Mol Med* 2003; **9**: 135–138.
17 Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG: Genetic associations in large versus small studies: an empirical assessment. *Lancet* 2003; **361**: 567–571.
18 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; **33**: 177–182.
19 Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT: Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 2005; **437**: 1365–1369.
20 Zhang W, Duan S, Kistner EO *et al*: Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet* 2008; **82**: 631–640.
21 Li J, Burmeister M: Genetical genomics: combining genetics with gene expression analysis. *Hum Mol Genet* 2005; **14** (Spec no 2): R163–R169.
22 Stranger BE, Nica AC, Forrest MS *et al*: Population genomics of human gene expression. *Nat Genet* 2007; **39**: 1217–1224.
23 Chasman DI: On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet Epidemiol* 2008; **32**: 658–668.
24 Yu K, Li Q, Bergen AW *et al*: Pathway analysis by adaptive combination of P-values. *Genet Epidemiol* 2009; **33**: 700–709.
25 Peng G, Luo L, Siu H *et al*: Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet* 2010; **18**: 111–117.
26 Cantor RM, Lange K, Sinsheimer JS: Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet* 2010; **86**: 6–22.
27 Efron B, Tibshirani R: On testing the significance of sets of genes. *Ann Appl Stat* 2007; **1**: 107–129.
28 Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci USA* 2005; **102**: 13544–13549.
29 Goeman JJ, Buhlmann P: Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007; **23**: 980–987.
30 Yan X, Sun F: Testing gene set enrichment for subset of genes: Sub-GSE. *BMC Bioinformatics* 2008; **9**: 362.
31 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)