PSEA: Phenotype Set Enrichment Analysis—A New Method for Analysis of Multiple Phenotypes

Janina S. Ried,[1] Angela Döring,[2,3] Konrad Oexle,[4] Christa Meisinger,[3] Juliane Winkelmann,[4,5,6] Norman Klopp,[7,8] Thomas Meitinger,[4,6] Annette Peters,[3] Karsten Suhre,[9,10,11] H.-Erich Wichmann,[2,12,13] and Christian Gieger[1]*

[1] *Institute of Genetic Epidemiology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany*
[2] *Institute of Epidemiology I, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany*
[3] *Institute of Epidemiology II, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany*
[4] *Institute of Human Genetics, MRI, Technische Universität München, Munich, Germany*
[5] *Department of Neurology, MRI, Technische Universität München, Munich, Germany*
[6] *Institute of Human Genetics, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany*
[7] *Research Unit of Molecular Epidemiology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany*
[8] *Hannover Unified Biobank, Hannover Medical School, Hannover, Germany*
[9] *Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany*
[10] *Faculty of Biology, Ludwig-Maximilians-Universität, Munich, Germany*
[11] *Department of Physiology and Biophysics, Weill Cornell Medical College, Doha, Qatar*
[12] *Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany*
[13] *Klinikum Grosshadern, Munich, Germany*

Most genome-wide association studies (GWAS) are restricted to one phenotype, even if multiple related or unrelated phenotypes are available. However, an integrated analysis of multiple phenotypes can provide insight into their shared genetic basis and may improve the power of association studies. We present a new method, called "phenotype set enrichment analysis" (PSEA), which uses ideas of gene set enrichment analysis for the investigation of phenotype sets. PSEA combines statistics of univariate phenotype analyses and tests by permutation. It does not only allow analyzing predefined phenotype sets, but also to identify new phenotype sets. Apart from the application to situations where phenotypes and genotypes are available for each person, the method was adjusted to the analysis of GWAS summary statistics. PSEA was applied to data from the population-based cohort KORA F4 ($N = 1{,}814$) using iron-related and blood count traits. By confirming associations previously found in large meta-analyses on these traits, PSEA was shown to be a reliable tool. Many of these associations were not detectable by GWAS on single phenotypes in KORA F4. Therefore, the results suggest that PSEA can be more powerful than a single phenotype GWAS for the identification of association with multiple phenotypes. PSEA is a valuable method for analysis of multiple phenotypes, which can help to understand phenotype networks. Its flexible design enables both the use of prior knowledge and the generation of new knowledge on connection of multiple phenotypes. A software program for PSEA based on GWAS results is available upon request. *Genet. Epidemiol.* 36:244–252, 2012. © 2012 Wiley-Periodicals, Inc.

Key words: genome-wide association; pleiotropy; permutation test

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Christian Gieger, Institute of Genetic Epidemiology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. E-mail: christian.gieger@helmholtz-muenchen.de

Received 2 September 2011; Revised 8 November 2011; Accepted 2 January 2012
Published online 14 March 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21617

INTRODUCTION

Most published genome-wide association studies (GWAS) are focused on one phenotype. Some studies present the results for several related phenotypes side by side in a comparing manner [Soranzo et al., 2009]. Combined analysis of multiple phenotypes is not widely applied at the moment. However, there are good reasons to analyze two or more phenotypes together. Multiple phenotypes can represent different aspects of one superior phenotype or various endpoints of biological mechanisms. The combined analysis of multiple phenotypes can help to understand their underlying genetic basis. For example, the iron metabolism and the hematopoiesis are known to be connected [Orkin and Zon, 2008]. A joint analysis of both iron and blood related traits can improve the understanding of shared pathways. Moreover, the analysis of multiple phenotypes can lead to improved power compared to association analysis of single phenotypes [Klei et al., 2008]. In many situations, several related or

© 2012 Wiley-Periodicals, Inc.

unrelated phenotypes are available for analysis. Then, the combined analysis of multiple phenotypes can improve the results of a study by using a larger fraction of the available information.

The methods for multiple phenotypes in GWAS are under development at the moment, proposing various approaches. Some authors developed multivariate algorithms [Ferreira and Purcell, 2009; Liu et al., 2009] that are of limited use if many phenotypes are under consideration. Moreover, they cannot be easily applied to situations where phenotype and genotype measurements for each person are not available. Other approaches use the results of GWAS on single phenotypes [Gupta et al., 2011; Huang et al., 2011; Yang et al., 2010]. The focus of these algorithms is hypothesis generation, mining the data for sets of phenotypes associated with genetic loci. In many cases both testing predefined phenotype combinations and generating new sets by screening for new phenotype combinations are of interest. Therefore, we developed a method, which may be used for testing and for generating hypotheses. Moreover, we made it applicable to measurements of genotypes and phenotypes per person as well as to aggregated single phenotype GWAS results. The current methods for multiple phenotypes address either single nucleotide polymorphism (SNP) level [Yang et al., 2010] or gene level [Huang et al., 2011] effects on multiple phenotypes. For genome-wide scans calculations based on SNP level imply a large multiple-testing burden. Therefore, we focused on the gene-based approach. We developed a set enrichment analysis for sets of phenotypes, borrowing ideas from gene set enrichment analysis on GWAS data [Ackermann and Strimmer, 2009; Wang et al., 2007, 2010]. Known methodological elements of different gene set enrichment analysis approaches were combined and adapted to the specific situation of investigation of phenotype sets. The basic principle is that for each phenotype set the statistics of the corresponding univariate phenotype analyses are combined to a score and this score is tested for enrichment by permutation over phenotype values or SNPs. To our knowledge, this is the first set enrichment approach for the analysis of multiple phenotypes. The method was applied to a panel of iron-related and blood count traits. Several of these phenotypes are known to be connected via the genesis of red blood cells. That made it a useful example for testing our method. Moreover, we exploited the fact that many genes were previously published to be associated with at least one of these phenotypes (Fig. 1) as we used published associations for the evaluation of the results of our method.

This paper demonstrates that the proposed method, called "phenotype set enrichment analysis" (PSEA), is a valid method for the analysis of gene effects on multiple phenotypes. The method enables testing hypotheses of shared genetic basis due to analysis of predefined phenotype sets. At the same time, PSEA may be used to derive new hypotheses in terms of newly defined phenotype sets. Moreover, we show that the analysis of genetically connected phenotypes can lead to an improved power, as it benefits from all available phenotypes in contrast to the common single phenotype GWAS approach. The application either to genotype and phenotype measurements per person or to GWAS results makes it widely applicable. A software implementation of PSEA for application on GWAS results can be obtained from the authors upon request.

In the following, a phenotype set is regarded as a group of two or more quantitative phenotypes, which are approximately normally distributed. The phenotypes may or may not be selected using some criteria such as correlation.

MATERIAL AND METHODS

The aim of PSEA is to test if a predefined set of phenotypes is associated with a gene. To perform a phenotype set based enrichment approach, we modified methods already used for gene set enrichment analysis and its application to GWAS data [Efron and Tibshirani, 2007; Guo et al., 2009; Segre et al., 2010; Subramanian et al., 2005; Wang et al., 2007]. Parallels of gene set enrichment analysis and phenotype set enrichment analysis facilitated usage of the same methodological algorithms. Nevertheless, specific characteristics of the analysis of phenotype sets had to be regarded in the methods development. In the following description of the algorithm it is stated for each step which elements were borrowed from gene set analysis approaches and which methods were changed. First, the general algorithm for testing the enrichment of a phenotype set for a gene using phenotype and genotype measurements per person is described. Afterwards, two extensions are presented, namely the identification of new phenotype sets that are likely to be enriched for a gene and the use of GWAS test statistics instead of phenotype and genotype measurements per person. The algorithm for PSEA and the extensions were implemented in the programming language C. A schematic overview of the PSEA algorithm is given in the Figure S1.

In the following, $p_1,p_2,\ldots,p_{\mathrm{Npheno}}$ are $\mathrm{N}_{\mathrm{pheno}}$ different continuous phenotypes that are approximately normally distributed and PS is a phenotype set that is a selection of N_{PS} phenotypes from $p_1,p_2,\ldots,p_{\mathrm{Npheno}}$.

STEP 1: GENE-BASED STATISTICS

To test genes for enrichment of a phenotype set, a gene-based statistic is required for each phenotype. We calculated positive test statistics of an association test for all SNPs mapped to a gene (e.g., χ^2-test statistic for testing the effect estimate in a linear regression). These statistics were combined to a gene-wise test statistic per phenotype. Analogous to gene set enrichment methods for GWAS data [Guo et al., 2009; Wang et al., 2007], we selected the maximal test statistic of all SNPs mapped to a gene as the gene-based statistic $t(gene,p_m)$. All SNPs in the transcribed region and a surrounding area of a 110 kb upstream and 40 kb downstream were mapped to a gene. This definition for assigning SNPs to genes was chosen as 99% of the expected cis-eQTL are located within this interval [Segre et al., 2010; Veyrieras et al., 2008].

STEP 2: ENRICHMENT SCORE

In the next step we determined an enrichment score (ES) for each combination of phenotype set and gene. The ES was calculated as sum of the gene-based statistics of all phenotypes in the phenotype set:

$$ES(PS, gene) = \sum_{p_m \in PS} t(gene,\, p_m). \tag{1}$$

This is a modification for positive test statistics of the widely used maxmean statistic [Efron and Tibshirani, 2007], which was presented as sumstat score for gene set enrichment [Tintle et al., 2009].

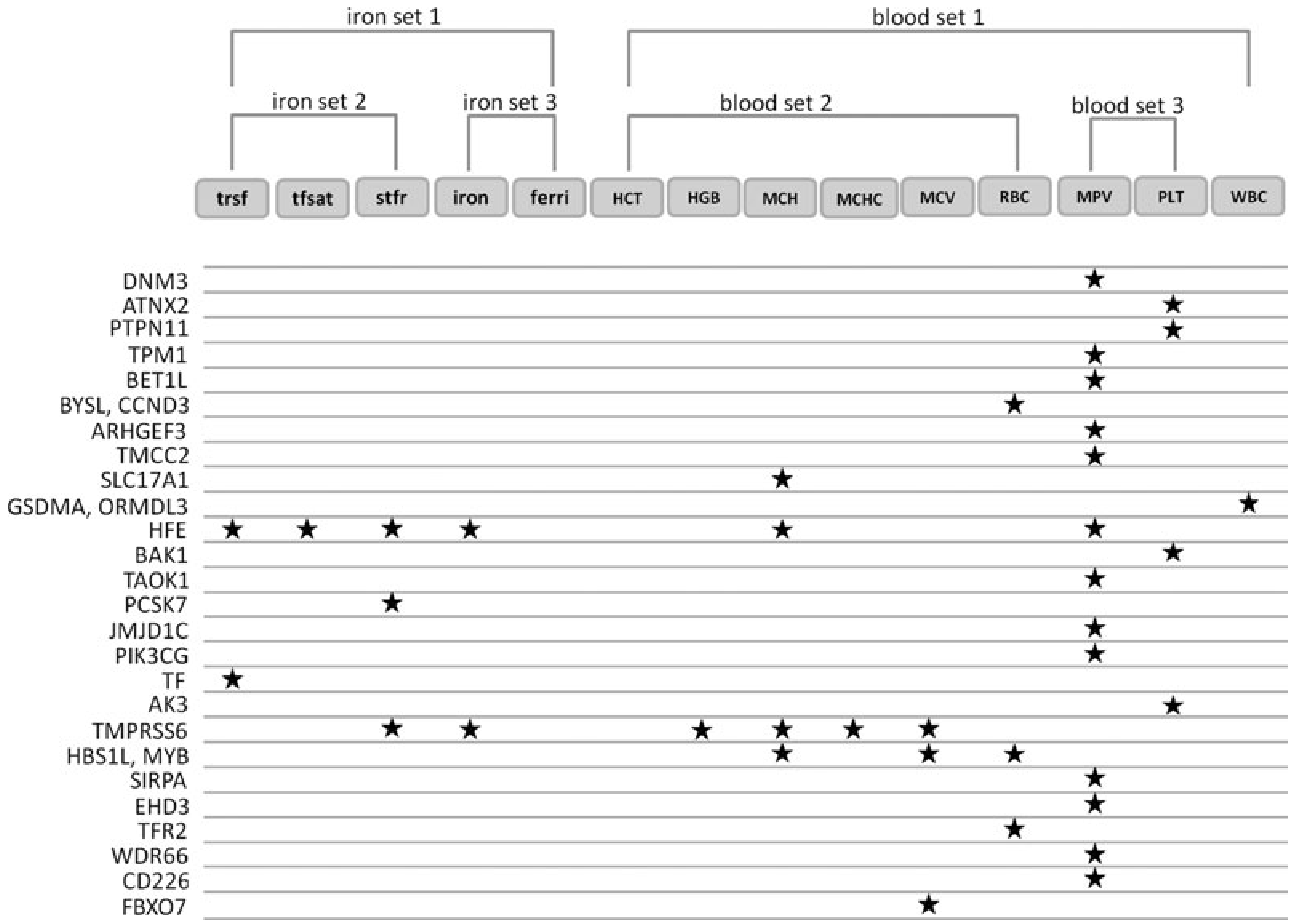

Fig. 1. Phenotype sets used for PSEA and overview of known genes. This figure summarizes all 14 phenotypes and the predefined phenotype sets used in PSEA. All genes that have been reported to be significantly associated with at least one phenotype in previous GWAS or meta-analysis including multiple cohorts are indicated. For each gene, the phenotypes with published associations are marked with an asterisk. (trsf, transferrin; tfsat, transferrin saturation; stfr, soluble transferrin receptor; ferri, ferritin; other abbreviations as given in the text).

STEP 3: PERMUTATION

We decided to apply the so-called "self-contained" test for PSEA. The "self-contained" approach tests whether a phenotype set is enriched compared to sets of random phenotypes [Ackermann and Strimmer, 2009; Wang et al., 2010]. It ensures that the result of PSEA for one phenotype set is independent of other phenotype sets under consideration.

If individual level phenotypes and genotypes are available, random phenotypes can be generated by permutation of the values over all individuals. This was also proposed for different gene set enrichment approaches [Wang et al., 2007]. A special requirement for PSEA is that one has to preserve the correlation structure of the phenotypes. Otherwise, sets of phenotypes and sets of permutated phenotypes would not be comparable (see the discussion section). Therefore, all phenotypes were permuted in the same way. Based on each permutation of the phenotypes the SNP-wise test statistics could be calculated. Analogous to the calculations for the original phenotype values, gene-based statistics per gene (compare step 1) and the permutation ES (compare step 2) were derived,

$$ES^{(j)}(PS, gene) = \sum_{p_m \in PS} t(gene, p_m^{(j)}). \qquad (2)$$

Thereby, the index (j) indicates a permutation with $1 \leq j \leq N_{\mathrm{perm}}$. To make the ES comparable between different phenotype sets and genes, we calculated the normalized ES (NES) by standardization of ES with the mean and standard deviation of the permutation ES. This standardization method was used in various gene set enrichment approaches [Guo et al., 2009; Wang et al., 2007].

STEP 4: STATISTICAL SIGNIFICANCE AND MULTIPLE TESTING CORRECTIONS

Statistical testing was performed using empirical distributions generated by permutation of phenotype values. The permutation P-value of the test if a phenotype set (PS*) is enriched for a gene (*gene**) is the fraction of permutations for which the permutation-based ES is larger than the original ES,

$$P(PS^*, gene^*) = \frac{\#_j\{ES(PS^*, gene^*) \leq ES^{(j)}(PS^*, gene^*)\}}{N_{perm}}. \quad (3)$$

Thereby, the index j indicates the permutation ($1 \leq j \leq N_{\text{perm}}$) and $\#_j$ means count over all permutations. The significance level must be corrected for the number of phenotype sets times the number of genes. As the number of permutations is limited, the P-value is not continuous but raises in steps of $1/N_{\text{perm}}$. Alternatively, the false discovery rate (FDR) and the family wise error rate (FWER) could be used for estimation (see section Methods in Supporting Information).

EXTENSION 1: IDENTIFICATION OF NEW PHENOTYPE SETS

If only predefined phenotype sets are used, one could miss an important phenotype set. This is especially true as different phenotype sets may be enriched for different genes. Testing all possible combinations of sets would raise the number of tests and is therefore often not feasible. With the identification of new phenotype sets parallel to the testing of predefined sets we enable to analyze promising associations.

For each gene we applied a threshold criterion based on single phenotype association results. All phenotypes that had gene-based univariate test statistics higher than a predefined threshold were regarded as newly identified phenotype set for this gene. We decided to apply a P-value threshold of $5{\times}10^{-4}$. This threshold is used only for identification of a phenotype set. Testing for enrichment of the newly identified set was done in the same way as for the predefined sets. Therefore, we could use a threshold that was less stringent than the univariate significance level. With this we aimed to balance between including phenotypes on which the gene has no effect and missing phenotypes on which the gene has an effect.

EXTENSION 2: PSEA BASED ON GWAS RESULTS

We developed a modification of PSEA that can use test statistics per SNP and phenotype from GWAS results if phenotype and genotype measurements for each person are not available. The permutation strategy was the major element that had to be adapted. Instead of permuting phenotype levels over all individuals, the test statistics were permuted over all SNPs (Guo et al. proposed a similar approach for gene set enrichment analysis [Guo et al., 2009]). To preserve the correlation structure of the phenotypes, the vector of the SNP-wise test statistics was permuted in the same way for all phenotypes. Gene-based test statistics and permutation ES were calculated using the permutation as described above. Testing was performed with P-value estimation (as presented in (3)). There are two impacts that have to be considered for this permutation scheme: (A) the distribution of test statistics might be influenced by associations of the phenotypes with other genes and (B) the permutation of SNP-wise test statistics destroys the LD structure. In other words, (A) means that the vector of SNP-wise test statistics might include more high test statistics than expected under the distribution of the null hypotheses of no SNP phenotype association. In permutations, these high test statistics could lead to higher permutation ES and therefore reduce the number of identified enrichments. Of course, this is highly dependent on the phenotypes in the considered phenotype set, especially on their strength of genetic associations and number of associated SNPs. The effect of the destroyed LD structure in the permutations could also lead to higher permutation ES. That is because of the gene test statistics, which were calculated as maximum of all SNP test statistics. By chance the maximal test statistic of independent SNPs will be higher than the maximal test statistic of dependent SNPs.

REAL DATA APPLICATION

PSEA was applied to a random sample of 1,814 unrelated individuals of the population-based cohort KORA F4. We replicated our results in a random sample of 1,644 unrelated individuals of the independent cohort KORA F3 [Wichmann et al., 2005] (see study description in Supporting Information 1). The PSEA, using phenotype and genotype measurements per person, required four types of input data: phenotype sets, phenotype values for at least all elements of the phenotype sets, genotypes, and a SNP-gene mapping.

Phenotypes. The method was applied to a set of 14 phenotypes (Fig. 1): Five traits related to the iron metabolism (iron, ferritin, transferrin, transferrin saturation, soluble transferrin receptor) and nine traits related to blood cells including six red blood cell traits (hematocrit (HCT), hemoglobin (HGB), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), red blood cell count (RBC)), one white blood cell trait (white blood cell count (WBC)), and two platelet traits (mean platelet volume (MPV), platelet count (PLT)) (details of measurement methods and population statistics in Tables SI and SII). Outliers were excluded if they differed more than three standard deviations from the mean value. Residuals of a linear regression on each log-transformed phenotype concerning age and sex were calculated and taken as phenotypic input.

Phenotype sets. Six phenotype sets were tested for enrichment; three included various combinations of iron traits and three included combinations of blood traits (Fig. 1). Newly identified phenotype sets that show a significant enrichment are described in the result section.

Genotypes. The analysis was restricted to SNPs that passed quality control criteria: minor allele frequency higher than 5%, call rate higher than 95% for genotyped SNPs, and imputation quality higher than 0.4 for imputed SNPs (see details of genotyping in Supporting Information 1).

SNP-gene mapping. Gene information was downloaded from the UCSC genome browser (http://genome.ucsc.edu/). A total of 20,801 genes passed quality control and were mapped to SNPs by their position (see details in Supporting Information 2). In the following, the term gene is also used for the region in which SNPs were mapped to a gene, transcribed region of a gene with the flanking region (of 110 kb upstream and 40 kb downstream). Such gene regions may be overlapping and some SNPs were mapped to several genes. We called the group of genes that were overlapping in at least one SNP *gene group*. For our data, the 20,801 genes lead to 2,319

TABLE I. Number of genes that were identified by PSEA to be associated with at least one predefined phenotype set. The results of PSEA on individual genotype and phenotype measurements in KORA F4 and the replication in KORA F3 are presented, as well as the results for PSEA on GWAS results F4

		PSEA on genotypes/phenotypes				PSEA on GWAS results	
		Discovery KORA F4		Replication KORA F3		KORA F4	
		Identified genes	Prev. assoc. genes	Identified genes	Prev. assoc. genes	Identified genes	Prev. assoc. genes
P-value	Gene	272	43	52	35	58	22
<0.001	Gene group	70	15	6	6	17	7

gene groups. For the analysis of the LD effect, we generated a SNP-gene mapping that included only independent SNPs that are approximately in linkage equilibrium with each other (LD pruning with PLINK [Purcell et al., 2007]).

GWAS results. GWAS were calculated based on the described phenotype input and genotypes. SNPTEST v2.2.0 [Marchini et al., 2007] was used for the calculation of the GWAS on the residuals for single phenotypes (QQ plots for all GWAS results in Fig. S2).

RESULTS

The selection of iron-related phenotypes and blood traits enabled the usage of published results from large GWAS and meta-analyses for the evaluation of PSEA results. We compared the results of enriched phenotype sets per gene with published results of single phenotype GWAS. A gene is named *previously associated gene* for a trait if a SNP located in the gene (110 kb upstream to 40 kb downstream of the transcript) was published in a meta-analysis with a genome-wide significant *P*-value [Benyamin et al., 2009; Chambers et al., 2009; Kullo et al., 2010; Oexle et al., 2011; Soranzo et al., 2009; Tanaka et al., 2010]. The previously associated genes named in the mentioned literature were presented in Figure 1 (detailed information in Table SIII). Enrichment of phenotype sets is neither limited to previously associated genes nor do all previously associated genes show enrichment of a phenotype set. However, if the enrichment of a phenotype set is significant for a gene that was previously published to be associated with elements of this phenotype set, it would be more likely a correct finding than the enrichment for a random selected gene. For the results presented below, 1,000 permutations were performed. As the step size of the *P*-values was determined by the number of permutations, we took the lowest *P*-value unequal to zero (0.001 for 1,000 permutations) as significance level and called a phenotype set enriched for a gene if the *P*-value was lower than 0.001. As more than 50 tests were taken forward for replication, we set the replication *P*-value to 0.001.

RESULTS OF PSEA ON GENOTYPES AND PHENOTYPES PER PERSON

To evaluate the PSEA results, we considered the absolute count of genes that were identified by PSEA to be associated with at least one phenotype set (shortly named: "identified genes"). The number of previously associated genes among these genes was used to assess the number of presumably true positive findings (Table I). Furthermore, we regarded the number of corresponding gene groups, as this number accounted for the overlapping gene definition. Sixteen percent (43/272) of the identified genes were previously associated genes and therefore presumably true positive findings. Similar results were observed for gene groups (15/70≈21% presumably true positive findings). We must be aware that the testing correction for the *P*-value criterion was possibly not strict enough. That fact could lead to a higher false positive rate. The replication step increased the percentage of presumably true positive findings to 67%. In terms of gene groups it was even 100%. But the increase in true positive rate was coincident with a decrease in the absolute number of identified previously associated genes. The detailed results of significantly enriched and replicated phenotype sets identified by PSEA are presented in Table II.

A considerably high amount of identified genes also include SNPs that were previously published to be associated with a single phenotype. In other words, these findings were likely to be true positive.

IDENTIFICATION OF NEW PHENOTYPE SETS

The significant replicated results included one new phenotype set. The set "new set" including iron, soluble transferrin receptor, transferrin saturation, MCH, and MCV was enriched for *TMPRSS6*. Altogether, 296 new phenotype sets were identified with PSEA. A total of 162 of which were significantly enriched. Apart from the results presented in Table I, many genes were identified to be associated only with newly defined phenotype sets. The percentage of gene groups that include a previously associated gene among these identified gene groups is quite low with 1.9% (data not shown). Nevertheless, the replicated result showed that the identification of new phenotype sets can give valuable insight into phenotypic networks.

COMPARISON OF PSEA WITH SINGLE PHENOTYPE GWAS

We compared the performance of PSEA with results of GWAS in KORA F4 on single phenotypes (Table II, more detailed information in Table SIV). For genes of two gene groups, the lowest *P*-value in KORA F4 GWAS for all traits was not genome-wide significant (for all SNPs mapped to the gene, genome-wide significance level of 5×10^{-8}). These genes would not have been found in GWAS in KORA

TABLE II. Results of PSEA: significantly enriched and replicated phenotype sets. Gene groups are separated by background color. If the enrichment of a phenotype set was significantly enriched and replicated for more than one gene of a gene group, the gene with the highest NES is presented. The results for previously associated genes that were part of a gene group for which a significantly enriched and replicated phenotype set was identified are also reported in the table. Genes are marked with an asterisk if no SNP in the gene region has been published with an association of blood or iron traits before. The significance level for discovery and replication was <0.001. For each gene we indicated the *P*-value and trait that showed the strongest association signal of all blood or iron traits and of all SNPs in the gene in KORA F4 single phenotype GWAS. Gene names were indicated with bold letters if the minimal *P*-value of the KORA F4 GWAS was not genome-wide significant (*P*-value $< 5 \times 10^{-8}$)

		PSEA		GWAS results			
		Discovery KORA F4	Replication KORA F3	KORA F4 ($N = 1814$)			
		($N = 1814$)	($N = 1644$)	Blood traits		Iron traits	
Gene	Pheno.-set	*P*-value	*P*-value	Phenotype	*P*-value	Phenotype	*P*-value
TMCC2	set_blood3	<0.001	<0.001	MPV	4.10×10^{-7}	Iron	0.0057
ARHGEF3	set_blood3	<0.001	<0.001	MPV	3.03×10^{-12}	Iron	0.0017
TF	set_iron1	<0.001	<0.001	PLT	0.0025	Transferrin	1.86×10^{-41}
TF	set_iron2	<0.001	<0.001	PLT	0.0025	Transferrin	1.86×10^{-41}
HFE	set_iron1	<0.001	<0.001	MCH	1.41×10^{-5}	trans. sat.	7.61×10^{-9}
HIST1H1C	set_iron1	<0.001	<0.001	MCH	1.41×10^{-5}	trans. sat.	7.61×10^{-9}
HFE	set_iron2	<0.001	<0.001	MCH	1.41×10^{-5}	trans. sat.	7.61×10^{-9}
HIST1H1A	set_iron2	<0.001	<0.001	MCH	1.41×10^{-5}	trans. sat.	7.61×10^{-9}
PCSK7	set_iron1	<0.001	<0.001	MCHC	0.0054	stfr	2.25×10^{-9}
APOC3*	set_iron1	<0.001	<0.001	HCT	0.0017	stfr	5.22×10^{-10}
PCSK7	set_iron2	<0.001	<0.001	MCHC	0.0054	stfr	2.25×10^{-9}
APOC3*	set_iron2	<0.001	<0.001	HCT	0.0017	stfr	5.22×10^{-10}
TMPRSS6	new set[a]	<0.001	<0.001	MCV	1.64×10^{-4}	stfr	1.04×10^{-6}
TMPRSS6	set_iron1	<0.001	<0.001	MCV	1.64×10^{-4}	stfr	1.04×10^{-6}
C22orf33	set_iron1	<0.001	<0.001	MCV	1.64×10^{-4}	stfr	1.04×10^{-6}
TMPRSS6	set_iron2	<0.001	<0.001	MCV	1.64×10^{-4}	stfr	1.04×10^{-6}
C22orf33	set_iron2	<0.001	<0.001	MCV	1.64×10^{-4}	stfr	1.04×10^{-6}

stfr, soluble transferrin receptor; trans. sat., transferrin saturation.

[a] The newly identified phenotype set "new set" consist of iron, soluble transferrin receptor, transferrin saturation, MCH, and MCV.

F4 on single phenotypes but were identified by PSEA with enrichment of a phenotype set.

COMPARISON WITH PUBLISHED RESULTS

PSEA revealed that the set of MPV and PLT is enriched for gene groups including *ARHGEF3* or *TMCC2*. These two genes were previously published in a large meta-analysis ($N > 13{,}500$) for an association with MPV. The alleles of the corresponding SNPs that increased the MPV were found to decrease the PLT. *ARHGEF3* were shown to be involved in the regulation of platelet counts and volume (intracellular signaling) [Soranzo et al., 2009]. The region including *TMCC2* was reported to be associated with MPV but no candidate gene was identified [Soranzo et al., 2009]. The results of PSEA showed that the set of all five iron traits (iron_set1) and the set of transferrin-related traits (iron_set3: soluble transferrin receptor, transferrin, and transferrin saturation) were significantly enriched and replicated for four gene groups including the genes *TF*, *HFE*, *TMPRSS6*, and *PCSK7*. These findings were comparable to associations that were published in a large meta-analysis on different iron traits. *HFE* and *TF* were reported to be associated with transferrin [Benyamin et al., 2009] and *HFE* and *TMPRSS6* with iron [Benyamin et al., 2009; Tanaka et al., 2010]. Moreover, *HFE* as well as *TMPRSS6* were shown to have an indirect effect on the soluble transferrin receptor via the transferrin saturation. In contrast to that, the effect of *PCSK7* on soluble transferrin receptor was presumed to be more direct [Oexle et al., 2011]. The gene product of *TMPRSS6* is known to be involved in the regulation of levels of the peptide hormone hepcidin, which is an important master regulator of iron homeostasis in humans [Soranzo et al., 2009]. Additionally, *TMPRSS6* was identified in a large meta-analysis on blood traits to be associated with MCH, MCV, and MCHC. The new identified phenotype set includes five of the six mentioned traits for which *TMPRSS6* was reported. The findings of PSEA correspond to the published results from large meta-analyses.

RESULT OF PSEA BASED ON GWAS RESULT

PSEA was applied to KORA F4 GWAS results with 1,000 permutations of SNP-based test statistics. In the analysis of PSEA on GWAS results, 58 genes (17 gene groups) were detected for which phenotype sets were significantly enriched (*P*-value < 0.001; Table I; Table SVI). The comparison with the results of PSEA on genotypes and phenotypes showed that all these genes, except two, were detected by PSEA on genotypes and phenotypes as well. Vice versa, the *P*-values of PSEA based on GWAS results for the replicated findings of PSEA on genotypes (Table II) were considerably low (Table SV). In spite of the fact that the absolute number of identified genes is lower in PSEA based on GWAS results, we observed that the percentage of possibly true positive findings (22/58≈38%) was higher than for PSEA on genotypes. Despite the possible impacts on the permutation scheme, PSEA on GWAS still finds a valuable amount of presumably true findings.

DISCUSSION

IDENTIFICATION OF LOCI ASSOCIATED WITH MULTIPLE PHENOTYPES AND NEW PHENOTYPE SETS

The comparison of the phenotype set enrichments at several genes identified by PSEA with the results of published large meta-analyses confirmed that a considerably high amount of the findings of PSEA were likely to be true positive findings. The application of PSEA to the given data identified several genes that were not genome-wide significant in KORA F4 GWAS for the single phenotypes GWAS. Therefore, we concluded that for this situation PSEA could identify more loci that have an effect on multiple phenotypes than single phenotype GWAS. From phenotype sets that were significantly enriched for one or several genes, we could gain information regarding the connection of phenotypes. In our application these connections were already known. However, in other situations the enrichment of phenotype sets may help to understand the connection of phenotypes. The results of PSEA on GWAS data revealed similar results, even though the estimated *P*-values were a bit higher. This was mainly caused by the needed modifications of the permutation scheme, which is discussed in a separate section in this discussion.

The application showed that the identification of new sets provides additional information. For *TMPRSS6*, we gained the information that this gene has an effect on MCH and MCV apart from the effect on iron parameters. The identification of new phenotype sets in PSEA uses a fixed level of test statistics for single phenotype association. A data driven estimation of this threshold may be a point for further development.

These findings demonstrated the usefulness of PSEA and its valuable results. To draw the whole picture, we discuss methodological characteristics and limitations of PSEA in the following.

EFFECT OF SNP-GENE MAPPING

The SNP-gene mapping defines which SNPs are analyzed as one locus in the PSEA. Due to the overlapping definition of genes, the loci are not independent. To regard this overlapping structure of genes in the evaluation of PSEA results, one can consider gene groups as done in the result section. The SNP-gene mapping was designed to cover not only the transcribed region but also most cis-eQTLs. We think that this mapping is reasonable in terms of creating a good representation of each gene. For nongenome-wide applications of PSEA, the SNP-gene mapping could be reduced to independent genes of interest. Moreover, PSEA can be applied to any other SNP-gene mapping if there are good reasons for a modification.

Apart from different SNP-gene mapping methods, one could also extend PSEA on SNP level. Technically, this approach would be possible analogous to the gene-based analysis by using single SNP statistics (methods step 1). There are two main situations where the SNP-based PSEA could gain more information than the gene-based PSEA: first, PSEA on SNP level enables to find multiple independent loci that have an effect on a phenotype set in one gene. Second, SNPs with an effect on a phenotype set could be identified, even if they are not mapped to a gene. The main drawback of the SNP level calculation is that it would strongly increase the number of tests. Moreover, many SNPs are in high LD and therefore dependent of each other. To get reliable results, it would be reasonable to reduce the analysis to independent SNPs. We think that the additional computational effort (more tests, LD analysis) is greater than the possible findings. As an extension, we could think of using a SNP level PSEA to analyze genes for which gene level PSEA identified significant enrichment.

EFFECT OF PERMUTATION

The design of the permutation strategy has an important effect on the results of PSEA. It is obvious that the number of permutations has a direct impact on the PSEA results. Especially, the significance level for the *P*-value criterion is dependent on the number of permutations as it determines the gradation and therefore the lowest possible *P*-value. With 1,000 permutations it is not possible to apply Bonferroni correction for 20,801 genes and six phenotype sets (correction for new identified phenotype sets not included) as the corrected significance level $0.05/(6 \times 20{,}801) = 4{\times}10^{-7}$ is below the lowest possible *P*-value of 0.001. This can lead to false positive findings. To reduce the false positive rate, one could increase the number of permutation, but that would lead to increased computation time and memory consumption. Alternatively, we showed that a replication step can be used to reduce the false positive rate.

One important aspect of the permutation strategy is that the correlation of phenotypes is conserved. For the application of PSEA to individual level genotypes and phenotypes, the phenotypes were permuted over all individuals but for all phenotypes in the same way. For application of PSEA to GWAS results, the test statistics of each phenotype were permuted over all SNPs but again for all phenotypes in the same way. The conservation of correlation of phenotypes in the permutation is important as the correlation of the phenotypes in the phenotype set has an influence on the ES and permutation ES. The permutation test compares ES of phenotype sets based on original and permuted phenotypes. Different correlation structures in the original data and its permutation would change the results. In other words, the permutation strategy ensures that the result for a phenotype set is not influenced by modified correlation structure of the phenotypes in the set.

PERMUTATION SCHEME OF PSEA ON GWAS RESULTS

There are two aspects that make the permutation scheme of PSEA on GWAS results less optimal than the permutation scheme of PSEA on genotypes and phenotypes: (A) possibly inflated test statistics distribution by association of phenotypes with other genes and (B) destroyed LD structure by SNP permutation. For aspect (B), we considered PSEA on GWAS results for a pruned list of SNPs. With pruned GWAS results more genes were found than with PSEA on not pruned GWAS results. Apart from the increased absolute number of genes, the number also rose of identified genes that were previously published (Table SVII). Anyway, the percentage of previously published genes in all identified genes was even higher in PSEA on not pruned GWAS. The destroyed LD structure reduced the absolute number of identified genes but PSEA on GWAS still led to interesting results. The aspect (A) is highly dependent on the phenotypes under consideration. We saw in our data that the percentage of genes that were identified in PSEA on genotypes and with PSEA on GWAS (pruned data) decreased with the number of associated genes and strength of association of the phenotypes in the phenotype set under consideration in the single phenotype GWAS (Table SVIII). But in many cases especially phenotypes with no strong effect on a single phenotype would be interesting to test for phenotype set enrichment. For those phenotypes the effect of (A) would be small. Anyway, the results also showed that even with not pruned data and inflated test statistic distributions PSEA on GWAS results can identify interesting gene phenotype set relations.

COMPUTER INTENSITY

In terms of computer intensity of the permutation strategies PSEA on GWA results is clearly less computer intensive than PSEA on individual genotypes and phenotypes. The algorithm for PSEA on measurements per individual was implemented in the programming language C with MPI parallelization. The program was executed on the Edinburgh Parallel Computing Centre (EPCC) supercomputing platform HECToR phase 2a (12,288-processor Cray XT4), within a project of DEISA (Distributed European Infrastructure for Supercomputing Applications). On 100 nodes, each with four processes, a genome-wide run with 2.8 million SNPs, 28k genes, 14 phenotypes for 1,814 individuals and six phenotype sets with 1,000 permutations took around 105 min per cohort. The algorithm for PSEA on GWA results was implemented in the programming language C. A genome-wide run for the mentioned phenotypes and phenotype sets on 2.18 millon SNPs (only SNPs with good quality) takes approximately 38 hr on one core of an Intel core i7-975 extreme 3.33 GHz, 24 GB RAM Linux computer. Additional reduction in computation time can be achieved by lowering the number of genes. In many situations one can focus on some genes. The results of PSEA on individual level data are independent of the number of genes under consideration. The same is true for PSEA on GWAS results as long as GWAS results or at least results on a large set of independent SNPs are used.

OTHER STATISTICS COMBINING APPROACHES

PSEA belongs to the methods that combine statistics of univariate analysis. In comparison with multivariate analysis such approaches require in general fewer assumptions about the phenotypes [Yang et al., 2010]. Therefore, they can be transferred more easily to different situations such as the use of either categorical or continuous phenotypes. PSEA could easily be modified for the analyses of binary traits or combinations of binary and quantitative traits. Besides PSEA, at least two other algorithms of that kind have been published recently [Huang et al., 2011; Yang et al., 2010]. Yang et al. [Yang et al., 2010] proposed a variation of O'Briens method [O'Brien, 1984] to combine univariate GWAS results, which were realized on SNP basis in contrast to our gene-based approach. The program PRIMe [Huang et al., 2011] identifies pleiotropic regions by scanning GWAS results of multiple phenotypes for low *P*-values, whereupon the LD structure is taken into account. PRIMe can identify different phenotype sets for different genes but it is not designed for testing a panel of predefined phenotype sets. Therefore, prior knowledge cannot easily be integrated in the analysis.

CONCLUSION

PSEA was demonstrated to be a valid method for the analysis of multiple phenotypes. Independence of number and correlation structure of phenotypes makes it applicable to various situations. The analysis of multiple phenotypes can lead to identification of new loci. Assumptions of shared genetic basis can be tested by predefined phenotype sets, whereas the internal identification of new phenotype sets can reveal unexpected connections.

In addition to genome-wide scanning for phenotype set enrichment, the method can be used for testing selected genes for enrichment with predefined phenotype sets. Criteria for the selection of such candidate genes could be small (not significant) *P*-values in a previous GWAS or prior knowledge from other external sources. We thus think that PSEA can be a tool for mining GWAS results for hidden associations and give insight into biological networks.

ACKNOWLEDGMENTS

We thank the DEISA Consortium (http://www.deisa.eu), co-funded through the EU FP6 project RI-031513 and the FP7 project RI-222919, for support within the DEISA Extreme Computing Initiative. Moreover, we thank the DEISA-group of LRZ for technical support, especially Dr. Kamen Beronov, Siew Hoon Leong, and Markus Michael Müller. The KORA Augsburg studies were financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany and supported by grants from the German Federal Ministry of Education and Research (BMBF). Part of this work was financed by the German National Genome Research Network (NGFN). Our research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ.

REFERENCES

Ackermann M, Strimmer K. 2009. A general modular framework for gene set enrichment analysis. BMC Bioinformatics 10:47.

Benyamin B, McRae AF, Zhu G, Gordon S, Henders AK, Palotie A, Peltonen L, Martin NG, Montgomery GW, Whitfield JB,Visscher PM. 2009. Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. Am J Hum Genet 84(1):60–65.

Chambers JC, Zhang W, Li Y, Sehmi J, Wass MN, Zabaneh D, Hoggart C, Bayele H, McCarthy MI, Peltonen L, Freimer NB, Srai SK, Maxwell PH, Sternberg MJ, Ruokonen A, Abecasis G, Jarvelin MR, Scott J, Elliott P, Kooner JS. 2009. Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nat Genet 41(11):1170–1172.

Efron B, Tibshirani R. 2007. On testing the significance of sets of genes. Ann Appl Stat 1(1):107–129.

Ferreira MA, Purcell SM. 2009. A multivariate test of association. Bioinformatics 25(1):132–133.

Guo YF, Li J, Chen Y, Zhang LS, Deng HW. 2009. A new permutation strategy of pathway-based approach for genome-wide association study. BMC Bioinformatics 10:429.

Gupta M, Cheung CL, Hsu YH, Demissie S, Cupples LA, Kiel DP, Karasik D. 2011. Identification of homogeneous genetic architecture of multiple genetically correlated traits by block clustering of genome-wide associations. J Bone Miner Res 26(6):1261–1271.

Huang J, Johnson AD, O'Donnell CJ. 2011. PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. Bioinformatics 27(9):1201–1206.

Klei L, Luca D, Devlin B, Roeder K. 2008. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol 32(1):9–19.

Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. 2010. A genome-wide association study of red blood cell traits using the electronic medical record. PLoS One 5(9):e13011.

Liu J, Pei Y, Papasian CJ, Deng HW. 2009. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet Epidemiol 33(3):217–227.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–913.

O'Brien PC. 1984. Procedures for comparing samples with multiple endpoints. Biometrics 40(4):1079–1087.

Oexle K, Ried JS, Hicks AA, Tanaka T, Hayward C, Bruegel M, Gogele M, Lichtner P, Muller-Myhsok B, Doring A, Illig T, Schwienbacher C, Minelli C, Pichler I, Fiedler GM, Thiery J, Rudan I, Wright AF, Campbell H, Ferrucci L, Bandinelli S, Pramstaller PP, Wichmann HE, Gieger C, Winkelmann J, Meitinger T. 2011. Novel association to the proprotein convertase PCSK7 gene locus revealed by analysing soluble transferrin receptor (sTfR) levels. Hum Mol Genet 20(5):1042–1047.

Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132(4):631–644.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575.

Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D. 2010. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 6(8): e1001058.

Soranzo N, Spector TD, Mangino M, Kuhnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, Li M, Salo P, Voight BF, Burns P, Laskowski RA, Xue Y, Menzel S, Altshuler D, Bradley JR, Bumpstead S, Burnett MS, Devaney J, Doring A, Elosua R, Epstein SE, Erber W, Falchi M, Garner SF, Ghori MJ, Goodall AH, Gwilliam R, Hakonarson HH, Hall AS, Hammond N, Hengstenberg C, Illig T, Konig IR, Knouff CW, McPherson R, Melander O, Mooser V, Nauck M, Nieminen MS, O'Donnell CJ, Peltonen L, Potter SC, Prokisch H, Rader DJ, Rice CM, Roberts R, Salomaa V, Sambrook J, Schreiber S, Schunkert H, Schwartz SM, Serbanovic-Canic J, Sinisalo J, Siscovick DS, Stark K, Surakka I, Stephens J, Thompson JR, Volker U, Volzke H, Watkins NA, Wells GA, Wichmann HE, Van Heel DA, Tyler-Smith C, Thein SL, Kathiresan S, Perola M, Reilly MP, Stewart AF, Erdmann J, Samani NJ, Meisinger C, Greinacher A, Deloukas P, Ouwehand WH, Gieger C. et al. 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet 41(11):1182–1190.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545–15550.

Tanaka T, Roy CN, Yao W, Matteini A, Semba RD, Arking D, Walston JD, Fried LP, Singleton A, Guralnik J, Abecasis GR, Bandinelli S, Longo DL, Ferrucci L. 2010. A genome-wide association analysis of serum iron concentrations. Blood 115(1):94–96.

Tintle NL, Borchers B, Brown M, Bekmetjev A. 2009. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. BMC Proc 3(Suppl 7): S96.

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4(10):e1000214.

Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81(6):1278–1283.

Wang K, Li M, Hakonarson H. 2010. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11(12):843–854.

Wichmann HE, Gieger C, Illig T. 2005. KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67(Suppl 1):S26–S30.

Yang Q, Wu H, Guo CY, Fox CS. 2010. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet Epidemiol 34(5):444–454.