PCA-based GRS analysis enhances the effectiveness for genetic correlation detection

Yan Zhao*, Yujie Ning*, Feng Zhang, Miao Ding, Yan Wen, Liang Shi, Kunpeng Wang, Mengnan Lu, Jingyan Sun, Menglu Wu, Bolun Cheng, Mei Ma, Lu Zhang, Shiqiang Cheng, Hui Shen, Qing Tian, Xiong Guo and Hong-Wen Deng

Corresponding authors: Feng Zhang, Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, P. R. China. Tel.: +86-2982655091; Fax: +86-2982655332; E-mail: fzhxjtu@mail.xjtu.edu.cn; Hong-Wen Deng, Center for Bioinformatics and Genomics, Department of Global Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University, New Orleans, Louisiana, USA. Tel.: 504-988-1310; Fax: 504-988-1310; E-mail: hdeng2@tulane.edu
*The two authors contributed equally to this work.

Abstract

Genetic risk score (GRS, also known as polygenic risk score) analysis is an increasingly popular method for exploring genetic architectures and relationships of complex diseases. However, complex diseases are usually measured by multiple correlated phenotypes. Analyzing each disease phenotype individually is likely to reduce statistical power due to multiple

Yan Zhao is currently working as a postgraduate student at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Yujie Ning is currently working as a lecturer at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Feng Zhang is currently working as a professor at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Miao Ding is currently working as a postgraduate student at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Yan Wen is currently working as a lecturer at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Liang Shi is currently working as a postgraduate student at Health Science Center, Xi'an Jiaotong University.

Kunpeng Wang is currently working as a postgraduate student at Health Science Center, Xi'an Jiaotong University.

Mengnan Lu is currently working as a postgraduate student at Health Science Center, Xi'an Jiaotong University.

Jingyan Sun is currently working as a postgraduate student at Health Science Center, Xi'an Jiaotong University.

Menglu Wu is currently working as a postgraduate student at Health Science Center, Xi'an Jiaotong University.

Bolun Cheng is currently working as a doctor student at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Mei Ma is currently working as a postgraduate student at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Lu Zhang is currently working as a postgraduate student at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Shiqiang Cheng is currently working as a postgraduate student at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Hui Shen is currently working as an associate professor at the Center for Bioinformatics and Genomics, Department of Global Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University.

Qing Tian is currently working as an associate professor at the Center for Bioinformatics and Genomics, Department of Global Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University.

Xiong Guo is currently working as a professor at the Key Laboratory of Trace Elements and Endemic Diseases of National Health Commission of the People's Republic of China, School of Public Health, Health Science Center, Xi'an Jiaotong University.

Hong-Wen Deng is currently working as a professor at the Center for Bioinformatics and Genomics, Department of Global Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University.

Submitted: 17 May 2018; Received (in revised form): 30 July 2018

testing correction. In order to conquer the disadvantage, we proposed a principal component analysis (PCA)-based GRS analysis approach. Extensive simulation studies were conducted to compare the performance of PCA-based GRS analysis and traditional GRS analysis approach. Simulation results observed significantly improved performance of PCA-based GRS analysis compared to traditional GRS analysis under various scenarios. For the sake of verification, we also applied both PCA-based GRS analysis and traditional GRS analysis to a real Caucasian genome-wide association study (GWAS) data of bone geometry. Real data analysis results further confirmed the improved performance of PCA-based GRS analysis. Given that GWAS have flourished in the past decades, our approach may help researchers to explore the genetic architectures and relationships of complex diseases or traits.

Key words: bioinformatics; principal component analysis; genetic risk score; correlation analysis; complex diseases

Introduction

Over the past decades, large amounts of genome-wide association studies (GWASs) have been conducted and a plethora of susceptibility loci have been identified for complex diseases or traits [1]. In the meantime, multiple novel approaches were proposed for genetic studies utilizing published GWAS data. Genetic risk score (GRS, also known as polygenic risk score) analysis is increasingly popular recently. GRS is a summation of genotypic scores of disease-associated variants weighted by their effect sizes, estimated by genetic association studies [2]. The power of GRS is generally higher than that of individual causal variant through combining the genetic effects of multiple causal variants into a single variable. It is especially useful for genetic studies of complex diseases determined by multiple variants with moderate or weak effects. With published disease susceptibility variants, GRS analysis is widely used for evaluating the effects of susceptible variants on disease risks and for exploring the genetic relationships of various complex diseases or traits [3, 4].

The correlations of different diseases were massively explored in recent years, especially from the perspective of heredity [5]. The correlations of different diseases usually indicate certain common etiological risk factors and biological changes. This could enlighten some common pathological features and therapeutic targets for the diseases. Therefore, revealing the relationships of different diseases is helpful for prevention, treatment and mechanism studies of target diseases [6]. GRS analysis is commonly applied to search for shared genetic aetiology among various diseases. Briefly, a GWAS of a 'base' disease is first conducted in a 'discovery' sample. The identified risk alleles and their effect sizes of the 'base' disease are then used to calculate GRS in an independent 'target' sample for a 'target' disease. Then, the genetic relationship between the 'base' disease and 'target' disease can be evaluated by statistical analysis (for example correlation analysis) using the calculated GRS as the proxy of the 'base' disease. For instance, Purcell et al. successfully discovered shared genetic components between schizophrenia and bipolar disorder using GRS analysis [7].

It is common that multiple correlated phenotypes are measured for complex diseases in practice. For example, hyperlipidemia status can be defined by total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol and triglycerides. The simple method is to analyze each of disease phenotypes individually [8]. However, without incorporating the genetic information of multiple related phenotypes, analyzing each disease phenotype individually is likely to miss the susceptibility loci with moderate or weak pleiotropic effects. Furthermore, a multiple testing correction procedure is mostly required,

which will reduce statistical power. Similarly, GRS analysis is often applied to evaluate the correlation between a 'target' disease and a 'base' disease with multiple correlated phenotypes. However, researchers usually conduct GRS analysis between the 'target' disease and individual phenotype of the 'base' disease separately, which will affect the statistical power because of multiple testing correction, and may miss the moderate or weak pleiotropic effects of susceptibility loci underlying the multiple related phenotypes of the 'base' disease (Figure 1A).

Joint analysis of the genetic information of multiple correlated phenotypes can help to discover pleiotropic variants and increase the statistical power for detecting the causal variants with moderate or weak genetic effects [9–12]. One approach for the joint analysis is to subject phenotypic data to a dimension reduction technique, such as principal component analysis (PCA). The extracted principal components (PCs) are the linear combination of individual disease phenotypes. Generally, each PC denotes a direction of genetic effects of causal variants. The top several PCs account for most of the phenotypic variance of target diseases. Researchers can then analyze the reduced number of extracted PCs rather than each individual phenotype. PCA has been widely used for joint analysis of multiple disease phenotypes in genetic linkage and association studies to detect disease-associated loci [13–15]. To the best of our knowledge, no study has explored the performance of PCA for enhancing the power of GRS analysis.

In this study, we developed a PCA-based GRS analysis approach for studying complex diseases with multiple correlated phenotypes. In our approach, PCA is first applied to the GRS matrix of multiple disease phenotypes for extracting a set of orthogonal GRS components. The top several extracted PCs are then included into correlation analysis instead of original GRS (Figure 1B). Simulation studies and a real Caucasian GWAS data were used to evaluate the performance of PCA-based GRS analysis.

Materials and methods

Primary arithmetic steps

Step 1—SNP selection for 'base' disease

For each phenotype of the 'base' disease, the independent causal variants and their genetic effects (such as beta or odds ratio) are driven from previous genetic association studies of the 'base' disease, typically GWAS.

Step 2—GRS calculation

Let E_i denote the genetic effect parameter (beta or logarithm of odds ratio) of the i th single nucleotide polymorphism (SNP) ($i = 1, 2, \dots, n$) associated with the j th phenotype ($j = 1, 2, \dots, m$) of

Figure 1. Comparison of traditional GRS analysis (A) with PCA-based GRS analysis (B) when testing the correlation between a 'base' disease with multiple correlated phenotypes and a 'target' disease. PC1: the first principal component of GRS matrix; PC2: the second principal component of GRS matrix; PC3: the third principal component of GRS matrix.

the 'base' disease; C_i denotes the number of risk allele of the i th SNP ($C_i = 0, 1$ or 2) derived from the genotype data of the 'target' disease. For the u th subject ($u = 1, 2, \dots, l$), the weighted GRS of the j th phenotype of the 'base' disease can be calculated by

$$GRS_{uj} = \sum_{i=1}^n E_i C_i$$

Step 3—PCA of GRS matrix

Let X denote the $l \times m$ GRS matrix for the m phenotypes of the 'base' disease. PCA is then applied to X for extracting orthogonal PCs. Briefly, let P_{uk} ($k = 1, 2, \dots, m$) denote the k th PC extracted from matrix X for the u th subject. P_{uk} is calculated by $P_{uk} = Z_k X_u$, where Z_k denotes the eigenvector of the k th PC from matrix X . X_u denotes the GRS vector of matrix X for the u th individual. After PCA, the original GRS data is transformed into a new coordinate system, such that a large part of GRS variance lies on the top several extracted PCs.

Step 4—Correlation analysis

The top several extracted PCs of GRS matrix are included into correlation analysis to evaluate the genetic relationships between the 'base' disease and 'target' disease. Without loss of generality, the first principal component (PC1) of GRS matrix is selected for correlation analysis in our simulation studies and real data analyses. Other selecting approaches can also be used.

For instance, we can select the top several PCs explaining a predefined proportion of the variance of GRS matrix.

Simulation studies

Simulation studies were conducted to evaluate the performance of PCA-based GRS analysis and traditional GRS analysis. Without loss of generality, we supposed that there were a 'target' disease and a 'base' disease with three measured phenotypes. GWAsimulator software [16] (<http://biostat.mc.vanderbilt.edu/GWAsimulator>) was used to generate the genotype and qualitative phenotype data of the 'target' disease and 'base' disease, respectively. GWAsimulator implements a rapid moving-window algorithm and can simulate genotype and phenotype data based on real Illumina HumanHap550 chip data [16]. 5000 SNP loci were simulated for each chromosome. Among the simulated chromosomes, we randomly selected five chromosomes to simulate causal loci.

Using GWAsimulator, we generated various scenarios, including sample sizes, relative risks of simulated causal loci, two-way interaction effects, inheritance models and numbers of common causal loci shared by the 'target' disease and each phenotype of the 'base' disease. For two-way interaction effects simulation, two of the five causal loci were randomly simulated with interaction effects for the 'target' disease and

Table 1. Parameter settings of simulation studies

	Model 1	Model 2	Model 3	Model 4	Model 5
Sample sizes	2000	3000	4000	5000	6000
Relative risks of causal loci	1.1	1.2	1.3	1.4	1.5
Two-way interaction effects ^a	1.0/1.0/1.0/1.0	1.2/1.2/ 1.5/1.5	1.5/1.5/2.0/2.0	2.0/2.0/2.5/2.5	2.5/2.5/3.0/3.0
Numbers of common loci between 'target' disease and each phenotype of 'base' disease	2	3	4	5	-
Inheritance models ^b	D	M	R	-	-

Note: ^a Interaction effects between two causal loci. Let D_{ij} denote two-way interaction effect between the first locus with i ($i = 1, 2$) copy of the risk allele and the second locus with j ($j = 1, 2$) copy of the risk allele. In the table above, the four decimals in each model denote D_{11} , D_{12} , D_{21} and D_{22} , respectively.

^b D: dominance model; M: multiplicative model; R: recessive model.

^c The default parameters are highlighted in italic.

^d For parameters with less than five settings, the blanks are filled by dashes.

each phenotype of the 'base' disease, respectively. For common causal loci simulation, the numbers of common causal loci shared by the 'target' disease and each phenotype of the 'base' disease were controlled to simulate different correlation levels between the 'base' and 'target' diseases as well as three phenotypes of the 'base' disease. The population prevalence of the 'base' and 'target' diseases was set to 0.1 in GWASimulator. For type I error rate simulations, the relative risks of all simulated causal loci were set to 1.0 under various scenarios. Additionally, we conducted a simulation study to evaluate the impact of noisy loci on the performance of GRS analysis. Without loss of generality, we selected all SNPs with $P < 10^{-3}$ for GRS calculation. Other simulation parameters were assigned to default values. Detailed description of simulation parameters are shown in Table 1.

During each simulation, the most significant SNP locus was selected from each of the five chromosomes containing simulated disease loci and used for calculating GRS by PLINK [17]. PCA was implemented by R (<http://www.r-project.org/>). t-tests were performed to evaluate the significance of GRS difference between simulated case and control samples of the 'target' disease. For traditional GRS analysis, significant correlations were detected at $P < 0.0167$ (0.05/3) after multiple testing correction. For PCA-based GRS analysis, the PC1 of GRS matrix was included into correlation analysis. 2000 simulations were conducted for each parameter setting. The power and type I error rates were calculated as the proportions of significant correlations detected in 2000 simulations, respectively.

Application to real data

To illustrate the performance of PCA-based GRS analysis, we applied both traditional GRS analysis and PCA-based GRS analysis to a real GWAS data of bone geometry ('target' disease). This GWAS sample included 2286 Caucasian subjects with bone area data at hip, spine and ulna and radius. Bone areas were measured using Hologic 4500W dual energy X-ray absorptiometry (Hologic Inc., Bedford, MA, USA) and adjusted for age, sex, height and weight using linear regression model. Genome-wide SNP genotyping was performed using Affymetrix Human SNP Array 6.0. We selected obesity and educational attainment-related phenotypes as the positive and negative 'base' diseases for bone areas, respectively. The 2017 GWAS summaries of body mass index (BMI) and 2015 GWAS summaries of waist-to-hip ratio (WHR) [18] were obtained from the GIANT consortium (http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files). The GWAS summaries of educational attainment-related phenotypes (including educational

attainment, cognitive performance, highest level math class completed and self-reported math ability) were collected from the Social Science Genetic Association Consortium (<https://www.thessgac.org/data>). The SNPs with $P < 10^{-5}$ in the GWAS of obesity and educational attainment-related phenotypes were selected for GRS calculation and subsequent correlation analysis with bone areas. For traditional GRS analysis, significant correlations were detected at $P < 0.025$ (0.05/2) for obesity-related phenotypes and $P < 0.0125$ (0.05/4) for educational attainment-related phenotypes after multiple testing correction.

Results

Simulation results

All original t-test results are summarized in Tables S1–S6. Figure 2 summarizes the simulation results of sample sizes (Figure 2A), relative risks of simulated causal loci (Figure 2B), effects of two-way interaction (Figure 2C) and numbers of common disease loci shared by the 'target' disease and each phenotype of the 'base' disease (Figure 2D). We observed that PCA-based GRS analysis generally outperformed traditional GRS analysis approach under various scenarios simulated by this study. Furthermore, as shown in Figures 2B and 2C, PCA-based GRS analysis appeared to have greater advantage under the models with smaller relative risks and two-way interaction effects (increased the power by 0.1 compared to traditional method). The simulation results of inheritance models are presented in Figure 3. Under the various inheritance models simulated by this study, PCA-based GRS analysis showed higher power than traditional GRS analysis approach, especially for recessive model. Additionally, we found that both PCA-based GRS analysis and traditional GRS analysis showed reduced power when noisy loci were included into GRS calculation. But the power of PCA-based GRS analysis (63%) was higher than that of the traditional GRS analysis (53%) in our simulation study. We also evaluated the type I error rates of PCA-based GRS analysis and traditional GRS analysis approach through simulating all causal loci without phenotypic effects. The type I error rates of both methods were all close to nominal 0.05 under various scenarios simulated by this study.

Real data analysis results

Genetic correlations among obesity as well as educational attainment-related phenotypes are shown in Table 2. Correlation analysis results of bone areas versus obesity-related

Figure 2. Simulation results of sample sizes (A), relative risks of simulated causal loci (B), effects of two-way interaction (C) and numbers of common causal disease loci shared by the 'target' disease and each phenotype of the 'base' disease (D). T-GRS: traditional GRS analysis; P-GRS: PCA-based GRS analysis.

Figure 3. Simulation results of inheritance models. T-GRS denotes traditional GRS analysis; P-GRS denotes PCA-based GRS analysis.

phenotypes and bone areas versus educational attainment-related phenotypes are shown in Table 3. After combining the genetic information of BMI and WHR, PCA-based GRS analysis detected significant correlation between spine total bone area and the PC1 of BMI and WHR. In contrast, traditional GRS analysis failed to detect significant correlation for spine total bone area versus BMI or WHR. For bone areas versus educational attainment-related phenotypes, traditional GRS analysis identified a significant correlation for ulna and radius total bone area versus educational attainment and a suggesting correlation for hip total bone area versus educational attainment. After combining the genetic information of four educational attainment-related phenotypes, PCA-based GRS analysis did not identify any correlation between bone areas and the PC1 of educational attainment-related phenotypes.

Discussion

GRS analysis is an increasingly popular method for exploring genetic architectures and genetic relationships of complex diseases [7, 19]. GRS is capable of combining the genetic effects of multiple causal loci into a single variable. Therefore, the power of GRS is superior over that of a single locus. However, complex diseases are usually measured by multiple correlated phenotypes in practice; unfortunately, researchers usually analyze each

Table 2. Genetic correlations among obesity as well as educational attainment-related phenotypes

Phenotype pairs	Correlation coefficient	P-value
BMI-WHR	0.5859	<0.0001
EA-CP	0.9046	<0.0001
CP-HM	0.7979	<0.0001
CP-MA	0.7782	<0.0001
EA-HM	0.9499	<0.0001
EA-MA	0.8395	<0.0001
HM-MA	0.9483	<0.0001

Note: BMI: body mass index; WHR: waist-to-hip ratio; EA: educational attainment; CP: cognitive performance; HM: highest-level math class completed; MA: self-reported math ability.

disease phenotype individually [8]. Analyzing each disease phenotype individually requires multiple testing correction, leading to reduced statistical power for GRS analysis, and is likely to miss potential pleiotropic effects. To address this issue, we proposed a PCA-based GRS analysis. This method can maximize the GRS variance of multiple correlated disease phenotypes; meanwhile, it can reduce the number of statistical tests. We observed that PCA could increase the statistical power of GRS analysis under various scenarios simulated by this study. The results of real data analysis further confirmed the improved performance of PCA-based GRS analysis.

It is worth nothing that as shown by Figures 2B and 2C, the improved performance of PCA-based GRS analysis appeared to be greater for the causal loci with moderate or weak genetic effects. One possible explanation is that due to the stringent threshold of significance after multiple testing correction, the moderate but meaningful correlations may be missed by single phenotype GRS analysis. In contrast, PCA-based GRS analysis incorporates the genetic effects of multiple disease phenotypes into single test statistic. The moderate correlation signals in single phenotype GRS analysis are likely to present strong correlation signals in our PCA-based GRS analysis reducing the impact of multiple testing corrections. Since complex diseases

Table 3. Correlation analysis results of real GWAS data of bone areas

Bone phenotypes	Traditional GRS				PCA-based GRS	
	BMI		WHR		r	P
	r	P	r	P		
Hip bone area	0.005	0.818	0.032	0.121	−0.021	0.31
Spine bone area	−0.015	0.482	0.036	0.087	−0.047	0.025
Ulna and radius bone area	−0.015	0.461	0.019	0.353	−0.034	0.104

Bone phenotypes	Traditional GRS				PCA-based GRS	
	EA		CP		r	P
	r	P	r	P		
Hip bone area	−0.049	0.019	−0.04	0.173	0.038	0.071
Spine bone area	−0.036	0.084	−0.01	0.173	0.007	0.741
Ulna and radius bone area	−0.065	0.002	−0.02	0.134	0.01	0.621

Note: BMI: body mass index; WHR: waist-to-hip ratio; EA: educational attainment; CP: cognitive performance; HM: highest-level math class completed; MA: self-reported math ability; r: correlation coefficient; P: P-value. Significant P-values and corresponding correlation coefficients are highlighted in bold.

are usually measured by multiple correlated phenotypes and determined by a group of causal loci, PCA-based GRS analysis may help to discover novel correlations between diseases.

This study aims to evaluate the performance of dimension reduction approach for enhancing the power of GRS studies of complex diseases with multiple related phenotypes. Given that PCA is a classical dimension reduction approach and is widely used for joint analysis of multiple correlated disease phenotypes [15, 20, 21], we used PCA to improve traditional single phenotype GRS analysis approach. Both simulation studies and real Caucasian data analysis supported that PCA could enhance the power of GRS analysis for detecting the correlations of complex diseases with multiple related phenotypes. Additionally, other dimension reduction approaches such as canonical correlation analysis [22], as well as other multivariate methods such as MultiPhen [23], and mixed effects models [12, 24] can also be applied to GRS analysis for joint analysis of multiple phenotypes. Comparative studies have been done for some of these approaches. For instance, Galesloot et al. performed a comprehensive study comparing the performance of multiple multivariate genetic association analysis methods using simulated data, providing useful information for researches to select proper association testing approach. They suggested that not one single method could always perform the best under all simulation scenarios [25]. Similarly, further comparative studies with extensive simulations and large scale GWAS data are warranted to evaluate the performance of various approaches for GRS correlation analysis of complex diseases with multiple related phenotypes.

Given the objective of this study, we considered one common situation that GRS was calculated from a few of significant SNPs with moderate genetic effects. For instance, Hammerschlag et al. conducted a GWAS of insomnia complaints. They calculated a polygenic score from six genome-wide significant SNP loci [26]. Therefore, without loss of generality, we simulated five causal loci with moderate genetic effects for each disease phenotype in this study. Additionally, the proportions of causal variants shared between the 'base' and 'target' diseases were controlled to simulate different correlation levels between the 'base' and 'target' diseases. The high correlation levels are generally common for real complex diseases in practice. For example, Zhu et al. evaluated the genetic correlation levels (denoted as correlation coefficient R_g) of multiple immune diseases. They detected

significant genetic correlations among different immune diseases, such as asthma and allergic diseases ($R_g = 0.75$), Crohn's disease and ulcerative colitis ($R_g = 0.57$) [27]. Additionally, Hammerschlag et al. observed significant genetic correlations between insomnia and other psychiatric traits, such as insomnia and anxiety disorders ($R_g = 0.59$), insomnia and depressive symptoms ($R_g = 0.53$) and insomnia and neuroticism ($R_g = 0.44$) [26]. Besides the simulated causal loci controlled by us, random polygenic effects were also simulated by GWASimulator software for each phenotype [16], which were not shared between the 'base' disease and 'target' disease. Therefore, we believe that the simulated proportions of common causal variants are generally acceptable in practice.

To evaluate the performance of PCA-based GRS analysis, we applied both traditional GRS analysis and PCA-based GRS analysis to a real GWAS data of bone areas. No significant correlation of bone areas and BMI/WHR was detected by traditional GRS analysis. After combining the genetic information of BMI and WHR, PCA-based GRS analysis observed significant negative correlation ($P = 0.025$) between spine bone area and the PC1 of BMI and WHR. Negative correlations between bone phenotypes and obesity-related phenotypes have been reported in previous studies [28–31]. For educational attainment-related phenotypes, traditional GRS analysis detected a significant negative correlation between ulna and radius bone area and educational attainment ($P = 0.002$). However, after combining the genetic information of four educational attainment-related phenotypes, PCA-based GRS analysis did not detect significant correlation between bone areas and the PC1 of four educational attainment-related phenotypes. To the best of our knowledge, no genetic correlation has been reported between bone areas and educational attainment, supporting the improved performance of PCA-based GRS analysis.

Using both simulation studies and real data analysis, we found that PCA could enhance the power of GRS analysis for detecting the correlations of complex diseases with multiple related phenotypes. To the best of our knowledge, no such work has been conducted before. In addition, PCA-based GRS analysis has two advantages over traditional GRS analysis when they are applied to identify correlation between a 'base' disease with several phenotypes and a 'target' disease. First, multiple testing corrections are usually required by traditional single phenotype

GRS analysis, the moderate but meaningful correlations may be missed due to the stringent significant thresholds after multiple testing corrections. In contrast, through incorporating the genetic information of multiple related phenotypes of the 'base' disease into single test statistic, PCA-based GRS analysis is more powerful to detect the moderate or weak correlation signals. Second, pleiotropic susceptibility variants are common for complex diseases or traits, especially for multiple related phenotypes of one disease. Compared with traditional single phenotype GRS analysis approach, which individually analyzes each of multiple related phenotypes, PCA-based GRS analysis should be more powerful to capture the pleiotropic effects of susceptibility loci underlying multiple related disease phenotypes, providing more useful information for GRS analysis of complex diseases.

Real data analysis found that PCA-based GRS performed well under both moderate (BMI and WHR) and strong (four educational attainment-related phenotypes) correlation levels among the multiple phenotypes of the 'base' diseases. This may attribute to the advantage of PCA for multiple phenotypes analysis. The top PC of GRS can capture major genetic information of multiple related phenotypes of the 'base' disease. However, it is worth noting that more PCs should be included into correlation analysis if the correlation levels of the multiple phenotypes of the 'base' disease were low. For instance, we can select top several PCs explaining large part of the GRS variance of the 'base' disease.

In conclusion, we developed a PCA-based GRS analysis approach. The results of both simulation and real data analysis supported the improved performance of PCA-based GRS analysis compared to traditional GRS analysis. PCA-based GRS analysis is more suitable for the complex diseases having multiple phenotypes and causal loci with moderate or weak genetic effects. Given that GWAS have flourished in the past decades, our approach may help researchers to explore the genetic architectures and relationships of complex diseases or traits.

Key Points

- We developed a PCA-based GRS analysis approach.
- PCA-based GRS analysis is able to simultaneously analyze multiple disease phenotypes.
- Simulation studies demonstrated that our approach is significantly superior over traditional GRS analysis.
- Real data analyses illustrated the application of our proposed approach.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

This study was supported by the National Natural Scientific Foundation of China (81472925, 81673112) and the Key projects of international cooperation among governments in scientific and technological innovation (2016YFE0119100). This study was partially supported by grants from the National Institutes of Health (R01AR057049, R01AR059781, P20 GM109036, R01MH107354, R01MH104680, R01GM109068,

AR069055 and U19 AG055373 to Q.T., H.S. and H-W.D.); the Edward G. Schlieder Endowment fund; and the Tsai and Kung endowment fund to Tulane University.

References

1. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;**42**:1001–6.
2. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013;**9**: e1003348.
3. Kawai VK, Levinson RT, Adefurin A, et al. A genetic risk score that includes common type 2 diabetes risk variants is associated with gestational diabetes. *Clin Endocrinol (Oxf)* 2017;**87**:149–55.
4. Shahid SU, Shabana, Cooper JA, et al. Genetic risk analysis of coronary artery disease in Pakistani subjects using a genetic risk score of 21 variants. *Atherosclerosis* 2017;**258**:1–7.
5. Ripke S. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* 2013;**45**:984–94.
6. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015;**47**:1236–41.
7. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009;**460**:748.
8. Willer CJ, Sanna S, Jackson AU, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 2008;**40**:161–9.
9. Solovieff N, Cotsapas C, Lee PH, et al. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 2013;**14**: 483–95.
10. Stephens M. A unified framework for association analysis with multiple related phenotypes. *PLoS One* 2013;**8**:e65245.
11. Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. *J Prob Stat* 2012;**2012**:652569.
12. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 2014;**11**:407–9.
13. He LN, Al E. Genomewide linkage scan for combined obesity phenotypes using principal component analysis. *Ann Hum Genet* 2008;**72**:319–26.
14. Karasik D, Cheung CL, Zhou Y, et al. Genome-wide association of an integrated osteoporosis-related phenotype: is there evidence for pleiotropic genes? *J Bone Miner Res* 2012;**27**:319–30.
15. Karasik D, Cupples LA, Hannan MT, et al. Genome screen for a combined bone phenotype using principal component analysis: the Framingham study. *Bone* 2004;**34**:547–56.
16. Li C, Li M. GWASimulator: a rapid whole-genome simulation program. *Bioinformatics* 2008;**24**:140.
17. 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**:559–75.
18. Shungin D, Winkler TW, Croteau-Chonka DC, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 2015;**518**:187–96.
19. Guo W, Samuels JF, Wang Y, et al. Polygenic risk score and heritability estimates reveals a genetic relationship between ASD and OCD. *Eur Neuropsychopharmacol* 2017;**27**: 657–66.

20. He LN, Liu YJ, Xiao P, et al. Genomewide linkage scan for combined obesity phenotypes using principal component analysis. *Ann Hum Genet* 2008;**72**:319–26.
21. Rainwater DL, Mahaney MC, Vandeberg JL, et al. A quantitative trait locus influences coordinated variation in measures of ApoB-containing lipoproteins. *Atherosclerosis* 2004;**176**:379–86.
22. Ferreira MA, Purcell SM. A multivariate test of association. *Bioinformatics* 2009;**25**:132–3.
23. O'Reilly PF, Hoggart CJ, Pomyen Y, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* 2012;**7**:e34861.
24. Korte A, Vilhjálmsson BJ, Segura V, et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* 2012;**44**:1066–71.
25. Galesloot TE, Van SK, Kiemeneij LA, et al. A comparison of multivariate genome-wide association methods. *PLoS One* 2014;**9**:e95923.
26. Hammerschlag AR, Stringer S, de Leeuw CA, et al. Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat Genet* 2017;**49**:1584–92.
27. Zhu Z, Lee PH, Chaffin MD, et al. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nat Genet* 2018;**50**:857–64.
28. Goulding A, Taylor RW, Jones IE, et al. Overweight and obese children have low bone mass and area for their weight. *Int J Obes Relat Metab Disord* 2000;**24**:627–32.
29. Afghani A, Goran MI. Racial differences in the association of subcutaneous and visceral fat on bone mineral content in prepubertal children. *Calcif Tissue Int* 2006;**79**:383–8.
30. Dimitri P, Wales JK, Bishop N. Fat and bone in children: differential effects of obesity on bone size and mass according to fracture history. *J Bone Miner Res* 2010;**25**:527–36.
31. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 2014;**11**:407–9.