PAPA: a flexible tool for identifying pleiotropic pathways using genome-wide association study summaries

academic.oup.com/bioinformatics/article/32/6/946/1744094

Abstract

Summary: Pleiotropy is common in the genetic architectures of complex diseases. To the best of our knowledge, no analysis tool has been developed for identifying pleiotropic pathways using multiple genome-wide association study (GWAS) summaries by now. Here, we present PAPA, a flexible tool for pleiotropic pathway analysis utilizing GWAS summary results. The performance of PAPA was validated using publicly available GWAS summaries of body mass index and waist-hip ratio of the GIANT datasets. PAPA identified a set of pleiotropic pathways, which have been demonstrated to be involved in the development of obesity.

Availability and implementation: PAPA program, document and illustrative example are available at http://sourceforge.net/projects/papav1/files/.

Contact:fzhxjtu@mail.xjtu.edu.cn

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

Issue Section:

GENETICS AND POPULATION ANALYSIS

1 Introduction

Pleiotropy describes the genetic phenomenon of a single gene affecting multiple phenotypes. It can be explained by single gene product having various biological functions, or acting as a signal factor implicated in the development of different phenotypes. Pleiotropy is common in the genetic architectures of complex diseases (Sivakumaran *et al.*, 2011). A total of 16.9% of genes recorded in the NHGRI Catalog of published genome-wide association studies (GWAS) have pleiotropic effects (Sivakumaran *et al.*, 2011). Clarifying the molecular mechanism of pleiotropy is helpful for pathogenetic studies and drug development of human complex diseases.

Recent achievements in GWAS provide a good opportunity for systematical pleiotropy studies of complex diseases. A simple approach is to analyse disease phenotypes separately using univariate approaches. The study results of different disease phenotypes are compared, which may result in low statistical power. To address this issue, multiple pleiotropy analysis approaches and tools have been proposed (Lee *et al.*, 2012; Nyholt, 2014). A group of genes with pleiotropic effects were identified for complex diseases (Elliott *et al.*, 2013; International Schizophrenia *et al.*, 2009). However, current pleiotropic mechanism studies of complex diseases mostly focus on individual pleiotropic genes, which are sometimes functionally independent. Some individual pleiotropic genes participate in multiple cellular processes. The biological functions of a part of genes remain elusive now. Therefore, identifying individual pleiotropic genes often provides limited information for genetic studies of complex diseases.

Inspired by the gene set enrichment analysis (GSEA) of microarray data (Subramanian *et al.*, 2005), pathway-based GWAS were proposed (Wang *et al.*, 2007) and successfully applied in the genetic studies of complex diseases (Zhang *et al.*, 2010). However, to the best of our knowledge, no analysis tool has been developed for identifying pleiotropic pathways using GWAS summary results by now.

In this study, we extended the pathway analysis algorithm proposed by Wang *et al.* (2007), and developed a pleiotropic pathway analysis tool PAPA. We applied PAPA to public available GWAS summaries of body mass index (BMI) and waist-hip ratio (WHR) of the Genetic Investigation of ANthropometric Traits datasets (Speliotes *et al.*, 2010; Heid *et al.*, 2010).

2 Methods

2.1. Implementation

Step 1 – Assigning association testing statistics to genes

We suppose that GWAS summaries of *M* genes and *N* phenotypes were available. Let S_{ij} denote the association testing statistic (for instance, chi-square values for qualitative traits) of *j*th SNP for *i*th phenotype ($i = 1,2,\ldots,N$). SNPs are assigned to genes by distance. A physical distance of 500 kb is used to connect a SNP to a gene in this study. For *i*th phenotype and *m*th gene ($m = 1,2,\ldots,M$), we select the largest S_{ij} from the SNPs assigned to the gene as the score of the gene, denoted as S_{im}. All genes are ranked by sorting their scores S_{im} from largest to smallest (Sri1≥Sri2≥....SriMSi1r≥Si2r≥....SiMr), which is denoted as Sri=[Sri1,Sri2,....SriM]Sir=[Si1,rSi2r,....SiMr].

Step 2 – Calculating enrichment scores

For a given pathway *P* consisting of MPMP genes, let GvGv denote the *v*th gene (*v* = 1,2,...,MPMP) of pathway *P.* Let ESPiESiP denote the enrichment score (ES) of pathway *P* for *i*th phenotype. ESPiESiP is calculated by Kolmogorov–Smirnov-like running sum statistic (Wang *et al.*, 2007):
ESPi=max1≤v≤M {∑Gu∈P,u≤v|Sriu|NRi−∑Gu∉P,u≤v1M−MP},where NRi=∑Gu∈P|Sriu|.ESiP=max1≤v≤M{∑Gu∈P,u≤v|Siur|NRi−∑Gu∉P,u≤v1M−MP},where

Step 3 – Permutation and centralization

To obtain the null distribution of ESPiESiP, permutations were conducted through circular genome permutation (Cabrera *et al.*, 2012). For *k*th permutation, let ESPnullikESikPnull denote the ES value of pathway *P* for *i*th phenotype, calculating from permutated data. After *K* times permutations, we can obtain the null distribution of ESPiESiP, denoted as ESPnulli=[ESPnulli1,ESPnulli2,...,ESPnulliK]ESiPnull=[ESi1Pnull,ESi2Pnull,...,ESiKPnull] . For pathway *P*, we calculate the centered ES (CES) of observed data (denote as CESPCESP) and permutated data (denote as CESPnull=[CESPnull1,CESPnull2,...,CESPnullK]CESPnull=[CES1Pnull,CES2Pnull,...,CESKPnull]) of *N* phenotypes, defined by CESP=∑i=1NESPimean(ESPnulli)×wi and CESPnullk=∑i=1NESPnullikmean(ESPnulli)×wi,CESP=∑i=1NESiPmean(ESiPnull)×wiandCESkPnull=∑i=1NESikPnullme

Where *k* (*k* = 1,2,...,*K*) denotes *k*th permutation. wiwi is the weight parameter of *i*th phenotype. For instance, wiwi can be assigned as the proportion of GWAS samples of *ith* phenotype in total samples.

Step 4 – Calculating normalized CES

The normalized CES (NCES) of pathway *P* is defined by NCESP=CESP−mean(CESPnull)SD(CESPnull).NCESP=CESP−mean(CESPnull)SD(CESPnull).

The null distribution of NCESPNCESP, which is denoted as NCESPnull=[NCESPnull1,NCESPnull2,...,NCESPnullK]NCESPnull=[NCES1Pnull,NCES2Pnull,...,NCESKPnull], can be calculated from *K* permutations using the following formula, NCESPnullk=CESPnullk−mean(CESPnull)SD(CESPnull),NCESkPnull=CESkPnull−mean(CESPnull)SD(CESPnull),

where *k* (*k* = 1,2,...,*K*) denotes *k*th permutation. After normalization, the NCES values of pathways with different sizes can be directly compared with each other (Wang *et al.*, 2007).

Step 5 – Calculating P values

Statistical testing *P* value of each pathway is calculated as the proportion of NCESPNCESP being smaller than NCESPnullNCESPnull in *K* times permutations.

2.2. Application to GWAS summaries of BMI and WHR

GWAS summaries of BMI were obtained from Speliotes *et al.* (2010), containing 123 865 study subjects of European ancestry. GWAS summaries of WHR were collected from Heid *et al.* (2010), including 77 167 study subjects of European ancestry. 3269 pathways or gene ontology categories collecting from the Molecular Signatures Database of GSEA were analyzed (Subramanian *et al.*, 2005). The weighting parameters of BMI and WHR were 0.62 and 0.38, respectively. 1000 permutations were conducted by PAPA to calculate the empirical *P* value of each pathway.

3 Results and discussion

As shown by Supplementary Table S1, the top three significant pathways functionally involved in adipocyte differentiation, Wnt signalling and synthesis of bile acids and bile salts, which have been demonstrated to be involved in the development of obesity (Mori *et al.*, 2012; Prinz *et al.*, 2015). The computational cost of PAPA program is affordable. The GWAS datasets of BMI and WHR were analyzed (1000 permutations) on a Dell computer with Intel Xeon CPU E5620 (2.4 GHz) and 4 GB memory. PAPA spent 98 h to complete data analysis.

In this study, we extended the GSEA algorithm (Subramanian *et al.*, 2005; Wang *et al.*, 2007), and developed a pleiotropic pathway analysis tool PAPA. Because of integrating GWAS results and prior knowledge of biological pathways, PAPA may provide novel clues for clarifying the pleiotropic mechanisms of human complex diseases. It should be noted that the definition of biological pathways may affect the performance of PAPA. In PAPA package, we provided two pathway gene annotation files, which were collected from public pathway database, including KEGG Pathway Database, Reactome Pathway Database, BioCarta, Ambion GeneAssist Pathway Atlas, Gene Ontology and GSEA Molecular Signatures Database.

Funding

The study was supported by National Natural Scientific Fund of China (81472925).

Conflict of Interest: none declared.

References

Author notes

Associate Editor: Oliver Stegle