Bioinformatics, 32(6), 2016, 946–948
doi: 10.1093/bioinformatics/btv668
Advance Access Publication Date: 14 November 2015
Applications Note

Genetics and population analysis

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

Yan Wen, Wenyu Wang, Xiong Guo and Feng Zhang*

Key Laboratory of Trace Elements and Endemic Diseases of Ministry of Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, People's Republic of China

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on August 5, 2015; revised on November 3, 2015; accepted on November 7, 2015

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.

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 jth SNP for ith 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 ith phenotype and mth 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 ($S^r_{i1} \geq S^r_{i2} \geq \ldots S^r_{iM}$), which is denoted as $S^r_i = [S^r_{i1}, S^r_{i2}, \ldots S^r_{iM}]$.

Step 2 – Calculating enrichment scores

For a given pathway P consisting of M_P genes, let G_v denote the vth gene ($v = 1,2,\ldots,M_P$) of pathway P. Let ES^P_i denote the enrichment score (ES) of pathway P for ith phenotype. ES^P_i is calculated by Kolmogorov–Smirnov-like running sum statistic (Wang *et al.*, 2007):

$$ES^P_i = \max_{1\leq v\leq M}\left\{\sum_{G_u\in P, u\leq v}\frac{|S^r_{iu}|}{N_{Ri}} - \sum_{G_u\notin P, u\leq v}\frac{1}{M-M_P}\right\}, \text{where } N_{Ri} = \sum_{G_u\in P}|S^r_{iu}|.$$

Step 3 – Permutation and centralization

To obtain the null distribution of ES^P_i, permutations were conducted through circular genome permutation (Cabrera *et al.*, 2012). For kth permutation, let ES^{Pnull}_{ik} denote the ES value of pathway P for ith phenotype, calculating from permutated data. After K times permutations, we can obtain the null distribution of ES^P_i, denoted as $ES^{Pnull}_i = [ES^{Pnull}_{i1}, ES^{Pnull}_{i2}, \ldots, ES^{Pnull}_{iK}]$. For pathway P, we calculate the centered ES (CES) of observed data (denote as CES^P) and permutated data (denote as $CES^{Pnull} = [CES^{Pnull}_1, CES^{Pnull}_2, \ldots, CES^{Pnull}_K]$) of N phenotypes, defined by

$$\begin{aligned} CES^P &= \sum_{i=1}^{N}\frac{ES^P_i}{mean(ES^{Pnull}_i)} \times w_i \text{ and } CES^{Pnull}_k \\ &= \sum_{i=1}^{N}\frac{ES^{Pnull}_{ik}}{mean(ES^{Pnull}_i)} \times w_i, \end{aligned}$$

Where k ($k = 1,2,\ldots,K$) denotes kth permutation. w_i is the weight parameter of ith phenotype. For instance, w_i 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

$$NCES^P = \frac{CES^P - mean(CES^{Pnull})}{SD(CES^{Pnull})}.$$

The null distribution of $NCES^P$, which is denoted as $NCES^{Pnull} = [NCES^{Pnull}_1, NCES^{Pnull}_2, \ldots, NCES^{Pnull}_K]$, can be calculated from K permutations using the following formula,

$$NCES^{Pnull}_k = \frac{CES^{Pnull}_k - mean(CES^{Pnull})}{SD(CES^{Pnull})},$$

where k ($k = 1,2,\ldots,K$) denotes kth 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 $NCES^P$ being smaller than $NCES^{Pnull}$ 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

Cabrera,C.P. *et al.* (2012) Uncovering networks from genome-wide association studies via circular genomic permutation. *G3*, **2**, 1067–1075.

Elliott,K.S. *et al.* (2013) Evaluation of the genetic overlap between osteoarthritis with body mass index and height using genome-wide association scan data, *Ann. Rheum. Dis.*, **72**, 935–941.

Heid,I.M. *et al*. (2010) Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution, *Nat. Genet.*, **42**, 949–960.

International Schizophrenia,C. *et al.* (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.

Lee,S.H. *et al*. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, **28**, 2540–2542.

Mori,H. *et al*. (2012) Secreted frizzled-related protein 5 suppresses adipocyte mitochondrial metabolism through WNT inhibition. *J. Clin. Invest.*, **122**, 2405–2416

Nyholt,D.R. (2014) SECA: SNP effect concordance analysis using genome-wide association summary results. *Bioinformatics*, **30**, 2086–2088.

Prinz,P. *et al*. (2015) Plasma bile acids show a positive correlation with body mass index and are negatively associated with cognitive restraint of eating in obese patients. *Front. Neurosci.*, **9**, 199.

Speliotes,E.K. *et al*. (2010) Association analyses of 249 796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, **42**, 937–948.

Sivakumaran,S. *et al*. (2011) Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.*, **89**, 607–618.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Wang,K. *et al*. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.

Zhang,L. *et al*. (2010) Pathway-based genome-wide association analysis identified the importance of regulation-of-autophagy pathway for ultradistal radius BMD. *J. Bone. Miner. Res.*, **25**, 1572–1580.