Bioinformatics, 2018, 1–3
doi: 10.1093/bioinformatics/bty061
Advance Access Publication Date: 8 February 2018
Applications Note

Genetics and population analysis

Improving SNP prioritization and pleiotropic architecture estimation by incorporating prior knowledge using graph-GPA

Hang J. Kim[1], **Zhenning Yu**[2], **Andrew Lawson**[2], **Hongyu Zhao**[3,4,5] **and Dongjun Chung**[2,*]

[1]Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221, USA, [2]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29425, USA, [3]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA, [4]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA and [5]Department of Genetics, Yale School of Medicine, New Haven, CT 06520, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on July 19, 2017; revised on January 10, 2018; editorial decision on February 4, 2018; accepted on February 6, 2018

Abstract

Summary: Integration of genetic studies for multiple phenotypes is a powerful approach to improving the identification of genetic variants associated with complex traits. Although it has been shown that leveraging shared genetic basis among phenotypes, namely pleiotropy, can increase statistical power to identify risk variants, it remains challenging to effectively integrate genome-wide association study (GWAS) datasets for a large number of phenotypes. We previously developed graph-GPA, a Bayesian hierarchical model that integrates multiple GWAS datasets to boost statistical power for the identification of risk variants and to estimate pleiotropic architecture within a unified framework. Here we propose a novel improvement of graph-GPA which incorporates external knowledge about phenotype–phenotype relationship to guide the estimation of genetic correlation and the association mapping. The application of graph-GPA to GWAS datasets for 12 complex diseases with a prior disease graph obtained from a text mining of biomedical literature illustrates its power to improve the identification of risk genetic variants and to facilitate understanding of genetic relationship among complex diseases.

Availability and implementation: graph-GPA is implemented as an R package 'GGPA', which is publicly available at http://dongjunchung.github.io/GGPA/. DDNet, a web interface to query diseases of interest and download a prior disease graph obtained from a text mining of biomedical literature, is publicly available at http://www.chunglab.io/ddnet/.

Contact: chungd@musc.edu

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

1 Introduction

Genome-wide association studies (GWAS) have identified over 37 000 single nucleotide polymorphisms (SNPs) associated with at least one disease or trait (https://www.ebi.ac.uk/gwas/). However, the identification of genetic variants associated with complex traits still remains challenging due to their large number and small effect sizes, namely polygenicity (Manolio *et al.*, 2008). Recently, it has been shown that integration of GWAS results can improve statistical power to identify risk variants by leveraging shared genetic basis between distinct phenotypes, namely pleiotropy (Manolio *et al.*, 2008). In order to address these challenges and opportunities, Chung *et al.* (2017) recently proposed graph-GPA, a Bayesian hierarchical model to integrate GWAS datasets for multiple phenotypes using a hidden Markov random field (MRF) architecture. Simulation studies and real data analyses showed that graph-GPA improves statistical power to identify risk genetic variants compared with traditional separate analyses and previous approaches developed to leverage pleiotropy such as GPA (Chung *et al.*, 2014). In addition, graph-GPA also estimates the pleiotropic architecture in the form of a phenotype graph. However, the original graph-GPA was based on uninformative prior distribution for the phenotype graph and did not allow to incorporate prior knowledge from external sources. In order to address this limitation, we have developed an improved version of graph-GPA that allows to incorporate a prior phenotype graph, which can potentially improve both the association mapping and the pleiotropic architecture estimation.

2 Materials and methods

graph-GPA is implemented as an R package 'GGPA', which provides model fitting, association mapping, and pleiotropic architecture estimation functionalities with a simple and intuitive user interface. The R package 'GGPA', is publicly available at its GitHub webpage (http://dongjunchung.github.io/GGPA/). graph-GPA takes GWAS summary statistics (phenotype-genotype association *P*-values) for multiple phenotypes as input. This will potentially allow the wide application of graph-GPA because summary statistics are often readily available for many genetic studies. In order to take into account pleiotropic structure, graph-GPA utilizes an MRF architecture to model latent indicators for phenotype-genotype association. Specifically, the pleiotropic architecture is represented as a phenotype graph, where each node corresponds to a phenotype and an edge between two phenotypes represents the genetic correlation between them. In 'GGPA', parameters are efficiently estimated using a Metropolis-Hastings within a Gibbs sampler while computational efficiency is further facilitated by utilizing the R package 'Rcpp', a seamless interface between R and C++ (Supplementary Material).

The original implementation of graph-GPA did not allow to incorporate prior knowledge on genetic relationship among phenotypes. Our new implementation allows the incorporation of prior knowledge in the form of a phenotype graph. Specifically, if there exists external source information that supports potential genetic correlation between two phenotypes, then the corresponding edge is 'forced in' in the phenotype graph during the Markov chain Monte Carlo (MCMC) procedures (Supplementary Material). This allows the prior knowledge to guide the estimation of phenotype graph. More importantly, this approach also improves prioritization of SNPs with information sharing to more accurate directions. Note that this 'forcing in' does not necessarily 'add' an edge to the final phenotype graph provided to users because the final phenotype graph is determined based on both the estimated phenotype graph and the MRF coefficient estimates. For example, if the GWAS data indicate no or negligible genetic correlation between two phenotypes whose edge is forced in, its MRF coefficient estimate would be close to zero and the corresponding edge would be omitted from the final phenotype graph. Hence, the final pleiotropic architecture estimation and the association mapping results are determined as a compromise between prior knowledge and the information in genetic studies.

In this article, we construct a prior disease graph based on the pattern of gene sharing between diseases in the PubMed literature mining. To facilitate users' convenience to utilize this workflow, we developed 'DDNet' (http://www.chunglab.io/ddnet), a web interface that allows users to query a list of diseases and download the corresponding disease graph, which can be directly fed into the R package 'GGPA'. Alternatively, a prior disease graph can also be constructed using other sources of information, such as shared symptoms, comorbidities, coexpression, and biological similarities, among others (Supplementary Material).

3 Examples

To demonstrate the usefulness of graph-GPA, we analyzed the summary statistics from genetic studies for 12 complex diseases, including attention deficit-hyperactivity disorder (ADHD), autism spectrum disorder (ASD), bipolar disorder (BPD), major depressive disorder (MDD), schizophrenia (SCZ), Crohn's disease (CD), ulcerative colitis (UC), systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), type 1 diabetes (T1D), type 2 diabetes (T2D) and coronary artery disease (CAD) (Supplementary Material). We generated a prior disease graph from the literature mining using the web interface 'DDNet'. In this graph, we observed edges among all the five psychiatric disorders, among all the autoimmune diseases, and between type 1 and 2 diabetes (Supplementary Material). To evaluate quality of this prior information, we implemented gene set enrichment analyses for top ranking genes predicted to be associated with each disease in the literature mining. We found that these genes are enriched for the corresponding disease and also for the pathways related to each disease (Supplementary Material). Figure 1 shows the disease graphs estimated using graph-GPA without and with the prior disease graph. We can see that incorporating the prior knowledge changes the estimated phenotype graph structure. For example, when we incorporated the prior disease graph, the edge between BPD and MDD was added while the edge between CD and T2D was removed. Note that the strong pleiotropy between BPD and MDD has been reported, e.g. Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* (2013).

We further evaluated the effects of incorporating prior knowledge on the association mapping results. Overall, incorporating the prior disease graph does not significantly change the identified SNPs associated with each disease (Supplementary Material). However, incorporation of the prior disease graph impacts identification of the SNPs shared between disease pairs corresponding to edges that were newly added or removed, e.g. those between BPD and MDD and between CD and T2D. Specifically, when the prior disease graph was incorporated, the number of SNPs shared between BPD and MDD increased from 20 to 36 SNPs while the number of SNPs shared between CD and T2D decreased from 38 to 26 SNPs. We further evaluated the functional impacts of these changes using the overall and tissue-specific functional scores for each genomic loci profiled in GenoCanyon (Lu *et al.*, 2015) and GenoSkyline (Lu *et al.*, 2016),

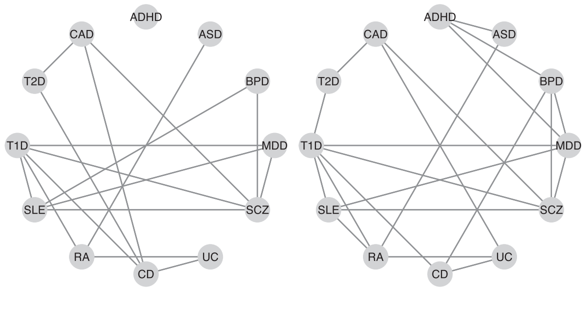

Fig. 1. Pleiotropic architectures estimated using graph-GPA without (left) and with (right) using a prior disease graph obtained from a text mining of PubMed literature

respectively. We found that when the prior disease graph was incorporated, both the GenoCanyon and GenoSkyline scores increased for both the BPD–MDD and CD–T2D pairs (Supplementary Material). This might imply that incorporating a prior disease graph can boost statistical power to identify risk variants and also remove potential false positives.

Finally, we found that when a prior graph is incorporated, findings are more reproducible in an independent validation GWAS dataset (Supplementary Material). More importantly, our simulation studies and real data analyses indicate that incorporating a prior graph also improves accuracy of phenotype graph estimation when there exists collinearity among phenotypes (Supplementary Material).

Funding

This work has been supported by the National Institutes of Health [R01 GM122078 and R21 CA209848 awarded to D.C.].

Conflict of Interest: none declared.

References

Chung,D. *et al.* (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet.*, **10**, e1004787.

Chung,D. *et al.* (2017) graph-GPA: a graphical model for prioritizing GWAS results and investigating pleiotropic architecture. *PLoS Comput. Biol.*, **13**, e1005388.

Cross-Disorder Group of the Psychiatric Genomics Consortium. *et al.* (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.*, **45**, 984–994.

Lu,Q. *et al.* (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.*, **5**, 10576.

Lu,Q. *et al.* (2016) Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genet.*, **12**, e1005947.

Manolio,T.A. *et al.* (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Investig.*, **118**, 1590–1605.