BIOINFORMATICS APPLICATIONS NOTE
Vol. 31 no. 2 2015, pages 262–264
doi:10.1093/bioinformatics/btu644

Genome analysis
Advance Access publication September 25, 2014

PINBPA: Cytoscape app for network analysis of GWAS data

Lili Wang[1], Takuya Matsushita[2], Lohith Madireddy[2], Parvin Mousavi[1] and Sergio E. Baranzini[2,*]
[1]School of Computing, Queen's University, 25 Union Street, Goodwin Hall, Kingston, Ontario K7L 3N6, Canada and [2]Department of Neurology, University of California San Francisco, 675 Nelson Rising Lane, Room 215, San Francisco, CA 94158, USA

Associate Editor: John Hancock

ABSTRACT

Summary: Protein interaction network-based pathway analysis (PINBPA) for genome-wide association studies (GWAS) has been developed as a Cytoscape app, to enable analysis of GWAS data in a network fashion. Users can easily import GWAS summary-level data, draw Manhattan plots, define blocks, prioritize genes with random walk with restart, detect enriched subnetworks and test the significance of subnetworks via a user-friendly interface.

Availability and implementation: PINBPA app is freely available in Cytoscape app store.

Contact: pmousavi@cs.queensu.ca and sebaran@cgl.ucsf.edu

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

Received on August 8, 2014; revised on September 18, 2014; accepted on September 22, 2014

1 INTRODUCTION

Genome-wide association studies (GWAS) continue to be a widely used approach to detect genetic associations with a phenotype of interest in well-defined populations. As of September 15, 2014, almost 2000 publications have reported associations of >13 000 single-nucleotide polymorphisms (SNPs) with close to 200 phenotypes in the GWAS catalog (Welter *et al*., 2014). The successful record of this genomic mapping strategy includes the identification of dozens or even hundreds of susceptibility alleles in common diseases, such as multiple sclerosis (MS), type 1 and type 2 diabetes, lymphomas, leukemias and metabolic disorders. Despite the unquestionable utility of this method, most of the data generated by GWAS are neglected because of the heavy emphasis devoted to eliminate false discoveries (type I error). Typically, a stringent threshold ($P\text{-}value < 5 \times 10^{-8}$) is applied to minimize type I error, thus inevitably increasing the proportion of false-negative results (type II error). Although this is a necessary tactic to effectively evaluate studies testing up to millions of markers individually, a number of methods that analyze groups of markers simultaneously (thus potentially increasing statistical power) have recently emerged (Huang Da *et al*., 2009; Khatri *et al*., 2012; Lee *et al*., 2012; Wang *et al*., 2007; Yaspan *et al*., 2011). These approaches, collectively known as pathway analysis, aim at identifying functional relationships among associated signals. Given that susceptibility to complex human diseases is likely a result of genes operating as part of functional modules rather than individual effects (Lage *et al*., 2007), pathway analysis methods hold promise in discovering additional associations from existing GWAS data.

The most recent class of pathway analysis methods is network based, and they largely overcome the assumptions of independence and preselection of reference database that limited its predecessors. Network-based analyses commonly use a scaffold of protein interactions to build connections between gene products, where nodes represent proteins and edges represent physical or functional interactions between pairs of proteins. Rather than focusing on individual markers, network-based analysis methods take into account multiple loci in the context of molecular pathways. Owing to this critical feature, these methods can afford to use sub genome-wide statistical significance and yet increase the power to detect new associations and functional relationships between genes in complex traits. Several network-based methods have been proposed to identify active modules (subnetworks) in a given network, such as DAPPLE (Rossin *et al*., 2011), dmGWAS (Jia *et al*., 2011) and NIMMI (Akula *et al*., 2011).

The original protein interaction network-based pathway analysis (PINBPA) method was first developed in the context of MS research (Baranzini *et al*., 2009) and most recently used to successfully identify novel associations by the International MS Genetics Consortium (International Multiple Sclerosis Genetics Consortium, 2013). In this article, we introduce the PINBPA app for Cytoscape (Shannon *et al*., 2003).

2 IMPLEMENTATION AND FEATURES

PINBPA has been implemented as an app for Cytoscape 3.0 and later versions using Java. Additionally, R scripts are called via Rserve inside Cytoscape for plotting.

Like many other pathway analysis methods, PINBPA requires gene-level summary statistics (*P*-values), as those generated by the popular tool VEGAS (Liu *et al*., 2010). As shown in Figure 1, the PINBPA app directly reads the VEGAS output as input file (but other formats are also possible), and has six features: (i) generates a gene-wise Manhattan plot of the GWAS; (ii) sorts all genes by their genomic coordinates and defines association blocks at any user-defined threshold ($P\text{-}value < 0.05$ by default); (iii) annotates the user-selected PPI network with imported gene-wise GWAS *P*-values; (iv) generates a subnetwork of only significant genes (first-order networks) exceeding a user-defined threshold, and tests the statistical significance of the sub-networks using random permutations; (v) runs network smoothing (an optional gene prioritization scheme) using a

*To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

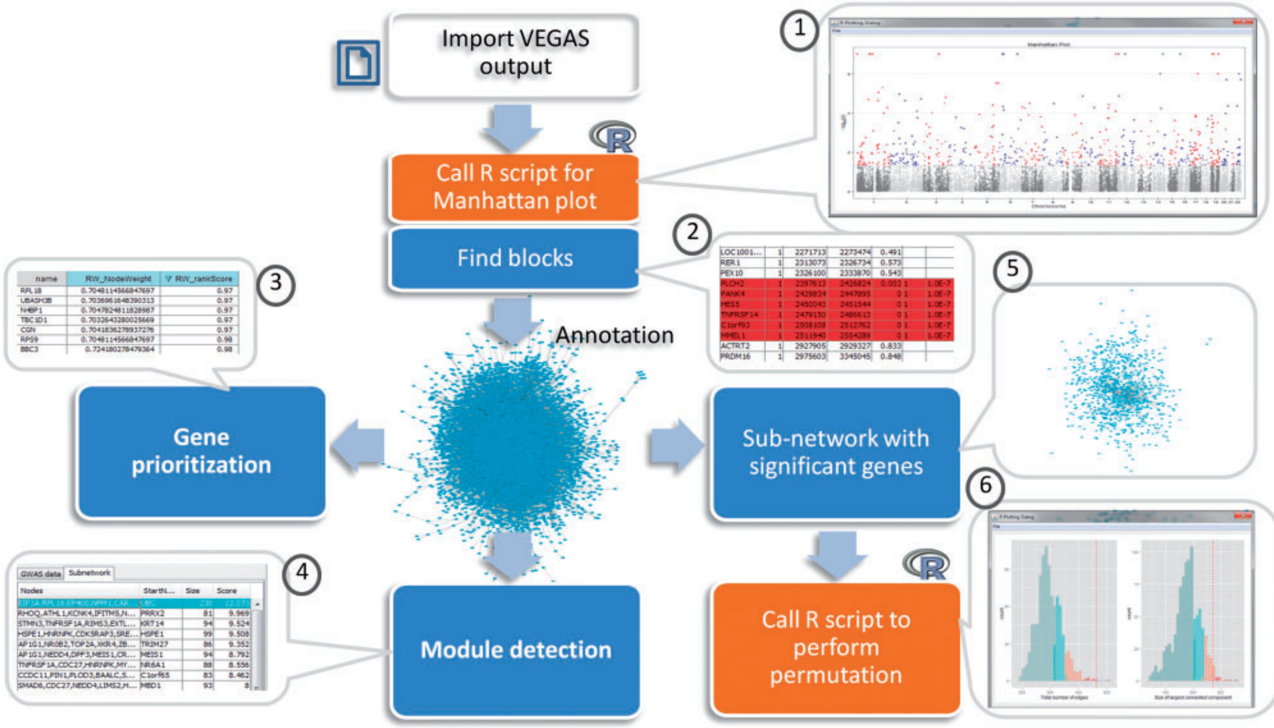

Fig. 1. Workflow of PINBPA. The blue blocks denote functions implemented in java, and the orange blocks denote functions calling R scripts

random walk with restart algorithm (Kohler *et al.*, 2008); and (vi) detects subnetworks enriched in significant genes using either unweighted (Ideker *et al.*, 2002) or weighted z-scores (Whitlock, 2005). The network z-score is adjusted using 1000 permutations and then a greedy algorithm (Supplementary File) is applied to search for the n optimal subnetworks.

3 DISCUSSION

PINBPA is the first Cytoscape app designed for network analysis of GWAS data. Through a user-friendly interface, the PINBPA app enables genomic researchers and biologists with no computational expertise to run a powerful and otherwise complex analytical pipeline. PINBPA is flexible in many ways. First, it enables the use of any network as scaffold for subsequent analysis. For example, a user may choose to run PINBPA using a smaller but high confidence PIN, or a larger network, including computationally predicted interactions. Additionally, other networks such as metabolic, gene co-expression or text mining can be used. Second, while direct import from VEGAS is available, any method can be used to convert SNP-level and gene-level *P*-values. Third, the user can select the significance threshold to define association blocks. Although some experimentation may be needed, this significance threshold is typically related to the power of the GWAS in question (the larger the GWAS, the more relaxed the *P*-value for block definition). Fourth, significant (first-order) networks can be generated at any user-selected statistical threshold. Fifth, an optional network smoothing step allows the user to weight the most significant associations differently. Finally, the size and the significance of enriched networks generated by the greedy algorithm can also be selected, thus providing maximal control over the final output.

In summary, the Cytoscape PINBPA app offers a simple interface to perform a complex set of analyses and can be used under a wide variety of scenarios, thus empowering genomic scientists to conduct post-GWAS analysis in a streamlined, rigorous and reproducible fashion.

Funding: S.E.B. is a Harry Weaver Neuroscience scholar of the National MS Society.

Conflict of interest: none declared.

REFERENCES

Akula,N. *et al.* (2011) A network-based approach to prioritize results from genome-wide association studies. *PLoS One*, **6**, E24220.

Baranzini,S.E. *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, **18**, 2078–2090.

Huang Da,W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (**Suppl. 1**), S233–S240.

International Multiple Sclerosis Genetics Consortium. (2013) Network-based multiple sclerosis pathway analysis with gwas data from 15,000 cases and 30,000 controls. *Am. J. Hum. Genet.*, **92**, 854–865.

Jia,P. *et al.* (2011) dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, **27**, 95–102.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Kohler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Lee,P.H. *et al*. (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics*, **28**, 1797–1799.

Liu,J.Z. *et al*. (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet*., **87**, 139–145.

Rossin,E.J. *et al*. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLos Genet*., **7**, e1001273.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

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

Welter,D. *et al*. (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., **42**, D1001–D1006.

Whitlock,M.C. (2005) Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *J. Evol. Biol*., **18**, 1368–1373.

Yaspan,B.L. *et al*. (2011) Genetic analysis of biological pathway data through genomic randomization. *Hum. Genet*., **129**, 563–571.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016