BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 18 2009, pages 2444–2446
doi:10.1093/bioinformatics/btp431

Genetics and population analysis

Path: a tool to facilitate pathway-based genetic association analysis

David Zamar, Ben Tripp, George Ellis and Denise Daley*

James Hogg iCAPTURE Center, University of British Columbia (UBC), Vancouver, BC, Canada V6Z1Y6

Received on March 9, 2009; revised on July 11, 2009; accepted on July 13, 2009

Advance Access publication July 23, 2009

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Traditional methods of genetic study design and analysis work well under the scenario that a handful of single nucleotide polymorphisms (SNPs) independently contribute to the risk of disease. For complex diseases, susceptibility may be determined not by a single SNP, but rather a complex interplay between SNPs. For large studies involving hundreds of thousands of SNPs, a brute force search of all possible combinations of SNPs associated with disease is not only inefficient, but also results in a multiple testing paradigm, whereby larger and larger sample sizes are needed to maintain statistical power. Pathway-based methods are an example of one of the many approaches in identifying a subset of SNPs to test for interaction. To help determine which SNP–SNP interactions to test, we developed Path, a software application designed to help researchers interface their data with biological information from several bioinformatics resources. To this end, our application brings together currently available information from nine online bioinformatics resources including the National Center for Biotechnology Information (NCBI), Online Mendelian Inheritance in Man (OMIM), Kyoto Encyclopedia of Genes and Genomes (KEGG), UCSC Genome Browser, Seattle SNPs, PharmGKB, Genetic Association Database, the Single Nucleotide Polymorphism database (dbSNP) and the Innate Immune Database (IIDB).

Availability: The software, example datasets and tutorials are freely available from http://genapha.icapture.ubc.ca/PathTutorial.

Contact: ddaley@mrl.ubc.ca

1 INTRODUCTION

The introduction of high-throughput single nucleotide polymorphism (SNP) genotyping methods has given rise to large-scale genome-wide association studies (GWAS) to identify common SNPs associated with complex traits. Until recently, the primary focus of most of these studies has been the discovery of single-SNP associations. However, single-SNP analyses are limited to identifying a subset of the most significant SNPs that account for only a small fraction of the total phenotypic variance. As the number of hypotheses that may be tested increases exponentially with the number of SNPs included in a study, biological information from the literature is commonly utilized in the development of hypotheses.

For these kinds of large studies, the simple task of storing, retrieving and visualizing results of an analysis has become surprisingly challenging. Although several software applications, such as PLINK (Purcell *et al.*, 2007), were designed to help analyze genetic association data and subsequently help to store and visualize results, none was designed to retrieve information from several bioinformatics resources and to conveniently integrate this knowledge with the results from a genetic association study.

We were, therefore, motivated to develop Path, a software application designed to help researchers interface their data with biological information from several bioinformatics resources. This information may be used to help generate biologically plausible hypotheses for testing gene–gene interactions. The Path software is a first-step bioinformatics approach to investigate gene–gene interactions in genetic association studies. Examples of the type of information retrieved and the bioinformatics resources accessed by Path are shown in Table 1.

2 FUNCTIONALITY

As input, Path requires a dataset in the LINKAGE pre-makeped format (Terwilliger and Ott, 1994) and a data file in QTDT format (Abecasis *et al.*, 2000). Additionally, one may supply a file with single-SNP association results. If association results are not supplied, the application initially performs a single-SNP analysis. Thereafter, a simple graphical user interface is used to explore the data along with the information retrieved from all nine bioinformatics resources and to conduct studies on the SNP–SNP interactions of the user's choice. Version 3.0.13 of the software application, UNPHASED (Dudbridge, 2003, 2006, 2008) is used for all analyses. The imported data and results of the analysis are stored in a local database.

Analogous to PLINK, our software also provides the means to analyze and store genetic association data and to visualize results with charts, plots and summary tables. A summary page that can be easily accessed and queried through the user interface is provided for each SNP. Entries for each SNP include basic background information, such as function, gene, chromosome, etc., and a summary of the results of single-SNP associations. Each SNP entry also provides several links to other data, such as KEGG pathway (Kanehisa *et al.*, 2006, 2008 and Kanehisa and Goto, 2000), and to previous association study results. To facilitate the selection of SNPs to test for gene-gene interactions, Path automates the SNP to gene annotation. This allows the user to easily visualize association results

*To whom correspondence should be addressed.

Table 1. Bioinformatics resources accessed by Path

Resource Name	URL	Description	Extracted Information
National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov	A resource for molecular biology information.	The SNP function and gene it belongs to.
Online Mendelian Inheritance in Man (OMIM)	http://www.nslij-genetics.org/search_omim.html	Archive of human genes and genetic phenotypes.	List of known patterns of disease inheritance and genes with prior substantial evidence for association with disease.
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg	A collection of manually drawn pathway maps representing current knowledge concerning several networks of molecular interactions and reactions.	Biological pathways and corresponding diagrams that each gene is involved in.
UCSC Genome Browser	http://genome.ucsc.edu	Archive of reference sequences and working draft assemblies for a large collection of genomes.	Genome browser page link for each gene.
Seattle SNPs	http://pga.gs.washington.edu	SNP variation discovery resource.	Links to the sequencing and genotyping information for each gene.
PharmGKB	http://www.pharmgkb.org	Collection of relationships among drugs, diseases and genes, including their genetic variations and gene products.	PharmGKB page link for each gene.
Genetic Association Database	http://geneticassociationdb.nih.gov	Archive of composite information about genetic linkage data and genetic association data from published reports.	Links to results from published association studies.
The Single Nucleotide Polymorphism database (dbSNP)	http://www.ncbi.nlm.nih.gov/projects/SNP	A public-domain archive for a broad collection of SNPs	dbSNP page link for each SNP.
The Innate Immune Database (IIDB)	http://db.systemsbiology.net/cgi-bin/GLUE/U54/IIDBHome.cgi	A repository of genomic annotations and experimental data for over 2000 genes associated with immune response behavior in the mouse genome.	IIDB page link for each gene.

for SNPs and genes in the context of KEGG pathways. Path will display the selected KEGG pathway and highlight the genes with genotypes in the selected pathway. Path includes visualization tools that interface with KEGG pathways and the users genetic association results, facilitating the exploration of genetic associations in the context of genetic pathways. Path guides users with simple point and click interfaces in the selection of SNPs to test for gene-gene interactions. In addition, the linkage disequilibrium (LD) plot of the gene containing the specified SNP is provided in the SNP summary page. The LD plots are generated by using the Haploview (Barrett *et al.*, 2005) software.

The majority of the bioinformatics information provided by our application is accessed through external links; therefore, connection to the Internet is required. These links are not automatically generated when a dataset is imported, because they may already exist and take time to create (depending on the speed of your Internet connection). Instead, we have incorporated an update option that may be periodically run by the user to maintain an up-to-date database of links to external resources.

Studies of gene–gene interactions are carried out by using a simple point-and-click interface. Results of analysis of single-SNP associations and of gene–gene interactions are calculated by using UNPHASED. After a gene–gene interaction has been submitted for testing, a link to the analysis results produced by the UNPHASED application is returned. Most GWAS typically include several hundred thousand SNPs, therefore, to help users single out SNPs that they want to include in a gene–gene interaction test, we have implemented a filtering option that enables the user to work with a subset of SNPs whose single-SNP association *P*-values fall below a chosen threshold. It took 124 s to import a dataset of 10 000 SNPs on 162 individuals using an Intel Core 2 Duo 2.4 GHz CPU.

3 FUTURE DIRECTIONS

We will extend our application to include information based on ontology and gene expression profiles. Due in part to the current partial identification and understanding of locus control regions, our software is limited in that it does not extract information on SNPs that may regulate a genetic pathway (i.e. promoters or locus control regions); thus our application, at present, does not account for such SNPs. To remedy this, we will include pathway information on SNPs that fall outside gene regions.

ACKNOWLEDGEMENTS

We would like to thank Scott Tebbutt for his useful comments and discussions and also to Thea VanRossum for her help in testing out Path.

Funding: AllerGen, a National Centre of Excellence Network (Canada); the Canadian Institutes of Health Research (CIHR) (grant number KTB88288).

Conflict of Interest: none declared.

REFERENCES

Abecasis,G.R. *et al*. (2000) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet*., **66**, 279–292.

Barrett,J.C. *et al*. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Dudbridge,F. (2003) Pedigree disequilibrium tests for multilocus haplotypes. *Genet. Epidemiol*., **25**, 115–121.

Dudbridge,F. (2006) UNPHASED user guide. *Technical Report 2006/5,* MRC Biostatistics Unit, Cambridge.

Dudbridge,F. (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered*., **66**, 87–98.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic. Acids Res*., **36**, D480–D484.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Terwilliger,J.D. and Ott,J. (1994) *Handbook of Human Genetic Linkage*. John Hopkins University Press, Baltimore, p. 15.