BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 20 2009, pages 2762–2763
doi:10.1093/bioinformatics/btp448

Systems biology

The SNP ratio test: pathway analysis of genome-wide association datasets

Colm O'Dushlaine*, Elaine Kenny, Elizabeth A. Heron, Ricardo Segurado, Michael Gill, Derek W. Morris and Aiden Corvin

Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Ireland

Received on March 24, 2009; revised on June 9, 2009; accepted on July 16, 2009

Advance Access publication July 20, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: We present a tool that assesses the enrichment of significant associations from genome-wide association studies (GWAS) in a pathway context. The SNP ratio test (SRT) compares the proportion of significant to all SNPs within genes that are part of a pathway and computes an empirical *P*-value based on comparisons to ratios in datasets where the assignment of case/control status has been randomized. We applied the SRT to a Parkinson's disease GWAS dataset, using the KEGG database, revealing significance for Parkinson's disease and related pathways.

Availability: https://sourceforge.net/projects/snpratiotest/

Contact: codushlaine@gmail.com; colm.odushlaine@tcd.ie

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

1 INTRODUCTION

Analysis of GWAS data has expanded our understanding of complex diseases, but typically only a small fraction of genetic variance is explained by even large studies and many of the findings map to non-genic regions. This may reflect the underlying genetic models including for example, locus heterogeneity, small effects or epistasis. Pathway analysis may be robust to these effects and increase power by summarizing combined effects of all SNPs within a pathway in an attempt to make biologically meaningful interpretations of the data (Askland *et al*., 2009; Dinu *et al*., 2007; Lesnick *et al*., 2007; Wang *et al*., 2007). This approach also provides additional information relating to function over and above single SNP associations which may be helpful in interpreting the data. Pathway-based analyses of genomic data are potentially powerful, if as has been suggested, the joint action of variants of small effect clustering within biological pathways plays a major role in predisposing to complex genetic disorders (Lesnick *et al*., 2007). Even very large genome-wide association studies (GWAS) may lack power to identify small SNP effects, but these may be detectable at a pathway level. An example may be autism where pathway analysis of GWAS data has implicated molecular mechanisms involved in neuronal cell adhesion extending beyond the two cadherin genes implicated by the SNP analysis (Wang *et al*., 2009). The SNP ratio test (SRT) uses both significant and non-significant SNPs within a pathway to construct a ratio and compares this ratio to a distribution of ratios based on GWAS results using randomized phenotypes. The SRT is similar to methods such as gene set enrichment analysis (GSEA) (Subramanian *et al*., 2005) and the set-based test offered in PLINK (Purcell *et al*., 2007) in that it tests for enrichment of statistically associated SNPs in a pathway, also using empirical *P*-values. As the SRT uses all SNPs in the pathways, it can account for situations in which extensive LD, stretching beyond the gene/pathway of interest, generates false positives for that pathway. Single SNP association does not allow for the influence that differences in linkage disequilibrium (LD) (e.g. between studies or SNP arrays) may have on the identification of truly associated variants. Thus, the magnitude of any one-association statistic is not key, but rather the number of significant SNPs above what would be expected by chance is key, making the SRT more robust to false-positives at a SNP level. In addition, application of the SRT is at a pathway level rather than at a gene level, precluding the need to adjust for factors such as pathway/gene size (Wang *et al*., 2007). The SRT is also extremely easy to implement, working with PLINK inputs and outputs.

2 METHODS AND DATASETS

For a GWAS dataset, all SNPs are individually tested in the standard fashion for association with phenotype/disease (e.g. trend test), resulting in a list of significant and non-significant SNPs, (where significant is defined as the *P*-value being below or equal to a specified threshold, giving a total of *M* significant SNPs). A subset of these SNPs annotated as arising within genes within pathways ($_{PG}$SNPs) are then analyzed. KEGG (Kanehisa and Goto, 2000) ($N = 212$ pathways, Release 48.0, October 2008) was used here to define the pathways, but in principle any pathway dataset may be used. Alternatively, custom pathways may be specified to test specific hypotheses. SNP data were obtained from dbSNP (b129_SNPContigLocusId_36_3.bcp table) and genes annotated in this file were merged with KEGG genes to create a file linking KEGG and SNP information. For a given pathway, *W*, the ratio is then defined as: $r_w = \#$ significant SNPs in $W / \#$ SNPs in W.

For a given GWAS dataset and, in this case, KEGG pathway, the SRT uses simulated datasets to estimate the significance of a given pathway. The SRT accepts files in PLINK binary format and allows the user to prepare randomized phenotype datasets. The simulated datasets are constructed from the original dataset, preserving the original case/control ratio but randomizing the assignment of case/control status among individuals. The same individuals are used, maintaining the same LD structure. This minimizes spurious findings arising from LD because, even if LD were leading to an excess of significance for a pathway (e.g. 1 truly unassociated SNP in an LD block may give rise to a significant *P*-value by chance, leading to spuriously significant *P*-values across an LD block) this LD block would

*To whom correspondence should be addressed.

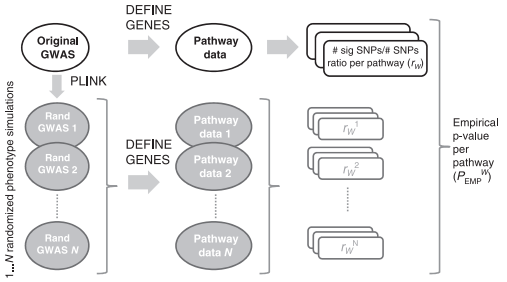

Fig. 1. Illustration of the SRT.

be identical across all datasets using randomized phenotypes. A total of N such datasets are simulated. In each of the simulated datasets, for each pathway, the ratio in Equation (1) is computed ($r_W^1 \ldots r_W^N$), where in each simulated dataset the lowest M P-values are defined as significant. Use of the M most significant SNPs rather than re-applying a P-value threshold in simulations should prevent any inflation in empirically significant pathways due to an excess of false positive SNPs in the original GWAS (due to e.g. genotyping error, or other bias). The empirical P-value for a particular pathway, $P_{\mathrm{EMP}}^W = (s+1)/(N+1)$, where s is the number of simulated datasets (in this case randomized phenotype simulations) that produce a ratio greater than or equal to the original ratio (North *et al*., 2002) (Fig. 1; Supplementary Fig. 2).

Note that the SRT does not correct for multiple testing at a pathway level. Multiple-testing correction of the pathway-level P-values is still required, although this is non-trivial due to the lack of independence between pathways. However, the multiplicity problem is greatly reduced relative to a SNP-level analysis. We applied the SRT to a Parkinson's disease GWAS. The CIDR dataset [CIDR: Genome Wide Association Study in Familial Parkinson Disease (phs000126.v1.p1) (13 May 2008)] consisted of a total of 344 301 SNPs, genotyped in 900 cases and 867 controls.

3 RESULTS AND DISCUSSION

We conducted standard association analysis in PLINK (Purcell *et al*., 2007) for both the original and 1000 randomized phenotype datasets. The association tests for the original dataset resulted in 17 773 nominally significant SNPs (unadjusted $P \leq 0.05$) with a genomic inflation factor of 1.03. A quantile–quantile plot is shown in Supplementary Fig. 1.

We applied the SRT to investigate associations with Parkinson's disease for 212 KEGG pathways in the CIDR dataset. Looking at pathways that rank highly for the SRT, there is strong evidence supporting the roles of these pathways in the etiology of Parkinson's disease; for example 'Parkinson's disease' (hsa05020) (Supplementary Fig. 2), 'Neurodegenerative Disorders' (hsa01510), 'Neuroactive ligand–receptor interaction' (hsa04080). Using more stringent thresholds ($P \leq 0.01$ and $P \leq 0.005$), hsa01510 remains significant suggesting that this pathway may be enriched for variants of small and larger effect (Supplementary Table 1).

As with other methods (Askland *et al*., 2009; Lesnick *et al*., 2007; Wang *et al*., 2007), the SRT is less equipped to identify significant pathways when there is a paucity of $_{\mathrm{PG}}$SNPs; if there are no significant $_{\mathrm{PG}}$SNPs, the ratio is always 0 for the original data. Thus, all simulations will at least equal this ratio, resulting in an empirical P-value of 1. These limits reflect both the limits of pathway annotation and the power of the GWAS dataset used. This pattern clearly depends on the pathway and GWAS datasets used but we note that, in the Parkinson's dataset, only pathways with at least 20 $_{\mathrm{PG}}$SNPs are observed to be significant (Supplementary Fig. 3). Thus, adequate SNP coverage of a pathway is essential for that pathway to be effectively tested.

Note the P-value chosen can reflect the disease model; choosing a stringent P-value (e.g. <0.001) to define 'associated' SNPs, tests a hypothesis that only highly associated SNPs are enriched in a pathway. If a pathway becomes significant using a more stringent cut-off, it reflects a role for a smaller number of variants of larger effect in the disease. Alternatively, if a pathway becomes significant using a less stringent cut-off, it may reflect a role for more variants of smaller effect. If a pathway is significant under a range of thresholds, it may suggest that a number of different genetic models underlie the disease, as is the case for many complex diseases. We believe that a less stringent cut-off is reasonable where the user lacks information regarding the underlying disease model.

While designed primarily to work with PLINK, the SRT may easily be used with datasets derived from other applications. The primary advantages of the SRT are first that it avoids issues arising from LD because it uses the same LD structure—by using the same SNPs—in all simulations, and thus only pathways with additional significant SNPs, not merely arising from LD, are reported as significant. Second that it uses individual level data in its simulations, maximizing the information available in testing pathway hypotheses. The SRT can be used to test a wide variety of pathway-based hypotheses, in addition to specific user-defined ones, in existing GWAS datasets and in datasets emerging from next-generation sequencing initiatives.

ACKNOWLEDGEMENTS

The authors thank Peter Holmans and Richard Anney for comments.

Funding: Health Research Board of Ireland (HRB), Irish Research Council for Science, Engineering and Technology (IRCSET) and Science Foundation Ireland (SFI).

Conflict of Interest: none declared.

REFERENCES

Askland,K., *et al.* (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet.*, **125**, 63–79.

Dinu,I. *et al*. (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.

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

Lesnick,T.G. *et al*. (2007) A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet.*, **3**, e98.

North,B.V. *et al*. (2002) A note on the calculation of empirical P values from Monte Carlo procedures. *Am. J. Hum. Genet.*, **71**, 439–441.

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.

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.

Wang,K. *et al*. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, **459**. [Epub ahead of print, doi:10.1038/nature07999, May 28, 2009]