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

Colm O’Dushlaine\textsuperscript{1,*}, Elaine Kenny\textsuperscript{1}, Elizabeth A. Heron\textsuperscript{1}, Ricardo Segurado\textsuperscript{1}, Michael Gill\textsuperscript{1}, Derek W. Morris\textsuperscript{1}, Aiden Corvin\textsuperscript{1}

\textsuperscript{1}Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Ireland

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/snpratitest/

Contact: codushlaine@gmail.com

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 for Parkinson’s disease and related pathways.

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

Contact: codushlaine@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

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 (M\textsubscript{gsNP}s) 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 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 \((r_W^{original}/r_W^{GWAS})\), 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, \(P_{EMPIRICAL} = (s+1) / (N+1)\), where \(s\) is the number of simulated datasets (in this case randomised phenotype simulations) that produce a ratio greater than or equal to the original ratio (North, et al., 2002) (Figure 1) (Supplementary Figure 2).

Figure 1: illustration of the SRT

Note that the SRT does not correct for multiple testing at a pathway level. Multiplex 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) (May 13, 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 1,000 randomized phenotype datasets. The association tests for the original dataset resulted in 17,773 nominally significant SNPs (unadjusted \(p<0.05\)) with a genomic inflation factor of 1.03. A quantile-quantile plot is shown in Supplemental Figure 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 Figure 2), “Neurodegenerative Disorders” (hsa01510), “Neuroactive ligand-receptor interaction” (hsa04080). Using more stringent thresholds (\(p<0.01\), \(p<0.005\)), hsa01510 (hsa01510), “Neuroactive ligand-receptor interaction” (hsa04080).

acknowledgements

Funding: Health Research Board of Ireland (HRB), Irish Research Council for Science, Engineering & Technology (IRCSET) and Science Foundation Ireland (SFI). We thank Peter Holmans and Richard Anney for comments.

REFERENCES


