BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 13 2012, pages 1797–1799
doi:10.1093/bioinformatics/bts191

Genetics and population analysis

Advance Access publication April 17, 2012

INRICH: interval-based enrichment analysis for genome-wide association studies

Phil H. Lee[1,2,3], Colm O'Dushlaine[3], Brett Thomas[1] and Shaun M. Purcell[1,2,3,4,*]

[1]Analytic and Translational Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, MA 02114, [2]Department of Psychiatry, Harvard Medical School, Boston, MA 02115, [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142 and [4]Center for Statistical Genetics, Mount Sinai School of Medicine, New York, NY 10029, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Here we present INRICH (INterval enRICHment analysis), a pathway-based genome-wide association analysis tool that tests for enriched association signals of predefined gene-sets across independent genomic intervals. INRICH has wide applicability, fast running time and, most importantly, robustness to potential genomic biases and confounding factors. Such factors, including varying gene size and single-nucleotide polymorphism density, linkage disequilibrium within and between genes and overlapping genes with similar annotations, are often not accounted for by existing gene-set enrichment methods. By using a genomic permutation procedure, we generate experiment-wide empirical significance values, corrected for the total number of sets tested, implicitly taking overlap of sets into account. By simulation we confirm a properly controlled type I error rate and reasonable power of INRICH under diverse parameter settings. As a proof of principle, we describe the application of INRICH on the NHGRI GWAS catalog.

Availability: A standalone C++ program, user manual and datasets can be freely downloaded from: http://atgu.mgh.harvard.edu/inrich/.

Contact: shaun@atgu.mgh.harvard.edu

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

Received on October 26, 2011; revised on March 28, 2012; accepted on April 12, 2012

1 INTRODUCTION

Multi-locus approaches (often known as pathway or gene-set enrichment analysis methods) can be used to ask whether sets of single-nucleotide polymorphisms (SNPs), often defined by groups of functionally related genes, are in aggregate more highly associated with a phenotype than expected by chance. Consideration of the biological relationships among the 'top hits' in a genome-wide association study (GWAS) can provide orthogonal evidence, over and above the functionally agnostic analysis of the number, statistical significance and/or variance explained of those hits. For example, that a GWAS has three independent SNPs with *P*-values around at 1×10^{-6} is in itself unremarkable. However, if the associated regions independently map to three of a small set of functionally related genes, this will be very unlikely to occur by chance: consequently, we would likely wish to put more weight on these associations. As well as providing additional statistical evidence to sub-threshold association results, another use of gene-set analysis can be called *in silico* fine-mapping, or prioritizing specific genes in loci that contain multiple genes with equivalent association evidence. For example, of 10 associated genes within a block of strong linkage disequilibrium (LD), we may find that only one shows above-chance relatedness to genes that appear in other, statistically independent association intervals. All other things being equal, one would presumably consider that gene as more likely to be causally related compared with the other nine. Furthermore, the identity of the particular enriched gene-sets may offer insights into disease mechanism and biology, although this will be contingent on the gene-sets' accuracy, comprehensiveness and relevance to the phenotype's underlying biology.

Over the past few years, several gene-set methods for GWAS have been developed ((Holmans *et al.*, 2009; Wang *et al.*, 2007)). Still, there clearly exist challenges and limitations to be addressed ((Hong *et al.*, 2009)). Desirable properties of a gene-set test include the following: (i) robust, and so able to calculate experiment-wide significance, with adjustment for common biases due to gene size, LD within and between genes, etc.); (ii) flexible, with application to (summary) data from different sources, such as GWAS, from imputed data, copy number variant (CNV) studies, targeted sequencing, from tables in manuscripts, etc.; and (iii) computationally manageable, allowing genome-wide analysis in a reasonable time on a single machine.

Here, we describe the gene-set enrichment analysis tool INRICH (INterval enRICHment analysis) that aims to satisfy the above properties. INRICH takes a set of independent, nominally associated genomic *intervals* and then tests for the enrichment of predefined gene-sets. An 'interval' will typically correspond to a genomic region of SNP association defined by LD from a genome-wide scan, although intervals could also represent, e.g. deletion or duplication events observed in cases, regions identified as homozygous-by-descent, etc.

2 METHODS

We describe the method implemented in INRICH, focussing on the case of SNP association from GWAS data. Specifically, analysis follows the following three steps:

2.1 Interval data generation

INRICH takes disease-associated genomic intervals as input—for example, all GWAS SNPs (and the other, local SNPs in LD) that are associated with

*To whom correspondence should be addressed.

a phenotype at $P < 1 \times 10^{-4}$. Either PLINK ((Purcell *et al.*, 2007)) LD-clumping or tag SNP selection commands (or similar tools) can be used to define such independent regions of association, which ensures that multiple, adjacent SNPs that potentially tag the same causal variant are analyzed as one independent association unit. Due to space limitation, we provide a detailed instruction manual on the data generation and testing procedure at our website (http://atgu.mgh.harvard.edu/inrich/).

2.2 Overlapping interval/gene merging

It is not uncommon for functionally related genes to show physical clustering, and therefore yield an inflated false positive rates for such gene-sets if dependent signals are assumed to be independent ((Holmans *et al.*, 2009; Hong *et al.*, 2009)). To avoid this potential bias due to multi-counting physically clustered genes belonging to the same set, we merge overlapping genes belonging to the same gene-set. We also merge overlapping testing intervals to ensure that testing units are statistically independent from each other.

2.3 Set-based enrichment tests

The primary enrichment statistic E for each gene-set is the number of *intervals* that overlap at least one 'target' gene (i.e. gene in the tested set), which we refer to as the *interval* mode. An alternative test instead counts the number of target *genes* that overlap at least one interval, which is useful for analyzing structural variation data (e.g. CNVs) that typically span large genomic regions and therefore are likely to disrupt multiple, non-overlapping genes. We call this test setting as the *target* mode. We use a permutation approach, described below, to calculate empirical significance P-values for each gene-set.

Suppose that input data I includes k intervals, $I = \{i_1, ..., i_k\}$, and target gene-set T includes m genes, $T = \{t_1, ..., t_m\}$.

(1) Null interval set R is generated by randomly assigning intervals to genomic locations with the constraints that each null interval $r_i \in R$ approximately matches to the original interval $I_i \in I\,(i = 1, ..., k)$ in terms of the number of SNPs and overlapping genes; we also ensure approximately similar SNP density per kilobase. Supplementary Figure S1 illustrates the three matching criteria.

(2) Corresponding to the selected testing mode as described above, the null enrichment statistic E is calculated as the number of overlapping intervals (or genes) between target gene-set T and randomly matched null set R.

(3) Steps (1) and (2) are repeated N times to generate a distribution of the enrichment statistics for target gene-set T under the null hypothesis.

(4) The empirical P-value for T is the proportion of N replicates where the enrichment statistic is as large as that of original interval set I.

(5) Multiple testing correction is achieved via a second, nested round of permutation to assess the null distribution of the minimum empirical P-value across all tested gene-sets.

This permutation procedure, therefore, respects the relationship between gene size and the probability of chance overlap, namely that large genes are more likely to be hit by chance. As previously reported, large genes are not representative of all genes in terms of function ((Raychaudhuri *et al.*, 2010)).

INRICH also presents global enrichment statistics G_p that test for an excess of enriched genes at nominal gene-set $P = 0.001$, 0.01 and 0.05. This test is based on the number of unique genes within an association interval that are in at least one nominally enriched gene-set. The empirical significance of G_P is evaluated within the same permutation procedure as described above.

3 DATA ANALYSIS AND SUMMARY

We first conducted a simulation study to assess the Type I error rates of INRICH using two GWAS datasets: HapMap III (CEU + TSI; $n = 200$), and schizophrenia case/control study ($n = 1468$; (Lieberman *et al.*, 2005). Tested parameter settings include different enrichment statistics (i.e. 'interval' or 'target' mode), LD-clumping r^2 measures ($r^2 = 0.2$), as well as significant P-value thresholds to define associated regions (1×10^{-3} and 5×10^{-3}). Under each setting, we repeated the following procedures 200 times, and calculated the average type I error rate: (i) Generate random phenotype labels for subjects; (ii) Apply standard χ^2 association analysis on individual SNPs; and (iii) Run INRICH on the association results using the KEGG gene-sets ((Kanehisa *et al.*, 2010)). We also conducted the same simulation study using two commonly used gene-set enrichment approaches: GenGen ((Wang *et al.*, 2007); i.e. GSEA tool specifically designed for GWAS) and the hypergeometric test. Compared with these methods, the average type I error rates of INRICH did not exceed the nominal 5% level. In contrast, under some conditions, the hypergeometric test yielded a type I error rate as high as 100%. We also considered the power under conditions where the hypergeometric test is valid, and confirmed that INRICH gives a comparably good power to the hypergeometric test (see Supplementary Table S3 for details). Phenotype-permutation-based gene-set enrichment methods (such as GenGen) provide statistically rigorous tests, but are computationally very demanding (particularly if based on imputed datasets, or complex family-based association tests, etc.). In contrast, other gene-set enrichment methods based on summary data alone (such as the hypergeometric test) are not computationally intensive, but can be very anti-conservative, as our simulations show, due to unwarranted assumptions of independence. We argue that INRICH is well-placed between these two poles, providing an efficient yet robust middle-ground.

As a proof of concept to demonstrate the performance of INRICH under the alternative hypothesis, we applied INRICH to the summary association data from the NHGRI (National Human Genome Research Institute) GWAS catalog ((Hindorff *et al.*, 2009)). First, we downloaded a list of 4689 SNPs that are associated with 411 complex diseases/traits at a P-value $< 1 \times 10^{-5}$ (download date: March 4, 2011). This analysis focused on 236 diseases/traits that have at least five-associated SNPs. For each phenotype, LD-independent intervals were generated around the associated SNPs using PLINK, and enrichment test was conducted using 3182 Gene Ontology (GO) terms (gene-set size between 5 and 200 genes; (The Gene Ontology Consortium, 2000)) and 10^6 replicates in the first round of permutation and 10^4 in the second. We excluded all genes and intervals mapping to the broad MHC region (chr6: 25–35 Mb): in practice because this region contains so many genes, it is unlikely to improve the power of gene-set enrichment analysis in most cases. After multiple testing correction, 47 disorders were predicted with at least one significantly enriched GO term at $\alpha = 5\%$. Many of the associations were consistent with known pathology of examined complex diseases/traits. For example, Type II diabetes-associated intervals were most significantly enriched for genes involved in glucose homeostasis (corrected $P = 0.001$) and Crohn's disease-associated intervals enriched for regulation of activated T cell proliferation (corrected $P = 0.003$).

In summary, we have implemented a new gene-set enrichment method in the INRICH package, based on a constrained reshuffling of associated intervals, to test whether more genes from particular sets are contained in those intervals than expected by chance. Importantly, we preserve the properties of the original data while reshuffling, in terms of the number, SNP density and gene-density. We have shown appropriate type I error rates, even when correcting for hundreds of partially overlapping gene-sets. Preliminary application to the NHGRI GWAS catalogue indicates good power to detect true signals. INRICH was recently applied to a large GWAS of bipolar disorder, implicating calcium ion channel genes as enriched ((Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011)). Practically, INRICH is fast, applicable without individual genotype data and freely available either as a command-line tool or with a GUI.

ACKNOWLEDGEMENT

The authors thank Dr Peter Holmans for insightful comments.

Funding: We gratefully acknowledge support from NIMH/NIH grant U01MH0855513.

Conflict of Interest: none declared.

REFERENCES

Hindorff,L. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Holmans,P. *et al.* (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.*, **85**, 13–24.

Hong,M.G. *et al.* (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.*, **126**, 289–301.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Lieberman,J. *et al.* (2005) Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N. Engl. J. Med.*, **353**, 1209–1223.

Psychiatric GWAS Consortium Bipolar Disorder Working Group. (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.*, **43**, 977—983.

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.

Raychaudhuri,S. *et al.* (2010) Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet.*, **6**, e1001097.

The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

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