BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2496–2497
doi:10.1093/bioinformatics/btu326

Genome analysis

Advance Access publication May 10, 2014

SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci

Kamil Slowikowski[1,2], Xinli Hu[2] and Soumya Raychaudhuri[3,4,*]

[1]Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA 02138, USA, [2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston MA 02215, [3]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA and [4]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: We created a fast, robust and general C++ implementation of a single-nucleotide polymorphism (SNP) set enrichment algorithm to identify cell types, tissues and pathways affected by risk loci. It tests trait-associated genomic loci for enrichment of specificity to conditions (cell types, tissues and pathways). We use a non-parametric statistical approach to compute empirical *P*-values by comparison with null SNP sets. As a proof of concept, we present novel applications of our method to four sets of genome-wide significant SNPs associated with red blood cell count, multiple sclerosis, celiac disease and HDL cholesterol.

Availability and implementation: http://broadinstitute.org/mpg/snpsea

Contact: soumya@broadinstitute.org

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

Received on December 12, 2013; revised on April 3, 2014; accepted on May 5, 2014

1 INTRODUCTION

As genome-wide association studies (GWAS) continue to find disease alleles, investigators seek to identify the set of pathways and tissue types affected by these alleles, and the physiological conditions under which they act (Elbers *et al*., 2009; Lango Allen *et al*., 2010; Raychaudhuri, 2011; Wang *et al*., 2013; Yaspan and Veatch, 2011). For example, we have previously presented statistical methods to identify immune cell types for further functional investigation by finding cell type-specific expression of genes in linkage disequilibrium (LD) with autoimmune disease-associated single-nucleotide polymorphisms (SNPs) (Hu *et al*., 2011). Presumably, alleles influence disease risk through pathways specific to these cell types.

We sought a general implementation of these methods to leverage data from high-throughput functional assays that assess genome-wide transcription, protein binding, epigenetic modifications and other functional parameters across diverse cellular conditions and tissue types. Each of these diverse data types can be represented as a continuous matrix of genes and *conditions* (e.g. cell types, tissues, pathways, experimental conditions). Databases such as Gene Ontology (GO) (Botstein *et al*., 2000) offer expert-defined pathways and complementary gene annotations that can be represented as binary values.

Investigators have already described strategies to assess enrichment of GWA results for pathways or gene sets but not for condition specificity (Holden *et al*., 2008; Weng *et al*., 2011). In contrast to these methods, we do not require genotypes, *P*-values, *a priori* gene sets or pathways or *a priori* definitions of gene–SNP associations. We require only a list of SNP identifiers, use LD structures to identify plausibly influential genes and use a simple sampling approach to identify the conditions they influence.

SNPsea is a general algorithm to identify the conditions relevant to a trait by assessing the genes within associated loci for enrichment of condition specificity.

2 METHODS

For a given set of SNPs, SNPsea tests genes implicated by LD, in aggregate, for enrichment of specificity to a condition in a given matrix of genes and conditions. The matrix must be normalized so that conditions are comparable.

First, we identify genes implicated by each SNP using LD from reference genomes. Second, we calculate a specificity score for each condition with these genes. Finally, we compare these scores with scores obtained with null sets of matched SNP sets to calculate an empirical *P*-value for each condition (see Supplementary Notes for algorithm details).

We empirically calculate *P*-values because we previously found that analytical distributions can result in inaccurate *P*-values (Hu *et al*., 2011). SNP linkage intervals, gene densities, gene sizes and gene functions are correlated across the genome and are challenging to model analytically.

We used C++ for fast computation of *P*-values because Python was prohibitively slow. The online reference manual details compilation and installation procedures; we also provide executable files for immediate use on select platforms.

2.1 Multiple genes implicated by LD

Accurate analyses must address the critical issue that SNPs from GWA studies frequently implicate more than one gene (50% of GWAS Catalog SNPs, Supplementary Fig. S2).

We defined LD intervals with SNPs from the 1000 Genomes Project (EUR) (Genomes Project Consortium, 2010) and a previously described strategy (Supplementary Fig. S1) (Rossin *et al*., 2011). A SNP implicates genes overlapping its LD interval, defined by the furthest SNPs in a 1 Mb window with $r^2 > 0.5$. To ensure the associated genes are included, we extend each interval to the nearest recombination hotspots with

*To whom correspondence should be addressed.

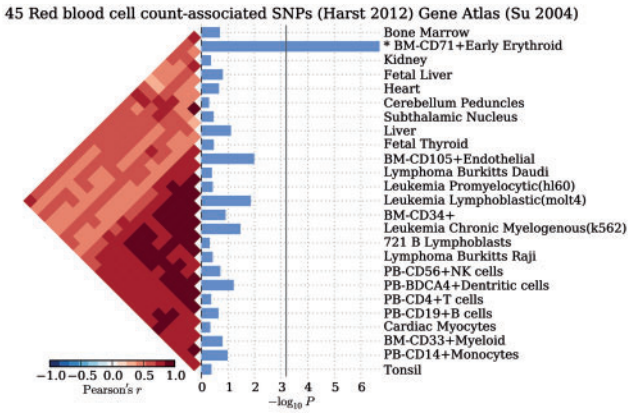

Fig. 1. Empirical *P*-values for specificity to each condition. 25 of 79 tissues (Gene Atlas) are shown. Adjacent: Pearson correlation coefficients for pairs of expression profiles ordered by hierarchical clustering with UPGMA

recombination rate >3 cM/Mb (HapMap3) (Myers *et al.*, 2005). We merge SNPs with shared genes into a single locus.

By default, we assume that each associated locus harbours a single influential gene rather than multiple genes. We provide an alternative scoring method to account for multiple genes (Supplementary Notes) that produces similar results in four traits we tested (Supplementary Fig. S4).

Because interval lengths depend on the choice of r^2 threshold, we looked for an effect of this choice (Supplementary Fig. S3). The significant result for the Gene Atlas and blood cell count SNPs is robust to different thresholds. Similarly, the choice of r^2 threshold has little effect on the significant GO enrichment result for these SNPs.

2.2 Type I error estimates

We tested 10 000 sets of 100 randomly selected LD-pruned SNPs. For each condition (tissue or GO term), we observed appropriate proportions of *P*-values <0.5, 0.1, 0.05, 0.01 and 0.005 (Supplementary Fig. S5).

3 EXAMPLES

We used SNPsea to identify tissues relevant to blood cell count by testing 45 genome-wide significant SNPs (van der Harst *et al.*, 2012) with expression data (Gene Atlas) for 17 581 genes across 79 human tissues (Su *et al.*, 2004). Bone marrow CD71+ early erythroid cells are significantly enriched for cell type-specific expression of the genes within the trait-associated loci ($P = 2 \times 10^{-7}$) (Fig. 1).

The genes in these loci are enriched for the term *hemopoiesis* (GO:0030097) ($P = 2 \times 10^{-5}$) (Supplementary Fig. S6), suggesting that blood cell count may be influenced by the genes expressed specifically in early erythroid cells and involved in forming blood cellular components.

We provide additional examples for SNPs associated with multiple sclerosis, celiac disease and HDL cholesterol. Each includes Gene Atlas and GO enrichments, r^2 comparisons and comparisons of results assuming a single or multiple causal genes (Supplementary Figs S7–9).

Funding: The National Institutes of Health (5K08AR055688, 1R01AR062886-01, 1U01HG0070033, T32 HG002295/HG/NHGRI, 7T32HG002295-10), the Arthritis Foundation and the Doris Duke Foundation.

Conflict of Interest: none declared.

REFERENCES

Botstein,D. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Elbers,C.C. *et al.* (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet. Epidemiol.*, **33**, 419–431.

Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Holden,M. *et al.* (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, **24**, 2784–2785.

Hu,X. *et al.* (2011) Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.*, **89**, 496–506.

Lango Allen,H. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Myers,S. *et al.* (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **301**, 321–324.

Raychaudhuri,S. (2011) Mapping rare and common causal alleles for complex human diseases. *Cell*, **147**, 57–69.

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.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci.*, **16**, 6062–6067.

van der Harst,P. *et al.* (2012) Seventy-five genetic loci influencing the human red blood cell. *Nature*, **492**, 369–375.

Wang,X. *et al.* (2013) Association of polymorphisms in the Chr18q11.2 locus with tuberculosis in Chinese population. *Hum. Genet.*, **132**, 691–695.

Weng,L. *et al.* (2011) SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, **12**, 99.

Yaspan,B.L. and Veatch,O.J. (2011) Strategies for pathway analysis from GWAS data. *Curr. Protoc. Hum. Genet.*, **Chapter 1**, Unit1.20.