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

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1093/bioinformatics/btu326</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:12987235">http://nrs.harvard.edu/urn-3:HUL.InstRepos:12987235</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci

Kamil Slowikowski¹,², Xinli Hu² and Soumya Raychaudhuri³,⁴,*

¹Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA 02138, USA, ²Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston MA 02215, ³Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA 02115, USA and ⁴Program 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/snps
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

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 r² > 0.5. To ensure the associated genes are included, we extend each interval to the nearest recombination hotspots with...
recombination rate \( >3 \text{cM/Mb} \) (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 this choice (Supplementary Fig. S4).

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 (\( P = 2 \times 10^{-5} \)) (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


