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

Kamil Slowikowski, Xinli Hu, Soumya Raychaudhuri

April 2, 2014

Contents

Supplementary Note **2**
Algorithm Details . . . 2
Data . . . 5
Commands . . . 6

Supplementary Figures **7**
Supplementary Figure 1: Determining SNP linkage intervals . . . 7
Supplementary Figure 2: Counting genes in GWAS SNP linkage intervals . . . 8
Supplementary Figure 3: Choosing the r^2 threshold for linkage intervals . . . 9
Supplementary Figure 4: Comparison of 'single' and 'total' options . . . 10
Supplementary Figure 5: Type 1 error estimates . . . 11

Additional Examples **12**
Supplementary Figure 6: Red blood cell count GO enrichment . . . 13
Supplementary Figure 7: Multiple sclerosis . . . 14
Supplementary Figure 8: Celiac disease . . . 15
Supplementary Figure 9: HDL cholesterol . . . 16

References **17**

Supplementary Note

Algorithm Details

SNPsea tests if genes implicated by risk loci (e.g., those discovered through genome-wide association (GWA) studies) are specifically expressed in some conditions over others, and if this specificity is statistically significant. The program requires two inputs:

1. A list of SNP identifiers: rs123, 12:456, ...
2. A matrix of genes and conditions, such as:
   - expression profiles of different cell types.
   - Ontology terms and presence/absence 1/0 values for each gene in each term.

For example, SNPsea can be used to find tissues or cell types whose function is likely to be influenced by genes in risk loci. If the genes in risk loci are used in relatively few cell types, we hypothesize that they are likely to affect those cell types' unique functions. This assumes that expression specificity is a good indicator of a gene's importance to the unique function of the cell type.

For a given set of SNPs associated to some phenotype, SNPsea tests whether all implicated genes, in aggregate, are enriched for specificity to a condition in a user-provided matrix of genes and conditions/annotations. The algorithm consists of three steps:

- **Step 1: Assigning genes to each SNP**
  - We use linkage disequilibrium (LD) to identify the genes implicated by each SNP.
- **Step 2: Calculating specificity scores**
  - We look up implicated genes in a user-provided matrix and calculate a specificity score for each annotation/condition based on the values of these genes.
- **Step 3: Testing significance**
  - We compare the specificity scores to a null distribution of scores obtained with random sets of matched SNP sets and compute an empirical P-value.

Step 1: Assigning genes to each SNP

Accurate analyses must address the critical issue that SNPs frequently implicate a region with multiple different genes (**Supplementary Figure 2**). The challenge is to find evidence to show which of those genes are associated with a given trait.

We determine the genes plausibly implicated by each trait-associated SNP using a previously described strategy (**Supplementary Figure 1** and [1]). First, we define the linkage interval for a given SNP as the span between the furthest correlated SNPs $r^2 > 0.5$ (EUR) within a 1 Mb window [2]. Next, we extend the interval to the nearest recombination hotspots with recombination rate >3 cM/Mb [3]. To address the case when no genes overlap an interval, we provide an option for SNPsea to extend the interval up- and downstream (by default 10 Kb).

Most frequently, we find multiple genes ($m_k > 1$) in a single SNP locus k. We expect many loci with multiple genes because of regions with high LD across long stretches of a chromosome. Less frequently, a locus has a single gene ($m_k = 1$), and loci with no genes ($m_k = 0$) are discarded.

After each SNP has been assigned an interval and a set of genes overlapping the interval, we merge SNPs with shared genes into a single locus to avoid multiple-counting of genes.

Step 2: Calculating specificity scores

SNPsea uses different algorithms for matrices with continuous or binary values. By default, SNPsea assumes one gene in each associated locus is associated with the given trait. We also include the option to assume all genes within a locus are associated. We compare results of the two options with four phenotypes (**Supplementary Figure 4**).

1. The `'--score single'` option (default) assumes that a single gene in each locus is associated with the given phenotype. For each condition, we choose the gene in each locus with the greatest specificity to that condition.

2. The `'--score total'` option assumes that all genes in a SNP's linkage interval are associated. We account for all linked genes when calculating scores.

Specificity for a matrix of continuous values Before running SNPsea, a matrix with continuous values must be normalized so that columns are directly comparable. *It is not appropriate to use this method on a "raw" matrix of expression values.*

We extend an approach we have previously described in detail [4]. Let A denote a continuous gene expression matrix with m genes and n conditions. First, we normalize the expression of each gene by dividing each value by the L2 norm of the gene's values across all conditions.

$$A'_{i,j} = \frac{A_{i,j}}{\sqrt{A_{i,1}^2 + A_{i,2}^2 + \cdots + A_{i,n}^2}}$$

The resulting matrix A' has values $A'_{i,j}$ between 0 and 1 indicating specificity of gene i to condition j. A value $A'_{i,j} = 1$ indicates that gene i is exclusively expressed in condition j, and $A'_{i,j} = 0$ indicates that gene i is not expressed in condition j.

Next, we transform A' to a matrix A'' of non-parametric condition-specificity percentiles as follows. For each condition j, we rank the values of $A'_{,j}$ in ascending order and divide them by the number of genes m, resulting in percentiles between 0 and 1 where a lower value indicates greater specificity to the given condition.

$$A''_{i,j} = \frac{\mathrm{Rank}_j(A'_{i,j})}{m}$$

Locus scores for a matrix of continuous values We create a new matrix P, where each value $P_{k,j}$ is a score for a SNP locus k and a condition j. We define the locus scores $P_{,j}$ for a single condition j to be approximately uniformly distributed for a set of randomly selected loci under the null hypothesis of no association to the condition. We make the assumption that, for a set of genes in a given SNP locus I_k,

the values $A''_{i \in I_k,j}$ are random, independent, and approximately uniformly distributed. If there is an actual association to a condition j, we will observe an unexpectedly small value for $P_{,j}$.

`'--score single' (default)` This approach tests for the association of one gene in each SNP locus to condition j.

For each locus-condition pair (k, j), we choose the single gene i in locus k with greatest specificity to condition j among the m_k genes in the locus, as previously described in Hu *et al.* [4]. Let g_k denote this most specific gene, so that $A''_{g_k,j} = \text{Min}_{i \in I_k}(A''_{i,j})$ where I_k denotes the set of genes in locus k. If we assume values of $A''_{i \in I_k,j}$ are uniformly distributed for a given condition j and genes $i \in I_k$, then the probability to randomly draw a value equal to or less than $A''_{g_k,j}$ is as follows:

$$P_{k,j} = 1 - (1 - \text{Min}_{i \in I_k}(A''_{i,j}))^{m_k}$$

`'--score total'` This approach tests for the association of all genes in each SNP locus to condition j — we consider this model to be unlikely in most situations. The product (log sum) of uniform values between (0,1) follows a gamma distribution [5]. If the genes I_k have no specificity to a condition j, then the values $A''_{i \in I_k}$ are approximately uniformly distributed. So, we compute the probability to randomly draw values $A''_{i \in I_k}$ with a smaller product as the upper tail of the gamma distribution:

$$P_{k,j} \quad = \quad \int_x^{\infty} \Gamma(m_k, 1) \;\; \text{for} \;\; x = \sum_{i \in I_k} -\text{ln} A''_{i,j}$$

Locus scores for a matrix of binary values Let B denote a binary matrix (1=present, 0=absent) with m genes and n conditions. Let m_j denote the number of genes present in condition j. Let m_k denote the number of genes in locus k and $m_{k,j} \leq m_k$ denote the number of genes in locus k that are present in condition j.

We provide two options to calculate locus scores. By default, we account for presence or absence of any of the m_k genes in condition j, as shown below (`'--score single'`). Alternatively, we account for the number of genes in a given locus (`'--score total'`).

	`'--score single'`	`'--score total'`
$p(x) = \frac{\binom{m_j}{x}\binom{m-m_j}{m_k-x}}{\binom{m}{m_k}}$	$P_{k,j} = \begin{cases} 1 - p(0) & m_{k,j} > 0 \\ 1 & m_{k,j} = 0 \end{cases}$	$P_{k,j} = \begin{cases} 1 - \sum_{x=0}^{m_{k,j}-1} p(x) & m_{k,j} > 0 \\ 1 & m_{k,j} = 0 \end{cases}$

Condition specificity scores For both continuous and binary matrices, we define a specificity score S_j for each condition j as the aggregate of $P_{k,j}$ values across SNP loci:

$$S_j = \sum_k -\text{log} P_{k,j}$$

Step 3: Testing significance

Analytical p-values We previously found that aggregating the $P_{k,j}$ scores and determining a P-value analytically from a distribution results in inaccurate p-values [4]. $A''_{i,j}$ values may be relatively uniform genome-wide, but proximate genes often have shared functions. The genome has a complex correlation structure of linkage disequilibrium, gene density, gene size and function that is challenging to model analytically. We use the sampling strategy described below instead.

Permutation p-values For each condition, we use a sampling approach to calculate an empirical p-value. This is the tail probability of observing a condition-specificity score greater or equal to S_j. We obtain the distribution empirically with null SNP sets.

We compute specificity scores S for random SNP sets. Each SNP in a null set is matched to a SNP in the user's set on the number of linked genes. To adequately sample genes from the entire genome, we sample SNP sets from a list of LD-pruned SNPs (subset of SNPs in 1000 Genomes Project) [6].

For each condition j, we calculate an exact permutation p-value [7]. Let a_j denote the number of sampled SNP sets (e.g. 10,000) and let b_j denote how many null specificity scores are greater than or equal to the user's score S_j:

$$p_j = \frac{b_j + 1}{a_j + 1}$$

We implemented adaptive sampling to calculate p-values efficiently. As each condition is tested for significance, we increase the number of iterations to resolve significant p-values and save computation by using fewer iterations for less significant p-values. Two options allow the user to control the adaptive sampling:

1. `'--max-iterations N'` The maximum number of iterations for each condition. We stop testing a condition after sampling N SNP sets.

2. `'--min-observations N'` The minimum number of observed null specificity scores greater than or equal to S_j required to stop sampling SNP sets for a condition j.

Data

Please find the data required to reproduce this analysis here: http://dx.doi.org/10.6084/m9.figshare.871430

Gene Atlas gene expression matrix

We downloaded the data from BioGPS: http://plugins.biogps.org/download/gnf1h-gcrma.zip

We averaged the expression values for tissue replicates. For each gene, we selected the single probe with the largest minimum value. Finally, we converted the file to GCT format.

Gene Ontology binary presence/absence matrix

We downloaded the OBO file from Gene Ontology (data-version: 2013-06-29, CVS revision: 9700):

http://www.geneontology.org

For each gene, we climbed the hierarchy of ontology terms and applied parental terms. If a gene is annotated with some term T, we also add all of the terms that are parents of T. We copy terms between

homologous genes using Homologene data (http://www.ncbi.nlm.nih.gov/homologene). If a mouse gene is annotated with some term and the human homolog is not, then we copy the term to the human gene. We discard all GO terms assigned to fewer than 100 or to more than 1000 genes. This leaves us with a matrix of 19,111 genes and 1,751 terms.

1000 Genomes Project

We downloaded a filtered (diallelic and 5 or more copies of the minor allele) set of markers from the BEAGLE website and calculated pairwise LD (EUR) for all SNPs in a 1 Mb sliding window:

http://bochet.gcc.biostat.washington.edu/beagle

Commands

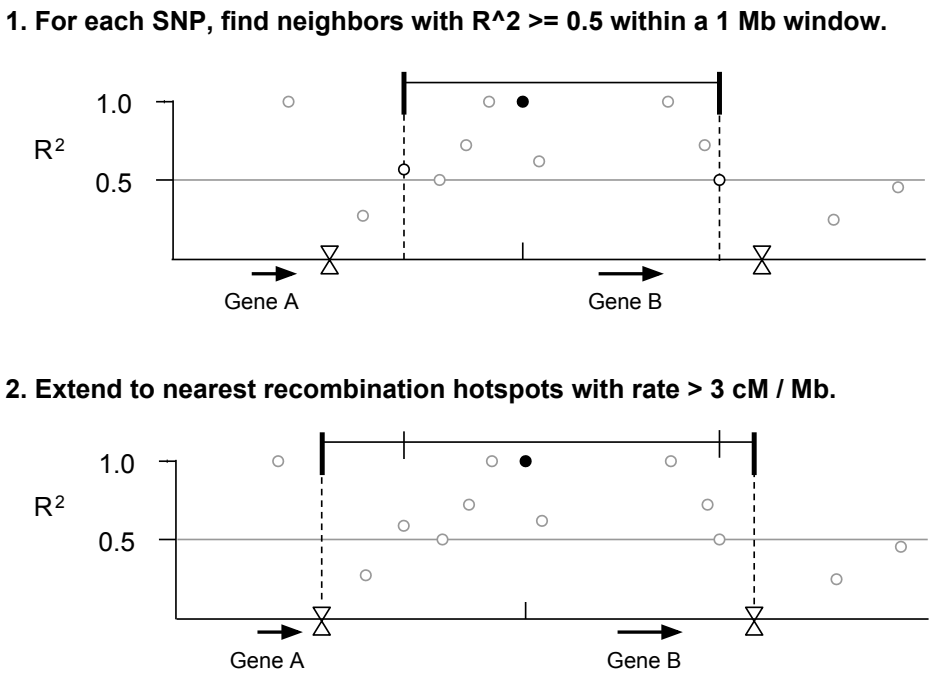
More details: **http://www.broadinstitute.org/mpg/snpsea/SNPsea_manual.html**

Analysis	Figures

`snpsea`
`--snps HDL_Teslovich2010.txt`
`--gene-matrix GeneAtlas2004.gct.gz`
`--gene-intervals NCBIgenes2013.bed.gz`
`--snp-intervals TGP2011.bed.gz`
`--null-snps Lango2010.txt.gz`
`--score single`
`--out <out>`
`--slop 10e3`
`--threads 2`
`--null-snpsets 0`
`--min-observations 100`
`--max-iterations 1e6`
	Horizontal bar plot of p-values:
	`snpsea-barplot <out>`
	
	Heatmap of conditions and loci:
	`snpsea-heatmap <out>`
	
	Type 1 error rates for each condition:
	`snpsea-type1error <out>`

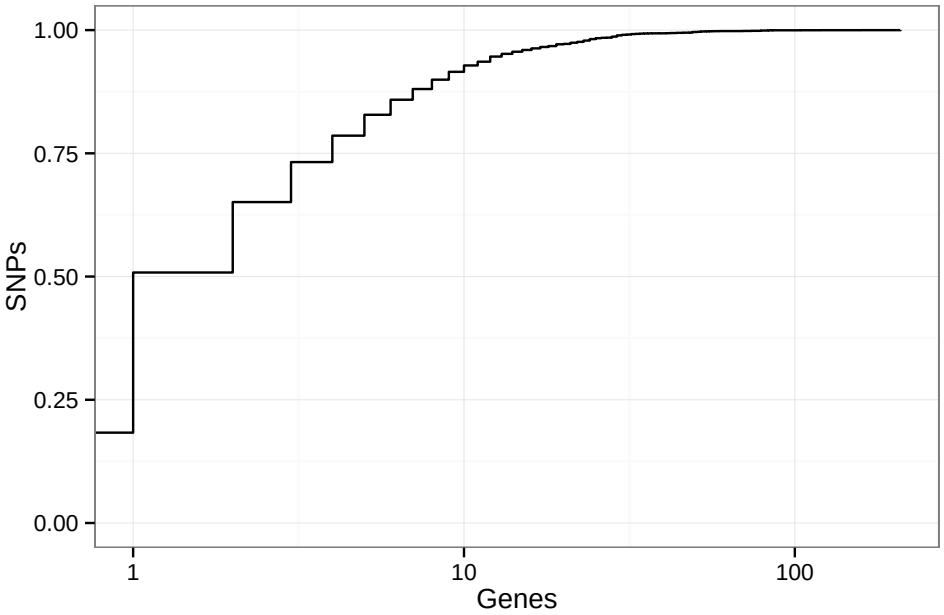
Supplementary Figures

Supplementary Figure 1: Determining SNP linkage intervals



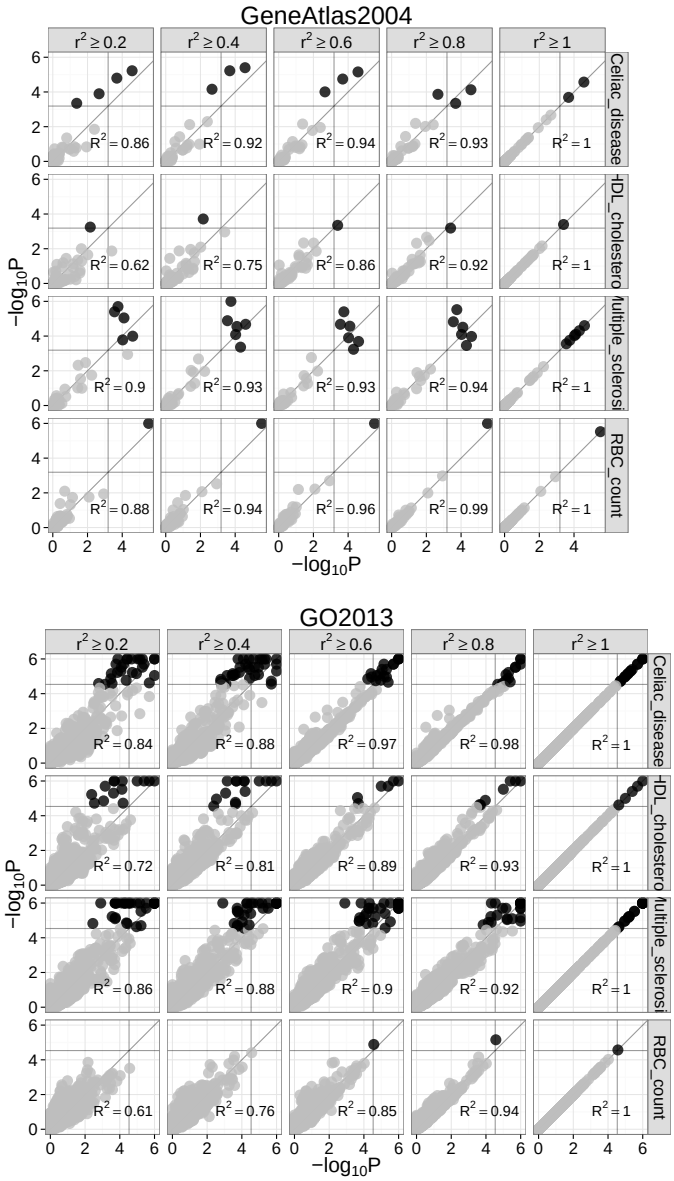
We calculated r^2 values for all pairs of SNPs within a 1 Mb sliding window along each chromosome. Next, we assigned each of the SNPs from The 1000 Genomes Project Phase I [2] to a linkage interval by identifying each SNP's furthest upstream and downstream neighbors with $r^2 \geq 0.5$. Finally, we extended each interval to recombination hotspots reported by HapMap [3] with recombination rate >3 cM/Mb.

Supplementary Figure 2: Counting genes in GWAS SNP linkage intervals



A cumulative density plot of the number of genes overlapped by the linkage intervals of GWAS SNPs. We downloaded the GWAS Catalog SNPs on January 17, 2014 and selected the 11,561 SNPs present in the 1000 Genomes Project [2]. Of these SNPs, 2,119 (18%) of them have linkage disequilibrium (LD) intervals that overlap no genes, and 3,756 (32%) overlap a single gene. The remaining 50% of SNPs overlap 2 or more genes. This illustrates the critical issue that many SNPs implicate more than one gene.

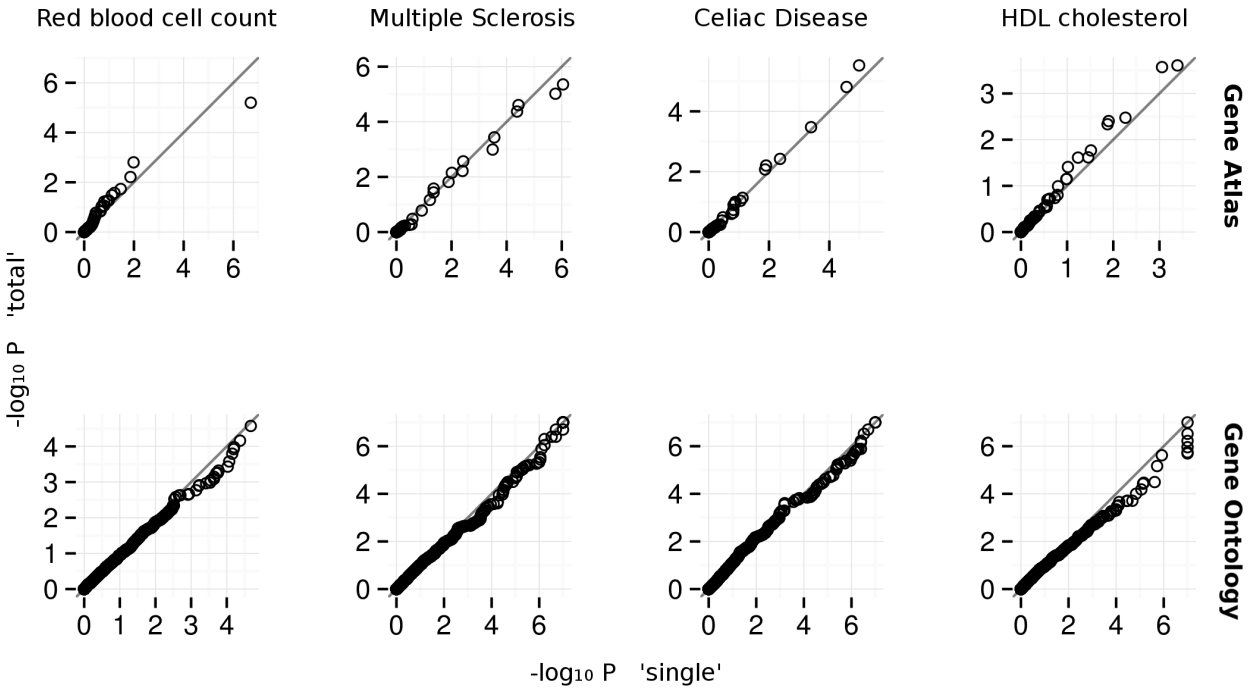
Supplementary Figure 3: Choosing the r^2 threshold for linkage intervals



Gene Atlas and Gene Ontology (top and bottom). Each subplot has $-\log_{10}P$ for $r^2 = 1$ on the x-axis and $-\log_{10}P$ on the y-axis for the r^2 threshold marked above. Grey lines are significance thresholds after correction testing multiple conditions (cell types, GO annotations). Black points are significant and grey are not. We used the `'--score single'` option. Red blood cell count SNPs are enriched for *hemopoiesis* (GO:0030097) ($P = 2 \times 10^{-5}$) for linkage intervals with $r^2 = (0.6, 0.8, 1.0)$. This result falls below the multiple testing threshold at $r^2 \geq 0.4$, but remains significant at $r^2 \geq 0.5$ (see main text).

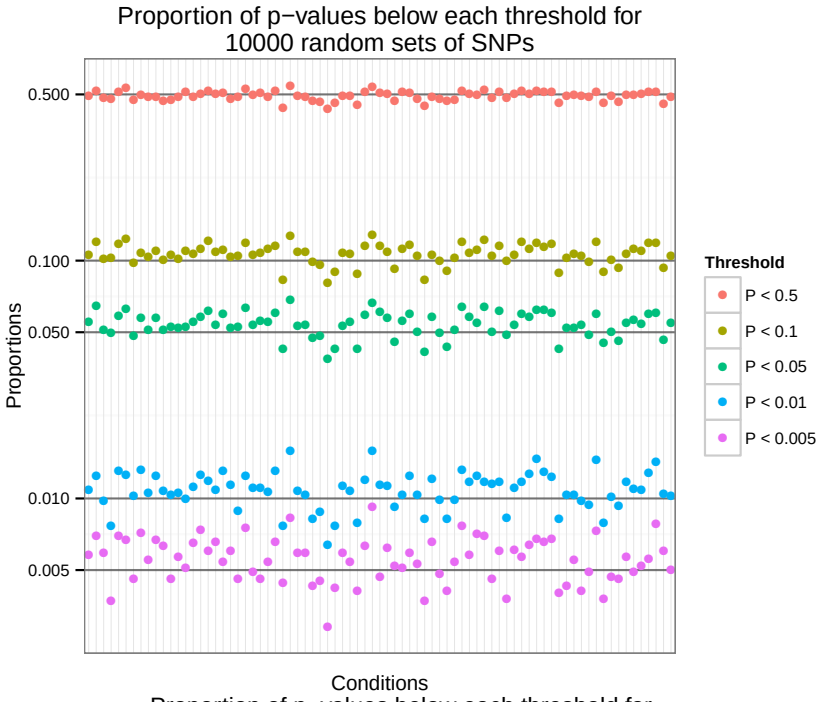
To investigate if the choice of r^2 threshold influences SNPsea results, we repeated the analysis of four traits using 5 different thresholds ($r^2 \geq 0.2$, 0.4, 0.6, 0.8, 1.0). SNPsea results seem to be robust to the choice of threshold, mostly due to the fact that we extend linkage intervals for each SNP to the nearest recombination hotspots as described in **Supplementary Figure 1**. We chose to use $r^2 \geq 0.5$ for our analysis and for our provided data files due to this result, and also due to previous experience [1].

Supplementary Figure 4: Comparison of 'single' and 'total' options



Quantile-quantile plots for Gene Atlas [8] and Gene Ontology (top and bottom). The x and y axes are $-\log_{10}P$ for `'--score single' and '--score total'` SNPsea options, respectively. The `'single'` and `'total'` methods are described on page 3. The P-values appear similar between methods.

Supplementary Figure 5: Type 1 error estimates



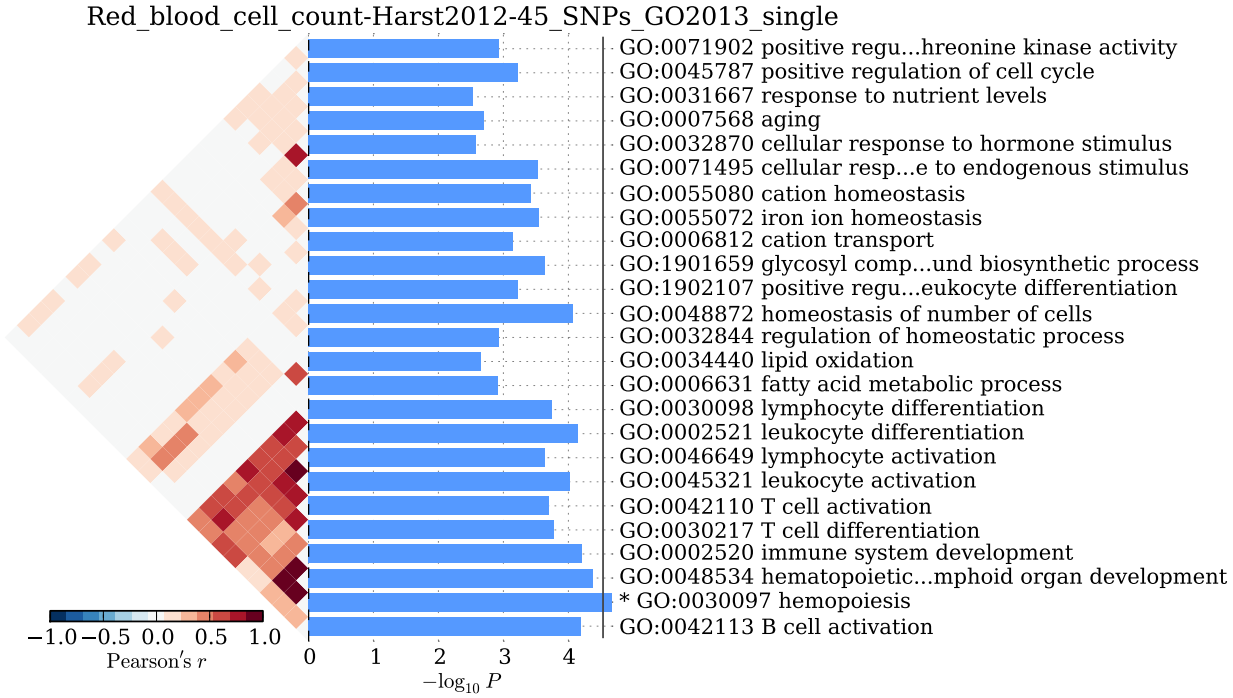
We sampled 10,000 sets of 100 SNPs uniformly from a list of LD-pruned SNPs [6]. We tested each of the 10,000 sets for enrichment of tissue-specific expression in the Gene Atlas [8] gene expression matrix (top) and for enrichment of annotation with Gene Ontology terms (bottom). For each condition, we show the proportion of the 10,000 enrichment p-values that are below the given thresholds. We observe that the p-values are near the expected values, so the type 1 (false positive) error rate is well-calibrated.

Additional Examples

We tested SNPsea with the three additional phenotypes listed below with genome-wide significant SNPs ($P \leq 5 \times 10^{-8}$). When multiple SNPs implicated the same genes, we merged them into a single locus. We tested each phenotype with the Gene Atlas and GO matrices with the `'--score single'` option. Below we show the number of significantly enriched conditions we found for each phenotype.

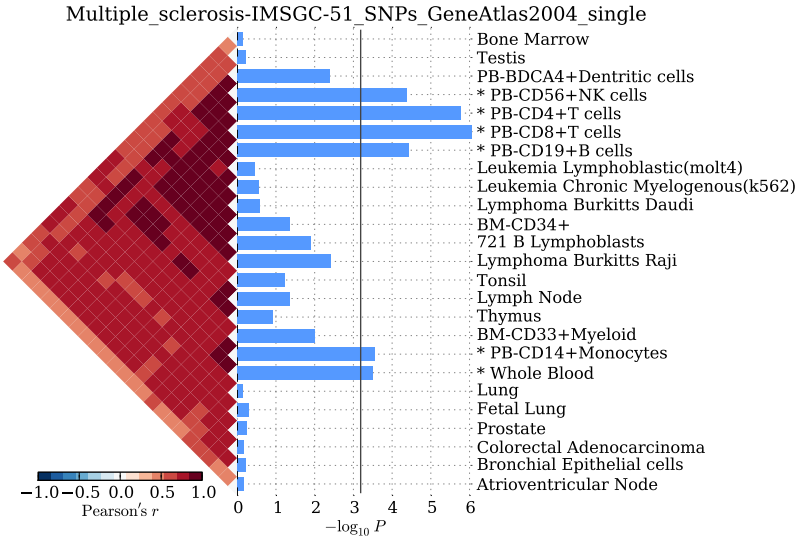
Phenotype	SNPs	Loci	GO	Gene Atlas	Reference	
Red blood cell count	45	45	1	1	Table 1	van der Harst, *et al.* 2012 [9]
Multiple sclerosis	51	47	52	6	Supp. Table A	IMSGC WTCCC 2011 [10]
Celiac disease	35	34	28	3	Table 2	Trynka, *et al.* 2011 [11]
HDL cholesterol	46	46	13	1	Supp. Table 2	Teslovich, *et al.* 2010 [12]

Supplementary Figure 6: Red blood cell count GO enrichment

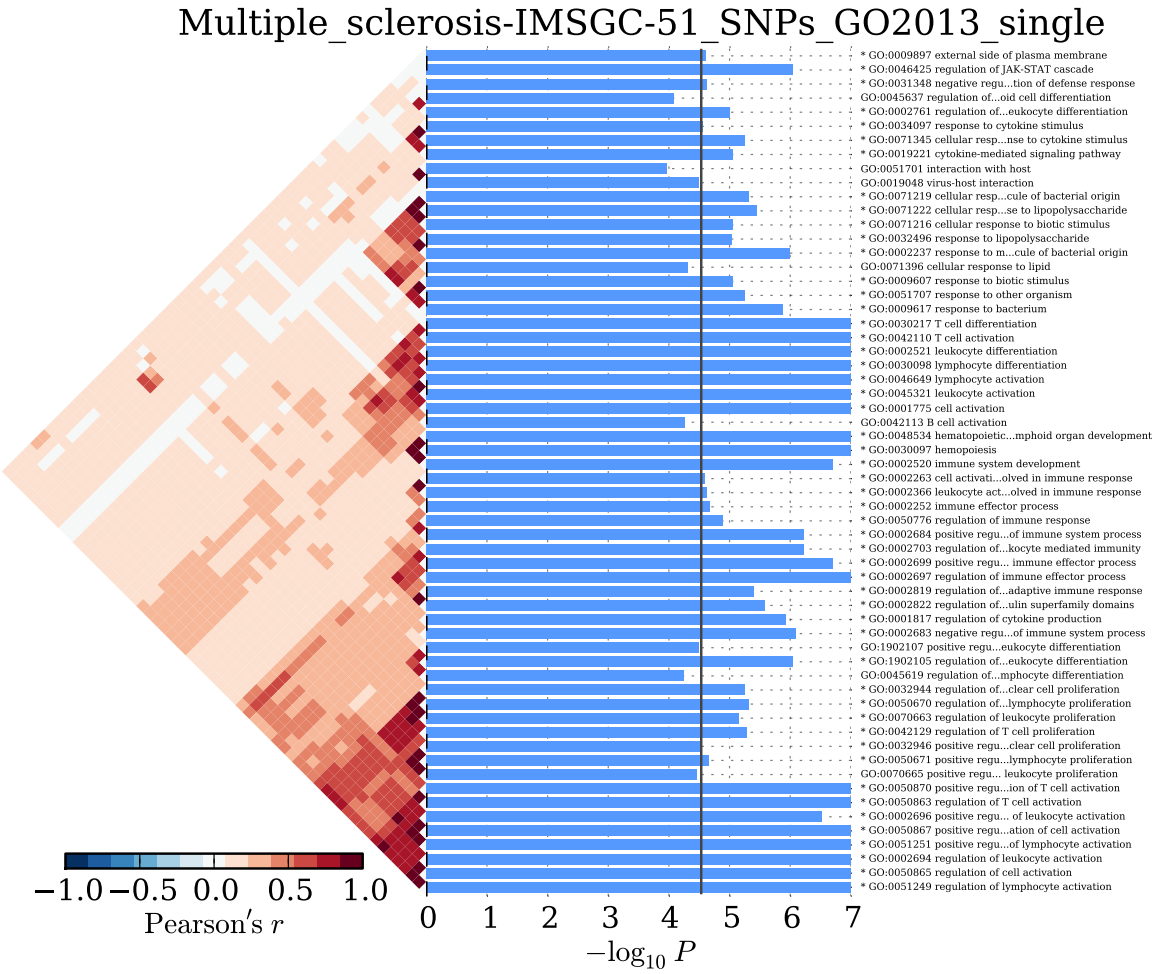


We observed significant enrichment for *hemopoiesis* (2×10^{-5}). The top 25 terms are shown.

Supplementary Figure 7: Multiple sclerosis

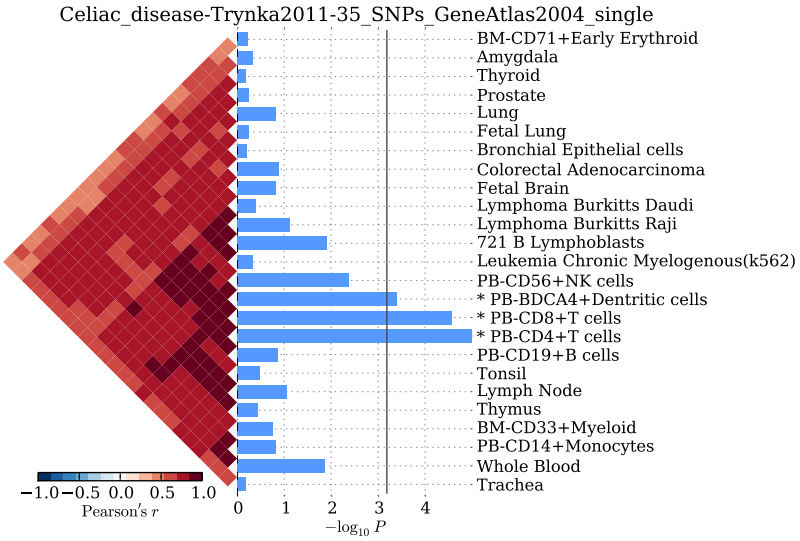


We observed significant enrichment for 6 cell types. The top 25 of 79 are shown.

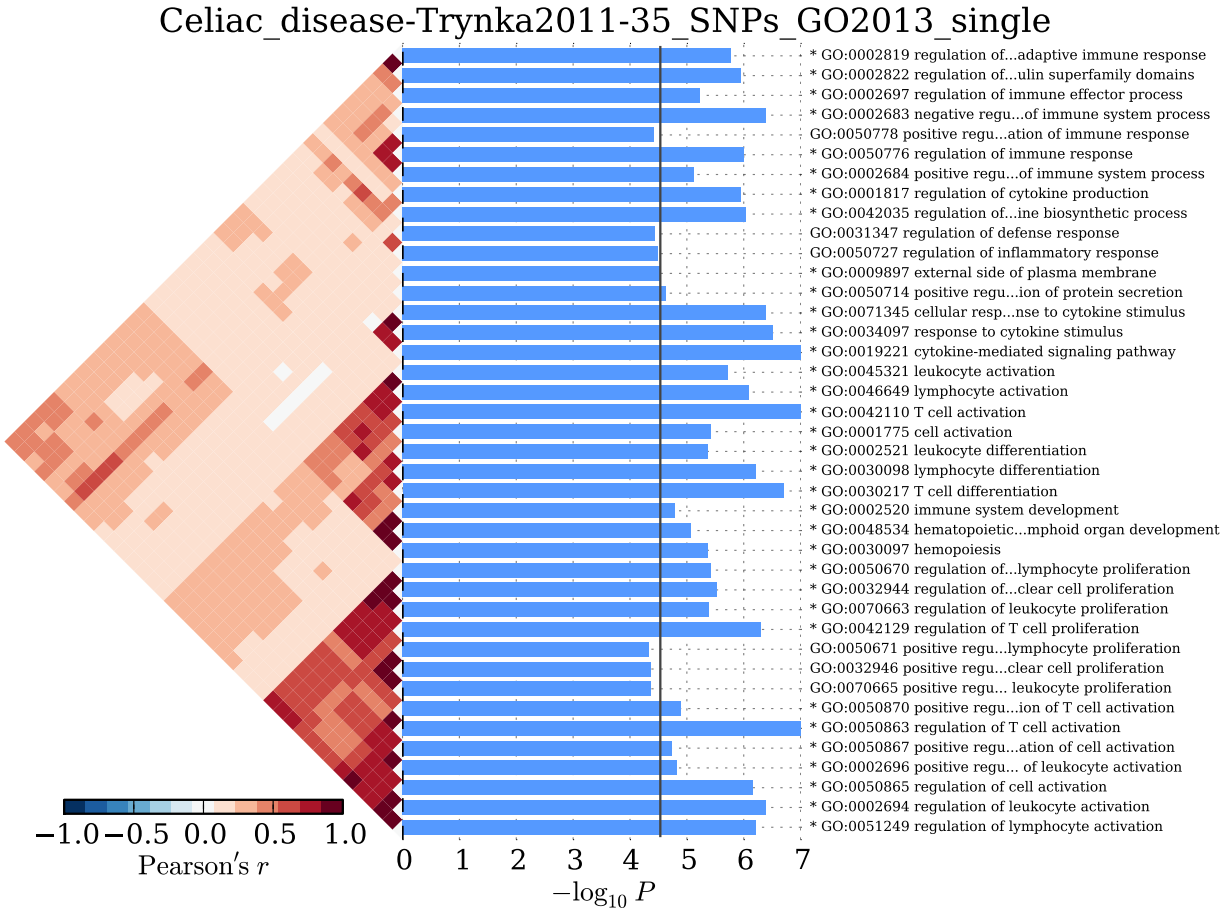


We observed significant enrichment for 52 Gene Ontology terms. The top 60 terms are shown.

Supplementary Figure 8: Celiac disease

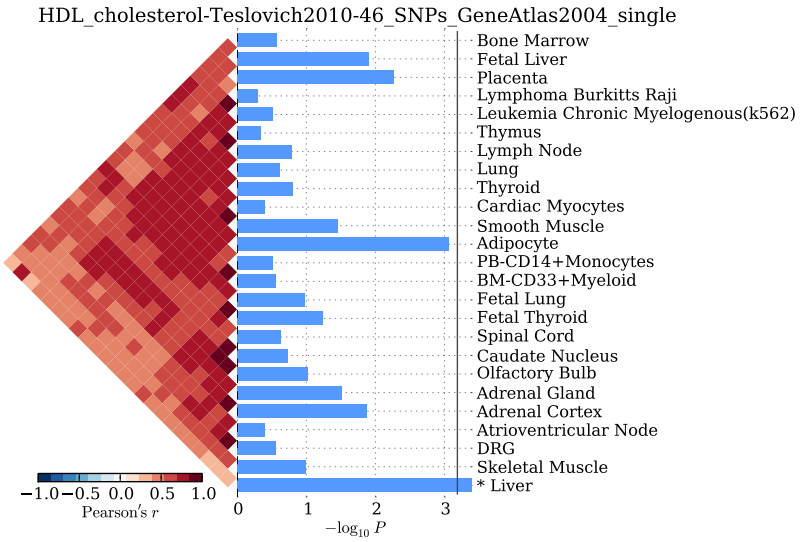


We observed significant enrichment for 3 cell types. The top 25 of 79 are shown.

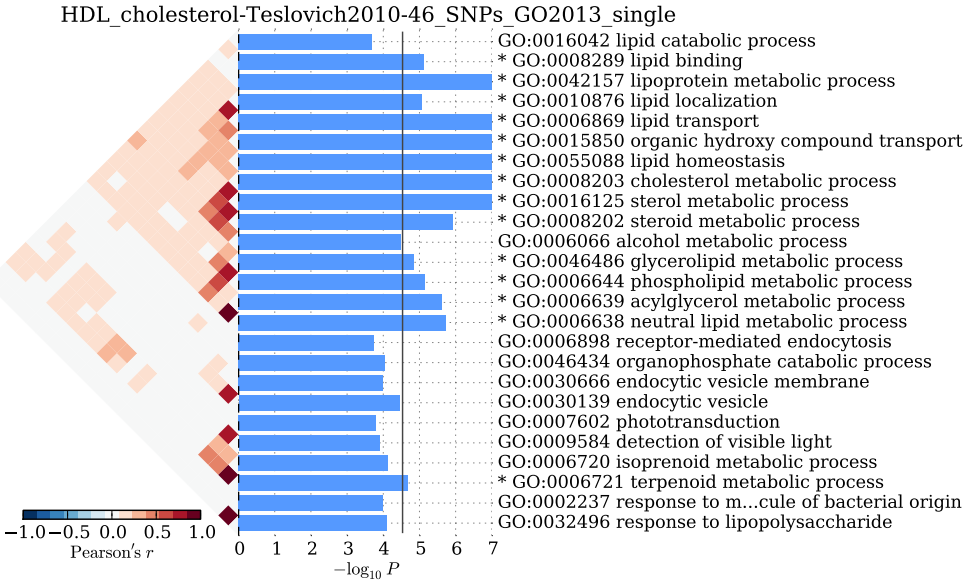


We observed significant enrichment for 28 Gene Ontology terms. The top 40 terms are shown.

Supplementary Figure 9: HDL cholesterol



We observed significant enrichment for liver tissue-specific gene expression. The top 25 of 79 are shown.



We observed significant enrichment for 13 Gene Ontology terms. The top 25 terms are shown.

References

[1] Elizabeth J. Rossin, Kasper Lage, Soumya Raychaudhuri, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*, 7(1):e1001273, January 2011.

[2] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012.

[3] Simon Myers, Leonardo Bottolo, Colin Freeman, et al. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, October 2005.

[4] X. Hu, H. Kim, E. Stahl, et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *The American Journal of Human Genetics*, 89(4):496âĂŞ506, 2011.

[5] Norman Lloyd Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate discrete distributions.* Wiley, Hoboken, N.J., 2005.

[6] Hana Lango Allen, Karol Estrada, Guillaume Lettre, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, October 2010.

[7] Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[8] Andrew I. Su, Tim Wiltshire, Serge Batalov, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, April 2004. PMID: 15075390 PMCID: PMC395923.

[9] Pim van der Harst, Weihua Zhang, Irene Mateo Leach, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429):369–375, December 2012.

[10] The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, August 2011.

[11] Gosia Trynka, Karen A. Hunt, Nicholas A. Bockett, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics*, 43(12):1193–1201, December 2011.

[12] Tanya M. Teslovich, Kiran Musunuru, Albert V. Smith, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, August 2010.