SNPsea Documentation

Release 1.0.3

Kamil Slowikowski

June 17, 2015

Contents

Github project: https://github.com/slowkow/snpsea

CHAPTER 1

Introduction

SNPsea is an algorithm to identify cell types and pathways likely to be affected by risk loci. It requires a list of SNP identifiers and a matrix of genes and conditions.

Genome-wide association studies (GWAS) have discovered multiple genomic loci associated with risk for different types of disease. SNPsea provides a simple way to determine the types of cells influenced by genes in these risk loci.

Suppose disease-associated alleles influence a small number of pathogenic cell types. We hypothesize that genes with critical functions in those cell types are likely to be within risk loci for that disease. We assume that a gene's specificity to a cell type is a reasonable indicator of its importance to the unique function of that cell type.

First, we identify the genes in linkage disequilibrium (LD) with the given trait-associated SNPs and score the gene set for specificity to each cell type. Next, we define a null distribution of scores for each cell type by sampling random SNP sets matched on the number of linked genes. Finally, we evaluate the significance of the original gene set's specificity by comparison to the null distributions: we calculate an exact permutation p-value.

SNPsea is a general algorithm. You may provide your own:

1. Continuous gene matrix with gene expression profiles (or other values).
2. Binary gene annotation matrix with presence/absence 1/0 values.

We provide you with three expression matrices and one annotation matrix. See Data.

The columns of the matrix may be tissues, cell types, GO annotation codes, or other *conditions*.

Note: Continuous matrices *must* be normalized before running SNPsea. That is, columns must be directly comparable to each other. For example, you might consider quantile normalization.

If you benefit from this method, please cite:

> Slowikowski, K. et al. SNPsea: an algorithm to identify cell types, tissues, and pathways affected by risk loci. Bioinformatics (2014).

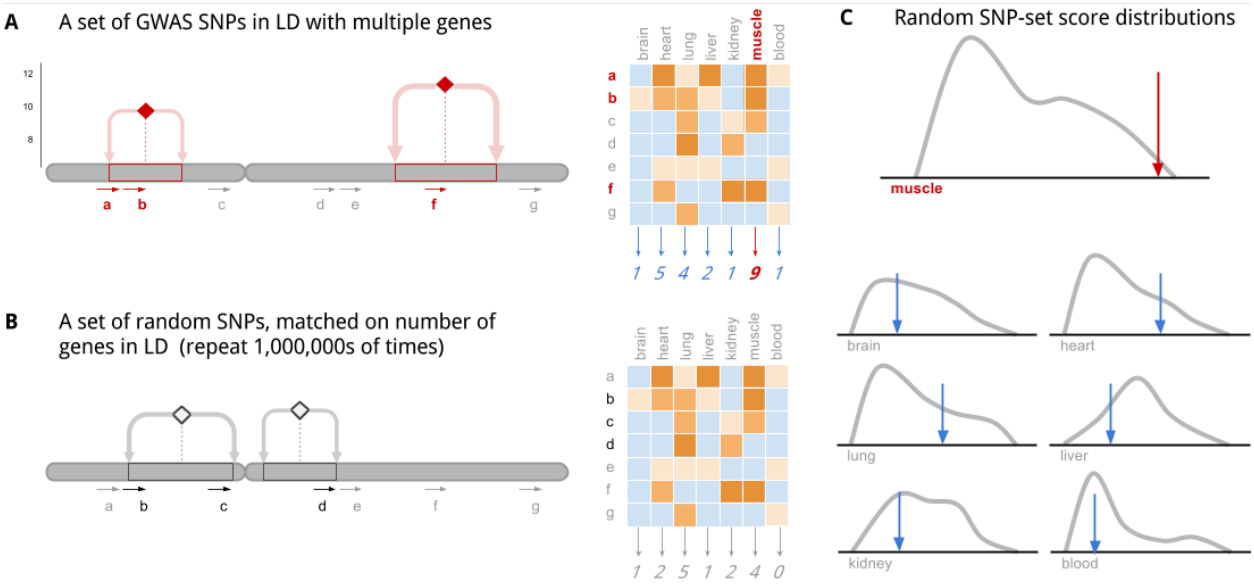
See the first description of the algorithm and additional examples here:

> Hu, X. et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. The American Journal of Human Genetics 89, 496–506 (2011).

CHAPTER 2

Visual Summary

2.1 Cartoon



This cartoon illustrates the key ideas of the algorithm:

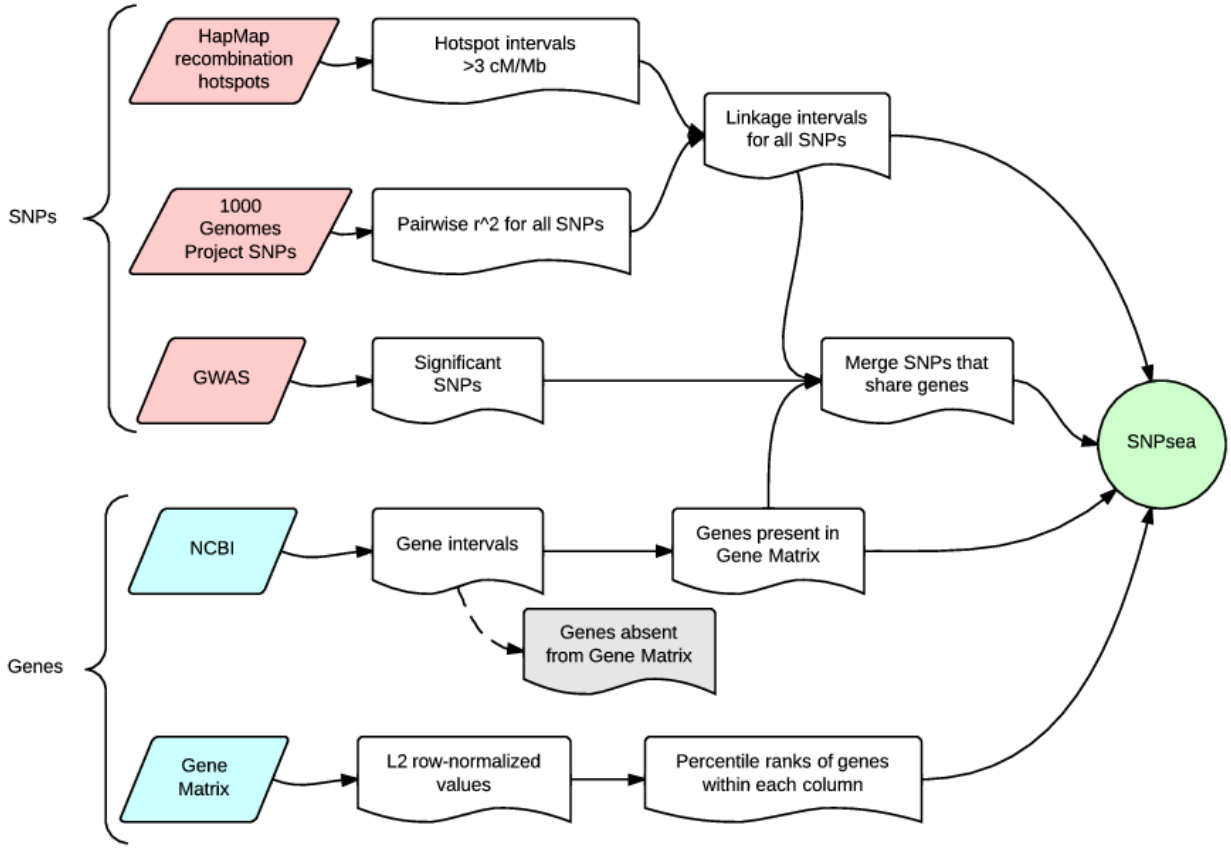
A| Step 1. Each SNP in a set of disease-associated SNPs is in linkage disequilibrium (LD) with multiple genes. The genes are scored, in aggregate, for specificity to each tissue.

B| Step 2: The algorithm is repeated with random null SNP sets that are not associated with any phenotype. These have been selected from an LD-pruned list of SNPs, so the whole genome is covered.

C| Step 3: The random SNP set scores form the null distributions which allows us to determine statistical significance for enrichment of specificity to a particular tissue/cell-type/condition.

2.2 Flow Chart

This flow chart shows the input data required to perform the analysis, and a summary of the intermediate steps.


SNPs
HapMap recombination hotspots
Hotspot intervals >3 cM/Mb
1000 Genomes Project SNPs
Pairwise r^2 for all SNPs
Linkage intervals for all SNPs
GWAS
Significant SNPs
Merge SNPs that share genes
SNPsea
Genes
NCBI
Gene intervals
Genes present in Gene Matrix
Genes absent from Gene Matrix
Gene Matrix
L2 row-normalized values
Percentile ranks of genes within each column

CHAPTER 3

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:
    - Gene expression profiles of multiple 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.

3.1 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 Rossin *et al.* 2011). 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 (1000 Genomes Consortium 2012).

Next, we extend the interval to the nearest recombination hotspots with recombination rate >3 cM/Mb (Myers *et al.* 2005). 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.

3.1.1 Two score options

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'` method (default option) 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'` method assumes that all genes in a SNP's linkage interval are associated. We account for all linked genes when calculating scores.

3.2 Step 2: Calculating specificity scores

SNPsea uses different algorithms for matrices with continuous or binary 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.*

3.2.1 Specificity for a matrix of continuous values

We extend an approach we have previoulsy described in detail (Hu *et al.* 2011). 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 genes values in different 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 descending 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.

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

3.2.2 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. The locus scores $P_{,j}$ for a single condition j are approximately uniformly distributed for a set of randomly selected loci under the following assumption: for the 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. We'll come back to this assumption later when testing significance in Step 3 below.

3.2.3 `'--score single' (default)`

This approach assumes one gene in each SNP locus is associated with the trait.

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.* (Hu *et al.* 2011). 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 of obtaining a value equal to or less than $A''_{g_k,j}$ is as follows:

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

3.2.4 `'--score total'`

This assumes all genes in a given SNP locus are associated with a trait — we consider this model to be unlikely in most situations. We compute the probability of observing values $A''_{i \in I_k}$ for some locus k as the product of percentiles. This assumes $A''_{i \in I_k}$ values are uniformly distributed.

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

3.2.5 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'`
$P_{k,j} = \begin{cases} 1 - p(0) & m_{k,j} > 0 \\ 1 & m_{k,j} = 0 \end{cases}$	$P_{k,j} = \begin{cases} 1 - \sum_{x=0}^{m_{k,j}-1} p(x) & m_{k,j} > 0 \\ 1 & m_{k,j} = 0 \end{cases}$

where

$$p(x) = \frac{\binom{m_j}{x}\binom{m-m_j}{m_k-x}}{\binom{m}{m_k}}$$

3.2.6 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:

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

3.3 Step 3: Testing significance

3.3.1 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 (Hu *et al.* 2011). $A''_{i,j}$ 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.

3.3.2 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) (Lango Allen *et al.* 2010).

For each condition j, we calculate an exact permutation p-value (Phipson *et al.* 2010). 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.

3.4 Example

Suppose we have a gene expression matrix A:

```
> A1 = read.table(text = "
2.55 0.05 3.28 1.11
2.63 4.53 4.66 3.89
0.61 3.31 2.49 4.59
0.82 1.27 4.47 2.31
```

```
4.91 1.23 0.51 0.95")
> A1
    V1   V2   V3   V4
1 2.55 0.05 3.28 1.11
2 2.63 4.53 4.66 3.89
3 0.61 3.31 2.49 4.59
4 0.82 1.27 4.47 2.31
5 4.91 1.23 0.51 0.95
```

Compute the specificity (L2 norm) of each gene (row) to each condition (column):

```
> A2 = t(apply(A1, 1, function(row) row / sqrt( sum(row ^ 2) )))
> A2
             V1          V2         V3        V4
[1,] 0.59293508 0.01162618 0.76267727 0.2581012
[2,] 0.32801918 0.56499121 0.58120508 0.4851690
[3,] 0.09818755 0.53278820 0.40079837 0.7388211
[4,] 0.15607783 0.24173030 0.85081451 0.4396827
[5,] 0.94873958 0.23766796 0.09854525 0.1835647
```

Rank the genes in each condition and convert to percentiles:

```
A3 = apply(A2, 2, function(col) rank(-col) / length(col))
> A3
      V1  V2  V3  V4
[1,] 0.4 1.0 0.4 0.8
[2,] 0.6 0.2 0.6 0.4
[3,] 1.0 0.4 0.8 0.2
[4,] 0.8 0.6 0.2 0.6
[5,] 0.2 0.8 1.0 1.0
```

Notice that gene 3 has the greatest specificity (0.74) to condition V4, so it is assigned the lowest percentile rank (0.2).

Compute the locus scores for a SNP locus k that overlaps genes 2 and 4, assuming that a single gene (either 2 or 4 but not both) is associated with the trait:

```
> genes = c(2, 4)
> P = apply(A3[genes, ], 2, function(col) 1 - (1 - min(col)) ^ length(col))
> P
  V1   V2   V3   V4
0.84 0.36 0.36 0.64
```

Notice that the SNP locus k is most specific to conditions V2 and V3 (0.36), and this is because:

- gene 2 has the lowest specificity percentile (0.2) in condition V2
- gene 4 has the lowest specificity percentile (0.2) in condition V3

CHAPTER 4

Installation

On Linux 64-bit, you may use the provided executable

This runs on kernel 2.6.18 and newer: https://github.com/slowkow/snpsea/releases

Otherwise, you must build the executable from source

The source code is available: https://github.com/slowkow/snpsea

Mac: To compile C++ code with the required dependencies, you need XCode and MacPorts: http://guide.macports.org/#installing.xcode

Install the dependencies:

```
# Ubuntu
sudo apt-get install build-essential libopenmpi-dev libgsl0-dev

# Mac
#   First, install port (MacPorts): http://www.macports.org/
#   Next, use it to install the dependencies:
sudo port selfupdate && sudo port install gcc48 openmpi gsl

# Broad Institute
#   Add this line to ~/.my.bashrc or ~/.my.cshrc
use .gcc-4.8.1 .openmpi-1.4 .gsl-1.14
```

Download and compile the code:

```
#   Clone with git; easily get updates with 'git pull':
git clone https://github.com/slowkow/snpsea.git
cd snpsea

#   If you don't have git:
curl -LOk https://github.com/slowkow/snpsea/archive/master.zip
unzip master.zip; cd snpsea-master

cd src; make               #   Compile.
cp ../bin/snpsea* ~/bin/   #   Copy the executables wherever you like.
```

4.1 C++ Libraries

To compile SNPsea, you will need a modern C++ compiler that supports c++0x and the dependencies listed below. I compiled successfully with gcc versions 4.6.3 (the default version for Ubuntu 12.04) and 4.8.1.

intervaltree

a minimal C++ interval tree implementation

Eigen

Eigen is a C++ template library for linear algebra: matrices, vectors, numerical solvers, and related algorithms.

OpenMPI

MPI is a standardized API typically used for parallel and/or distributed computing. Open MPI is an open source, freely available implementation.

GSL - GNU Scientific Library

The GNU Scientific Library (GSL) is a numerical library for C and C++ programmers.

GCC, the GNU Compiler

The GNU Compiler Collection is a compiler system produced by the GNU Project supporting various programming languages.

4.2 Python Packages

To plot visualizations of the results, you will need Python 2.7 and the packages listed below.

Instructions: Install with pip:

```
pip install docopt numpy pandas matplotlib
```

Note: The packages available on the Ubuntu repositories may be outdated and might fail to work. So, avoid using `apt-get` for these dependencies.

docopt

Command-line interface description language.

numpy

NumPy is the fundamental package for scientific computing with Python.

pandas

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

matplotlib

matplotlib is a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

Note: On a server with no display, please edit your matplotlibrc file to use the `Agg` backend:

```
perl -i -pe 's/^(\s*(backend).*)$/#$1\n$2:Agg/' ~/.matplotlib/matplotlibrc
```

Otherwise, you may see an error message like this:

```
_tkinter.TclError: no display name and no $DISPLAY environment variable
```

4.3 R Packages

Some visualizations use R and ggplot2 instead of Python and matplotlib.

Instructions: Start a session in R and run:

```
install.packages(c("data.table", "reshape2", "gap", "ggplot2"))
```

data.table

: Extension of data.frame for fast indexing, fast ordered joins, fast assignment, fast grouping and list columns.

reshape2

: Flexibly reshape data: a reboot of the reshape package.

gap

: Genetic analysis package.

ggplot2

: An implementation of the Grammar of Graphics.

CHAPTER 5

Data

```
cd snpsea
curl -LOk http://files.figshare.com/1504037/SNPsea_data_20140520.zip
unzip SNPsea_data_20140520.zip
```

Download the compressed archive with data required to perform this analysis (138M). The direct link to the zip shown above may be out of date and fail to load. If so, please visit the link below instead:

http://dx.doi.org/10.6084/m9.figshare.871430

Contents of the compressed archive with data:

```
Celiac_disease-Trynka2011-35_SNPs.gwas
HDL_cholesterol-Teslovich2010-46_SNPs.gwas
Multiple_sclerosis-IMSGC-51_SNPs.gwas
Red_blood_cell_count-Harst2012-45_SNPs.gwas

GeneAtlas2004.gct.gz  # Gene Atlas 2004 Affymetrix expression matrix
ImmGen2012.gct.gz     # ImmGen 2012 Affymetrix expression matrix
FANTOM2014.gct.gz     # FANTOM5 2014 CAGE matrix
GO2013.gct.gz         # Gene Ontology 2013 binary annotation matrix

NCBIgenes2013.bed.gz  # NCBI gene intervals
Lango2010.txt.gz      # LD-pruned SNPs
TGP2011.bed.gz        # 1000 Genomes Project SNP linkage intervals
```

5.1 SNP sets

Phenotype	SNPs	Loci	Reference
Celiac disease	35	34	Table 2 (Trynka, et al. 2011)
HDL cholesterol	46	46	Supp. Table 2 (Teslovich, et al. 2010)
Multiple sclerosis	51	47	Supp. Table A (IMSGC WTCCC 2011)
Red blood cell count	45	45	Table 1 (Harst et al. 2012)

5.1.1 Celiac_disease-Trynka2011-35_SNPs.gwas

35 SNPs associated with Celiac disease taken from Table 2. Positions are on hg19. All SNPs have $P \leq 5e-8$.

> Trynka G, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011;43(12):1193-201.

5.1.2 HDL_cholesterol-Teslovich2010-46_SNPs.gwas

46 SNPs associated with HDL taken from Supplementary Table 2. Positions are on hg19. All SNPs have $P \leq 5e-8$.

> Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707-13.

5.1.3 Multiple_sclerosis-IMSGC-51_SNPs.gwas

51 SNPs associated with Multiple Sclerosis taken from Supplementary Table A. Positions are on hg19. All SNPs have $P \leq 5e-8$.

> Sawcer S, Hellenthal G, Pirinen M, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476(7359):214-9.

5.1.4 Red_blood_cell_count-Harst2012-45_SNPs.gwas

45 SNPs associated with red blood cell count (RBC) taken from Table 1. Positions are on hg19. All SNPs have $P \leq 5e-8$.

> van der Harst P, Zhang W, Mateo leach I, et al. Seventy-five genetic loci influencing the human red blood cell. Nature. 2012;492(7429):369-75.

5.2 Gene matrices

Type	Genes	Conditions	Species	Reference
Affy	17581	79 tissues	homo sapiens	GeneAtlas 2004
Affy	15139	249 cells	mus musculus	ImmGen 2012
CAGE	18502	533 cells	homo sapiens	FANTOM5 2014
Binary	19111	1751 terms	homo sapiens, mus musculus	Gene Ontology 2013, Homologene

5.2.1 GeneAtlas2004.gct.gz

Gene expression data for 79 human tissues from GSE1133. 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.

> Su AI et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A, 2004 Apr 9;101(16):6062-7.

5.2.2 GO2013.gct.gz

A GCT formatted gene matrix with 1,751 annotation terms (1s and 0s indicating presence or absence of the gene in a Gene Ontology term).

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

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. 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.

5.2.3 ImmGen2012.gct.gz

Gene expression data for 249 blood cell types from GSE15907. We averaged cell type replicates. For each gene, we selected the single probe with the largest minimum.

5.2.4 FANTOM2014.gct.gz

CAGE data for 533 human cell types from FANTOM5. We averaged cell type replicates. We discarded CAGE entries with 0 or multiple corresponding NCBI Entrez IDs. Then, we summed the CAGE entries for each gene.

5.3 LD-pruned SNPs and Genomic Intervals

5.3.1 Lango2010.txt.gz

A list of SNPs that span the whole genome, pruned by linkage disequilibrium (LD). SNPsea samples null SNP sets matched on the number of genes in the user's SNP set from this list. See this paper for more information:

> Lango allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832-8.

5.3.2 NCBIgenes2013.bed.gz

All human start and stop positions taken from:

> ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2refseq.gz

5.3.3 TGP2011.bed.gz

Linkage intervals for a filtered set of SNPs from the 1000 Genomes Project Phase 1 (May 21, 2011). 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. The linkage intervals were extended to the nearest HapMap recombination hotspot with >3 cM/Mb recombination rate (*Supplementary Figure 1*).

CHAPTER 6

Usage

Here is a Bash script with a usage example:

```
options=(
    --snps              Red_blood_cell_count-Harst2012-45_SNPs.gwas
    --gene-matrix       GeneAtlas2004.gct.gz
    --gene-intervals    NCBIgenes2013.bed.gz
    --snp-intervals     TGP2011.bed.gz
    --null-snps         Lango2010.txt.gz
    --out               out
    --slop              10e3
    --threads           4
    --null-snpsets      0
    --min-observations  100
    --max-iterations    1e7
)
snpsea ${options[*]}
```

SNPsea will test SNPs associated with Red blood cell count for tissue-specific expression of linked genes across 79 human tissues in the Gene Atlas expression matrix. Each tissue will be tested up to 10 million times with matched random SNP sets, or testing will stop for a tissue if 100 matched SNP sets achieve a higher specificity score than the user's SNPs.

6.1 Options

All input files may optionally be compressed with `'gzip <http://www.gzip.org/>'__.

6.1.1 Required

```
--snps ARG                Text file with SNP identifiers in the first
                          column. Instead of a file name, you may use
                          'randomN' with an integer N for a random SNP list
                          of length N.

--gene-matrix ARG         Gene matrix file in GCT format. The Name column
                          must contain the same gene identifiers as in
                          --gene-intervals.

--gene-intervals ARG      BED file with gene intervals. The fourth column
                          must contain the same gene identifiers as in
```

```
                        --gene-matrix.

--snp-intervals ARG     BED file with all known SNP intervals. The fourth
                        column must contain the same SNP identifiers as
                        in --snps and --null-snps.

--null-snps ARG         Text file with names of SNPs to sample when
                        generating null matched or random SNP sets.
                        These SNPs must be a subset of --snp-intervals.

--out ARG               Create output files in this directory. It will be
                        created if it does not already exist.
```

6.1.2 Optional

```
--condition ARG         Text file with a list of columns in --gene-matrix
                        to condition on before calculating p-values. Each
                        column in --gene-matrix is projected onto each
                        column listed in this file and its projection is
                        subtracted.

--slop ARG              If a SNP interval overlaps no gene intervals,
                        extend the SNP interval this many nucleotides
                        further and try again.
                        [default: 10000]

--threads ARG           Number of threads to use.
                        [default: 1]

--null-snpsets ARG      Test this many null matched SNP sets, so you can
                        compare your results to a distribution of null
                        results.
                        [default: 0]

--min-observations ARG  Stop testing a column in --gene-matrix after
                        observing this many null SNP sets with
                        specificity scores greater or equal to those
                        obtained with the SNP set in --snps. Increase
                        this value to obtain more accurate p-values.
                        [default: 25]

--max-iterations ARG    Maximum number of null SNP sets tested for each
                        column in --gene-matrix. Increase this value to
                        resolve smaller p-values.
                        [default: 10000]
```

6.2 Input File Formats

6.2.1 `--snps ARG`

You must provide one or more comma-separated text files. SNP identifiers must be listed one per line. Lines starting with # are skipped. If the file has no header, the first column is assumed to contain SNP identifiers. Otherwise, SNPsea looks for a column named (case-sensitive) `SNP` or `snp` or `name` or `marker`.

```
head Red_blood_cell_count-Harst2012-45_SNPs.gwas

# Harst et al. 2012
# doi:10.1038/nature11677
# PMID: 23222517
# 45 SNPs associated with red blood cell count (RBC) taken from Table 1.
# Positions are on hg19. SNPs are included if $P \le 5e-8$.
CHR POS SNP P
chr1    40069939    rs3916164   3e-10
chr1    158575729   rs857684    4e-16
chr1    199007208   rs7529925   8e-09
chr1    248039451   rs3811444   5e-10
```

Instead of providing a file with SNPs, you may use "randomN" like this:

```
--snps random20
```

to sample 20 random SNPs from the **"–snp-intervals"** file.

6.2.2 --gene-matrix ARG

You must provide a single gene matrix that must be in GCT format.

```
zcat GeneAtlas2004.gct.gz | cut -f1-4 | head

#1.2
17581  79
Name   Description  Colorectal_Adenocarcinoma  Whole_Blood
1      A1BG         115.5                      209.5
2      A2M          85                         328.5
9      NAT1         499                        1578
10     NAT2         115                        114
12     SERPINA3     419.5                      387.5
13     AADAC        125                        252.5
14     AAMP         2023                       942.5
```

6.2.3 --condition ARG (Optional)

You may provide column names present in the **"–gene-matrix"** file, one per line. The matrix will be conditioned on these columns before the analysis is performed to help you identify secondary signals independent of these columns. Binary (0, 1) matrices will not be conditioned.

```
head conditions.txt

Whole_Blood
```

6.2.4 --gene-intervals ARG

You must provide gene intervals in BED format with a fourth column that contains the same gene identifiers as those present in the Name column of the **"–gene-matrix"** GCT file. Only the first four columns are used.

```
zcat NCBIgenes2013.bed.gz | head

chr1  10003485   10045555   64802      NMNAT1
chr1  100111430  100160096  54873      PALMD
```

```
chr1  100163795  100164756  100129320  HMGB3P10
chr1  100174205  100232185  391059     FRRS1
chr1  10027438   10027515   100847055  MIR5697
chr1  100308165  100308317  100270894  RPL39P9
chr1  100315632  100389578  178        AGL
chr1  100433941  100435837  730081     LOC730081
chr1  100435344  100492534  23443      SLC35A3
chr1  100503669  100548932  64645      HIAT1
```

6.2.5 `--snp-intervals ARG`

SNP linkage intervals must be specified in BED format and include a fourth column with the SNP identifiers. The linkage intervals assigned to the trait-associated SNPs you provide with **"–snps"** are taken from this file.

```
zcat TGP2011.bed.gz | head

chr1    0   254996  rs113759966
chr1    0   254996  rs114420996
chr1    0   254996  rs114608975
chr1    0   254996  rs115209712
chr1    0   254996  rs116400033
chr1    0   254996  rs116504101
chr1    0   254996  rs12184306
chr1    0   254996  rs12184307
chr1    0   254996  rs138808727
chr1    0   254996  rs139113303
```

6.2.6 `--null-snps ARG`

The null SNPs file must have one SNP identifier per line. Only the first column is used. The identifiers must be a subset of the identifiers in **"–snp-intervals"**.

```
zcat Lango2010.txt.gz | head

rs58108140  chr1    10583
rs180734498 chr1    13302
rs140337953 chr1    30923
rs141149254 chr1    54490
rs2462492   chr1    54676
rs10399749  chr1    55299
rs189727433 chr1    57952
rs149755937 chr1    59040
rs77573425  chr1    61989
rs116440577 chr1    63671
```

6.3 Output Files

The usage example shown above produces the following output files:

```
out/
    args.txt
    condition_pvalues.txt
    null_pvalues.txt
```

```
    snp_condition_scores.txt
    snp_genes.txt
```

6.3.1 args.txt

The command line arguments needed to reproduce the analysis.

```
cat args.txt

# SNPsea v1.0.2
--snps              Red_blood_cell_count-Harst2012-45_SNPs.gwas
--gene-matrix       GeneAtlas2004.gct.gz
--gene-intervals    NCBIgenes2013.bed.gz
--snp-intervals     TGP2011.bed.gz
--null-snps         Lango2010.txt.gz
--out               out
--score             single
--slop              100000
--threads           8
--null-snpsets      0
--min-observations  100
--max-iterations    10000000
```

Repeat the analysis:

```
snpsea --args args.txt
```

6.3.2 condition_pvalues.txt

The p-values representing enrichment of condition-specificity for the given SNPs.

```
head condition_pvalues.txt | column -t

condition                  pvalue     nulls_observed  nulls_tested
Colorectal_Adenocarcinoma  0.933555   280             300
Whole_Blood                0.521595   156             300
BM-CD33+Myeloid            0.159772   111             700
PB-CD14+Monocytes          0.103264   154             1500
PB-BDCA4+Dentritic_cells   0.0606256  187             3100
PB-CD56+NK_cells           0.194009   135             700
PB-CD4+T_cells             0.428571   128             300
PB-CD8+T_cells             0.531561   159             300
PB-CD19+B_cells            0.226819   158             700
```

6.3.3 null_pvalues.txt

If the argument for **"–snps"** is the name of a file, the p-values for null matched SNP sets. You can compare these null results to the results for your trait-associated SNPs.

If the argument for **"–snps"** is "randomN" where N is some integer, like "random20" the p-values for random unmatched SNP sets, each with N SNPs.

The fifth column is the replicate index. The number of replicates performed is specified with **"–null-snpsets INT"**.

```
head null_pvalues.txt | column -t

ColorectalAdenocarcinoma  0.056     84   1500  0
WholeBlood                0.236667  71   300   0
BM-CD33+Myeloid           0.55      55   100   0
PB-CD14+Monocytes         0.59      59   100   0
PB-BDCA4+Dentritic_Cells  0.59      59   100   0
PB-CD56+NKCells           0.71      71   100   0
PB-CD4+Tcells             0.383333  115  300   0
PB-CD8+Tcells             0.128571  90   700   0
PB-CD19+Bcells            0.168571  118  700   0
BM-CD105+Endothelial      0.386667  116  300   0
```

6.3.4 snp_genes.txt

Each SNP's linkage interval and overlapping genes. If a SNP is not found in the reference file specified with **"–snp-intervals"**, then the name of the SNP will be listed and the other columns will contain NA.

```
head snp_genes.txt | column -t

chrom  start      end        snp         n_genes  genes
chr4   55364224   55408999   rs218238    0        NA
chr6   139827777  139844854  rs590856    0        NA
NA     NA         NA         rs99999999  NA       NA
chr6   109505894  109651220  rs1008084   2        8763,27244
chr10  71089843   71131638   rs10159477  1        3098
chr2   111807303  111856057  rs10207392  1        55289
chr16  88831494   88903796   rs10445033  4        353,2588,9780,81620
chr7   151396253  151417368  rs10480300  1        51422
chr12  4320955    4336783    rs10849023  2        894,57103
chr15  76129642   76397903   rs11072566  4        26263,92912,123591,145957
```

6.3.5 snp_condition_scores.txt

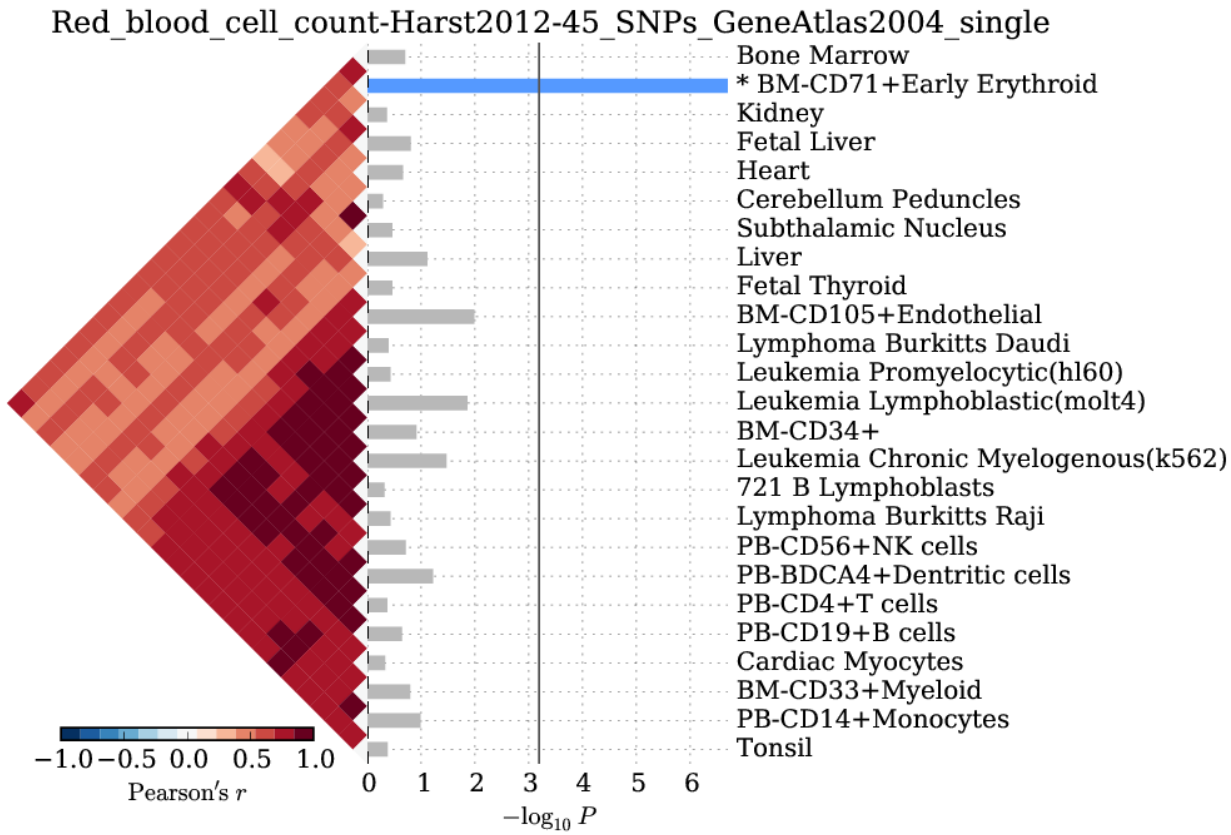
Each SNP, condition, gene with greatest specificity to that condition, and score for the SNP-condition pair, adjusted for the number of genes overlapping the given SNP's linkage interval.

```
head snp_condition_scores.txt | column -t

snp        condition                  gene   score
rs9349204  Colorectal_Adenocarcinoma   10817  0.693027
rs9349204  Whole_Blood                896    0.285864
rs9349204  BM-CD33+Myeloid            896    0.236487
rs9349204  PB-CD14+Monocytes          29964  0.340561
rs9349204  PB-BDCA4+Dentritic_cells    29964  0.411727
rs9349204  PB-CD56+NK_cells           896    0.0356897
rs9349204  PB-CD4+T_cells             896    0.38182
rs9349204  PB-CD8+T_cells             896    0.332008
rs9349204  PB-CD19+B_cells            29964  0.255196
```

Output Visualizations

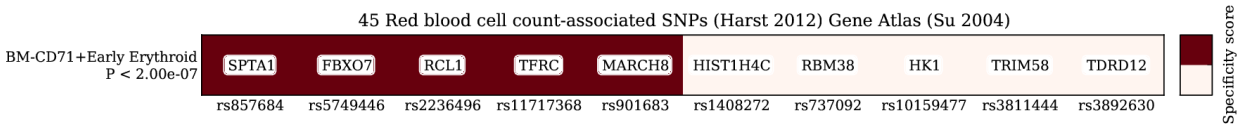
7.1 View enrichment of tissue-specific gene expression



A horizontal bar plot of negative log10 p-values for a test of 45 red blood cell count-associated SNPs for enrichment of tissue-specific expression in profiles of 79 human tissues and cells.

```
python bin/snpsea-barplot out
```

7.2 View the most specifically expressed gene for each SNP-tissue pair



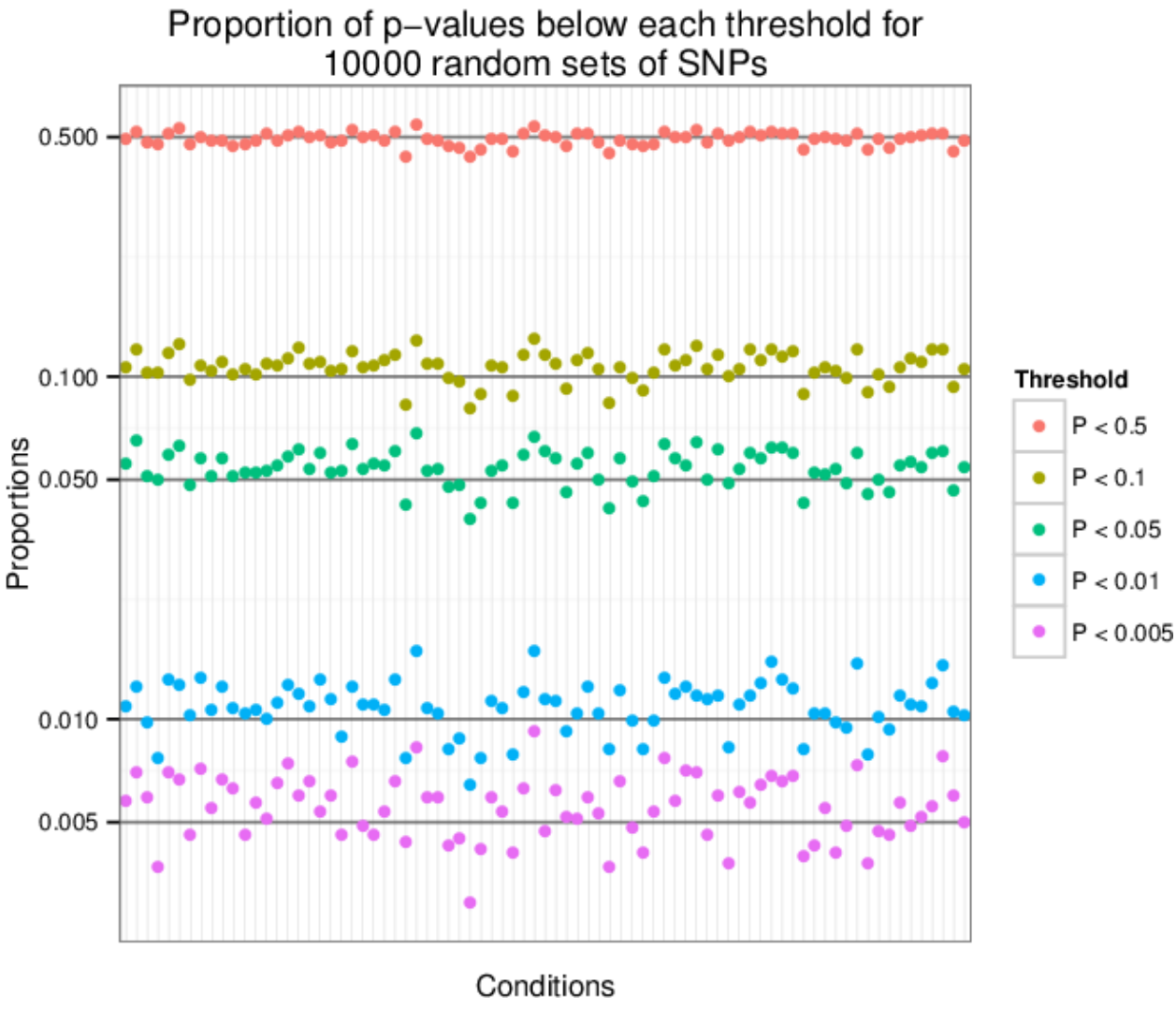
A heatmap exposing the contributions of specifically expressed genes within each SNP linkage interval to the specificity scores of each tissue.

```
python bin/snpsea-heatmap out
```

7.3 View the type 1 error rate estimates for each tissue

A scatter plot of the observed proportion of p-values under various thresholds after repeating the analysis with 10,000 random SNP sets.

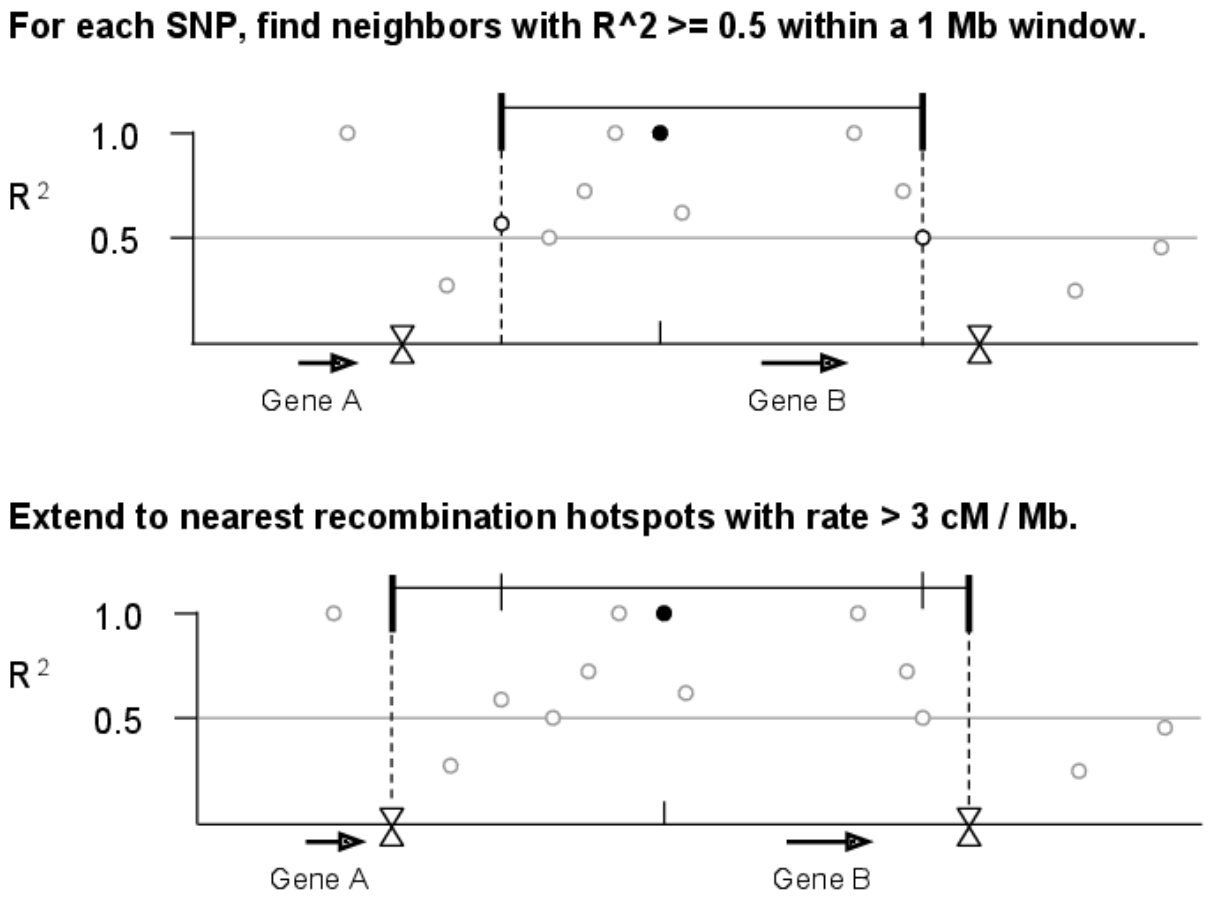
```
Rscript bin/snpsea-type1error out
```


Proportion of p−values below each threshold for
10000 random sets of SNPs
0.500
0.100
0.050
0.010
0.005
Proportions
Conditions
Threshold
P < 0.5
P < 0.1
P < 0.05
P < 0.01
P < 0.005

CHAPTER 8

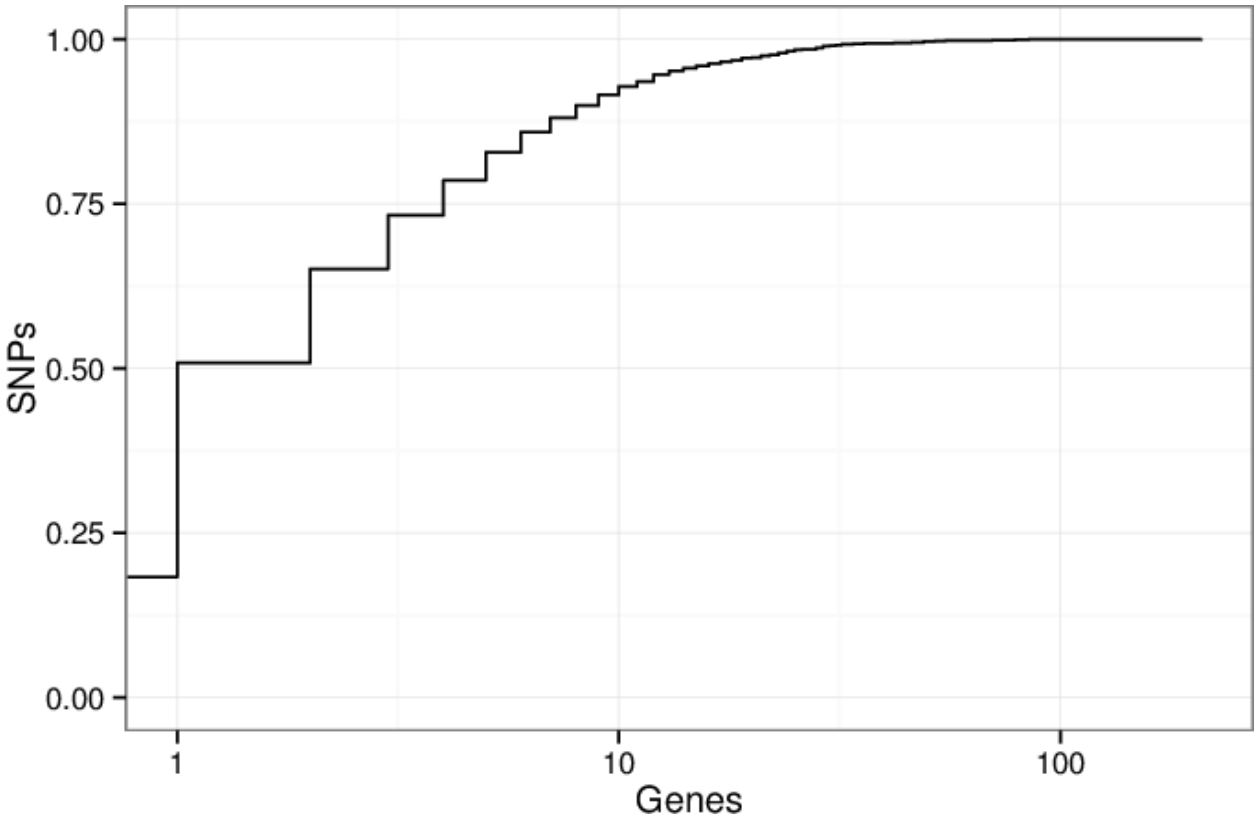
Supplementary Figures

8.1 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 (1000 Genomes Consortium 2012) 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 (Myers *et al.* 2005) with recombination rate >3 cM/Mb.

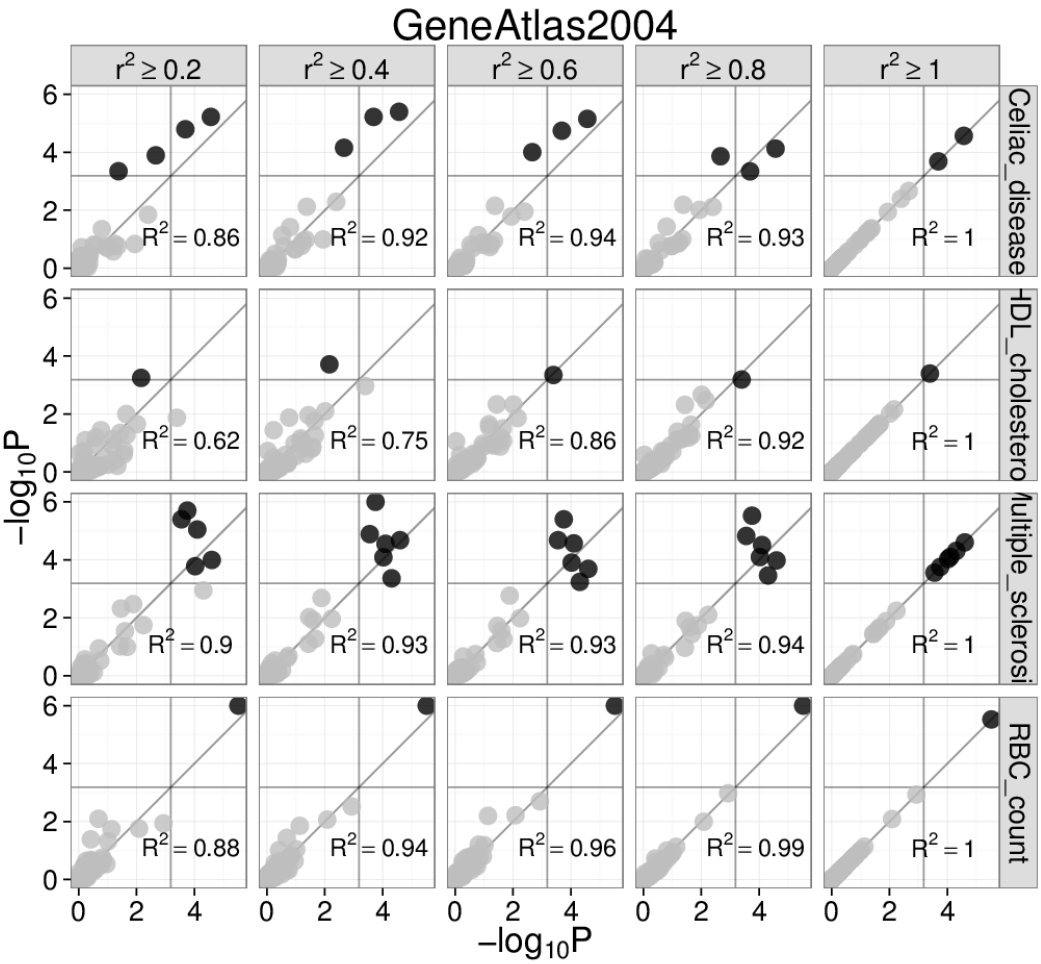
8.2 Supplementary Figure 2: Counting genes in GWAS SNP linkage intervals

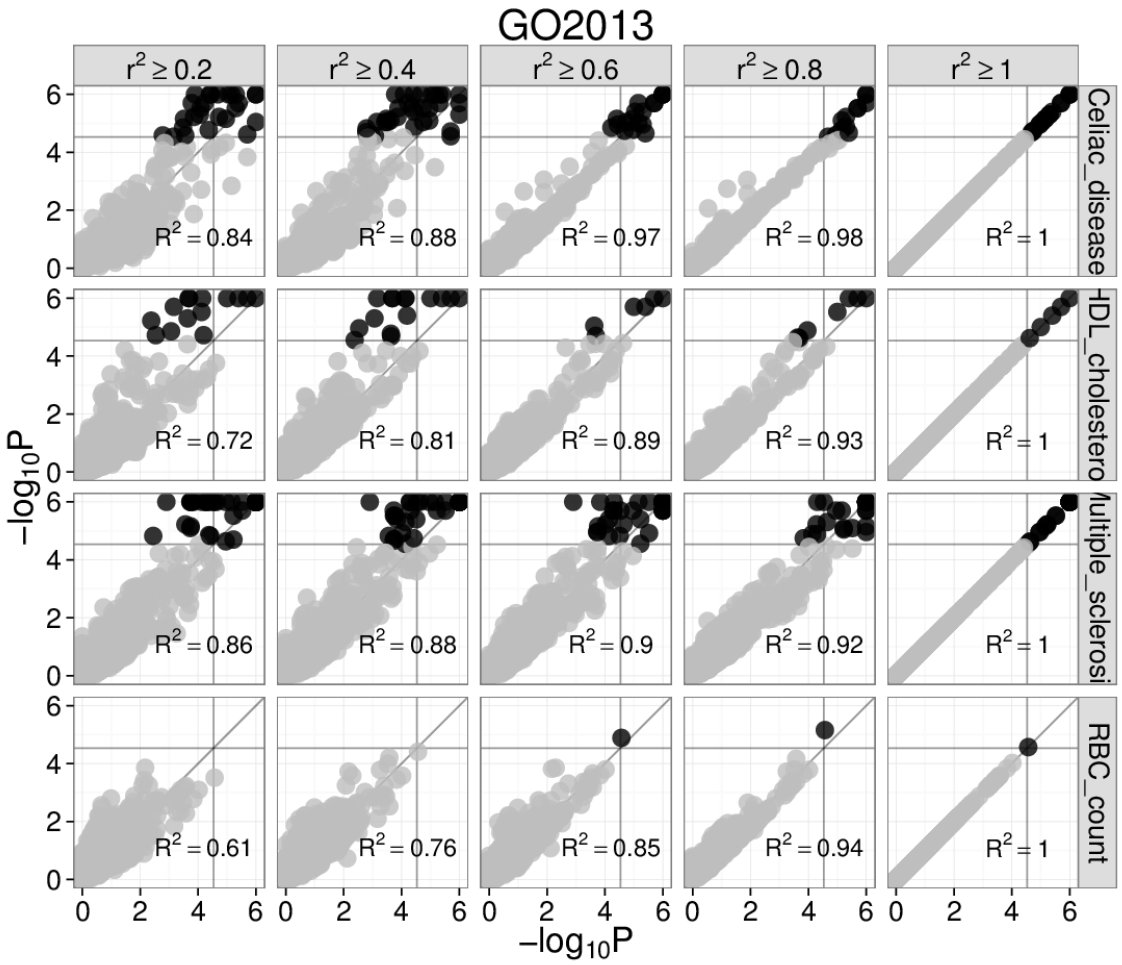


A cumulative density plot of the number of genes overlapped by the linkage intervals of GWAS SNPs. We downloaded the GWAS Catalog SNPs on January 17, 2014 and selected the 11,561 SNPs present in the 1000 Genomes Project (1000 Genomes Consortium 2012). Of these SNPs, 2,119 (18%) of them have linkage disequilibrium (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.

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

We chose to use $r^2 \geq 0.5$ due to previous experience (Rossin *et al.* 2011). To investigate if this choice influences SNPsea results, we repeated the analysis of 45 red blood cell count-associated SNPs (Van der Harst *et al.* 2012) using 5 different thresholds ($r^2 \geq 0.2$, 0.4, 0.6, 0.8, 1.0). We also did this for SNPs associated with multiple sclerosis, celiac disease, and HDL cholesterol.


GeneAtlas2004
r² ≥ 0.2
r² ≥ 0.4
r² ≥ 0.6
r² ≥ 0.8
r² ≥ 1
Celiac_disease
HDL_cholestero
Multiple_sclerosi
RBC_count
R² = 0.86
R² = 0.92
R² = 0.94
R² = 0.93
R² = 1
R² = 0.62
R² = 0.75
R² = 0.86
R² = 0.92
R² = 1
R² = 0.9
R² = 0.93
R² = 0.93
R² = 0.94
R² = 1
R² = 0.88
R² = 0.94
R² = 0.96
R² = 0.99
R² = 1
−log10P
−log10P



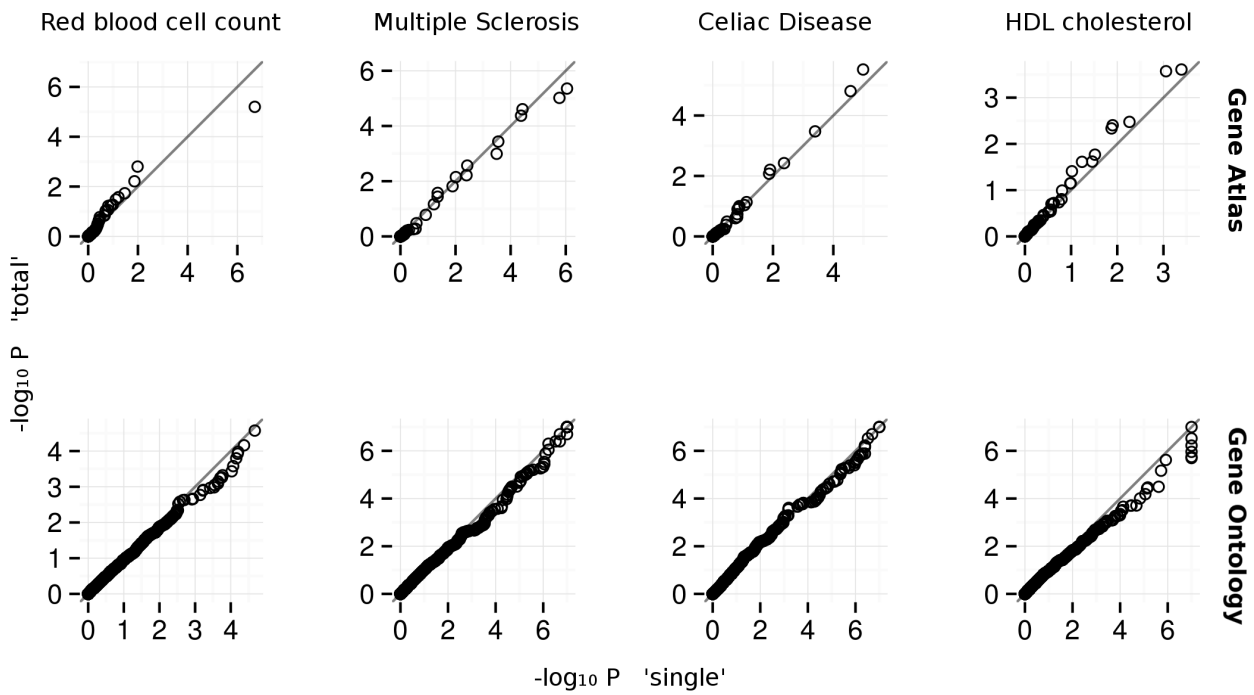
Gene Atlas and Gene Ontology (left and right). Each subplot has $-\log_{10}P$ for $r^2 = 1$ on the x-axis and $-\log_{10}P$ on the y-axis for the r^2 threshold marked above. Grey lines are significance thresholds after correction testing multiple conditions (cell types, GO annotations). Black points 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).

8.4 Supplementary Figure 4: Each trait-associated locus harbors a single associated gene

Quantile-quantile plots for Gene Atlas (Su *et al.* 2004) 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 . The P-values appear similar between methods.

8.5 Supplementary Figure 5: Type 1 error estimates

We sampled 10,000 sets of 100 SNPs uniformly from a list of LD-pruned SNPs (Lango Allen *et al.* 2010). We tested each of the 10,000 sets for enrichment of tissue-specific expression in the Gene Atlas (Su *et al.* 2004) 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.

8.5.1 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. The adjacent heatmaps show Pearson correlation coefficients for all pairs of conditions.

Phenotype	SNPs	Loci	Reference
Multiple sclerosis	51	47	Supp. Table A (IMSGC WTCCC 2011)
Celiac disease	35	34	Table 2 (Trynka, *et al.* 2011)
HDL cholesterol	46	46	Supp. Table 2 (Teslovich, *et al.* 2010)

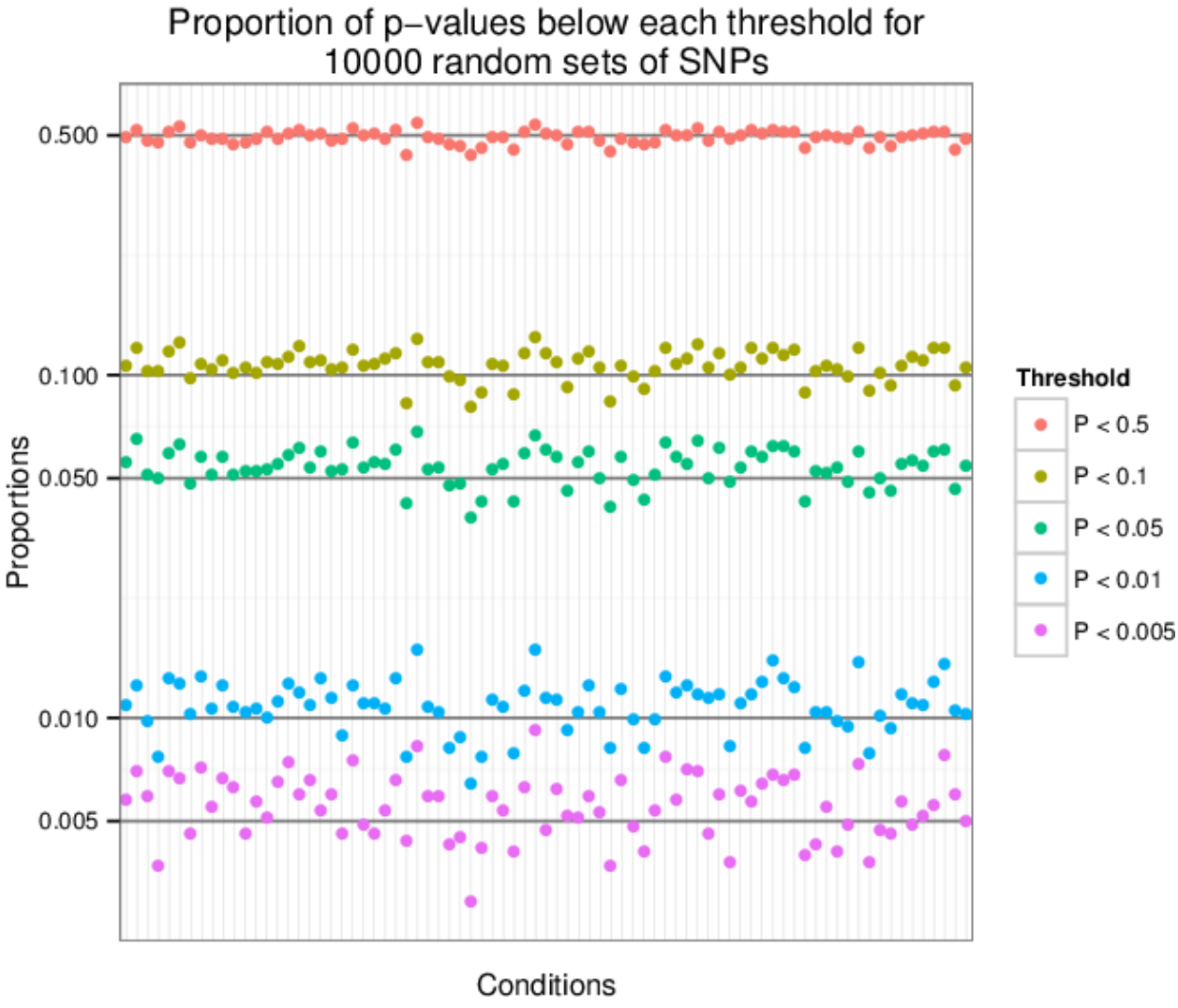
8.6 Supplementary Figure 6: Red blood cell count GO enrichment

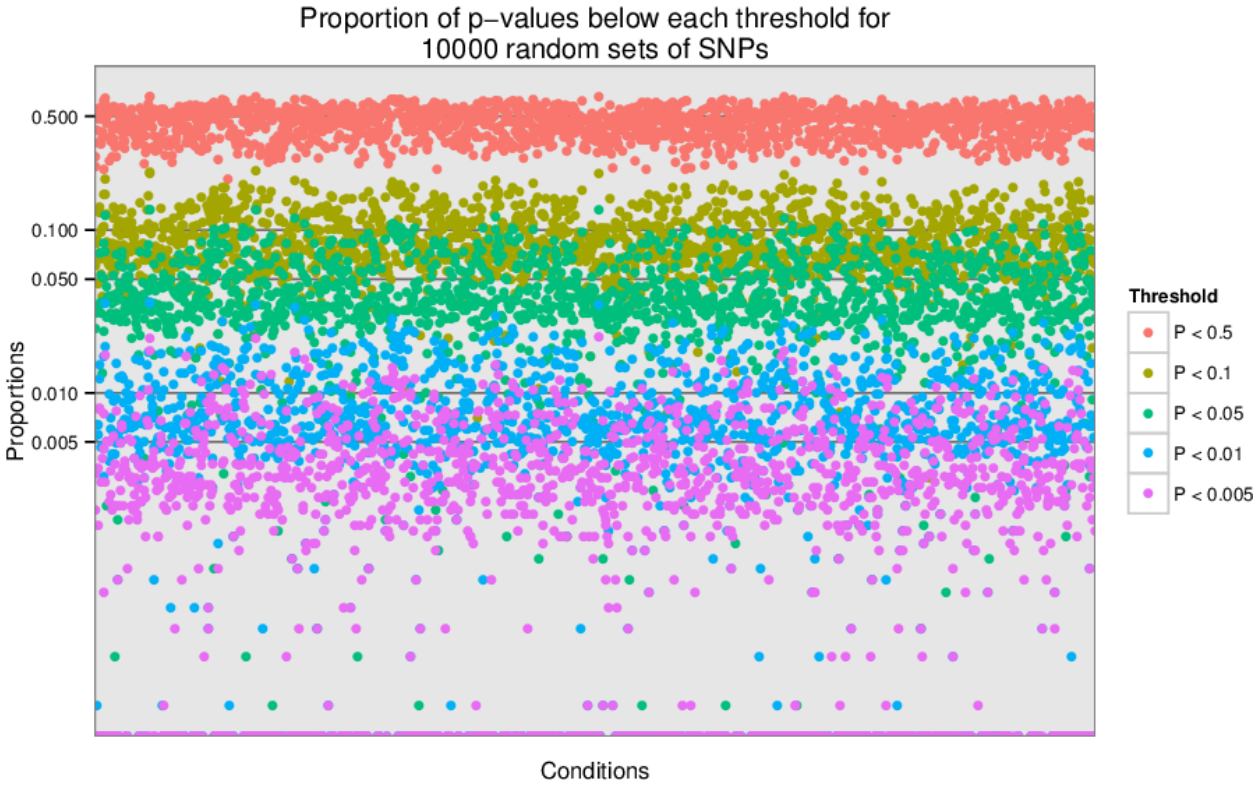
We observed significant enrichment for *hemopoiesis* (2×10^{-5}). The top 25 terms are shown.

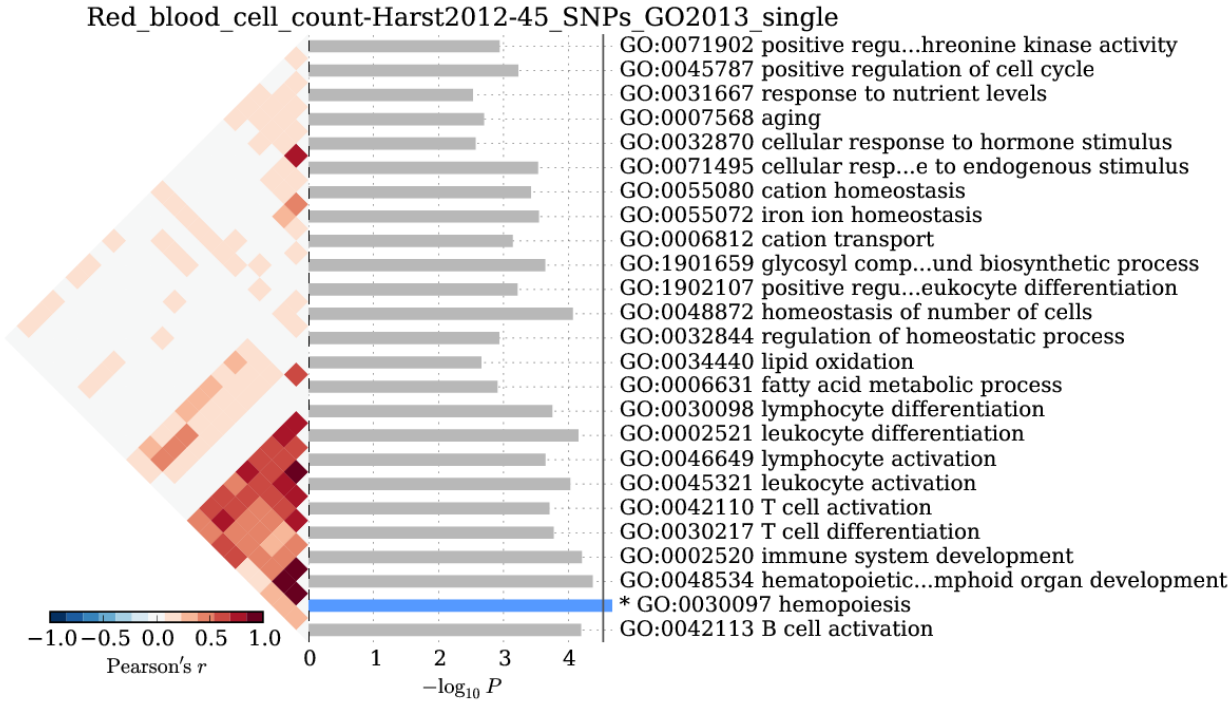
8.7 Supplementary Figure 7: Multiple sclerosis

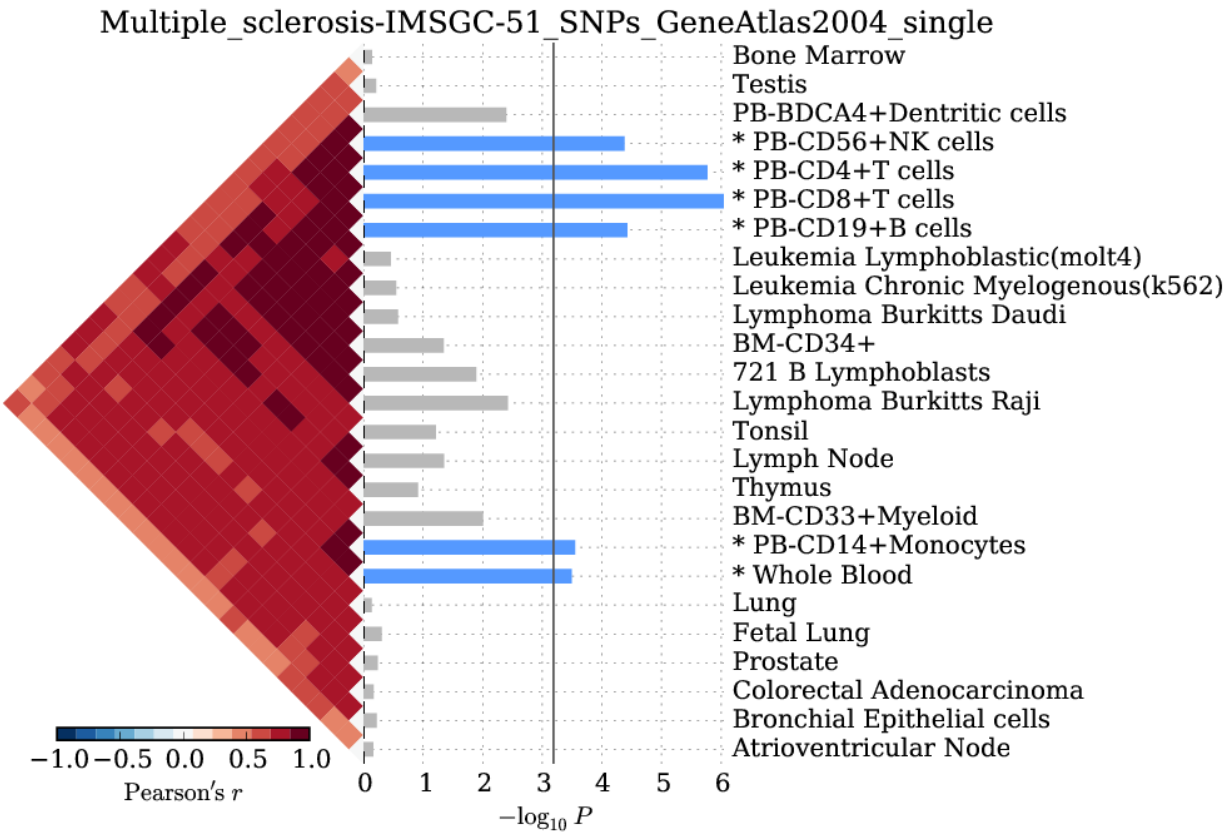
We observed significant enrichment for 6 cell types. The top 25 of 79 are shown.

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


Proportion of p−values below each threshold for
10000 random sets of SNPs
0.500
0.100
0.050
0.010
0.005
Proportions
Conditions
Threshold
P < 0.5
P < 0.1
P < 0.05
P < 0.01
P < 0.005


Proportion of p−values below each threshold for
10000 random sets of SNPs
0.500
0.100
0.050
0.010
0.005
Proportions
Conditions
Threshold
P < 0.5
P < 0.1
P < 0.05
P < 0.01
P < 0.005


Red_blood_cell_count-Harst2012-45_SNPs_GO2013_single
GO:0071902 positive regu...hreonine kinase activity
GO:0045787 positive regulation of cell cycle
GO:0031667 response to nutrient levels
GO:0007568 aging
GO:0032870 cellular response to hormone stimulus
GO:0071495 cellular resp...e to endogenous stimulus
GO:0055080 cation homeostasis
GO:0055072 iron ion homeostasis
GO:0006812 cation transport
GO:1901659 glycosyl comp...und biosynthetic process
GO:1902107 positive regu...eukocyte differentiation
GO:0048872 homeostasis of number of cells
GO:0032844 regulation of homeostatic process
GO:0034440 lipid oxidation
GO:0006631 fatty acid metabolic process
GO:0030098 lymphocyte differentiation
GO:0002521 leukocyte differentiation
GO:0046649 lymphocyte activation
GO:0045321 leukocyte activation
GO:0042110 T cell activation
GO:0030217 T cell differentiation
GO:0002520 immune system development
GO:0048534 hematopoietic...mphoid organ development
* GO:0030097 hemopoiesis
GO:0042113 B cell activation
−1.0 −0.5 0.0 0.5 1.0
Pearson's r
0 1 2 3 4
−log10 P



8.8 Supplementary Figure 8: Celiac disease

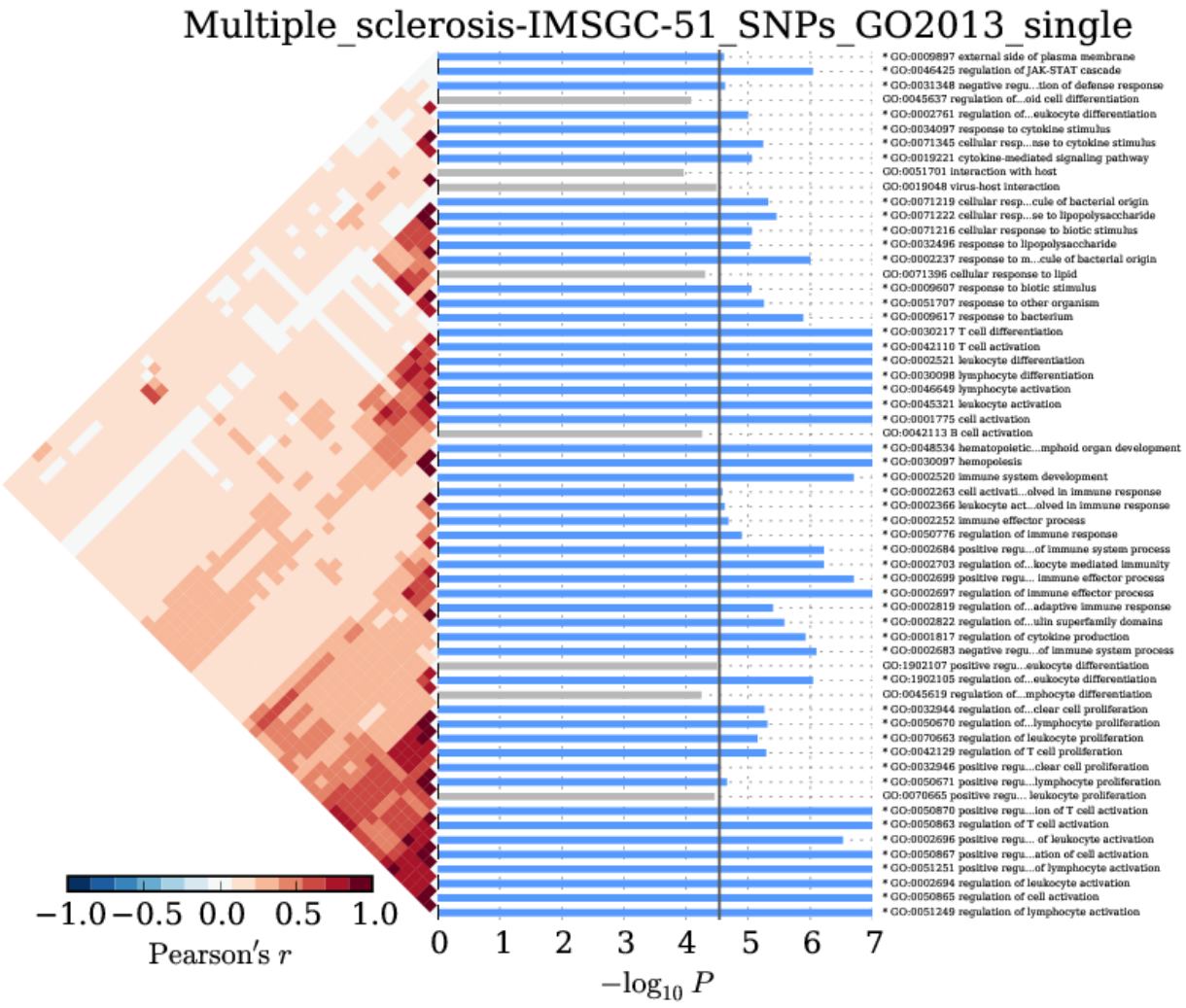
We observed significant enrichment for 3 cell types. The top 25 of 79 are shown.

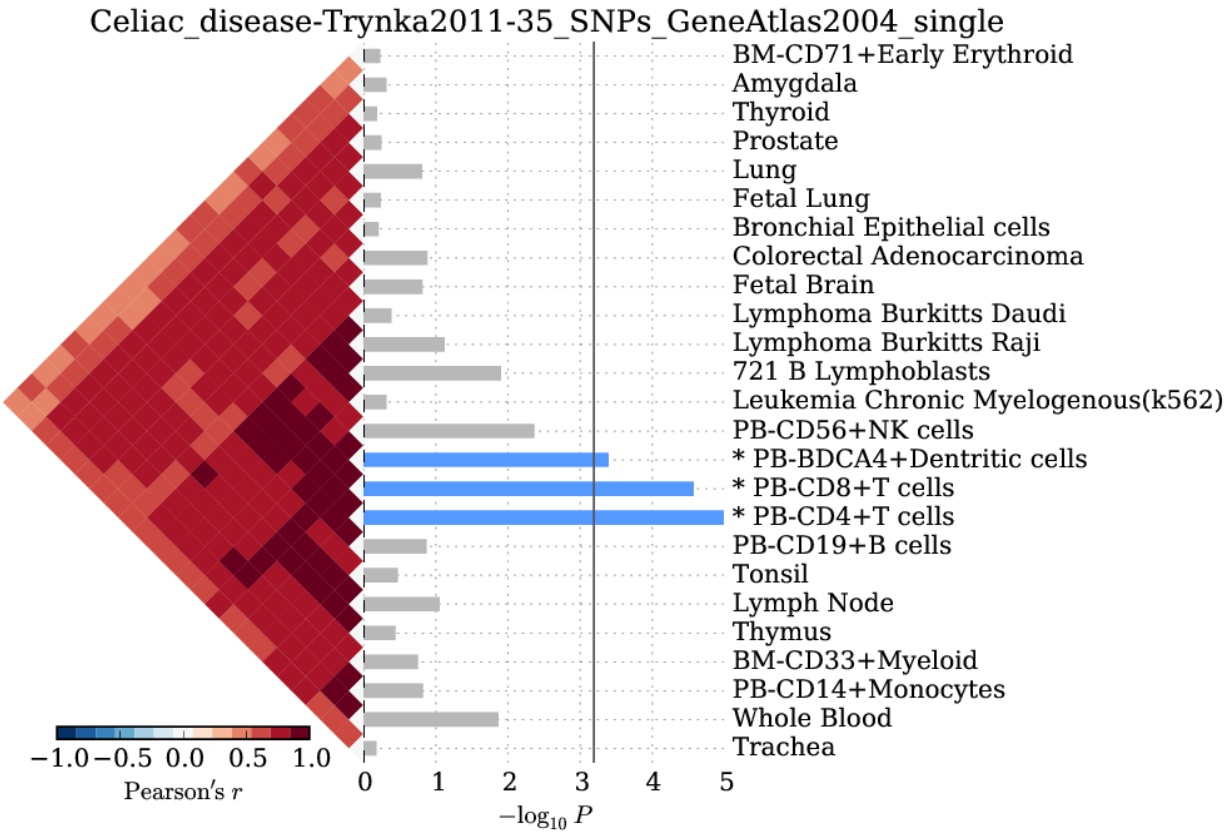
We observed significant enrichment for 28 Gene Ontology terms. The top 40 terms are shown.

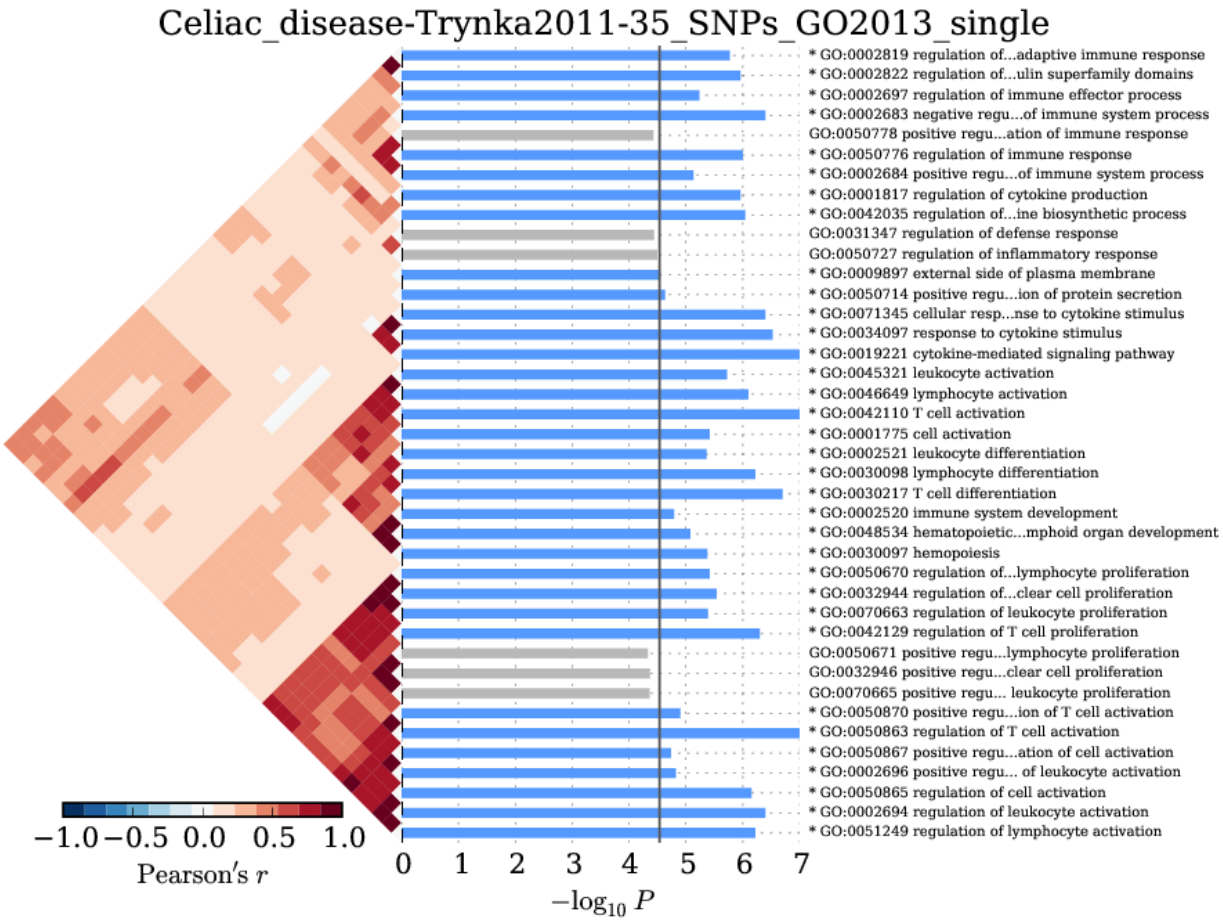
8.9 Supplementary Figure 9: HDL cholesterol

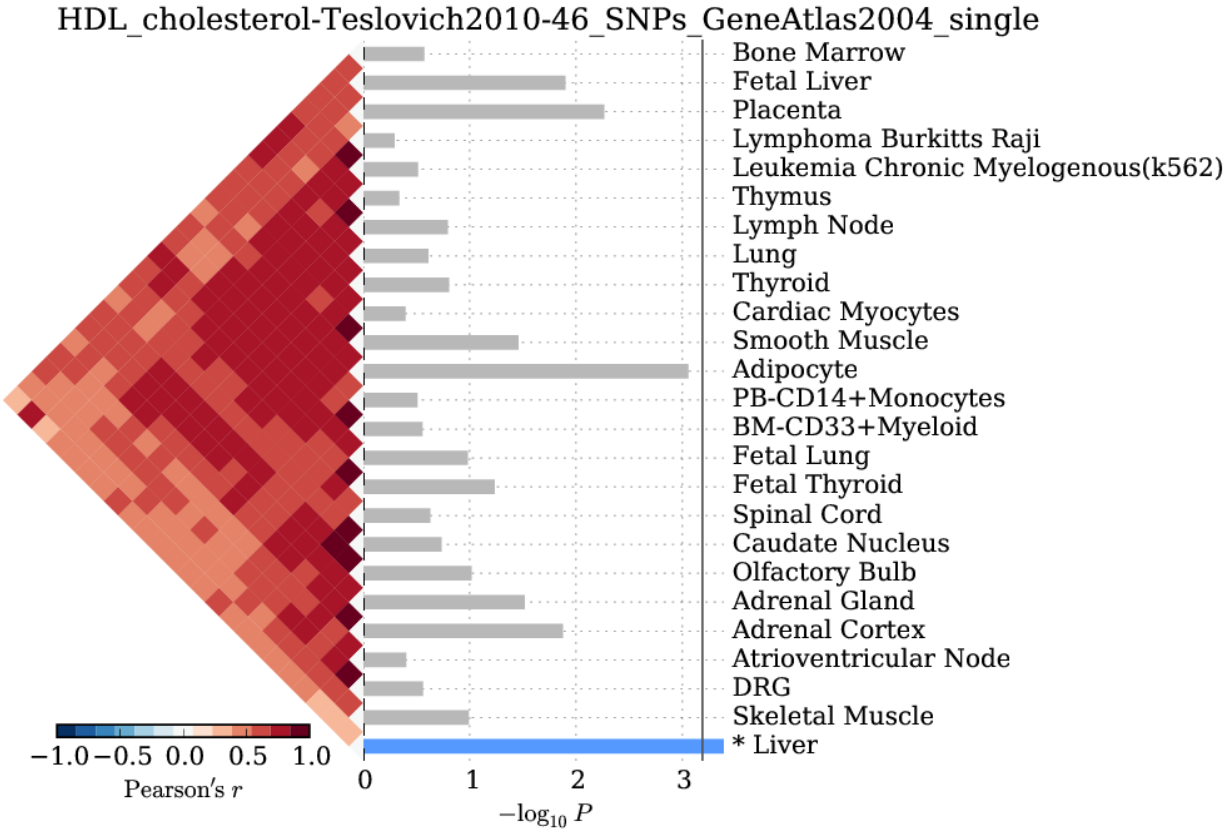
We observed significant enrichment for liver tissue-specific gene expression. The top 25 of 79 are shown.

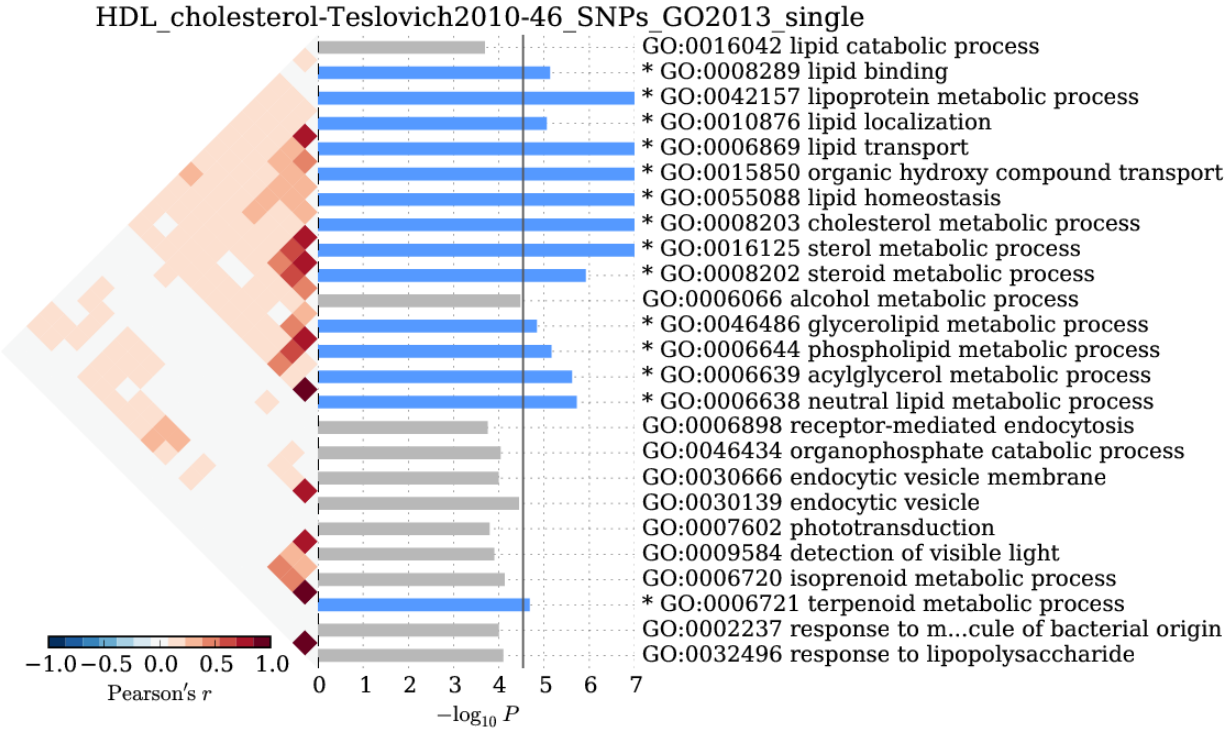
We observed significant enrichment for 13 Gene Ontology terms. The top 25 terms are shown.


Multiple_sclerosis-IMSGC-51_SNPs_GO2013_single
* GO:0009897 external side of plasma membrane
* GO:0046425 regulation of JAK-STAT cascade
* GO:0031348 negative regu...tion of defense response
GO:0045637 regulation of...oid cell differentiation
* GO:0002761 regulation of...eukocyte differentiation
* GO:0034097 response to cytokine stimulus
* GO:0071345 cellular resp...nse to cytokine stimulus
* GO:0019221 cytokine-mediated signaling pathway
GO:0051701 interaction with host
GO:0019048 virus-host interaction
* GO:0071219 cellular resp...cule of bacterial origin
* GO:0071222 cellular resp...se to lipopolysaccharide
* GO:0071216 cellular response to biotic stimulus
* GO:0032496 response to lipopolysaccharide
* GO:0002237 response to m...cule of bacterial origin
GO:0071396 cellular response to lipid
* GO:0009607 response to biotic stimulus
* GO:0051707 response to other organism
* GO:0009617 response to bacterium
* GO:0030217 T cell differentiation
* GO:0042110 T cell activation
* GO:0002521 leukocyte differentiation
* GO:0030098 lymphocyte differentiation
* GO:0046649 lymphocyte activation
* GO:0045321 leukocyte activation
* GO:0001775 cell activation
GO:0042113 B cell activation
* GO:0048534 hematopoietic...mphoid organ development
* GO:0030097 hemopoiesis
* GO:0002520 immune system development
* GO:0002263 cell activati...olved in immune response
* GO:0002366 leukocyte act...olved in immune response
* GO:0002252 immune effector process
* GO:0050776 regulation of immune response
* GO:0002684 positive regu...of immune system process
* GO:0002703 regulation of...kocyte mediated immunity
* GO:0002699 positive regu... immune effector process
* GO:0002697 regulation of immune effector process
* GO:0002819 regulation of...adaptive immune response
* GO:0002822 regulation of...ulin superfamily domains
* GO:0001817 regulation of cytokine production
* GO:0002683 negative regu...of immune system process
GO:1902107 positive regu...eukocyte differentiation
* GO:1902105 regulation of...eukocyte differentiation
GO:0045619 regulation of...mphocyte differentiation
* GO:0032944 regulation of...clear cell proliferation
* GO:0050670 regulation of...lymphocyte proliferation
* GO:0070663 regulation of leukocyte proliferation
* GO:0042129 regulation of T cell proliferation
* GO:0032946 positive regu...clear cell proliferation
* GO:0050671 positive regu...lymphocyte proliferation
GO:0070665 positive regu... leukocyte proliferation
* GO:0050870 positive regu...ion of T cell activation
* GO:0050863 regulation of T cell activation
* GO:0002696 positive regu... of leukocyte activation
* GO:0050867 positive regu...ation of cell activation
* GO:0051251 positive regu...of lymphocyte activation
* GO:0002694 regulation of leukocyte activation
* GO:0050865 regulation of cell activation
* GO:0051249 regulation of lymphocyte activation
−1.0 −0.5 0.0 0.5 1.0
Pearson's r
0 1 2 3 4 5 6 7
$-\log_{10} P$


Celiac_disease-Trynka2011-35_SNPs_GeneAtlas2004_single
BM-CD71+Early Erythroid
Amygdala
Thyroid
Prostate
Lung
Fetal Lung
Bronchial Epithelial cells
Colorectal Adenocarcinoma
Fetal Brain
Lymphoma Burkitts Daudi
Lymphoma Burkitts Raji
721 B Lymphoblasts
Leukemia Chronic Myelogenous(k562)
PB-CD56+NK cells
* PB-BDCA4+Dentritic cells
* PB-CD8+T cells
* PB-CD4+T cells
PB-CD19+B cells
Tonsil
Lymph Node
Thymus
BM-CD33+Myeloid
PB-CD14+Monocytes
Whole Blood
Trachea
−1.0 −0.5 0.0 0.5 1.0
Pearson's r
0 1 2 3 4 5
−log10 P


Celiac_disease-Trynka2011-35_SNPs_GO2013_single
* GO:0002819 regulation of...adaptive immune response
* GO:0002822 regulation of...ulin superfamily domains
* GO:0002697 regulation of immune effector process
* GO:0002683 negative regu...of immune system process
GO:0050778 positive regu...ation of immune response
* GO:0050776 regulation of immune response
* GO:0002684 positive regu...of immune system process
* GO:0001817 regulation of cytokine production
* GO:0042035 regulation of...ine biosynthetic process
GO:0031347 regulation of defense response
GO:0050727 regulation of inflammatory response
* GO:0009897 external side of plasma membrane
* GO:0050714 positive regu...ion of protein secretion
* GO:0071345 cellular resp...nse to cytokine stimulus
* GO:0034097 response to cytokine stimulus
* GO:0019221 cytokine-mediated signaling pathway
* GO:0045321 leukocyte activation
* GO:0046649 lymphocyte activation
* GO:0042110 T cell activation
* GO:0001775 cell activation
* GO:0002521 leukocyte differentiation
* GO:0030098 lymphocyte differentiation
* GO:0030217 T cell differentiation
* GO:0002520 immune system development
* GO:0048534 hematopoietic...mphoid organ development
* GO:0030097 hemopoiesis
* GO:0050670 regulation of...lymphocyte proliferation
* GO:0032944 regulation of...clear cell proliferation
* GO:0070663 regulation of leukocyte proliferation
* GO:0042129 regulation of T cell proliferation
GO:0050671 positive regu...lymphocyte proliferation
GO:0032946 positive regu...clear cell proliferation
GO:0070665 positive regu... leukocyte proliferation
* GO:0050870 positive regu...ion of T cell activation
* GO:0050863 regulation of T cell activation
* GO:0050867 positive regu...ation of cell activation
* GO:0002696 positive regu... of leukocyte activation
* GO:0050865 regulation of cell activation
* GO:0002694 regulation of leukocyte activation
* GO:0051249 regulation of lymphocyte activation
−1.0 −0.5 0.0 0.5 1.0
Pearson's r
0 1 2 3 4 5 6 7
−log10 P


HDL_cholesterol-Teslovich2010-46_SNPs_GeneAtlas2004_single
Bone Marrow
Fetal Liver
Placenta
Lymphoma Burkitts Raji
Leukemia Chronic Myelogenous(k562)
Thymus
Lymph Node
Lung
Thyroid
Cardiac Myocytes
Smooth Muscle
Adipocyte
PB-CD14+Monocytes
BM-CD33+Myeloid
Fetal Lung
Fetal Thyroid
Spinal Cord
Caudate Nucleus
Olfactory Bulb
Adrenal Gland
Adrenal Cortex
Atrioventricular Node
DRG
Skeletal Muscle
* Liver
−1.0 −0.5 0.0 0.5 1.0
Pearson's r
0 1 2 3
−log10 P


HDL_cholesterol-Teslovich2010-46_SNPs_GO2013_single
GO:0016042 lipid catabolic process
* GO:0008289 lipid binding
* GO:0042157 lipoprotein metabolic process
* GO:0010876 lipid localization
* GO:0006869 lipid transport
* GO:0015850 organic hydroxy compound transport
* GO:0055088 lipid homeostasis
* GO:0008203 cholesterol metabolic process
* GO:0016125 sterol metabolic process
* GO:0008202 steroid metabolic process
GO:0006066 alcohol metabolic process
* GO:0046486 glycerolipid metabolic process
* GO:0006644 phospholipid metabolic process
* GO:0006639 acylglycerol metabolic process
* GO:0006638 neutral lipid metabolic process
GO:0006898 receptor-mediated endocytosis
GO:0046434 organophosphate catabolic process
GO:0030666 endocytic vesicle membrane
GO:0030139 endocytic vesicle
GO:0007602 phototransduction
GO:0009584 detection of visible light
GO:0006720 isoprenoid metabolic process
* GO:0006721 terpenoid metabolic process
GO:0002237 response to m...cule of bacterial origin
GO:0032496 response to lipopolysaccharide
−1.0 −0.5 0.0 0.5 1.0
Pearson's r
0 1 2 3 4 5 6 7
−log10 P

CHAPTER 9

References

1. Elizabeth J. Rossin, Kasper Lage, Soumya Raychaudhuri, Ramnik J. Xavier, Diana Tatar, Yair Benita, Chris Cotsapas, Mark J. Daly, and International Inflammatory Bowel Disease Genetics Constortium. 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, Gil McVean, and Peter Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. Science, 310(5746):321–324, October 2005.

4. Xinli Hu, Hyun Kim, Eli Stahl, Robert Plenge, Mark Daly, and Soumya Raychaudhuri. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. The American Journal of Human Genetics, 89(4):496 506, 2011.

5. Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature, 467(7317):832– 838, October 2010.

6. 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.

7. Pim van der Harst, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S. Paul, Ulrich Elling, Hooman Allayee, Xinzhong Li, et al. Seventy-five genetic loci influencing the human red blood cell. Nature, 492(7429):369–375, December 2012.

8. Andrew I. Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, 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. The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Con- sortium. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature, 476(7359):214–219, August 2011.

10. Gosia Trynka, Karen A. Hunt, Nicholas A. Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F. Bakker, Maria Teresa Bardella, Leena Bhaw-Rosun, Gemma Castillejo, 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.

11. Tanya M. Teslovich, Kiran Musunuru, Albert V. Smith, Andrew C. Edmondson, Ioannis M. Stylianou, Masahiro Koseki, James P. Pirruccello, Samuli Ripatti, Daniel I. Chasman, Cristen J. Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature, 466(7307):707–713, August 2010.