SNPsea Reference Manual

Kamil Slowikowski

May 19, 2014

# 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](#data).

The columns of the matrix may be tissues, cell types, GO annotation codes, or other *conditions*. Continuous matrices *must* be normalized before running SNPsea: columns must be directly comparable to each other.

If you benefit from this method, please cite:

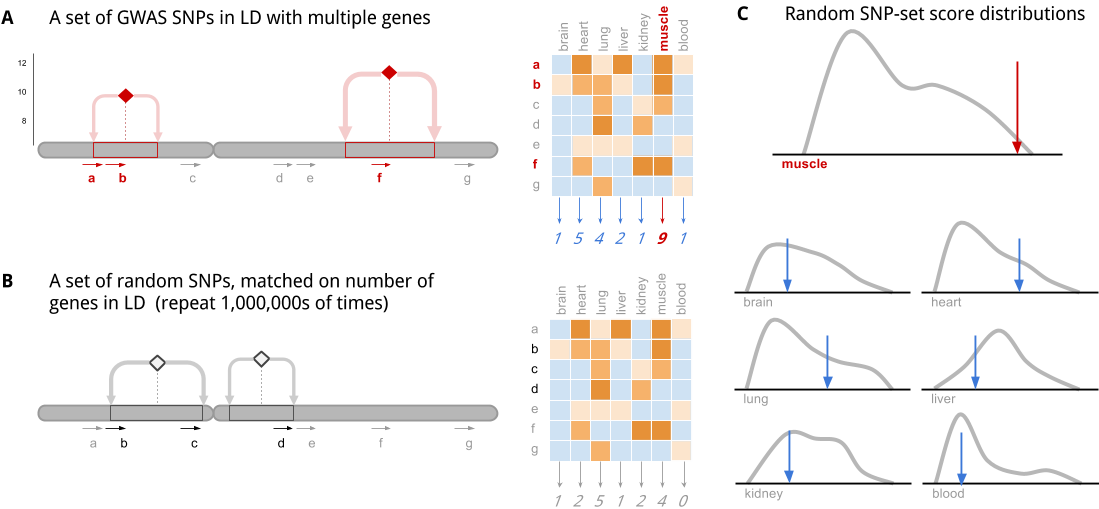
Slowikowski, K. et al. **SNPsea: an algorithm to identify cell types, tissues, and pathways affected by risk loci.** Bioinformatics (2014). doi:[10.1093/bioinformatics/btu326](http://bioinformatics.oxfordjournals.org/content/early/2014/05/10/bioinformatics.btu326)

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). [PubMed](http://www.ncbi.nlm.nih.gov/pubmed/21963258)

## Visual Summary

### Cartoon



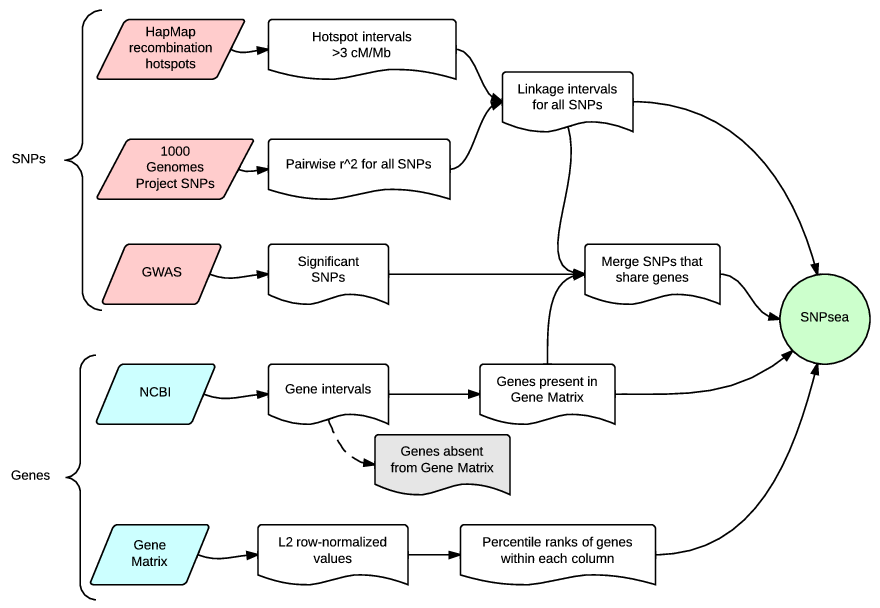
This cartoon illustrates the key ideas of the algorithm:

**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.

### Flow Chart



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

## 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 -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 Rossin *et al.* 2011). First, we define the linkage interval for a given SNP as the span between the furthest correlated SNPs (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 in a single SNP locus . 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 , and loci with no genes 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.

#### 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.

### 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*.

#### Specificity for a matrix of continuous values

We extend an approach we have previoulsy described in detail (Hu *et al.* 2011). Let denote a continuous gene expression matrix with genes and conditions. First, we normalize the expression of each gene by dividing each value by the L2 norm of the genes values in different conditions.

The resulting matrix has values between 0 and 1 indicating specificity of gene to condition . A value indicates that gene is exclusively expressed in condition , and indicates that gene is not expressed in condition .

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

#### Locus scores for a matrix of continuous values

We create a new matrix , where each value is a score for a SNP locus and a condition . The locus scores for a single condition 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 , the values are random, independent, and approximately uniformly distributed. We’ll come back to this assumption later when testing significance in Step 3 below.

#### ’--score single’ (default)

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

For each locus-condition pair , we choose the single gene in locus with greatest specificity to condition among the genes in the locus, as previously described in Hu *et al.* (Hu *et al.* 2011). Let denote this most specific gene, so that where denotes the set of genes in locus . If we assume values of are uniformly distributed for a given condition and genes , then the probability of obtaining a value equal to or less than is as follows:

#### ’--score total’

This assumes all genes in a given SNP locus are associated with a trait — in most circumstances, we consider this model to be unlikely in most situations. We compute the probability of observing values for some locus as the product of percentiles. This assumes values are uniformly distributed.

#### Locus scores for a matrix of binary values

Let denote a binary matrix (1=present, 0=absent) with genes and conditions. Let denote the number of genes present in condition . Let denote the number of genes in locus and denote the number of genes in locus that are present in condition .

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

|  |  |
| --- | --- |
| `’--score single’ | ’--score total’ |
|  |  |

where

$$p(x)=\dfrac{{m\_{j} \choose x}{m-m\_{j} \choose m\_{k}-x}}{{m \choose m\_{k}}}$$

#### Condition specificity scores

For both continuous and binary matrices, we define a specificity score for each condition as the aggregate of values across SNP loci:

### Step 3: Testing significance

#### Analytical p-values

We previously found that aggregating the scores and determining a -value analytically from a distribution results in inaccurate p-values (Hu *et al.* 2011). 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 . We obtain the distribution empirically with null SNP sets.

We compute specificity scores 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 , we calculate an exact permutation p-value (Phipson *et al.* 2010). Let denote the number of sampled SNP sets (e.g. 10,000) and let denote how many null specificity scores are greater than or equal to the user’s score :

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

# 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.

## 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

### SNP sets

|  |  |  |  |
| --- | --- | --- | --- |
| Phenotype | SNPs | Loci | Reference |
| Celiac disease | 35 | 34 | Table 2 ([Trynka, *et al.* 2011](http://www.ncbi.nlm.nih.gov/pubmed/22057235)) |
| HDL cholesterol | 46 | 46 | Supp. Table 2 ([Teslovich, *et al.* 2010](http://www.ncbi.nlm.nih.gov/pubmed/20686565)) |
| Multiple sclerosis | 51 | 47 | Supp. Table A ([IMSGC WTCCC 2011](http://www.ncbi.nlm.nih.gov/pubmed/21833088)) |
| Red blood cell count | 45 | 45 | Table 1 ([Harst *et al.* 2012](http://www.ncbi.nlm.nih.gov/pubmed/23222517)) |

#### Celiac\_disease-Trynka2011-35\_SNPs.gwas

35 SNPs associated with Celiac disease taken from Table 2. Positions are on hg19. All SNPs have .

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. [Pubmed](http://www.ncbi.nlm.nih.gov/pubmed/22057235)

#### HDL\_cholesterol-Teslovich2010-46\_SNPs.gwas

46 SNPs associated with HDL taken from Supplementary Table 2. Positions are on hg19. All SNPs have .

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. [Pubmed](http://www.ncbi.nlm.nih.gov/pubmed/20686565)

#### Multiple\_sclerosis-IMSGC-51\_SNPs.gwas

51 SNPs associated with Multiple Sclerosis taken from Supplementary Table A. Positions are on hg19. All SNPs have .

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. [Pubmed](http://www.ncbi.nlm.nih.gov/pubmed/21833088)

#### 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 .

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. [Pubmed](http://www.ncbi.nlm.nih.gov/pubmed/23222517)

### Gene matrices

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type | Genes | Conditions | Species | Reference |
| Affy | 17581 | 79 tissues | homo sapiens | [GeneAtlas 2004](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1133) |
| Affy | 15139 | 249 cells | mus musculus | [ImmGen 2012](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15907) |
| CAGE | 18502 | 533 cells | homo sapiens | [FANTOM5 2014](http://fantom.gsc.riken.jp/5/data/) |
| Binary | 19111 | 1751 terms | homo sapiens, mus musculus | [Gene Ontology](http://www.geneontology.org) 2013, [Homologene](http://www.ncbi.nlm.nih.gov/homologene) |

#### GeneAtlas2004.gct.gz

Gene expression data for 79 human tissues from [GSE1133](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=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](http://www.broadinstitute.org/cancer/software/genepattern/gp_guides/file-formats/sections/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. [PubMed](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1133)

#### GO2013.gct.gz

A [GCT](http://www.broadinstitute.org/cancer/software/genepattern/gp_guides/file-formats/sections/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](http://www.geneontology.org) (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 , we also add all of the terms that are parents of . We copy terms between homologous genes using [Homologene](http://www.ncbi.nlm.nih.gov/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.

#### ImmGen2012.gct.gz

Gene expression data for 249 blood cell types from [GSE15907](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15907). We averaged cell type replicates. For each gene, we selected the single probe with the largest minimum.

#### FANTOM2014.gct.gz

CAGE data for 533 human cell types from [FANTOM5](http://fantom.gsc.riken.jp/5/data/). 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.

### LD-pruned SNPs and Genomic Intervals

#### 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. [Pubmed](http://www.ncbi.nlm.nih.gov/pubmed/20881960)

#### NCBIgenes2013.bed.gz

All human start and stop positions taken from:

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

#### TGP2011.bed.gz

Linkage intervals for a filtered set of SNPs from the [1000 Genomes Project](http://www.1000genomes.org/) 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](http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3) website and calculated pairwise LD (EUR) for all SNPs in a 1 Mb sliding window. The linkage intervals were extended to the nearest [HapMap](http://hapmap.ncbi.nlm.nih.gov/downloads/) recombination hotspot with >3 cM/Mb recombination rate.

## C++ Libraries

To compile SNPsea, you will need a modern C++ compiler that supports [c++0x][cxx0x] and the dependencies listed below.

See: [Installation](#installation)

[intervaltree](https://github.com/slowkow/intervaltree)

a minimal C++ interval tree implementation

[Eigen](http://eigen.tuxfamily.org)

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

[OpenMPI](http://www.open-mpi.org)

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](http://www.gnu.org/software/gsl)

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

[GCC, the GNU Compiler](http://gcc.gnu.org)

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

I use [c++11] features in my code, so you must use a compiler that supports them. I compiled successfully with versions 4.6.3 (the default version for Ubuntu 12.04) and 4.8.1.

## Python Packages

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

**Instructions:** Install with [pip](http://www.pip-installer.org):

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](http://docopt.org/)

Command-line interface description language.

[numpy](http://www.numpy.org)

NumPy is the fundamental package for scientific computing with Python.

[pandas](http://pandas.pydata.org)

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](http://matplotlib.org)

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](http://matplotlib.org/users/customizing.html) 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

## 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](http://cran.r-project.org/web/packages/data.table)

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

[reshape2](http://cran.r-project.org/web/packages/reshape2)

Flexibly reshape data: a reboot of the reshape package.

[gap](http://cran.r-project.org/web/packages/gap)

Genetic analysis package.

[ggplot2](http://cran.r-project.org/web/packages/ggplot2)

An implementation of the Grammar of Graphics.

# Usage Example

Here is a [Bash](http://www.gnu.org/software/bash/manual/bashref.html) 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.

## Options

All input files may optionally be compressed with [gzip](http://www.gzip.org/).

### 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.

### 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]

## Input File Formats

### --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.

### --gene-matrix ARG

You must provide a single gene matrix that must be in [GCT](http://www.broadinstitute.org/cancer/software/genepattern/gp_guides/file-formats/sections/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

### --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

### --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](http://www.broadinstitute.org/cancer/software/genepattern/gp_guides/file-formats/sections/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

### --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

### --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

## 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

### 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

### 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

### 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

### 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

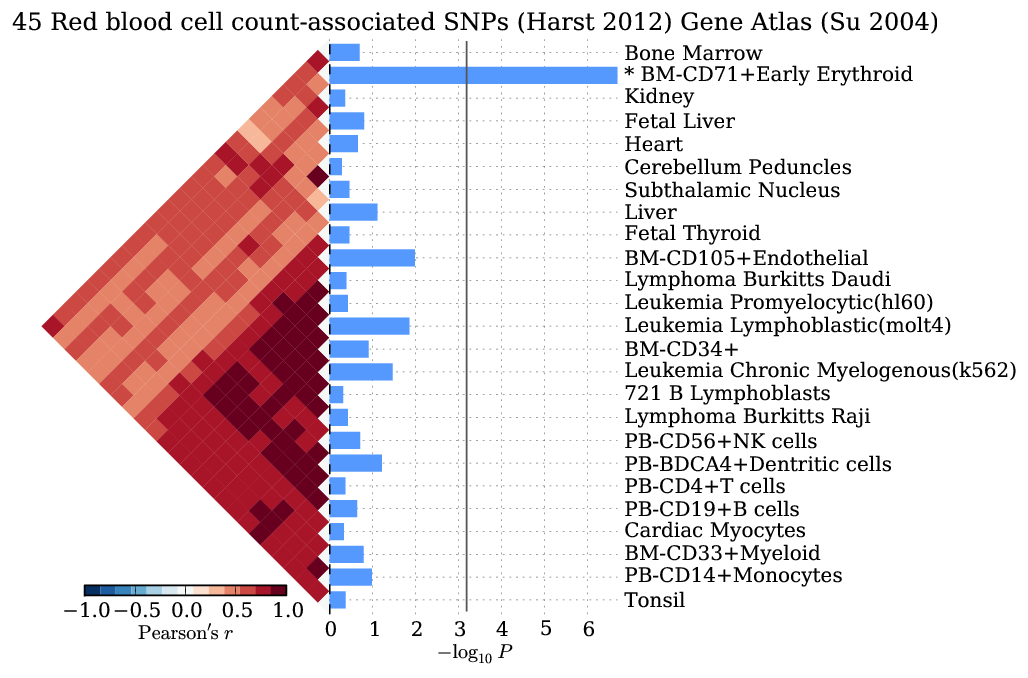
### snp\_condition\_scores.txt

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

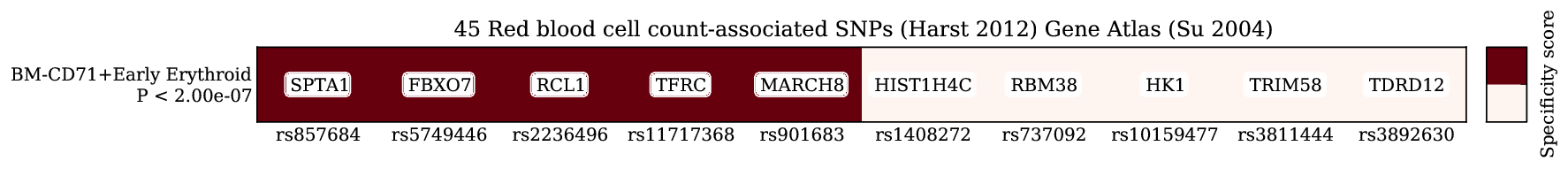
### View enrichment of tissue-specific gene expression



A horizontal bar plot of negative log10 p-values for a test of 37 LDL-associated SNPs for enrichment of tissue-specific expression in profiles of 79 human tissues.

python bin/snpsea-barplot out

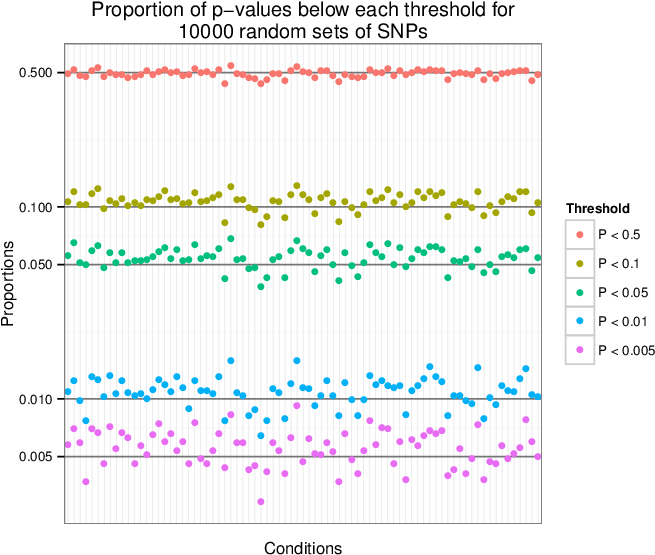
### View the most specifically expressed gene for each SNP-tissue pair



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

### View the type 1 error rate estimates for each tissue

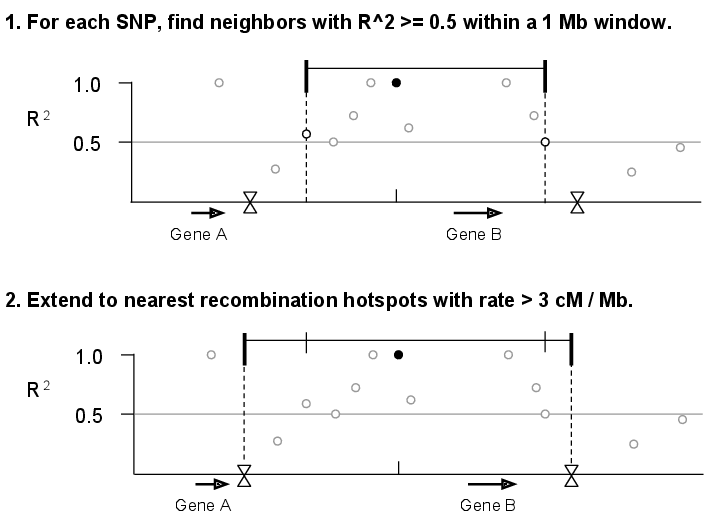


A scatter plot of the observed proportion of p-values under various thresholds after repeating the analysis with 10,000 random SNP sets.

Rscript bin/snpsea-type1error out

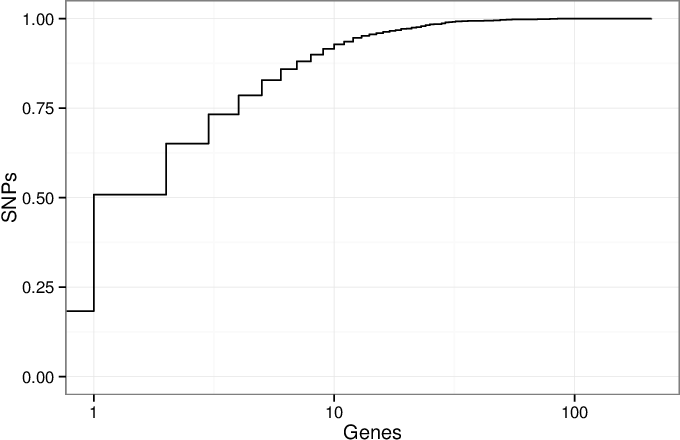
# Supplementary Figures

#### Supplementary Figure 1: Determining SNP linkage intervals



We calculated 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 . Finally, we extended each interval to recombination hotspots reported by HapMap (Myers *et al.* 2005) 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 (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.

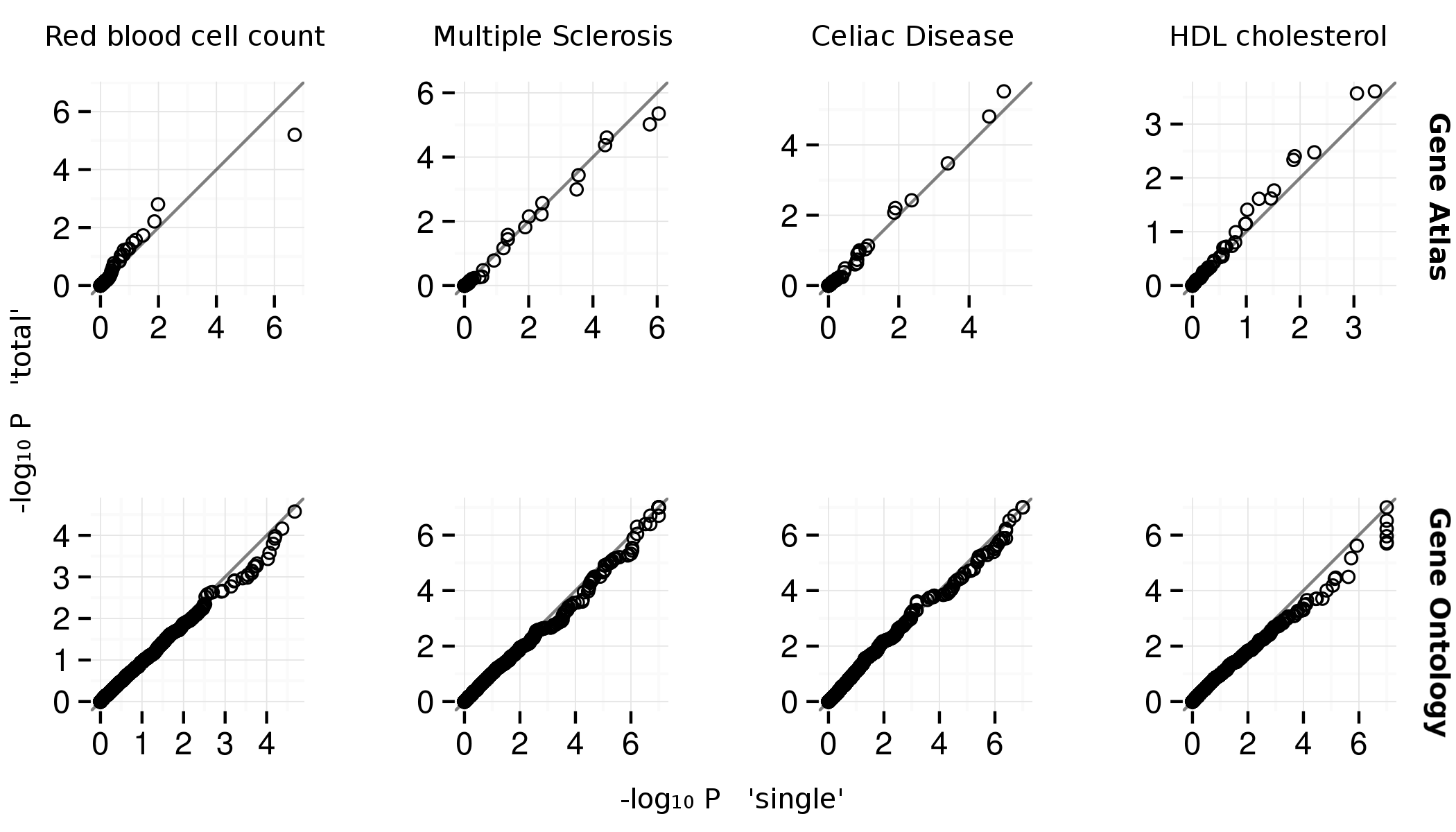
#### Supplementary Figure 3: Choosing the threshold for linkage intervals

We chose to use 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 ( 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.

|  |  |
| --- | --- |
|  |  |

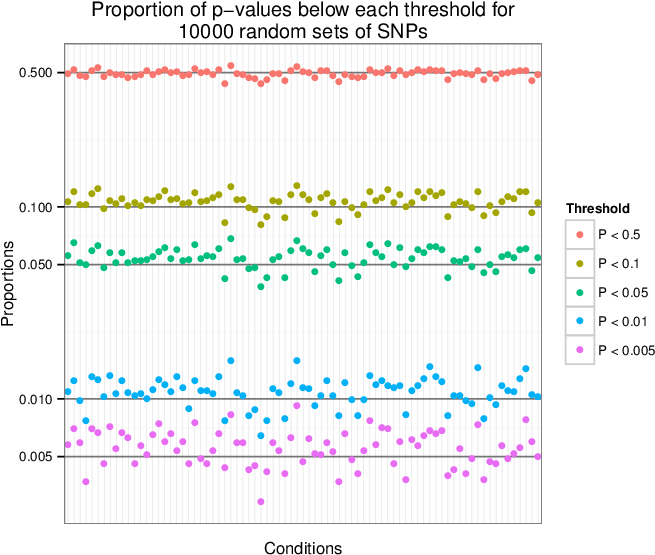
Gene Atlas and Gene Ontology (left and right). Each subplot has for on the x-axis and on the y-axis for the threshold marked above. Grey lines are significance thresholds after correction testing multiple conditions (cell types, GO annotations). Black points are significant and grey are not. We used the ’--score single’ option. Red blood cell count SNPs are enriched for *hemopoiesis* (GO:0030097) () for linkage intervals with . This result falls below the multiple testing threshold at , but remains significant at (see main text).

#### 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 for ’--score single’ and ’--score total’ SNPsea options, respectively. The ’single’ and ’total’ methods are described . The -values appear similar between methods.

#### Supplementary Figure 5: Type 1 error estimates





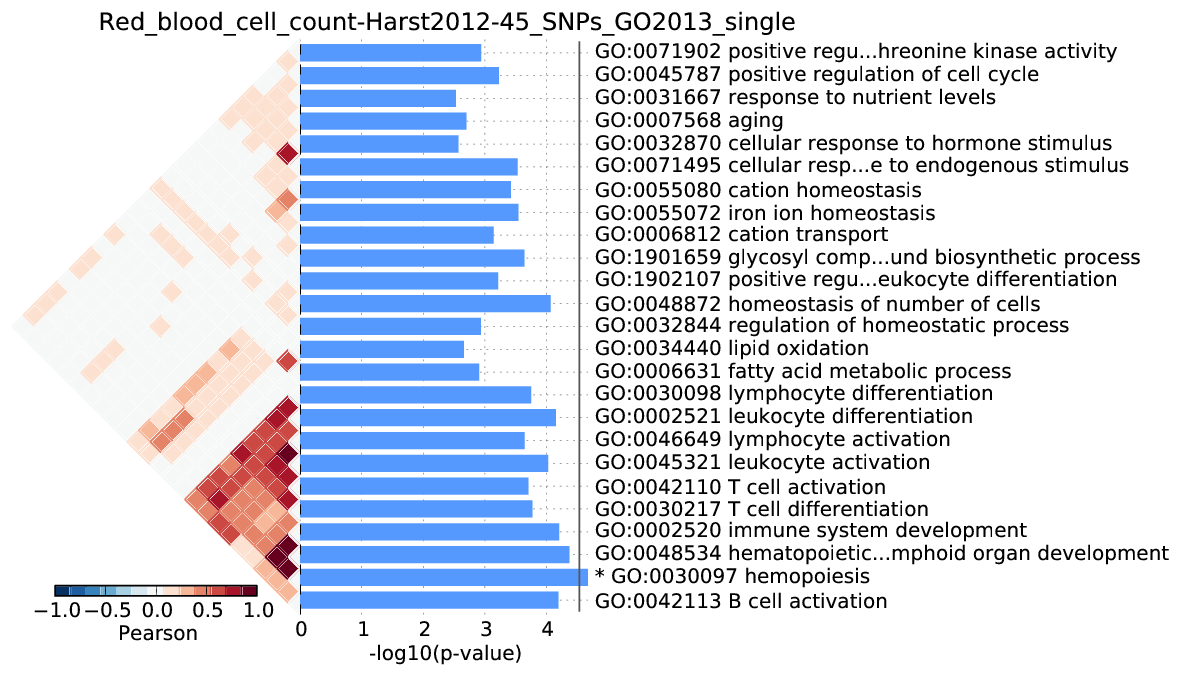
We sampled 10,000 sets of 100 SNPs uniformly from a list of LD-pruned SNPs (Lango Allen *et al.* 2010). We tested each of the 10,000 sets for enrichment of tissue-specific expression in the Gene Atlas (Su *et al.* 2004) gene expression matrix (top) and for enrichment of annotation with Gene Ontology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

# Additional Examples

We tested SNPsea with the three additional phenotypes listed below with genome-wide significant SNPs . 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.

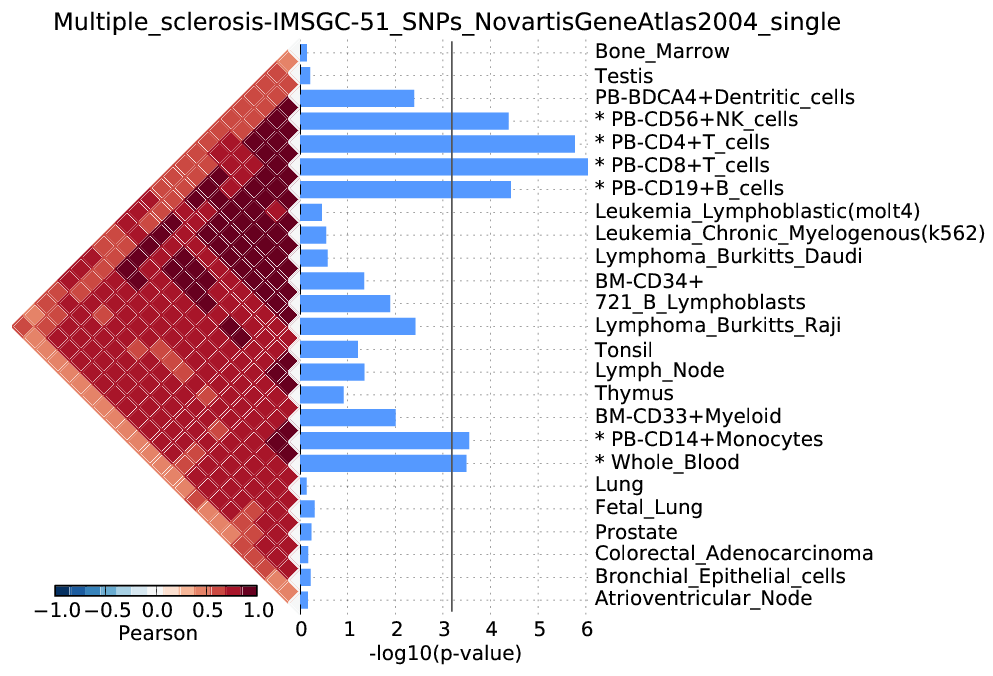
|  |  |  |  |
| --- | --- | --- | --- |
| Phenotype | SNPs | Loci | Reference |
| Multiple sclerosis | 51 | 47 | Supp. Table A (IMSGC WTCCC 2011) |
| Celiac disease | 35 | 34 | Table 2 (Trynka, *et al.* 2011) |
| HDL cholesterol | 46 | 46 | Supp. Table 2 (Teslovich, *et al.* 2010) |

#### Supplementary Figure 6: Red blood cell count GO enrichment

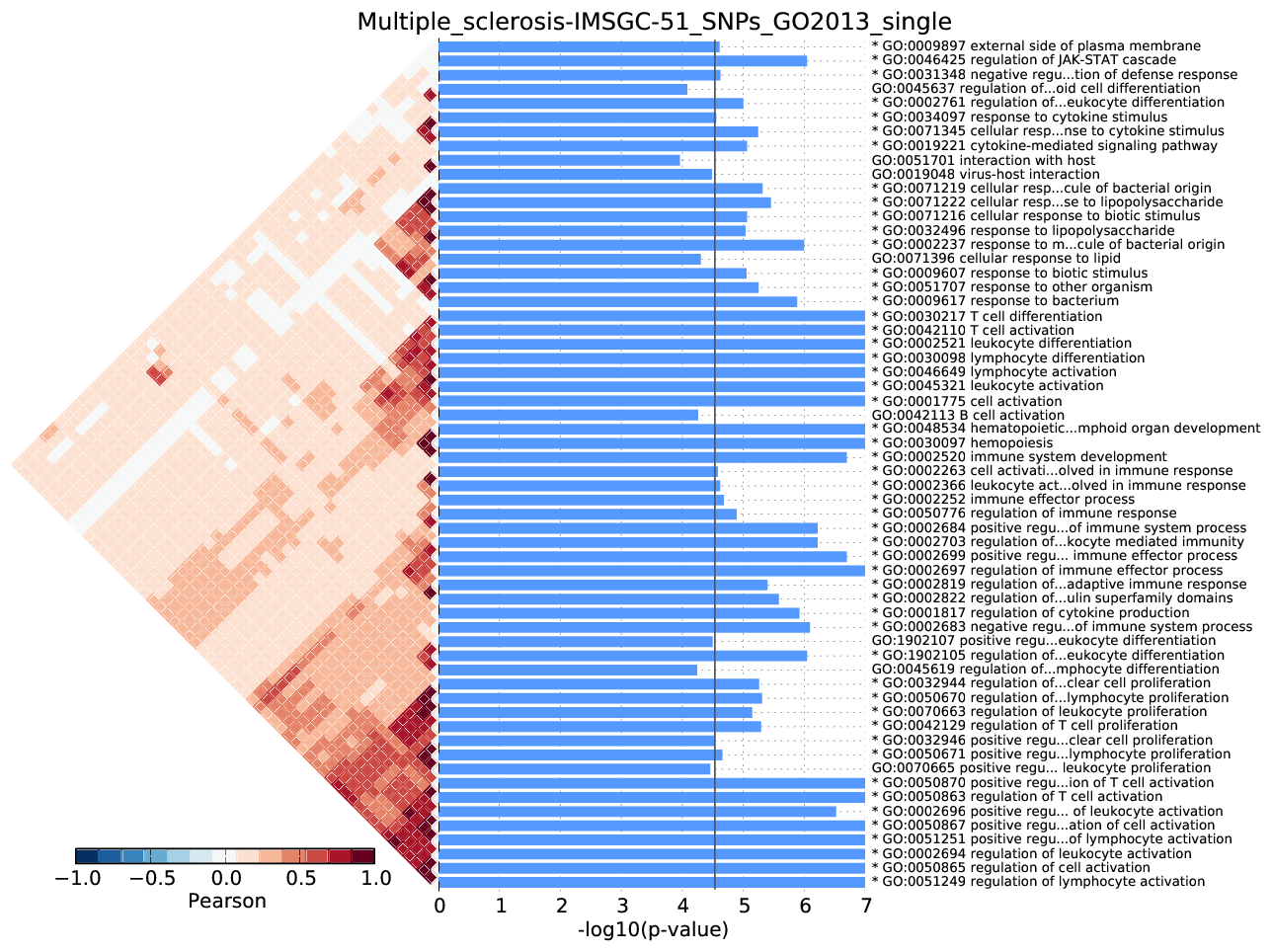


We observed significant enrichment for *hemopoiesis* . The top 25 terms are shown.

#### Supplementary Figure 7: Multiple sclerosis

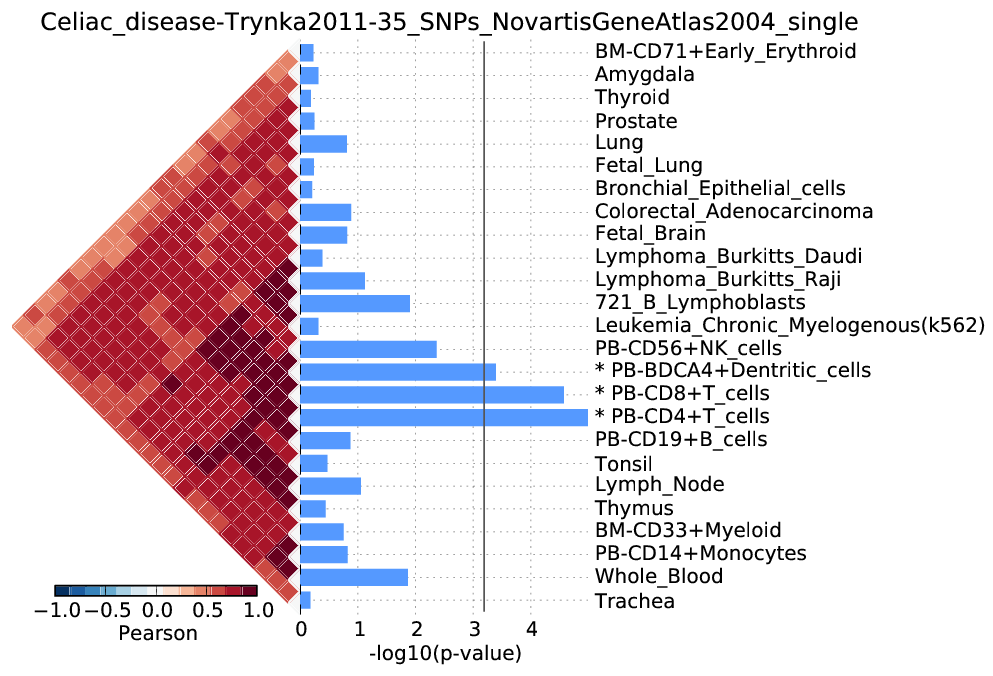


We observed significant enrichment for 6 cell types. The top 25 of 79 are shown.

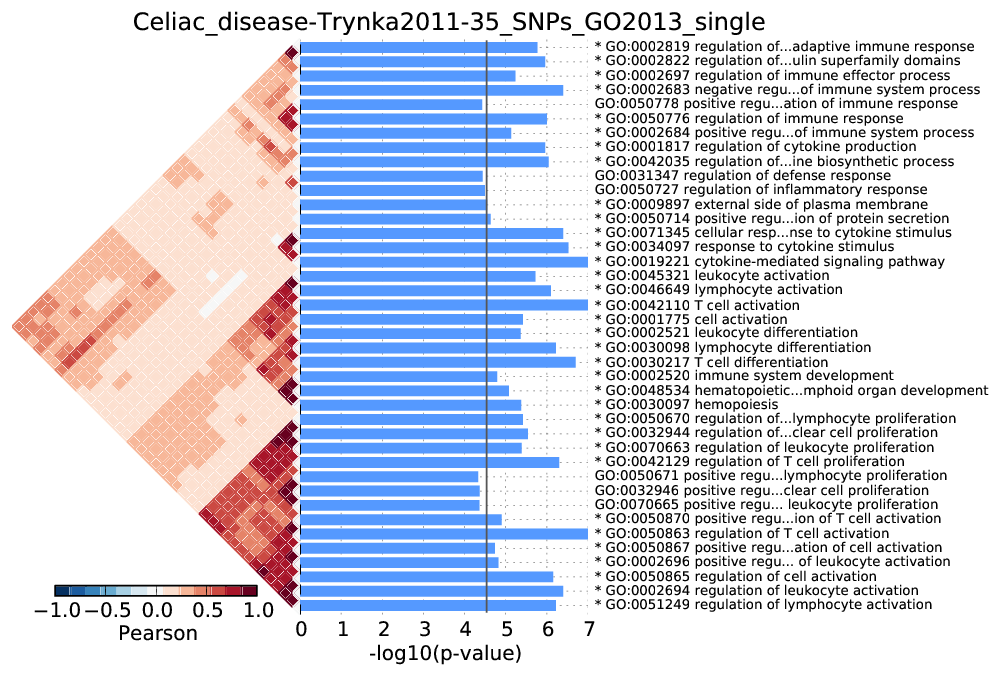


We observed significant enrichment for 52 Gene Ontology terms. The top 60 terms are shown.

#### Supplementary Figure 8: Celiac disease

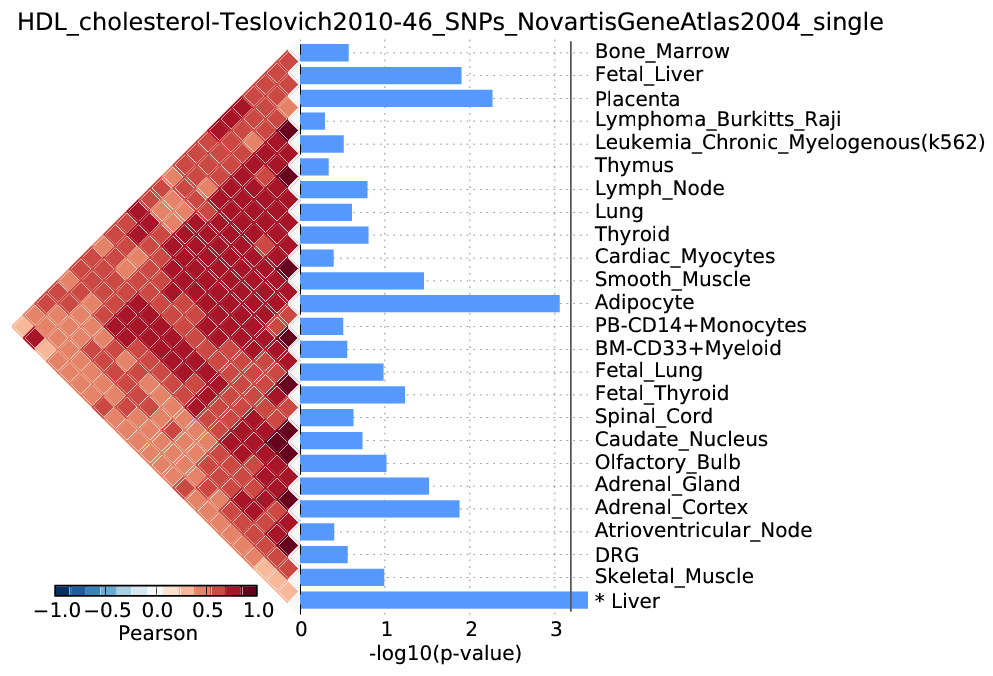


We observed significant enrichment for 3 cell types. The top 25 of 79 are shown.

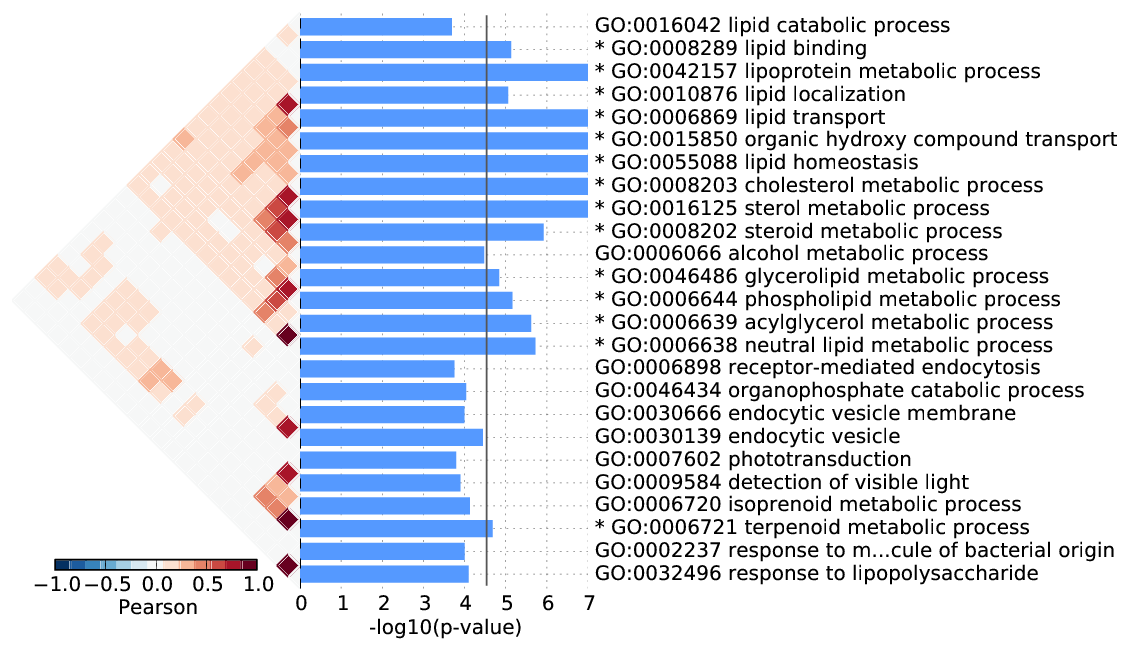


We observed significant enrichment for 28 Gene Ontology terms. The top 40 terms are shown.

#### 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, 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. X. Hu, H. Kim, E. Stahl, R. Plenge, M. Daly, and S. 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 Ri- vadeneira, 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.