

# SNPsea Reference Manual

Kamil Slowikowski

February 13, 2014

## Contents

<b>Introduction</b>	<b>2</b>
Contact . . . . .	2
Visual Summary . . . . .	3
<b>Installation</b>	<b>3</b>
Data . . . . .	4
C++ Libraries . . . . .	7
Python Packages . . . . .	8
R Packages . . . . .	9
<b>Usage</b>	<b>10</b>
Example . . . . .	10
Options . . . . .	10
Input File Formats . . . . .	12
Output Files . . . . .	14
Output Visualizations . . . . .	17

## Introduction

SNPsea is a general algorithm to identify cell types, tissues, and pathways likely to be affected by risk loci.

For example, with a gene expression matrix containing expression profiles for multiple cell types, we identify genes in linkage disequilibrium with trait-associated SNPs and score them for specificity to each cell type. We compare the score to a null distribution by sampling random SNP sets matched on the number of linked genes. To evaluate significance, we calculate an exact permutation p-value.

This implementation is generalized, so you may provide (1) a continuous gene matrix with gene expression (or any other values) or (2) a binary gene matrix with presence/absence 1/0 values.

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

In general, this analysis is appropriate when you are interested in testing for enrichment of condition-specificity of genes linked to a set of trait-associated SNPs.

If trait-associated alleles impact a small number of pathogenic tissues or cell types, we hypothesize that the subset of genes with critical functions in those pathogenic cell types are likely to be within trait-associated loci.

We assume that a gene's specificity to a given cell type or condition is a reasonable indicator of the gene's importance to its function.

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.** Manuscript in progress.

See additional examples:

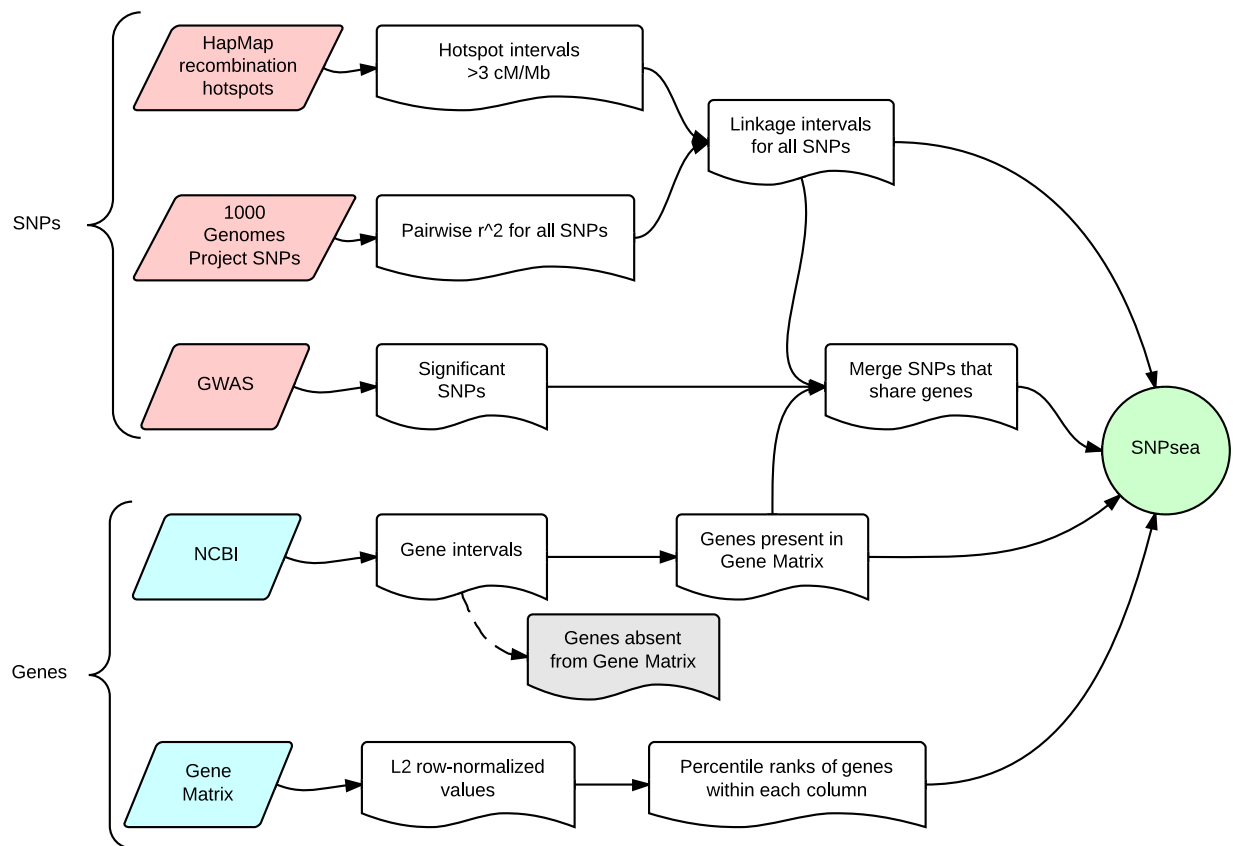
Hu, X. et al. **Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets.** The American Journal of Human Genetics 89, 496–506 (2011). [PubMed](#)

## Contact

Please contact me with questions and comments: [slowikow@broadinstitute.org](mailto:slowikow@broadinstitute.org)

## Visual Summary

### Flow Chart



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

## Installation

### Linux 64-bit

Download the binary and run it: <https://github.com/slowkow/snpsea/releases>

### Mac

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

### All other platforms

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

Install the dependencies:

*# Ubuntu*

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

*# Mac*

*# Install MacPorts: <http://www.macports.org/>*

*# Then do this:*

```
sudo port selfupdate
```

```
sudo port install gcc48 openmpi gsl
```

*# Broad Institute*

*# You can add this to your .my.bashrc or .my.cshrc*

```
use .gcc-4.8.1 .openmpi-1.4 .gsl-1.14
```

Download and compile the code:

*# Clone with git, so you can get updates with 'git pull'*

```
git clone https://github.com/slowkow/snpsea.git
```

```
cd snpsea
```

*# or if you don't have git*

```
curl -LOk https://github.com/slowkow/snpsea/archive/master.zip
```

```
unzip master.zip
```

```
cd snpsea-master
```

*# Compile.*

```
cd src
```

```
make
```

*# Move the executables wherever you like.*

```
mv ../bin/snpsea* ~/bin/
```

Download the required data and run SNPsea:

```
mkdir ../snpsea/data
```

```
cd ../snpsea/data
```

```
curl -LOk http://files.figshare.com/1307287/SNPsea_data_20131204.zip
```

```
unzip SNPsea_data_20131204.zip
```

```
snpsea
```

## Data

Download the compressed archive with data required to perform this analysis here (138M):

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

## GWAS SNPs

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 gene expression matrix  
GO2013.gct.gz # Gene Ontology 2013 gene annotation matrix  
ImmGen2012.gct.gz # ImmGen 2012 gene expression matrix

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

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

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

doi:10.1038/ng.998  
PMID: 22057235

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.

<http://www.ncbi.nlm.nih.gov/pubmed/22057235>

## HDL\_\_cholesterol-Tesolvich2010-46\_\_SNPs.gwas

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

doi:10.1038/nature09270  
PMID: 20686565

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.

<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  $P \leq 5e-8$ .

doi:10.1038/nature10251

PMID: 21833088

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.

<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  $P \leq 5e-8$ .

doi:10.1038/nature11677

PMID: 23222517

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.

<http://www.ncbi.nlm.nih.gov/pubmed/23222517>

### **GeneAtlas2004.gct.gz**

Gene expression data for 79 human tissues from GSE1133. Replicates for each tissue profile were averaged. For each gene, the single probe with the largest minimum was selected.

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

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1133>

### **GO2013.gct.gz**

A GCT formatted gene matrix with 1s and 0s indicating presence or absence of genes in Gene Ontology annotations. 19,111 genes in 1,751 Gene Ontology annotations.

<http://www.geneontology.org/>

### **ImmGen2012.gct.gz**

Gene expression data for 249 blood cell types from GSE15907. Replicates for each cell type profile were averaged. For each gene, the single probe with the largest minimum was selected.

Immunological Genome Project

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15907>

### **NCBIgenes2013.bed.gz**

All human start and stop positions taken from:

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

### **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.

<http://www.ncbi.nlm.nih.gov/pubmed/20881960>

### **TGP2011.bed.gz**

Linkage intervals for a filtered set of SNPs from the 1000 Genomes Project Phase 1 (May 21, 2011). SNP genotypes were obtained from the BEAGLE release v3 website and processed to create linkage intervals for each SNP. The linkage intervals were extended to the nearest HapMap recombination hotspot with >3 cM/Mb recombination rate.

<http://www.1000genomes.org/>

[http://bochet.gcc.biostat.washington.edu/beagle/1000\\_Genomes.phase1\\_release\\_v3/](http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/)

<http://hapmap.ncbi.nlm.nih.gov/downloads/>

### **C++ Libraries**

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

See: [Installation](#)

[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.

I use `c++0x` features in my C++ 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`:

```
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
```

## 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.

## Usage

### Example

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[*]}
```

SNPs 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](#).

### Required

<code>--snps ARG</code>	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.
<code>--gene-matrix ARG</code>	Gene matrix file in GCT format. The Name column must contain the same gene identifiers as in <code>--gene-intervals</code> .
<code>--gene-intervals ARG</code>	BED file with gene intervals. The fourth column must contain the same gene identifiers as in <code>--gene-matrix</code> .

`--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. Only the first column is used.

```
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 <= 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](#) 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 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+Dendritic_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+Dendritic_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	name	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
```

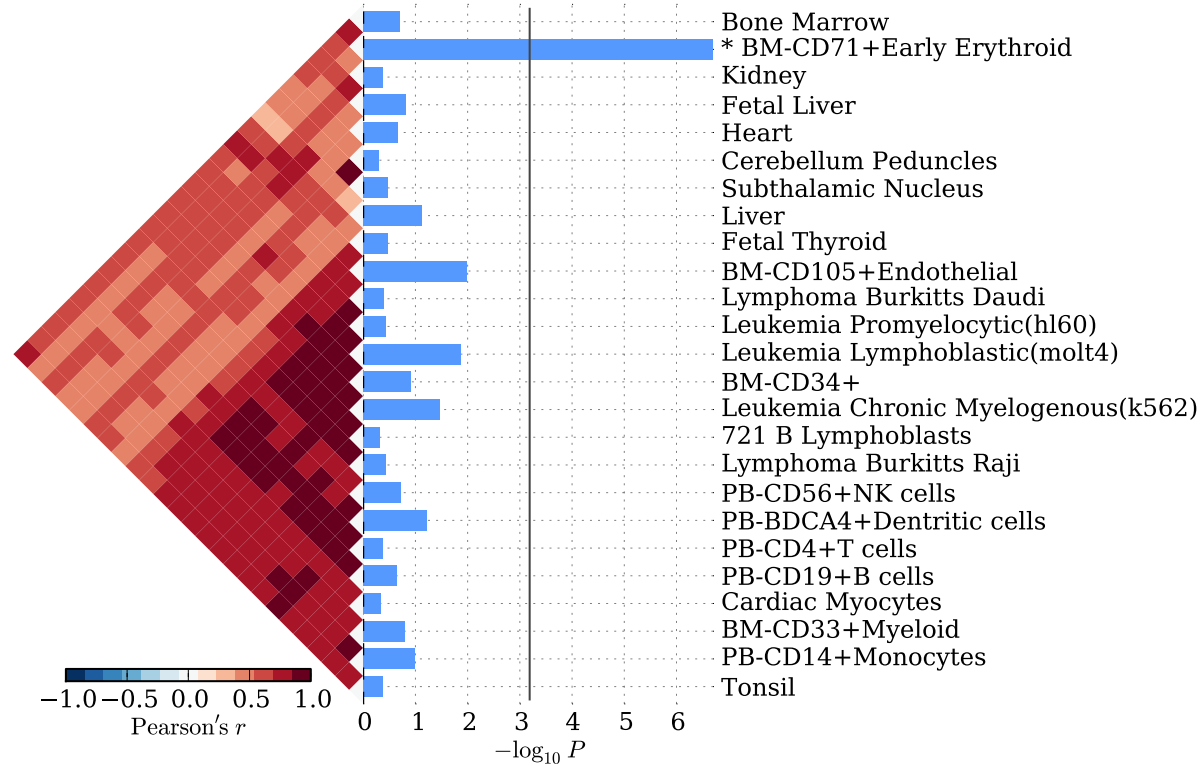
snp	condition	gene	score
rs9349204	Colorectal_Adenocarcinoma	10817	0.693027
rs9349204	Whole_Blood	896	0.285864
rs9349204	BM-CD33+Myeloid	896	0.236487
rs9349204	PB-CD14+Monocytes	29964	0.340561
rs9349204	PB-BDCA4+Dendritic_cells	29964	0.411727
rs9349204	PB-CD56+NK_cells	896	0.0356897
rs9349204	PB-CD4+T_cells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



Output Visualizations

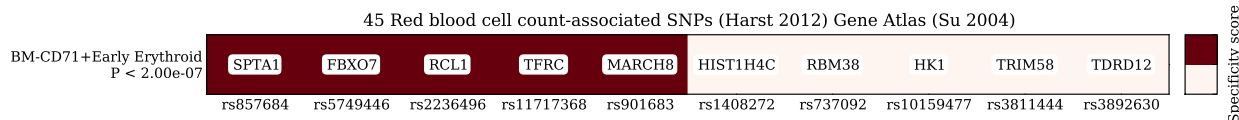
View enrichment of tissue-specific gene expression

45 Red blood cell count-associated SNPs (Harst 2012) Gene Atlas (Su 2004)



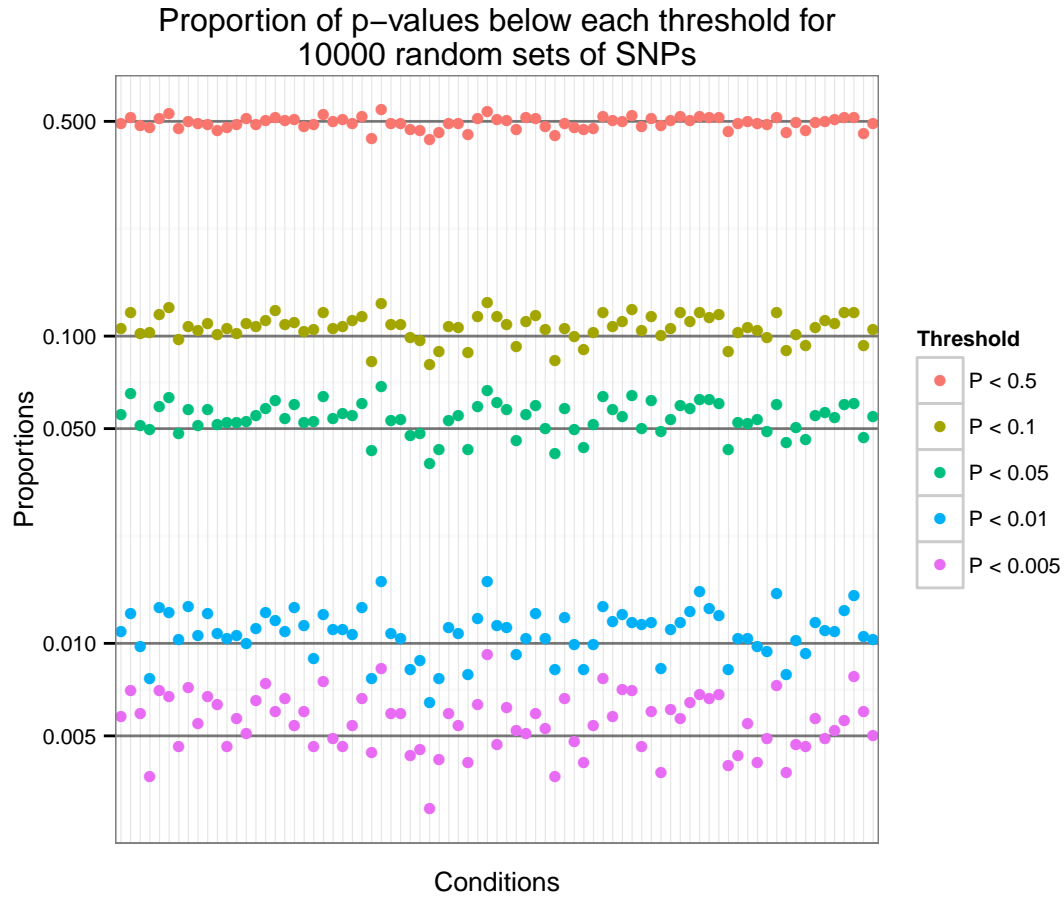
`python bin/snpsea-barplot out`

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



`python bin/snpsea-heatmap out`

View the type 1 error rate estimates for each tissue



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