# Documentation for Software GARFIELD

Valentina Iotchkova, Matthias Geihs, Graham Ritchie

February 9, 2015

## 1 Introduction

Software GARFIELD (GWAS analysis of regulatory and functional information enrichment with LD correction) implements the non-parametric functional enrichment analysis approach described in [Iotchkova et al, 2015] using C++ for data pre-processing, enrichment estimation and significance testing and R for visualisation. It provides a tool for assessing the enrichment of association analysis signals in 1005 features extracted from ENCODE, GENCODE and Roadmap Epigenomics projects, including genic annotations, chromatin states, histone modifications, DNaseI hypersensitive sites and transcription factor binding sites, among others, in a number of publicly available cell lines.

### 1.1 Method Overview

Genome-Wide Association Studies (GWAS) have been increasingly fruitful in discovering genotype-phenotype associations. The mechanisms underlying these associations, however, are still largely unknown as only a small fraction of these SNPs are known to directly alter protein-coding genes. The interpretation of functional consequences of non-coding variants has been greatly enhanced by large-scale efforts to identify regulatory genomic regions (e.g ENCODE). However, robust methods are still lacking to systematically evaluate the contribution of these regions to genetic variation implicated in diseases or quantitative traits.

We have developed a novel approach, named GARFIELD, that leverages GWAS findings with regulatory or functional annotations to find features relevant to a phenotype of interest. It performs greedy pruning of GWAS SNPs (LD $r^2 > 0.1$) and then annotates them based on functional information overlap. Next, it quantifies Fold Enrichment (FE) at various GWAS significance cut-offs and assesses them by permutation testing, while matching for minor allele frequency, distance to nearest transcription start site and number of LD proxies ($r^2 > 0.8$). Within this framework, GARFILED accounts for major sources of confounding that current methods do no offer.

### 1.2 How to cite

Valentina Iotchkova, Graham Ritchie, Matthias Geihs, Josine Min, Ian Dunham, Ewan Birney and Nicole Soranzo. *GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction.* In preparation

### 1.3 Contacts

For issues or further questions about GARFIELD you can get in touch at vi1@sanger.ac.uk.

## 2 Download and Installation

We provide source code for your own compilation, which can be downloaded as a compressed tarball from

`http://www.ebi.ac.uk/birney-srv/GARFIELD/`

or by opening a terminal and typing

```
wget http://www.ebi.ac.uk/birney-srv/GARFIELD/package/garfield.tar.gz
```

We further provide all data necessary for running GARFIELD for analysis of genome-wide association studies in European populations. Those can be downloaded from

`http://www.ebi.ac.uk/birney-srv/GARFIELD/`

or by typing the following in a terminal

```
wget http://www.ebi.ac.uk/birney-srv/GARFIELD/package/garfield-data.tar.gz
```

**Note:** You should download both files in the same directory or you will need to modify the paths for the data files in the `garfield` script.

### 2.1 Installation

To decompress and extract the files in the software bundle in a terminal from the location where it is downloaded type

```
tar -xvf garfield.tar.gz
```

In order to compile GARFIELD a C++ compiler is required. Additionally an R distribution needs to already be pre-installed for the end figure creation. Please make sure you have both of them before proceeding any further.

To compile the code type

```
cd garfield
make
```

This would create two executables `garfield-prep` and `garfield-perm`. Further details of how to run them will be given in the **Running GARFIELD** section.

In order to use the data files, they must also be decompressed. To do so execute

```
cd ../
tar -xvf garfield-data.tar.gz
```

## 2.2 License



Author : Valentina Iotchkova <vi@sanger.ac.uk>
Author : Matthias Geihs <mg18@sanger.ac.uk>



## 2.3 Tree structures

```
garfield
|----- garfield
|----- garfield-perm.cpp
|----- garfield-plot-function.R
|----- garfield-plot.R
|----- garfield-prep.cpp
|----- LICENSE
|----- makefile
|----- README
```

```
garfield-data
|----- annotation
|   |----- chr1
|   |----- ...
|   |----- chr22
|   |----- link_file.txt
|----- maftssd
|   |----- chr1
|   |----- ...
```

```
|  |----- chr22
|----- output
|  |----- cd-meta
|  |  |----- garfield.cd-meta.Chromatin_States.pdf
|  |  |----- garfield.cd-meta.FAIRE.pdf
|  |  |----- garfield.cd-meta.Footprints.pdf
|  |  |----- garfield.cd-meta.Genic.pdf
|  |  |----- garfield.cd-meta.Histone_Modifications.pdf
|  |  |----- garfield.cd-meta.Hotspots.pdf
|  |  |----- garfield.cd-meta.Peaks.pdf
|  |  |----- garfield.cd-meta.TFBS.pdf
|  |  |----- garfield.perm.cd-meta.out
|  |  |----- garfield.prep.cd-meta.out
|  |----- GIANT_HEIGHT
|     |----- garfield.GIANT_HEIGHT.Chromatin_States.pdf
|     |----- garfield.GIANT_HEIGHT.FAIRE.pdf
|     |----- garfield.GIANT_HEIGHT.Footprints.pdf
|     |----- garfield.GIANT_HEIGHT.Genic.pdf
|     |----- garfield.GIANT_HEIGHT.Histone_Modifications.pdf
|     |----- garfield.GIANT_HEIGHT.Hotspots.pdf
|     |----- garfield.GIANT_HEIGHT.Peaks.pdf
|     |----- garfield.GIANT_HEIGHT.TFBS.pdf
|     |----- garfield.perm.GIANT_HEIGHT.out
|     |----- garfield.prep.GIANT_HEIGHT.out
|----- pval
|  |----- cd-meta
|  |  |----- chr1
|  |  |----- ...
|  |  |----- chr22
|  |----- GIANT_HEIGHT
|     |----- chr1
|     |----- ...
|     |----- chr22
|----- tags
   |----- r01
   |  |----- chr1
   |  |----- ...
   |  |----- chr22
   |----- r08
      |----- chr1
      |----- ...
      |----- chr22
```

# 3 Input files

GARFIELD requires a number of input files containing the GWAS, LD and annotation data as well as minor allele frequencies (MAF) and distances to nearest transcription start site (TSS) needed for the analysis. We have pre-processed data on all variants from the UK10K study, which can be used for the enrichment analysis of European sample GWAS. This data is contained in the `garfield-data.tar.gz` file. The size of the compressed tarball is 5.9Gb which amounts to 83Gb after unpacking spread into five sub-directories `annotation, pval, maftssd, tags` and `output`.

If you have downloaded the data somewhere else than the same folder as the software package you would need to change the `$DATADIR` paths set in the `./garfield/garfield` file.

If you are interested in running enrichment analysis for non-European samples then allele frequencies and LD need to be re-calculated for them and put into the types of files described in the next subsections.

## 3.1 GWAS data

Summary statistics from a GWA study of interest need to be provided in a separate subfolder of the `./garfield-data/pval` folder, e.g. `./garfield-data/pval/GIANT_HEIGHT`, in files split by chromosome and named `chr1, chr2,` etc.. The files must contain no header and have genomic position (build 37) in the first column and P-value from association analysis in the second column using a space as a column delimiter. In addition, files must be numerically sorted. An example by the supplied in the package `./garfield-data/pval/GIANT_HEIGHT/chr22` file is given below

```
16381711 0.9274
16966809 0.4753
16970900 0.5239
16983606 0.4753
16989290 0.4689
17030792 0.2353
```

The GIANT Height association analysis summary statistics have been downloaded from

`http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`

and processed by us and added here as an example together with a GWAS on Crohn's disease downloaded from

`http://www.ibdgenetics.org/downloads.html`

Tree structure

```
garfield-data
|----- pval
|  |----- cd-meta
|  |  |----- chr1
|  |  |----- ...
```

```
|  |  |----- chr22
|  |----- GIANT_HEIGHT
      |----- chr1
      |----- ...
      |----- chr22
```

## 3.2 Annotation data

The annotation data provided can be found in the `./garfield-data/annotation` folder. It is again split in different `chr` files organised as numerically sorted two-column space-delimited files containing position of variant and annotations, where the annotations are provided in 0/1 format and stacked together with no delimiters in the second column. Order of the annotations corresponds to the file `link_file.txt` provided in the same folder, which contains additional information on each annotation. A dummy example of the file format is given below

```
16051237 00000000000000000000001000001000000000000000000000000000000011111
16051249 00000000000000000000001000001000000000000000000000000000000000000
16051250 00000000000000000000001000001000000000000000000000000000000000001
```

Tree structure

```
garfield-data
|----- annotation
   |----- chr1
   |----- ...
   |----- chr22
   |----- link_file.txt
```

## 3.3 LD data

For the GARFIELD enrichment analysis we need two types of LD information - low $r^2(>= 0.1)$ for ensuring independence of the variants to be tested and high $r^2(>= 0.8)$ for further annotation of variants. This data is provided in the `./garfield-data/tags/r01` and `./garfield-data/tags/r08` folders, respectively. Files are again split by chromosome and numerically sorted. They contain two space separated columns, where the first column contains the genomic position (build 37) of each UK10K variant and the second contains lists of comma separated genomic positions of the LD tags of each variant in the first column. An example is given below

```
16057013 16052080,16053863,16055070
16057248 16433985
16057287 16555281
16057320 16053791,16554886
16057417 16051249,16051497,16055122,16114555
16060308 16341445
16060348 16053767
16060517 16052239,16052513,16054667,16054740,16055942,16069707,16097764
16060732 16052837
16060741 16056839
```

Tree structure

```
garfield-data
|----- tags
   |----- r01
   |  |----- chr1
   |  |----- ...
   |  |----- chr22
   |----- r08
      |----- chr1
      |----- ...
      |----- chr22
```

### 3.4 MAF, TSS distance

Finally, for testing we need to match variants by MAF and distance to nearest TSS. This data is provided in the `./garfield-data/maftssd` folder, again split into space separated numerically sorted `chr` files. An example is given below, where the first column contains the genomic position of each variant, the second it's MAF calculated from the UK10K sequence data and the third contains it's distance to the nearest TSS.

```
16051237 0.000145677788349 -10919
16051249 0.106493504342 -10907
16051477 0.00320565708155 -10679
16051497 0.36456744798 -10659
16052080 0.104518066225 -10076
16052107 0.000285031428574 -10049
16052216 0.000143040758483 -9940
```

Tree structure

```
garfield-data
|----- maftssd
   |----- chr1
   |----- ...
   |----- chr22
```

## 4 Running GARFIELD

The file `./garfield/garfield` contains all necessary commands and shows how to run the software for the GIANT Height data example. To execute simply open a terminal, go to the `garfield` software directory and type

```
./garfield
```

If you follow the same directory structure as described above, to run analysis for your own GWAS data you only need to put it in a subfolder of `./garfield-data/pval` and then change the

`$INPUTNAME` variable in the `garfield` file to the name of this subfolder.

In more detail, the enrichment analysis is performed with the use of three separate tools

1. `garfield-prep` for preparing the data to be used for analysis
2. `garfield-perm` for calculating fold enrichment and associated empirical p-values
3. `garfield-plot.R` for visualisation of the computed results.

## 4.1 garfield-prep

prepares the input data for `garfield-perm` and prints it to stdout.

Usage:

`./garfield-prep prunetags clumptags maftssd pvalue annot exclude`

- `prunetags` is a file with LD tags format like `./garfield-data/tags/r01/chr22` which is to be used for LD pruning
- `clumptags` is a file with LD tags format like `./garfield-data/tags/r08/chr22` which is to be used for further annotation and feature matching for the permutation testing
- `maftssd` is a file like `./garfield-data/maftssd/chr22` which contains the MAF and TSS distances for feature matching in the permutation testing
- `pvalue` is a file with format as in `./garfield-data/pval/GIANT_HEIGHT/chr22` which contains the p-values from the association analysis
- `annot` is a file with format at in `./garfield-data/annotation/chr22` which contains the annotations of all genetic variants
- `exclude` is a comma separated string of integer values of annotation indices not to be additionally inheriting the annotations of their high LD proxies. This may be useful for very broad annotations such as intergenic regions and repressed segmentation states. An example would be `1,2,3`, which corresponds to the first three annotations in the `./garfield-data/annotation/link_file.txt`. Conversely if no such annotations are to be present `-1` should be used.

The output of the tool is space separated and contains the remaining independent variants after greedy pruning with the following fields

`snpid pvalue clump_tag_count maf tssd a1a2...`

where `snpid` is the genomic position of a variant, `pvalue` is its association analysis p-value, `clump_tag_count` is the number of high LD proxies that variant has, `maf` is its MAF, `tssd` is its distance to the nearest TSS, `a1a2...` are all annotations of that variant after additionally adding the annotations of all high LD proxies with format as in the `./garfield-data/annotation/chr22` file. A sample output is given below

```
723819 0.9018 40 0.0408471 2498 11010010101111110011111111111111001111010011
768448 0.3574 3 0.10815 1462 00000000000000000000000000000000000000000001
779322 0.1321 44 0.12471 -3731 11100110011111100101011111101110101110100000
```

## 4.2 garfield-perm

computes fold enrichment and empirical p-values. For calculating empirical p-values, the genetic variants are put into bins based on number of high LD proxies (`clump_tag_count`), minor allele frequency (MAF) and transcription start site (TSS) distance. Within each bin, p-values between SNPs are permuted and for each permutation fold enrichment is calculated and compared to original fold enrichment.

Usage:

```
garfield-perm -n permuts -a annots -p pthreshs -pt ptheshstest -q nqs -i prepfile \
-o outfile [-g] [-m minit] [-t sigthresh] [-progress]
```

- `-i prepfile` specifies the input file, where `prepfile` is a file created by the `garfield-prep`tool
- `-n` specifies the number of permutations to run, e.g. -n 10000
- `-a` specifies the number of annotations provided, e.g. -a 1005
- `-p` specifies the GWAS p-value thresholds at which to compute fold enrichment, e.g. -p 1e-1,1e-2,1e-3
- `-pt` specifies the GWAS p-value thresholds to be tested against, e.g. -pt 1e-2,1e-3 . Note these must be a subset of the threshold specified with the `-p` option
- `-q` specifies number of binning quantiles for each of the binning dimensions: number of high LD tags, minor allele frequency and nearest transcription start site distance, e.g. `-q n7,m10,t10` will create 7 LD tag bins, 10 MAF bins and 10 TSS bins, respectively, resulting in 7*10*10=700 bins altogether
- `-o outfile` specifies the output file
- `-g` is an optional flag which sets the program to perform greedy p-value estimation, where if at a certain permutation a significant p-value (given by the optional `-t sigthresh`) can no longer be obtained it will stop performing further permutations
- `-t sigthresh` is an option to set the significance threshold for the greedy algorithm. Default value is 1. Suggested value for the provided annotation data is 0.0001.
- `-m minit` is an option to set the minimum number of iterations for the greedy algorithm. Default value in 100 permutations.
- `-progress` shows an intermediate progress bar of the number and percentage of annotations analysed and estimated time to completion. Note: when using with the `-g` option it does not produce a reliable estimate of remaining time.

The output of the tool is space separated and contains the results from the permutation testing along with some summaries and has the following format

```
Index PThresh FE EmpPval NAnnotThesh NAnnot NThresh N Annotation Celltype Tissue Type Category
0 1e-05 2.68463 -1 30 5095 182 82981 AG10803_footprints.txt AG10803 skin footprints Footprints
0 1e-08 2.71446 0.2 12 5095 72 82981 AG10803_footprints.txt AG10803 skin footprints Footprints
1 1e-05 2.46084 -1 36 6670 182 82981 AoAF_footprints.txt AoAF blood_vessel footprints Footprints
1 1e-08 2.59186 0.2 15 6670 72 82981 AoAF_footprints.txt AoAF blood_vessel footprints Footprints
2 1e-05 4.21029 -1 24 2599 182 82981 CD20+_footprints.txt CD20+ blood footprints Footprints
2 1e-08 6.20823 0 14 2599 72 82981 CD20+_footprints.txt CD20+ blood footprints Footprints
3 1e-05 4.13142 -1 39 4304 182 82981 CD34+_Mobilized_footprints.txt CD34+ blood footprints Footprints
3 1e-08 5.62333 0 21 4304 72 82981 CD34+_Mobilized_footprints.txt CD34+ blood footprints Footprints
4 1e-05 1.53121 -1 17 5062 182 82981 fBrain_footprints.txt fBrain fetal_brain footprints Footprints
4 1e-08 1.82144 0.6 8 5062 72 82981 fBrain_footprints.txt fBrain fetal_brain footprints Footprints
5 1e-05 2.50792 -1 25 4545 182 82981 fHeart_footprints.txt fHeart fetal_heart footprints Footprints
5 1e-08 2.53578 0.3 10 4545 72 82981 fHeart_footprints.txt fHeart fetal_heart footprints Footprints
```

where

- `Index` shows the ID of the annotation from the link file

  ```
  ./garfield-data/annotation/link_file.txt
  ```

- `PThresh` is a GWAS threshold used for enrichment analysis testing
- `FE` is the observed fold enrichment for that annotation at that threshold
- `EmpPval` is the empirical p-value of the significance of the observed enrichment. This will be set to -1 if analysis was not performed but the threshold was only used for fold enrichment calculation
- `NAnnotThresh` is the number of (independent) annotated variants with the considered annotation passing the GWAS significance threshold `PThresh` (after pruning)
- `NAnnot` is the total number of (independent) annotated variants with the given annotation (after pruning)
- `Nthresh` is the number of (independent) variants passing the GWAS significance threshold `PThresh` (after pruning)
- `N` is the total number of LD pruned variants.
- `Annotation` specifies a unique name for the annotation
- `Celltype` specifies the cell type of the annotation (if applicable)
- `Tissue` specifies the tissue of the annotation (if applicable)
- `Type` specifies the subtype of the annotation
- `Category` specifies whether the annotation is a Chromatin state, transcription factor binding site, genic annotation, etc...

### 4.3 garfield-plot.R

produces the final figures from the table of results. It has the following usage

```
Rscript garfield-plot.R perm.out nperm output_path_prefix plot_title min thresh
```

- `perm.out` is an output file from the `garfield-perm` tool
- `nperm` is the number of permutations used for the functional enrichment significance testing
- `output_path_prefix` is a path prefix for the output files and figures
- `plot_title` is the title label for the figures. If you do not want a title "" should be used here
- `min` is the minimum number of variants at a certain threshold to be used for filtering the data before plotting
- `thresh` is the significance threshold to be used for plotting. Value of zero sets it to the default value of 0.05/498.

The output of the function is a set of figures for different classes of functional annotations. Sample output files can be found in the `./garfield-data/output/GIANT_HEIGHT` folder, which are a result from the enrichment analysis of the GIANT Height GWAS data.

## 5 Output

### 5.1 Tree structure

The final output is created in subfolders of the `./garfield-data/output/` directory and contains a total of 10 files: an output from the `garfield-prep` step, a final results file from the `garfield-perm` step and 8 figures, one for each of the different types of annotations used in our analysis, namely genic annotations, chromatin segmentation states, transcription factor binding sites, histone modifications and open chromatin data (FAIRE, DHS Hotspots, peaks and footprints).

```
garfield-data
|----- output
   |----- cd-meta
   |  |----- garfield.cd-meta.Chromatin_States.pdf
   |  |----- garfield.cd-meta.FAIRE.pdf
   |  |----- garfield.cd-meta.Footprints.pdf
   |  |----- garfield.cd-meta.Genic.pdf
   |  |----- garfield.cd-meta.Histone_Modifications.pdf
   |  |----- garfield.cd-meta.Hotspots.pdf
   |  |----- garfield.cd-meta.Peaks.pdf
   |  |----- garfield.cd-meta.TFBS.pdf
   |  |----- garfield.perm.cd-meta.out
   |  |----- garfield.prep.cd-meta.out
```

```
|----- GIANT_HEIGHT
   |----- garfield.GIANT_HEIGHT.Chromatin_States.pdf
   |----- garfield.GIANT_HEIGHT.FAIRE.pdf
   |----- garfield.GIANT_HEIGHT.Footprints.pdf
   |----- garfield.GIANT_HEIGHT.Genic.pdf
   |----- garfield.GIANT_HEIGHT.Histone_Modifications.pdf
   |----- garfield.GIANT_HEIGHT.Hotspots.pdf
   |----- garfield.GIANT_HEIGHT.Peaks.pdf
   |----- garfield.GIANT_HEIGHT.TFBS.pdf
   |----- garfield.perm.GIANT_HEIGHT.out
   |----- garfield.prep.GIANT_HEIGHT.out
```

## 5.2 Results output file

An example of the `./garfield-data/output/cd-meta/garfield.perm.cd-meta.out` file is shown below

```
Index PThresh FE EmpPval NAnnotThesh NAnnot NThresh N Annotation Celltype Tissue Type Category
0 1e-05 2.68463 -1 30 5095 182 82981 AG10803_footprints.txt AG10803 skin footprints Footprints
0 1e-08 2.71446 0.2 12 5095 72 82981 AG10803_footprints.txt AG10803 skin footprints Footprints
1 1e-05 2.46084 -1 36 6670 182 82981 AoAF_footprints.txt AoAF blood_vessel footprints Footprints
1 1e-08 2.59186 0.2 15 6670 72 82981 AoAF_footprints.txt AoAF blood_vessel footprints Footprints
2 1e-05 4.21029 -1 24 2599 182 82981 CD20+_footprints.txt CD20+ blood footprints Footprints
2 1e-08 6.20823 0 14 2599 72 82981 CD20+_footprints.txt CD20+ blood footprints Footprints
3 1e-05 4.13142 -1 39 4304 182 82981 CD34+_Mobilized_footprints.txt CD34+ blood footprints Footprints
3 1e-08 5.62333 0 21 4304 72 82981 CD34+_Mobilized_footprints.txt CD34+ blood footprints Footprints
4 1e-05 1.53121 -1 17 5062 182 82981 fBrain_footprints.txt fBrain fetal_brain footprints Footprints
4 1e-08 1.82144 0.6 8 5062 72 82981 fBrain_footprints.txt fBrain fetal_brain footprints Footprints
5 1e-05 2.50792 -1 25 4545 182 82981 fHeart_footprints.txt fHeart fetal_heart footprints Footprints
5 1e-08 2.53578 0.3 10 4545 72 82981 fHeart_footprints.txt fHeart fetal_heart footprints Footprints
```

where

- `Index` shows the ID of the annotation from the link file

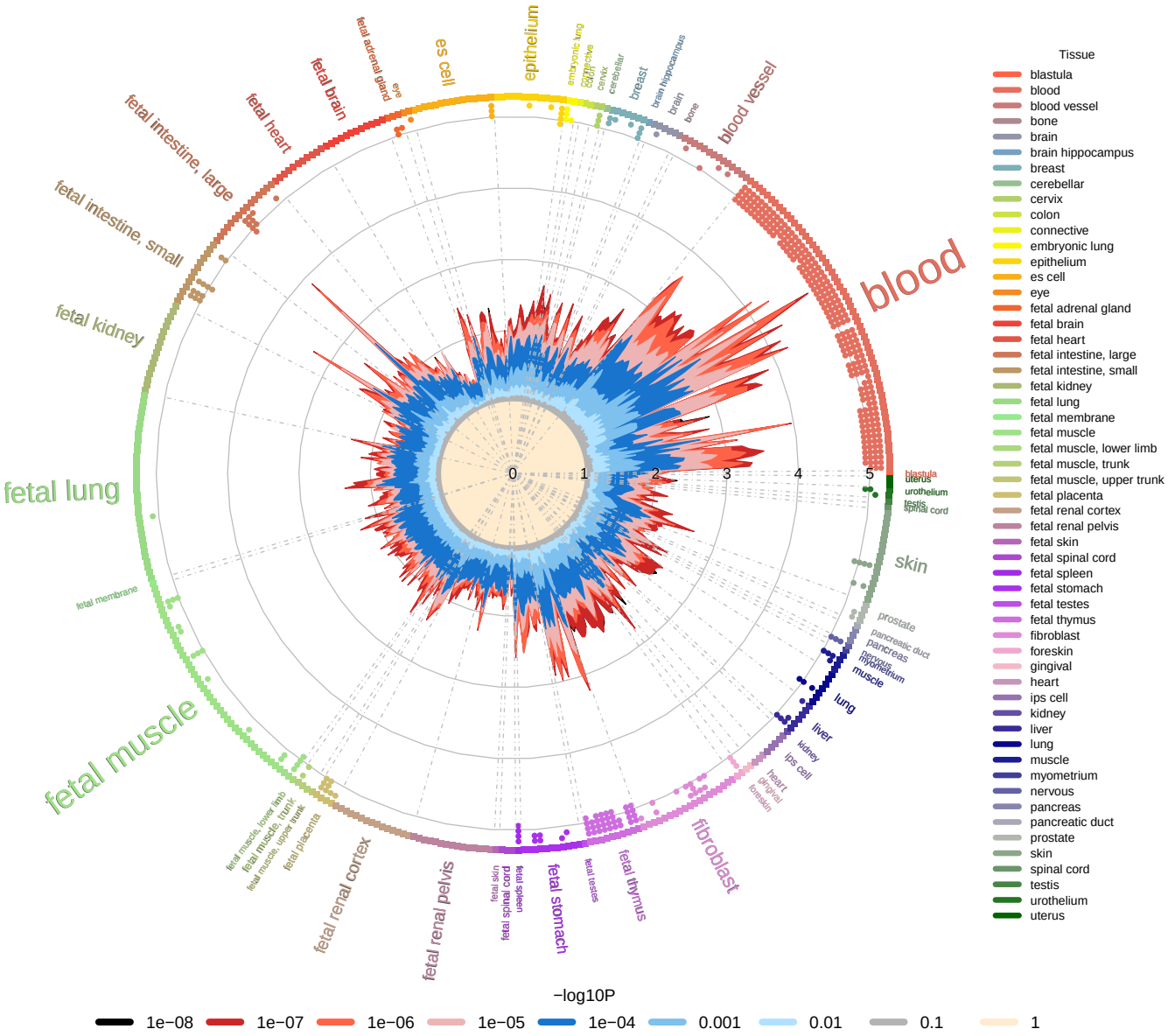
  `./garfield-data/annotation/link_file.txt`

- `PThresh` is a GWAS threshold used for enrichment analysis testing
- `FE` is the observed fold enrichment for that annotation at that threshold
- `EmpPval` is the empirical p-value of the significance of the observed enrichment. This will be set to -1 if analysis was not performed but the threshold was only used for fold enrichment calculation
- `NAnnotThresh` is the number of (independent) annotated variants with the considered `Annotation` passing the GWAS significance threshold `PThresh` (after pruning)
- `NAnnot` is the total number of (independent) annotated variants with the given annotation (after pruning)

- `Nthresh` is the number of (independent) variants passing the GWAS significance threshold `PThresh` (after pruning)
- `N` is the total number of LD pruned variants.
- `Annotation` specifies a unique name for the annotation
- `Celltype` specifies the cell type of the annotation (if applicable)
- `Tissue` specifies the tissue of the annotation (if applicable)
- `Type` specifies the subtype of the annotation
- `Category` specifies whether the annotation is a Chromatin state, transcription factor binding site, genic annotation, etc...

## 5.3 Figures

`./garfield-data/output/cd-meta/garfield.cd-meta.Hotspots.pdf`




Enrichment of Crohn's Disease variants in DNaseI Hypersensitive sites (broad peaks) from ENCODE and Roadmap Epigenomics data. Radial plot shows the fold enrichment in each cell type (dots on the outside of the circle sorted by tissue) for each GWAS significance threshold between $10^{-8}$ and all (independent) variants (shown by inner colours and bottom legend). Furthermore, small dots on the outer side of the plot show if the observed enrichment is significant (if there is a dot present) or not (if there isn't) for thresholds $10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}$ in direction from outside to inside.

# References

Valentina Iotchkova, Graham Ritchie, Matthias Geihs, Josine Min, Ian Dunham, Ewan Birney and Nicole Soranzo. *GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction.* In preparation