# FM-pipeline

This is a pipeline for finemapping using GWAS summary statistics, implemented in Bash as a series of steps to furnish an incremental analysis. As depicted in the diagram below



*LocusZoom plot showing Regional association for chr1:39114617-39614617*

where our lead SNP rs4970634 is in LD with many others, the procedure attempts to identify causal variants from region(s) showing significant SNP-trait association.

The process involves the following steps, 1. Extraction of effect (beta)/z statistics from GWAS summary statistics (.sumstats), 2. Extraction of correlation from the reference panel among overlapped SNPs from 1 and the reference panel containing individual level data. 3. Information from 1 and 2 above is then used as input for finemapping.

The measure of evidence is typically (log10) Bayes factor (BF) and associate SNP probability in the causal set.

Software included in this pipeline are listed in the table below.

| Name | Function | Input | Output | Reference |
| --- | --- | --- | --- | --- |
| CAVIAR | finemapping | z, correlation matrix | causal sets and probabilities | Hormozdiari, et al. (2014) |
| CAVIARBF | finemapping | z, | BF and | Chen, et al. |

| | | correlation matrix | probabilities for all configurations | (2015) |
|---|---|---|---|---|
| GCTA | joint/conditional analysis | .sumstats, reference data | association results | Yang, et al. (2012) |
| FM-summary | finemapping | .sumstats association results | updated results | Huang, et al. (2017) |
| JAM | finemapping | beta, individual reference data | Bayes Factor of being causal | Newcombe, et al. (2016) |
| LocusZoom | regional plot | partial .sumstats | .pdf/.png plots | Pruim, et al. (2010) |
| fgwas | functional GWAS | | | Pickrell (2014) |
| finemap | finemapping | z, correlation matrix | causal SNPs and configuration | Benner, et al. (2016) |

so they range from regional association plots via LocusZoom, joint/conditional analysis via GCTA, functional annotation via fgwas to dedicated finemapping software including CAVIAR, CAVIARBF, an adapted version of FM-summary, R2BGLiMS/JAM and finemap. One can optionally use a subset of these for a particular analysis by specifying relevant flags from the pipeline's settings.

## INSTALLATION

On many occasions, the pipeline takes advantage of the GNU parallel.

Besides (sub)set of software listed in the table above, the pipeline requires qctool 2.0, PLINK 1.9, and the companion program LDstore from finemap's website need to be installed.

The pipeline itself can be installed in the usual way,

```
git clone https://github.com/jinghuazhao/FM-pipeline
```

The setup is in line with summary statistics from consortia where only RSid are given for the fact that their chromosomal position may be changed over different builds. To remedy this, we use information from UCSC, e.g.,

```
wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/snp150.txt.gz
gunzip -c snp150.txt.gz | \
awk '{split($2,a,"_");sub(/chr/,"",a[1]);print a[1],$4,$5}' | \
sort -k3,3 > snp150.txt
```

Note that JAM requires Java 1.8 so call to Java -jar inside the function needs to reflect this, not straightforward with `install_github()` from `devtools` but one needs to clone the package, modify the R source code and then use

```
git clone https://github.com/pjnewcombe/R2BGLiMS
### change java to java-1.8 in R2BGLiMS/R/R2BGLiMS.R
R CMD INSTALL R2BGLiMS
```

Implementations have been done for the finemapping software along with LocusZoom and GCTA; support for fgwas is still alpha tested. To facilitate handling of grapahics, e.g., importing them into Excel, pdftopng from xpdf is used.

We use Stata and Sun grid engine (sge) for some of the data preparation, which would become handy when available.

## USAGE

Before start, settings at the beginning of the script need to be changed and only minor change is expected after this. The syntax of pipeline is then simply

```
bash fmp.sh <input>
```

## Inputs

### --- GWAS summary statistics ---

These include the following columns,

| Column | Name | Description |
|---|---|---|
| 1 | SNP | RSid |
| 2 | A1 | Effect allele |
| 3 | A2 | Other allele |
| 4 | freqA1 | A1 frequency |
| 5 | beta | effect estimate |
| 6 | se | standard error of effect |
| 7 | P | P-value |
| 8 | N | sample size |
| 9* | chr | chromosome |
| 10* | pos | position |

This format is in line with joint/conditional analysis by GCTA. Note the last two columns are not always available but can be obtained from UCSC as above; see below for example use.

### --- Reference panel ---

The pipeline uses a reference panel in a .GEN format, taking into account directions of effect in both the GWAS summary statistics and the reference panel. Its development will facilitate

summary statistics from a variety of consortiua as with reference panels such as the HRC and 1000Genomes.

A .GEN file is required for each region, named such that chr{chr}_{start}_{end}.gen, together with a sample file. For our own data, a utility program in Stata is written to generate such files from their whole chromosome counterpart using SNPinfo.dta.gz which has the following information,

| chr | snpid | rsid | pos | FreqA2 | info | type | A1 | A2 |
|-----|-------|------|-----|--------|------|------|----|----|
| 1 | 1:54591_A_G | rs561234294 | 54591 | .0000783 | .33544 | 0 | A | G |
| 1 | 1:55351_T_A | rs531766459 | 55351 | .0003424 | .5033 | 0 | T | A |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

Optionally, a file is specified which contains samples to be excluded from the reference panel; one leaves it unspecified when not needed. In line with qctool -excl-samples option, it contains a list of individuals corresponding to ID_2 of the sample file rather than ID_1 and ID_2.

### --- The lead SNPs ---

Given these, one can do away with Stata and work on a text version for instance SNPinfo.txt. An auxiliary file called `st.bed` contains chr, start, end, rsid, pos, r corresponding to the lead SNPs specified and r is a sequence number of region. As GCTA conditional/joint analysis requires whole chromosome reference the counterpart is HRC.do. Note in this case the snpid and rsid variables are called rsid and RSnum instead; both porgrams filter SNPs on minor allele count and measure of imputation quality. As it is very slow, we use .bgen instead see the section on WHOLE-GENOME CONDITIONAL/JOINT ANALYSIS below.

## Outputs

The output will involve counterpart(s) from individual software, i.e., .set/post, caviarbf, .snp/.config, .jam/.top

| Software | Output type | Description |
|----------|-------------|-------------|
| CAVIAR | .set/.post | causal set and probabilities in the causal set/posterior probabilities |
| CAVIARBF | .caviarbf | causal configurations and their BFs |
| FM-summary | .txt | additional information to the GWAS summary statistics |
| JAM | .jam/.top | the posterior summary table and top models containing selected SNPs |
| finemap | .snp/.config | top SNPs with largest |

> log10(BF) and top
> configurations as with their
> log10(BF)

It is helpful to examine directions of effects together with their correlation which is now embedded as with finemap.

## WHOLE-GENOME CONDITIONAL/JOINT ANALYSIS

As the pipeline works on regsions defined by lead SNPs, it is desirable to have a genomewide counterpart and currently this is possible with GCTA and we have a script called gcta-slct.sh which accepts a single sumstats file, and only a minor change is required, namely,

```
gcta-slct.sh <input>
```

At the end of the script, it also shows how the relevant information was generated. As it is very time-consuming for interactive use, on our system we resort to sge, e.g.,

```
qsub -cwd -e HRC.err -o HRC.out -pe make 10 -q all.q /genetics/bin/gcta-
slct.sh HRC
```

so the job is sent to the clusters instead.

The use of gene list from the analysis can be fed into VEGASv2 as shown with vegas2v2.sh but some changes are required for the command-line version described at the of the script. We don't have experiences with the web-based interface from https://vegas2.qimrberghofer.edu.au/. Nevertheless, as indicated in the original VEGAS paper (Liu et al. 2010), `If a gene contains only one causal variant, then the inclusion of a large number of nonsignificant markers into the gene-based test will dilute this gene's significance.`

However, more broadly software in PW-pipeline can be used and in terms of LD information PASCAL will be useful.

## EXAMPLE

We show how to set up for BMI GWAS summary data as reported by the GIANT consortium, Locke, et al. (2015),

```
# GWAS summary statistics
wget
http://portals.broadinstitute.org/collaboration/giant/images/1/15/SNP_gwas_mc
_merge_nogc.tbl.uniq.gz
gunzip -c SNP_gwas_mc_merge_nogc.tbl.uniq.gz |
awk 'NR>1' | \
join -11 -23 - snp150.txt | \
awk '($9!="X" && $9!="Un")' > bmi.txt

# A list of 97 SNPs
R --no-save <<END
```

```
library(openxlsx)
xlsx <- "https://www.nature.com/nature/journal/v518/n7538/extref/nature14177-
s2.xlsx"
snps <- read.xlsx(xlsx, sheet = 4, colNames=FALSE, skipEmptyRows = FALSE,
cols = 1, rows = 5:101)
snplist <- sort(as.vector(snps[,1]))
write.table(snplist, file="97.snps", row.names=FALSE, col.names=FALSE,
quote=FALSE)
END

# st.bed
grep -w -f 97.snps snp150.txt | \
sort -k1,1n -k2,2n | \
awk -vflanking=250000 '{print $1,$2-flanking,$2+flanking,$3,$2,NR}' > st.bed
```

where we download the GWAS summary statistics adding SNP positions in build 37 rather than 36. The list of SNPs can also be used to generate st.bed as above.

We illustrate use of 1000Genomes data, available as FUSION LD reference panel, with 1KG.sh to generate SNPinfo.dta.gz and st.do to generate the required data.

## ACKNOWLEDGEMENTS

## SOFTWARE AND REFERENCES

### CAVIAR

Hormozdiari F, et al. (2014) Identifying Causal Variants at Loci with Multiple Signals of Association. Genetics, 44, 725–731

### CAVIARBF

Chen W, et al. (2015) Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics 200:719-736.

### FM-summary

Huang H, et al (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 547, 173–178, doi:10.1038/nature22969

### GCTA

Yang J, et al. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44:369-375

## JAM

Newcombe PJ, et al. (2016) JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. Genet Epidemiol 40:188–201

## LocusZoom

Pruim RJ, et al. (2010) LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics 2010 September 15; 26(18): 2336.2337

## fgwas

Pickrell JK (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. bioRxiv 10.1101/000752

## finemap

Benner C, et al. (2016) FINEMAP: Efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493-1501

Benner C, et al. (2017) Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. Am J Hum Genet 101(4):539-551

## VEGAS paper

Liu JZ, et al. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. Am J Hum Genet 87:139–145.

## GIANT paper

Locke AE, et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518(7538):197-206. doi: 10.1038/nature14177