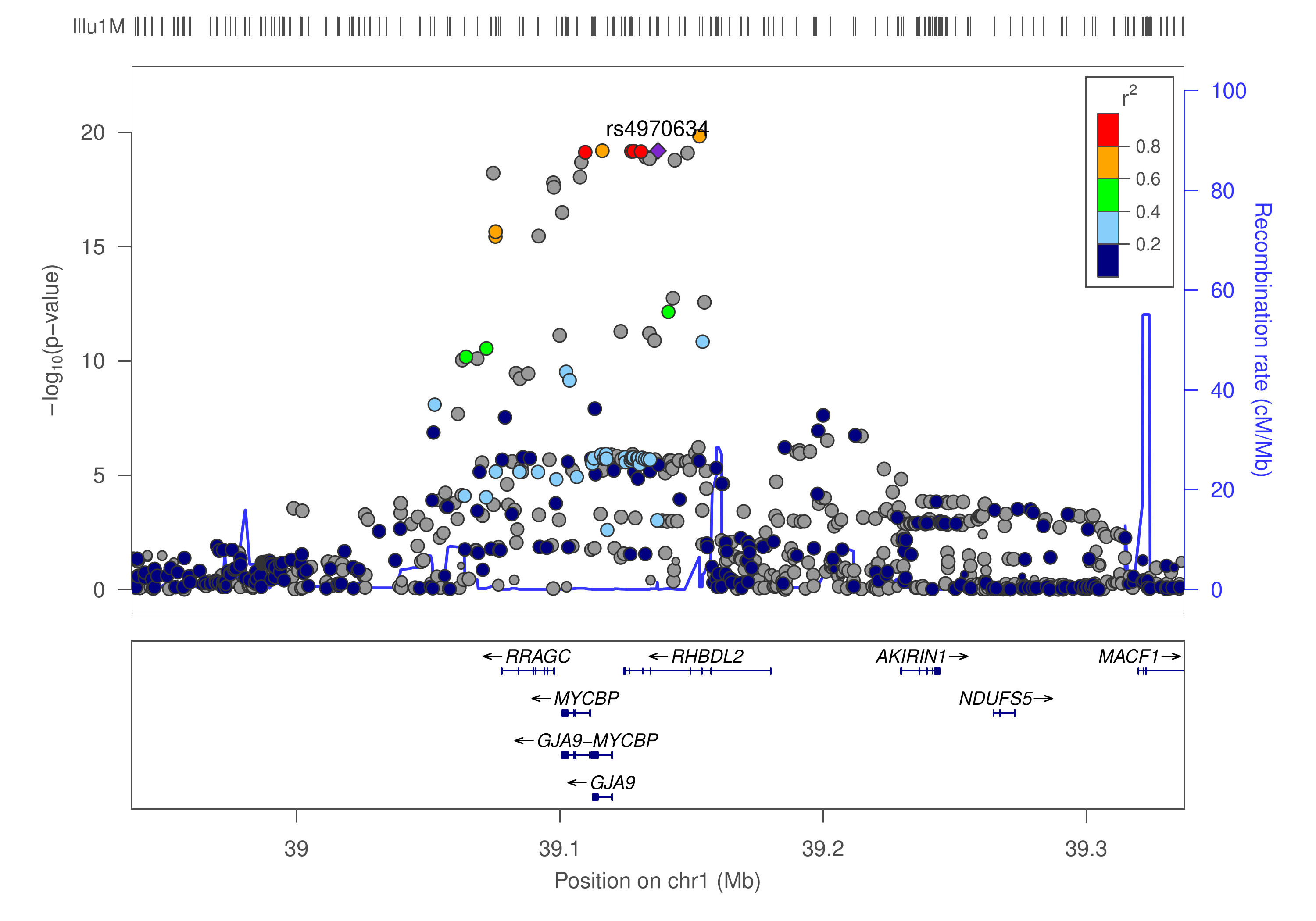
# FM-pipeline

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, correlation matrix | BF and probabilities for all configurations | Chen, et al. (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](http://www.gnu.org/software/parallel/).

Besides (sub)set of software listed in the table above, the pipeline requires [GTOOL](http://www.well.ox.ac.uk/%7Ecfreeman/software/gwas/gtool.html), [PLINK](https://www.cog-genomics.org/plink2) 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, i.e.,

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](https://www.xpdfreader.com/) is used.

## 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 and lead SNPs ---

The **first input file** will be GWAS summary statistics with 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-vale |
| 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 normally required as they are obtained from UCSC (use\_UCSC=1). However, for one of our analyses UCSC does not have all the coordinates as in the GWAS summary statistics, so we set use\_ucsc==0 and provide st.bed containing the (chr, start, end, pos, rsid, r) sextuplets.

The **second input file** is a list of SNPs for which finemapping will be conducted.

A header is required for neither file.

### --- 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](files/p0.do) is written to generate such files from their whole chromosome counterpart using SNPinfo.dta.gz which has the following information,

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| chr | rsid | RSnum | 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 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

Given these, one can do away with Stata and work on a text version for instance SNPinfo.txt. When option stbed=1 in the settings, it only generates st.bed which contains chr, start, end, RSid, pos corresponding to the lead SNPs specified.

Optionally, a file is specified which contains sample to be excluded from the reference panel; one leaves it unspecified when not needed

## 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 the correlation of them, e.g., for use with finemap, the code [here](files/finemap-check.R) is now embedded in the pipeline.

## EXAMPLES

1. We use GWAS on 2-hr glucose level as reported by the MAGIC consortium, Saxena, et al. (2010). The GWAS summary data is obtained as follows,

* wget ftp://ftp.sanger.ac.uk/pub/magic/MAGIC\_2hrGlucose\_AdjustedForBMI.txt  
  awk -vOFS="\t" -vN=15234 '(NR>1){print $0, N}' MAGIC\_2hrGlucose\_AdjustedForBMI.txt > 2hrglucose.txt
* For two SNPs contained in [2.snps](files/2.snps), the Stata program [p0.do](files/p0.do) generates [Extract.sh](files/Extract.sh) excluding SNPs in [exc3\_122844451\_123344451.txt](files/exc3_122844451_123344451.txt) and [exc3\_122881254\_123381254.txt](files/exc3_122881254_123381254.txt). The command to call is
* bash fmp.sh 2hrglucose.txt

1. Next 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' > 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
* so the GWAS summary statistics from GIANT is almost ready (we only drop the header) as with the list of 97 SNPs. The positions of these SNPs were in build 36 while we used build 37.

In both cases, the GWAS summary data are used togther with the reference panel in .GEN format to furnish the finemapping analysis.

## ACKNOWLEDGEMENTS

The work was motivated by finemapping analysis at the MRC Epidemiology Unit and inputs from authors of GCTA, finemap, JAM, FM-summary as with participants in the Physalia course Practical GWAS Using Linux and R are greatly appreciated. In particular, the Stata program p0.do was adapted from code originally written by Dr Jian'an Luan.

## SOFTWARE AND REFERENCES

[**CAVIAR**](https://github.com/fhormoz/caviar)

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

[**CAVIARBF**](https://bitbucket.org/Wenan/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**](https://github.com/hailianghuang/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**](cnsgenomics.com/software/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**](https://github.com/pjnewcombe/R2BGLiMS)

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

[**LocusZoom**](http://locuszoom.sph.umich.edu/)

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

[**fgwas**](https://github.com/joepickrell/fgwas)

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

[**finemap**](http://www.christianbenner.com/#)

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

**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

**MAGIC paper**

Saxena R, et al. (2010) Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet 42:142-148