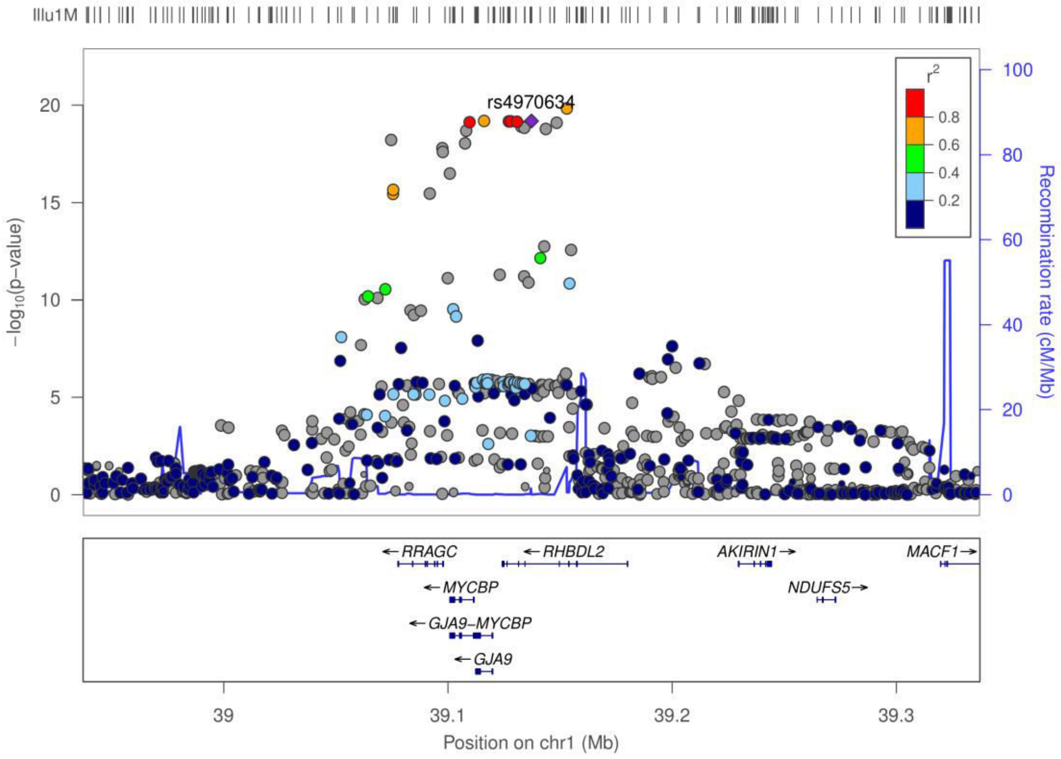
FM-pipeline

FineMapping analysis using GWAS summary statistics

INTRODUCTION

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.

Information on whole-genome analysis, which could be used to set up the regions, are described at the repository's wiki page.

INSTALLATION

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

Option	Name	Function	Input	Output	Reference
CAVIAR	CAVIAR	finemapping	z, correlation matrix	causal sets and probabilities	Hormozdiari, et al. (2014)
CAVIARBF	CAVIARBF	finemapping	z, correlation matrix	BF and probabilities for all configurations	Chen, et al. (2015)
GCTA	GCTA	joint/conditional analysis	.sumstats, reference data	association results	Yang, et al. (2012)
FM_summary	FM-summary	finemapping	.sumstats	posterior probability & credible set	Huang, et al. (2017)
JAM	JAM	finemapping	beta, individual reference data	Bayes Factor of being causal	Newcombe, et al. (2016)
LocusZoom	LocusZoom	regional plot	.sumstats	.pdf/.png plots	Pruim, et al. (2010)
fgwas	fgwas	functional GWAS	.sumstats	functional significance	Pickrell (2014)
finemap	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.

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.

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 XpdfReader 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.

```
# software flags: 1=enable

export clumping=0
export CAVIAR=0
export CAVIARBF=0
export FM_summary=0
export GCTA=0
export JAM=1
export LocusZoom=0
export fgwas=0
export finemap=1
# parallel processes when available
export threads=1
# Default location and reference data
export FM_location=/genetics/bin/FM-pipeline
# GEN files named chr{chr}_{start}_{end}.gen.gz
export GEN_location=$FM_location/1KG/LD-blocks
# sample file
export sample_file=$FM_location/1KG/EUR.sample
# Complementary files for fgwas
export fgwas_location_1kg=/genetics/data/fgwas/1000-genomes-genetic-maps

## OTHER SETTINGS

export PATH=/genetics/bin:/usr/local/bin:$PATH:/genetics/data/software/bin
export R_LIBS=/usr/local/lib/R/site-library/:/genetics/bin/R:/usr/local/lib64/R/library:/genetics/data/software/R
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/lib64/R/lib:/genetics/data/software/lib
```

The syntax of the pipeline is then simply

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

Inputs

--- GWAS summary statistics ---

The input will be GWAS summary statistics described at https://github.com/jinghuazhao/SUMSTATS.

This format is in line with joint/conditional analysis by GCTA.

--- 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, st.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
...	...	...	...	...	...	...	...	...

Note that unlike fmp.sh, the utility program uses qctool-1.4 for its more comprehensive options. 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.

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
GCTA	.jma.cojo	joint/conditional analysis results

JAM	.jam/.top/.cs	posterior summary table, top models containing selected SNPs and credible sets
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 when finemap is also called.

In addition, we have implemented clumping using PLINK with options comparable to those used in depict (e.g. description in PW-pipeline).

EXAMPLE

Files `bmi.txt` and `97.snps` are described in https://github.com/jinghuazhao/SUMSTATS.

--- 1000Genomes panel using approximately indepdent LD blocks ---

This is available as FUSION LD reference panel, with 1KG.sh to generate `SNPinfo.dta.gz` and st.do to generate the script Extract.sh for the required data.

We then proceed with.

```
awk '{gsub(/chr/,"",$0);if(NR==1) {print "chr","start","end","region"} else
print $1,$2,$3,$4}' 1KG/EUR.bed > st.bed
cp bmi.txt 1KG
cp fmp.sh 1KG.sh
# modify 1KG.sh to use the 1KG panel
1KG.sh 1KG
```

and the results will be in `1KG.out`.

--- HRC panel ---

File `97.snps` is used to build `st.bed` and the analysis proceed as follows,

```
# 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
cp fmp.sh HRC.sh
# modify HRC.sh to use the HRC panel
export GEN_location=/scratch/tempjhz22/LDcalc/HRC
HRC.sh HRC
```

and the results will be in `HRC.out`.

ADDITIONAL TOPICS

The wiki page has the following information,

- Whole-genome conditional/joint analysis

- Whole genome analysis using approxmiately independent LD blocks.

RELATED LINK

Credible sets are often described, see https://github.com/statgen/gwas-credible-sets

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 utility program in Stata was adapted from p0.do (which is still used when LD_MAGIC is enabled) originally written by Dr Jian'an Luan and computeCorrelationsImpute2forFINEMAP.r by Ji Chen from the MAGIC consortium who also provides code calculating the credible set based on finemap configurations. Earlier version of the pipeline also used GTOOL.

SOFTWARE AND REFERENCES

CAVIAR (Causal Variants Identification in Associated Regions)

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

CAVIARBF (CAVIAR Bayes Factor)

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 (Genome-wide Complex Trait Analysis)

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 (Joint Analysis of Marginal statistics)

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 (Functional genomics and genome-wide association studies)

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