# Exercise for Fine-mapping and visualisation in Post-GWAS Analysis

Jean-Tristan Brandenburg Scott Hazelhurst
Sydney Brenner Institute for Molecular Bioscience
University of the Witwatersrand, Johannesburg

February 2023

## 1 Overview

### 1.1 Organisation

There are three sessions of exercises: data management using R (optional, but recommended), visualisation of GWAS and fine-mapping.

### 1.2 Data description

### 1.3 Description

The data are a combination of available data and simulated data. The genotypes are from individuals in the 1000 Genomes Project (varied populations), after quality control in build hg19. Positions of the array have been extracted from raw data and imputed using Sanger imputation server using African data sets as a reference panel. 500 individuals have been sampled and phenotypes have been simulated using effects extracted from GWAS result in the GWAS catalogue. A simulated diabetes phenotype and the *gcta* tool uses. GWAS was done using *gemma* and *plink*. Simulation and GWAS Data was done using h3abionet/h3agwas workflow (for *future* reference, the data that we used can be found here h3abionet/h3agwas-example, but don't download it as we'll give more specific instructions later.

### 1.4 Data and material access

Data are available on google drive and `https://github.com/jeantristanb/course_postgwas_visuFinemap_2023`.

- The repository contains material you will need for exercises.

  ```
  git clone https://github.com/jeantristanb/course_postgwas_visuFinemap_2023.git
  ```

  The repository of data is split by exercise: Q-Q plot, Manhattan plot, Fine-Mapping, Meta Analysis; Phenotype by genotype and regional plot.

## 1.5 Software requirement

We're expecting that you've installed a standard Linux distribution and have R at least 3.6 (we've tested on Ubuntu 20 and 22 with R 3.6 and 4.1). This requires some Linux and R packages. We also require *LocusZoom* and *finemap*. Because the downloads are quite big we haven't put it in the docker container, and we've also broken the steps down into several steps to make sure that we can pick up if anything has gone wrong. Please change directory to the `course_postgwas_visuFinemap_2023/install` directory.

- Install Linux and R packages `sudo source install.sh`
- Download and install *finemap*. Do each line in turn to make sure that everything works

```
wget -c http://www.christianbenner.com/finemap_v1.4.1_x86_64.tgz
tar -xzf finemap_v1.4.1_x86_64.tgz
sudo cp finemap_v1.4.1_x86_64/finemap_v1.4.1_x86_64 /usr/local/bin/finemap
sudo chmod a+rx /usr/local/bin/finemap
```

### Locus Zoom Stand Alone (optional)

Locus zoom is a use full software to do regional plot, but need to download a file of 14 GB and installation need a space requirements of 88.5 GB on your computer after untar

- Download and install LocusZoom. It is big so be patient. We recommend that you do this per classroom. Download one copy and then share.
  - Change directory to where you want to install.
  - Download
    `wget -c https://statgen.sph.umich.edu/locuszoom/download/locuszoom_1.4.tgzw`
    or
    `wget -c https://www.bioinformatics.africa/locuszoom_1.4.tgzw`
  - Untar LocusZoom: `tar -xf locuszoom_1.4.tgz`
    If you get an error message, the `locuszoom_1.4.tgz` file is corrupt. Delete it and download again.
  - Put this in your your PATH.
    Execute: `pwd`
    You will see your directory name
    Edit your `.bashrc` file so that the following line is at the end, replacing the directory name shown for the text DIRECTORY below.

```
export PATH=$PATH:DIRECTORT
```

A Docker file can be found in `docker` folder contained software used without locus zoom due to size of data.

# 2 Data management with R (optional, but highly recommended)

**Objective** open a data set with **R**, analyse the structure of the data set, sub-select the data set. We will use a data set that you can download and uncompress

```
wget https://www.dropbox.com/s/yqf9krvsu0fxu87/course_h3abionet_2023.gemma.gz
```

```
gunzip course_h3abionet_2023.gemma.gz
```

You can find in `exercise/Datamanagement` information about the data set

### Analyse the data

1. Run **R**
2. open your data set using *read.table* (base function) and *fread* from **data.table R-library**, compare behaviours of two function.
3. using R function : *summary*, *header*, *nrow*, *ncol* described your data-set :
   - how many columns?
   - how many rows or positions in your data set?
   - determine header for column of interest : chromosome, positions, allele1, allele0, beta, standart error and p-value
4. do you have minor allele frequency less than 0.01?
5. extracted position from chromosome 17 and around positions 78727734 with a windows of 1 MB
6. find positions with min $p$-value (using *which.min*).

# 3 Exercise: visualisation

## 3.1 QQ-plot

**Objective:** plot a quantile-quantile plot.

**Data:** data are available `exercise/QQplot/`, folder data contains one file with a subsample of summary statistics

### Build your own QQ-plot

1. open file using *fread* from **data.table** lib
2. find the observed $p$-value: extract the $p$-value, transform using $-\log 10$ (- see *log* function) and sort using *sort* function
3. get expected $p$-value: using $-\log 10$ and *ppoints* function
4. using plot function to plot observed and expected $p$-values.
5. save your figure using *jpeg*, *png* or *tiff* function followed by *dev.off()* (*pdf* due to high number of point should be avoided)

**Build a QQ-plot useing a fancy R-library** Build a QQ-plot using **fastman** library or **qqman** library.

**Computed inflation factors** Inflation factors can be used measure bias of your QQ-Plot, usinng the formula:
$\chi^2 = q^2_\chi(1-P,1)$ with $P$ p-value and $q_{\chi^2}$ quantile function of $\chi^2$ with 1 degrees of freedom and $\lambda = \text{median}(\chi^2)/0.456$
Using R compute your own inflation factors where $q^2_\chi$ for function is *qchisq* function in **R** :

- open data set using *fread*
- extract summary statistics
- using P-value computed inflation factors.

### Using web interface LocusZoom V2

Using LocuZoom v2 and data at url `https://my.locuszoom.org/gwas/91333/`, visualise QQ-plot and inflation factors gave by web interface.

## 3.2 Manhattan plot

**Objective:** Plot a Manhattan plot. Manhattan plots represent the $p$ values of the entire GWAS on a genomic scale.

**Data** Data are available at `exercise/Manhattan/`, folder data contained two file with two sub sample of summary statistics

**Exercise** Use the following R libraries:

- using 1 data set, use *manhattan* function from **qqman** library
- using 1 data set, *fastman* function from **fastman** library: highlight lead SNPs with $p < 5 \times 10^{-8}$
- using 2 data sets and *gmirror* function from **hudson** library, highlight lead SNPs with $p < 5 \times 10^{-8}$.

**Using web interface LocusZoom V2** Using LocusZoom v2 and data at url : `https://my.locuszoom.org/gwas/91333/`: visualise Manhattan plot and lead SNPs.

## 3.3 Regional plot

**Objective:** do regional plot using LocusZoom.

### 3.3.1 Using web interface LocusZoom V2

**RPTOR region** Using locuszoom v2 at url `https://my.locuszoom.org/gwas/91333/region/?chrom=17&start=78507626&end=79007626`,

- What is genes around?
- what is lead positions?
- What previous genome wide association studies had been found for the locus?
- Change LD population, what change of your plot?

### 3.3.2 Exercise with *LocusZoom* stand alone - Optional -:

**Data:** Data are genetics data in plink format and summary statistics around region of interest. Data can be found in `exercise/Regional_plot`

If you have been able to download LocusZoom successfully you can do this exercise by running LocusZoom on your computer. Otherwise, you can go to `https://my.locuszoom.org` and sign in with a Google account and you can do this exercise using the web version

- LocusZoom uses as input summary statistics in format epact, with header : *#CHROM BEGIN END MARKER_ID PVALUE* using **R** or **shell** command (*head*, *grep*, *awk*)
- plot a locus zoom on chromosome 17 wit start 77727734 and end 79727734 with lead locus : 78727734
- defined parameter of locus zoom build (–build), LD database (–source and –pop) and GWAS catalog (–gwas-cat) to plot your regional plot.

**Exercise with Locus Zoom Stand Alone and your own LD – optional:**

- Run LocusZoom using your own LD data – the regional plot will give you a better relation between lead SNPs and SNPs in LD if you use your own data.
  - use *PLINK* to build LD using *–ld-snp* option with lead SNPs
  - the PLINK output file needs to small reformatting to make it suitable for LocusZoom.

    Add `snp1 snp2 dprime rsquare` as header (you need to add NA in *dprime* column).
  - run your new file with (*–ld*)
  - defined other parameters of LocusZoom build (*–build*) and *GWAS catalog* (*–gwas-cat*) to plot your regional plot.

## 3.4 Plot phenotype in function of your genotype

**Objective:** observation of distribution of phenotype or residual between your 3 genotype (homozygous reference, heterozygous and homozygous alternative)

**Data** : genotype data and phenotype data from position that we want to plot `exercise/PhenoPlotGeno/`.

**Exercise**

- Extract genotype of chromosome 17 positions 78727734 using *plink* in tab format ( *–recode tab*)
- Open R
- read phenotype
- read genotype
- merge genotype and phenotype.
- using *boxplot* function, plot phenotype by genotype.

**Exercise – optional** using *an_plotboxplot.r* plot your phenotype in function of your genotype, compared $p$-value with $p$-value of your GWAS what is difference?

### 3.5 Meta Analysis: forest plots – optional

A forest plot is a graphical display of estimated results from a number of GWASs studies addressing the same question, the result of GWAS using same positions and phenotypes.

**Data** Summary statistics result of association in different populations and meta analyse result can be found in `excercise/MetaAnalyse`

**Exercise**

- Open each file of association result and meta analyse
- Format header of each file selected, select positions of interest
- Concatenate different data and add column contains population information
- Build your forest plot using **ggplot2** with function : *geom_point*, *geom_errorbarh*, *scale_y_continuous*
- Build your forest plot using *forestplot* function from **forestplot**

## 4 Exercise: Fine-Mapping

An overview of the *finemap* program can be found at `http://www.christianbenner.com/`

**Objectives** Run a fine-mapping software based on Bayesian methods to identify number of causal/lead variant in the region and credible interval sets at 95 %.

**Data set** We can use the same data described in the regional plot and additional data can be found in `exercise/Finemapping`.

**Exercise** Using *finemap*, find the lead SNPs and credible interval set, using data in a 100kb window of chromosome 17 starting at position 78727734.

- Summary statistics :
  - open R
  - open summary statistics : how many positions?
  - open file contain positions files (bim ) : how many positions ?
  - Find positions in common to the genotype and summary statistics data : merge both file using *merge* and all=F : how many positions are left
  - Input of finemap : Change header the header in the summary statistics file: `rsid chromosome positio` and write file in a new file
  - Input of plink : write a bed file contain as column `chromosome pos pos rsid` without header. more information about bed file here
- Build a square-LD file using *plink*, in command line :
  - *–extract range file_bed* , where file_bed is the file_bed created in previous steps (obtain common positions with summary statistics)
  - *–r2 square0 yes-really* as argument will give you matrix in square computed on all positions.
  - format output of *plink* file with replacing tabulation by space (example using *tr* of bash)
- Build a config file for *finemap*.
- Run *finemap* to find a credible set of 0.95 (*–prob-cred-set*) and number causal SNPs maximum of 5 (*–n-causal-snps*)

Example of config file of *finemap*:

```
z;ld;snp;config;cred;log;n_samples
dataset1.z;dataset1.ld;dataset1.snp;dataset1.config;dataset1.cred;dataset1.log;5363
```

Using output file of *finemap* software, answer the following:

- how many causal SNPs are found by finemap?
- how many SNPs are in credible sets?
- Optional : which SNPs have been previous described?