Day 3 PM: Practical

Approximate conditional analyses using GCTA and fine-mapping

In this exercise you will be using the GCTA software to perform assess the evidence for multiple distinct association signals for kidney function at the *UMOD* locus (chromosome 16, at ~20Mb) through approximate conditional analysis. More information about the GCTA software can be found at: http://cnsgenomics.com/software/gcta/#Overview.

The data provided consists of genome-wide meta-analysis summary statistics for estimated glomerular filtration rate (eGFR) in more than 100,000 individuals of European ancestry, and a reference data set of genotype data, obtained from European ancestry individuals from the 1000 Genomes Project, to approximate the structure of LD in the *UMOD* locus.

There are two sets of input files required for this exercise:

1. Genome-wide association summary statistics from the eGFR meta-analysis: `METAL.txt`. This file includes one row per SNP, with several columns required for GCTA: the SNP ID (MarkerName), effect allele and other allele (Allele1 and Allele2), effect allele frequency (Freq1), effect size (Effect) and corresponding standard error (StdErr), p-value (P-value) and sample size (TotalSampleSize).
2. Binary PED files of reference genotype data: `EUR.UMOD.bed, EUR.UMOD.bim, and EUR.UMOD.fam`

The data for this practical are stored in the directory DAY3-PM, so you should begin by opening a terminal window and moving to that location.

1. Summarise association summary statistics across the UMOD locus.

You can use LocusZoom, a web-based interface, to generate "signal plots" that summarise association summary statistics across a region of interest. Direct Google Chrome (or another internet browser) to the LocusZoom website: http://locuszoom.org/. You should click on the "Single Plot" button, which will open a webpage to enter the relevant data to generate a signal plot.

For "Path to Your File", click on "Browse" and select the `METAL.txt` file in the DAY3-PM directory. Type "P-value" and "MarkerName" in the p-value and marker name boxes. For Column Delimiter, select "Whitespace".

Ignore the next two columns, and in "Region", select chromosome 16, and for starting and ending positions, type 19.892332 and 20.892332: these define 500kb up- and downstream of the lead SNP at the locus.

The remaining options will impact on the display, for example the reference for LD between SNPs across the region (default European from 1000 Genomes) but for this practical, you can move to the bottom of the screen and click “Plot Data”. It will appear that nothing is happening, but be patient! It might take several minutes to generate the plot.

When the plot is generated, make sure you understand the data being presented. The lead SNP at the locus is indicated by the purple diamond, with all other SNPs coloured according to the extent of LD with the lead SNP. Do you think there could be evidence for multiple association signals at this locus?

2. Perform approximate conditional analysis in the UMOD locus

We can use GCTA to perform approximate conditional analysis at the UMOD locus to assess the evidence for multiple distinct association signals. When performing approximate conditional analysis, it is essential that you provide genome-wide association summary statistics, since these are used to estimate the heritability of the trait under investigation. The specific region in which the conditional analysis is to be performed can be defined from the reference data set: in this case, the binary PED files EUR.UMOD include genotype data only in the interval from 19.892332Mb to 20.892332Mb of chromosome 16. The approximate conditional analysis can only be performed on SNPs that are present in both the association summary statistic file and the reference data set.

To identify index SNPs representing distinct association signals, we can use the option `--cojo-slct`. This option implements backward selection, selecting the “best” set of SNPs explaining the association at the locus at a specified significance threshold, provided by the option `--cojo-p`. The full command to perform the approximate conditional analysis is, which should be entered on one continuous line:

```
gcta64 --bfile EUR.UMOD --cojo-file METAL.txt --cojo-slct --cojo-p 0.00001
--out UMOD
```

The command will produce several output files with the root “UMOD”. The index SNPs for the distinct association signals are given in the file UMOD.jma.cojo. You should see that there are two index SNVs at the UMOD locus. The summary file provides association summary statistics from the unconditional analysis (b, se and p) and the conditional analysis (bJ, bJ_se and pJ), and a summary of the LD between the variants (LD_r) as measured by the correlation coefficient *r*.

Whilst this analysis provides the index SNPs for each distinct association signal, it doesn’t provide the association summary statistics for each signal. To obtain the association summary statistics for the distinct signal indexed by chr16:20392332, we must condition out the effects of all other index SNPs at the locus, in this case just chr16:20353815. To do this, we must create a file, called `snplist.txt` that lists the SNP that we want to condition out, i.e. chr16:20353815. We can do this with the command:

```
echo chr16:20353815 > snplist.txt
```

We can then generate association summary statistics after conditioning out the effect of this SNP using the command (type on one continuous line):

```
gcta64 --bfile EUR.UMOD --cojo-file METAL.txt --cojo-cond snplist.txt --out
UMOD.chr16:20392332
```

Association summary statistics for the distinct signal indexed by chr16:20392332 are given in the file `UMOD.chr16:20392332.cma.cojo`. The file provides summary statistics from the unconditional analysis (b, se and p) and after conditioning (bC, bC_se and pC).

You can produce a signal plot for the association signal using LocusZoom. For "Path to Your File", click on "Choose file" and select the `UMOD.chr16:20392332.cma.cojo` file in the DAY3-PM directory. Type "pC" and "SNP" in the p-value and marker name boxes. For Column Delimiter, select "Whitespace". As before, ignore the next two columns, and in "Region", select chromosome 16, and for starting and ending positions, type 19.892332 and 20.892332, which will focus on the same region as in the unconditional analysis.

In the same way, you could obtain association summary statistics for the signal indexed by chr16:20353815 by conditioning out the effects of chr16:20392332. Note that if a region contains three index SNPs, say rs1, rs2, and rs3, the process to obtain association summary statistics for each distinct signal is a little more complex. For example, to obtain association summary statistics for the signal indexed by rs1, you must condition out the effects of both rs2 and rs3, so both SNPs would be included in the file specified by the `--cojo-cond` option. Similarly, to obtain association summary statistics for the signal indexed by rs2, you must condition on rs1 and rs3. With more index SNPs, the obtain association summary statistics for the signal indexed by one SNP, you must condition out the effect of all other index SNPs at the locus.

3. Fine-map the chr16:20392332 association signal

We will next fine-map the association signal indexed by chr16:20392332 using summary statistics from the approximate conditional analysis and Wakefield's approach. First, we will calculate the Bayes' factor in favour of association for each SNP in the signal using a pre-prepared analysis script, bfcal, which you can use by typing the command:

```
./bfcal
```

You will be prompted to enter the name of the GCTA cma.cojo file in which the conditional association summary statistics are reported, in this case `UMOD.chr16:20392332.cma.cojo`. The program will produce an output file named `bfcal.txt` that contains the following columns: SNP ID, chromosome, position, log10 Bayes' factor, and the posterior probability of association.

We can sort this file by posterior probability, from largest to smallest, using the command:

```
sort -g -r -k5 bfcal.txt > bfcal.sort.txt
```

We can then view the first few lines of the file using the command:

```
head bfcal.sort.txt
```

To form a 95% (or 99%) credible set, we include SNPs until the total posterior probability exceeds 0.95 (or 0.99). How many SNPs would be in the 95% credible set?