Intro to bioinformatics

Gibran Hemani

LD Clumping

LD clumping is a commonly used technique. It provides a reasonable solution to the problem that often you only want the best SNP from a GWAS signal, not all the SNPs in the region.

LD clumping is performed as follows.

1. Identify the best SNP for that exceeds a p-value threshold (`--clump-p1`)
2. Find all the SNPs within some radius (`--clump-kb`) of the best SNP that exceed another p-value threshold (`--clump-p2`)
3. Calculate the LD r-square between the best SNP and each of the remaining SNPs in the region from step 2.
4. All the SNPs that are above some r-square threshold (`--clump-r2`) form a clump for the best SNP, and are discarded from being potential future best SNP candidates
5. Retain the best SNP from this round and its clump, then find the next remaining best SNP and repeat from step 2.

The basic idea is that if a SNP is in LD with the top SNP in the region then we assume that it only appears to have an effect because it is associated with the top SNP.

We can use plink to perform LD clumping. Let's navigate to the folder that contains our GWAS results

```
cd results/gwas
```

Performing clumping on the BMI GWAS

```
../../software/plink_mac \
--bfile ../../data/genetic/geno_qc \
--clump precomputed/bmi_adjusted.assoc.linear.add.gz \
--clump-p1 5e-8 \
--clump-p2 5e-8 \
--clump-kb 1000 \
--clump-r2 0.01 \
--out bmi
```

The flags being used here are

- `--clump precomputed/bmi_adjusted.assoc.linear.add.gz` specifies the file with the results. All that is required is a column with SNPs (labelled SNP in the header) and a column with P values (labelled P in the header)
- `--clump-p1 5e-8` specifies to only allow best SNPs with this minimum p-value
- `--clump-p2 5e-8` specifies that SNPs to include in the clump list are only retained if they surpass this p-value
- `--clump-kb 1000` specifies the radius around which to look for clumps (1Mb in this case)
- `--clump-r2 0.01` specifies to include SNPs with r-squared greater than 0.01 in the clump for the best SNP.

This creates a file called `bmi.clumped`. It has one line for every clump, with the best SNP specified in the `SNP` column. The SNPs that were within the radius, had the p2 p-value, and were in LD above the threshold, are listed in the `SP2` column.

You can do the same thing with the other GWAS results too.

Looking up a SNP

Let's first look at the UCSC genome browser. Choose one of the retained SNPs and use it as a search term here

http://genome.ucsc.edu/cgi-bin/hgGateway

Zooming out may help to see more information. The UCSC genome browser has a huge amount of information that is organised in 'tracks'. A track is a layer of information that is mapped to a genome build. For example, the common SNPs track shows SNP positions within the window that is visible. There is a 'RefSeq Genes' track which shows the orientation, structure and position of genes.

Note that the SNP you chose likely is listed in the 'NHGRI-EBI Catalog of Published Genome-Wide Association Studies' track. If you click on a SNP there it will tell you information about the publication.

Getting more information

A useful resource for finding out more information about a particular SNP is the Haploreg website.

http://www.broadinstitute.org/mammals/haploreg/haploreg.php

Enter your SNP into this search box. What this database does is it provides information about the SNP that you searched for, for example whether it lies in DNA motifs, which tissues it might be relevant to, etc; but it also shows the same information for SNPs that are in LD with the SNP you specified. This is valuable because it could be that your SNP is not the causal variant, and it is simply in LD with the true causal variant which will have some relevant annotations.