**Summary-based methods practical: colocalization analysis**

In this practical we are going to use colocalization analysis to appraise whether 1) protein level of CSF1 share the same causal variant with years of schooling within the LDLR region; and 2) protein level of MUC16 share the same causal variant with asthma within the MUC16 region.

**Files you will need:**

1. R script: coloc-practice.R
2. Input file for GWAS summary data of LDLR expression and LDL-c: LDLR-LDL-c.txt
3. Input file for GWAS summary data of protein level of MUC16 and asthma: MUC16-Asthma.txt

**R packages you will need:**

1. R version > 3.3.0
2. Coloc which can be installed within R using the following command

install.packages("coloc")

**Key things before start running coloc:**

1. Please make sure you select **all SNPs** within **a genomic region** (no matter whether they are significant or not). We used +- 500Kb window from the top association signal (top pQTL) here.
2. Please check the **number of SNPs** you used for the coloc analysis, we recommend at least more than 50 test SNPs.
3. Since colocalization have 5 hypotheses, and our major research interests is to distinguish hypothesis 4 (H4) from hypothesis 3 (H3). To save time, we should only consider running coloc analysis when there are association signals (e.g. the lead SNP has a **P-value < 1x10-5**

for both the protein and disease trait of interest). Otherwise, the coloc analysis will favour the H0, H1 or H2 hypotheses, which is not of interest.

1. We have created two R functions to help you running coloc analysis more easily:
   1. The function “coloc.analysis” is for coloc analysis between a quantitative trait (e.g. protein level) and a binary trait (e.g. asthma).
   2. The function “coloc.analysis.quant” is for coloc analysis between two quantitative traits
   3. Since coloc is a Bayesian method, so we need to set up prior probabilities:
      1. p1 prior probability a SNP is associated with trait 1 (expected proportion of SNPs associated with protein expression), default 1x10-4
      2. p2 prior probability a SNP is associated with trait 2 (expected proportion of SNPs associated with disease trait), default 1x10-4
      3. p12 prior probability a SNP is associated with both traits (expected proportion of SNPs associated with both traits), default 1x10-5

**ANALYSIS 1: colocalization analysis of protein level of CSF1on years of schooling**

Colocalization analysis can be applied to all types of GWAS data. In the first analysis, we will apply the method to two quantitative traits, protein level of CSF1 and years of schooling.

**A. Test whether protein level of CSF1 and years of schooling share the same variant within the CSF1 region.**

1. Call the R function “coloc.analysis.quant”
   1. What kind of parameters we need to run coloc between two quantitative traits?
2. Estimate the probability of all 5 hypotheses

**B. Interpret the results**

1. Do the number of SNPs included in this analysis more than 50? Check “nsnps”
2. Does the result suggest colocalization between CSF1 and years of schooling (**PP.H4>80%**)?

Note: notice that the sum of all 5 probabilities will be 1.

**ANALYSIS 2: colocalization analysis of protein level of MUC16 and asthma**

Colocalization analysis can also be applied to binary trait. In the first analysis, we will apply the method to two one quantitative trait and one binary trait, protein level of CSF1 and asthma.

**A. Test whether protein level of MUC16 and asthma share the same variant within the CSF1 region.**

1. Call the R function “coloc.analysis”
   1. What kind of parameters we need to run coloc between a quantitative trait and a binary trait? Is there any difference between “coloc.analysis” and “coloc.analysis. quant”.
2. Estimate the probability of all 5 hypotheses

**B. Interpret the results**

1. Do the number of SNPs included in this analysis more than 50? Check “nsnps”
2. Do the results suggest colocalization between CSF1 and years of schooling (**PP.H4>80%**)? What is this result imply?

**Core assumptions and practical limitations of colocalization analysis**

1) There is only one causal variant within the genomic region we tested.

2) More than 50 SNPs have been included in the analysis

3) All SNPs within the genomic region have been included in the analysis.