GGSB PRELIM QUESTION # 5

Problem. GWAS have successfully identified a large number of variants associated with a range of complex diseases. In most cases, however, the causal variants and their target genes remain unknown. A common strategy is to leverage expression QTL (eQTL) data to help narrow down causal variants and genes.

a) Suppose a GWAS variant associated with the disease trait of interest is also associated with the expression of a gene X in some tissue. Can we conclude that X is a causal gene of the disease?

b) A strategy to better integrate eQTL and GWAS data is colocalization analysis. Read the paper presenting the method coloc[1]. Given association data of two traits in one region, coloc tests if there is a common causal variant in the region for both traits. State the five hypothesis coloc is evaluating. Suppose we are given the prior probability a SNP is associated with trait 1, p_1, with trait 2, p_2, and with both traits p_{12}. Let S be a configuration: it is a pair of binary vector of whether a SNP is causal to trait 1 and trait 2 (see Figure 1 for example). What is the prior probability of S when S belongs to each of the five hypothesis?

c) The evidence of SNP association of a trait can be expressed as Bayes factor (BF), which compares two models, the SNP is associated with the trait vs. not associated. Suppose we know BF of each SNP in the region with respect to the two traits. Derive the posterior probability of colocalization in terms of the BFs of SNPs.

d) It was found that results of coloc are sensitive to prior parameters. A program Enloc is designed to address this problem[2]. Explain its main ideas.

e) A naive approach to colocalization is: perform fine-mapping of both traits separately, and this allows one to assess how often a SNP is a causal variant to both traits. It would be the product of Posterior Inclusion Probabilities (PIPs) of the SNP with respect to the two traits. Explain the shortcoming of this approach, in light of Enloc.

f) The key parameter of Enloc is called α_1. Define the parameter, and explain why it is important in your own language.

g) Explain how Enloc is related to coloc, and why it may provide some advantages.

References

1. Giambartolemi, C. *et al*. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014)
2. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, 1-25 (2017)