Notes-Lecture-12

Below we list some facts that you are expected to know from Lecture 12.

Unless otherwise specified, we assume an additive genetic model for the phenotypes.

$$Y = \sum_{l}^{M} \beta_l X_l + \epsilon$$

where l indexes the SNP and M is the number of SNPs.

A useful trick is to partition the sum into different classes.

$$Y = \sum_{l \in \text{class1}}^{M} \beta_l X_l + \sum_{l \in \text{class2}}^{M} \beta_l X_l + \epsilon$$

where class1 and class2 are two different classes of SNPs. Class1 could be for example all coding variants and class2 non coding variants.

Prediction

To predict a complex trait, all we need to know are the β_l's and the person's genotype data. We saw in class several ways to estimate them.

- GWAS output (fitting the phenotype on one SNP at a time)
- Fitting all SNPs simulataneously
  - Ridge Regression (least squares minimizing the sum of β^2)
  - LASSO (least squares minimizing the sum of $|\beta|$)
  - Elastic Net (least squares minimizing both the sum of β^2 and the sum of $|\beta|$)
  - BLUP (Best linear unbiased preditor can be calculated with GCTA and is equivalent to Ridge Regression)
  - BSLMM (this has two one sparse component with variants of large effect sizes and a polygenic background)

Performance of prediction

We use the correlation or the square of the correlation between observed and predicted values (pred_perf_R2) as a measure of performance of the predictor.

Correlation is useful to compare to heritability but in general, it may not be a good measure of performance. A predictor with good correlation may be very different to the observed value. For example, if we take the predictor and multiply by a constant factor and or add/subtract a constant number, the correlation does not change but the difference between predictor and observed can change dramatically.

Pred_perf_R2 of the predictor cannot exceed the heritability of the trait. In other words, the narrow sense heritability of the trait determines an upper bound to how well we can predict the trait.

We mentiond that the best performing model's expected pred_perf_R2 cannot be larger than the heritability of the trait (although since the actual pred_perf_R2 is noisy and the h2 is estimated with noise, we may see R2 greater than heritability estimates).

Prediction of gene expression traits

For gene expression traits, we can define two classes of SNPs, one in the vicinity of the gene's transcription start site (TSS), say within 1Mb of, and another class with variants farther from the gene. We call those local and distal regions.

We partitioned the heritability of genes into local and distal components and found that at sample sizes under 1000, we can get reasonable estimates of local heritability but distal heritability is not reliable (confidence intervals are very large).

PrediXcan, SMR, and TWAS

PrediXcan is a method that uses predicted expression levels to find genes that are associated with complex traits. TWAS is a method based on the same idea but was implemented using BSLMM prediction models whereas PrediXcan uses elastic net models.

LD Contamination

We didn't go through this in class.