Permutations for model selection of genetic regulatory pathways

To study phenotypes of economic importance, it is important to understand the genetic architecture underlying its regulation. These genetic pathways can be formalized as a combination of quantitative trait loci (QTL) that have additive and epistatic e↵ects. Given all genetic variants in a population of diverse individuals (e.g., approx. 28 million variants in the BioEnergy Science Center's Populus association mapping population (n=882)), the possible combinations of multiple-QTL models to assess is practically intractable. The exploration of multiple-QTL models falls into the larger field of model selection, where it's critical to determine how to traverse models. Permutation testing is a common method to provide test statistics under a null hypothesis of no QTL, and a means to penalize multiple QTL models.

Challenge

Utilizing the DOE's BioEnergy Science Center's (BESC's) Poplar Genome-Wide Association Study (GWAS) Genotype Dataset, publicly available at their Download page(see Table 1, and a simulated phenotype available at the Data Challenge's web page (see Table 2).

Description	Name	Size
Poplar variants (in VCF format)	SNP dataset	78G

Table 1: Genotype Data – variants Populus individuals

Description	Name	Size
Simulated phenotypes coded in two columns containing individual ID and corresponding quantitative phenotype	BESC_phenotype_DataChallenge.tsv	25k

Table 2: Phenotype Data – simulated quantitative trait for genotyped Populus individuals

The challenge is to empirically determine or estimate the distribution of test statistics under the null hypothesis of no genetic regulation from permutation tests of genotype and phenotype data of the Populus association mapping population on reasonable timescales; this provides accurate significance thresholds that enable biological interpretation. To this end, the following aims should be addressed.

1. Determine significance thresholds for genetic e↵ects from permutations of an n-dimensional genome scan.
    a. additive e↵ects of a 1-dimensional scan
    b. additive and epistatic e↵ects of a n-dimensional scan, for n >= 2

2. Apply significance thresholds to multiple-QTL model selection.
3. Can the number of permutations be reduced by identifying patterns in permutation result? (e.g., adaptive permutations? fixed density patterns?)