Human genetics in the 21st century:
Using bioinformatics to link genotype and phenotype

High School Student Program 2015
Bioinformatics and Research Computing, Whitehead Institute, http://barc.wi.mit.edu

Introduction

Each row includes phenotypes and genotype about one individual (plant or animal or person, for example). Columns B-G show the phenotypes of six traits. Phenotypes are generally described as H and L (so can be considered as High or Low) and are color coded. Columns H-AA show rows of genotypes at different loci. Each genotype is represented by 20 different tag SNPs. For simplification we assume that there are only 2 common alleles for each SNP. Each SNP is represented as two letters. One refers to a location on the chromosome coming from the mother and one refers to the location on the chromosome coming from the father. The order of the alleles is not significant, so a heterozygous SNP is always represented as "AB". The SNP notation includes "A", the allele with a sequence that matches the reference genome, or "B", the allele with a sequence that differs from the reference genome. The SNPs have been color coded too. **Your goal is to find the locus (SNP; region of the genome) that is associated with each trait.** In more detail, when the trait is H, is the SNP always AA or is it always BB? How about when the trait is L? In this simulated data, alleles AB may encode the same trait as alleles AA or as alleles BB. The color will help you find the relationships.

Exercise step 1: Sheet "Demo_visual"

One trait at a time, sort the table by the trait column and look across the SNP columns. Which SNP column follows a similar pattern to the trait pattern? Record this genotype-phenotype connection. You will find it because the colors on the SNP column will be ordered too. If you don't see it right away try sorting the trait first A-Z and then Z-A.

Exercise step 2: Sheet "Demo_numerical"

This sheet contains the same information as "Demo_visual" but
- Traits L and H and have been replaced by -1 and 1, respectively.
- Alleles AA, AB, and BB have been replaced by 1, 0, and -1, respectively.

We can use these numbers to calculate correlation between each trait column and each SNP column. The color-coded results of this correlation analysis are in a table at the bottom of the sheet. For each trait, what is the SNP with the highest correlation (most yellow background)? Clicking on that cell of the correlation table will show the formula used (at the top of the page). Note that we're showing the absolute value of each correlation. Do these results agree with the results of the "Demo_visual" results?