Population Genetics Project #6

Friedline et al. [**?**] investigated the role of natural selection in the rapid spread of gypsy moth (*Lymantria dispar*) across Eastern North America. Their analysis focused on analysis of polymorphisms at 91,468 SNPs identified through double digest restriction-site associated DNA sequencing (ddRADseq). As part of their work, they also collected data on pupal mass (Mass), pupal development time (PD), and total development time (TDT). This allows us to perform a genome-wide association analysis that assesses the relationship, if any, between individual SNP loci and each of the phenotypes. You'll find two data sets on the course web site:

- `gypsymoth.csv`: A CSV file containing information about each individual in the sample (population of origin, sample label, phenotype, genotype at each SNP locus).
- `gypsymoth-relatedness.csv`: A CSV file containing the relatedness matrix for individuals included in the data set.

The data set is a subset of the data included in the paper. Specifically, I filtered the data to include only (a) loci that were scored in more than 100 individuals, (b) individuals that had more than 4000 of the remaing SNP loci scored, and (c) loci that were scored in all of the remaining individuals. The resulting data set includes 141 individuals scored at 218 loci. Using these data, answer the following questions:

1. What loci are associated with each of the three traits (pupal mass, pupal development time, and total development time), and what is the heritability associated with genetic differences at those loci?
2. Is there evidence that any of the SNP loci are associated with variation in more than one of the traits?
3. Given your answer to #2, would you expect to see a change in pupal development time, total development time, or both if natural selection led to a change in pupal mass? Why or why not?

4. When you run the `R` script (see "Hints" below), the table will list the estimated allelic effect and heritability for the 20 loci with the largest magnitude of allelic effect. If you sum the heritability associated with each locus across all 20 loci, you'll probably find that the sum is greater than 1. I've been telling you that quantitative geneticists treat variation at multiple loci as the sum of effects across loci. How do you reconcile a heritability bigger than 1, which is impossible, with summing effects across loci?

Hints

- Use the script `gwas.R` to run your GWAS. You may need to install `rstan`.[1] For each analysis, you'll only need to change one thing, the phenotype that you're analyzing. You'll see the names of the phenotypes in columns 3-5 of `dat` after you've run the script or in columns 3-5 of the CSV file (if you open it as a spreadsheet).

- To resolve the paradox mentioned in the final question, you'll need to think about the biological implication of adding effects across loci and what it means if adding effects doesn't work. This is stretching you a bit, because a complete answer would involve a genetic concept that is well established, one that you have probably encountered before if you have already taken a genetics course, but not a concept that we've discussed in class.

[1] You also might not need to install it. You may have installed it as part of `rstanarm`. It's been so long since I installed `rstan` that I don't remember.