Project Progress Report: W3500 Independent Biological Research

Kyelee Ruth Fitts

Background: The (chi-squared distributed) test statistic Q_x is used in population genetics to determine whether a polygenic trait shows signals of selection across different populations, given the effect sizes (determined by genome-wide association studies, or GWAS) and frequencies of each locus that contributes to the trait. We expect populations that have drifted apart to have no selection under the null model, but if Q_x is significant, then we know that the differences in genetic value (weighted average of change in allele frequencies from the ancestral frequency multiplied by the effect size) among the different populations is not due to random processes like drift.

On the second page is the expression for the statistic in terms of the average effect size α, the ancestral allele frequency $\overline{p}$, and the current allele frequency p of our target population, with each parameter taken over loci l and l', as determined by Berg and Coop et al. [1]

The average effect size at each locus is of particular importance– not only does GWAS use it to determine which alleles are deemed significant to a trait, but it is also the weighting factor for loci in the expression for Q_x. The average effect size depends on the allele frequency, the homozygous effect size (effect size of the most frequent homozygote) and the dominance deviation (difference in effect size for the heterozygote deviating from $\frac{1}{2}$ of the homozygous effect size). On the next page is the expression for α, represented in terms of the dominance deviation D_l at each locus and the homozygous effect A_l at each locus (1).

Hypothesis: The use of α in this statistic becomes problematic in the presence of directional dominance, which is when alleles for a particular trait with a positive effect size are systematically dominant or when alleles with a negative effect size are systematically recessive. Quantitatively, this means the signs of the dominance deviation and homozygous effect sizes are the same for a large number of loci. In the presence of directional dominance, we hypothesize that there is a bias in the Q_x statistic: for alleles displaying directional dominance, systematically dominant alleles in our target population which have increased in frequency relative to other populations will tend to display smaller effect sizes, while systematically recessive alleles which have recently decreased in frequency relative to other populations will tend to display larger effect sizes. In this latter case, false positives for polygenic selection will occur, because GWAS will systematically choose recessive alleles that have recently decreased in frequency because of their larger effect sizes. This is problematic because our test statistic will then use loci that are skewed away from the ancestral genetic value, biasing the test.

Progress: Substituting (1) into (2), we can algebraically derive the an expansion of the statistic, as noted on the second page (3). Using this expansion, I have been creating simulations to verify graphically the effect that dominance has on the test statistic– I have attached some of the most important results from these simulations on the next page.

Figures a and b compare the expected cumulative distribution function of the Q_x statistic (in red) with the cdf of the simulated distribution. When the dominance deviation is zero, as expected the two are nearly identical. However, as dominance deviation increases, the distribution of the Q_x statistic shifts to the right. This is further shown in figure c, which plots the fraction of the simulated statistics above the value of the statistic expected for 5% of the expected distribution, which is chi-squared with 1 df. This plot also shows that the proportion of simulated Q_x values over the expected threshold increases as dominance deviation increases.

Figures d, e and f show how dominance affects the genetic value of loci over time. These figures are especially interesting because they show how specifically directional dominance affects genetic value. When the dominance deviation is equal to 0 in figure d, as expected the genetic values of the replicates drift randomly over time and the mean genetic value (red) is approximately zero. However, for positive dominance deviations (e), the average genetic value over time becomes negative, while for negative dominance deviations, it becomes positive. This indicates that for positive dominance the effect of α makes the genetic value for alleles which are increasing in frequency smaller, and decreasing in frequency larger, while for negative dominance deviations the effect is the opposite. This indicates, crucially, that for directionally recessive traits (like height) which have recently increased in frequency we expect the effect of multiplying α to increase the perceived genetic value of the allele, making it more likely to be chosen by GWAS as signifiant and skewing our test statistic Q_x.

Next steps: So far the work I've been doing this semester has been mainly in ascertaining the theoretical basis for this project and assembling these simulations to directly quantify the statistical bias in these tests due to dominance. I am currently working on a simulation that will break down the terms of the expansion to see what effect each term has separately on the distribution of the statistic. In addition, I have just gained access to the UK BioBank data in order to search for the bias within real human genetic data.

References

[1] Jeremy J. Berg. Polygenic Adaptation has Impacted Multiple Anthropometric Traits. *bioRxiv*, 2017.

Equations:

$$\alpha_\ell = \frac{1}{2}A_\ell + D_\ell\left(1 - 2p_{1\ell}\right). \tag{1}$$

$$Q_X = \frac{1}{V_A F_{ST}} \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{\ell\prime=1}^{L} \alpha_\ell \alpha_{\ell'} \left(p_{m\ell} - \overline{p}_\ell\right)\left(p_{m\ell\prime} - \overline{p}_{\ell\prime}\right) \tag{2}$$

$$\begin{aligned} &\sum_{l=1}^{L}(\frac{1}{2}A_l(p_{1l} - \epsilon_l))^2 + \sum_{l=1}^{L}\sum_{l\neq l'}^{L}(\frac{1}{4}A_l(p_{1l} - \epsilon_l)A_{l'}(p_{1l'} - \epsilon_{l'})) \\ &+\sum_{l=1}^{L} A_l D_l(1 - 2p_{1l})(p_{1l} - \epsilon_l)^2 + \sum_{l=1}^{L}\sum_{l\neq l'}^{L}(\frac{1}{2}A_l D_{l'}(1 - 2_{p1l'})(p_{1l} - \epsilon_l)(p_{1l'} - \epsilon_{l'}) \\ &+\frac{1}{2}D_l A_{l'}(1 - 2_{p1l})(p_{1l} - \epsilon_l)(p_{1l'} - \epsilon_{l'})) \\ &+\sum_{l=1}^{L}(D_l(1 - 2p_l)(p_{1l} - \epsilon_1))^2 + \sum_{l=1}^{L}\sum_{l\neq l'}^{L} D_l(1 - 2p_l)(p_{1l} - \epsilon_1)D_{l'}(1 - 2p_{l'})(p_{1l'} - \epsilon_{l'}) \end{aligned} \tag{3}$$

Table 1: From top to bottom, left to right: figures a, b, c, d, e, f.