

# Project Progress Report: W3500 Independent Biological Research

Kyelee Ruth Fitts

The (chi-squared distributed) test statistic  $Q_x$  is used in population genetics to determine whether alleles with significant effect sizes (usually determined by GWAS) contribute to selection on a polygenic trait. Specifically, given a certain number of populations, it measures whether the effect sizes of each locus within each population show a trend of selection for or against the polygenic trait by measuring whether the covariance of genetic value (allele frequency minus mean ancestral frequency) is nonzero over all loci with significant effect sizes.

$Q_x$  depends on the value of the average effect size at each locus, which in turn 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).

Directional dominance is when heterozygotes display greater effect sizes than expected— that is, the signs of the dominance deviation and homozygous effect sizes are the same. In the presence of directional dominance, we hypothesize that there is a bias in the  $Q_x$  statistic: for alleles displaying directional dominance, alleles which have recently increased in frequency (dominant alleles) will tend to display smaller effect sizes, and alleles which have recently decreased in effect sizes (recessive alleles) will tend to display larger effect sizes. In this latter case, false positives for polygenic selection will occur. Below is a theoretical justification for this hypothesis.

$Q_x$  can be expressed in terms of the effect size  $\alpha$  in this relationship [1]:

$$Q_x = \frac{1}{V_{AFST}} \sum_{m=1}^M \sum_{\ell=1}^L \sum_{\ell'=1}^L \alpha_{\ell} \alpha_{\ell'} (p_{m\ell} - \bar{p}_{\ell}) (p_{m\ell'} - \bar{p}_{\ell'}) \quad (1)$$

Where  $\alpha$  can be represented in terms of the dominance deviation  $D_l$  at each locus  $l$  and the homozygous effect  $A_l$  at each locus:

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

Substituting (2) into (1), we can algebraically derive the following expansion:

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

Where we can expect much of the bias in the  $Q_x$  statistic coming from dominance to come from the last two terms involving  $D_l$ . Since the semester has begun, I've been creating simulations to verify the effect that dominance has on the test statistic. On the second page, Figure 1 shows the distribution of the  $Q_x$  statistic without dominance as simulated by the expansion above, which matches a chi-squared distribution with degree of freedom equal to 1, as expected.

Figures 2, 3, and 4 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 further and further to the right. This is further shown in figure 5, 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. That is, the plot shows the proportion of  $Q_x$  values above the  $p = 0.05$  cutoff value of the expected chi-squared distribution for different dominance deviation values. As expected, this plot also shows that the proportion of simulated  $Q_x$  values over the expected threshold increases as the effect of dominance increases.

Finally, figures 6, 7, and 8 show how dominance affects the genetic value (that is, the change in allele frequency multiplied by the effect size  $\alpha$ ) over time. These figures are especially interesting because they show how specifically directional dominance can affect genetic value. When the dominance deviation is equal to 0 in figure 6, as expected the genetic values of the replicates drift randomly over time and the mean line (red) is approximately zero. However, for positive dominance deviations (fig. 7), the genetic value over time becomes negative, while for negative dominance deviations, the genetic value over time becomes positive (compare with the blue line at  $y=0$ ). This indicates, crucially, that the bias due to dominance will increase the perceived effect

size and result in false positives if alleles tend to be recessive (as with height) and decrease the perceived effect size for alleles displaying dominance, as hypothesized.

So far the work I've been doing this semester has been mainly in 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 human genetic data.

## References

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

Figure 1: Distribution of Qx Expansion (Chi-squared, df=1)

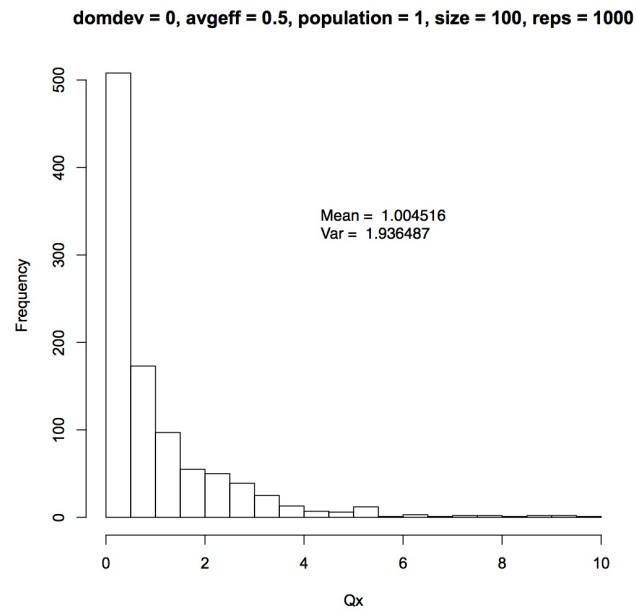


Figure 2: CDF of Qx Statistic vs expected cdf where dominance deviation = 0

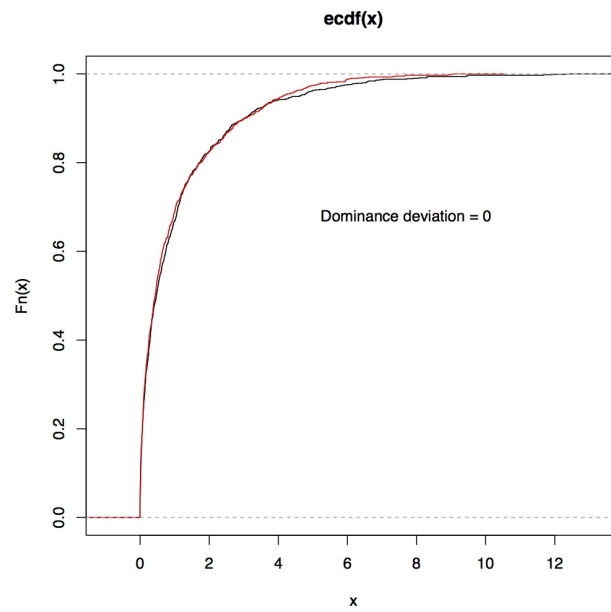


Figure 3: CDF of Qx Statistic vs expected cdf where dominance deviation = 0.5

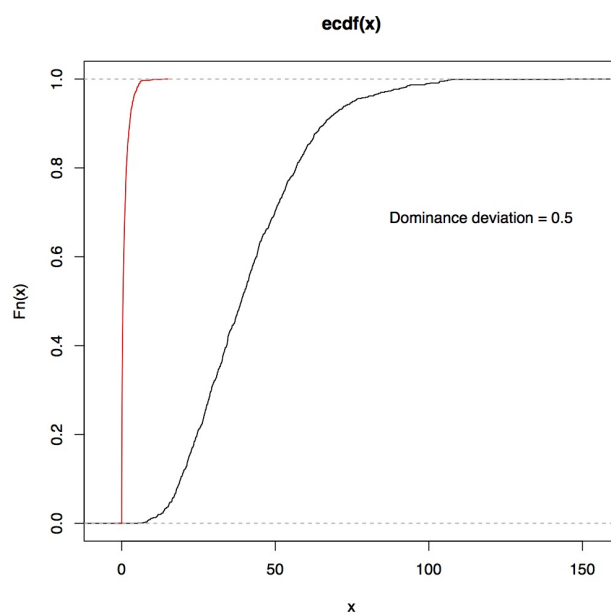


Figure 4: CDF of Qx Statistic vs expected cdf where dominance deviation = 1

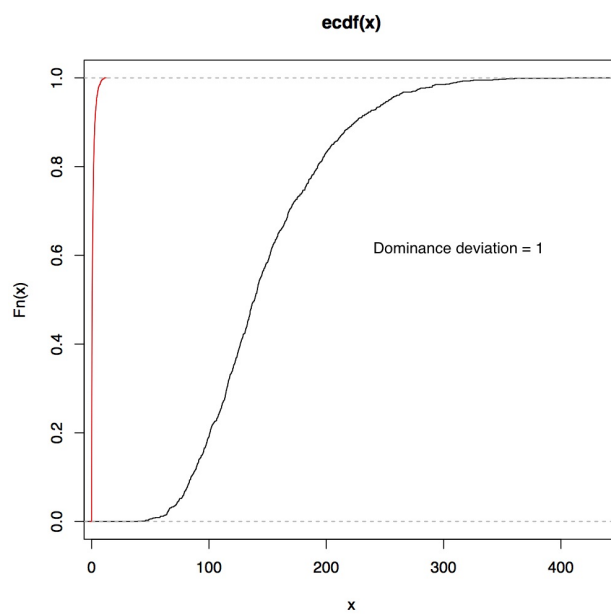


Figure 5: Genetic value of locus vs time, dominance deviation = 0

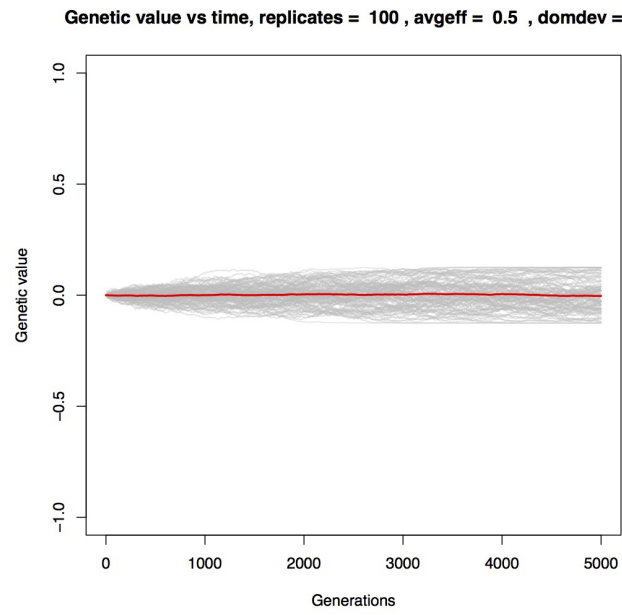


Figure 6: Genetic value of locus vs time, dominance deviation = 0.2

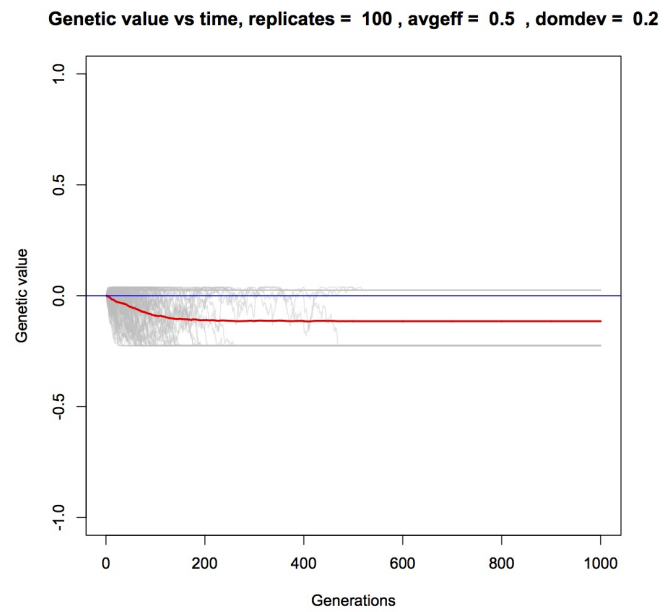


Figure 7: Genetic value of locus vs time, dominance deviation = -0.2

