Title here

Kyelee Ruth Fitts, Jeremy Berg, Guy Amster, Guy Sella

November 19, 2017

Abstract

Introduction

It has long been known that diversity on Earth is in large part due to the adaptation of organisms to varying environments. We now know that adaptation has a large genetic basis, and that genetic variation is the key to biological diversity. Thus, understanding how and why such genetic variation occurs is the ultimate goal of evolutionary biology. Quantifying genetic adaptation is the goal of population geneticists, and to this end finding methods to detect signals of selection within the genome is of particular interest. Change in variation over time can be attributed to many causes including genetic drift or population migration, so detecting signals of selection in particular is no trivial task.

Wright first introduced the parameter F_{st} as a measure comparing the variation of an allele in a subpopulation to that of the entire population [5]. Lewontin and Krakauer, using this parameter developed a novel statistical test based on the fact that under no selection, the variation of specific alleles over all subpopulations will not be different from that of the population F_{st} [7]. These conclusions were extended by Spitze, who coined the parameter Q_{st} as a measure of how the phenotypic value of an allele in a subpopulation compares to that of the entire population [2]. These early tests for selection utilized the ratio $\frac{Q_{st}}{F_{st}}$ as a test statistic for detecting signals of selection, where $F_{st} = Q_{st}$ is the null model, where no selection occurs.

Until recently, the study of detecting selection has been limited to large-effect alleles, when in fact numerous phenotypic traits of interest are affected by many different loci across the genome. Understanding how to find signals of selection for such polygenic traits has been the subject of much study in recent years. Increased compuational capabilities have given us to tools to be able to evaluate the effects of many thousands of loci on a trait, allowing us to evaluate selection in highly polygenic traits, such as height. One of the most important tools developed are Genome-Wide Association Studies (GWAS) which determine which loci across the genome have significant effects on a certain polygenic trait [6].

Berg introduces a comprehensive test statistic, called Q_x, to test for selection [1]. This statistic depends on F_{st} and a generalized analogy to Q_{st}, expressed in terms of α, or the effect size of each significant locus as determined by GWAS, and the allele frequencies p over loci l and l', summed over populations m. V_a is the additive genetic variance of the entire population.

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

The distribution of this statistic is expected to be chi-squared. We are particularly interested in the average effect of this statistic– 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 (A), or effect size of the most frequent homozygote, and the dominance deviation (D), which is the difference in effect size for the heterozygote deviating from $\frac{1}{2}$ of the homozygous effect size.

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

The use of α in this test statistic becomes problematic in the presence of directional dominance, when alleles with a postive effect size on the trait are systematically dominant and alleles

with a negative effect size on the trait are systematically negative. In terms of (2), this means that the signs of A and D are the same for a large number of loci.

In the presence of directional dominance, we hypothesized a bias in tests for polygenic adaptation that depends on α. Alleles which are systematically dominant and which have recently increased in frequency (large p, positive D) will tend to have smaller effect sizes, while alleles which are recessive which have recently decreased in frequency (small p, negative D) will tend to have larger effect sizes. In the latter case, the test for polygenic selection advanced by Berg (1) will tend to make false positive judgements for selection, because the statistic will be calculated over alleles which do not actually increase the effect size.

Height is one polygenic trait with many well-defined signficant alleles via GWAS [3]. It has also been shown to exhibit directional dominance [4]. We aim to quantify the hypothesized bias, first with simulated populations, then with height genotype data from the UK Biobank.

Theory/Methods

Using the expression for the test statistic Q_x (1), we can substitute the expression for alph (2) and manipulate the expression algebraically to derive the following expansion for the test statistic, in terms of the homozygous effect (A) and the dominance deviation (D):

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

We can loosly consider the single summation terms to be variances corresponding to the expansion of additive effects multiplied by additive effects, additive times dominance, and dominance times dominance, and the double summation terms as covariances of these quantities. We expect the inflation of the test statistic due to dominance effects to come from the last two terms. Note that when dominance is not present ($D = 0$) the expansion reduces to the expression Berg presents when α is treated as a constant [1].

Using this expansion, we have created simulations to characterize our hypothesized dominance bias. We used simulated populations under a couple of assumptions: first, that the values of the dominance deviations and homozygous effects are constant throughout the population. Second, that F_{st} for these simulated approximations can be roughly estimated by the number of generations elapsed over the population size. Third, that the distribution of allele frequencies after one generation can be approximated by a normal distribution centered at the ancestral frequency with variance $F_{st} * \epsilon * (1 - \epsilon)$ where ϵ is the ancestral frequency. (CITE THESE ASSUMPTIONS???)

All simulations were performed in R.

Results

Discussion

References

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

[2] K Spitze et al. Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. *Nature Genetics*, 1993.

[3] Michael C Turchin et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics*, 44:1015–1019, 2012.

[4] Peter K Joshi et al. Directional dominance on stature and cognition in diverse human populations. *Nature*, 523:459–462, 2015.

[5] Sewall Wright et al. The Genetical Structure of Populations. *Annals of Eugenetics*, 1949.

[6] Matthew Brown Peter Visscher et al. Five years of gwas discovery. *American Journal of Human Genetics*, 2012.

[7] Jesse Krakauer R.C. Lewontin et al. Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. *Nature Genetics*, 1973.