Project Proposal: W3500 Independent Biological Research

Kyelee Ruth Fitts

Diversity of life on Earth is in large part due to the adaptation of living organisms to the circumstances of their environments. As adaptation is a fundamentally genetic process, understanding the genetic mechanisms by which populations adapt to their environments is a fundamental goal of evolutionary biology. Until recently, the study of the genetics underlying adaptation has been focused on individual mutations of large effect size, when in fact we know that many important traits of interest are polygenic traits that are affected by many different loci.

In order to study such polygenic traits, genome-wide association studies (GWAS) have become increasingly important in the field of population genetics. Utilizing advanced genotyping technology and very large sample sizes, GWAS provides estimates for the extent of correlation between traits and loci as well as the effect sizes of significant alleles over the entire genome. GWAS in combination with statistically rigorous tests of polygenic adaptation have been used to detect significant and systematic correlations between the changes of allele frequencies at each site and the effect that each allele has on the trait in question– in effect, finding a footprint of natural selection against the noise of genetic drift. These tests for polygenic adaptation using GWAS data have been used to detect signals of selection in many anthropometric traits like height and waist-hip ratio, as well as disease phenotypes like type 2 diabetes[1, 4].

As these tests for polygenic adaptation become more and more ubiquitous, it is necessary to ensure that the underlying assumptions of the statistical tests used remain as free from bias as possible to prevent false signals of selection.

One important model for studying polygenic traits is the additive model, where

$$y = \mu + \alpha g + \epsilon \tag{1}$$

This model refers to the trait value, y, of each SNP in an individual. μ is the average phenotype of the population. g refers to the allelic dosage (i.e. 0, 1, or 2 copies of the allele), α is the average effect size of each allele on the phenotype, and ϵ is a residual term which captures both the effects of the environment and all other loci. GWAS use robust statistical methods to assess the evidence that an allele has a significant effect on the value of a certain quantitative trait, as well as to estimate the size of this effect.

Dominance is well understood in population genetics to mean that the effect an allele has on an organism's phenotype depends on the other allele at the locus. For instance, if a dominant (in the traditional sense) allele is represented by A and a recessive allele by a, then, the effect size of the allele a depends on whether or not there is an A at the site. Polygenic traits that exhibit dominance pose a problem for the assumption underlying tests of polygenic adaptation: that allele frequency and average effect size are uncorrelated.

A well-defined equation for average effect size is:

$$\alpha = a - a(2h - 1)(2q - 1) \tag{2}$$

Where a is half the phenotypic difference between homozygotes, h is the dominance coefficient, indicating how the phenotype of the heterozygote deviates from halfway between the homozygotes, and q is the frequency of the allele for which the effect size is being measured. Tests for polygenic adaptation assume that the average effect an allele has on a trait (α) is uncorrelated with the change in allele frequencies over time (q). However, in the presence of dominance, this assumption is violated in two ways, because 1) the allele frequency used in the model comes from the population on which the GWAS was performed and 2) because through directional dominance, a and h can be correlated. Specifically, if alleles that increase the effect size of a trait also tend to be recessive ($h < \frac{1}{2}$), then a GWAS is more likely to identify alleles that have recently increased in frequency (larger q) as significant due to their seemingly larger average α, leading to false positive signals of polygenic adaptation.

To investigate this bias, I plan to work with Dr. Jeremy Berg in the Sella Lab, approaching this problem in two steps: first, to derive mathematically an expression that can quantify the bias due to directional dominance in GWAS given known expressions and concepts in population genetics. The second step would be to use the expression derived in step one to measure this bias using real data.

Height is an anthropometric trait for which many studies have found signals of selection using evidence from GWAS [5] . However, other studies have shown that height is also subject to directional dominance [3] – a combination that makes the trait well-suited for the purposes of my research. Data will come from the recent UK Biobank study, which has gathered genetic data on about 500,000 participants from the UK [2] .

Some progress has already been made on this project. In the spring of 2017 we found using the UK Biobank data further evidence of directional dominance in height. Over the summer, I worked with Dr. Berg to begin developing a mathematical expression for the bias. I hope this semester to make significant progress on what I believe is a fascinating project in mathematics and biology.

References

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

[2] Clare Bycroft. Genome-wide genetic data on 500,000 UK Biobank participants. *bioRxiv*, 2017.

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

[4] Fernando Racimo. Detecting polygenic adaptation in admixture graphs. *bioRxiv*, 2017.

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