Project Proposal

W3500 Independent Biological Research

Kyelee Ruth Fitts

Fall 2017

Genome-wide association studies (GWAS) have become increasingly important in the field of population genetics as a method of detecting polygenic adaptation, or adaption via traits that are affected by many different loci instead of just one as per traditional Mendelian genetics. GWAS have been used to detect polygenic adaptation in anthropometric traits like height and waist-hip ratio, as well as disease phenotypes like type 2 diabetes[1, 4].

(Hmm, I think this next bit is a bit muddled. First, you talking about "this method of analysis", but you don't really explain what the method of analysis is. The basic idea behind most tests of polygenic adaptation (ours included) is that selection on a quantitative trait should cause subtle but coordinated shifts in allele frequency across all of the many sites which contribute to variation in the trait. These shifts are too small to detect individually against the background of genetic drift, but if we have an annotation of which alleles contribute to a given trait, and which know which allele is the trait increasing allele and which is the trait decreasing allele, then we can ask whether it is the case that the allele frequency changes across all of these many loci are systematically correlated with the effect that a given allele has on the trait. This sort of pattern is extremely unlikely under genetic drift, and is also what leads to divergence among populations in their average phenotypes, so observing it is thought to be pretty good evidence that the phenotype we're looking at (or something genetically correlated with it) has been impacted by directional selection.

Another point worth mentioning is that the potential bias is not so much with GWAS as with how GWAS data are typically applied in tests of polygenic adaptation. For example, GWAS does not assume anything about the relationship between an allele's additive effect on the trait and the frequency of the allele over time. That is an assumption of polygenic adaptation tests. GWAS is just a means for 1) identifying trait associated loci, and 2) estimating the size of their effects.)

As this method of analysis becomes more and more crucial to analyzing , it is necessary to ensure that the underlying assumptions of GWAS and its methods 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

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

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 e is a residual term which captures both the effects of the environment and all other loci. GWAS use robust statistical methods to obtain p-values that indicate whether a certain allele has a significant affect on the expression of a certain quantitative trait.

(Somewhere around here, I think you could add a second equation showing how the average effect depends on the difference between homozygotes and the dominance coefficient. This is the equation which appears on slides 8, 9, and 10 of Yuval's powerpoint. Tests of polygenic adaptation generally implicitly make an assumption that h (the dominance coefficient) is equal to 1/2 (i.e. that there is no additivity; note that when $h = 1/2$ the second term in that equation is 0, and the average effect of an allele is independent of its frequency in the population). So you could insert the equation

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

where a is half the difference in phenotype between homozygotes, h is the dominance coefficient, and q is the frequency of the allele in the population where the effect size is estimated. The issue for polygenic adaptation tests is that if the effect size is estimated in a population that is also included in the polygenic adaptation test (which is often unavoidable), AND there is directional dominance (i.e. a and h are correlated, and therefore $a(2h-1)$ tends to be systematically greater than or less than zero, depending on the direction of dominance, i.e. whether h is usually less than 1/2 or greater than 1/2 for trait increasing alleles), THEN the average effect (α) will tend to be larger at sites where the allele has recently increased (or decreased, again depending on the direction of dominance) in frequency. This violates the assumption of independence between effect size and direction of allele frequency change under neutrality in tests of polygenic adaptation, and is why directional dominance can potentially generate false signals in such tests adaptation. This is more than you want to explain, but perhaps you could include this second equation (which would more strongly justify showing the first I think) and try to give a brief summary. I've rearranged your test below a little bit as a first stab at expressing what I think we're trying to say here.)

However, polygenic adaptation tests assume as a null model that the average affect an allele has on a trait (α) is uncorrelated with patterns of allele frequency change over time, a condition which is violated when directional dominance is present. Specifically, if alleles that increase the effect size of a trait tend to be recessive, then GWAS will be more likely to identify alleles which have recently increased in frequency (due to their larger average effect size estimates relative to those which have decreased in frequency), and this may generate false positive signals in polygenic adaptation tests. (I swapped recessive in for dominant in the above paragraph as it more closely matches what the evidence suggests for height)

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.