

Detecting single-locus selection in admixture graphs

Fernando Racimo

March 10, 2018

Introduction

Here, we introduce a method to detect selective sweeps across the genome, when using many populations that are each related via a complex admixture graph. We made some slight modifications to the Q_B statistic from Racimo, Berg and Pickrell [3] which was originally meant to detect polygenic adaptation using admixture graphs. The new statistic - which we call S_B - does not need GWAS data and works with allele frequency data alone. It can be used to both to scan the genome for regions under strong single-locus positive selection, and to pinpoint where in the graph the selective event most likely took place.

Methods

We modified the Q_B statistic [3] to handle single-locus allele frequencies without effect sizes from an association study. We assume that the topology of the admixture graph relating a set of populations is known and that we have allele frequency data for all the populations we are studying.

For a single SNP, let \mathbf{p} be the vector of allele frequencies across populations. We then make a multivariate normal approximation to obtain a distribution with which we can model these frequencies [1, 2]:

$$\mathbf{p} \sim MVN(e, e(1-e)\mathbf{F}) \quad (1)$$

where \mathbf{F} is the neutral covariance matrix and e is the ancestral allele frequency of all populations. We can obtain a mean-centered version of the vector \mathbf{p} , which we call \mathbf{y} :

$$\mathbf{y} = \mathbf{p} - e \sim MVN(e, e(1-e)\mathbf{F}) \quad (2)$$

For an arbitrarily-defined vector \mathbf{b} with the same number of elements as there are populations:

$$\mathbf{y}^T \mathbf{b} \sim N(0, e(1-e)\mathbf{b}^T \mathbf{F} \mathbf{b}) \quad (3)$$

Our test statistic - which we call S_B - is then defined as:

$$S_B = \frac{(\mathbf{y}^T \mathbf{b})^2}{e(1-e)\mathbf{b}^T \mathbf{F} \mathbf{b}} \sim \chi_1^2 \quad (4)$$

The key is to choose a vector \mathbf{b} that represents a particular branch of our graph. For that, see the way the "branch vector" \mathbf{b} is constructed in Racimo, Berg and Pickrell [3]. Essentially, for a particular branch j , the elements of its corresponding branch vector \mathbf{b}_j are the ancestry contributions of that branch to each of the populations in the leaves of the graph.

If \mathbf{b} is chosen to be the vector corresponding to branch j when computing the statistic in equation 4, then significant values of the statistic $S_B(j)$ will capture deviations from neutrality in the graph that are attributable to a disruption that occurred along branch j .

If we only have a few genomes per population, we can increase power (at the cost of spatial genomic resolution and rigorous statistical interpretation) by combining information from several SNPs into windows, as was done for example, in Skoglund et al. [4]. In that case, we compute the average χ^2 statistic over all SNPs in each window and provide a new p-value for that averaged statistic.

References

- [1] Graham Coop, David Witonsky, Anna Di Rienzo, and Jonathan K Pritchard. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4):1411–1423, 2010.
- [2] George Nicholson, Albert V Smith, Frosti Jónsson, Ómar Gústafsson, Kári Stefánsson, and Peter Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):695–715, 2002.
- [3] Fernando Racimo, Jeremy J Berg, and Joseph K Pickrell. Detecting polygenic adaptation in admixture graphs. *Genetics*, pages genetics–300489, 2018.
- [4] Pontus Skoglund, Jessica C Thompson, Mary E Prendergast, Alissa Mittnik, Kendra Sirak, Mateja Hajdinjak, Tasneem Salie, Nadin Rohland, Swapan Mallick, Alexander Peltzer, et al. Reconstructing prehistoric african population structure. *Cell*, 171(1):59–71, 2017.