S'_B: a modified version of the S_B statistic that is more robust to low sample sizes

Fernando Racimo

October 2021

In Refoyo-Martinez et al. (2020), we defined the S_B statistic for a specific branch k of an admixture graph as:

$$S_B = \frac{((\mathbf{p} - \bar{p}\mathbf{1})^T \mathbf{b_k})^2}{\bar{p}(1-\bar{p}) \mathbf{b_k}^T \hat{\mathbf{F}} \mathbf{b_k}} \tag{1}$$

Here, $\mathbf{p}$ is the vector of sample allele frequencies across populations, $\hat{\mathbf{F}}$ is an estimate of the genome-wide allele frequency covariance matrix, and $\bar{p}$ is the mean allele frequency among populations. The elements of the branch vector $\mathbf{b_k}$ are the ancestry contributions of that branch to each of the populations in the leaves of the graph.

This statistic makes use of the sample allele frequencies $\mathbf{p}$ as an approximation the true population allele frequencies $\breve{\mathbf{p}}$, which may be particularly poor if a SNP under study has sequence data from a few individuals in a given population.

Conditional on the population allele frequency for a population j, the sample allele frequency is binomially distributed:

$$p_j | \breve{p_j} \sim Bin(2n_j, \breve{p_j}) \tag{2}$$

where n_j is the number of diploid individuals for which there is reliable genotype data at a particular SNP of interest in population j. We can approximate the above equation using a Normal distribution:

$$p_j | \breve{p_j} \sim Normal(\breve{p_j}, \breve{q_j}) \tag{3}$$

where $\breve{q_j}$ is equal to $\frac{\breve{p_j}(1-\breve{p_j})}{2n_j}$. Like $\breve{p_j}$, $\breve{q_j}$ will also not be known, and here we approximate it as $q_j = \frac{p_j(1-p_j)}{2n_j}$:

$$p_j | \breve{p_j} \sim Normal(\breve{p_j}, q_j) \tag{4}$$

Conditional on knowing the population allele frequencies for all populations, the sample allele frequencies for each population are independent of each other. In vector notation:

$$\mathbf{p} | \breve{\mathbf{p}} \sim MVN(\breve{\mathbf{p}}, diag(\mathbf{q})) \tag{5}$$

The population frequencies are, in turn, assumed to depend on some population-wide ancestral allele frequency e, as in Refoyo-Martinez et al. (2020):

$$\breve{\mathbf{p}} \sim MVN(e\mathbf{1}, \mathbf{e}(\mathbf{1}-\mathbf{e})\mathbf{F}) \tag{6}$$

where $\mathbf{1}$ is a vector of ones. If we make one further approximation and treat the variance of the conditional distribution as a constant that is not dependent on the mean, we can marginalize the population allele frequencies, and obtain:

$$\mathbf{p} \sim MVN(e\mathbf{1}, diag(\mathbf{q}) + \mathbf{e}(\mathbf{1}-\mathbf{e})\mathbf{F}) \tag{7}$$

We then mean-center the vector $\mathbf{p}$:

$$\mathbf{y} = \mathbf{p} - e\mathbf{1} \sim MVN(0, diag(\mathbf{q}) + \mathbf{e}(\mathbf{1}-\mathbf{e})\mathbf{F}) \tag{8}$$

We multiply the mean-centered vector by the branch vector b_k for a branch of interest and obtain:

$$\mathbf{y^T b} \sim Normal(0, \mathbf{b_k}^T diag(\mathbf{q})\mathbf{b_k} + e(1-e)\mathbf{b_k}^T \mathbf{F} \mathbf{b_k}) \tag{9}$$

Finally, we derive a statistic that follows a chi-squared distribution under neutrality:

$$\frac{((\mathbf{p} - \bar{p}\mathbf{1})^T \mathbf{b_k})^\mathbf{2}}{\mathbf{b_k}^T diag(\mathbf{q})\mathbf{b_k} + e(1-e)\mathbf{b_k}^T \mathbf{F} \mathbf{b_k}} \sim \chi_1^2 \tag{10}$$

If we use the mean sample frequency across populations $\bar{p}$ as an estimate of the ancestral frequency e, and also use the empirical covariance matrix $\hat{\mathbf{F}}$ as an estimate of the true covariance matrix $\mathbf{F}$, we can obtain a statistic that penalizes sites in which the number of sampled individuals for a given branch's subtended populations is low:

$$S'_B = \frac{((\mathbf{p} - \bar{p}\mathbf{1})^T \mathbf{b_k})^\mathbf{2}}{\mathbf{b_k}^T diag(\mathbf{q})\mathbf{b_k} + \bar{p}(1-\bar{p})\mathbf{b_k}^T \hat{\mathbf{F}} \mathbf{b_k}} \tag{11}$$