Supplementary Note: Fast and accurate inference of local ancestry in Latino populations

Baran, Pasaniuc, Sankararaman et al.

Power of admixture mapping as a function of accuracy

We show here that the power to detect a causal SNP using case-only admixture mapping can be quantified in terms of the correlation coefficient between the inferred and the true local ancestries. The main assumption that we need is that the error process is the same across the genome (just like in a case-control setting that assumes that the error process is the same in case and control groups). For simplicity of exposition, we assume a haploid model over a two-way admixed population. We note that the results below can be easily extended to diploid data and multi-way admixtures.

We consider case-only admixture mapping in a two-way admixture. The two ancestral populations are labelled 0 and 1. Consider the following generative model for the SNPs in a population formed by the admixture of 0 and 1. The admixture fraction of population 1 is f. At each SNP, we consider n alleles. The ancestries of each allele (denoted 0 or 1) are modelled as a random draw from Ber(f), a Bernoulli distribution with parameter f. There is an error process that flips the observed ancestries of each allele – we denote these error probabilities by $\epsilon_{0,1}, \epsilon_{1,0}$. Population 1 has a single causal SNP that has an ancestry relative risk $\lambda = 1 + K > 1$. Hence, at this causal SNP,

the ancestral alleles are drawn Ber($g = f(1+K)$) (assuming that K is small) and are then subject to the same error process.

Consider the case of no error. Assume we know f (an estimator of f formed by summing the ancestries across the SNPs would converge to f at the rate $O(\frac{1}{mn})$ where m is the number of SNPs. For a large number of SNPs, we can estimate f precisely) . Then the admixture statistic is

$$X_1 = \frac{\hat{p_1} - f}{\sqrt{\frac{f(1-f)}{n}}} \tag{1}$$

where $\hat{p_1}$ is the average ancestry across the n alleles. At a non-causal SNP, $X_1 = Z$ while at a causal SNP $X_1 = \sqrt{n}\frac{Kf}{\sqrt{f(1-f)}} + \sqrt{\frac{f(1+K)(1-f(1+K))}{f(1-f)}}Z \sim \mathbb{N}(\mu_1, {\sigma_1}^2)$ where $Z \sim \mathbb{N}(0,1)$. The power of this statistic is a monotonic function of $\sqrt{n}\lambda_1$ where λ_1 is the non-centrality parameter

$$\lambda_1 = \frac{\mu_1}{\sigma_1} = \frac{Kf}{\sqrt{f(1+K)(1-f(1+K))}} = \frac{Kf}{\sqrt{g(1-g)}} \tag{2}$$

Now for the case of errors in the inference of the ancestral alleles. Here each ancestral allele is drawn $\sim$ Ber($\tilde{f} = f(1-\epsilon_{1,0}) + (1-f)\epsilon_{0,1}$). We assume we know the $f, \epsilon_{0,1}, \epsilon_{1,0}$. In practice, when we estimate the global admixture fraction, we end up with an estimate of $f(1-\epsilon_{1,0}) + (1-f)\epsilon_{0,1}$ – so this is a reasonable assumption. Now the admixture statistic is

$$X_2 = \frac{\hat{p_2} - \tilde{f}}{\sqrt{\frac{\tilde{f}(1-\tilde{f})}{n}}} \tag{3}$$

This is a well-calibrated statistic – for non-causal SNPs, it has the usual standard normal distribution. For a causal snp, using the fact that the alleles are drawn from Ber($\tilde{g} = f(1+K)(1-\epsilon_{1,0}) + (1-$

$f(1+K))\epsilon_{0,1})$, we have $\sqrt{n}\frac{X_2-\mu_2}{\sigma_2} \sim \mathbb{N}(0,1)$ so that the non-centrality parameter

$$\lambda_2 = \frac{\mu_2}{\sigma_2} = \frac{Kf(1-\epsilon_{0,1}-\epsilon_{1,0})}{\sqrt{\tilde{g}(1-\tilde{g})}} \tag{4}$$

We can also compute the correlation between the true and the inferred alleles

$$\rho = \sqrt{\frac{g(1-g)}{\tilde{g}(1-\tilde{g})}}(1-\epsilon_{0,1}-\epsilon_{1,0}) \tag{5}$$

Plugging Equations 2 and 5 into Equation 4, we get

$$\lambda_2 = \lambda_1 \rho$$

Therefore the sample size needs to be increased by $\frac{n}{\rho^2}$ to obtain the same power as in the case of perfect inference.