Supplement 1: Admixslug Model Details

Benjamin Peter

January 27, 2022

Model Overview

Here, we present a graphical model for the joint estimation of contamination and a conditional site-frequency spectrum, implemented in a program called `admixslug`.

The model aims to combine information from both sequences and relatedness to other populations.

In brief, we assume we have NGS-data from L SNPs, and a set of one or more reference populations Z_i with reference genotypes at these loci.

Finally, we assume that for each SNP we have zero or more reads sampled from R disjoint read groups. The total number of reads from a particular read group at a particular SNP G_{sl} is n_{rsl}, and the random variable denoting the number of non-reference alleles is O_{rsl}. Each read group will have its own contamination rate c_r and error parameter e_r, that are estimated directly from the data.

We are primarily interested in estimating the latent states $\mathbf{Z}$ and $\mathbf{G}$, but we also estimate the transition matrix A between states (which, in turn, is informative about admixture proportion and times), the contamination and error rate for each read group, the substructure in each source τ_k, and the average drift since admixture from each source F_k.

Notation overview

To summarize, the notation is as follows:

- R, L denote the number of read groups and loci, respectively
- K is the number of SFS-bins
- n_{rl} the number of reads of readgroup r at SNP l.
- $\mathbf{O} = (O_{rli})$ the i-th read from read group r at SNP l
- $\mathbf{G} = (G_l)$ the genotype at SNP l
- $\mathbf{Z} = (Z_l)$ the SFS of SNP l
- F_k a parameter estimating coalescence sinc gene flow for SNP in SFS-entry k.
- τ_k the proportion of derived alleles in SFS-entry k.
- c_r proportion of contaminant reads in read group r
- e, b error rate, and reference bias
- $\theta = (c_r, e, b, \tau_k, F_k)$, the set of all parameters to be estimated

Model details

Error model

We consider sequencing error, contamination and reference bias.

The random variable X_{lri} reflects the base on the i-th sequence from read group r at SNP locus l, $X_{lri} = 0$ means the sequence carries the reference allele, and $X_{lri} = 1$ means the sequence carries the alt allele. O_{rli} is the base on the resulting sequencing read

$$\begin{aligned} P(O_{lri} = 0|X_{lri} = 0) &= 1 - e \\ P(O_{lri} = 1|X_{lri} = 0) &= e \\ P(O_{lri} = 1|X_{lri} = 1) &= b \\ P(O_{lri} = 1|X_{lri} = 1) &= 1 - b \end{aligned} \tag{1}$$

We can think of e as the sequencing error, and b as the reference bias + sequencing error.

Sequence model

There are two possible origins for each sequence X_{lri}, it is either a contaminant, or endogenous. Let $C_{lri} = 1$ mean that X_{lri} is contaminant, and $C_{lri} = 0$ mean it is not. Furthermore, let ψ_l be the alt allele frequency in a reference contamination panel, which we assume to be known. Let G_l be the genotype of the target individual at SNP l (which is either 0, 1 or 2).

$$\begin{aligned} P(X_{lri} = 0|C_{lri} = 1, \psi_l) &= 1 - \psi_l \\ P(X_{lri} = 1|C_{lri} = 1, \psi_l) &= \psi_l \\ P(X_{lri} = 0|C_{lri} = 0, G_l) &= 1 - \frac{G_l}{2} \\ P(X_{lri} = 1|C_{lri} = 0, G_l) &= \frac{G_l}{2} \end{aligned} \tag{2}$$

Faunal contamination For faunal contamination, we might expect always the ancestral allele. Let $\phi_l = 0$ mean the reference allele at locus l is ancestral, in which case faunal contamination would have the reference allele. Likewise, if the derived allele is ancestral, then $\phi_l = 1$.

If we add faunal contamination, we extend C_{lri} to an additional state, and add

$$\begin{aligned} P(X_{lri} = 0|C_{lri} = 2, \phi_l) &= 1 - \phi_l \\ P(X_{lri} = 1|C_{lri} = 2, \phi_l) &= \phi_l \end{aligned} \tag{3}$$

Contamination model

For read-group r, the probability that a read from that read group is a contaminant is

$$\begin{aligned} P(C_{lri} = 0|c_r) &= 1 - c_r \\ P(C_{lri} = 1|c_r) &= c_r \end{aligned} \tag{4}$$

independent of the locus.

Genotype model

We estimate the genotype given the conditional-SFS entry $Z_l = k$, F_k is the probability that both alleles are IBD, and τ_k is the probability that the individual has a derived allele at position k Thus

$$\begin{aligned} P(G_l = 0|Z_l = k, \tau_k, F_k) &= F_k(1 - \tau_k) + (1 - F_k)(1 - \tau_k)^2 \\ P(G_l = 1|Z_l = k, \tau_k, F_k) &= 2(1 - F_k)\tau(1 - \tau_k) \\ P(G_l = 2|Z_l = k, \tau_k, F_k) &= F_k\tau_k + (1 - F_k)\tau_k^2 \end{aligned} \tag{5}$$

Likelihood

We observe the data $\mathbf{O}$, and we know the parameters $\theta = (\tau_k, F_k, c_r, e, b)$, the contamination panel ψ and the conditional SFS $\mathbf{Z}$. The variables C_r, X_{lri} and G_l are latent variables we need to sum over.

$$P(\mathbf{O}|\theta, \psi, \mathbf{Z}) = \prod_{l,r,i} \sum_{X_{lri}=0}^{1} \sum_{C_{lri}=0}^{1} \sum_{G_l=0}^{2} P(O_{lri}|X_{lri})P(X_{lri}|C_{lri}, \psi_l, G_l)P(C_{lri}|c_r)P(G_l|Z_l, \tau_k, F_k) \tag{6}$$

Parameter estimation

We estimate parameters using the complete-data log-likelihood using an EM-algorithm.

$$\begin{aligned}\log P(\mathbf{O}, \mathbf{X}, \mathbf{C}, \mathbf{G}|\theta, \psi, \mathbf{Z}) = &\sum_{lri} \log P(O_{lri}|X_{lri}, e, b) \\ &+ \sum_{lri} \log P(C_{lri}|c_r) \\ &+ \sum_{lri} \log P(X_{lri}|C_{lri}, \psi_l, G_l) \\ &+ \sum_{l} \log P(G_l|Z_l, \tau_k, F_k)\end{aligned}$$

F-stats

$$F_4(Y, X_2; X_3, X_4) = \pi_{y3} + \pi_{y4} - \pi_{23} - \pi_{24} \tag{7}$$

$$F_3(O; Y, X_2) = \pi_{OY} + \pi_{O2} - \pi_{Y2} - \pi_{OO} \tag{8}$$

$$\tag{9}$$

from admixslug one can calculate π_{yj}:

$$\pi_{yj} = n_0(X_j) \times \tau_0(X_j) + n_1(X_j)(.5\tau_1 + .5(1 - \tau_2(X_j)) + n_2(X_j)(1 - \tau_2(X_j))$$

References