Supplement 1: Admixslug Model Details

Benjamin Peter, Arev

February 21, 2025

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.

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, we also assume we have an ancestral allele .

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 that is estimated directly from the data. In addition, there is an error and bias parameter that is shared between all libraries.

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,and reference bias

The random variable X_{lri} reflects the base on the i-th sequence from read group r at SNP locus l, and O_{rli} is the base on the resulting sequencing read. X is based on ancestral/derived alleles, whereas O is codified by reference/alternative alleles. We further have a variable W_l that is 1 if the reference allele is flipped, i.e.

$$W_l = \begin{cases} 0 & \text{if REF = Ancestral allele} \\ 1 & \text{if REF = Derived allele} \end{cases}$$

The eight possible cases are then:

Observed base O_{li}	base on seq X_{li}	flipped W_l	alleles	probability
0 (ref)	0 (anc)	0	ref = anc, alt = der	1 - e
1 (alt)	0 (anc)	0	ref = anc, alt = der	e
0 (ref)	1 (der)	0	ref = anc, alt = der	b
1 (alt)	1 (der)	0	ref = anc, alt = der	1 - b
0 (ref)	0 (anc)	1	ref = der, alt = anc	b
1 (alt)	0 (anc)	1	ref = der, alt = anc	1 - b
0 (ref)	1 (der)	1	ref = der, alt = anc	1 - e
1 (alt)	1 (der)	1	ref = der, alt = anc	e

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

This is implemented in `bwd_p_o_given_x`, which calculates a matrix of size R x 2 containing the entries $P(O_i|X_i = j)$

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 in diploid individuals).

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

In haploid regions (or more generally, in regions with different ploidy) we divide G_l by the ploidy instead.

Contamination model

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

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

independent of the locus.

Genotype likelihoods

The genotype likelihood for locus l can be written as $P(O_l|G_l) = \prod_{rj} P(O_{lrj}|G_l)$, where the product is over all reads aligning to this locus (double indexing because we multiply over all read-groups (indexed by r) and all reads per read-group (indicated by j).

The backwards probabilities

$$P(O_{lrj}|G_l) = P(O|C_{lrj} = 1)c_{lrj} + P(O|G_l, C_{lrj} = 0)(1 - c_{lrj}) \tag{3}$$

where

$$P(O|C_{lrj} = 1) = \begin{cases} \psi_l & \text{if } O = 1 \\ (1 - \psi_l) & \text{if } O = 0 \end{cases}$$

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

$$\begin{aligned} P(G_l = 0|Z_l = k, \tau_k, F_k) &= F_k(1 - \tau_k) + (1 - F_k)(1 - \tau_k)^2 \\ P(G_l = 1|Z_l = k, \tau_k, F_k) &= 2(1 - F_k)\tau(1 - \tau_k) \\ P(G_l = 2|Z_l = k, \tau_k, F_k) &= F_k\tau_k + (1 - F_k)\tau_k^2 \end{aligned} \tag{4}$$

Estimating all the τ_k is one of the main goals of `admixslug`, as they can be used to calculate F-statistics and other quantities of interest.

For example, if we compare with the Altai Neandertal and Denisova 3 genomes, we would have the following Z states:

$Z_l = k$	Altai	Denisova 3
0	0	0
1	0	1
2	0	2
3	1	0
4	1	1
5	1	2
6	2	0
7	2	1
8	2	2

This is implemented in `p_gt_diploid` and tested in `tests/test_slug.py:test_slug_p_gt_diploid`

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.

$$P(\mathbf{O}|\theta, \psi, \mathbf{Z}) = \prod_{l,r,i} \sum_{X_{lri}=0}^{1} \sum_{C_{lri}=0}^{1} \sum_{G_l=0}^{2} P(O_{lri}|X_{lri})P(X_{lri}|C_{lri}, \psi_l, G_l)P(C_{lri}|c_r)P(G_l|Z_l, \tau_k, F_k) \tag{5}$$

Forward Probabilities

Read probabilities

$$P(X_{lrj}|G_l, C_r, \psi_l) = P(X_{lrj}|C = 0)Pr(C = 0) + \sum P(X_{lrj}|C = 1)Pr(C = 1)$$

$$P(X_{lrj}|C = 0) = \sum_{g=0}^{2} P(X_{lrj}|G_{lrj} = g)P(G_l = g|Z_l)\frac{P(O_l|G_l = g)}{P(O_{lrj}|G_l = g)}$$

the ratio in the last equation is the probability of all other observations given the genotype

Backward Probabilities

Calculate the probability of all observations given a genotype (interpreted as function of the genotype $G_l = 0, 1, 2$

$$\begin{aligned}
P(O_l|G_l) &= \prod_{rj} P(O_{lrj}|G_l) \\
P(O_{lrj}|G_l) &= \sum_a P(O_{lrj}|X_{lrj} = a)P(X_{lrj} = a|G_l, C_r, \psi_l) \\
P(X_{lrj}|G_l, C_r, \psi_l) &= P(X_{lrj}|C_{lrj} = 0)P(C_{lrj} = 0) + P(X_{lrj}|\psi_l, C_{lrj} = 1)P(C_{lrj} = 1)]
\end{aligned}$$

Posterior

Posterior Genotypes The probability that genotype G_l is 0, 1, 2

$$P(G_l|O) \propto P(G_l|Z_l) \times \prod rj P(O_{lrj}|G_l)$$

Posterior Reads The probability that read X_{lrj} carries a derived allele

$$P(X_{lrj}|O) \propto P(X_{lrj}|C, G, Z, \psi)P(O_{lrj}|X_{lrj})$$

Posterior Contamination Calculate the posterior probability that read rij is contamination

$$P(C_{rij}) = \frac{\sum_a P(X = a|C_r = 1)P(O|X = a)P(C_r = 1)}{\sum_i \big[P(X = a|C = 1)P(O|X = a)P(C_r = 1) + P(X = a|C = 0)P(O|X = ia)P(C_r = 0)\big]}$$

where $a = 0, 1$

Parameter estimation

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

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

The corresponding Q-function is

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

Estimating e and b

Let $n_{a,b,c}$ be the number of reads where $O_{lri} = 0, X_{lri} = b, W_l = c$

$$\hat{e} = \frac{n_{1,0,0} + n_{1,1,1}}{n_{1,0,0} + n_{1,1,1} + n_{0,0,0} + n_{0,1,1}}$$
$$\hat{b} = \frac{n_{0,1,0} + n_{0,0,1}}{n_{0,1,0} + n_{0,0,1} + n_{1,1,0} + n_{1,0,1}}$$

Estimating c_k

$$c_r = \sum_{li} P(C|O, c')$$

, i.e. we average over the posterior contamination estimates from all reads in the read group

Estimating τ_k and F_k

Done numerically by optimizing

$$\begin{aligned} Q(\tau_k, F_k|\tau'_k, F'_k) &= \sum_l I[Z_l = k] \log P(G_l|Z_l, \tau_k, F_k) P(\mathbf{G}|\tau'_k, F'_k) \\ &= \sum_l I[Z_l = k] \sum_{g=0}^{2} \log P(G_l = g|Z_l = k, \tau_k, F_k) P(G = g|\tau'_k, F'_k) \end{aligned}$$

where I is an indicator function, $P(G = g|\tau'_k, F'_k)$ are the estimates from the previous iteration and $P(G_l = g|Z_l = k, \tau_k, F_k)$ are given by eq 4

Calculating F-statistics

We can calculate *some* F-statistics directly from the `admixslug` output. We use the estimates based on pairwise differences (Peter, 2016):

$$F_2(X, Y) = 2\pi_{xy} - \pi_{xx} - \pi_{yy} \tag{6}$$
$$F_3(X; Y, Z) = \pi_{xy} + \pi_{xz} - \pi_{yz} - \pi_{xx} \tag{7}$$
$$F_4(X, Y; Z, W) = \pi_{xz} + \pi_{yw} - \pi_{xw} - \pi_{yz}. \tag{8}$$

Assume we have L loci, and population a_{il}, d_{il} are the ancestral/derived counts in population i at locus l, respectively, such that $n_{il} = a_{il} + d_{il}$. Then

$$\pi_{ij} = \begin{cases} \frac{1}{L} \sum_l \frac{a_{il}d_{jl} + d_{il}a_{jl}}{n_{il}n_{jl}}, & \text{if } i \neq j \\ \frac{1}{L} \sum_l \frac{a_{il}d_{il}}{n_{il}(n_{il}-1)}, & \text{if } i = j \end{cases} \tag{9}$$

using the conditonal SFS from `admixslug`, we can write equivalently

$$\pi_{ij} = \begin{cases} \sum_k c_k \frac{a_{ik}d_{jk} + d_{ik}a_{jk}}{n_{ik}n_{jk}}, & \text{if } i \neq j \\ \sum_k c_k \frac{a_{ik}d_{ik}}{n_{ik}(n_{ik}-1)}, & \text{if } i = j \end{cases}, \tag{10}$$

where a_{ik}, d_{ik}, n_{ik} are now the counts of ancestral/derived/total alleles in population i and SFS-category k, and c_k is the (estimated) proportion of SNPs of this category.

These equations can directly be used to calculate π within and between pairs of reference populations. To calculate π between a reference population and the target individual, we use the estimator

$$\pi_{is} = \sum_k \frac{c_k}{n_i} \left[\tau_k a_i + (1 - \tau_k) d_i\right], \tag{11}$$

since τ_k is the expected proportion of SNPs carrying a derived allele in SFS-category k.

One caveat is that we cannot calculate πss, the heterozygosity in the target individual. Also, for references without heterozygosity (e.g. the chimp-outgroup, or pseudo-haploid reference individuals), π_{ii} cannot be calculated. By convention, we set these to zero. Thus, F-statistics involving these heterozygosities (F_2-statistics, and F_3-statistics using these individuals as samples) will be overestimated by a constant.

References

Benjamin M Peter. Admixture, Population Structure, and F-Statistics. *Genetics*, 202(4):1485–1501, April 2016. ISSN 1943-2631. URL `https://doi.org/10.1534/genetics.115.183913`.