Admixfrog model

benjamin Peter

March 6, 2019

Introduction

We aim to build an HMM that classifies a single *target* genome into regions that were derived from different diverged populations. This is motivated by admixture between modern humans, Neandertals or Denisovans, (which I will occasionally refer to as source populations) although the model is applicable to many other taxa.

Briefly, the hidden states are the sources of the individual, e.g. homozygous modern human (HH) homozygous Denisovan (DD) and homozygous Neandertal (NN), as well as the heterozygosu states HD, DN, HN where one haplotype each is derived from the sources.

We assume a homogeneous Markov model throughout, which we achieve by binning the genome in bins of constant map size (e.g. 0.005 cM), so it is possible that a bin has zero, one or many emissions, which are treated independently distributed given the hidden state. We assume that data comes in the form of biallelic SNP. We assume we have a sample of genotypes from each source, as well as reads covering the SNP from the target.

Emissions are the number of reads of the ancestral /derived allele, with the allele frequency in Humans / Neandertals / Denisovans assumed known.

The transition matrix is estimated using standard- Baum-Welch. The rest of this note deals with inferring the emission probabilities, and in particular estimating contamination.

Notation

Let

- R, L the number of read groups, and hidden-state bins, respectively
- S_l the number of SNP in bin l
- n_{rs} the number of reads of readgroup r at SNP s
- $O = (O_{rs})$ the set of derived-allele- read counts from library r at SNP s
- $Z = (Z_l)$ be the sequence of hidden states
- $R = (R_l)$ the number of contaminant reads at SNP l
- A_{sk}, B_{sk} the number of derived and ancestral allele counts in reference k at SNP s, (or more generally parameters of a Beta prior)
- T be a transition matrix
- F_k a measure of distance between source k and the introgressing population
- c_r proportion of contaminant reads in read group r
- p_{sc} the allele frequency of a contaminant at position s.
- e the sequencing error rate
- $\theta = (F_k, c_r, T, e)$, the set of all parameters to be estimated

The model can be summarized as

- $O_{rs}|G_s, n_{rs} \sim Binomial(O_{rs}; n_{rs}, p(e, c_r, p_{sc}, G_s))$
- $G_s|Z_l = k \sim Binomial(2, f_{sk})$
- $f_s|Z_l = k \sim Beta(a_k F_k, b_k F_k)$
- $Z_s|Z_{s-1} = k' \sim T_{k'}$
- $Z_0 \sim \alpha_0$

Where we want to estimate the c_r, F_k, T and α_0

1 Algorithm details

We factor the likelihood as

$$P(O|\theta) = \sum_{\mathcal{G},\mathcal{Z}} P(O|G, c, p_c, N)P(G|Z, A, B)P(Z|T)$$

Where $\mathcal{G}, \mathcal{Z}$ are the sets of all possible genotype assignments.

Under the Markov assumption, this can be further split up as

$$P(O|\theta) = P(Z_0) + \sum_{k=1}^{K}\sum_{g=0}^{2}\sum_{l=1}^{L} P(O_l|G_l, c, p_{cn}, N_l)P(G_l|Z_l = k, A_{lk}, B_{lk}, \tau)P(Z_l = k|Z_{l-1}, T)$$

we refer to these terms as the genotype likelihood, state likelihood and transition probability, respectively.

A standard trick is to assume the latent variables G, Z are known. Taking the log yields the complete data likelihood

$$\mathcal{L} = P(O, Z, G|\theta) = \log P(Z_0) + \sum_{l=1}^{L} \log P(O_l|G_l, c) + \sum_{l=1}^{L} \log P(G_l|Z_l = k, F) + \sum_{l=1}^{L} \log P(Z_l|Z_{l-1}, T) \tag{1}$$

The model is further improved by taking into account the each bin has S_L snps, and we have R distinct read groups.

$$\begin{aligned}
\mathcal{L} &= P(O, Z, G|\theta) \\
&= \log P(Z_0) \\
&+ \sum_{r=1}^{R}\sum_{s=1}^{S_L}\sum_{l=1}^{L} \log P(O_r|G_s, c) \\
&+ \sum_{s=1}^{S_L}\sum_{l=1}^{L} \log P(G_s|Z_l = k, F) \\
&+ \sum_{l=1}^{L} \log P(Z_l|Z_{l-1}, T)
\end{aligned} \tag{2}$$

The parameters we want to estimate are the contamination rate c_r for each read group, the F_{ST} for each source population, and the sequence of hidden states Z.

1.1 Genotype Likelihood

$$P(O_{rs}|G_s, c_r, n_{rs}) \sim Binom(O_{rs}; n_{rs}, p) \tag{3}$$

where $p = (1 - e)p' + e(1 - p')$ and $p' = c_{rs}p_c + (1 - c_{rs})G_s$

1.2 State Likelihood

We want to calculate the probability of a single locus, given the hidden state (i.e ancestry) is known. For homozygous ancestry, if the allele frequency of the introgressed individual were known to be p, the genotype probabilities would be $p, 2p(1-p), (1-p)^2$, respectively However, we do not observe p, but instead need to estimate it from a possibly distant proxy. There are three causes of uncertainty that we will take into account.

1. Sampling uncertainty from the reference.
2. Genetic drift between the reference and introgressing population
3. Population substructure in the reference population

1.2.1 Homozygous Hidden States

As we deal with a single hidden-state bin and locus, subscripts designating the genomic location are omitted in this section. We start with the simplest possible model, and add complexity in sequence.

Direct samples We wish to calculate the probability that the individual is homozygous, given the SNP of interest derives two haplotypes from a population whose allele frequency is known to be p. If, p is not known,

$$G \sim Binom(2, p) \tag{4}$$

however, we may need to infer it from a panel where we observe a', d' ancestral and derived alleles, respectively. In this case

$$p \sim Beta(d' + d_0, a' + a_0) \tag{5}$$

The prior d_0, a_0 can be justified by noting that even if we only observe ancestral alleles in our sample, there is still a small chance that some individuals might have a derived allele in this locus. There are differing opinions about the values of d_0, a_0. Balding & Nichols proposed a uniform prior, i.e. $a_0 = d_0 = 1$. By noting that the derived allele occurs proportionally to θ/p, one can justify $d_0 = 0, a_0 = 1$. A third option is an empirical Bayes prior: In the absence of data at a particular locus, a sensible guess for the allele frequency at this locus would be the genome wide average. Hence, we obtain an empirical prior by fitting a $Beta(d_0, a_0)$ distribution to all loci in the reference panel. In particular for ascertained data, this might be most suitable.

Writing $d = d' + d_0, a = a' + a_0$

$$\begin{aligned} P(G|d, a, d_0, a_0) &\sim Betabinom(2, d + d_0, a + a_0) \\ &= \binom{2}{G} \frac{B[G + d, 2 - G + a]}{B[d + a]} \end{aligned} \tag{6}$$

where $B[.]$ denotes the beta function.

Genetic drift between reference and target In many cases, the target individual will not be sampled immediately after admixture, but thousands of generations after. Thus, we need to take into account that the allele frequency might have changed.

Conceptually, genetic drift will increase the chance that the two alleles in the target are derived from the same ancestor: If the target is very very distant from the panel, the probability F that the two alleles are identical by descent approaches one. Conversely, if we have a very

close reference, that probability is zero.

$$\begin{aligned} P(G=2|a,d,F) &= \int P(G|p)P(p|a,d)dp \\ &= \int \left[Fp + (1-F)p^2\right] P(p|a,d)df \\ &= F\mathbb{E}\left[p|a,d\right] + (1-F)\mathbb{E}\left[p^2|a,d\right] \\ &= F\frac{d}{a+d} + (1-F)\frac{d(d+1)}{(a+d)(a+d+1)} \\ &= \frac{d^2+d+adF}{(a+d)(a+d+1)} \end{aligned} \tag{7}$$

Hence we have an excess of homozygous emissions. The probability for $G=0$ is obtained by switching A and D, and for $G=1$ by subtracting them from one:

$$P(G=1|a,d) = \frac{2ad(1-F)}{(a+d)(a+d+1)} \tag{8}$$

Drift to reference, population structure in source As a final complication, the reference may be subdivided. For example, we want to infer Neandertal introgression by using the Altai Neandertal
citepprufer2014 as a reference, despite it being substantially diverged. This will lead to additional uncertainty in the allele frequency, but should not change our belief in the average allele frequency. For this purpose, we introduce a parameter τ

$$p|a,d,\tau \sim Beta(d\tau, a\tau) \tag{9}$$

$$\begin{aligned} P(G=2|a,d,F,\tau) &= \int P(G|p)P(p|a,d,\tau)df \\ &= \int \left[Fp + (1-F)p^2\right] P(p|a,d,\tau)df \\ &= FE[p|a,d] + (1-F)E[p^2|a,d,\tau] \\ &= F\frac{d}{a+d} + (1-F)\frac{d(d\tau+1)}{(a+d)(a\tau+d\tau+1)} \\ &= \frac{\tau d^2+d+adF\tau}{(a+d)(\tau a+\tau d+1)} \end{aligned} \tag{10}$$

The probability for $G=0$ is again obtained by switching a and d, and for $G=1$ by subtracting them from one:

$$P(G=1|a,d,F,\tau) = \frac{2ad(1-F)\tau}{(a+d)(\tau a+\tau d+1)} \tag{11}$$

In the limit as $\tau=1$, we have the same as without subdivision. In the other limit as $\tau=0$, we get

$$P(G=2|a,d,F,\tau) = d/(a+d) \tag{12}$$

Drift to reference / subdivided source / alt A more complex model is

$$p|a,d,\tau \sim Beta(d'\tau + d_0, a'\tau + a_0) \tag{13}$$

1.2.2 Heterozygous States

For a heterozygous state, we know the allele frequencies in the two source pops, and we know they contribute one haplotype each. If we assume that the sources have allele frequencies p_i with prior d_i, a_i, respectively.

$$
\begin{aligned}
P(G_i = 1|\tau_i, F_i, d_i, a_i) &= \int pP(p|d_i, a_i, p\tau_i, F_i)df \\
&= E[f|d_i, a_i, F_i, \tau_i] \\
&= \frac{d_i}{a_i + d_i}
\end{aligned} \tag{14}
$$

and hence is independent from F and τ. Intuitively, this is because if we just sample one allele from a population, that population's allele frequency is the best guess for the state of our sample. The distance increases our uncertainty, but not the mean. Thus for sources S_i, S_j,

$$
\begin{aligned}
P(G = 0) &= \frac{a_i a_j}{(d_i + a_i)(d_j + a_j)} \\
P(G = 1) &= \frac{a_i d_j + a_j d_i}{(d_i + a_i)(d_j + a_j)} \\
P(G = 2) &= \frac{d_i d_j}{(d_i + a_i)(d_j + a_j)}
\end{aligned}
$$

1.2.3 Haploid States

On the sex chromosome, or when inbreeding is present, we further encounter haploid regions. Here, the heterozygous states have emission probability 0, and we just have one allele of the heterozygous case: F_{ST} does not matter.

$$
\begin{aligned}
P(G = 0) &= \frac{a}{a + d} \\
P(G = 1) &= \frac{d}{a + d}
\end{aligned} \tag{15}
$$

1.3 Estimating F and τ

$$
\begin{aligned}
Q(F, \tau|F', \tau') &= E[\log P(O, Z, G)P(G|Z, \theta')P(Z|\theta')] \\
&= \sum_{Z \in \mathcal{Z}, G \in \mathcal{G}} \log P(G|Z, A, B, \tau)P(G|Z, \theta')P(Z|O, \theta') \\
&= \sum_{k} \sum_{g=0}^{2} \sum_{l=1}^{L} \log P(G_t|Z_l = k, A_{lk}, B_{lk}, F)P(G_l|Z_l, F', O)P(Z_l = k|O, \theta')
\end{aligned} \tag{16}
$$

Since the F_k-parameters are independent for each homozygous state, these likelihoods are optimized numerically and independently.

$$
\hat{F}_k = \underset{F}{\operatorname{argmax}} \left[\sum_{g=0}^{2} \sum_{l=1}^{L} \log P(G_t|Z_l = k, A_{lk}, B_{lk}, F)P(G_l|Z_l, F', O)P(Z_l = k|O, \theta') \right] \tag{17}
$$

where $P(Z_l = k|O, \theta')$ is the output of the forward-backward algorithm, $P(G_s|Z_l = k, A_{sk}, B_{sk}, F_k)$ is given by (6). and

$$
P(G_i|Z_i = k, O) = \frac{P(O_i|G_i, c')P(G_i|Z_i = k, F')}{\sum_{g=0}^{2} P(O_i|G_i = g, c')P(G_i = g|Z = k, F'_k)} \tag{18}
$$

This equation follows from

$$
\begin{aligned}
P(G, Z|O) &= P(G|Z, O)P(Z|O) \\
&= P(G|Z, O_i)P(Z|O) \\
&= \frac{P(O_i|G)P(G|Z)}{P(O_i|Z)} P(Z|O)
\end{aligned} \tag{19}
$$

as $G_i|Z_i$ is independent of all observations except O_i and Bayes theorem.

1.4 Estimating Contamination

$$\begin{aligned} Q(c|c') &= E[\log P(O,Z,G)P(G|Z,\theta')P(Z|\theta')] \\ &= \sum_{Z\in\mathcal{Z}}\sum_{G\in\mathcal{G}} \log P(O|G,c)P(G|Z,\theta')P(Z|O,\theta') \\ &= \sum_{r=1}^{R}\sum_{k}\sum_{g=0}^{2}\sum_{s=1}^{S_l}\sum_{l=1}^{L} \log P(O_{lr}|G_s=g,c_r)P(G_s|Z_l=k,\theta',O)P(Z_l=k|O,\theta') \end{aligned} \tag{20}$$

We thus optimize

$$\hat{c}_r = \underset{c_r}{\operatorname{argmax}} \sum_{k}\sum_{g=0}^{2}\sum_{s=1}^{S_l}\sum_{l=1}^{L} \log P(O_{lr}|G_s=g,c_r)P(G_s|Z_l=k,\theta',O)P(Z_l=k|O,\theta') \tag{21}$$

using (18) and (3).

1.5 Estimationg Transitions and Initial State

T is estimated using standard Baum-welch, α_0 is set to the stationary distribution of T.

1.5.1 Transitions(HWE)

$$\hat{T}_{ij} = \frac{\sum_l P(X_l=i, X_{l+1}=j|\theta',O)}{P(O|\theta')}$$

However, the two haplotypes might change ancestry independently from each other. If the transitions from haploid states H_{ab} is known, then

$$T_i(j) = T_{i_1,i_2}(j_1,j_2) = \frac{H_{i_1}(j_1)H_{i_2}(j_2) + H_{i_1}(j_2)H_{i_2}(j_1)}{2}$$

i.e. there are always two possible transitions. We can introduce another inbreeding parameter f, so that when $i_1=i_2, j_1=j_2$ $T_i(j) = fH_i(j) + (1-f)H_i(j)^2$. Conversely, if there is only a single introgressed lineage, we would only ever expect haploid introgressed segments:

Compared to standard BW

$$\xi_{ij}(t) \propto P(O_{[1\ldots l]}, Z_l|\theta')T_{ij}P(O_{[l+2\ldots L]}|Z_{l+1},\theta')P(O_{l+1}|Z_{l+1}=k)$$

$$P(H_l=N, H_{l+1}=A, O|\theta) = P(O_{[1\ldots l]}, H_l|\theta')P(H_{l+1}|H_l)P(O_{[l+2\ldots L]}|H_{l+1},\theta')P(O_{l+1}|H_{l+1}) \tag{22}$$

$$= P(AN\to NN) + P(AA\to AN) + P(AA\to NN) \tag{23}$$

2 Prior

There are a few options for priors. For computational reasons, a beta prior is most suitable. One possibility is to use a $Beta[0,0]$ or $Beta[1,1]$, which can be done by adding psuedocounts to the observed data. More interesting is the assumption that the observed alleles were drawn from a neutral population with allele frequencies proportional to

$$P(f) \propto \frac{\theta}{x} \tag{24}$$

If we encounter a sample with a ancestral and d derived alleles, we have by Bayes' theorem

$$\begin{aligned} P(f|a,d) &\propto P(f)P(a,d|f) \\ &\propto \frac{\theta}{f}\binom{a+d}{d}f^d(1-f)^a \\ &\propto f^{d-1}(1-f)^a \\ &\sim Beta(d,a+1) \end{aligned} \tag{25}$$

I.e. knowing the ancestral allele is equivalent to adding an additional ancestral allele to the observation. If the ancestral allele is unknown, "mirroring this distribution"

For the folded SFS,

$$\begin{aligned} P(f|a,d) &\propto P(f)P(a,d|f) \\ &\propto \left[\frac{1}{f}+\frac{1}{1-f}\right] f^d(1-f)^a \\ &\propto \frac{1}{f(1-f)} f^d(1-f)^a \\ &\propto f^{d-1}(1-f)^{a-1} \\ &\sim Beta(d,a) \end{aligned} \tag{26}$$

In a finite population, alleles enter a population at a frequency of $\frac{1}{2N}$, and will become fixed if at frequency $1-\frac{1}{2N}$. Writing

$$\eta(N) = \frac{1}{2N\sum_{i=1}^{2N-1} 1/i}$$

We find that the shape of the SFS is well approximated by

$$P(f|a,d,N) \sim Beta(d+\eta, a+1-\eta) \tag{27}$$

$$\sim Beta(d+\eta, a+\eta) \tag{28}$$

for the unfolded and folded case, respectively. As the frequency cannot exceed the limits, we need to truncate this distribution:

$$\begin{aligned} E(f|a,d,N) &= \frac{B[\frac{1}{2N},a+1,d]-B[1-\frac{1}{2N},a+1,d]}{B[\frac{1}{2N},a,d]-B[1-\frac{1}{2N},a,d]} \\ &= \frac{a}{a+d}\frac{\int_\epsilon^{1-\epsilon} t^a(1-t)^{d-1}dt}{\int_\epsilon^{1-\epsilon} t^{a-1}(1-t)^{d-1}dt} \\ &= V\frac{a}{a+d} \end{aligned} \tag{29}$$

$$E(1-f|a,d,N) = 1 - V\frac{a}{a+d} = \left[1+\frac{a}{d}(1-V)\right]\frac{d}{a+d} \tag{30}$$

$$\tag{31}$$

The term in the squared parentheses is also denoted as V'. V in general will depend on

2.1 Track lengths

To a first approximation, the length of introgressed fragment is $L \sim Exp[rt]$, where r is a recombination rate and t is the time since the fragment introgressed. However, tracks may overlap or be adjacent in a diploid genome. A simple model to correct for this is to assume that after a track finishes, with some probability p a new introgressed track with the same distribution starts. Let us call this probability m. Then

$$L \sim Exp[(1-m)rt] \tag{32}$$

writing $K = \frac{1-is}{rt}$, the characteristic function of L is

$$\begin{aligned} C(s) &= E[\exp(si[(1-m)L + (1-m)m(L+L') + \dots])] \\ &= (1-m)K^{-1} + (1-m)mK^{-2} + (1-m)m^2K^{-3} + \dots \\ &= \frac{(1-m)}{K}\sum_{i=0}^{\infty}\left(\frac{m}{K}\right)^i \\ &= \frac{(1-m)}{K}\frac{1}{1-m/K} \\ &= \frac{1-m}{K-m} \\ &= \frac{rt(1-m)}{rt(1-m)-is} \end{aligned}$$

which is the characteristic function of an exponential distribution with parameter $rt(1-m)$.