Admixfrog model

Benjamin Peter

June 14, 2019

Model Overview

We present a Hidden Markov Model that classifies a single *target* genome into regions that were derived from different diverged “source” populations. This is motivated by admixture between modern humans and Neandertals, although the model is applicable to many other taxa.

We subdivide the diploid target genome into small bins of fixed length (either in genetic units if the recombination rate is known, or physical size otherwise). For each bin, we have a latent variable Z_i that designates the combination of ancestry sources, which we will denote as e.g. homozygous modern human (HH) homozygous Denisovan (DD) and homozygous Neandertal (NN), as well as the heterozygos states HD, DN, and HN, where one haplotype each is derived from the sources. Note that we do not assume phasing, so we do not distinguish HN and NH as distinct states. The vector of all Z_i will be denoted as $\mathbf{Z}$.

In each bin, there may be zero or more observed SNP G, which we assume mutate according to an infinite sites model. We assume that SNP in different bins are independent conditional on $\mathbf{Z}$. Within each bin, we similarly assume that SNP are independent conditional on Z_i, although we will investigate alternative models later. For a particular SNP, $P(G_i|Z_i)$ will depend on the allele frequency p_i in the relevant source(s), which is estimated from a reference sample from the source.

Finally, we assume that for each SNP we have reads sampled from zero or more read groups, which will carry at most two alleles. Each read group will have its own contamination and error parameters 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 rate for each read group, the substructure in each source τ_i, and the average drift since admixture from each source F_i.

Notation

Let

- R, L the number of read groups, and hidden-state bins, respectively
- K the number of sources
- S_l the number of SNP in bin l
- n_{rs} the number of reads of readgroup r at SNP s
- $\mathbf{O} = (O_{rsl})$ the set of derived-alleles from library r at SNP s in bin l
- $\mathbf{G} = (G_{sl})$ the set of genotypes of SNP s in bin l
- $\mathbf{Z} = (Z_l)$ be the sequence of hidden states
- a_{sk}, d_{sk} the number of ancestral and derived allele counts in reference k at SNP s, (or more generally parameters of a Beta prior)
- A be a transition matrix

- F_k a measure of distance between source k and the introgressing population
- c_r, e_r proportion of contaminant reads and error rate in read group r
- p_{sc} the allele frequency of a contaminant at position s.
- $\theta = (F_k, c_r, \tau_k, T, e_r)$, the set of all parameters to be estimated

The model can be summarized as

- $O_{rs}|G_s, n_{rs} \sim Binomial(O_{rs}; n_{rs}, p(e_r, c_r, p_{sc}, G_s))$
- $\psi \sim Bernoulli(F)$
- $G_s|Z_l = k \sim \psi Binomial(1, p_{sk}) + (1-\psi)Binomial(2, p_{sk})$
- $p_{sk}|Z_l = k \sim Beta(a_k F_k, d_k F_k)$
- $Z_s|Z_{s-1} = k' \sim Categorical(A_{k'})$

1 Algorithm details

The standard approach in the inference from a Hidden-Markov model is to assume the latent variables $\mathbf{G}, \mathbf{Z}$ are known, and then use an EM-algorithm to estimate parameters. The complete data log-likelihood is

$$\begin{aligned}\mathcal{L} = \log P(\mathbf{O}, \mathbf{Z}, \mathbf{G}|\theta) = &\sum_{r=1}^{R}\sum_{s=1}^{S_l}\sum_{l=1}^{L} \log P(O_{rsl}|G_{sl}, c_r, e_r) \\ &+ \sum_{l=1}^{L}\sum_{s=1}^{S_l} \log P(G_{sl}|Z_l = k, F_{Z_l}, \tau_{Z_l}) \\ &+ \sum_{l=1}^{L} \log P(Z_l|Z_{l-1}, T) + \log P(Z_0) \qquad (1)\end{aligned}$$

Transition and initial probabilities

The transition probabilities $P(Z_l|Z_{l-1})$ follow a homogeneous Markov chain based on defined by a transition matrix A. The initial probabilities $P(Z_0)$ are set to the leading eigenvector of A.

1.1 State Likelihood

We want to calculate $P(G_{sl}|Z_l = k, a_k, d_k, F_k, \tau_k)$; the probability of the genotype of a single locus, given the hidden state Z_l (i.e ancestry). This probability will differ slightly between homozygous and heterozygous ancestries: 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, p is unknown, and we need to take three causes of uncertainty into account.

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

1.1.1 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.

Known allele frequency If, the allele frequency p is known,

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

If p is unknown, we estimate it from a sample where we observe a' ancestral and d' derived alleles, respectively. In this case we model the allele frequency as

$$f \sim Beta(d' + d_0, a' + a_0) \tag{3}$$

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, particularly when the sample size is small. There are differing opinions about the values of d_0, a_0. For large samples, 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 $d_0 = 0, a_0 = 1$ (DICE model). The third option we use here is an empirical Bayes prior: In the absence of data at a particular locus, the genome wide site-frequency spectrum is a sensible guess for the allele frequency distribution of the focal locus. Hence, we obtain an empirical prior by modelling the (folded or unfolded) site-frequency spectrum as a $Beta(d_0, a_0)$ distribution. d_0 and a_0 are fitted using a method-of-moments approach from the genome-wide distribution of derived allele frequencies.

Writing $d = d' + d_0, a = a' + a_0$, the probability of G is then

$$\begin{aligned} P(G|d, a) &\sim Betabinom(2, d, a) \\ &= \binom{2}{G} \frac{B[G + d, 2 - G + a]}{B[d + a]} \end{aligned} \tag{4}$$

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 distant from the source, 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. Therefore,

$$\begin{aligned} P(G = 2|a, d, F) &= \int P(G|p)P(p|a, d)df \\ &= \int \left[Fp + (1 - F)p^2\right] P(p|a, d)df \\ &= F\mathbb{E}\left[p|a, d\right] + (1 - F)\mathbb{E}\left[p^2|a, d\right] \qquad (5) \\ &= F\frac{d}{a + d} + (1 - F)\frac{d(d + 1)}{(a + d)(a + d + 1)} \\ &= \frac{d^2 + d + adF}{(a + d)(a + d + 1)} \qquad (6) \end{aligned}$$

We note that the allle frequency distribution only enters through the first two moments, and hence any model that would us give the some moments for the allele frequency in the source will give the same results. Furthermore, if $F > 0$, this will result in 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:

$$P(G = 1|a, d) = 2(1 - F)\Big[\mathbb{E}\left[p|a, d\right] - \mathbb{E}\left[p^2|a, d\right]\Big] = \frac{2ad(1 - F)}{(a + d)(a + d + 1)} \tag{7}$$

Drift to reference, population structure in source As a final complication, the our sources might be only distantly related to the true introgressing individual. Alternatively, ascertainment bias or non-random sampling might result in a lower uncertainty in the reference

allele frequency than might be expected from the sample size.For example, we want to infer Neandertal introgression by using the Altai Neandertal
citepprufer2014 as a reference, despite it being substantially diverged from the introgressing Neandertal. Thus we follow Bading & Nichols by scaling the variance (but not mean) of the allele frequency distribution: For this purpose, we introduce a parameter τ

$$p|a, d, \tau \sim Beta(d\tau, a\tau) \tag{8}$$

which leads to

$$\begin{aligned}
P(G = 2|a, d, F, \tau) &= FE[p|a, d, \tau] + (1 - F)E[p^2|a, d, \tau] \\
&= F\frac{d}{a+d} + (1 - F)\frac{d(d\tau + 1)}{(a + d)(a\tau + d\tau + 1)} \\
&= \frac{\tau d^2 + d + adF\tau}{(a + d)(\tau a + \tau d + 1)} && (9) \\
P(G = 1|a, d, F, \tau) &= \frac{2ad(1 - F)\tau}{(a + d)(\tau a + \tau d + 1)} && (10)
\end{aligned}$$

$$\begin{aligned}
P(G = 1|a, d, F, \tau) &= \frac{2ad(1 - F)\tau}{(a + d)(\tau a + \tau d + 1)} && (11) \\
P(G = 2|a, d, F, \tau) &= \frac{\tau d^2 + d + adF\tau}{(a + d)(\tau a + \tau d + 1)} && (12)
\end{aligned}$$

1.1.2 Heterozygous States

For a heterozygous latent state, the model is considerably simpler: We know that there is exactly one haploid genome sampled from each source. If we assume that the sources have allele frequencies p_i (described by a Beta distribution with parameters d_i, a_i, respectively).

$$\begin{aligned}
P(G_i = 1|\tau_i, F_i, d_i, a_i) &= \int pP(f|d_i, a_i, p\tau_i, F_i)dp \\
&= E[p|d_i, a_i, F_i, \tau_i] \\
&= \frac{d_i}{a_i + d_i} && (13)
\end{aligned}$$

and hence is independent from both 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. Thus for sources with samples (a_i, d_i) and (a_j, d_j),

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

1.1.3 Haploid States

On the sex chromosome, or when inbreeding is present, we further encounter haploid regions. Here, G takes only values of 0 or 1, and reasoning analogous to the heterozygous case yields

$$\begin{aligned}
P(G = 0) &= \frac{a}{a + d} \\
P(G = 1) &= \frac{d}{a + d} && (14)
\end{aligned}$$

Genotype Likelihood

We assume a simple binomial model with error e, and known contamination:

$$P(O_{rs}|G_s, c_r, n_{rs}) \sim Binom(O_{rs}; n_{rs}, p) \tag{15}$$

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

2 Parameter estimation

2.1 Estimationg Transitions and Initial State

As the model is a homogeneous Hidden-Markov model, T can be estimated using the standard Baum-Welch algorithm. The initial distribution $P(Z_0) = \alpha_0$ is set to the stationary distribution of T after each iteration.

2.2 Estimating F and τ

The hierarchical nature of eq. 1 simplifies optimization considerably. To estimate τ and F, only the terms $P(G|Z)$ are needed. Likewise, to estimate c and e, only $P(O|G)$ is required.

The Q-function is

$$\begin{aligned}
Q(F,\tau|F',\tau') &= E[\log P(O,Z,G|F,\tau)P(G,Z|\tau',F',O)] \\
&= \sum_{Z\in\mathcal{Z},G\in\mathcal{G}} \log P(G|Z,F,\tau)P(G|Z,O,F',\theta')P(Z|O,F',\tau') \\
&= \sum_k \sum_{g=0}^{2}\sum_{l=1}^{L}\sum_{s=1}^{S_l} \log P(G_s=g|Z_l=k,F_k,\tau_k)P(G_s=g|Z_l=k,F',\tau',O)P(Z_l=k|O,F',\tau')
\end{aligned} \tag{16}$$

Since both F_k and τ_k only depend on the terms for the respective homozygous hidden states, we numerically optimize

$$(\hat{F}_k,\hat{\tau}_k) = \underset{F,\tau}{\operatorname{argmax}} \left[\sum_{g=0}^{2}\sum_{l=1}^{L}\sum_{s=1}^{S_l} \log P(G_s=g|Z_l=k,F_k,\tau_k)P(G_s=g|Z_l=k,F',\tau',O)P(Z_l=k|O,F',\tau')\right] \tag{17}$$

where $P(Z_l=k|O,\theta')$ is the output of the forward-backward algorithm, $P(G_s|Z_l=k,\tau_k,F_k)$ is given by (4). and

$$P(G_s|Z_l,O,F',\tau') = \frac{P(O_s|G_s,c')P(G_s|Z_l,F',\tau')}{\sum_{g=0}^{2} P(O_s|G_s=g,c')P(G_s=g|Z_l,F',\tau')} \tag{18}$$

This equation follows by applying Bayes theorem:

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

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

2.3 Estimating Contamination and Error

The update for contamination and error per read group are analogous:

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

We thus optimize for each read group

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

using (18) and (15).

3 Linkage disequilibrium

The previous derivation assumed that all SNP in a bin are independent conditional on the latent state Z. However, these SNP may be in linkage disequilibrium (LD), and thus the emission probabilities might be misspecified. A common solution is LD-pruning, i.e. taking one SNP each from each bin. Denote the i-th replicate as $\mathcal{L}^{(i)} = \log P(O^{(i)}, Z^{(i)}, G^{(i)}|\theta)$. Formally, we can repeat this procedure n times, and obtain the

$$\log P(O,Z,G|\theta)$$

4 Post-processing

5 additional notes to be removed

5.0.1 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

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

$$\begin{aligned} P(H_l=N,H_{l+1}=A,O|\theta) &= P(O_{[1\ldots l]},H_l|\theta')P(H_{l+1}|H_l)P(O_{[l+2\ldots L]}|H_{l+1},\theta')P(O_{l+1}|H_{l+1}) \quad (22) \\ &= P(AN\to NN) + P(AA\to AN) + P(AA\to NN) \quad (23) \end{aligned}$$

6 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

$$\begin{aligned} P(f|a,d,N) &\sim Beta(d+\eta, a+1-\eta) &(27)\\ &\sim Beta(d+\eta, a+\eta) &(28) \end{aligned}$$

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

$$\begin{aligned} E(f|a,d,N) &= \frac{B[\frac{1}{2N}, a+1, d] - B[1-\frac{1}{2N}, a+1, d]}{B[\frac{1}{2N}, a, d] - B[1-\frac{1}{2N}, a, d]} \\ &= \frac{a}{a+d}\frac{\int_\epsilon^{1-\epsilon} t^a(1-t)^{d-1}dt}{\int_\epsilon^{1-\epsilon} t^{a-1}(1-t)^{d-1}dt} \\ &= V\frac{a}{a+d} &(29)\\ E(1-f|a,d,N) &= 1 - V\frac{a}{a+d} = \left[1 + \frac{a}{d}(1-V)\right]\frac{d}{a+d} &(30)\\ & &(31) \end{aligned}$$

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

6.1 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 $rt(1-m)$.

6.2 SMC'

Under the SMC'-model, the rate is

$$r = 2N(1-m)\left(1 - exp\left(-\frac{t}{2N}\right)\right). \tag{33}$$

$$\approx t(1-m)\left(1-\frac{\tau}{2}\right) + O\left(N^{-3}\right) \tag{34}$$

where $\tau = t/2N$ is the admixture time in coalescence units. This follows from Liang & Nielsen and a Taylor expansion in N.

6.3 Ralph coop theory

Let $N(x)$ denote the number of IBD blocks of genetic length at least x shared by two individual chromosomes, and $N_n(x)$ the number of blocks inherited through a path of n meioses. $N(x) = \sum_n N_n(x)$ and

$$\mathbb{E}\left[N(x)\right] = \sum_n \mathbb{E}\left[N_n(x)\right]$$

. $K_n(x)$ denotes the number of pieces of length at least x after n meioses. $\mu(n)$ denotes the probability that the tract introgressed n meioses ago.

$$\mathbb{E}\left[N(x)\right] = \sum_n \mathbb{E}\left[\mu(n)K_n(x)\right]$$

.

Wen we consider introgression from a Neandertal, the same logic applies, except we stop when we enter a Neandertal population. I.e. n measures the number of meioses after the introgression event.

$$K_t(x) = (t(G-x)+22)exp(-xt)$$

6.4 Time of most recent gene flow

We estimate the time of the most recent gene flow from the length of the longest fragment: After T generations, a fragment has an exponential distribution with rate $r = T(1-M)(1-\tau/2)1/\text{Morgan} = T(1-m)(1-\tau/2)\frac{1}{100cM}$.

We approximate the joint distribution of introgressed tracts as independent exponentials. (In truth, they are likely positively correlated, hence this will be an overestimate). We have an expected number of $n = 2mg\bar{r}$ fragments, where m is the proportion of introgressed material, and g is the length of the genome, and $\bar{r}$ is the average rate of an introgressed fragment. Assuming n is large, and the longest fragment is L_0. will have likelihood:

$$\begin{aligned} P(L_0 = l|r, m, g, n) &= P(L_1 \leq l, \dots L_n \leq l|r_1, \dots r_n)P(L_0 = l|r_0) \\ &\approx \prod_{i=1}^{n} P(L_i < l|r_i)P(L_0 = l|r_0) \\ &\approx (1 - \exp(-rl))^{2mgr} r \exp(-rl) \qquad (35) \end{aligned}$$

This estimate is true if admixed fragments were i.i.d. exponentials. Thus, this estimator will likely overestimate the most recent time of gene flow as

1. if admixture is ongoing, $r_i \geq r$.
2. if admixture fragments lengths are positively correlated, $P(L_1 \leq l, \dots L_n \leq l|r_1, \dots r_n) < \prod_{i=1}^{n} P(L_i < l|r_i)$
3. the longest introgressed tract may not be the oldest one.

6.5 EB time of gene flow per fragment

The goal is to estimate, for each fragment, when it was introgressed.

$$P(T_i|L_i) \propto P(L_i|T_i)P(T_i)$$

$P(L_i)$ is estimated from the genome-wide distribution of tract lengths, and we assume it is gamma distributed. Therefore,

$$P(T_i|L) \sim \Gamma(a+1, b+L_i) \qquad (36)$$

We estimate the time of the most recent gene flow from the length of the longest fragment: After T generations, a fragment has an exponential distribution with rate $r = T(1-M)(1-\tau/2)1/\text{Morgan} = T(1-m)(1-\tau/2)\frac{1}{100cM}$.

We approximate the joint distribution of introgressed tracts as independent exponentials. (In truth, they are likely positively correlated, hence this will be an overestimate). We have an expected number of $n = 2mg\bar{r}$ fragments, where m is the proportion of introgressed material, and g is the length of the genome, and $\bar{r}$ is the average rate of an introgressed fragment. Assuming n is large, and the longest fragment is L_0. will have likelihood:

$$\begin{aligned} P(L_0 = l|r, m, g, n) &= P(L_1 \leq l, \dots L_n \leq l|r_1, \dots r_n)P(L_0 = l|r_0) \\ &\approx \prod_{i=1}^{n} P(L_i < l|r_i)P(L_0 = l|r_0) \\ &\approx (1 - \exp(-rl))^{2mgr} r \exp(-rl) \qquad (37) \end{aligned}$$

This estimate is true if admixed fragments were i.i.d. exponentials. Thus, this estimator will likely overestimate the most recent time of gene flow as

1. if admixture is ongoing, $r_i \geq r$.
2. if admixture fragments lengths are positively correlated, $P(L_1 \leq l, \dots L_n \leq l|r_1, \dots r_n) < \prod_{i=1}^{n} P(L_i < l|r_i)$
3. the longest introgressed tract may not be the oldest one.

6.6 $P(G_1, G_2, G_S|Z_l)$

The SNP in a single bin are potentially linked (although we assume they have the same ancestry). The basic model assumes that $P(G_1, G_2, G_S|Z_l) = \prod_S P(G_s|Z_l)$, which may be unrealistic. Let us assume that Z is homozygous, and that it's prior is $Beta(a_i, d_i)$. The joint distribution of two SNP i and j is then

$$P(G_0 = 2) = F\mathbb{E}\left[f_A\right] + (1-F)\mathbb{E}\left[f_A^2\right] \tag{38}$$

$$P(G_1 = 2, G_0 = 2) = F\mathbb{E}\left[f_{AB}\right] + (1-F)(1-r)\mathbb{E}\left[f_{AB}^2\right] + (1-F)r\mathbb{E}\left[f_A^2 f_B^2\right] \tag{39}$$

$$P(G_1 = 2|G_0 = 2) = \frac{F\mathbb{E}\left[f_A\right] + (1-F)\mathbb{E}\left[f_A^2\right]}{F\mathbb{E}\left[f_{AB}\right] + (1-F)(1-r)\mathbb{E}\left[f_{AB}^2\right] + (1-F)r\mathbb{E}\left[f_A^2 f_B^2\right]} \tag{40}$$

If we consider two loci, we assume they are completely linked, so that the haplotype frequencies are $f_{01} = f_0 f_1 + D_{01}$

$$\mathbb{E}\left[f_{AB}\right] = \mathbb{E}\left[f_A f_B\right] + \mathbb{E}\left[D_{AB}\right] \tag{41}$$

$$\mathbb{E}\left[f_{Ab}\right] = \mathbb{E}\left[f_A - f_A f_B\right] + \mathbb{E}\left[D_{01}\right] \tag{42}$$

$$\mathbb{E}\left[f_{aB}\right] = \mathbb{E}\left[f_B - f_A f_B\right] + \mathbb{E}\left[D_{01}\right] \tag{43}$$

$$\mathbb{E}\left[f_{ab}\right] = \mathbb{E}\left[1 + f_A f_B - f_A - f_B\right] + \mathbb{E}\left[D_{01}\right] \tag{44}$$

$$Cov(f_a^2, f_b^2) = \mathbb{E}\left[f_A^2 f_B^2\right] - \mathbb{E}\left[f_A^2\right]\mathbb{E}\left[f_B^2\right]$$

Further assuming $\mathbb{E}\left[((f_a - \mathbb{E}\left[f_a\right])^2, (f_b - \mathbb{E}\left[f_b\right])|f_a, f_b) = 0\right]$, $\mathbb{E}\left[f_a^2 f_b^2|f_a, f_b\right] - f_a^2 f_b^2 = 0$ $\mathbb{E}\left[f_a^2|f_b^2\right]$

The basic idea is that we have a bunch of fragments that may be the result from the same introgression event. A composite-likelihood estimator for each fragment assumes that all fragments at a particular position have entered the population at the same time, and are independent after that. This is not correct since the fragments at a position are correlated; hence one might expect to be overly confident in ones estimates, but point estimates are likely to be accurate.

The main issue is that one does not know the rate at wihc fragments deteriorate, since the local recombination rate will be unknown. One possibility is to assume that there is a single average recombination rate-offset factor, which can be learned from e.g. regressing the mean fragment length against introgression time. Based on that, a introgression-time distributon can be inferred.