Some draft notes

Benjamin Peter

February 8, 2020

1 additional notes to be removed

2 Post-processing

We assume that an introgressed fragment $f \in F$ (or parts thereof) are observed in multiple individuals $s \in S$, and we know it's distribution from a number of samples. We assume that each fragment entered the population at a time T_f, and approximate the likelihood as independent between samples:

$$P(L_f|A, T_f, k) = \prod_s P(L_{sf}|A_s, T_f, k) \approx \prod_s \sum_i P(L_{sfi}|A_s, T_f, k)$$

The likelihood $P(L_{sfi}|A_s, T_f, k)$ is written as

$$P(L_{sfi}|A_s, T_f, k) = \begin{cases} \frac{T_f - A_s}{k} \exp\left(-\frac{T_f - A_s}{k} L_{sfi}\right), & \text{if } T_f \geq A_s \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

To estimate all T, we assume that all fragments are independent
$P(T|A, k, L) = \prod_f P(T_f|A, k, L_f)$

$$P(T_f|L_f, A, k) = \prod_s \frac{P(L_{sf}|A_s, T_f, k)P(T_f|A, k)}{P(L_{sf}|A_s, k)} \tag{2}$$

which yields

$$P(T|L, A, k) = \prod_f \prod_s \frac{P(L_{sf}|A_s, T_f, k)P(T_f|A, k)}{P(L_{sf}|A_s, k)} \tag{3}$$

Let us denote l_{fsi}, the i-th estimate of the length of a

Given a set of samples S with ages A_s, and a draw of introgressed fragments for each sample f_i. We assume the following mixture model:

$$P(f_{is}|A_s) = \sum_k \pi_k P(f_i s|A_s, T_k) P(T_k, a_k, b_k)$$

$$f_{is}|Z_i = i, A_s \sim Exp[T_i - A_s] \tag{4}$$
$$T_i|Z_i = i \sim \Gamma[a_i, b_i] \tag{5}$$
$$Z_i \sim Multinomial(\pi) \tag{6}$$

2.0.1 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

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

$$P(H_l = N, H_{l+1} = A, O|\theta) = P(O_{[1\ldots l]}, H_l|\theta')P(H_{l+1}|H_l)P(O_{[l+2\ldots L]}|H_{l+1},\theta')P(O_{l+1}|H_{l+1}) \tag{7}$$

$$= P(AN \to NN) + P(AA \to AN) + P(AA \to NN) \tag{8}$$

3 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

$$P(f) \propto \frac{\theta}{x} \tag{9}$$

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

$$\begin{aligned} P(f|a,d) &\propto P(f)P(a,d|f) \\ &\propto \frac{\theta}{f}\binom{a+d}{d}f^d(1-f)^a \\ &\propto f^{d-1}(1-f)^a \\ &\sim Beta(d, a+1) \end{aligned} \tag{10}$$

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,

$$\begin{aligned} P(f|a,d) &\propto P(f)P(a,d|f) \\ &\propto \left[\frac{1}{f} + \frac{1}{1-f}\right]f^d(1-f)^a \\ &\propto \frac{1}{f(1-f)}f^d(1-f)^a \\ &\propto f^{d-1}(1-f)^{a-1} \\ &\sim Beta(d, a) \end{aligned} \tag{11}$$

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

$$P(f|a,d,N) \sim Beta(d+\eta, a+1-\eta) \tag{12}$$

$$\sim Beta(d+\eta, a+\eta) \tag{13}$$

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

$$\begin{aligned} E(f|a,d,N) &= \frac{B[\frac{1}{2N},a+1,d] - B[1-\frac{1}{2N},a+1,d]}{B[\frac{1}{2N},a,d] - B[1-\frac{1}{2N},a,d]} \\ &= \frac{a}{a+d}\frac{\int_\epsilon^{1-\epsilon} t^a(1-t)^{d-1}dt}{\int_\epsilon^{1-\epsilon} t^{a-1}(1-t)^{d-1}dt} \\ &= V\frac{a}{a+d} \end{aligned} \tag{14}$$

$$E(1-f|a,d,N) = 1 - V\frac{a}{a+d} = \left[1+\frac{a}{d}(1-V)\right]\frac{d}{a+d} \tag{15}$$

$$\tag{16}$$

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

3.1 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

$$L \sim Exp[(1-m)rt] \tag{17}$$

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 $rt(1-m)$.

3.2 SMC'

Under the SMC'-model, the rate is

$$r = 2N(1-m)\left(1 - exp\left(-\frac{t}{2N}\right)\right). \tag{18}$$

$$\approx t(1-m)\left(1-\frac{\tau}{2}\right) + O\left(N^{-3}\right) \tag{19}$$

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

3.3 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.

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

3.4 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:

$$\begin{aligned} P(L_0 = l|r, m, g, n) &= P(L_1 \leq l, \dots L_n \leq l|r_1, \dots r_n)P(L_0 = l|r_0) \\ &\approx \prod_{i=1}^{n} P(L_i < l|r_i)P(L_0 = l|r_0) \\ &\approx (1-\exp(-rl))^{2mgr} r\exp(-rl) \end{aligned} \tag{20}$$

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.

3.5 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,

$$P(T_i|L) \sim \Gamma(a+1, b+L_i) \tag{21}$$

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:

$$\begin{aligned} P(L_0 = l|r, m, g, n) &= P(L_1 \leq l, \dots L_n \leq l|r_1, \dots r_n) P(L_0 = l|r_0) \\ &\approx \prod_{i=1}^{n} P(L_i < l|r_i) P(L_0 = l|r_0) \\ &\approx (1 - \exp(-rl))^{2mgr} r \exp(-rl) \end{aligned} \tag{22}$$

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.

3.6 $P(G_1, G_2, G_S|Z_l)$

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

$$P(G_0 = 2) = F\mathbb{E}\left[f_A\right] + (1 - F)\mathbb{E}\left[f_A^2\right] \tag{23}$$

$$P(G_1 = 2, G_0 = 2) = F\mathbb{E}\left[f_{AB}\right] + (1 - F)(1 - r)\mathbb{E}\left[f_{AB}^2\right] + (1 - F)r\mathbb{E}\left[f_A^2 f_B^2\right] \tag{24}$$

$$P(G_1 = 2|G_0 = 2) = \frac{F\mathbb{E}\left[f_A\right] + (1 - F)\mathbb{E}\left[f_A^2\right]}{F\mathbb{E}\left[f_{AB}\right] + (1 - F)(1 - r)\mathbb{E}\left[f_{AB}^2\right] + (1 - F)r\mathbb{E}\left[f_A^2 f_B^2\right]} \tag{25}$$

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

$$\mathbb{E}\left[f_{AB}\right] = \mathbb{E}\left[f_A f_B\right] + \mathbb{E}\left[D_{AB}\right] \tag{26}$$

$$\mathbb{E}\left[f_{Ab}\right] = \mathbb{E}\left[f_A - f_A f_B\right] + \mathbb{E}\left[D_{01}\right] \tag{27}$$

$$\mathbb{E}\left[f_{aB}\right] = \mathbb{E}\left[f_B - f_A f_B\right] + \mathbb{E}\left[D_{01}\right] \tag{28}$$

$$\mathbb{E}\left[f_{ab}\right] = \mathbb{E}\left[1 + f_A f_B - f_A - f_B\right] + \mathbb{E}\left[D_{01}\right] \tag{29}$$

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

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.

4 admixfrog one-snp-per-bin

The standard likelihood is

$$\begin{aligned} P(\mathbf{O},\mathbf{Z},\mathbf{G}|\theta) = &\prod_{r=1}^{R}\prod_{s=1}^{S_l}\prod_{l=1}^{L} P(O_{rsl}|G_{sl},c_r,e_r) \\ &\times \prod_{l=1}^{L}\prod_{s=1}^{S_l} P(G_{sl}|Z_l=k,F_{Z_l},\tau_{Z_l}) \\ &\times \prod_{l=1}^{L} \log P(Z_l|Z_{l-1},A) \times P(Z_0). \end{aligned} \tag{30}$$

which assumes that SNP are independent given the ancestral state. Alternatively, we can assume that we sample exactly one SNP from each bin ($S_l = 1, \forall l$), in which case we would optimize:

$$\begin{aligned} P(\mathbf{O},\mathbf{Z},\mathbf{G}|\theta) = &\prod_{r=1}^{R}\prod_{l=1}^{L} P(O_{rl}|G_{l},c_r,e_r) \\ &\times \prod_{l=1}^{L} P(G_{l}|Z_l=k,F_{Z_l},\tau_{Z_l}) \\ &\times \prod_{l=1}^{L} \log P(Z_l|Z_{l-1},A) \times P(Z_0). \end{aligned} \tag{31}$$

using

$$\begin{aligned} \log P(\mathbf{O},\mathbf{Z},\mathbf{G}|\theta) = &\sum_{r=1}^{R}\sum_{l=1}^{L} \log P(O_{rl}|G_{l},c_r,e_r) \\ &+ \sum_{l=1}^{L} \log P(G_{l}|Z_l=k,F_{Z_l},\tau_{Z_l}) \\ &+ \sum_{l=1}^{L} \log P(Z_l|Z_{l-1},A) + P(Z_0). \end{aligned} \tag{32}$$

To sum over all possible samplings V, we could use

$$\sum_{v}^{V} \log P(O_v,Z_v,G_v|\theta) \tag{33}$$

$$. \tag{34}$$

(Peter and Slatkin, 2013; Pakes, 1971) Peter et al. (2010); Peter (2016); Huerta-Sánchez et al. (2014)

References

Emilia Huerta-Sánchez, Xin Jin, Zhuoma Bianba, Benjamin M. Peter, Nicolas Vinckenbosch, Yu Liang, Xin Yi, Mingze He, Mehmet Somel, Peixiang Ni, and others. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513):194–197, 2014.

A. G. Pakes. Branching Processes with Immigration. *Journal of Applied Probability*, 8(1):32–42, March 1971. ISSN 0021-9002. ArticleType: research-article / Full publication date: Mar., 1971 / Copyright © 1971 Applied Probability Trust.

Benjamin M. Peter. Admixture, Population Structure and F-Statistics. *Genetics*, page genetics.115.183913, January 2016. ISSN 0016-6731, 1943-2631.

Benjamin M. Peter and Montgomery Slatkin. Detecting Range Expansions from Genetic Data. *Evolution*, 67(11):3274–3289, 2013. ISSN 1558-5646.

Benjamin M Peter, Daniel Wegmann, and Laurent Excoffier. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular ecology*, 19(21):4648–4660, November 2010. ISSN 1365-294X.