1 Generative Damage model for ancient DNA

This is a summary of the model that John Novembre proposed as a generative model for ancient DNA on which inference can be carried out.

Say for individual i, π_i is the probability that it is ancient. so, if we observe a read r from the i th individual, then for that read, we can define a latent variable $Z_{r,i}$ such that

$$Z_{r,i}|\pi_i \sim Bern(\pi_i)$$

Then for each position p of that read (from the end of the read), we define another latent random variable

$$\begin{align} D_{r,p,i} &= 1 \text{ if damage / error} \tag{1} \\ &= 0 \text{ polymorphism} \tag{2} \\ \tag{3} \end{align}$$

Then we can assign the following probabilities of $D|Z$.

$$Pr\left[D_{r,p,i} = 0 | Z_{r,i} = 0\right] = \theta_\pi = 1/1000$$
$$Pr\left[D_{r,p,i} = 0 | Z_{r,i} = 1\right] = \theta_\pi = 1/1000$$
$$Pr\left[D_{r,p,i} = 1 | Z_{r,i} = 1\right] = f_1(p, \alpha)$$
$$Pr\left[D_{r,p,i} = 0 | Z_{r,i} = 1\right] = f_2(p)$$

where f_1 and f_2 are both decreasing functions with respect to p (which is the minimum distance from the end of the read), we sort of assume that we know f_2 from 1000 genomes data may be, because otherwise it would be hard to distinguish between f_1 (damage) and f_2 (sequencing error).

From here on, we can go in two directions. If the primary aim is to detect contamination in the sample, we can then say

$$X_{r,p,i}|D_{r,p,i} \sim Mult\left(1, p_{i1}, p_{i2}, \cdots, p_{ij}\right)$$

where

$$p_{ij} = \sum_{k=1}^{K} \omega_{ik}\theta_{kj}$$

.

where j is the signature pattern.

Another direction would be to define another latent random variable $M_{r,p,i}$ such that

$$Pr\left[M_{r,p,i} = 0 | D_{r,p,i} = 0\right] = 1$$

$$Pr\left[M_{r,p,i} = 1 | D_{r,p,i} = 1\right] = 1$$

$$Pr\left[M_{r,p,i} = 2 | D_{r,p,i} = 0\right] = \omega$$
$$Pr\left[M_{r,p,i} = 2 | D_{r,p,i} = 1\right] = 1 - \omega$$