1 First method

Assume there are N population samples out of which N_p are present day samples and N_a are ancient samples. We have the genotypic configuration of L segregating sites for all individuals. Assume that N_s are the number of source ancestors (usually ancient samples) and N_u are the number of unobserved ancients which is not known. We will start with a user defined value for N_u (easier to think of this as a number of clusters among the unobserved ancients). Let $f_{s,l}$ represent the allele frequency of the l th allele for the source ancestral population s.

$$X_{il} \sim Binomial(2, \sum_{s=1}^{N_s} \omega_{is} f_{sl} + \sum_{u=1}^{N_u} \omega_{iu} f_{ul})$$

for the i th sample, i runs from 1 to N and l th SNP where l runs from 1 to L. ω_{ik} represents mixture proportions for i th individual and s are the labels for the source ancestral populations and u be the labels for the unobserved ancestral populations or other effects (which may be technical as well). The solution model has unknowns $\omega_{i.}$ vectors for each i and $f_{u.}$ for the unobserved ancient ancestors.

To solve this problem, we use a EM algorithm procedure in the same lines as Skotte et al (2013). Following their approach, if we assume there is no contamination in the samples, we have a nice iterative scheme to update the parameters.

$$a^n_{ijk} = \frac{\omega^n_{ik} f^n_{jk}}{\sum_k \omega^n_{ik} f^n_{jk}} \qquad k = 1, 2, \cdots, N_s.N_s + 1, \cdots, N_u + N_s$$

$$b^n_{ijk} = \frac{\omega^n_{ik}(1 - f^n_{jk})}{\sum_k \omega^n_{ik}(1 - f^n_{jk})} \qquad k = 1, 2, \cdots, N_s.N_s + 1, \cdots, N_u + N_s$$

$$\omega^{(n+1)}_{ik} = \frac{1}{2L} \sum_{j=1}^{L} (a^n_{ijk} + b^n_{ijk}) \;\; \forall k \;\; \forall i$$

$$f^{(n+1)}_{is} = f^{(n)}_{is} \;\; \forall i \;\; s = 1, 2, \cdots, N_s$$

$$f^{(n+1)}_{iu} = \frac{\sum_{i=1}^{N} a^n_{ijk}}{\sum_{i=1}^{N} a^n_{ijk} + \sum_{i=1}^{N} b^n_{ijk}} \;\; \forall i \;\; u = 1, 2, \cdots, N_u$$

We used this iterative scheme and ran our model on both simulated data, using simulations as per the above model as well as the Wright Fisher model with drift and the Haak Human Origins data on the 403 samples of interest.