Ian Arriaga MacKenzie

Introduction

Online aggregation of genetic sequencing data, and the publicly available data produced, are invaluable tools in research and the clinic. These databases have numerous applications including prioritizing causal variants and leveraging common controls. However, summarizing individual-level genotype data can mask population structure, resulting in increased potential for confounding and reduced power. This limits the utility of these databases, especially for understudied and ancestrally diverse populations.

Summix is a method used to deconvolute ancestry and provide ancestry-adjusted allele frequencies from summary genetic data. The method uses a reference panel of allele frequencies specific to ancestral populations, and estimates ancestry within a target admixed (or homogeneous) population.

Methods

Simulations

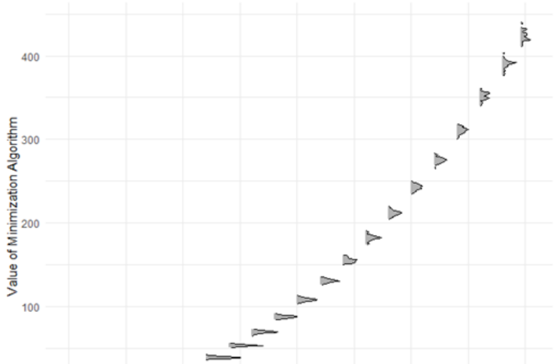
Data is simulated using 5 continental reference ancestries from 1000 Genomes (AFR, EAS, EUR, IAM, SAS) and 1 ”Unknown” ancestry. For each simulation replicate, the proportions of the 5 reference ancestries are random, while the proportion of the unknown ancestry is fixed, and the allele frequency for the unknown ancestry is random. These simulated admixed ancestries are created using the multinomial distribution with the homozygous/heterozygous genotype frequencies as parameters.

Summix Least Square Error

The minimization model defined by Summix is convex, and therefor solves to an absolute minimum when solved by SQP. The value of the final minimized solution is proportional to the amount of hidden ancestry present in the simulated admixed proportions which is not present in the reference panel.

Fixation Index

The fixation index is a measure between 0 and 1 which shows genetic variability between populations. With human populations, these values typically range from 0.01 (French-Spaniards) to 0.46 (Mbuti-Papuans). The F_{ST} is also shown to be proportional to the amount of hidden ancestry.



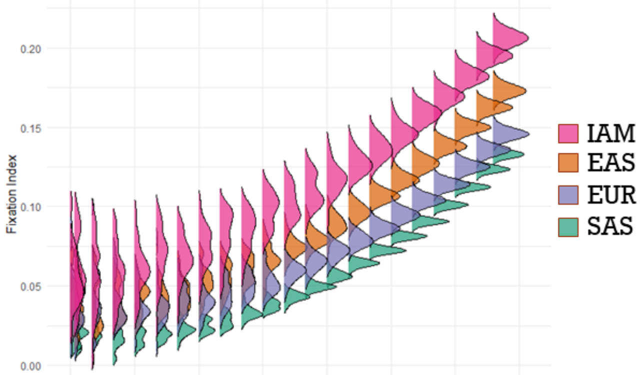


Figure 1. The Least Squares Error and F_{ST} values for simulations where the African ancestry was "unknown" and removed from the reference panel. There is a relationship between the error and F_{ST} and the amount of AFR ancestry present in the admixed population. This relationship is shown when the other 4 ancestries are "hidden".

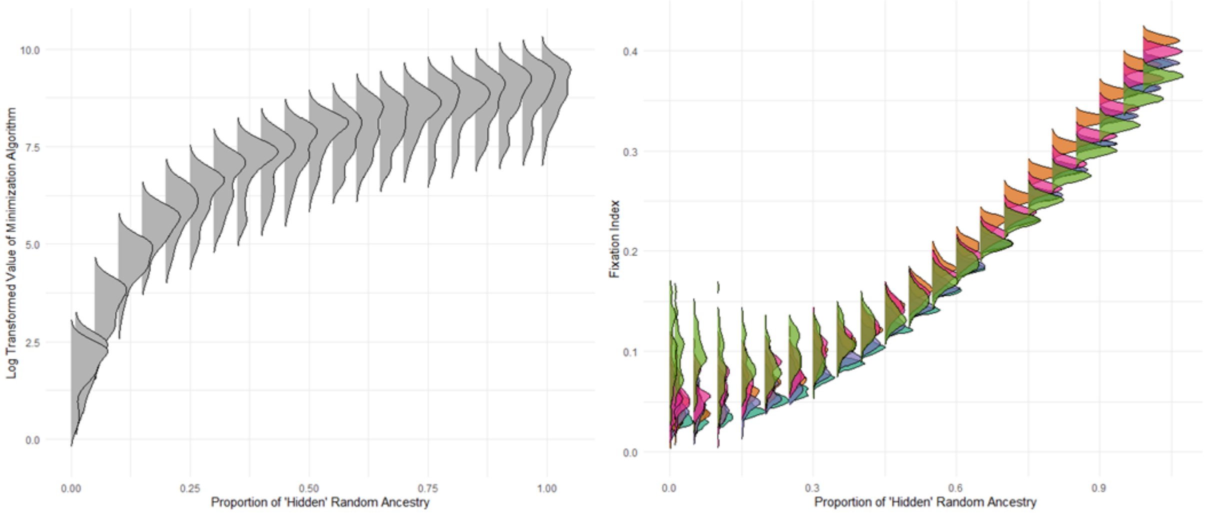


Figure 2. The log transformed least squares error and the F_{ST} values for a simulated population with a fixed unknown ancestry.

Weir and Cockerham's F_{ST}

Cockerham defined F_{ST} as the ratio of variance between populations to the total variance in the ancestral population, and Weir and Cockerham (WC) provide the following estimator:

$$\hat{F}_{ST}^{WC} = 1 - \frac{2\frac{n_1 n_2}{n_1+n_2}\frac{1}{n_1+n_2-2}\left[n_1\bar{p}_1\left(1-\bar{p}_1\right)+n_2\bar{p}_2\left(1-\bar{p}_2\right)\right]}{\frac{n_1 n_2}{n_1+n_2}\left(\bar{p}_1-\bar{p}_2\right)^2+\left(2\frac{n_1 n_2}{n_1+n_2}-1\right)\frac{1}{n_1+n_2-2}\left[n_1\bar{p}_1\left(1-\bar{p}_1\right)+n_2\bar{p}_2\left(1-\bar{p}_2\right)\right]}$$

This estimator is shown to be dependent on the ratio of sample sizes between the two populations, and therefor a different estimator (Hudson's) is recommended.

Hudson's F_{ST}

The estimator defined by Hudson et. all is:

$$\hat{F}_{ST}^{H} = \frac{\left(\bar{p}_1-\bar{p}_2\right)^2-\frac{\bar{p}_1\left(1-\bar{p}_1\right)}{n_1-1}-\frac{\bar{p}_2\left(1-\bar{p}_2\right)}{n_2-1}}{\bar{p}_1\left(1-\bar{p}_2\right)+\bar{p}_2\left(1-\bar{p}_1\right)}$$

F_{ST} across multiple SNPs

Bhatia et al. recommend using the ratio of averages to estimate F_{ST} across multiple SNPs, as opposed to the average of ratios.

Ratio of Averages:

$$R_n = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

Using these F_{ST} estimators, there seems to be a relationship between the Least Square Error and F_{ST} which can be used to estimate the proportion of hidden ancestry present in the admixed population.

Estimate Unknown AF

If we modify our original Summix minimization equation, we can include an "unknown" reference ancestry proportion and AF to be solved for:

$$f(\pi) = \left(AF_{OBS} - \sum_{k=1}^{K} (\pi_k \cdot AF_{ref,k}) - \pi_U \cdot AF_U \right)^2$$

subject to the constraints that $\sum \pi_k + \pi_U = 1$ and $0 \leq \{AF_{ref,k}, AF_U\} \leq 1$.

However, there are two limitations to note:

1. The optimal minimized solution (regardless of reference AFs) would be $AF_{OBS} = AF_U$ and $\pi_U = 1$. This will always minimize to 0, and therefor some limits should be placed on π_U.

2. If the AFs are allowed to vary by SNP, this could be computationally intensive. Currently our SQP model only has 5 variables, the inclusion of pi_U would be a 6th variable, but every SNP AF would be another variable. One possible solution would be to make the AF_U a vector across all SNPs, but that also would limit the utility of solving for the AFs.

GnomAD AFR Example - 2 Reference, 1 Unknown

If we model the gnomAD AFR ancestry with 2 reference ancestries (AFR and EUR) and 1 unknown ancestry, we can make substitutions regarding F_{ST} but the two variables of interest π_U and AF_U cannot be solved for.

$$f(\pi) = (AF_{OBS} - \pi_{AFR} \cdot AF_{AFR} - \pi_{EUR} \cdot AF_{EUR} - \pi_U \cdot AF_U)^2$$

$$F_{ST}^{OBS-AFR} = \frac{(AF_{OBS} - AF_{AFR})^2 - \frac{AF_{OBS}(1 - AF_{OBS})}{N_{OBS} - 1} - \frac{AF_{AFR}(1 - AF_{AFR})}{N_{AFR} - 1}}{AF_{OBS}(1 - AF_{AFR}) + AF_{AFR}(1 - AF_{OBS})}$$

$$F_{ST}^{OBS-EUR} = \frac{(AF_{OBS} - AF_{EUR})^2 - \frac{AF_{OBS}(1 - AF_{OBS})}{N_{OBS} - 1} - \frac{AF_{EUR}(1 - AF_{EUR})}{N_{EUR} - 1}}{AF_{OBS}(1 - AF_{EUR}) + AF_{EUR}(1 - AF_{OBS})}$$

Expectation-Maximization Algorithm

Binomial likelihood model for genotype as implemented in ADMIXTURE (originally from STRUCTURE):

$$\Pr(\text{1/1 for i at SNP j}) = \left[\sum_k q_{ik} f_{kj}\right]^2$$

$$\Pr(\text{1/2 for i at SNP j}) = 2\left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1 - f_{kj})\right]$$

$$\Pr(\text{2/2 for i at SNP j}) = \left[\sum_k q_{ik}(1 - f_{kj})\right]^2$$

where q is the ancestry proportion and f is the allele frequency. This gives the log likelihood of:

$$L(Q,F) = \sum_i \sum_j \left\{ g_{ij} \cdot ln\left[\sum_k q_{ik} f_{kj}\right] + (2 - g_{ij}) \cdot ln\left[\sum_k q_{ik}(1 - f_{kj})\right]\right\}$$

where g_{ij} is the observed number of alleles (0/1/2).

Modifying this, we can obtain the probability of having Allele 1 at SNP j, and the probability of not having Allele 1:

$$\Pr(\text{1 for i at SNP j}) = \left[\sum_k q_{ik} f_{kj}\right]^2 + \left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1 - f_{kj})\right]$$

$$\Pr(\text{2 for i at SNP j}) = \left[\sum_k q_{ik}(1 - f_{kj})\right]^2 + \left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1 - f_{kj})\right]$$

Therefor, the log likelihood for either having the allele or not can be written as:

$$L(Q,F) = \sum_i \sum_j \Bigg\{ h_{ij} \cdot ln\left(\left[\sum_k q_{ik} f_{kj}\right]^2 + \left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1 - f_{kj})\right]\right) + (1 - h_{ij}) \cdot ln\Bigg(\Bigg[\sum_k q_{ik}(1 - f_k$$

Where h_ij is whether the individual has the allele or not. This may or may not be applicable to overall allele frequencies, as this specifically targets individual observations instead of summary statistics.

We may be able to use the original Summix model as it is convex and can be solved for both ancestry proportions and unknown allele frequencies, if the other is fixed.

Block Relaxation Algorithm

This is Sequential Quadratic Programming. The log likelihoods stated above as well as the Summix model are both convex, and can solved with SQP.

By fixing the allele frequencies, we can solve for ancestry proportions. By fixing ancestry proportions, we can solve for allele frequencies. This process may be iterated until convergence.

Initialize the starting parameters with ancestry proportions of $\frac{1}{n}$. In our Afr/Eur/Unknown example each ancestry proportions would be $\frac{1}{3}$. In our 5 ancestry + unknown this would be $\frac{1}{6}$.

Initialize our starting unknown allele frequency. Either a uniform allele frequency across all SNPs or a random poisson/exponential distribution (reference alleles mirror exponential distribution).

Pick starting parameter to solve for. If solving for ancestry proportions, this can be done using SQP. If solving for unknown allele frequency, this can be done using:

$$AF_U = \frac{AF_{OBS} - \sum_{k=1}^{K} (\pi_k \cdot AF_{ref,k})}{\pi_u}$$

EM for Binomial mixture

We can treat the allele frequencies (which are usually calculated as allele count over allele number) as binomial mixture such that we have n alleles in a population of $N = 2 \cdot AC$ people. Therefor the probability of having n alleles in a two population example is:

$$P(n|N, \Theta) = \pi_1 \text{Binom}(n|N, \theta_1) + \pi_2 \text{Binom}(n|N, \theta_2)$$

or more generally

$$P(n|N, \Theta) = \sum_{k=1}^{K} \pi_k \text{Binom}(n|N, \theta_k)$$

where pi_k are the ancestry proportions, θ_k is the allele frequency of ancestral population, and n/N are the respective allele count/number.

Therefor the log-likelihood for our parameters is :

$$\mathcal{L}(\Theta|X, Z) = lnP(X, Z|\Theta)$$

Where Θ represents the set of π_k, θ_k, X represents the set of allele numbers/counts, and Z represents the set of SNPs from each respective ancestry. This model assumes that the SNPs are independent (LD) and that a given SNP comes from a respective ancestry. So the Auxiliary Function is:

$$Q(\Theta, \Theta_o) = E[ln\mathcal{L}(\Theta|X, Z)|X, \Theta_o]$$

and our expectation is:

$$E[lnP(X, Z, |\Theta)|X, \Theta_o] = \sum_{z_i=1}^{K} lnP(n_i, z_i|\Theta) \cdot P(z_i|n_i, \Theta_o)$$

where

$$P(z_i = k|n_i, \Theta_o) = \frac{P(z_i = k, n_i|\Theta_o)}{P(|\Theta_o)} = \frac{\pi_{k,o} \text{Binom}(n_i|N_i, \theta_{k,o})}{\sum_{l=1}^{K} \pi_{l,o} \text{Binom}(n_i|N_i, \theta_{l,o})}$$

We can then use the following expressions to update π and θ until convergence.

$$\pi_m = \frac{1}{S} \sum_{i=1}^{S} P(z_i = m|n_i, \Theta_o)$$

$$\theta_{m,S} = \frac{\sum_{i=1}^{S} n_i \cdot P\left(z_i = m | n_i, \Theta_o\right)}{\sum_{j=1}^{S} N_j \cdot P\left(z_j = m | n_j, \Theta_o\right)}$$

Application to Genetic Data

Previously, we had tried to use a binomial mixture which was modeled as a "jar full of coins" with unknown proportion of coins and probability of heads for different coins. This leads to the following mixture model:

$$P\left(n|N,\Theta\right)=\sum_{k=1}^{K}\pi_k \text{Binom}\left(n|N,\theta_k\right)$$

where the π_k and θ_k are estimated through an EM algorithm. We tried fixing the θ_k (allele frequencies) to try and solve for the π_k (ancestry proportions) but it didn't work. Instead, we adopt the model:

$$P\left(n|N,\Theta\right)=\text{Binom}\left(n\middle|N,\sum_{k=1}^{K}\pi_k\theta_k\right)$$

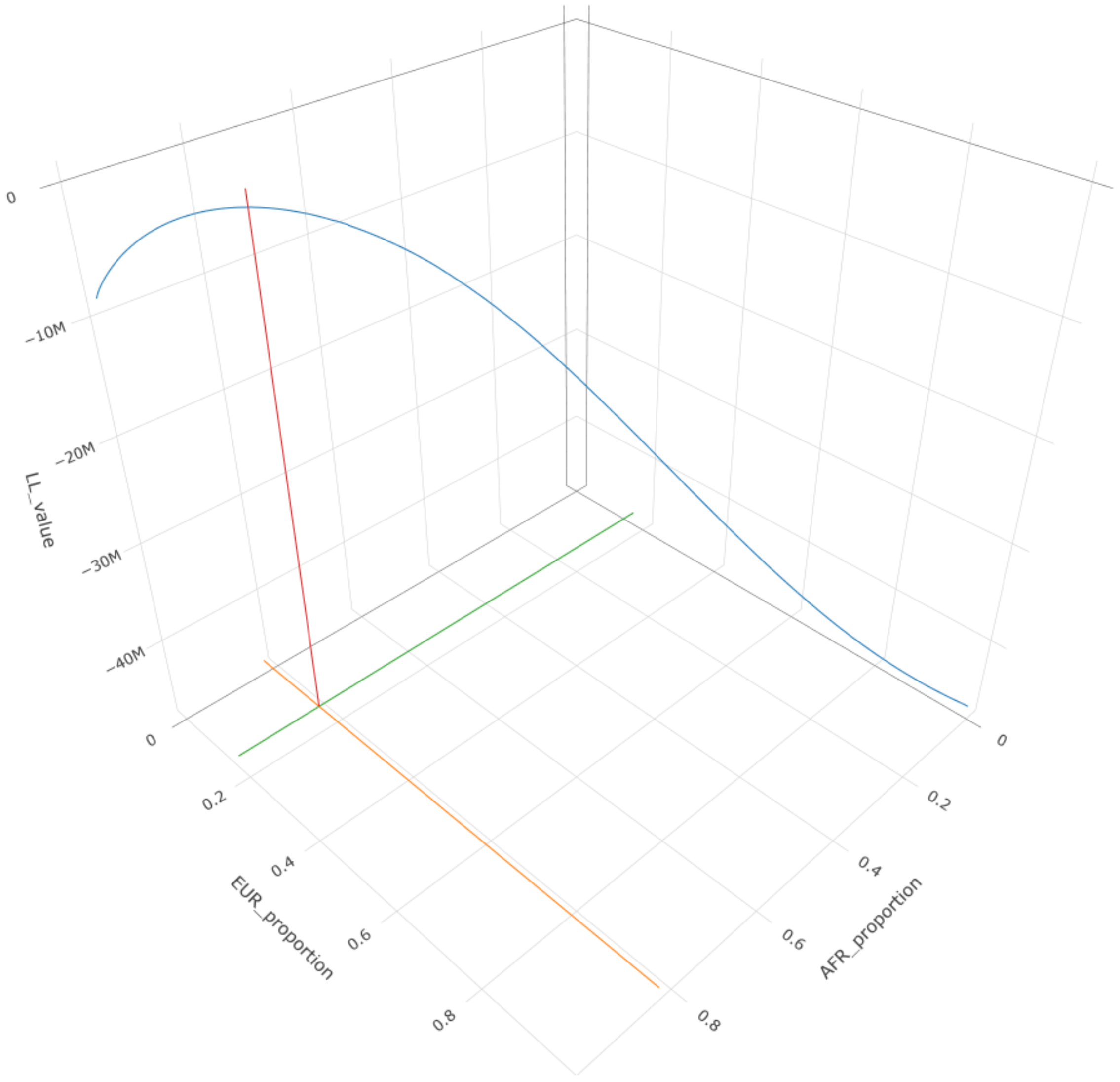
$$\ell(\Theta)=ln\mathcal{L}(\Theta|X)=\sum_{i=1}^{S}ln\left[\text{Binom}\left(n\middle|N,\sum_{k=1}^{K}\pi_k\theta_k\right)\right]$$

where
S is the set of SNPs
K are ancestries
π_k are ancestry proportions for k
n_i is the Allele Count for that SNP
N_i is the Allele Number for that SNP
θ_k is the Allele Frequency for that SNP

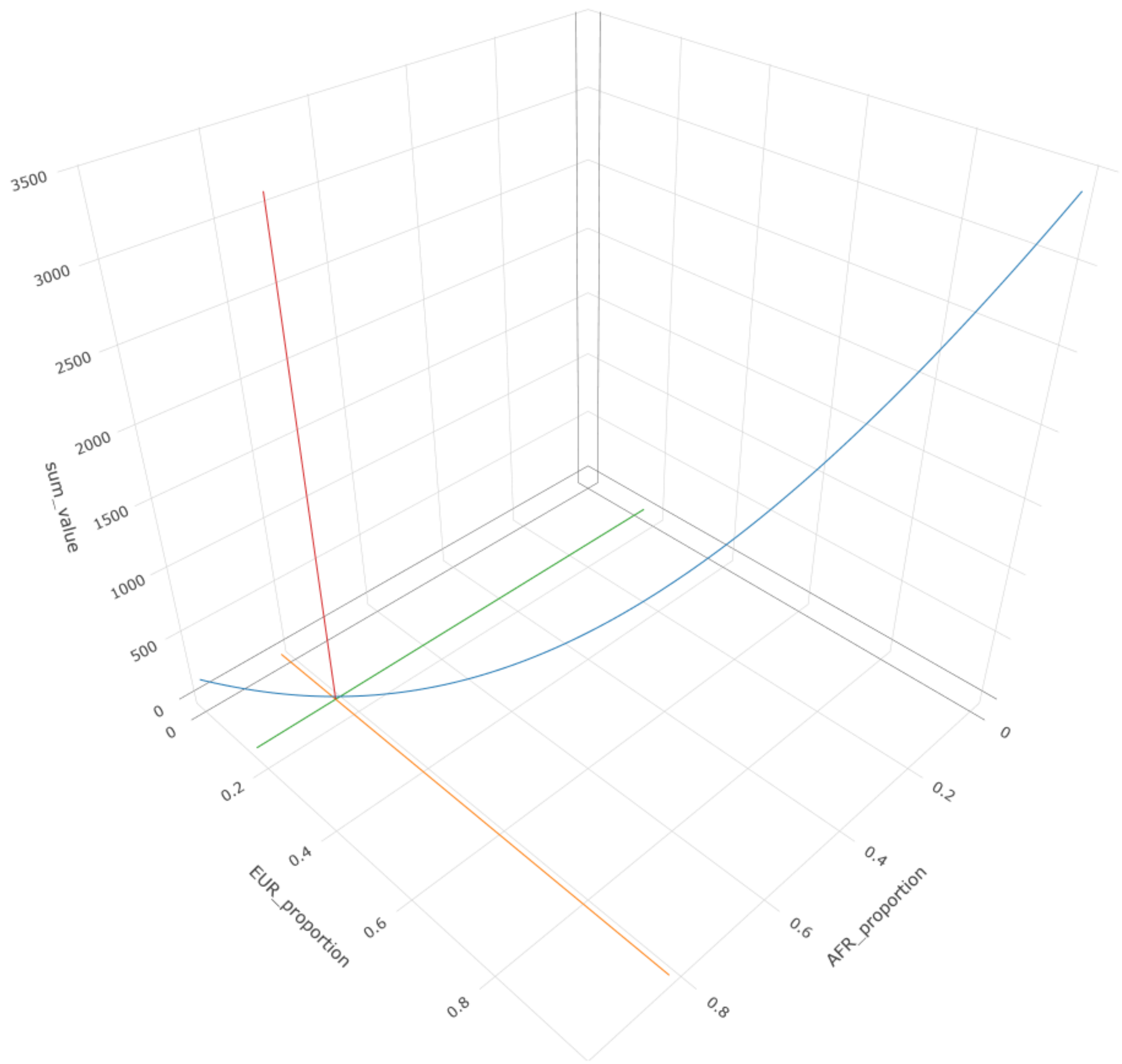
This returns the correct proportions as mentioned above, across all snps or random samples of snps. Written out for the gnomAD AFR example, this log-likelihood looks like:

$$\ell\left(\Theta|\text{Allele Count/Number, Set of Reference SNPS}\right)=ln\left[\text{Binom}\left(\text{Allele Count}\mid\text{Allele Number},\pi_{AFR}\text{AF}_{AFR}+\right.\right.$$

This gives us the following log-likelihood distribution, which has the shape we would expect for a log-likelihood of a binomial:



We can also compare this to the shape of our least-squares model, which is used in Summix and also solves to the same proportions:



Because our binomial mixture model has changed, I will need to re-evaluate whether the estimators I had been using the EM algorithm will work to solve for the maximum. The above plots were generated via grid-search with a precision of 0.001. In the meantime, the log-likelihood is continuous and twice differentiable, and if multiplied by -1, it is convex, so we can use it in SQP! The results from SQP using the log-likelihood as an objective function, and using the Summix package are below (across all 580K SNPs):

Log-Likelihood:
AFR: 0.8277273, **EUR**: 0.1722727, iterations: 31

Summix:
AFR: 0.828281, **EUR**: 0.171719. iterations: 15

Also, from preliminary testing, it seems that Summix is faster than using the LL as an objective function. More testing will need to be done, but I believe this is because in every iteration of SQP it is much faster to do matrix manipulation as in Summix, then to compute the binomial density across 580K snps as with LL. They are both still very fast, but I have noticed that Summix is 2-3 times faster.