Estimating sex-biased ancestry from uniparental markers in the presence of variable admixture proportions

We would like to quantify potential sex bias in admixture using uni-parental markers (Y chromosome- and mitochondrial haplotypes). Specifically, we consider admixed early-medieval individuals from England, for which we have estimated CNE-WBI admixture proportions. We would like to use the observed uniparental markers in these individuals, and the frequency of these haplogroups in the putative source populations, to compare the likelihood of drawing a male (or female) CNE haplogroup with the autosomal admixture proportion. In the following, we denote WBI as source 1, and CNE as source 2.

We first will derive a simple calculation of the likelihood for whether a given haplotype was inherited from one group or another group, based on the sample counts of that haplogroup in the given source population. We first assume that we know precisely the population haplogroup frequencies in both sources, denoted by for haplogroup in source (we use to denote the set of . The likelihood for an individual to carry haplogroup contributed from source is then simply:

But the true population haplogroup frequencies are not known. Rather, we have a finite number of individuals in both sources. Specifically, we use to denote the number of individuals with haplogroup in source (and to denote the entire matrix). The unobserved population frequency is now a latent variable following a conditional distribution, , which - following standard Bayesian computaton - takes the form of a beta distribution with parameters and , where :

which means that the full likelihood is a beta-binomial distribution with sample size 1, which can be written using beta functions:

We can now compute the probability that a given target haplogroup was contributed from Source using Bayes rule:

Here is a summary breakdown of the entire data into groups:

datj %>% dplyr::group\_by(Pop2) %>% dplyr::summarise(n = dplyr::n())

## # A tibble: 16 × 2  
## Pop2 n  
## <chr> <int>  
## 1 England\_BA 140  
## 2 England\_EMA 294  
## 3 England\_IA 304  
## 4 France\_MA 17  
## 5 Ireland\_BA 3  
## 6 Ireland\_MA 39  
## 7 NorthSea\_BA 37  
## 8 NorthSea\_IA 19  
## 9 NorthSea\_MA 209  
## 10 Scandinavia\_BA 4  
## 11 Scandinavia\_IA 7  
## 12 Scandinavia\_MA 180  
## 13 Scotland\_BA 28  
## 14 Scotland\_IA 25  
## 15 Wales\_BA 7  
## 16 Wales\_IA 7

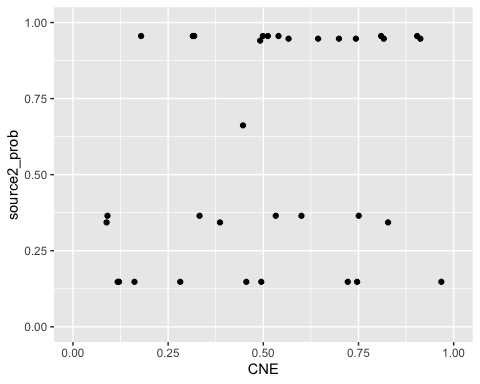
We now declare three groups of samples. First, a group of admixed individuals from England\_EMA, defined as individuals with CNE (continental North-European) admixture proportions between 2 and 98%. Then two source groups: one consisting of individuals from Britain before the study period plus early medieval individuals with less than 2% CNE admixture (Source 1), and one consisting of individuals from Britain with more than 98% CNE ancestry and those sampled from across the North Sea (Source 2).

# Y chromosomes

We can now get the Y haplogroup frequencies in the two sources, and compute the probability of that haplogroup to come from source 2:

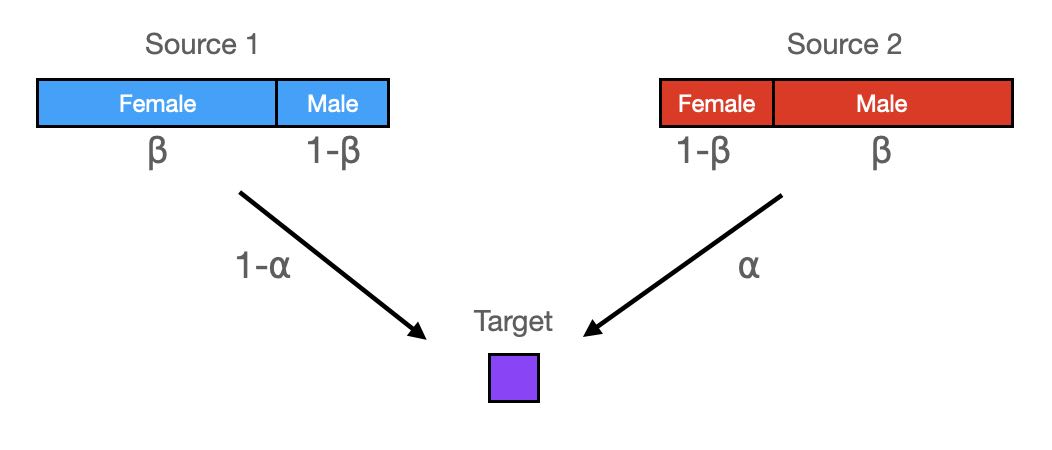
## # A tibble: 14 × 4  
## Y\_1 n1 n2 source2\_prob  
## <chr> <int> <int> <dbl>  
## 1 G 5 2 0.418  
## 2 I1 4 61 0.947  
## 3 I2a1a 3 3 0.590  
## 4 I2a1b 10 14 0.662  
## 5 J 1 4 0.782  
## 6 R1b 19 7 0.365  
## 7 R1b-L151 5 3 0.489  
## 8 R1b-L51 10 3 0.343  
## 9 R1b-P312 206 24 0.148  
## 10 R1b-U106 3 59 0.956  
## 11 E 0 3 0.852  
## 12 I 0 3 0.852  
## 13 N 0 1 0.742  
## 14 R1a 0 10 0.940

None of the probabilities is lower than 5% or higher than 95%, suggesting that none of these haplogroups by themself is strictly indicative of male ancestry from a specific source. Nevertheless, we can view these probabilities as a function of autosomal admixture proportion:



## Saving 5 x 4 in image

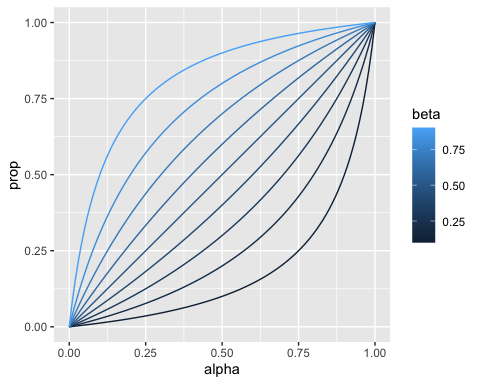
To understand this data, we set up a model with two parameters: The first parameter, , determines the amount of an individual’s CNE ancestry (Source 2). The second parameter, determines the proportion of males within Source 2, with being the proportion of males within Source 1. Here is a schematic:



A schematic of our sex-bias model and its two parameters and

This leads to a simple calculation of male Source2-related ancestry over total male ancestry, :

We can visualize this model as follows:



## Saving 5 x 4 in image

which shows that values of different from 0.5 can increase or decrease the male admixture proportion and deviate the curve from a model without sex-bias (a straight line).

In order to estimate from the data, we our above derived individual-based likelihood model of a haplogroup coming from Source2. The likelihood for a single individual with haplogroup H and an autosomal admixture estimate is

where we have used the individual-based posterior probability derived in equation .

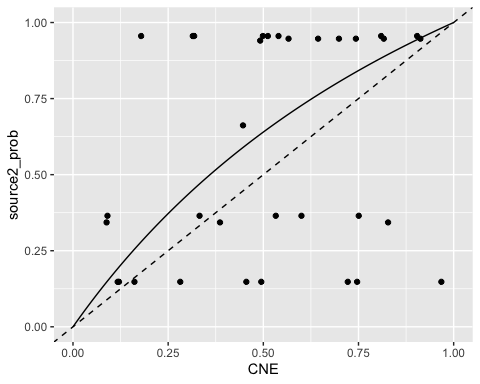
We can now compute the log-likelihood as a sum over all male admixed individuals

where index runs over all individuals with haplotype and autosomal admixture estimate .

We can estimate the maximum of that likelihood function under variation of numerically, and find

## $maximum  
## [1] 0.7242914  
##   
## $objective  
## [1] -20.51125

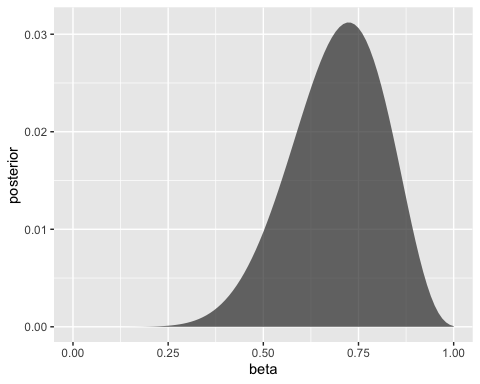
We can plot that model together with our frequency-based estimates:



## Saving 5 x 4 in image

So this would suggest a moderate male sex-bias in the CNE ancestry. But how significant is this? We can use our likelihood model to compute a posterior probability density over :

where the integral in the denominator can be approximated using a mesh over , and we have implicitly taken a uniform prior for over the interval from 0 to 1:



## Saving 5 x 4 in image

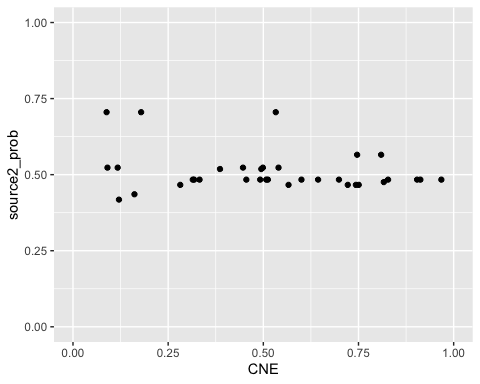
We can now see that the posterior distribution for comfortably includes the value , which means that the 95% credibility interval includes values below and above 0.5, suggesting models of male as well as female sex-bias being consistent with the data. We conclude that the Y haplotype data is not conclusive with respect to sex bias.

# Mitochondrial haplogroups

We can apply the same methodology to mitochondrial haplotypes, with the main difference being the interpretation of now describing the proportion of *female* Source2 ancestry over total female ancestry.

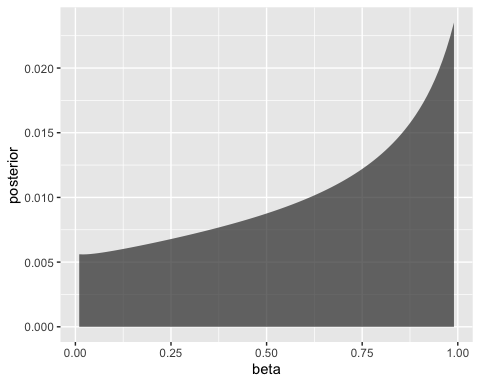
We get a similar table of posterior probabilitites for every MT haplogroup coming from source 2:

## # A tibble: 12 × 4  
## mtDNA\_simple n1 n2 source2\_prob  
## <chr> <int> <int> <dbl>  
## 1 H 206 134 0.484  
## 2 HV 5 9 0.705  
## 3 I 11 8 0.519  
## 4 J 52 47 0.565  
## 5 K 66 35 0.436  
## 6 N 2 1 0.489  
## 7 R 7 1 0.264  
## 8 T 54 41 0.523  
## 9 U 91 55 0.466  
## 10 V 18 11 0.476  
## 11 W 7 4 0.473  
## 12 X 11 5 0.418

which for our admixed individuals can be visualised again: 

## Saving 5 x 4 in image

but with all probabilities around 50%, the MT data conveys almost no information at all about source origin, since haplogroup frequencies are too similar in the two source populations.

Indeed, the posterior distribution of in this case yields: 

## Saving 5 x 4 in image

which is even broader than the already insignificant results from the Y chromosomal haplogroups, again showing that the data is consistent with models of both female and male sex-bias.

Note: The code used here can be found at <https://github.com/stschiff/AngloSaxon_Y-chrom_sex-bias>