Turkana demographic history inference report
Date: 09/04/2020
By: Tanya Phung (tanya.n.phung@gmail.com)
Github:
https://github.com/SexChrLab/Kenya_Selection_and_Demography/tree/master/Population_History

Methods

Obtaining neutral regions

We obtained neutral regions using the neutral explorer software (http://nre.cb.bscb.cornell.edu/nre/run.html) and converted to GRCh38 coordinates using liftOver.
Below are the parameters we used:

- Select Regions to Exclude:
    - Known Genes
    - Segmental Duplications
    - Gene Bounds
    - CNVs
    - Spliced ESTs
    - Self Chain
- Parameters:
    - Minimum region size (bp): 200
    - Distance to nearest gene: 0.4cM
    - Recombination rate (cM/Mb): 0.9
    - Genetic Map: HapMap
    - Human Diversity: YRI
    - Individuals: All
    - Mask: Strict
    - Min BG selection coefficient: 0.95
    - Chromosomes: 1-22
- Select Regions for which to Calculate % Overlap
    - Simple Repeats
    - Repeat Masker v3.27
    - 46 Way Conserved - Plac Mammal

Obtaining callable regions

We joint-genotyped emitting all sites. To obtain high quality sites, we employed filtering based on depth (DP) and the number of alleles that were genotyped. Specifically, we first calculated the mean depth across all sites (the field **DP** from the VCF file) (mean DP across all sites is

259.5). A site is considered high in quality if the site's DP is greater than 50% of the mean, which is 130. In addition, a site is considered high in quality if it is genotyped (the field **AN** from the VCF file) in at least 50% of the individuals. There are 110 individuals total or 220 alleles. Therefore, AN has to be at least 110.

Table 1. The number of sites after each filtering step

Criteria	Number of sites
All sites	2,863,768,058
Post DP filtering	2,676,216,026
Post AN filtering	2,664,308,333
Post filtering for neutral regions (NRE)	44,921,784

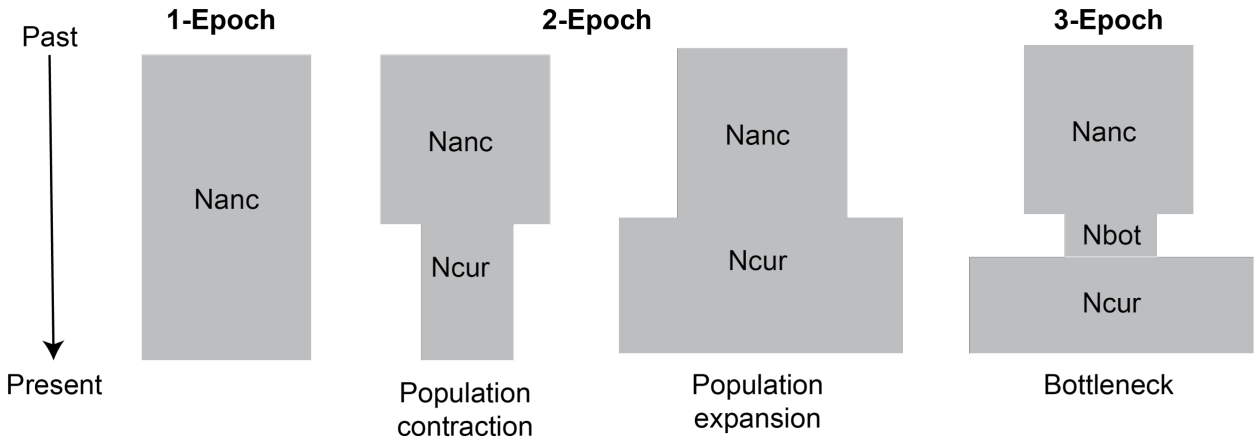
Table 2. The number of variants after each filtering step

Criteria	Number of variants
Biallelic SNPs	26,009,619
Post VQSR	16,855,346
Post HWE	16,854,369
Post 1000 genome pilot mask	16,712,825
Post filtering for neutral regions	356,517

Demographic inference with dadi

To infer demographic history, we used dadi with three models (**Figure 1**). The 1-Epoch model is a constant growth model. The 2-Epoch model includes an event: a population contraction or a population growth. The 3-Epoch model includes two events: for example, a population contraction followed by a population expansion (i.e. a bottleneck event).

Figure 1. Dadi was implemented for three models



In each model, 100 replications were run.

Results

Removing two individuals based on the principal component analysis and relatedness analysis

Principal component analysis showed that individuals A7 and A11 separate from the rest of the samples (**Figure 2**). In addition, we found that A7 and A11 showed parent-offspring relationship (**Figure 3**). Therefore, we removed these two individuals (A7 and A11) in our demographic history inference.

Figure 2. Principal component analysis on the autosomes showed separation of A7 and A11

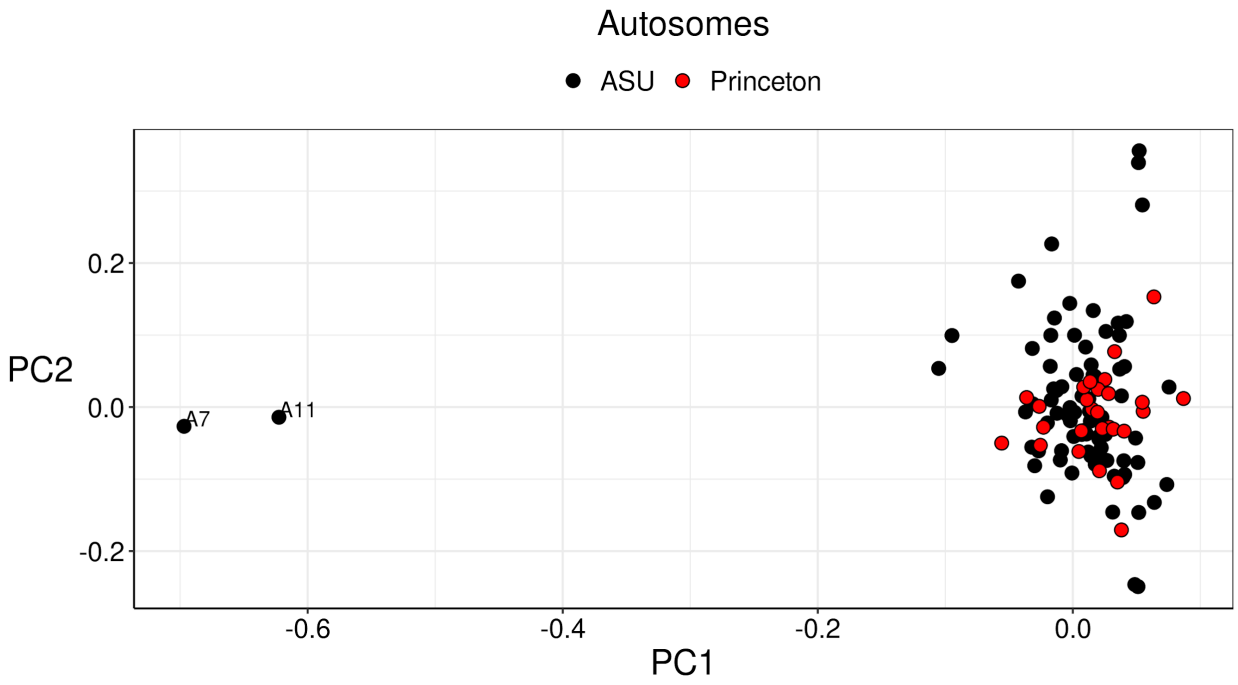
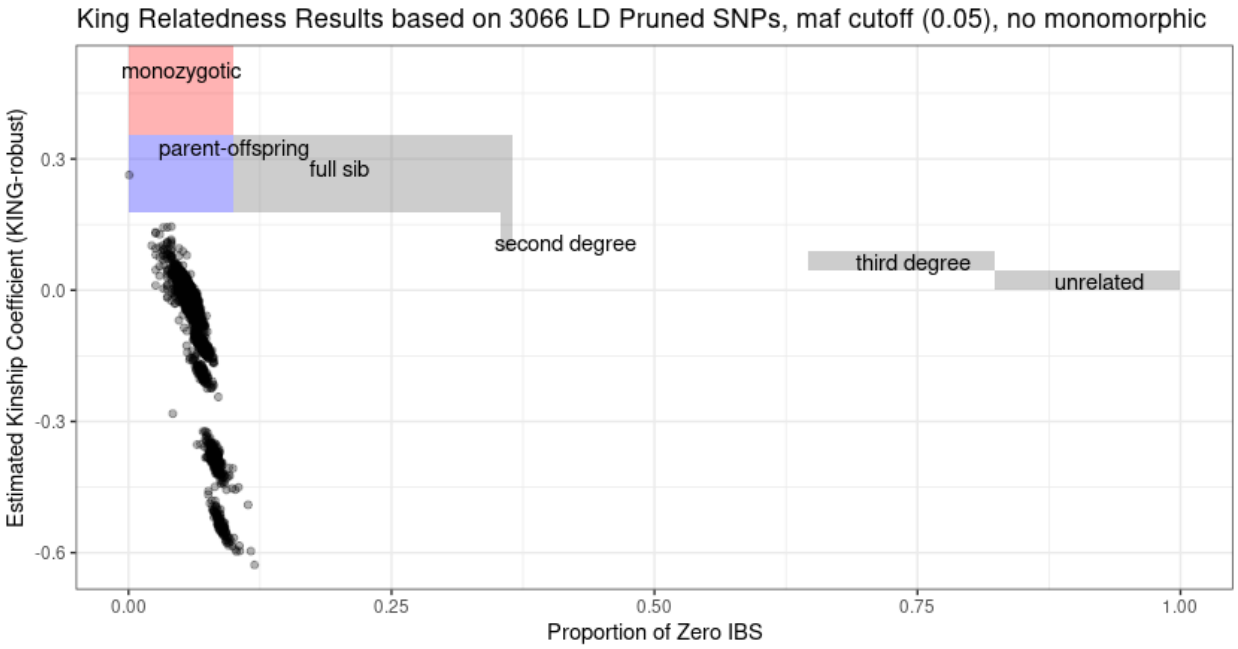


Figure 3. Relatedness analysis showed that A7 and A11 are parent-offspring



2-Epoch demographic model yields the best fits

The constant size model (1-Epoch) resulted in the worst likelihood (**Table 3**) and the site-frequency-spectrum generated using the parameters inferred from this model did not fit well to the empirical site-frequency-spectrum (**Figure 4**, yellow bars). Both the 2-Epoch model and the 3-Epoch model resulted in a similar likelihood and fit well to the empirical site-frequency-spectrum (**Table 3**, **Figure 4**). However, the 2-Epoch model is a more reasonable model because of the following reasons. First, even though the 2-Epoch model and the 3-Epoch model resulted in a similar likelihood, the 2-Epoch model is preferred because it contains fewer parameters and did not significantly improve the fit. Second, the parameters inferred from the 3-Epoch model is inconsistent with what we know about human populations. For example, the ancestral population size of 5 million seems tremendously large. In addition, the inferred time when the bottleneck occurs does not seem reasonable (with a conservative 20yr/generation time would be 9,115,260 years ago - well before human-chimp divergence). Interestingly, in the 3-Epoch model, the second event (going from the bottleneck size of around 15K to the current size of around 30K around 7000 generations ago) is consistent with the 2-Epoch model.

Table 3. Inferred parameters from three models

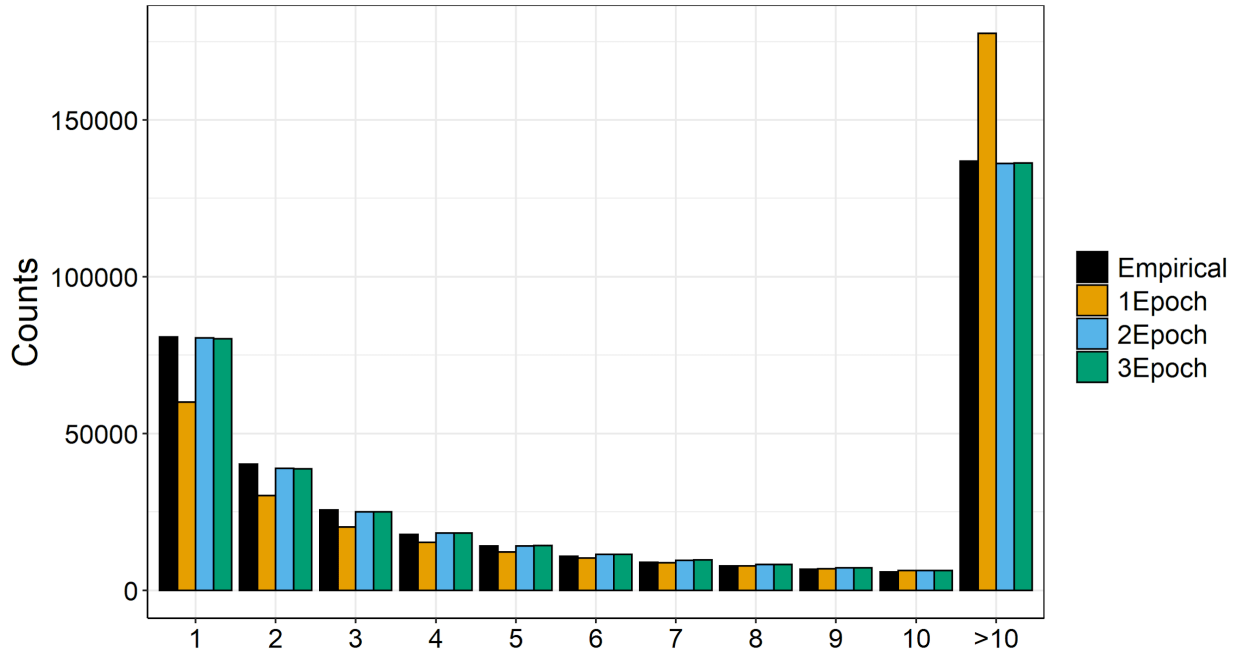
Model	N_anc	N_curr/N_anc	Time (in the past when event occurs)	Model_LL	Empirical_LL
1-Epoch	22,166	N/A	N/A	-12,580	-494
2-Epoch	15,436	2	6,978 generations ago	-999	-494

3-Epoch model parameters	Values
N_ancestral	5,444,112
N_bottleneck	15,422
N_current	30,842
Time when bottleneck occurs	455,763 generations ago

Time when bottleneck ends	7,105 generations ago
Model_LL	-1000.41

To visualize the fit of the different models to the empirical data, we used the inferred parameters to generate the site-frequency-spectrum. We observed that the data generated using the parameters from the 2-Epoch model resulted in a similar SFS to the empirical data (**Figure 4**).

Figure 4. Both the 2-Epoch model and the 3-Epoch model visually fit the empirical data



Conclusions

The 2-Epoch model yields a good fit to the empirical SFS. Since the 2-Epoch model is “simpler” (contains fewer parameters than the 3-Epoch model), the 2-Epoch model should probably work well.