Inference of population demographic history captures differing evolutionary signals based on the number of individuals in the dataset

Jonathan Mah[1] and Kirk Lohmueller[2, 3]

1. Bioinformatics Interdepartmental Program, University of California, Los Angeles
2. Department of Ecology and Evolutionary Biology, University of California, Los Angeles
3. Department of Human Genetics, University of California, Los Angeles

Background

Accurate estimation of population demographic history is central to population genetics[1] yet remains challenging due to the sensitivity of inference methods to sampling and model choice. The site-frequency spectrum (SFS) of synonymous variants, a widely used summary statistic of genetic variation, is particularly sensitive to demographic processes[2, 3, 4], but studies have shown that qualitative results from demographic inference can depend strongly on the number of individuals in the dataset[5, 6].

Specifically, SFS-based analysis[5, 6] of empirical human data suggests that smaller datasets yield evidence of a population contraction, whereas larger datasets yield evidence of a population expansion.

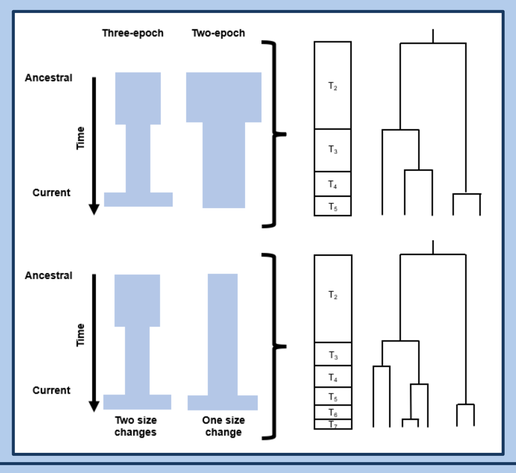


Figure 1. Schematic describing how the distribution of coalescent events and branch lengths relates to the number of individuals analyzed and the demographic scenario.

We hypothesize that the number of individuals analyzed in a dataset determines the distribution of coalescent tree properties across epochs of demographic history, resulting in differing evolutionary signals being captured. **Here, we assess how demographic inference, coalescent tree properties, and summary statistics of the SFS change based on the number of individuals analyzed. Our results highlight that demographic inference depends critically on the number of individuals analyzed and suggest that analyzing datasets at multiple sample sizes can reveal complementary aspects of population history.**

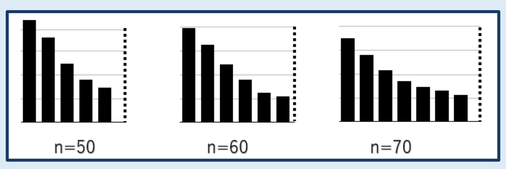
Data

We analyzed two simulated sets of 800 diploid individuals and one set of publicly available whole-genome sequencing data consisting of 305 human individuals of European descent.

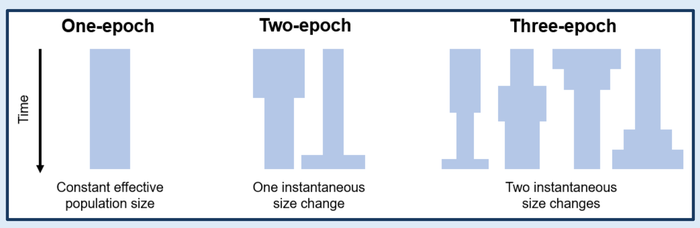
Our simulations were generated under the coalescent framework, MSPrime[7], and the forward-in-time framework, SLiM[8]. Our empirical data was downloaded via 1000Genomes[9].

Methods

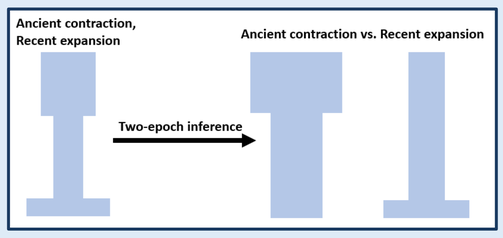
1. For each of our simulated and empirical datasets, we compute a site-frequency spectrum (SFS), i.e., a summary statistic of genetic variation. We analyze subsets of our datasets consisting of differing numbers of individuals, for up to 800 simulated or 300 empirical individuals.



2. Given an SFS, we infer the population demographic history by fitting three different demographic models to the data and assessing which model yields the best-fit parameters describing the timing and magnitude of population size change.



3. Lastly, we compare how demographic model inference and summary statistics of the SFS change depending on the number of individuals analyzed. We are especially interested in how two-epoch model inferences potentially capture differing evolutionary signals.



Results

Demographic inference shows an ancient population contraction when analyzing simulated data consisting of few individuals, and a recent population expansion when analyzing data consisting of numerous individuals.

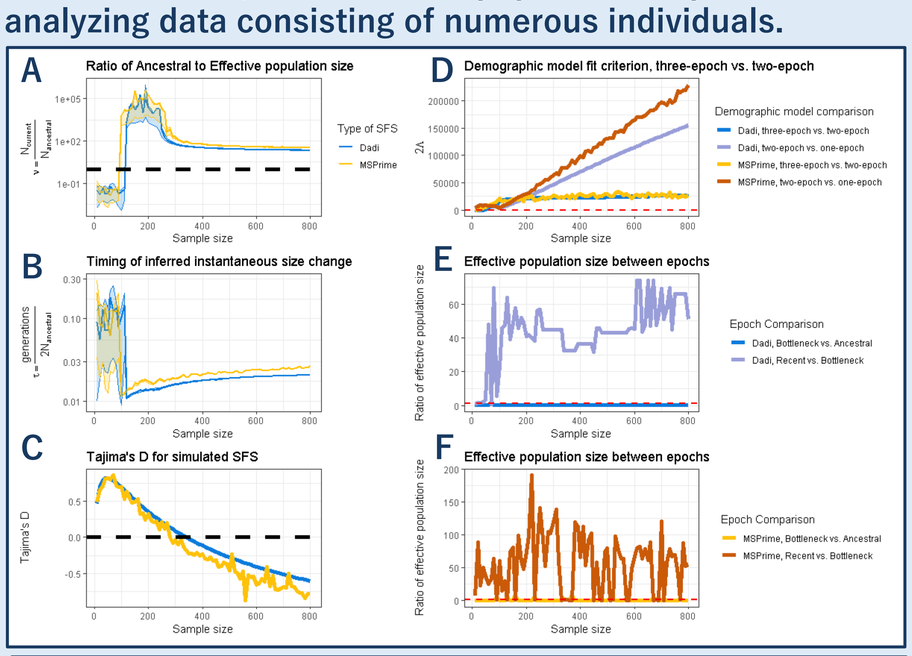


Figure 2. Inferred demographic parameters and SFS summary statistics for **simulated data**. The x-axis indicates the number of individuals analyzed. The y-axis indicates the **A**) size-change parameter, **B**) timing of inferred size change, **C**) Tajima's D of the simulated SFS, **D**) model comparison statistic 2Λ, and **E-F**) the size change between epochs in a three-epoch model.

Similarly to our analysis of simulated data, when analyzing empirical data, we inferred an ancient contraction at small sample sizes (n=10) and a recent expansion at larger sample sizes (n>=20).

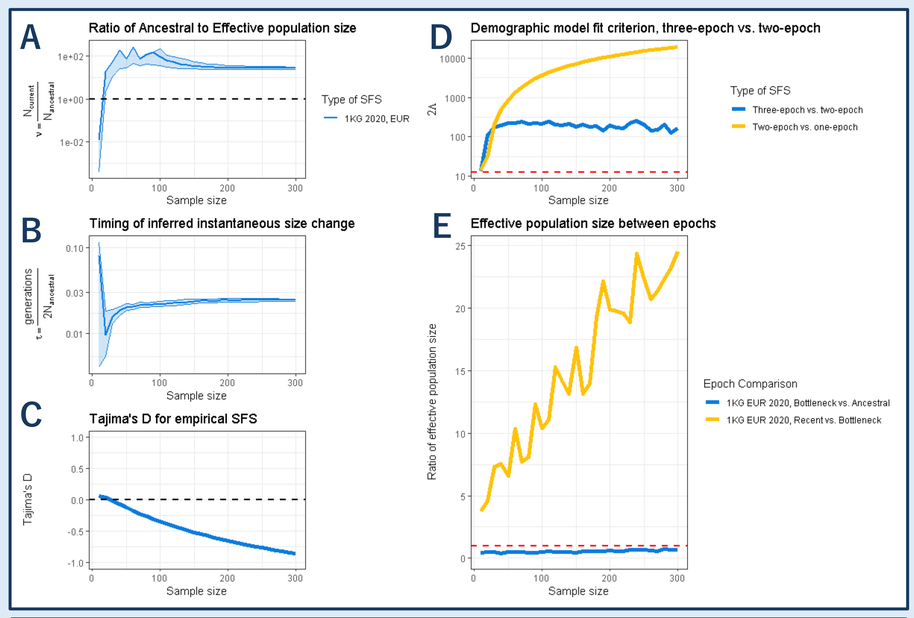


Figure 3. Inferred demographic parameters and SFS summary statistics for **empirical data**. The x-axis indicates the number of individuals analyzed. The y-axis indicates the **A**) size-change parameter, **B**) timing of inferred size change, **C**) Tajima's D of the empirical SFS, **D**) model comparison statistic 2Λ, and **E**) the size-change between epochs in a three-epoch model.

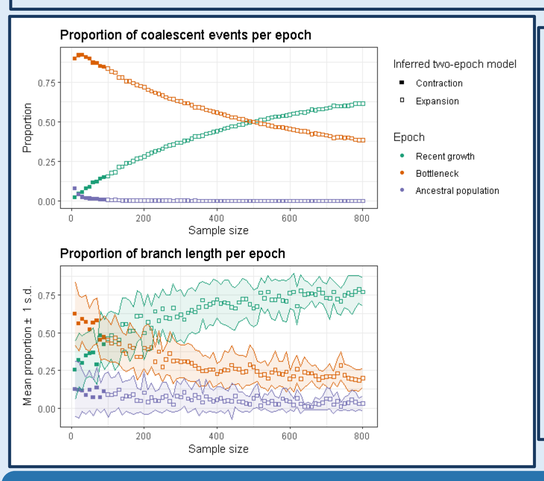


Figure 4. Coalescent summary statistics from SFSs simulated using MSPrime. The x-axis indicates the number of individuals, while the y-axis indicates the **A**) proportion of coalescent events per epoch, and **B**) proportion of branch lengths per epoch.

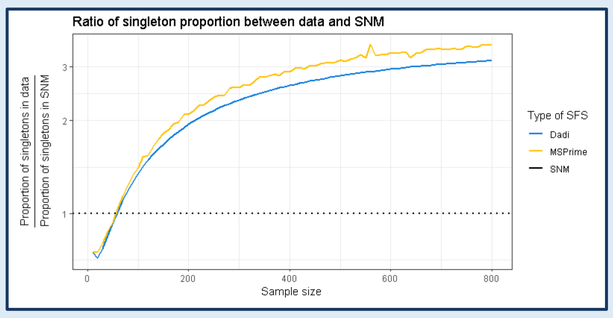


Figure 5. Ratio of the proportion of the SFS comprised of singletons between simulated data and the standard neutral model.

Conclusions

Conclusions:

- In both simulated and empirical data, when evaluating data consisting of few individuals, we infer an ancient population contraction. Conversely, when evaluating data consisting of numerous individuals, we infer a recent population expansion.
- Coalescent summary statistics reveal that when the highest mean proportion of branch lengths falls in the ancient epoch, we infer a contraction. Conversely, when the highest mean proportion of branch lengths falls in the recent epoch, we instead infer an expansion.
- Summary statistics of the SFS also suggest that, relative to a standard neutral model, smaller sample size SFSs more closely resemble population size change contractions and larger sample size SFSs resemble population expansions (e.g., higher proportion of rare variants or singletons).

References

1. Beichman 2018, Annual Reviews. 2) Wakeley and Hey 1997, Genetics. 3) Nielsen 2000, Genetics. 4) Gutenkunst et al. 2009, PLoS Genetics. 5) Kryukov et al. 2008, PNAS. 6) Gazave et al. 2014, PNAS. 7) Baumdicker et a. 2022, Genetics. 8) Haller and Messer 2023, The American Naturalist. 9) The 1000 Genomes Project Consortium 2015, Nature.

jonmah@g.ucla.edu

This work was supported by the Systems in Integrated Biology training grant (NIGMS T32 5T32GM008185), the Dissertation Year Award, and through the UCLA Graduate Programs in Biosciences.