A likelihood-based framework for demographic inference from genealogical trees

Caoqi Fan[1,2], Nicholas Mancuso[1,3,4], Charleston W. K. Chiang[1,2]

[1]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, KSOM, USC. [2]Department of Quantitative and Computational Biology, USC. [3]Biostatistics Division, Department of Population and Public Health Sciences, KSOM, USC. [4]Norris Comprehensive Cancer Center, KSOM, USC.

Email: caoqifan@usc.edu, nicholas.mancuso@med.usc.edu, charleston.chiang@med.usc.edu

Introduction

Accurately inferring the population history of humans not only has archaeological and historical significance, it also helps to properly account for population structure in association studies and reduce false positives in inferences about natural selection[1]. Traditional demographic inference methods ignore the vast majority of the genealogical information by trimming trees or using summarizing statistics. Here we introduce a genealogical likelihood (gLike) framework to compute the full likelihood of a given genealogical tree generated by any hypothesized demography, and then showcase its application in demography inference. Notably, gLike provides a unified methodology to infer all involved parameters (population sizes, migration rates, admixture times and proportions) in complex demographies.

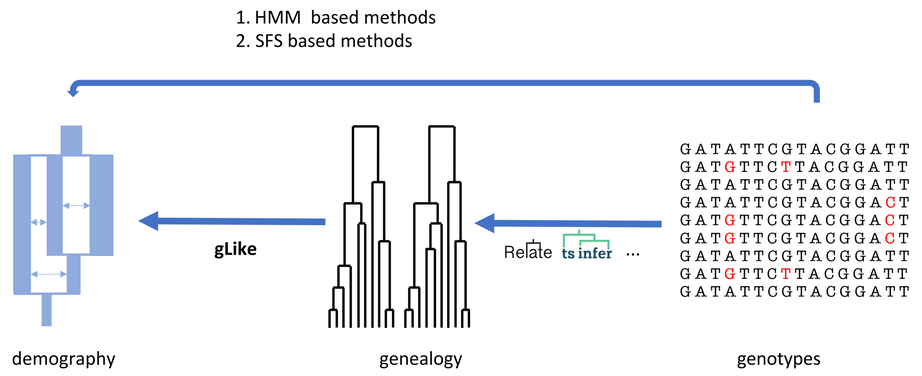


Figure 1. Demographic inference from the ARG. Unlike existing methods that infer demography from genotypes, gLike focuses on the parity between the flexible ARG structure and a general demography with various events – migrations, splits and admixtures.

Demography as Markov process

The demographic history of a group of populations emcompassing a variety of events can be generally described as a time-dependent Markov process of lineages. Specifically, each constant period of migration and coalescence is represented by a continuous Markov process such that the effective transition is the matrix exponential of the instantaneous rates, while each mass migration event has a discrete transition matrix.

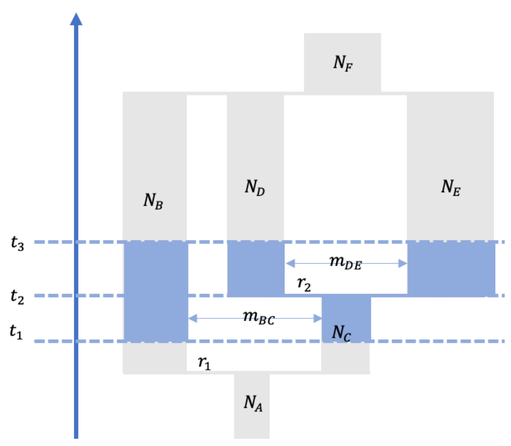
When two lineages are considered, the rule to derive instantaneous transition rates is given by (and zero if not mentioned below):

$$Q_{ii\to ii} = -n_i - 2\sum_{j\neq i} m_{ij} \qquad Q_{ii\to ij} = m_{ij} \qquad Q_{ij\to ik} = m_{jk}$$

$$Q_{ij\to ii} = m_{ji} \qquad Q_{ij\to kj} = m_{ik}$$

$$Q_{ij\to ij} = -\sum_{k\neq i} m_{ik} - \sum_{k\neq j} m_{jk} \qquad Q_{ij\to jj} = m_{ij}$$

Where i, j and k are different populations. The parameter *n* is inverse population sizes (that is, coalescent rates); *m*, migration rates; *r*, admixture proportions; *t*, times in generations ago. This lineage-pair transition matrix is of great importance because any higher-order transition matrix can be derived from it, assuming independence of non-coalescence between different lineage pairs. We have mathematically proved that this assumption always hold as the limit when the time interval is short.



$$\exp\left((t_2-t_1)\begin{array}{c}\\ BB\\ BC\\ CB\\ CC\end{array}\begin{array}{c}\begin{array}{cccc}BB & BC & CB & CC\end{array}\\ \begin{bmatrix}-n_B-2m_{BC} & m_{BC} & m_{BC} & 0\\ m_{BC} & -2m_{BC} & 0 & m_{BC}\\ m_{BC} & 0 & -2m_{BC} & m_{BC}\\ 0 & m_{BC} & m_{BC} & -n_C-2m_{BC}\end{bmatrix}\end{array}\right)$$

$$\begin{array}{c}\\ BB\\ BC\\ CB\\ CC\end{array}\begin{array}{c}\begin{array}{ccccccccc}BB & BD & BE & DB & DD & DE & EB & ED & EE\end{array}\\ \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & r_2 & 1-r_2 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & r_2 & 0 & 0 & 1-r_2 & 0 & 0\\ 0 & 0 & 0 & 0 & r_2^2 & r_2(1-r_2) & 0 & r_2(1-r_2) & (1-r_2)^2\end{bmatrix}\end{array}$$

$$\cdot \exp\big((t_3-t_2)[\cdots]\big)$$

Figure 2. Example showing a slice of demography represented by a Markov transition of lineage pair. A 9x9 matrix describing the continuous migration between *t2* and *t3* is ellipsized due to limited space.

Full likelihood of genealogical trees

The migration trajectory of lineages to explain a given genealogical tree structure is generally not unique. We use a directed graph to represent all possible states of lineages throughout history, while each edge is associated with a conditional probability computed from the Markov transition matrix. Special algorithm is developed to accelerate bulk computation of large number of states. The full likelihood of the tree is achieved by a breadth-first search from the present time.

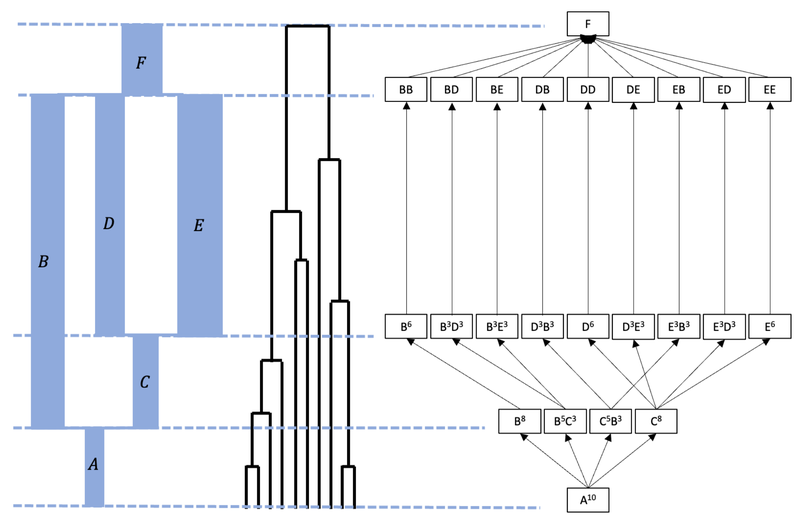


Figure 3. Example showing the graph of states of a genealogical tree with 10 haplotypes under a three-way admixture demography without migration.

Demography Inference on Simulated Data

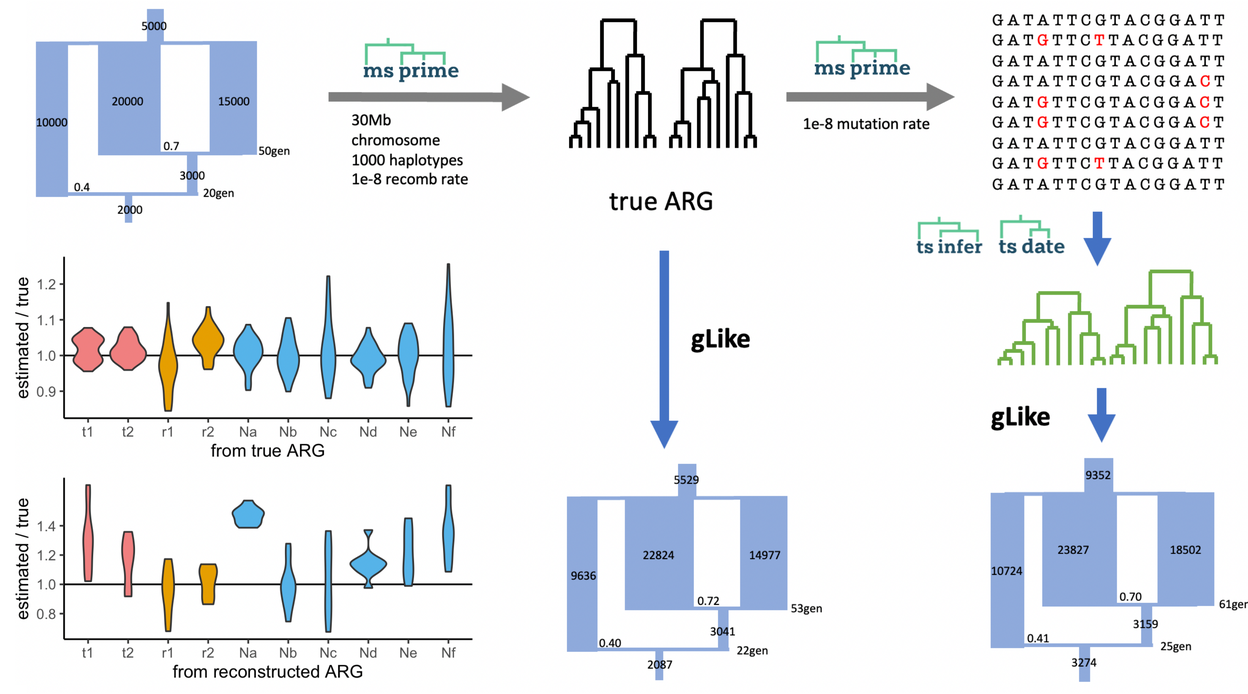


Figure 4. Performance of demography inference on the true and reconstructed ARGs. ARGs and genotypes are simulated by msprime[2] under a three-way admixture model with 10 parameters including population sizes, admixture times and proportions. A simple hill-climbing optimization was applied to maximize the genealogical likelihood function, demography was inferred based on the true ARG and the ARG reconstructed by tsinfer[3]+tsdate[4]. It took around 6 hours for all parameters to reach stable inferred values. Violin plots show the inference results in 30 replicate experiments. The deviation of estimated t1 and Na from true values are likely due to overestimation of very recent branch lengths common to existing ARG reconstruction tools.

References

1. Mathieson Iain et al. Differential confounding of rare and common variants in spatially structured populations. Nature Genetics, 44:243–246, 2012.
2. Baumdicker, et al. Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220.3 (2022): iyab229.
3. Kelleher, et al. Inferring whole-genome histories in large population datasets. Nature genetics 51.9 (2019): 1330-1338.
4. Wohns, et al. A unified genealogy of modern and ancient genomes. Science 375.6583 (2022): eabi8264.

Software Release

The gLike algorithm is implemented as a Python package with C extension modules and will be released on github along with the manuscript.