Software tools leveraging **site frequency spectra** based on the infinite sites model are widely used due to their simpler algorithms for evaluating sequence data, as compared to **finite sites mutation models**. Despite this, the relative impact remains underexplored. Species with **substantial population sizes**, such as *Anopheles gambiae*, display over 21% variable sites in their genome exceeding two states. We simulated DNA sequences using a finite sites model and analyzed these data with a site-frequency-spectra method (**momi2**) and a Bayesian estimator (**migrate-n**). The results suggest that we may need to potentially reevaluate conclusions for species with large effective populations.

Site frequency spectra based estimators vs. estimators using finite mutation models

Peter Beerli, Marzieh (Tara) Khodaei
Department of Scientific Computing, Florida State University, Tallahassee FL, USA

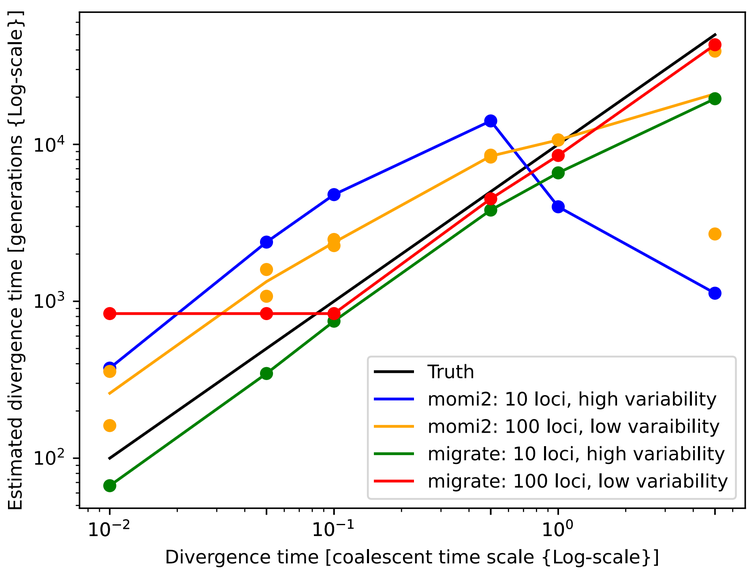
Introduction

Studying evolutionary processes depends heavily on the estimation of parameters of complex models that use DNA data as input. In recent years generation of genomic data has become relatively cheap and quick. These inferences usually depend on tree structures, such as species phylogenies or random coalescent trees of individual markers/chromosomes. Complex models, including those that estimate population sizes, immigration rates, and divergences times, often are declared so complicated that we cannot calculate the likelihood (the probability of the data given the tree or parameters). Alternatives to the full likelihood approaches, such as those based on summary statistics, became very common in the field of population genomics. In particular, the site frequency spectrum (a histogram of how many times a particular allele was found in the sample) has gained much momentum. Most applications of the site frequency spectrum assume that there are only two allele states, a feature shared with the infinite sites model where only one mutation can happen on a genealogy dividing the data into two alleles. The most common data in this context are single nucleotide polymorphisms. The number of variable sites (mutations) in a population depends on the number of individuals and the mutation rate. Even with low mutation rates but large sizes, the chance of seeing more than two alleles per site becomes sizeable. For example, the natural population of the malaria mosquito *Anopheles gambiae* has about 21% variables sites (Anopheles gambiae 1000 Genome project 2017). To adapt to standard site frequency methods, one must ignore either the third or fourth allele or discard the site.

Methods

We simulate 10 and 100 10000bp-loci for 20 individuals using a finite mutation model (Jukes and Cantor 1969) for two populations. This data was then analyzed using the true evolutionary model (Fig.1), estimating the time of the population split with the software *momi2*(Kamm et al. 2020) and *migrate-n* (Beerli et al. 2022)

Results



Conclusion

With high variability datasets, such as sequences of *Anopheles gambiae* (Anopheles gambiae 1000 Genome project 2017), methods that are based on site frequency spectra (assuming that the mutation per site happens only once) will deliver biased estimates of key parameters of the evolutionary model. We are working on a more detailed analysis and guidance. For example, for humans, the number of tri-allelic SNPs seems substantial (Phillips et al. 2020) but most likely will not change interpretation. Estimated evolutionary parameters from data of species with large population sizes will need careful examination.

Supplementary Material

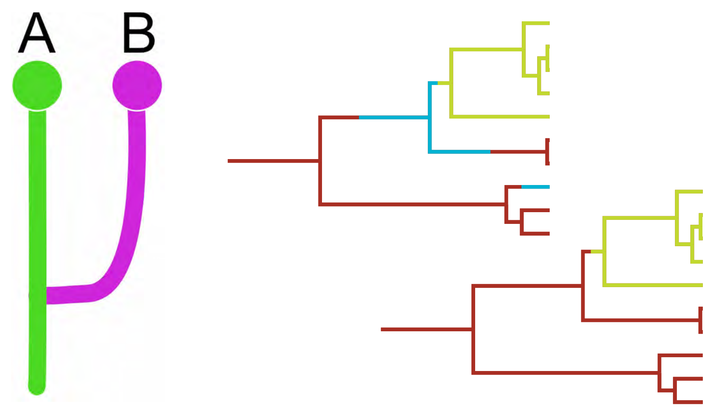


Figure 1: Simulation scenario, the genetrees (not shown) are embedded in this population tree. The left panel shows mutations on a genetree, (top) finite mutation model, (bottom) infinite mutation model.

Table 1: Frequency and number of variable sites is used in the analysis (this example is only for the simulations with a divergence time of 1.0 coalescent units and different population sizes from overall $\Theta = 0.002$ to $\Theta = 0.2$ ($Theta = 4N_e\mu$, mutation rate μ is per site). The table shows the total number of variable sites over 100,000 bp. Changing the divergence time does affect the number of SNPs, for example, of a $\Theta = 0.2$ the number. The rows with $\star$ are those used in the Results figure!

Population Size	Freq. Variable Sites	SNPs	Tri-Allelic	Tetra-Allelic
0.002	0.00888	888	0	0
0.005	0.02124	2124	4	0
0.010	0.03915	3915	36	0
$\star$ 0.020	0.07900	7900	124	1
0.050	0.19273	19273	847	15
0.100	0.36277	36277	3477	116
0.150	0.46862	46862	5872	291
$\star$ 0.200	0.60330	60330	10596	797

Supported by US National Science Foundation grant DBI 2019989.

Peter Beerli, Haleh Ashki, Somayeh Mashayekhi, and Michal Palczewski. Population divergence time estimation using individual lineage label switching. *G3 Genes|Genomes|Genetics*, 12(4), 02 2022. **Consortium Anopheles gambiae 1000 Genomes**. Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature, 552(7683):96-100, Dec 2017. **Thomas Jukes and Charles Cantor**. Chapter 24 - Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969. **Jack Kamm, Jonathan Terhorst, Richard Durbin, and Yun S. Song.** Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, 115(531):1472–1487, 2020. **C. Phillips, J. Amigo, A. O. Tillmar, M. A. Peck, M. de la Puente, J. Ruiz-Ramírez, F. Bittner, Š. Idrizbegović, Y. Wang, T. J. Parsons, and M. V. Lareu.** A compilation of tri-allelic snps from 1000 genomes and use of the most polymorphic loci for a large-scale human identification panel. *Forensic Science International: Genetics*, 46:1–12, 2023/06/08 2020.