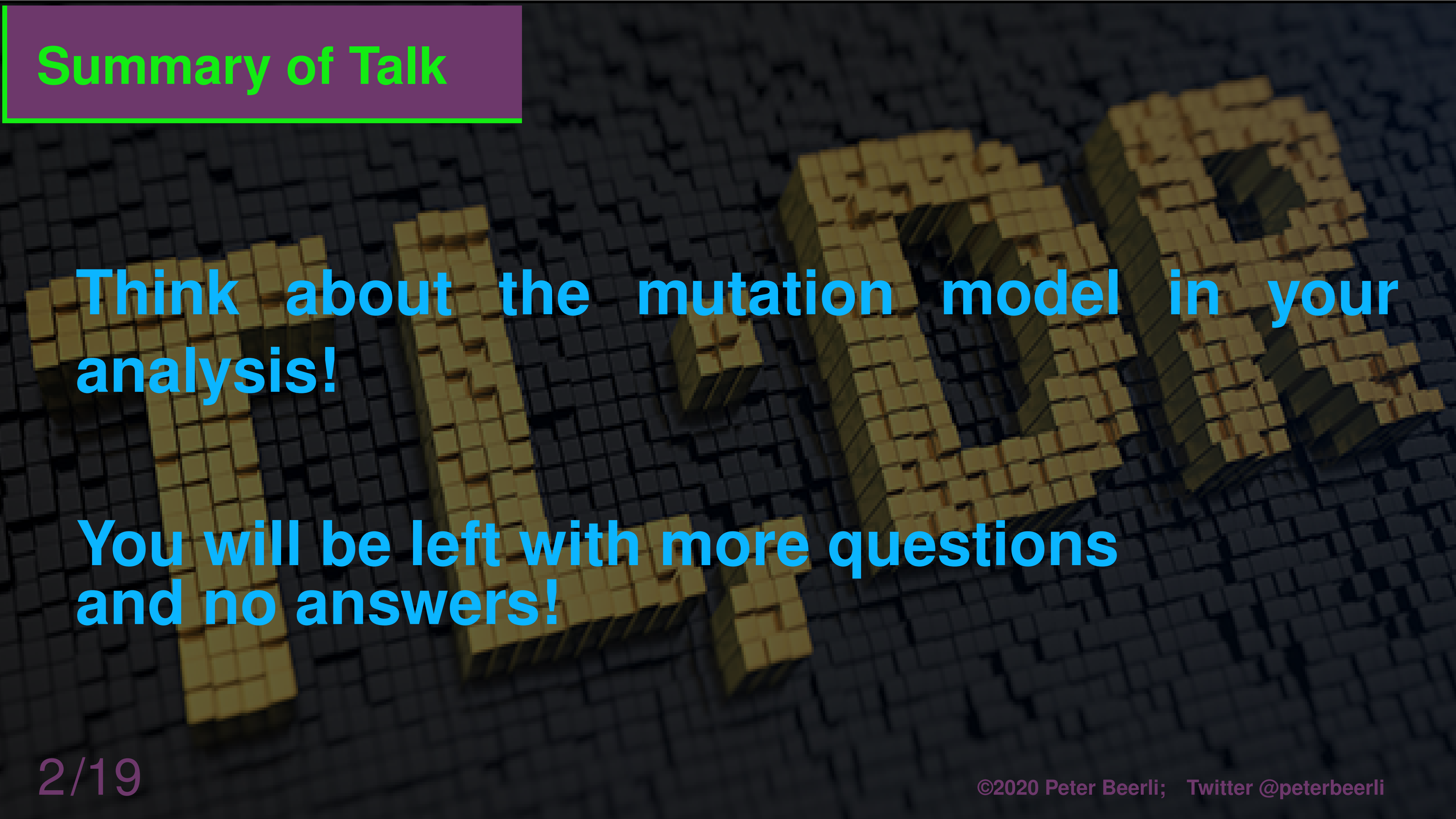
Population genomics using finite mutation models

Peter Beerli Scientific Computing, Florida State University Twitter:@peterbeerli

Summary of Talk



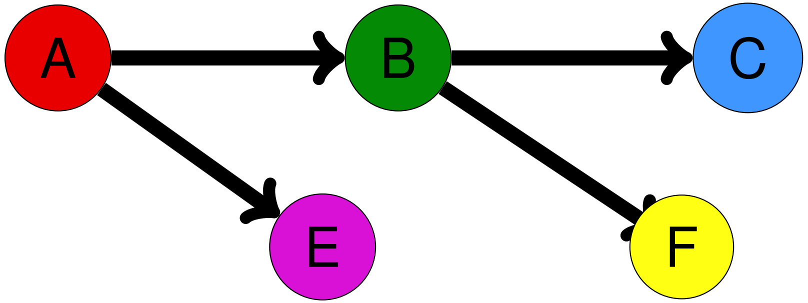
Think about the mutation model in your analysis!

You will be left with more questions and no answers!

Mutation model history

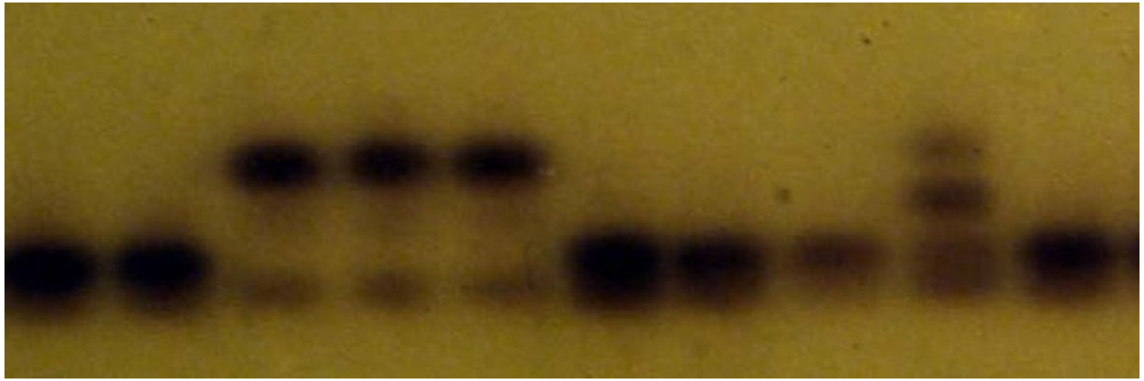
Early → Mid → Late

Early: Mutation models became a ”thing” when electrophoretic allozyme markers were used to look at variability in natural populations, starting with papers such by Lewontin and Hubby (∼1964). The **infinite allele model** became famous



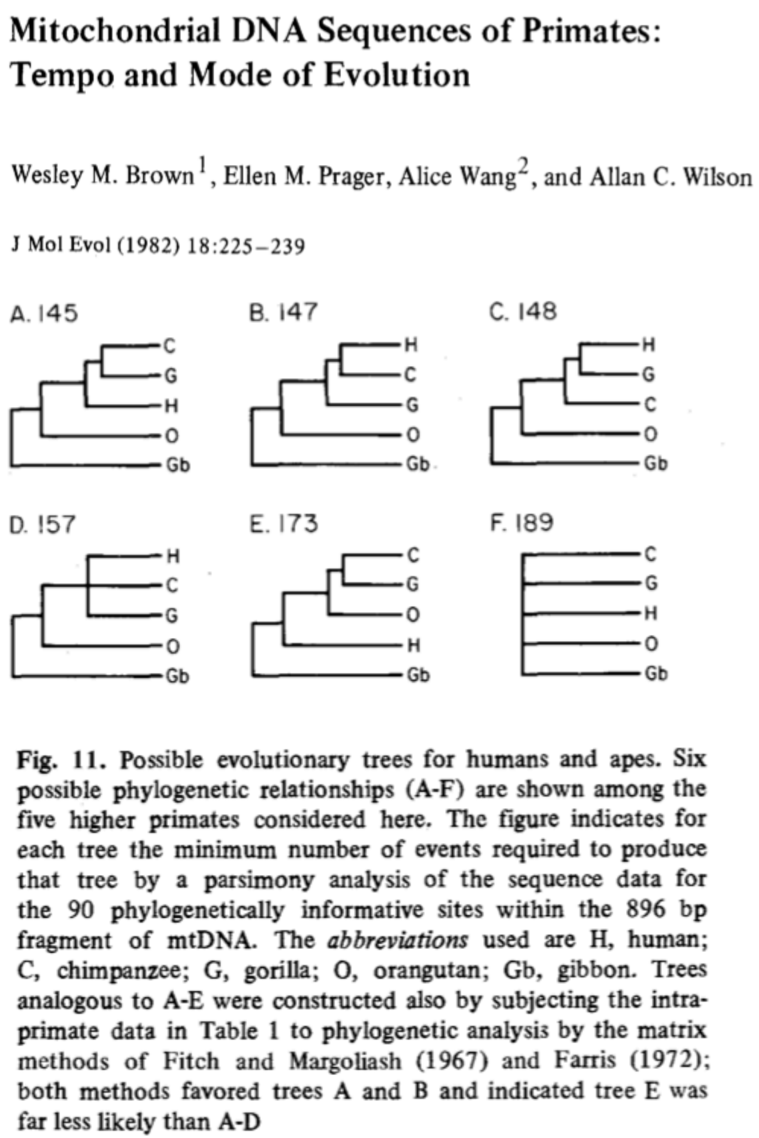
Mutation model history

Mid (popgen): These were the haydays of enzyme electrophoresis and many population genetic studies based on allozymes were generated, using the infinite allele model (Kimura and Crow, 1964) or a variant of the ladder model (Ohta and Kimura, 1973) that then became the standard for microsatellite data.



Mutation model history

Mid (phylogeny): Researchers started to. sequence DNA, such as 5S rRNA and mtDNA, and were able to work on phylogenetic trees of species; I guess that these were also the haydays for cladistic methods for DNA datasets because computers were ill equipped to run likelihood phylogenetic analyses using **finite mutation models.** Models that explicitly model the transition between nucleotides over the course of time, many variants created a considerable alphabet soup of models: JC69, K2P, F81, F84, HKY, TN93, GTR, ...



Mitochondrial DNA Sequences of Primates: Tempo and Mode of Evolution

Wesley M. Brown[1], Ellen M. Prager, Alice Wang[2], and Allan C. Wilson

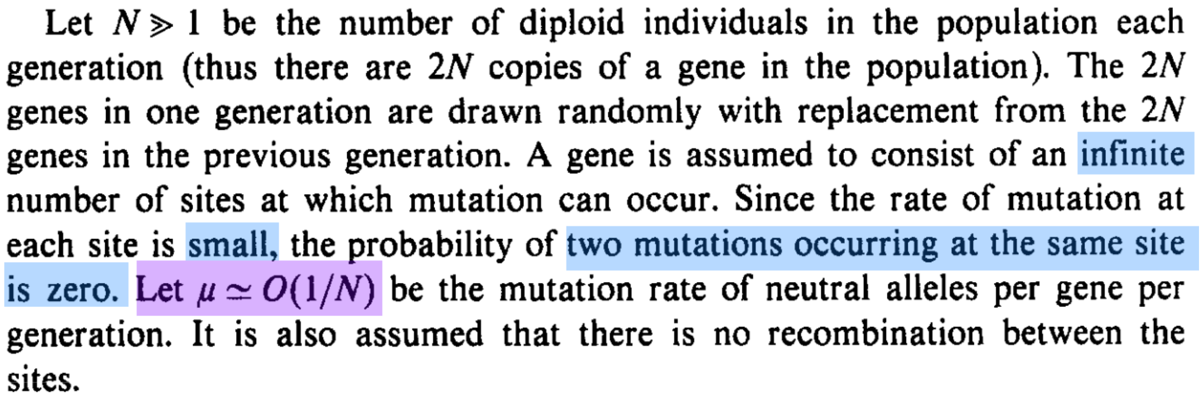
J Mol Evol (1982) 18:225–239

A. 145
C
G
H
O
Gb
B. 147
H
C
G
O
Gb
C. 148
H
G
C
O
Gb
D. 157
H
C
G
O
Gb
E. 173
C
G
O
H
Gb
F. 189
C
G
H
O
Gb

Fig. 11. Possible evolutionary trees for humans and apes. Six possible phylogenetic relationships (A-F) are shown among the five higher primates considered here. The figure indicates for each tree the minimum number of events required to produce that tree by a parsimony analysis of the sequence data for the 90 phylogenetically informative sites within the 896 bp fragment of mtDNA. The *abbreviations* used are H, human; C, chimpanzee; G, gorilla; O, orangutan; Gb, gibbon. Trees analogous to A-E were constructed also by subjecting the intra-primate data in Table 1 to phylogenetic analysis by the matrix methods of Fitch and Margoliash (1967) and Farris (1972); both methods favored trees A and B and indicated tree E was far less likely than A-D

Mutation model history

Mid - Late: Some population geneticists started to tinker with sequence data and used the **infinite sites model** (Kimura 1969). For example, Strobeck (1984) evaluated the population size of two Drosophila species assuming an infinite sites model.



Let $N \gg 1$ be the number of diploid individuals in the population each generation (thus there are $2N$ copies of a gene in the population). The $2N$ genes in one generation are drawn randomly with replacement from the $2N$ genes in the previous generation. A gene is assumed to consist of an infinite number of sites at which mutation can occur. Since the rate of mutation at each site is small, the probability of two mutations occurring at the same site is zero. Let $\mu \simeq O(1/N)$ be the mutation rate of neutral alleles per gene per generation. It is also assumed that there is no recombination between the sites.

Drosophila virilis: n=10, a_1=4, a_2=6

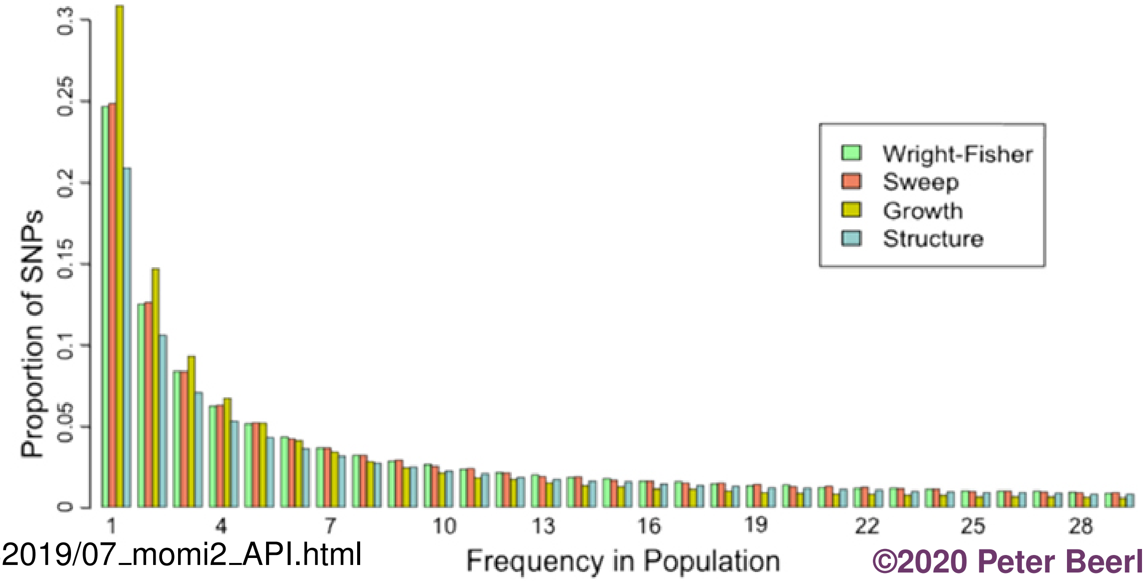
Infinite allele: $\theta_{\text{Ewens}} = 1.97$

Variable site: $\theta_{\text{Watterson}} = 0.35$

Infinite site: $\theta_{\text{Strobeck}} = 0.34$

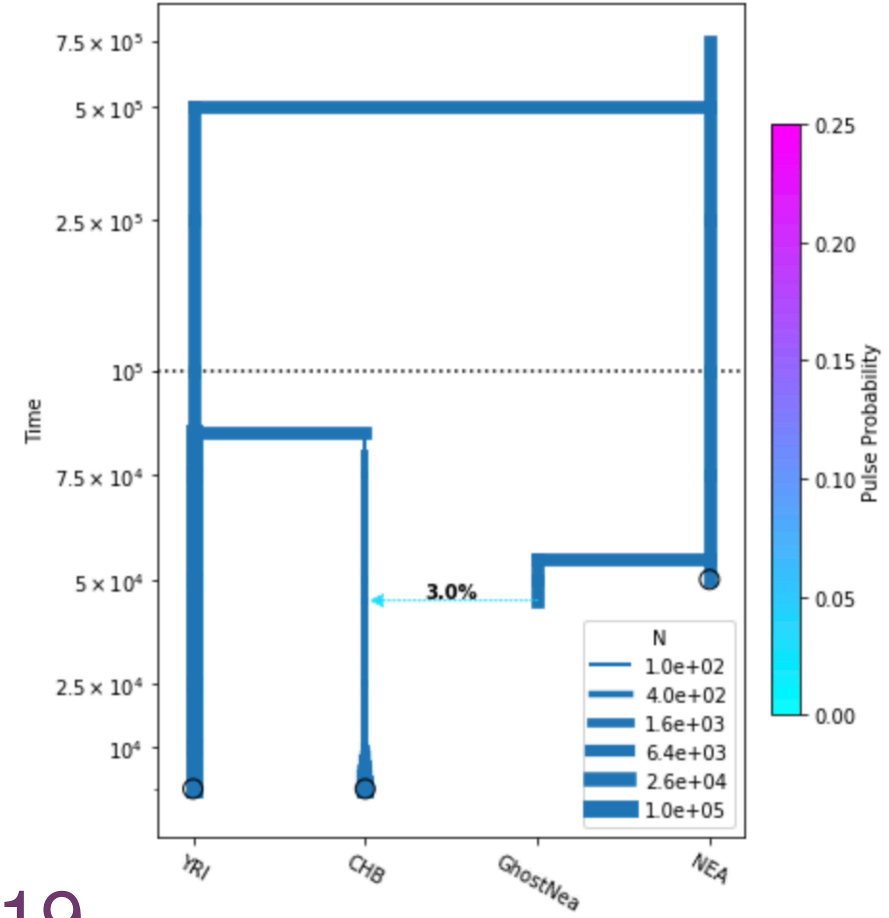
Mutation model history

Late: It seems that (almost) all phylogeneticists use the time reversible GTR finite mutation model now. In population genetics (aka population genomics) we used (and some still do ☺) for a short time finite mutation models in software that used Markov chain Monte Carlo methods. But then, it seems, many abandoned them because the programs were either not up to the task (memory problems) or then then too slow. Site frequency spectra were embraced and allowed to analyze very complex population models.



https://radcamp.github.io/IBS2019/07_momi2_API.html

Example of a site frequency spectrum method



Momi2: momi (MOran Models for Inference) is a Python package that computes the expected sample frequency spectrum (SFS) and uses it to fit demographic history. In short: we take SNP dataset with n population, create a multipopulation SFS, create a specific population model and find parameter setting of this particular model so that we can generate an expected SFS that is close the estimated SFS. We assume an infinite sites mutation model.

SNP ascertainment issues

SNPs and population parameters: Single nucleotide polymorphisms are usually reported as an ancestral allele and the alternative allele (2-state) this works fine under the assumption that populations are small and mutation rate is small.

Nextgen sequencing allows to retrieve SNPs relatively unbiased, with enough coverage we can find them all, and also if we do not remove lower frequency allele then we may get good estimates of a site frequency spectrum.

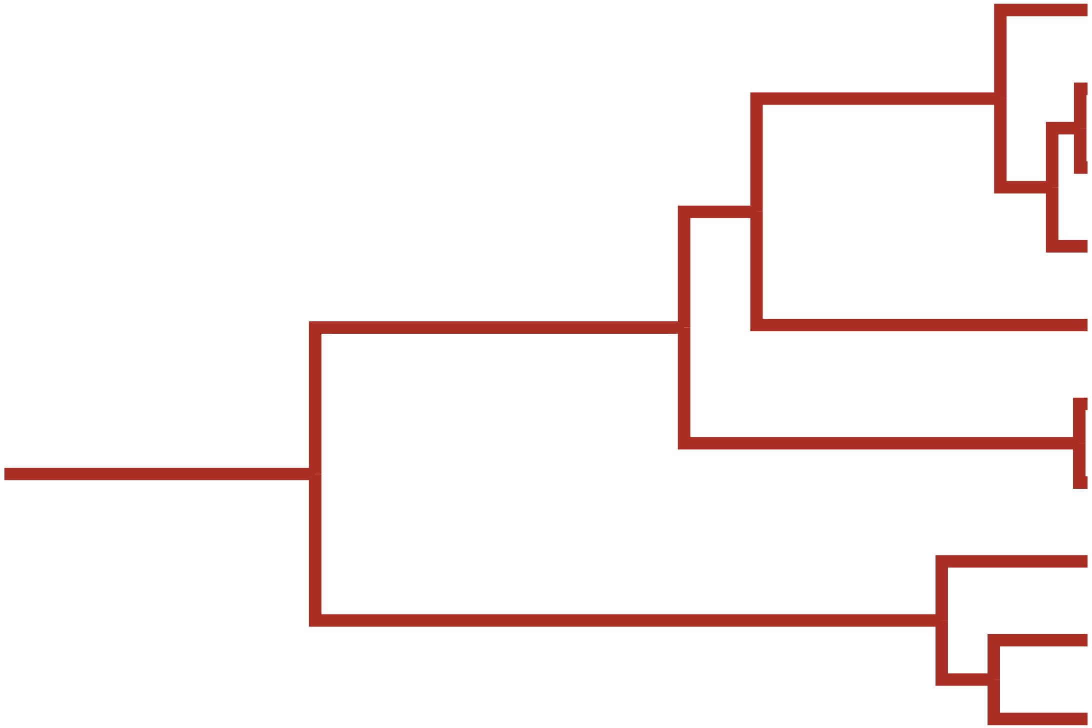
Removal of low frequency SNPs without correction will lead to errors.

Analyses in population genomics

Population model parameter estimation: The coalescent with many samples becomes rather intractable when we assume complicating forces such as gene flow, recombination, population size changes, admixture, population splitting. This was one of the reason for the development of methods that depend on the SFS, but one may wonder whether we have traded one problem with another untractable one.

Analyses in population genomics

A sample of a single population:

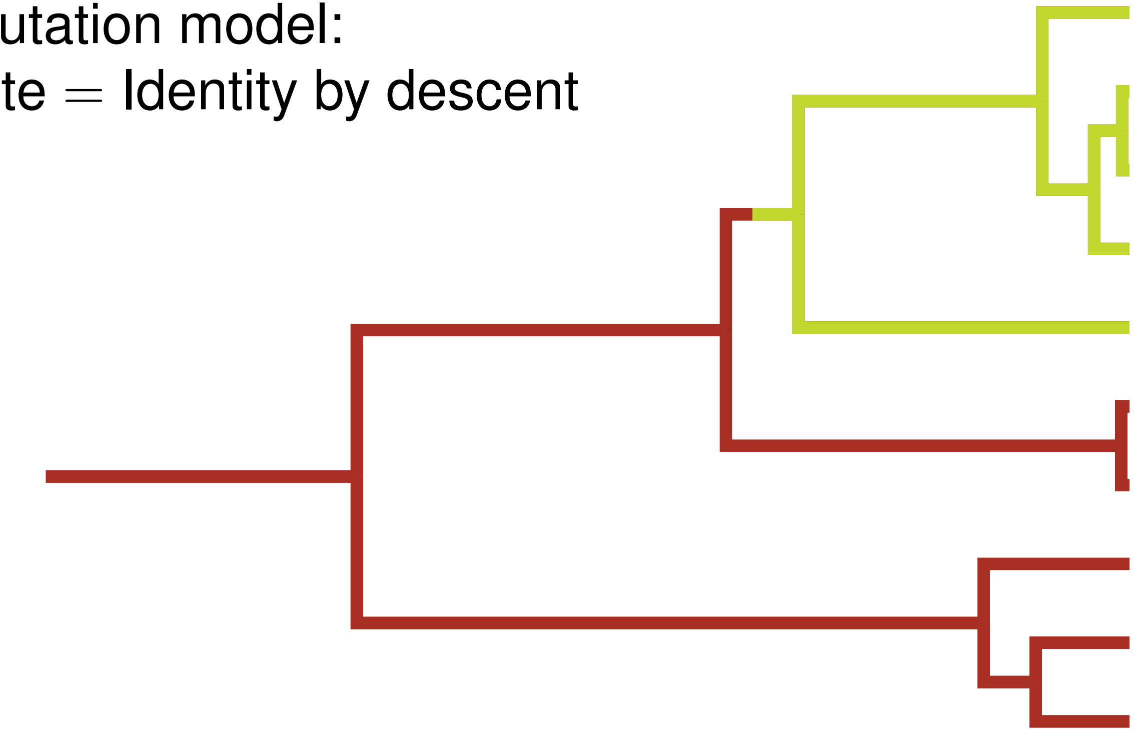


Analyses in population genomics

A sample of a single population:

Infinite mutation model:

allelic state = Identity by descent

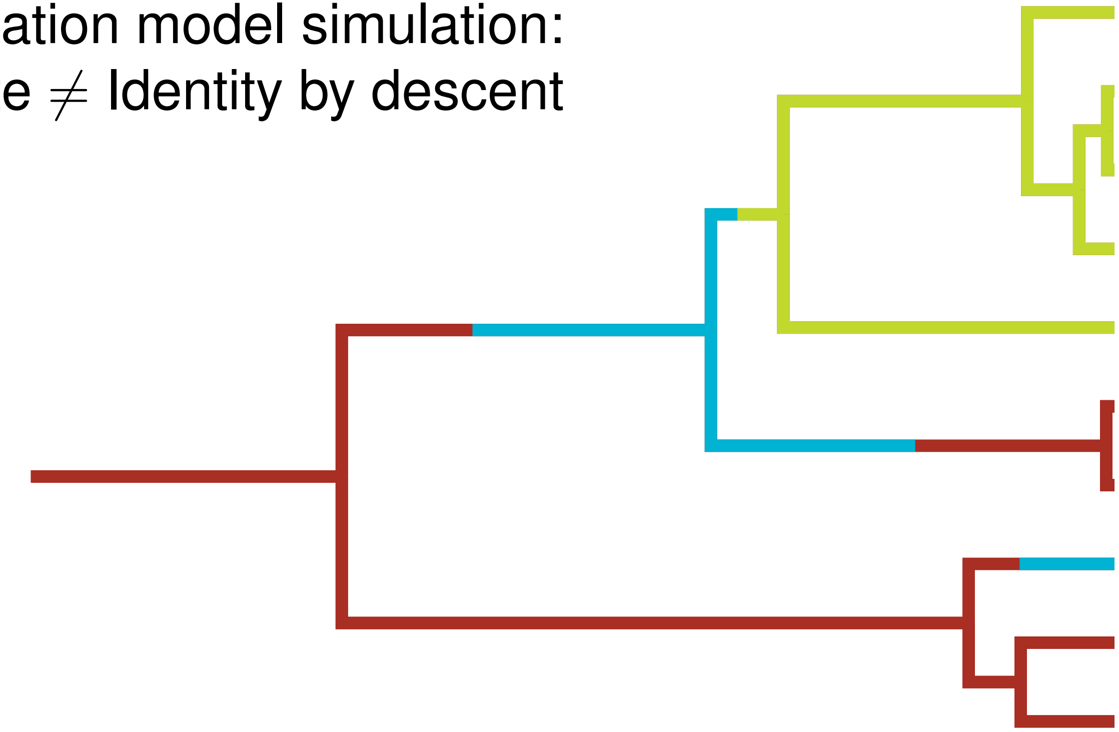


Analyses in population genomics

A sample of a single population:

Finite mutation model simulation:

allelic state $\neq$ Identity by descent



Analyses in population genomics

The last example shows three different alleles: Even if we would only see 2 alleles, then we may guess that the mutation rate with finite mutation models may be higher than with an infinite mutation model.

Variability is the measure for almost everything in population genetics! Low variability suggests low population size, with little variability we also assume that two populations are more similar, thus low gene flow, or recent divergence times, ...

Scenarios that we should explore, but I do not see lots of work on that.

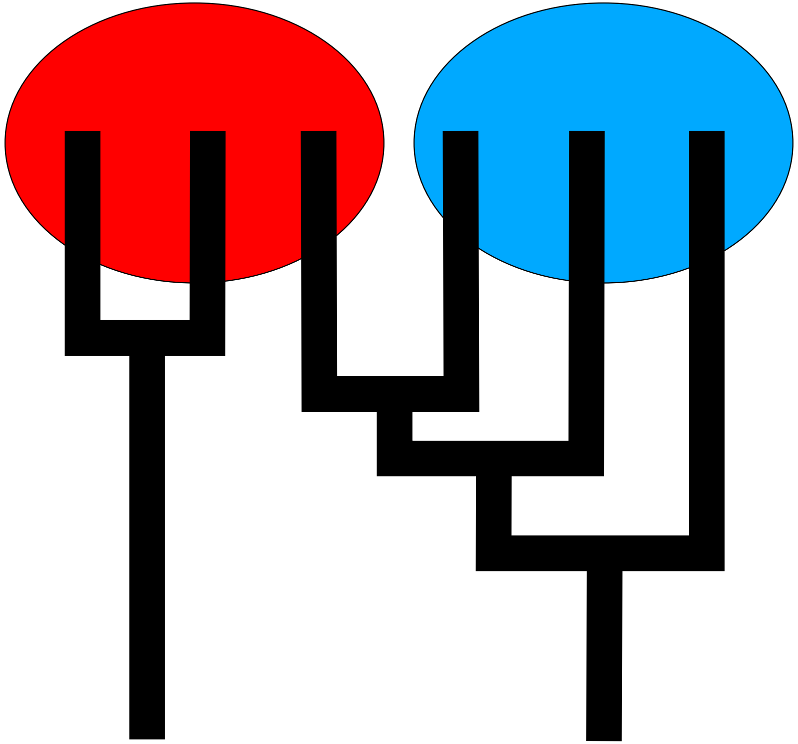
Infinite sites vs 2-allele SNPs vs finite sites

Population genomics using finite mutation models

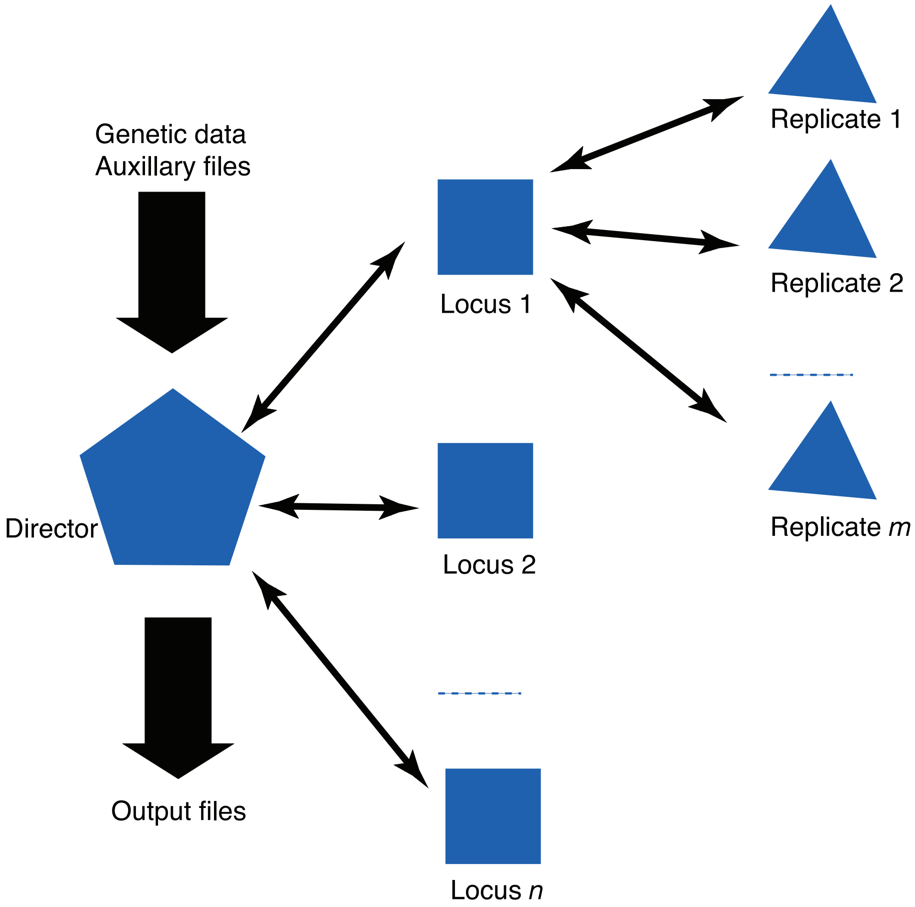
MIGRATE-N estimates population parameter for a large set of different models (there is still considerable room for improvement!)

For sequence data MIGRATE only allows for finite mutation models. It calculates likelihoods using them and approximates posterior distributions of the parameters.

Full coalescent likelihood methods, such as MIGRATE or LAMARC or IMA, use Markov chain Monte Carlo methods to approximate posterior distributions of the parameters of interest, this can take considerable amount of time and practice (Tutorials and a complete manual are good tools to understand these programs.)



Migrate runs in parallel



Genetic data
Auxillary files

Director

Output files

Locus 1

Locus 2

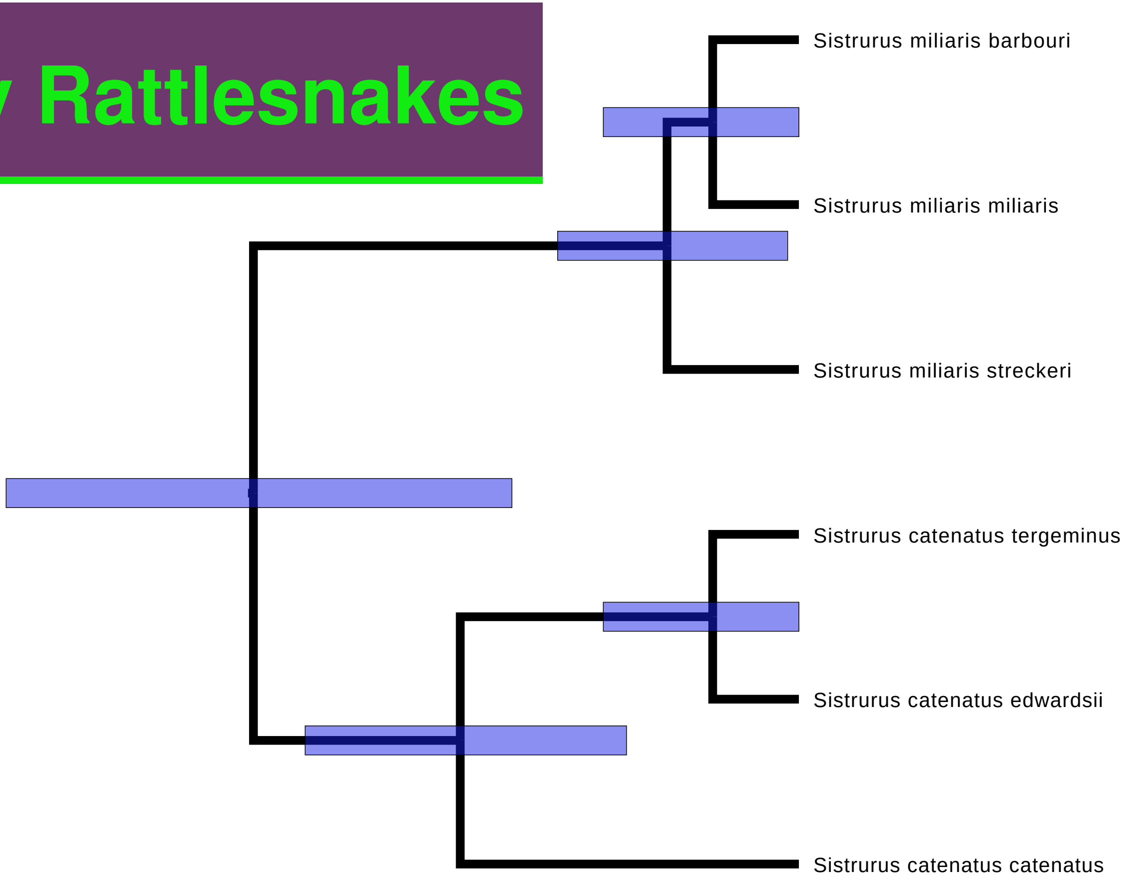
Locus *n*

Replicate 1

Replicate 2

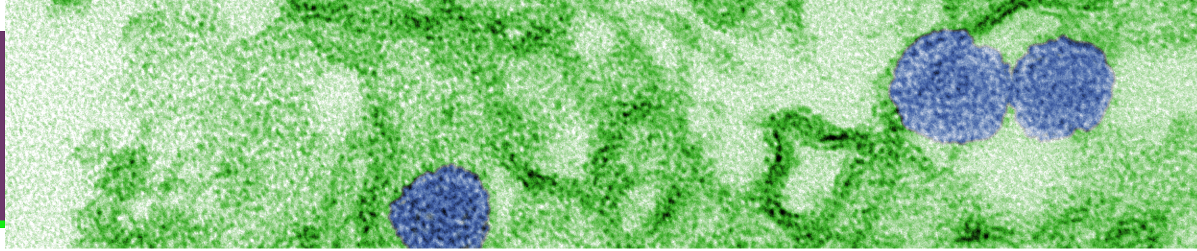
Replicate *m*

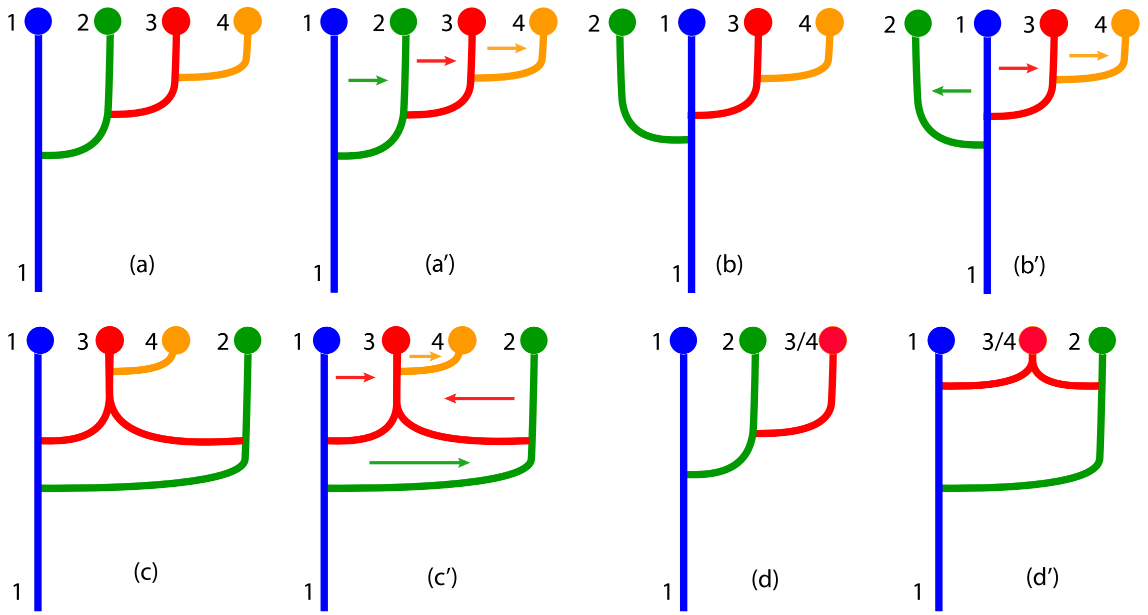
Pygmy Rattlesnakes


Sistrurus miliaris barbouri
Sistrurus miliaris miliaris
Sistrurus miliaris streckeri
Sistrurus catenatus tergeminus
Sistrurus catenatus edwardsii
Sistrurus catenatus catenatus

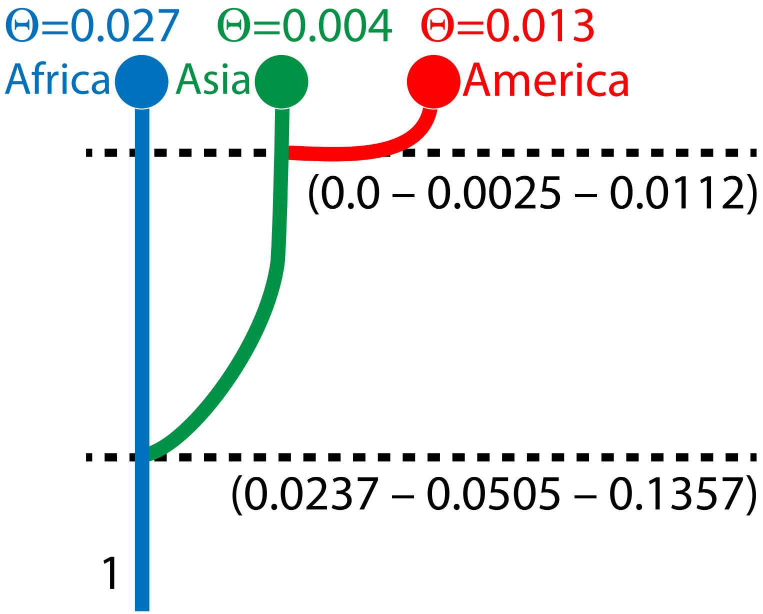
Model	Log(mL)	LBF	Model-probability
1: 3 species:	-15887.49	0.00	1.0000
2: 6 species:	-15961.95	-74.46	0.0000

Zika biogeography





1=Africa, 2=Asia, 3=Brazil, 4='Central' America



Thank you Chris and Erika

DBI 1564822

Hubby J. L., Lewontin R. C., 1966. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics 54: 577-594.

Motoo Kimura and James F. Crow. 1964. The Number of Alleles That Can Be Maintained in a Finite Population GENETICS April 1964 49: 725-738 [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4905530/]

Ohta, T., & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetical Research, 22(2), 201-204. doi:10.1017/S0016672300012994

Brown, W. M., Prager, E. M., Wang, A., and Wilson, A. C. 1982. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. J. Mol. Evol. 18:225-239.

Kimura, Motoo (1969). The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. Genetics. 61 (4): 893-903.

Strobeck C. Estimation of the neutral mutation rate in a finite population from DNA sequence data. Theor Popul Biol. 1983;24(2):160-172. doi:10.1016/0040-5809(83)90039-4

Jack Kamm, Jonathan Terhorst, Richard Durbin. & Yun S. Song (2019) Efficiently Inferring the Demographic History of Many Populations With Allele Count Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1635482

Beerli, P., Mashayekhi, S., Sadeghi, M., Khodaei, M., & Shaw, K. (2019). Population genetic inference with MIGRATE. Current Protocols in Bioinformatics, 68, e87. doi: 10.1002/cpbi.87; https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.87

Credit: ESO/C. Malin