

Statistical Algorithms in Population Genetics

Adam Pershan

The General Story

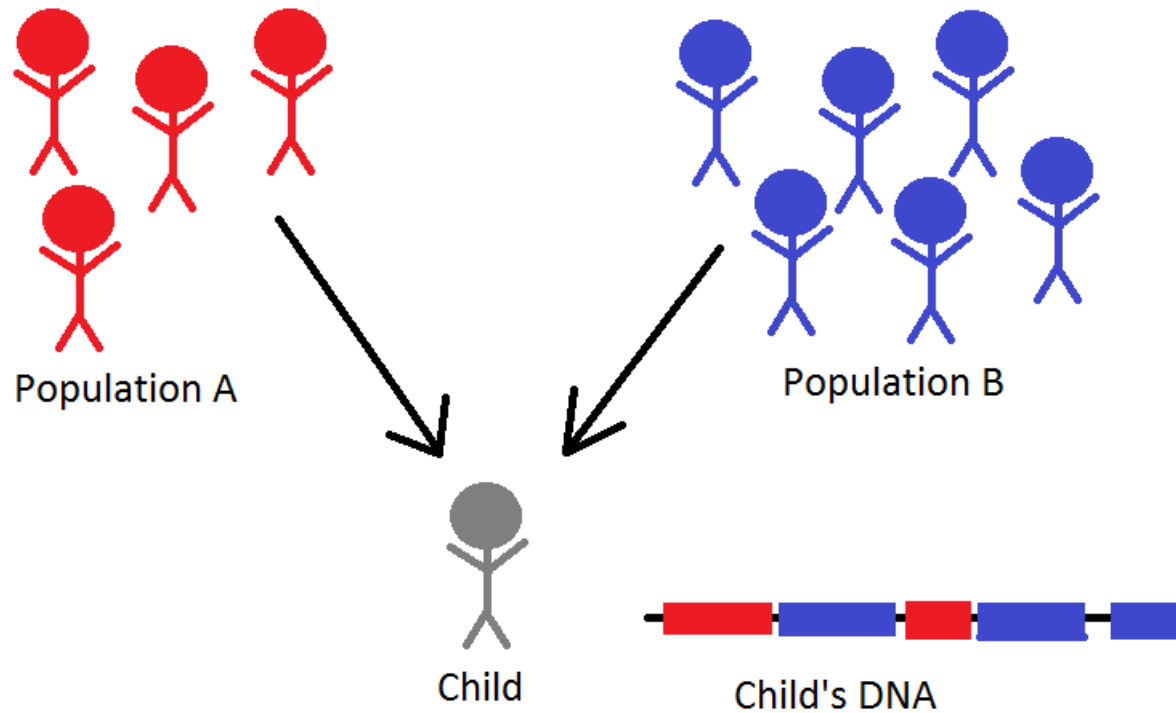
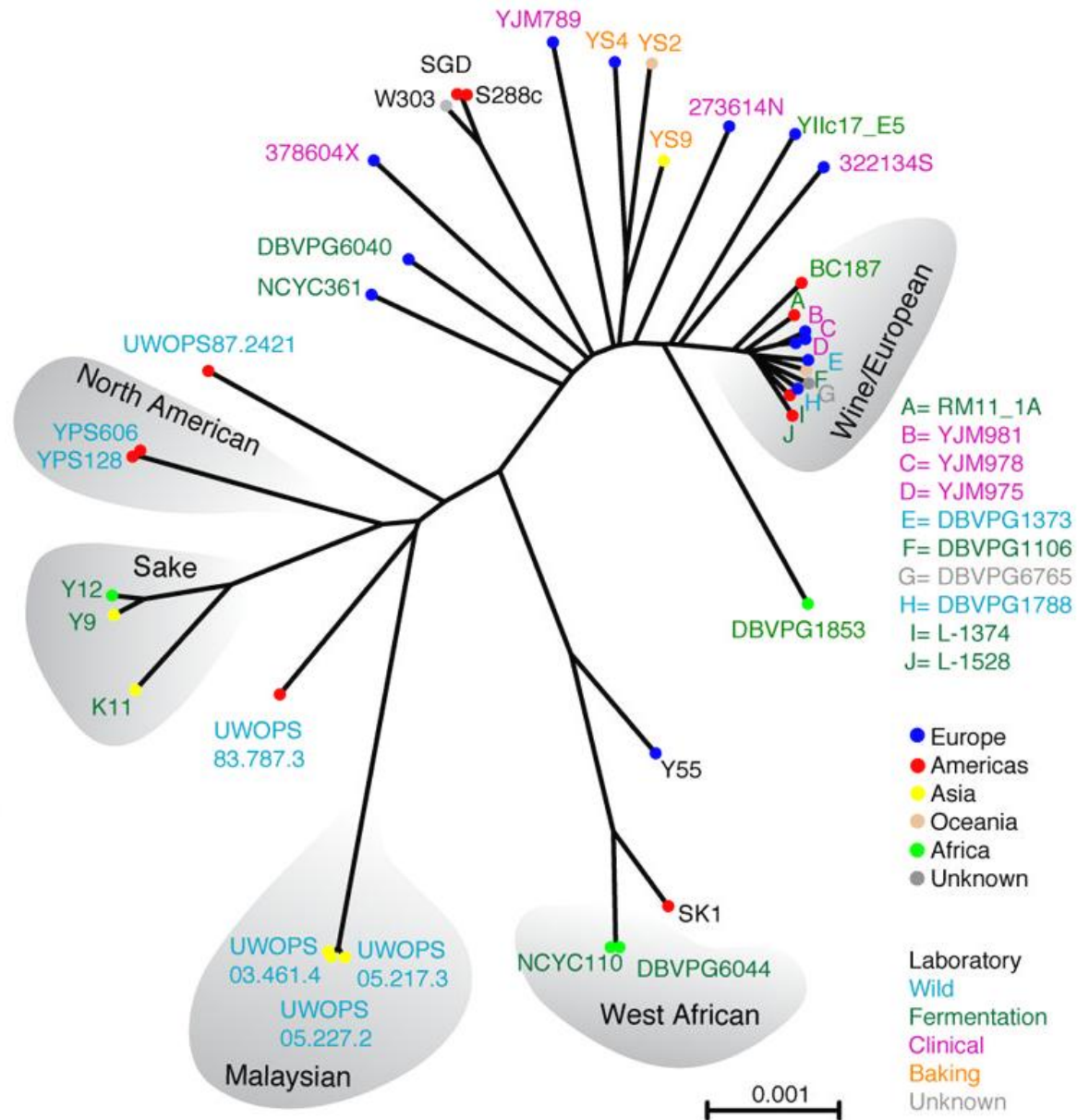




FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.



Why do we care?

- ▶ Interesting for studying human populations and evolution.
- ▶ Clinical ramifications for disease tracking: for example, Native American ancestry has been associated with a increased risk of relapse in children with some types of leukemia.
- ▶ Interesting for studying other organisms and their evolutionary tracks.

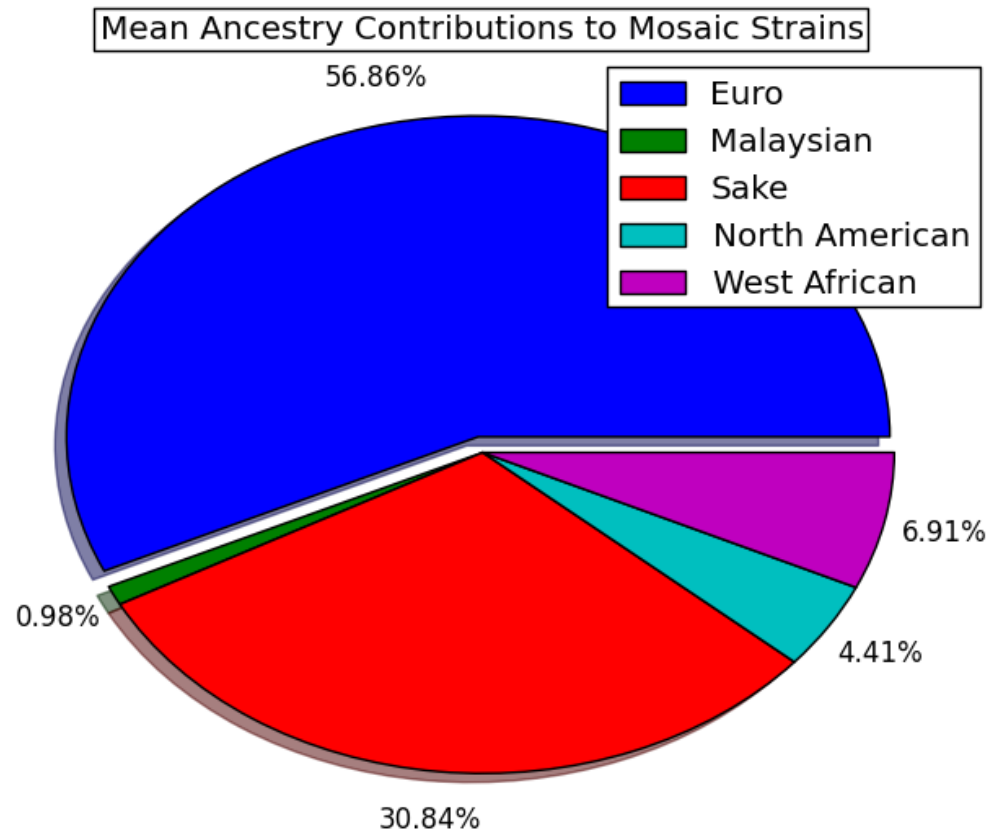


LAMP-LD

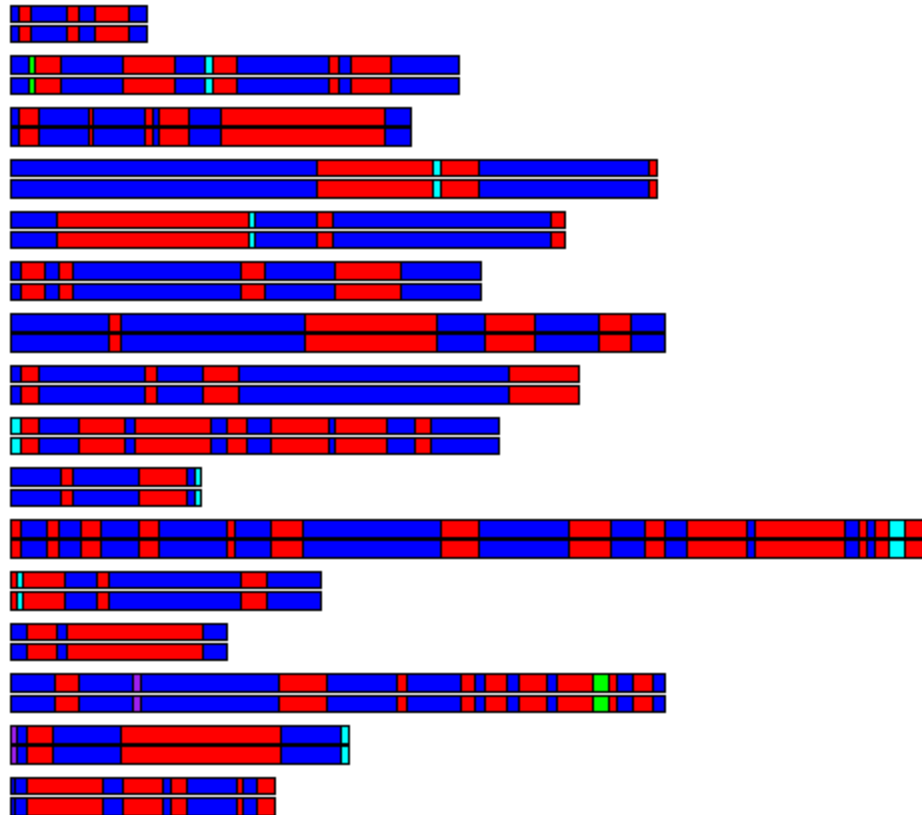
- ▶ Uses a Hidden Markov Model, where ancestry is the 'hidden state.'
- ▶ Used to infer local ancestry.
- ▶ Very fast, very accurate.
- ▶ Optimized for multiple-way admixture, like in our case with 5 ancestral populations.



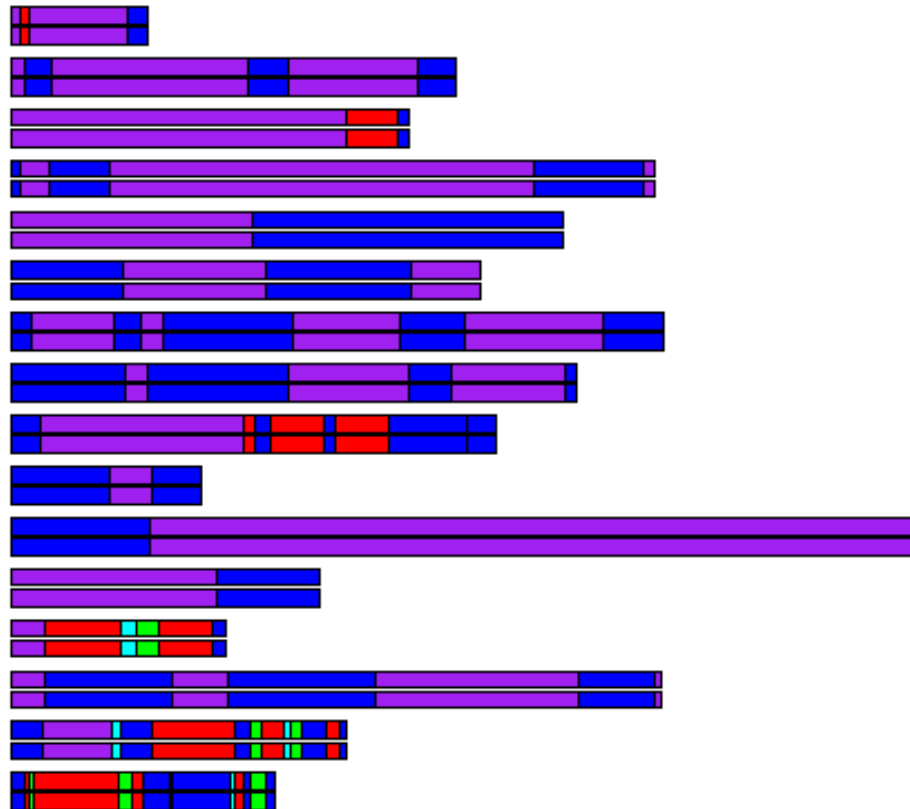
LAMP-LD output



LAMP-LD output



LAMP-LD output



Analysis

- ▶ Genetic recombinations are well modeled by a Poisson random variable, like any rare event with many trials (radioactive particle emissions or goals in a soccer match).¹
- ▶ Was first used to model the number of deaths due to horse kicks in the Prussian army in 1898.

1: <http://iopscience.iop.org/0295-5075/89/3/38007/>

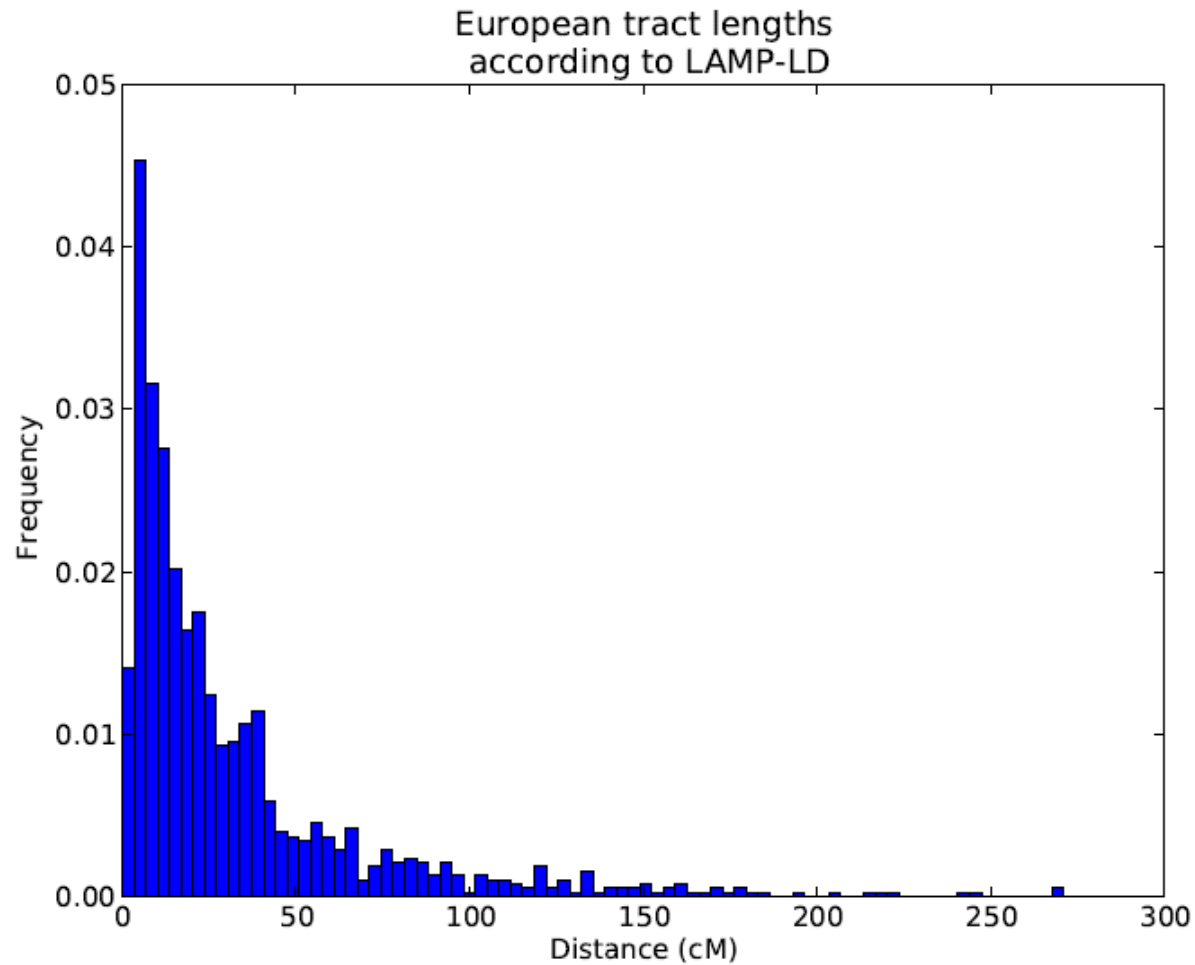


Analysis

- ▶ The time between ‘arrivals’ of a Poisson process follows an exponential distribution. $f(x; \lambda) = \lambda e^{-\lambda x}$
- ▶ Tract size depends on the number of recombinations that have occurred. In particular, a continuous tract is one in which no ancestry-breaking recombinations have occurred. Thus, tract length distribution should be exponential.



And it is!



A Question

- ▶ Can we use this distribution to approximate the number of generations since admixture?



Well...

- ▶ Possibly. But it also depends on the initial migration rates, something that's unknown in this case.
- ▶ It might also depend on the specifics of admixture. Continuous migration? Multiple waves?
- ▶ Bottom line – we know the exponential rate should be proportional to the number of generations since admixture, but numerical solutions depend on other parameters not readily available.



Rolloff to the rescue!

- ▶ Rolloff is a method that measures time since admixture.
- ▶ It does so by looking at the linkage disequilibrium between SNPs due to admixture.
- ▶ Definition: linkage disequilibrium (l_d) is the nonrandom association between two alleles such that certain combinations are more likely to occur together.



Why does Rolloff work?

- ▶ We plot admixture LD on the y-axis and genetic distance between SNPs on the x-axis.
- ▶ As two SNPs get farther apart, we expect there to be less admixture LD.
- ▶ The rate of decline of admixture LD is directly related to the number of generations since admixture, since that indicates how many recombinations have occurred between any two SNPs.

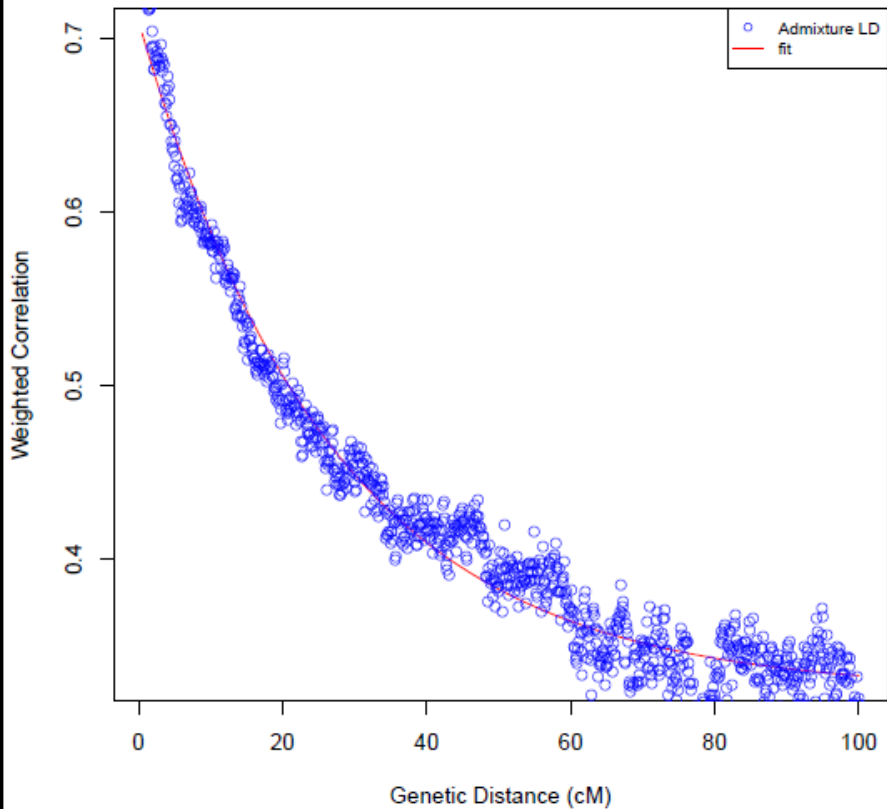


Why does Rolloff work?

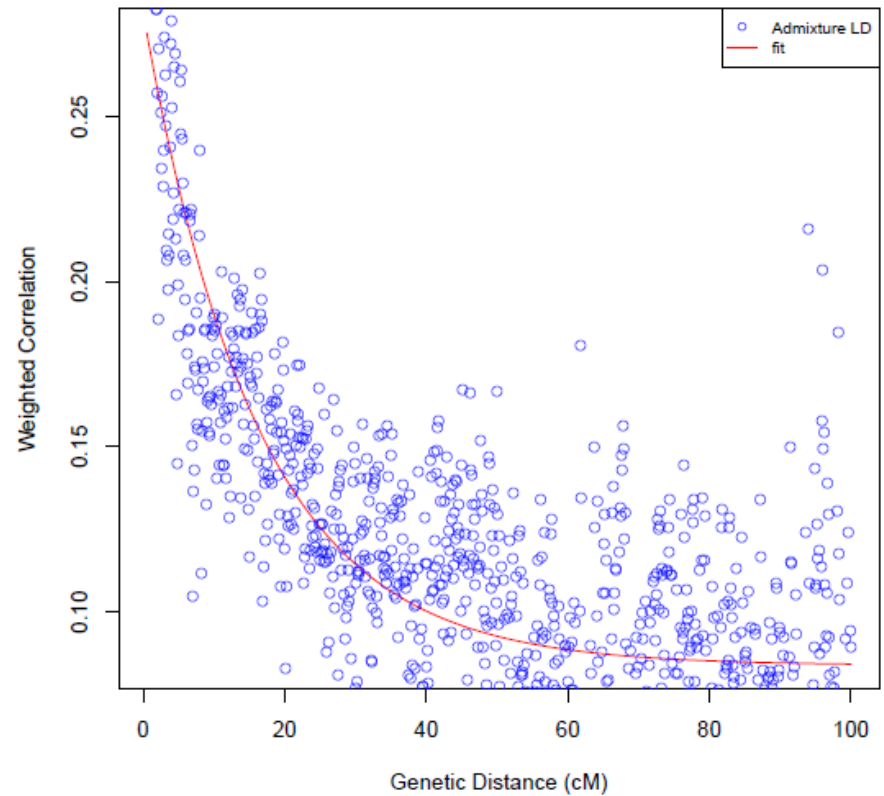
- ▶ Rolloff fits an exponential curve to a plot of admixture LD vs. distance, and uses the rate of exponential decline to calculate the generations since admixture.
- ▶ Side note: we're dealing with generations that involve recombination, something yeast do roughly every 50,000 asexual generations (~100,000 hours or ~11.5 years).



Rolloff outputs



European, Malaysian



North American, Sake

Pairwise Rolloff results

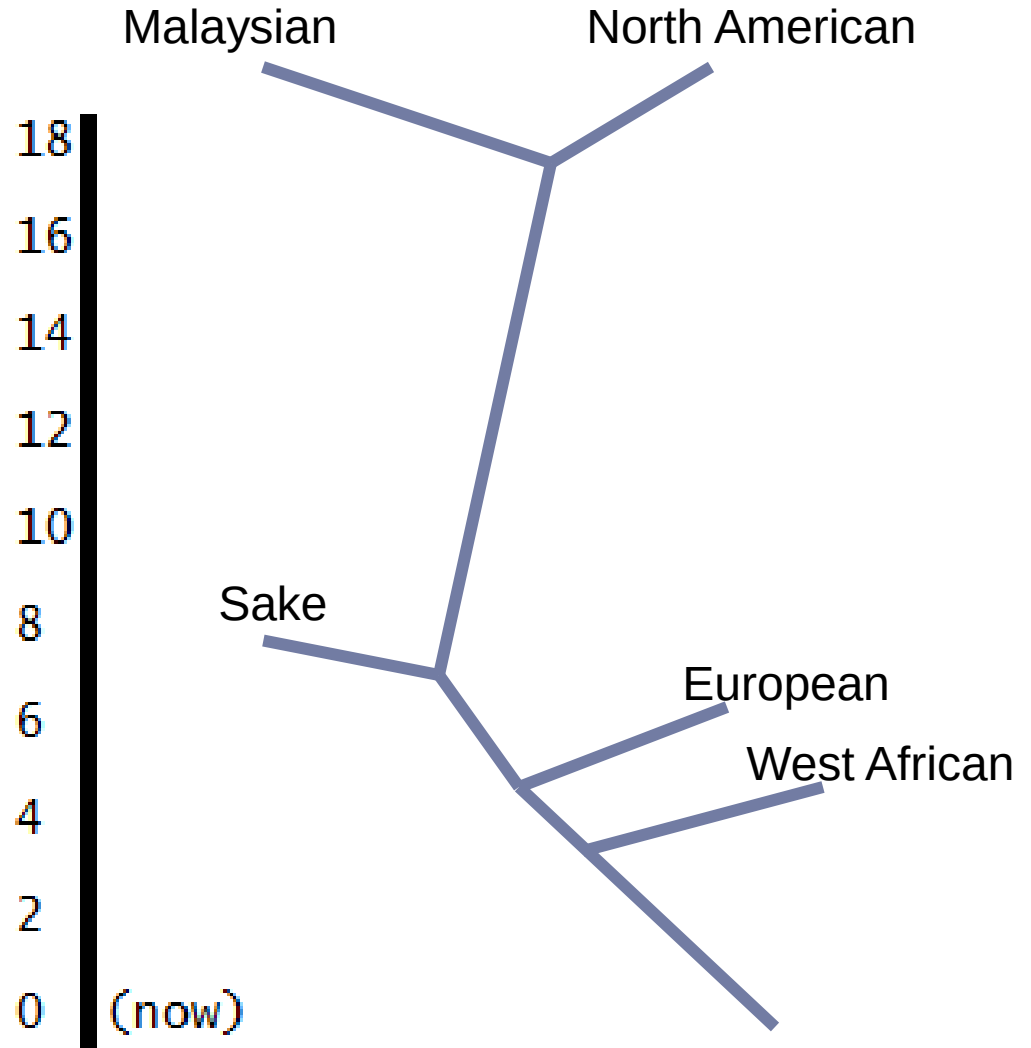
	Euro	Sake	Ma	Wa	Na
Euro		3.654	3.768	4.097	4.067
Sake			7.462	3.801	6.217
Ma				3.165	18.21
Wa					3.701
Na					

Data is in generations since admixture

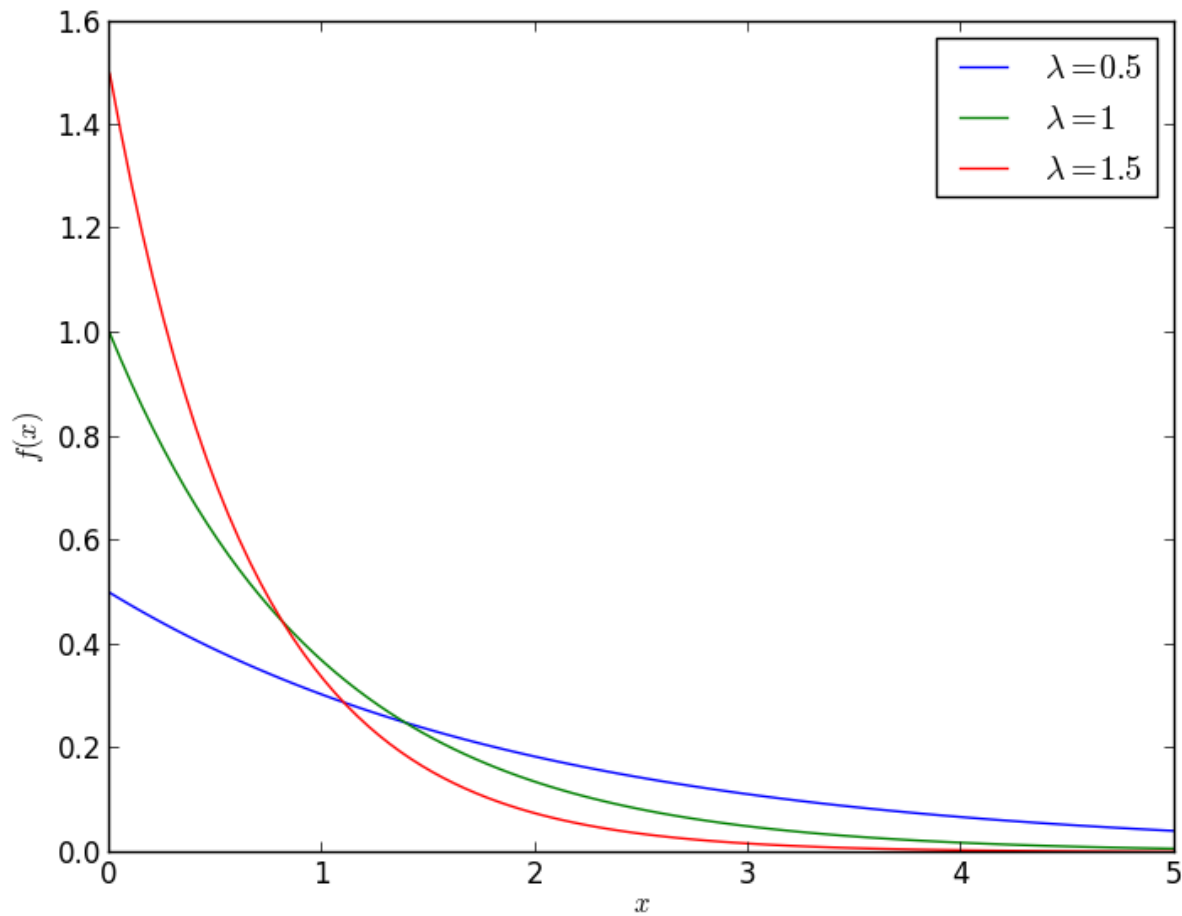


Inferred timeline

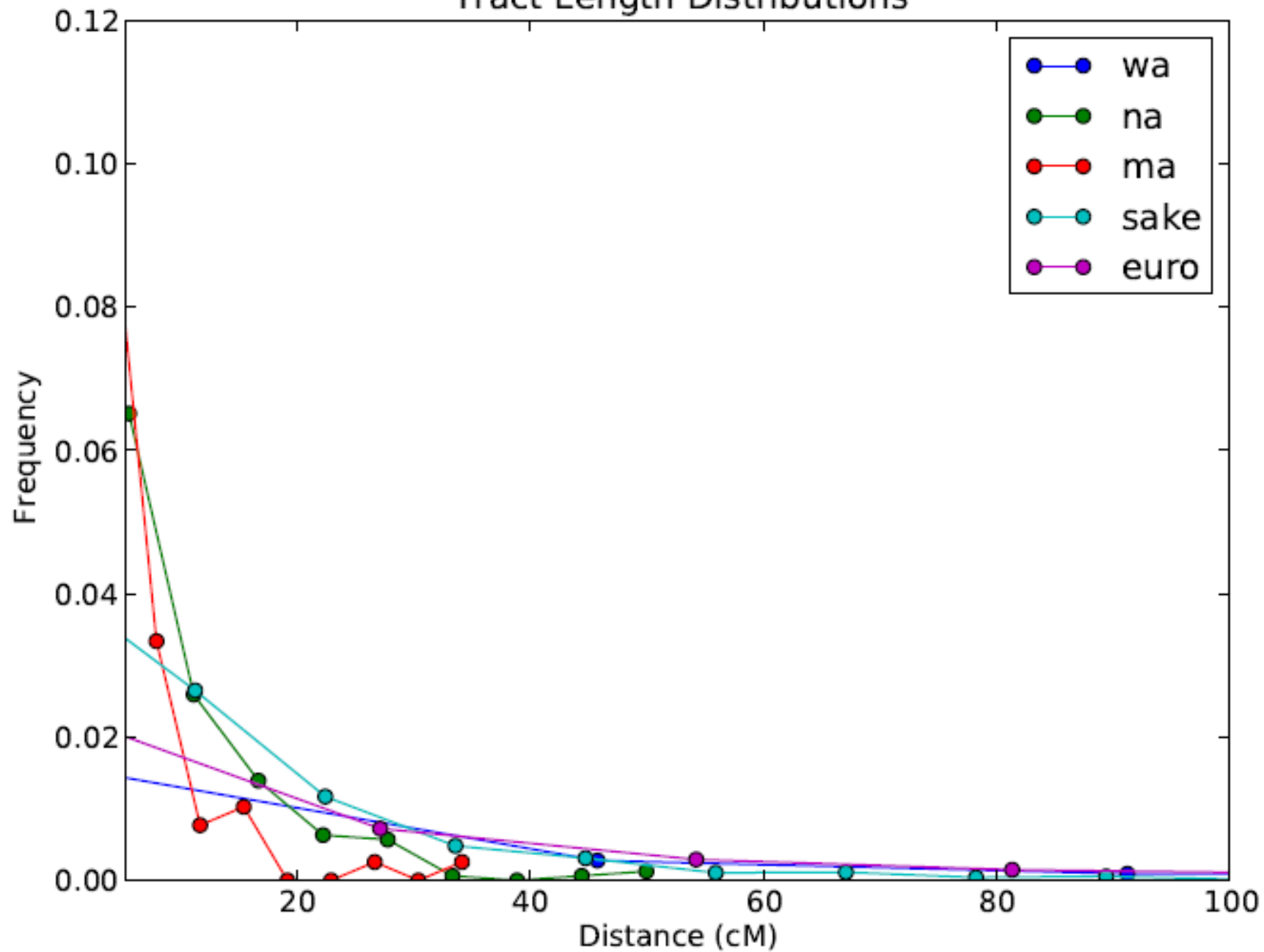
	Euro	Sake	Ma	Wa	Na
Euro		3.654	3.768	4.097	4.067
Sake			7.462	3.801	6.217
Ma				3.165	18.21
Wa					3.701
Na					



Exponential distributions, again



Tract Length Distributions



Thank you to my amazing mentors, Dr.
Kevin Chen and Liyang Diao!

