Parameter-rich demographic models

Fernando Racimo

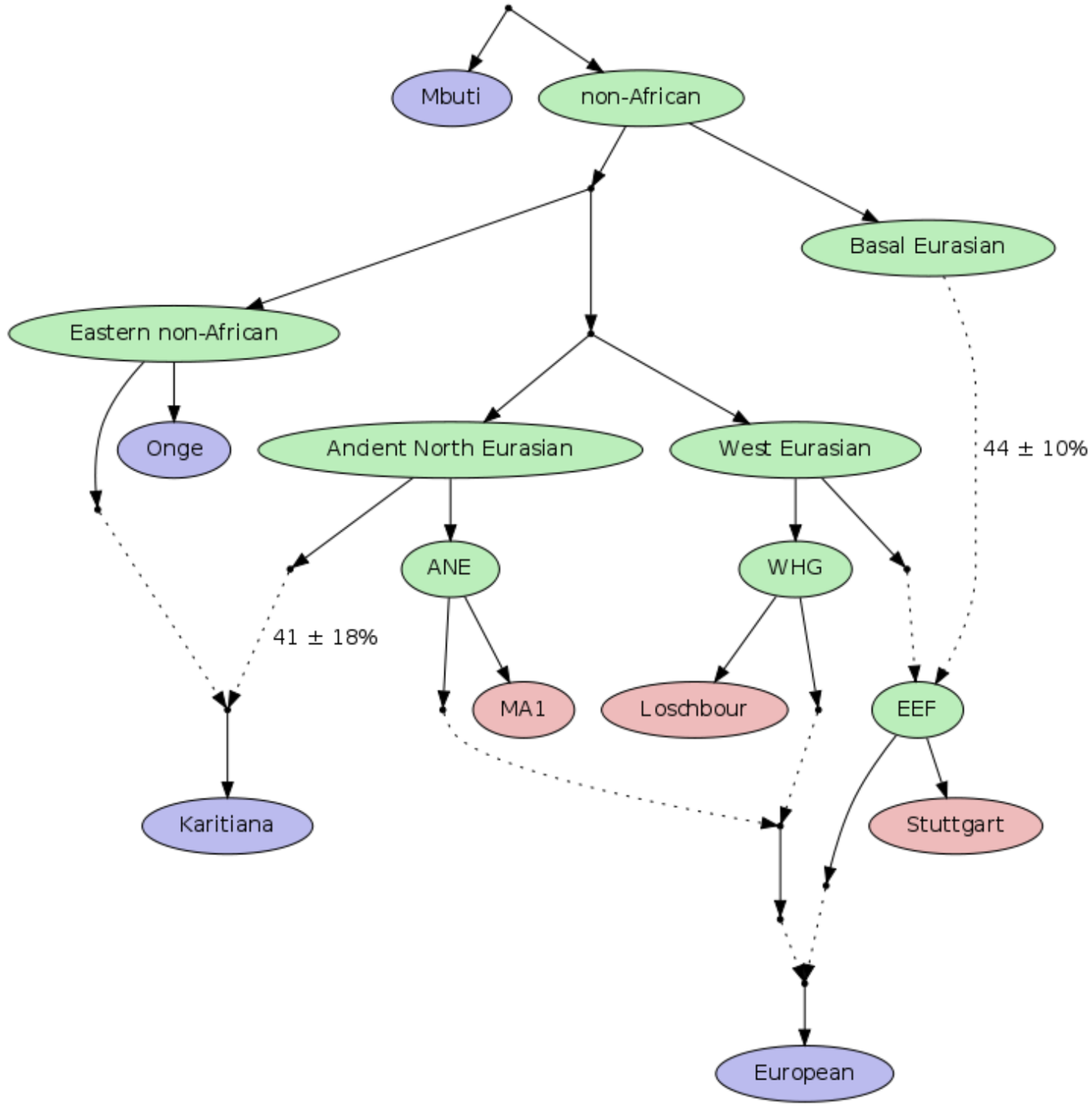
Parameter-rich demographic models

Fernando Racimo
Adelaide, 2018

Parameter-rich models

- Admixture graphs
- Explicit likelihood methods
- Approximate Bayesian Computation
- Deep Learning

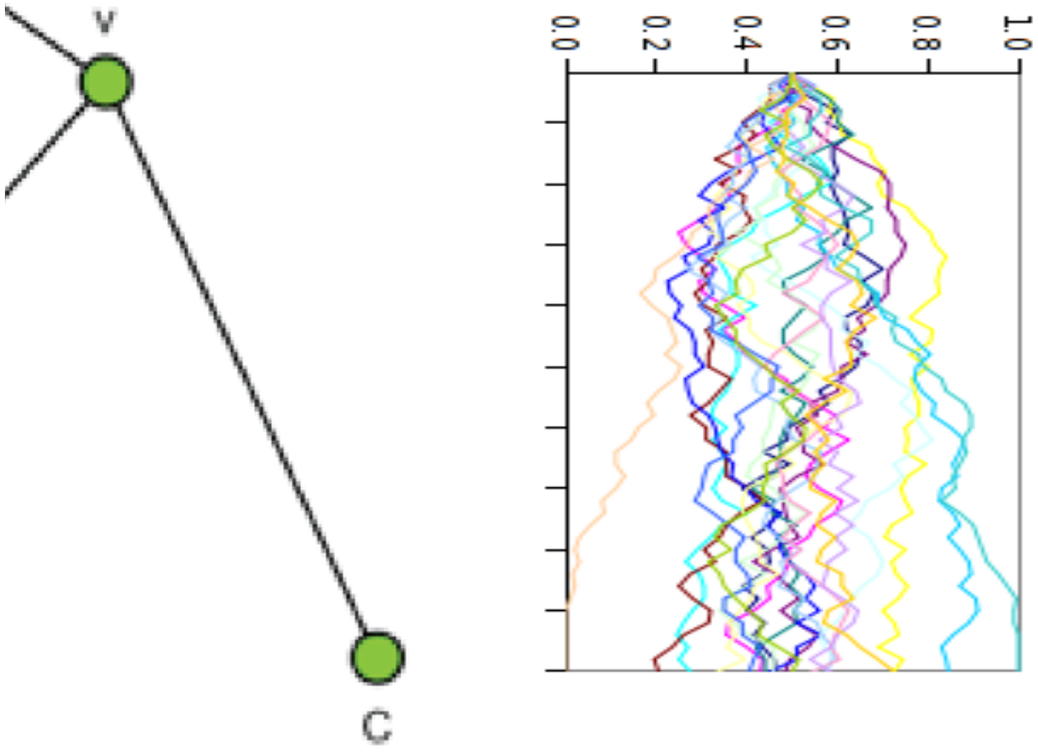
Admixture graphs



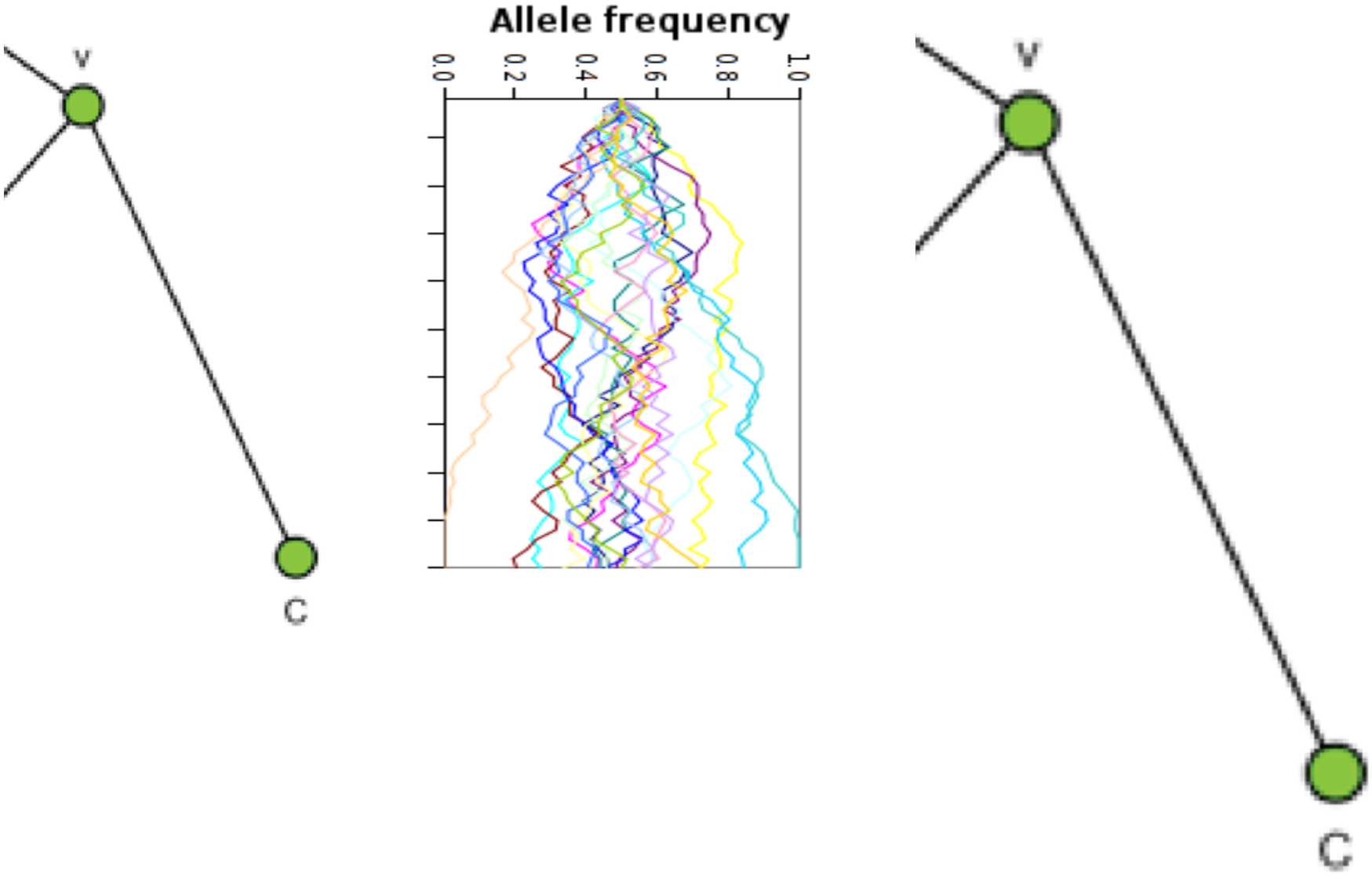
Lazaridis et al. (2014)

How to model genetic drift on a branch?

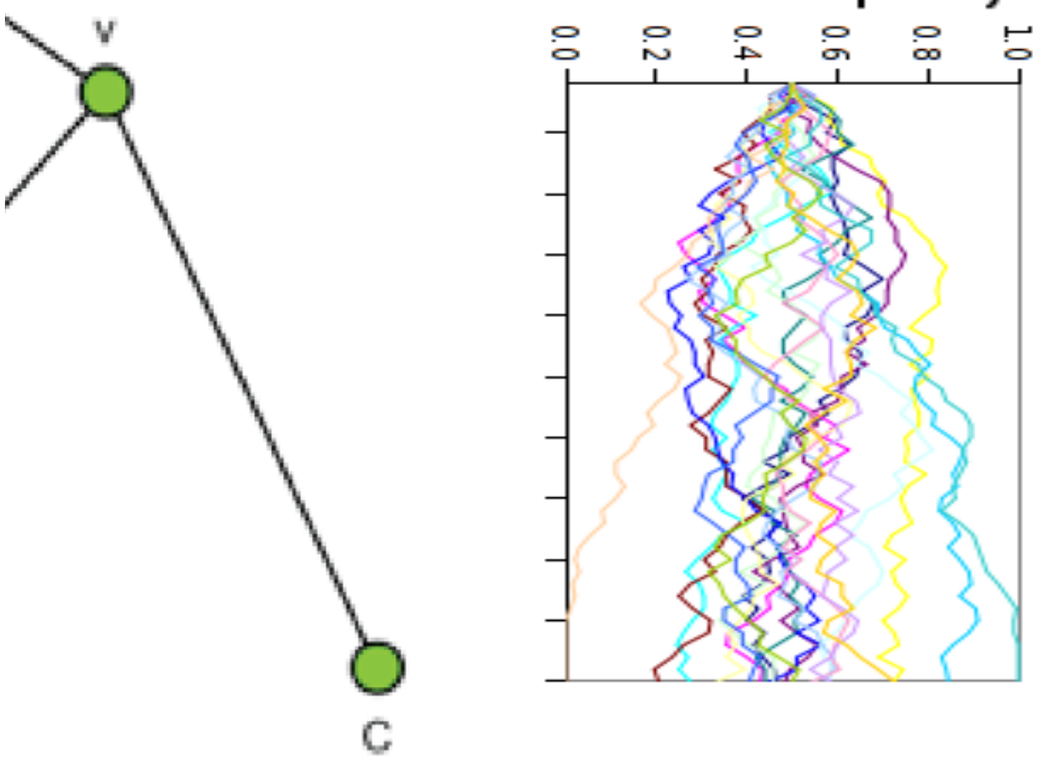
How to model genetic drift on a branch?

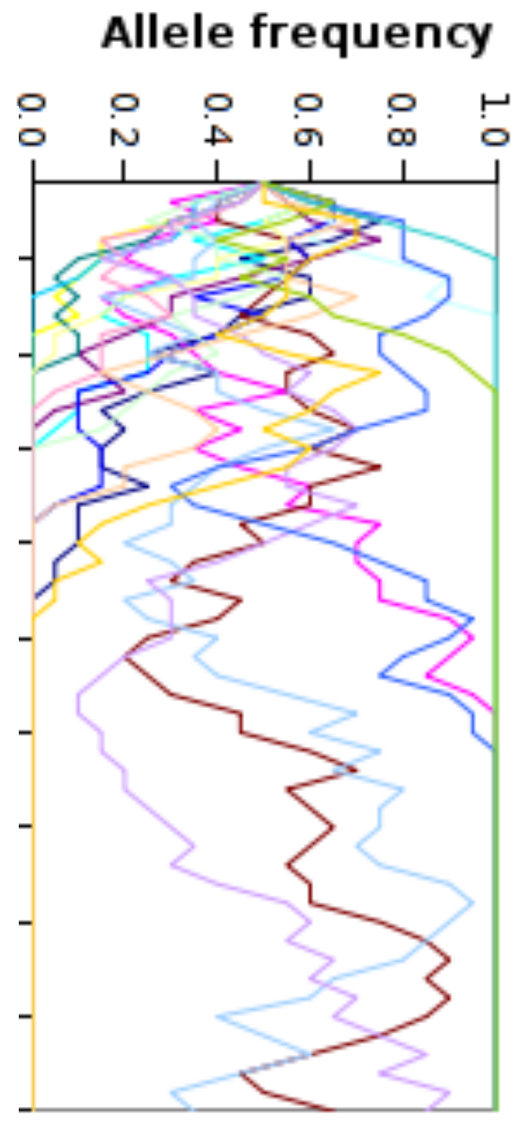


How to model genetic drift on a branch?

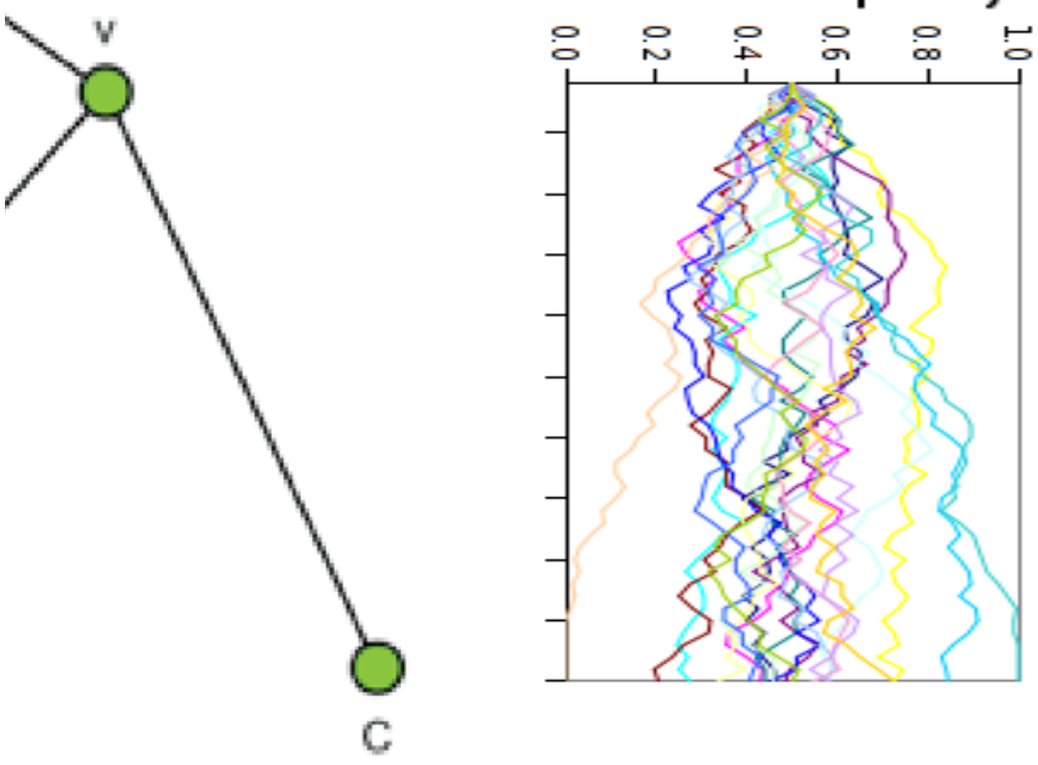


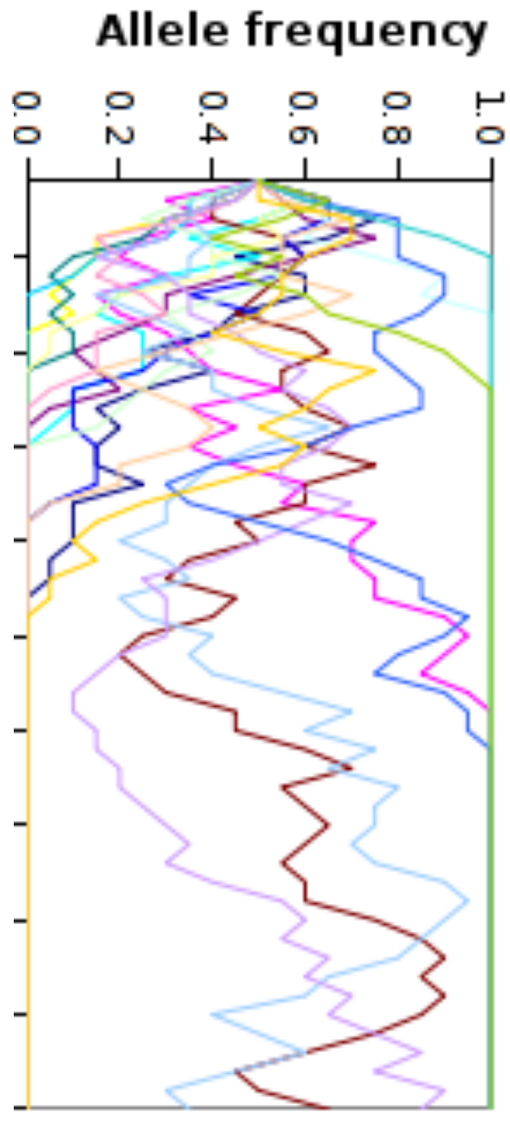
How to model genetic drift on a branch?





How to model genetic drift on a branch?

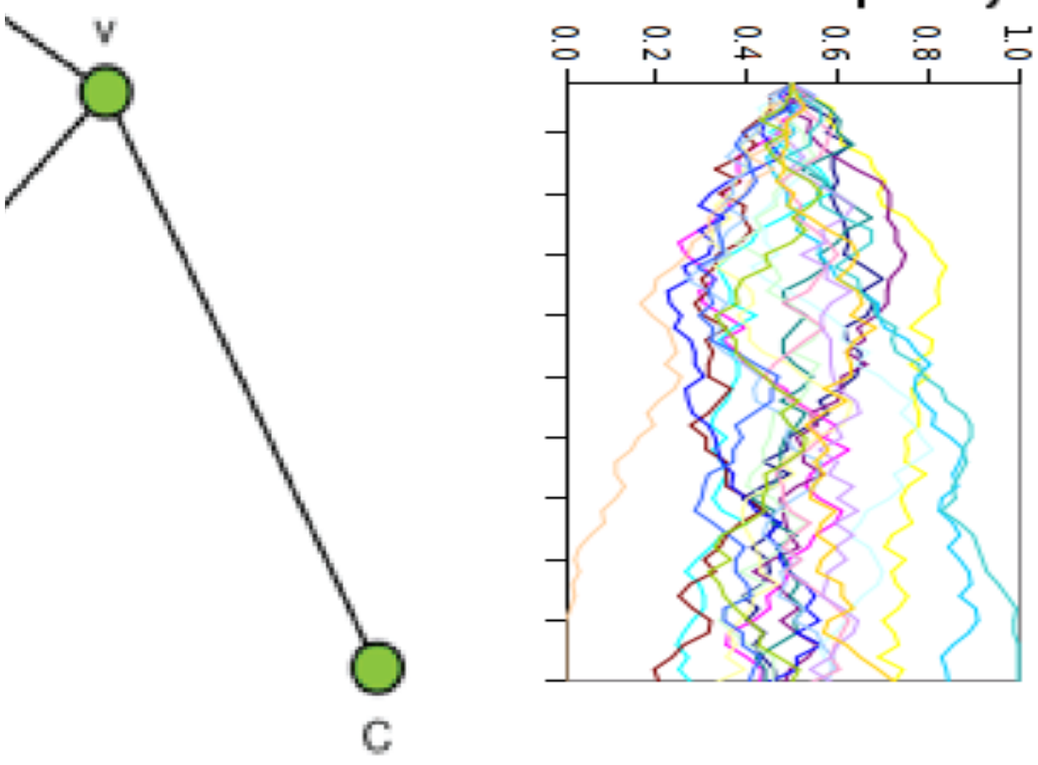


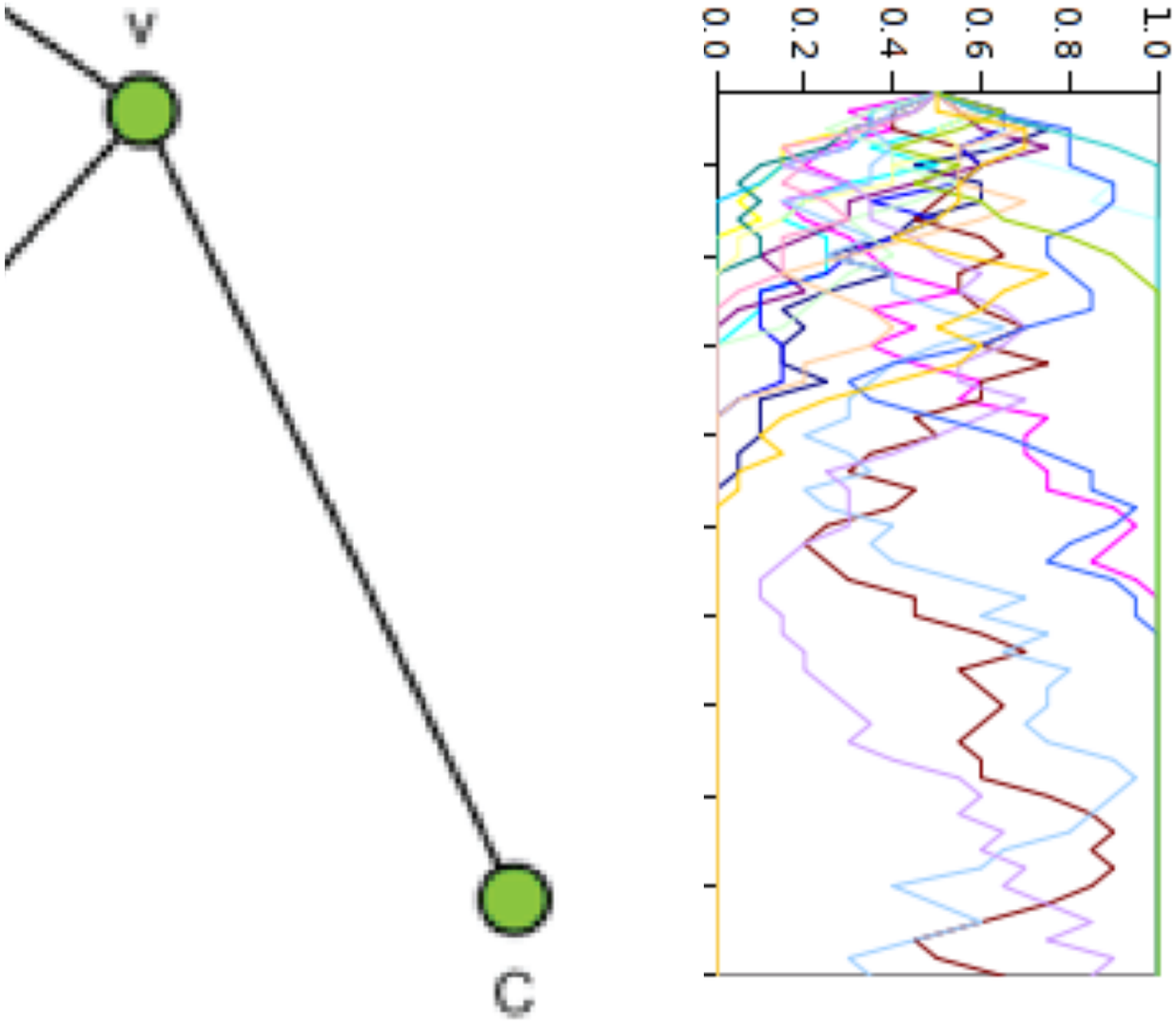


Reasons why a branch may be small:

- Short time-span
- Large population size

How to model genetic drift on a branch?





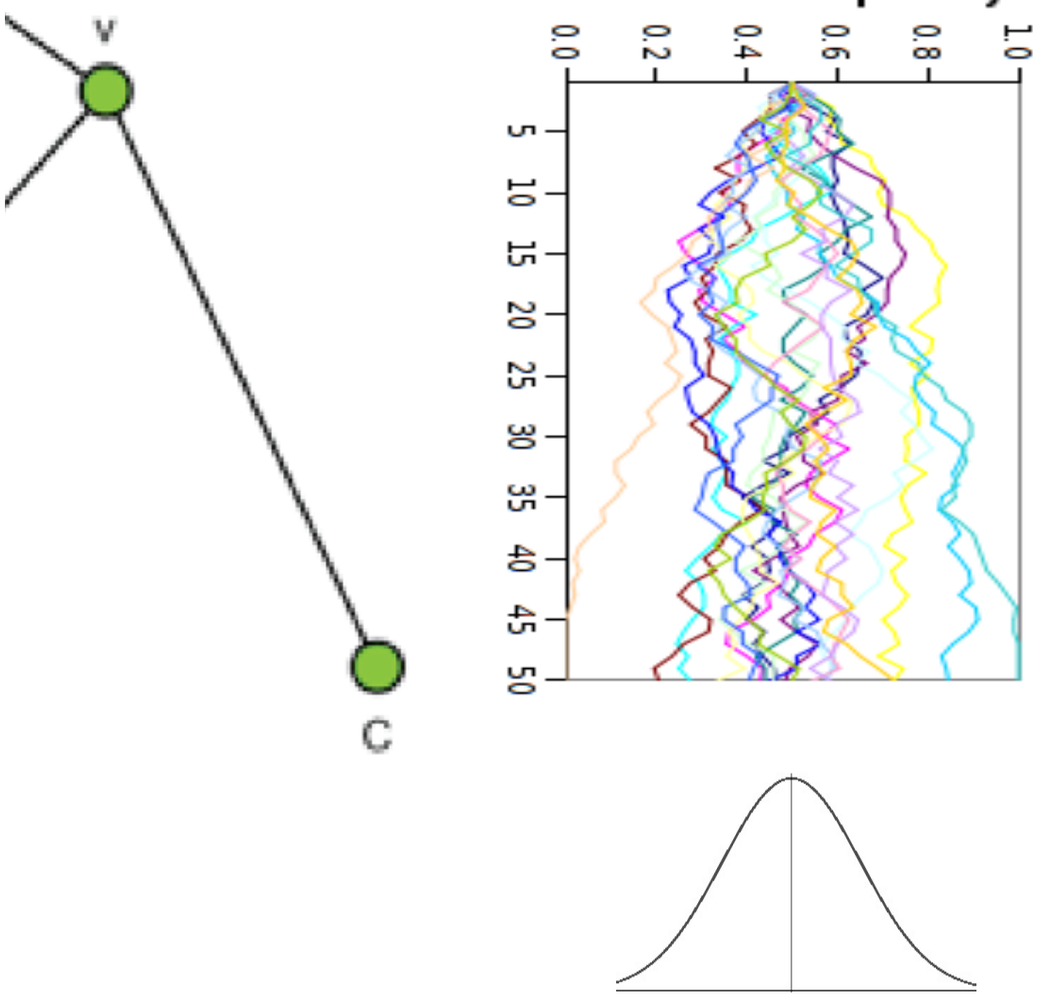
Reasons why a branch may be small:

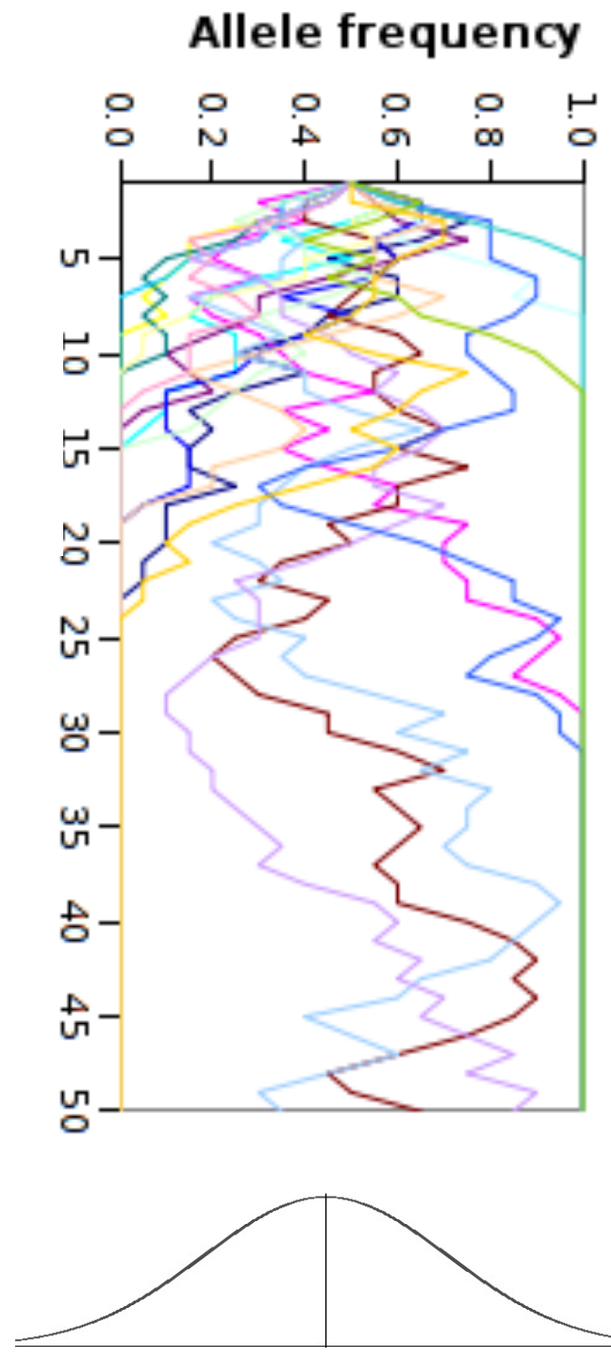
- Short time-span
- Large population size

Reasons why a branch may be large:

- Long time-span
- Small population size

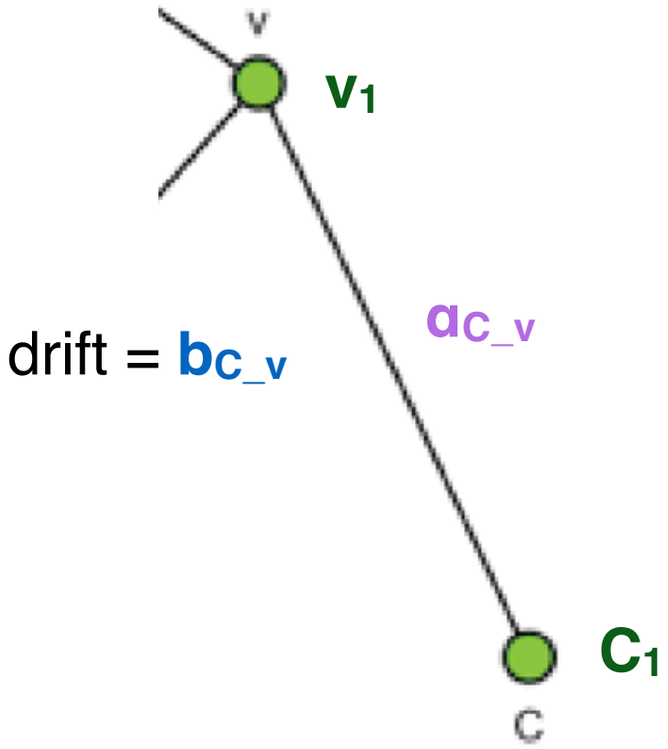
A Normal approximation to the Wright-Fisher model





Nicholson et al. (2002)

A Normal approximation to the Wright-Fisher model



for SNP 1 and branch C_v,

f(C_1 | v_1,b_{C_v}) = Normal(mu, var)

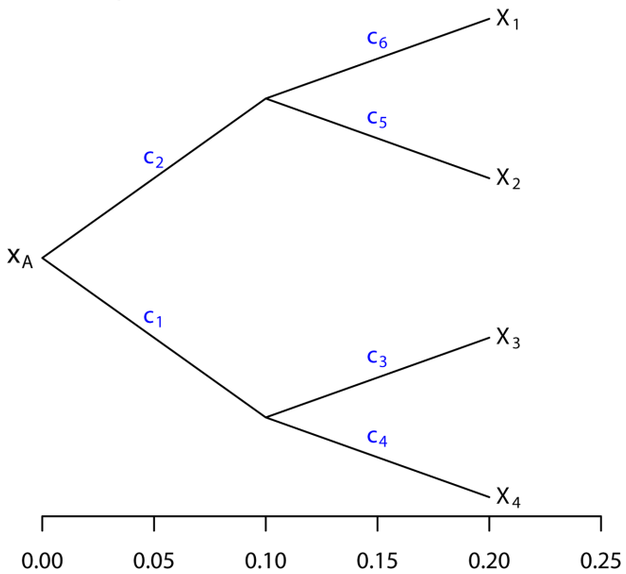
where mu = v_1

and var = b_{C_v}*v_1*(1-v_1)

Nicholson et al. 2002

TreeMix

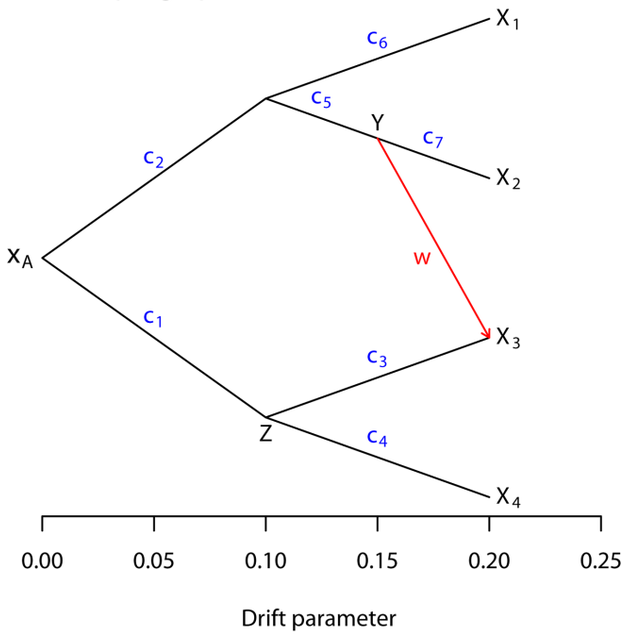
A. Example tree



B. Covariance matrix for tree in A.

	X_1	X_2	X_3	X_4
X_1	$c_2 + c_6$	c_2	0	0
X_2	c_2	$c_2 + c_5$	0	0
X_3	0	0	$c_1 + c_3$	c_1
X_4	0	0	c_1	$c_1 + c_4$

C. Example graph

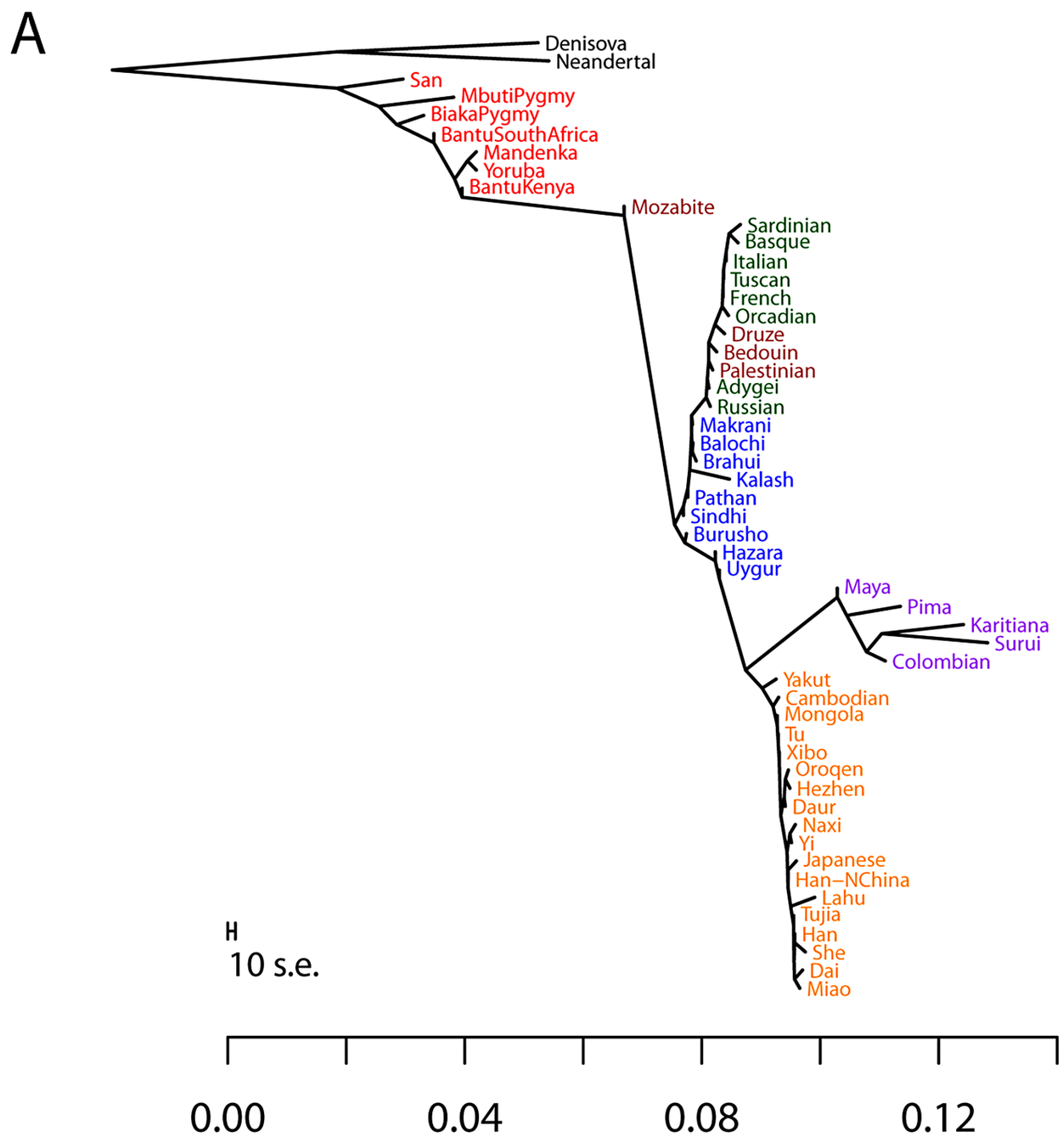


D. Covariance matrix for graph in C.

	X_1	X_2	X_3	X_4
X_1	$c_2 + c_6$	c_2	wc_2	0
X_2	c_2	$c_2 + c_5 + c_7$	$w(c_2 + c_5)$	0
X_3	wc_2	$w(c_2 + c_5)$	$w^2(c_2 + c_5)$ $+(1 - w)^2(c_1 + c_3)$	$(1-w)c_1$
X_4	0	0	$(1-w)c_1$	$c_1 + c_4$

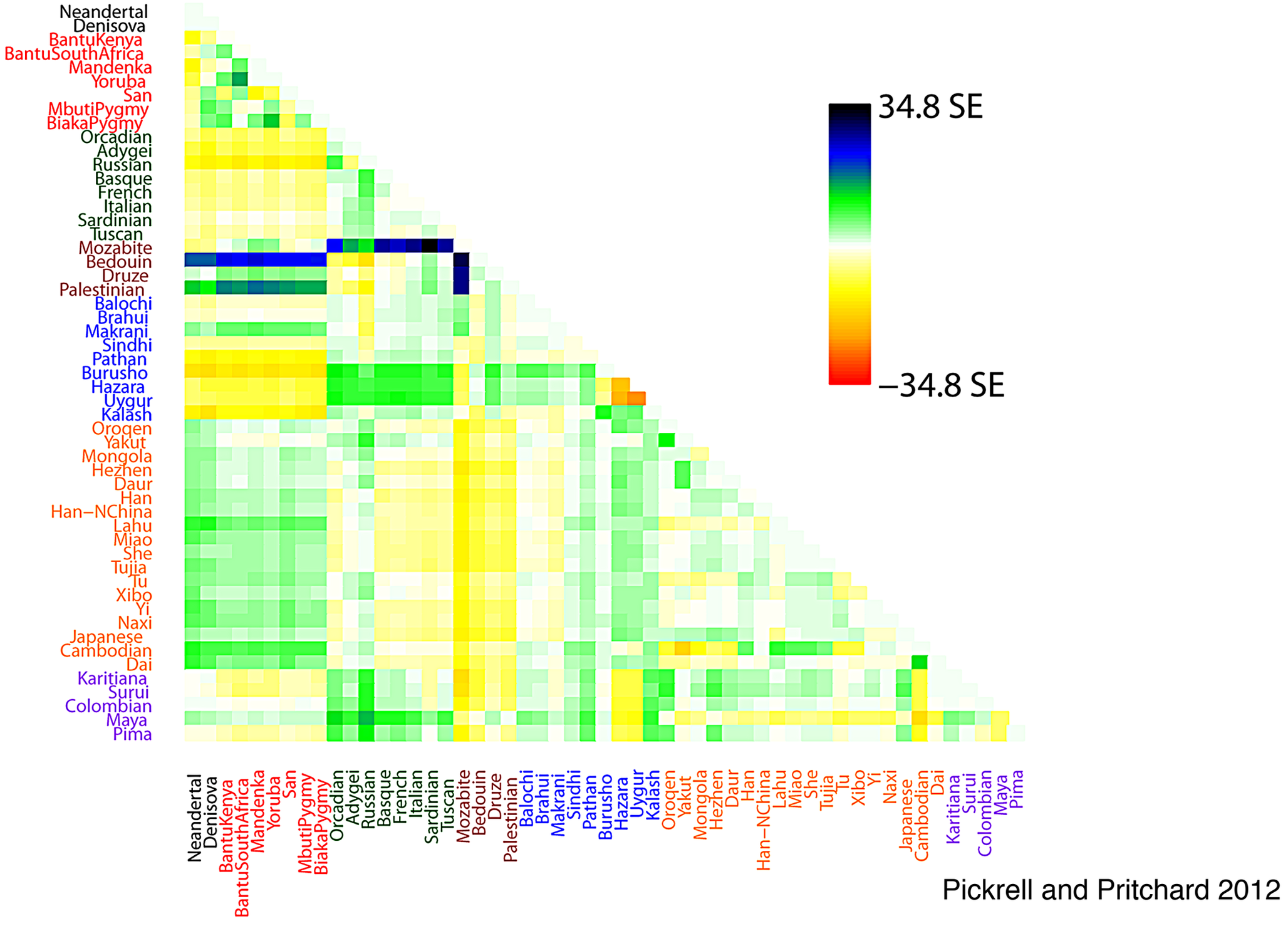
Pickrell and Pritchard 2012

TreeMix



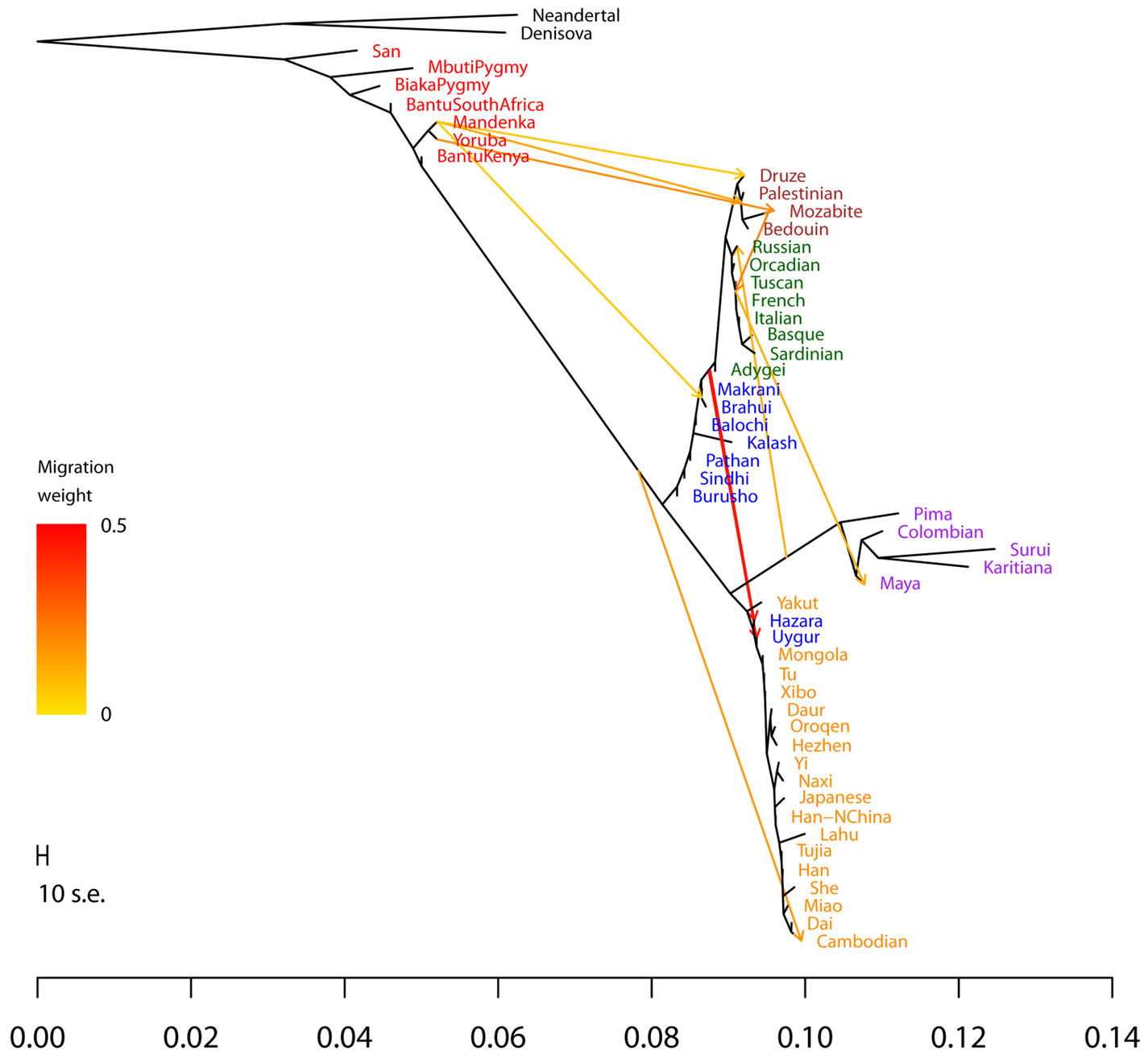
Pickrell and Pritchard 2012

TreeMix



Pickrell and Pritchard 2012

TreeMix

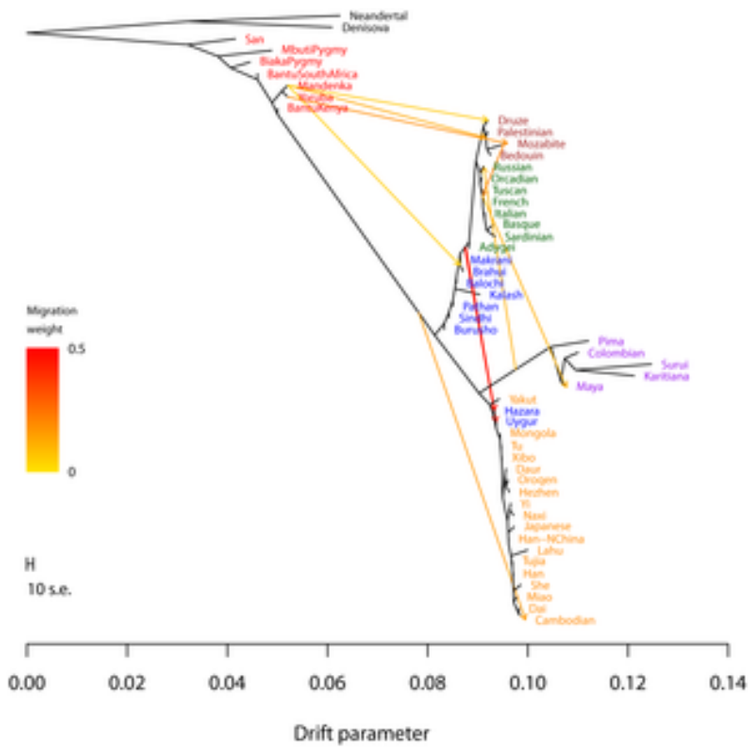


Pickrell and Pritchard 2012

Admixture graph methods

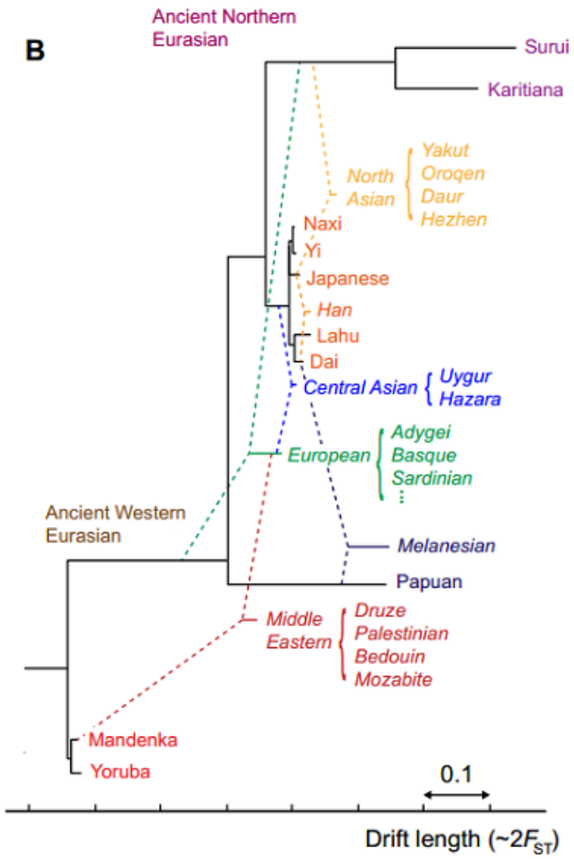
TreeMix

Pickrell and Pritchard 2012



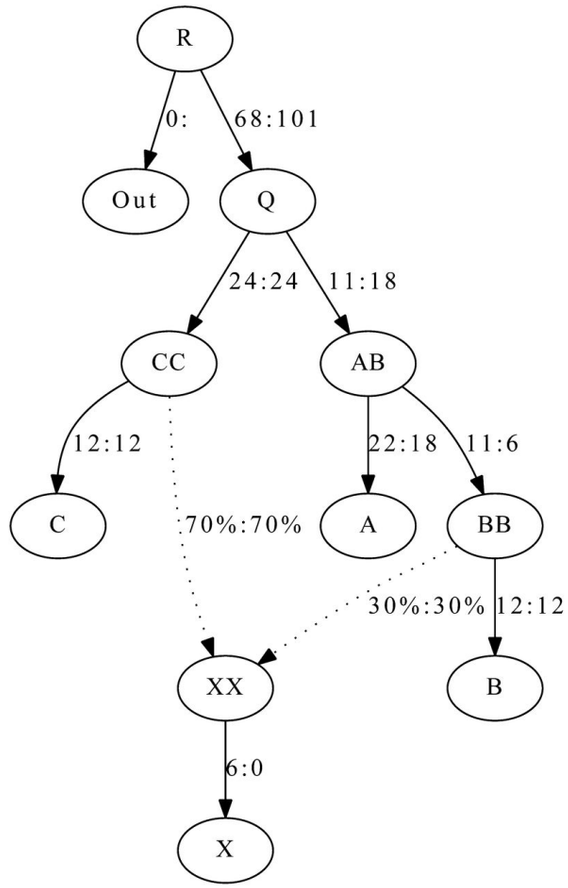
MixMapper

Lipson et al. 2013



qpGraph

Patterson et al. 2012



Less supervised

More supervised

R admixturegraph package

- Works with the list of F4 statistics outputted by AdmixTools
- Can easily build, manipulate and visualize graphs
- Can compute goodness-of-fit measures for particular graphs
- Can explore graph topologies to try to find the best-fitting graph

(a)

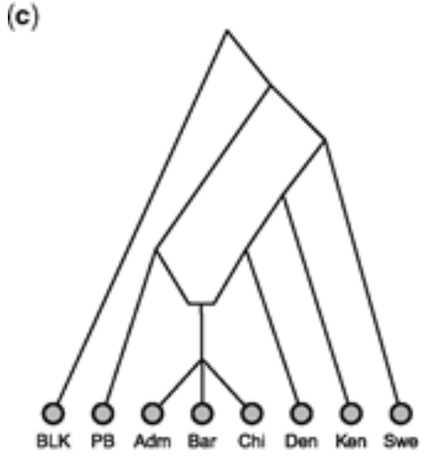
W	X	Y	Z	D	Z.value
BLK	PB	Swe	Adm	0.1258	12.8
BLK	PB	Ken	Adm	0.0685	5.9
BLK	PB	Den	Adm	0.0160	1.3
BLK	PB	Swe	Bar	0.1613	14.7
BLK	PB	Ken	Bar	0.1091	8.9
BLK	PB	Den	Bar	0.0573	4.3
BLK	PB	Swe	Chi	0.1786	17.7
BLK	PB	Ken	Chi	0.1278	11.3
BLK	PB	Den	Chi	0.0777	6.4

(b)

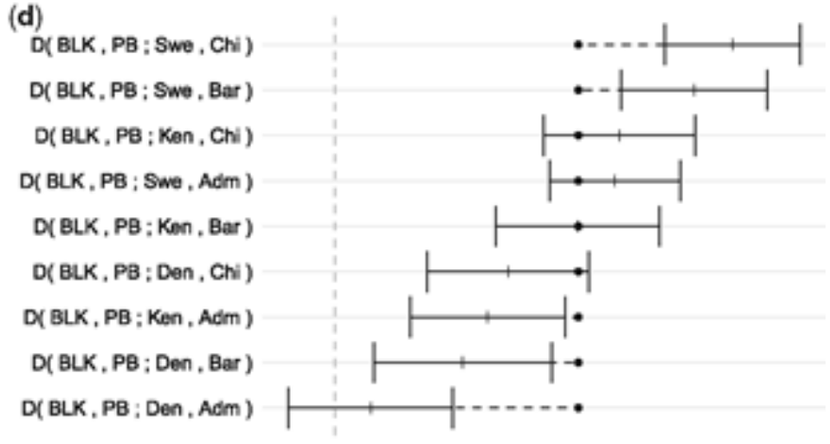
```
## Creating a graph
leaves <- c("BLK", "PB", "Adm", "Bar", "Chi", "Den", "Ken", "Swe")
inner_nodes <- c("Root", "x", "y", "abc", "ABC", "den", "ken", "swe")
edges <- parent_edges(c(edge("BLK", "Root"), edge("x", "Root"),
          edge("PB", "y"), edge("y", "x"), edge("Adm", "ABC"), edge("Bar", "ABC"),
          edge("Chi", "ABC"), edge("ABC", "abc"), admixture_edge("abc", "y", "den", "a"),
          edge("Den", "den"), edge("Ken", "ken"), edge("den", "ken"),
          edge("Swe", "swe"), edge("ken", "swe"), edge("swe", "x")
        ))
graph <- agraph(leaves, inner_nodes, edges)

## Plotting and fitting...
plot(graph)
plot(fit_graph(bears, graph))
```

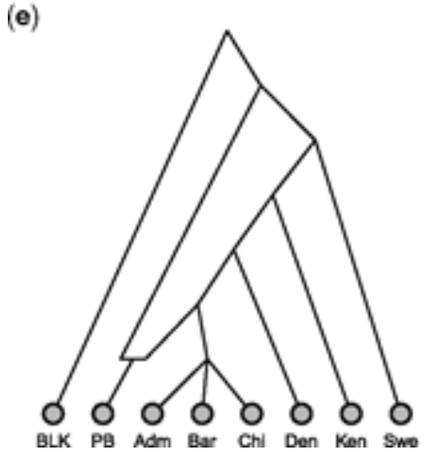
(c)



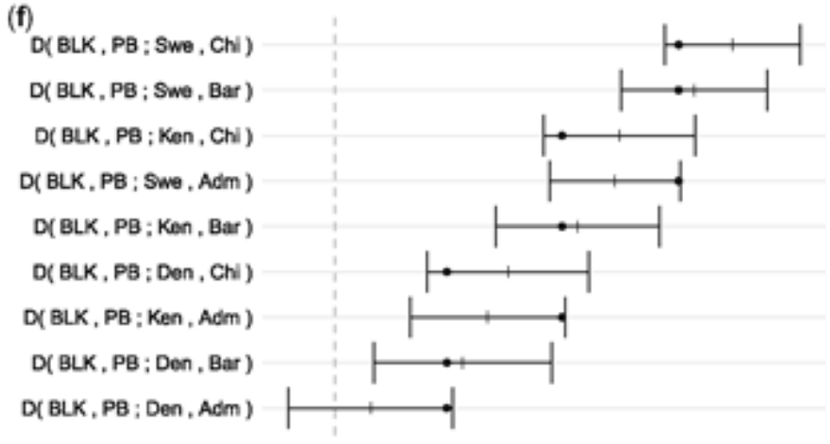
(d)



(e)



(f)

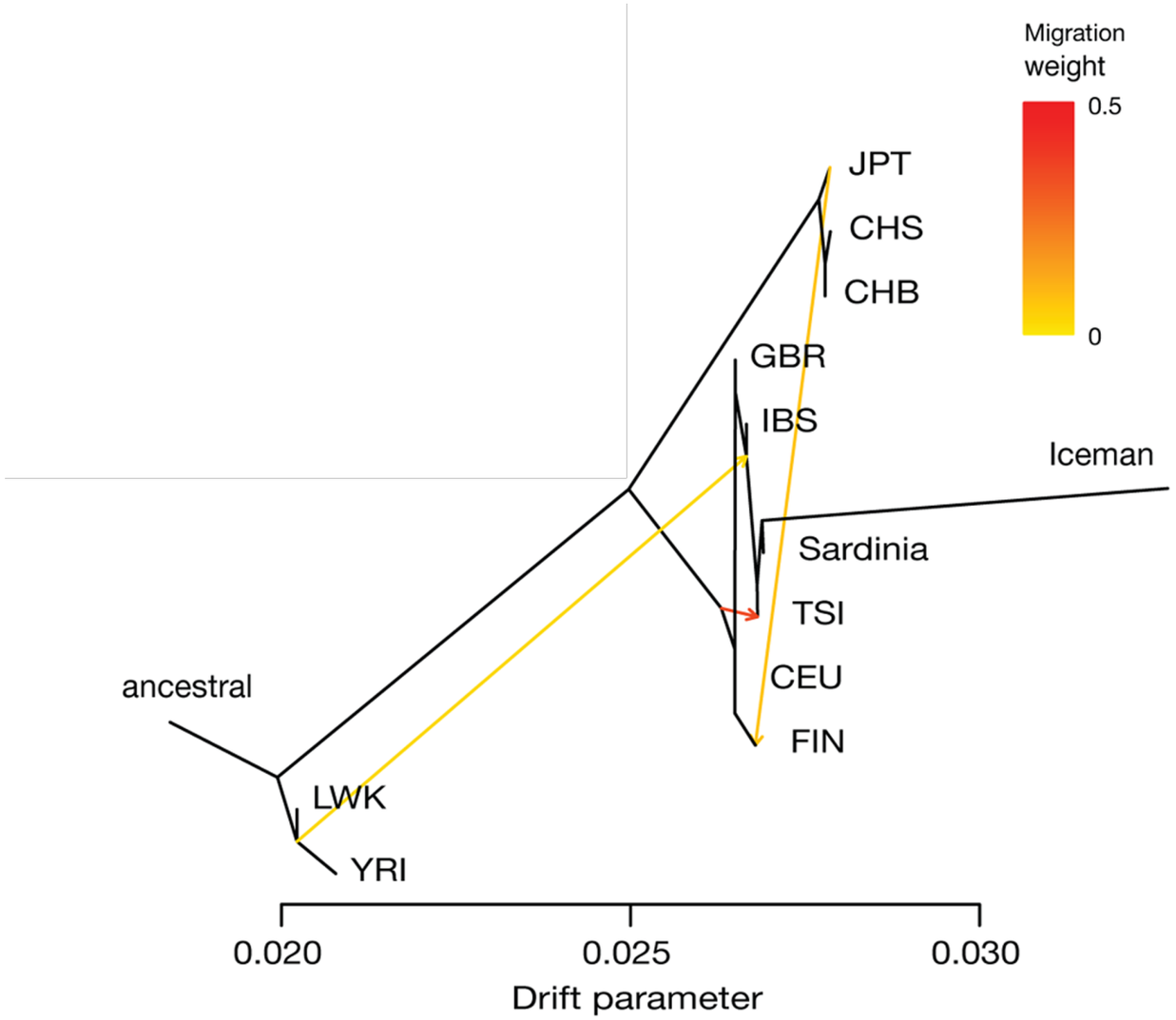


Leppälä et al. 2017

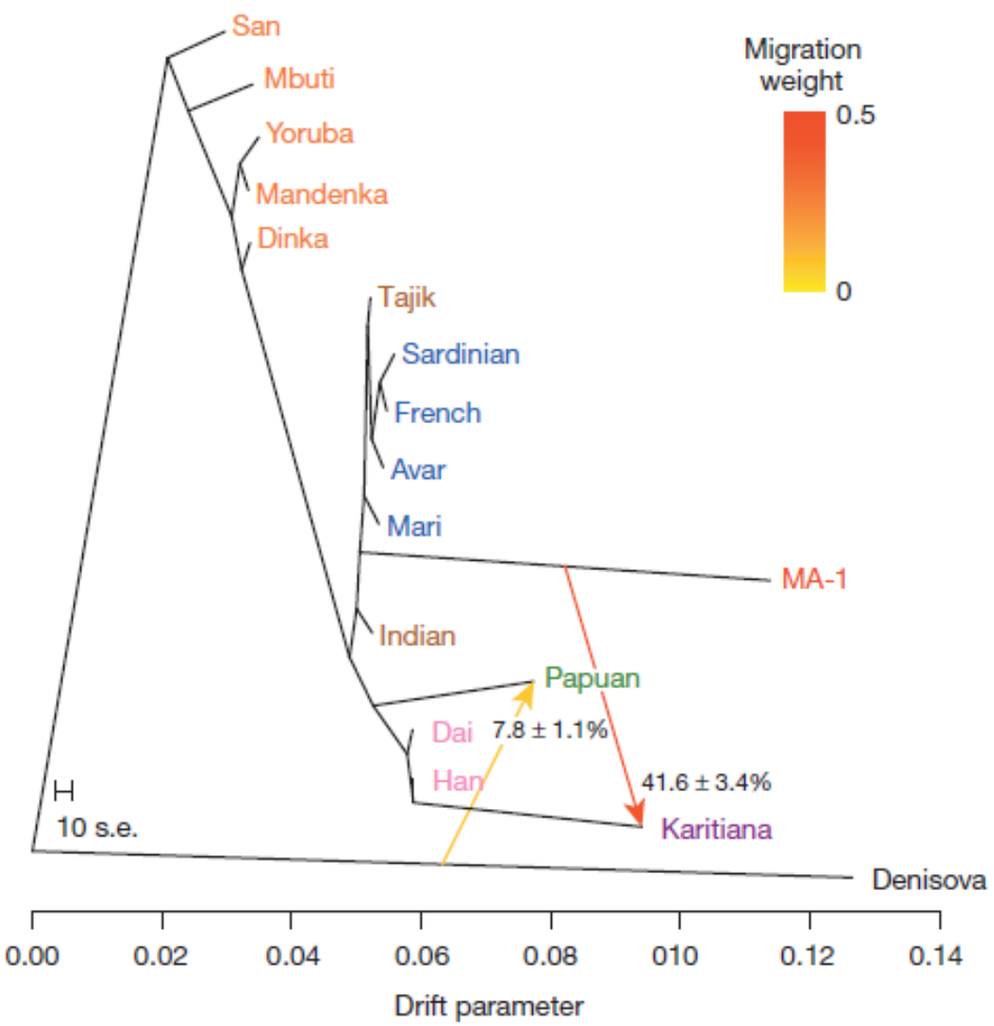
Admixture graph: limitations

- Cannot distinguish Ne from time -> conflated in one parameter: “drift”
- Cannot distinguish a single bout of admixture from multiple bouts or continuous migration
- Cannot model complex demographic processes: population size changes, population growth and decay, population structure, etc.

Caution: ancient DNA damage modeled as drift



Sikora et al. 2014



Raghavan et al. 2013

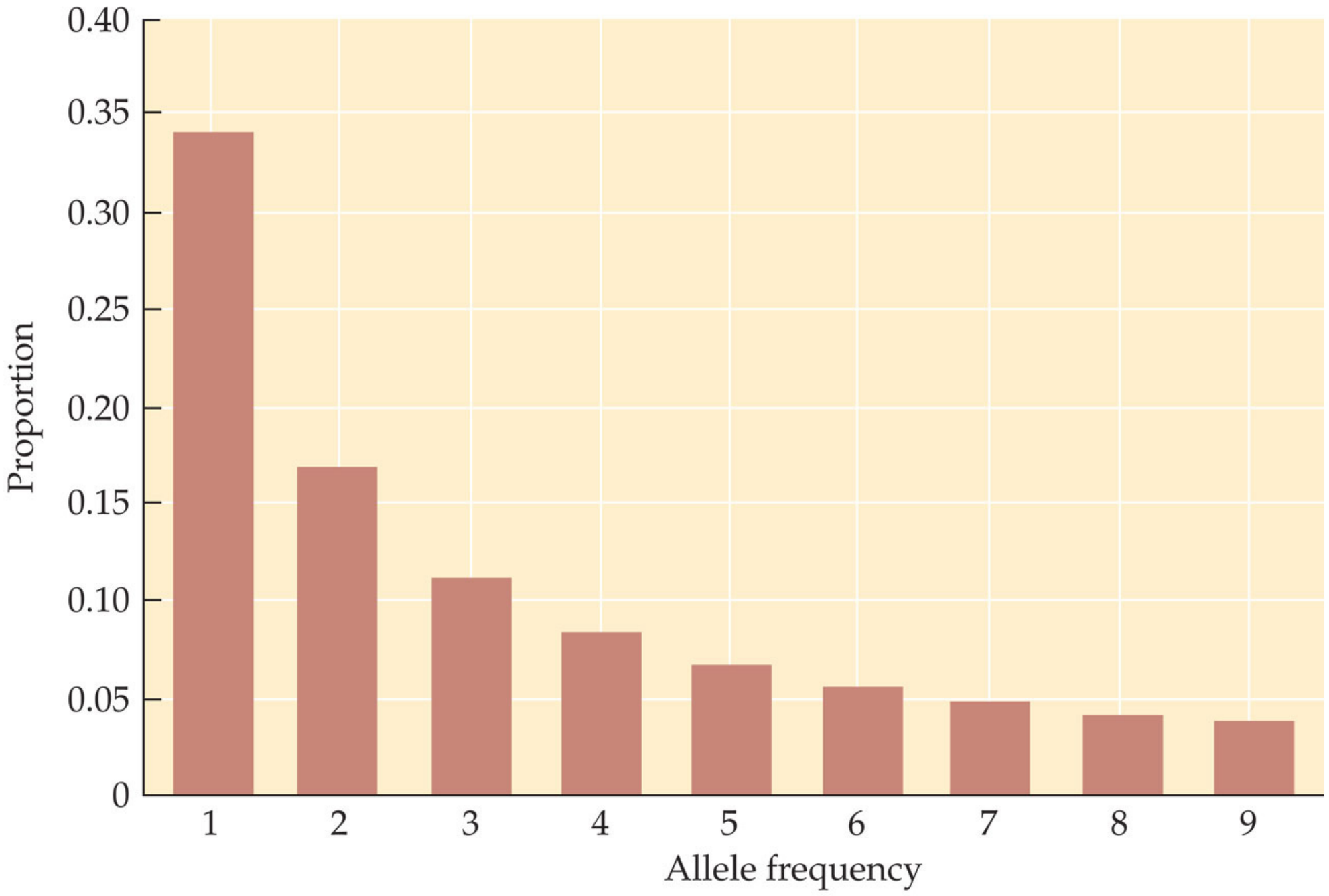
Estimating effective population sizes and generation time

- In general (with some exceptions) one needs whole-genome, non-ascertained, data to be able to properly distinguish effective population sizes from time.

- This is because new mutations occurring along individual population branches are informative about time, but are usually not considered with ascertained data

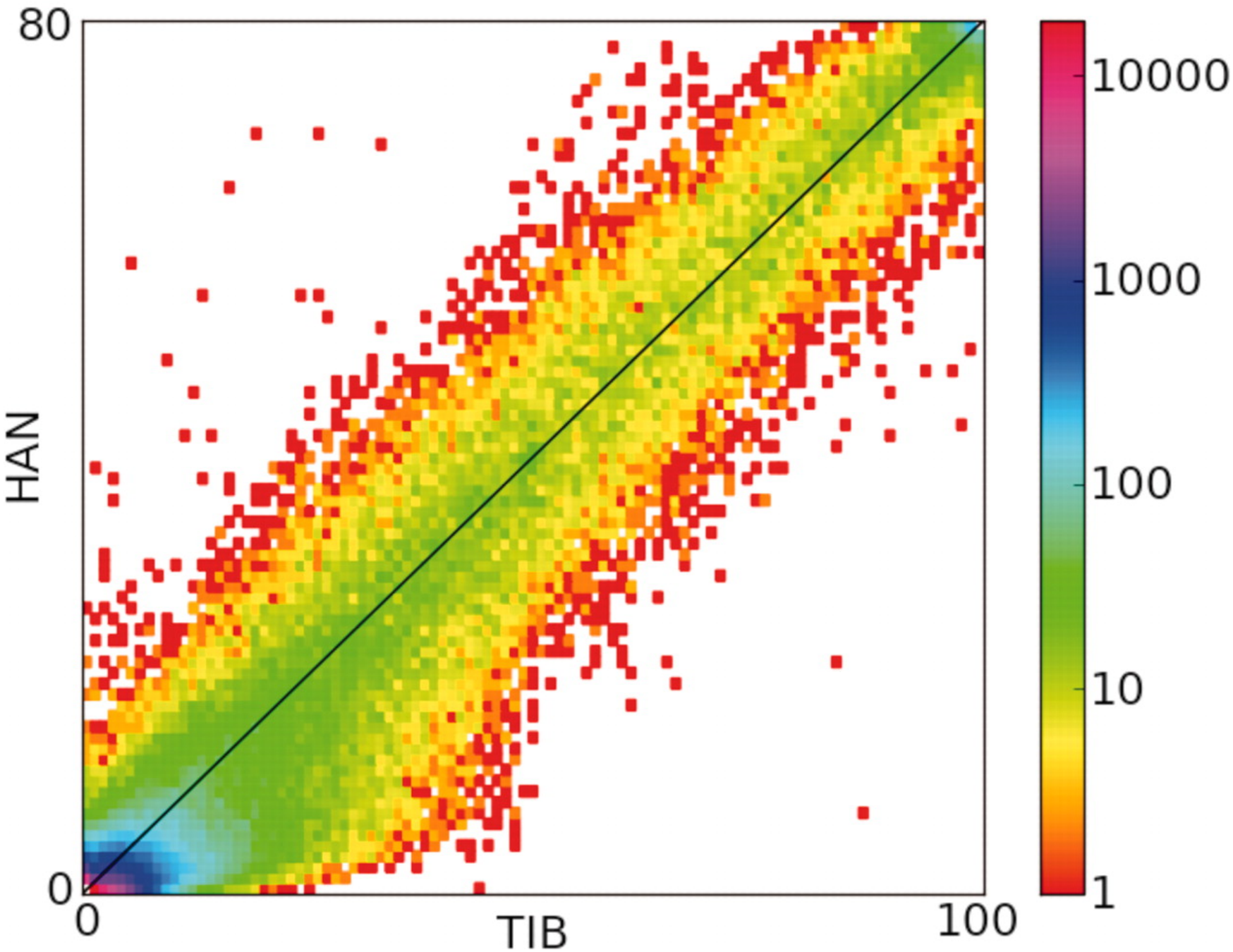
- Most methods try to model the multi-population site-frequency spectrum (SFS) using a **likelihood function** that relates the SFS to a particular demographic history

1-population SFS

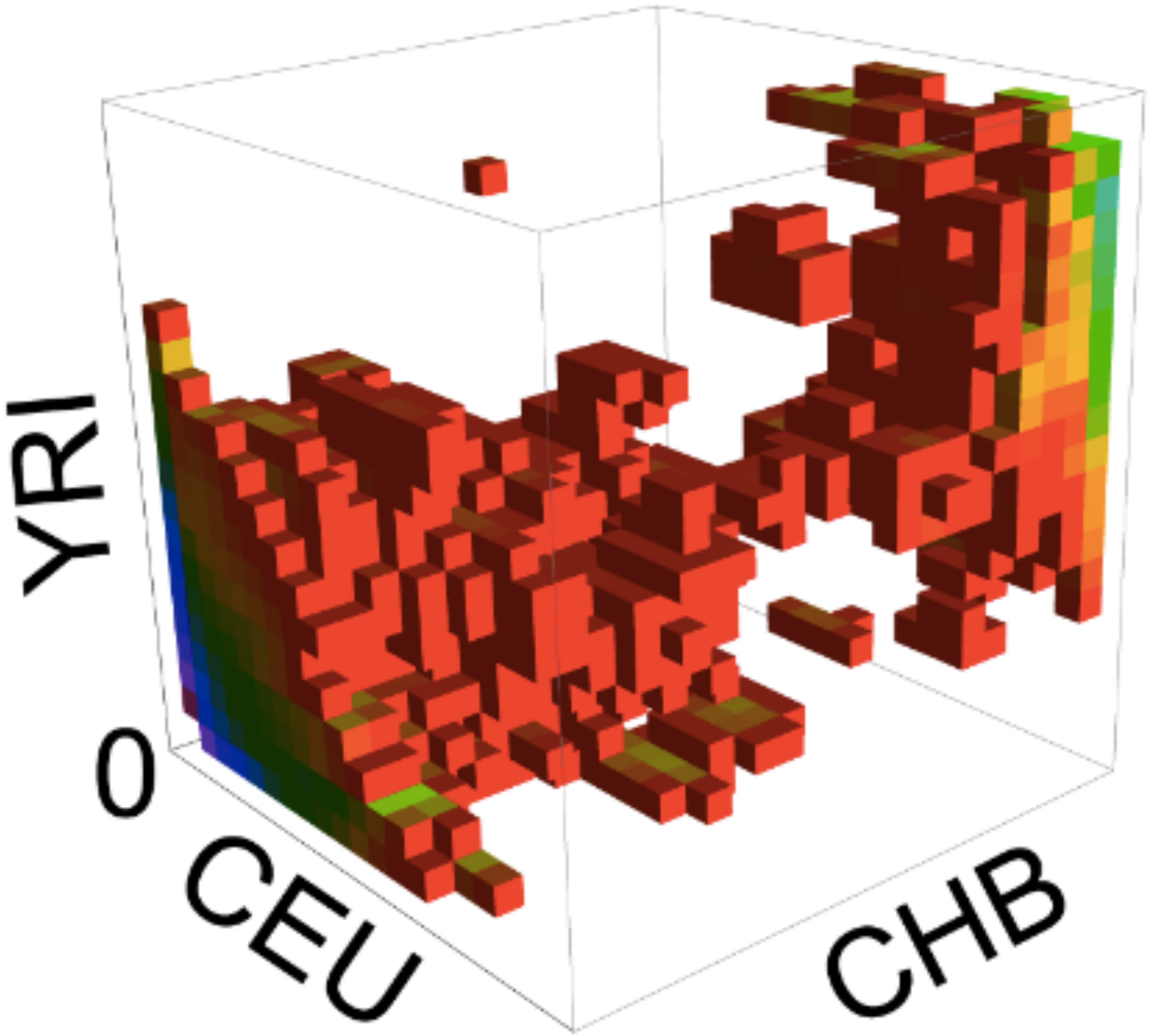


***INTRODUCTION TO POPULATION GENETICS,* Figure 3.9**

2-population SFS



3-population SFS



Explicit likelihood methods: the W-F diffusion equation

- The Wright-Fisher diffusion equation is a **continuos-time, continuous-space approximation** to the Wright-Fisher model

- Unlike the coalescent, this is a **forwards-in-time** approximation

- The solution to this equation gives the entries of the SFS under a particular demographic model

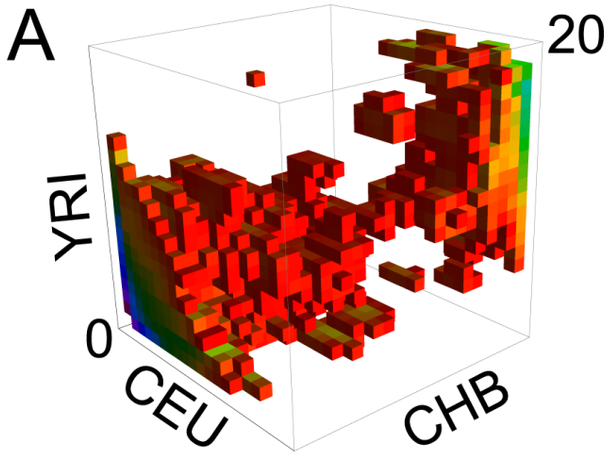
- Problem: **solution is hard for complex models**

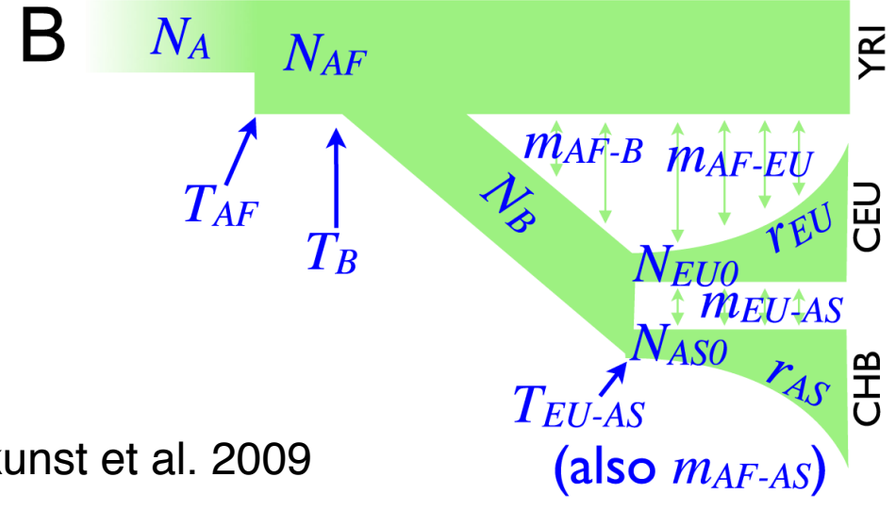
$$\frac{\partial}{\partial \tau}\phi = \frac{1}{2}\sum_{i=1,2,\ldots,P}\frac{\partial^2}{\partial^2 x_i}\frac{x_i(1-x_i)}{v_i}\phi$$
$$-\sum_{i=1,2,\ldots,P}\frac{\partial}{\partial x_i}\left(\gamma_i x_i(1-x_i)+\sum_{j=1,2,\ldots,P}M_{i\leftarrow j}(x_j-x_i)\right)\phi.$$

Explicit likelihood methods: dadi

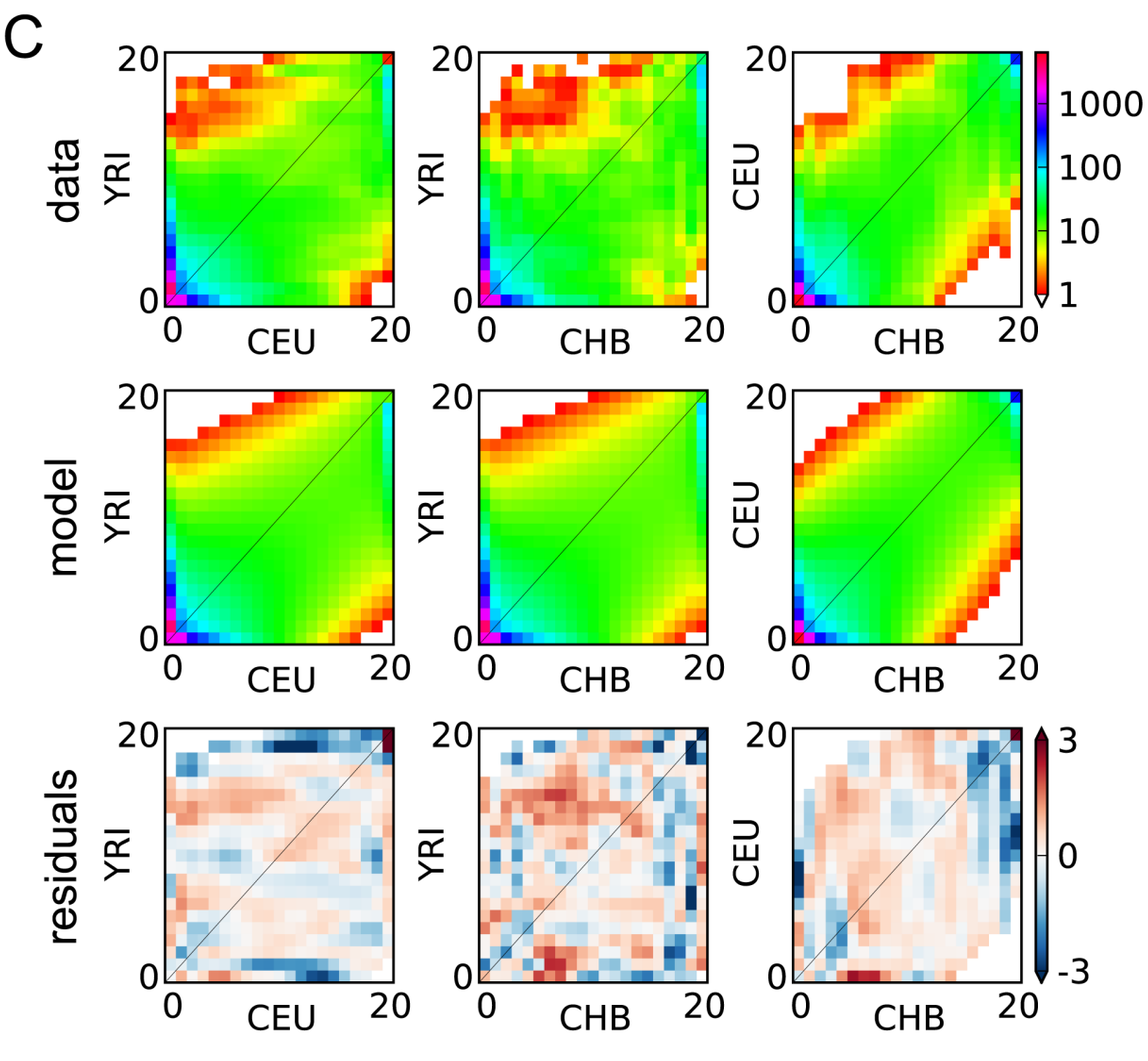
- Numerical solution to the W-F equation

- Can only handle 3 populations co-existing at the same time





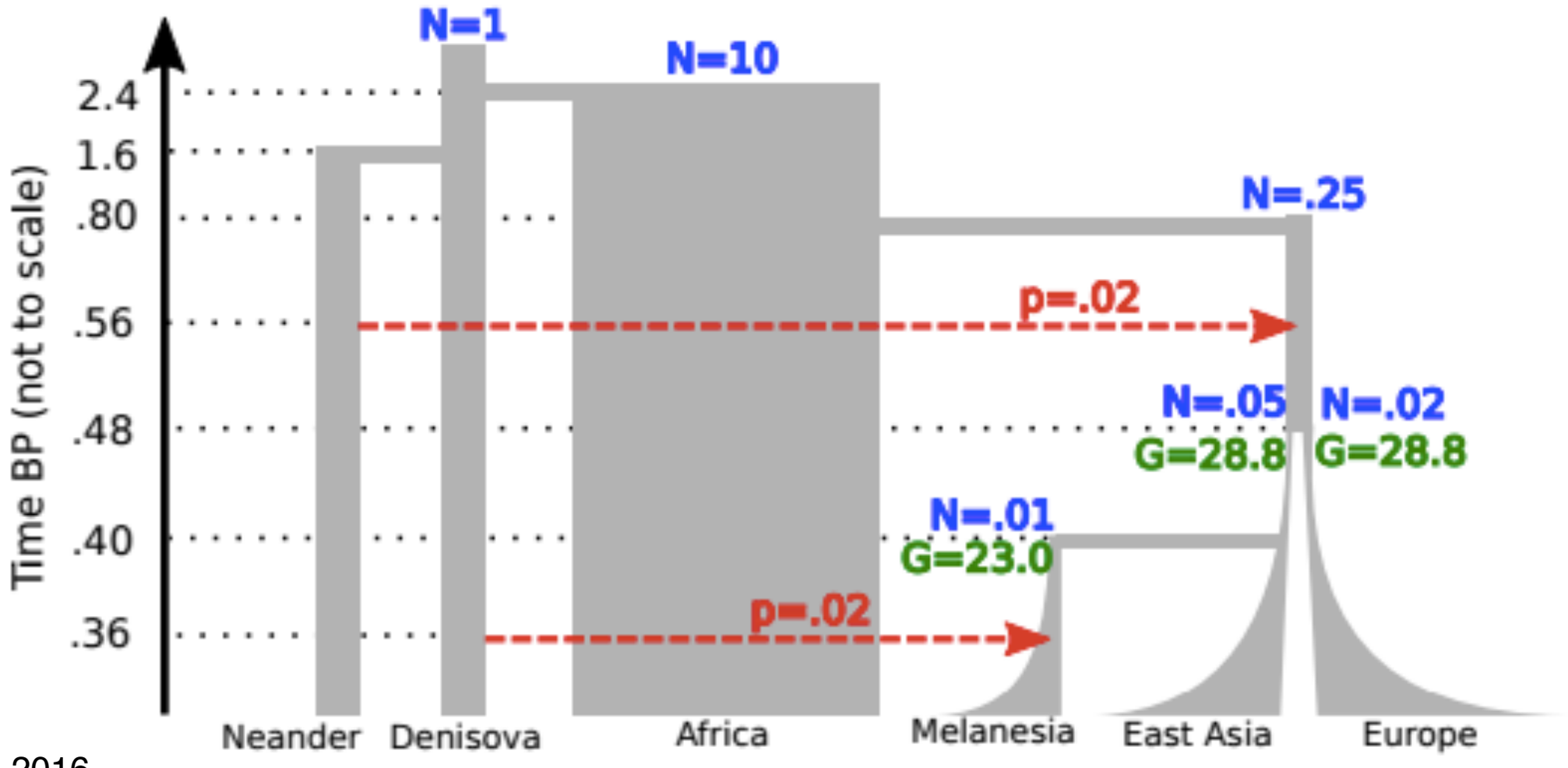
Gutenkunst et al. 2009



Explicit likelihood methods: momi

- Analytical solution to the W-F equation (via a graphical model)

- Can handle more than 3 populations



Kamm et al. 2016

Approximate Bayesian Computation (ABC)

- In many cases, there **may not be a known likelihood function** for a particular demographic model, or if there is, it may be **hard to solve**

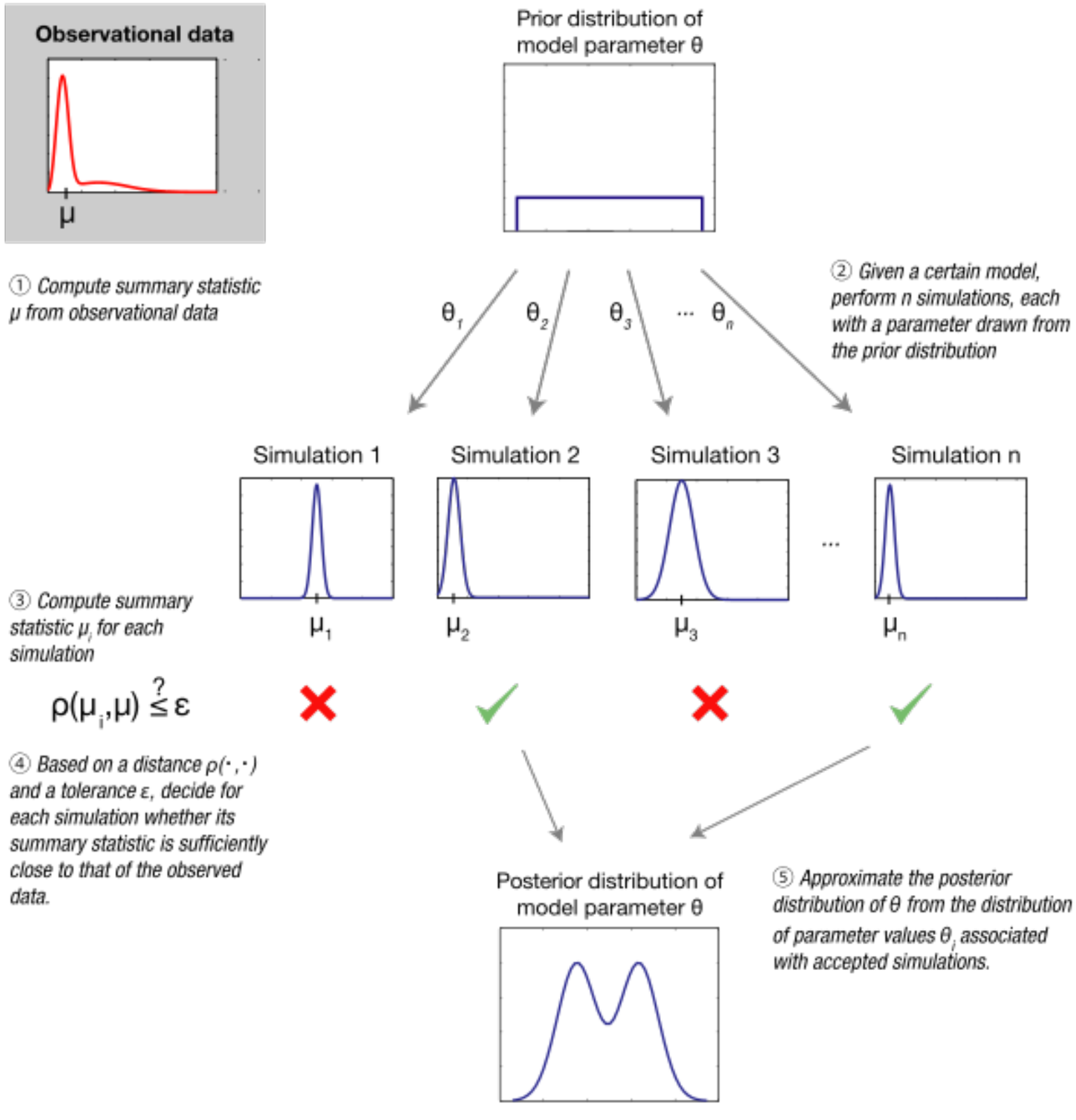
- One can then use **simulations** to perform **likelihood-free** inference

- Essentially, we can simulate a large set of different histories, and then try to find the histories that are **closest** to our data

- Problem: how do we measure “closeness” to our data?

- We can use **summary statistics** that can be computed on both the simulations and our data.

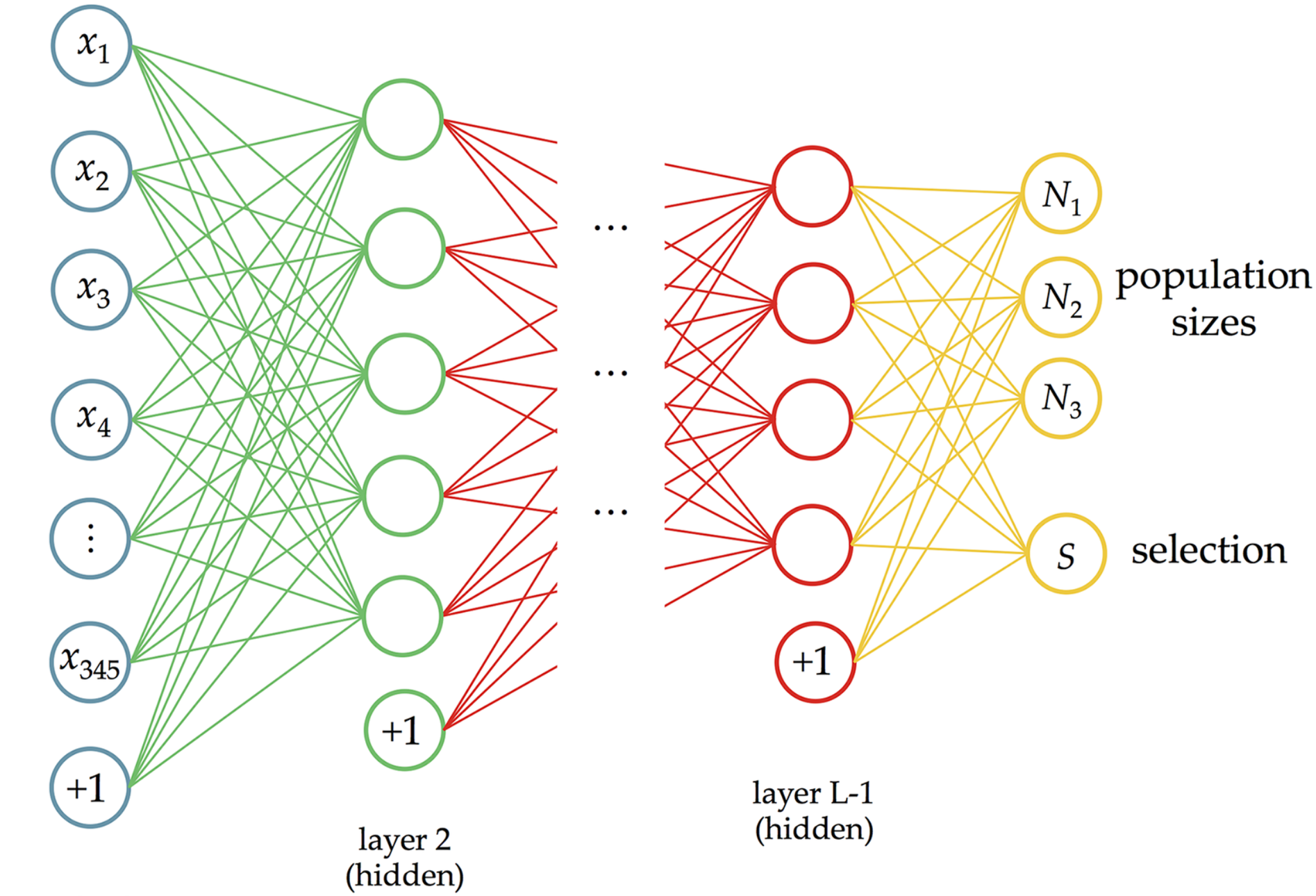
Approximate Bayesian Computation (ABC)



ABC: limitations

- Very slow: need to generate many simulations
- Little guarantee that parameter space has been extensively explored
- Often hard to find the correct summary statistics that are informative of particular demographic parameters
- We can’t use too many summary statistics or we run into the curse of dimensionality (alternative: use linear combinations of statistics)

Deep Learning: a new frontier in pop gen?



Sheehan et al. 2016

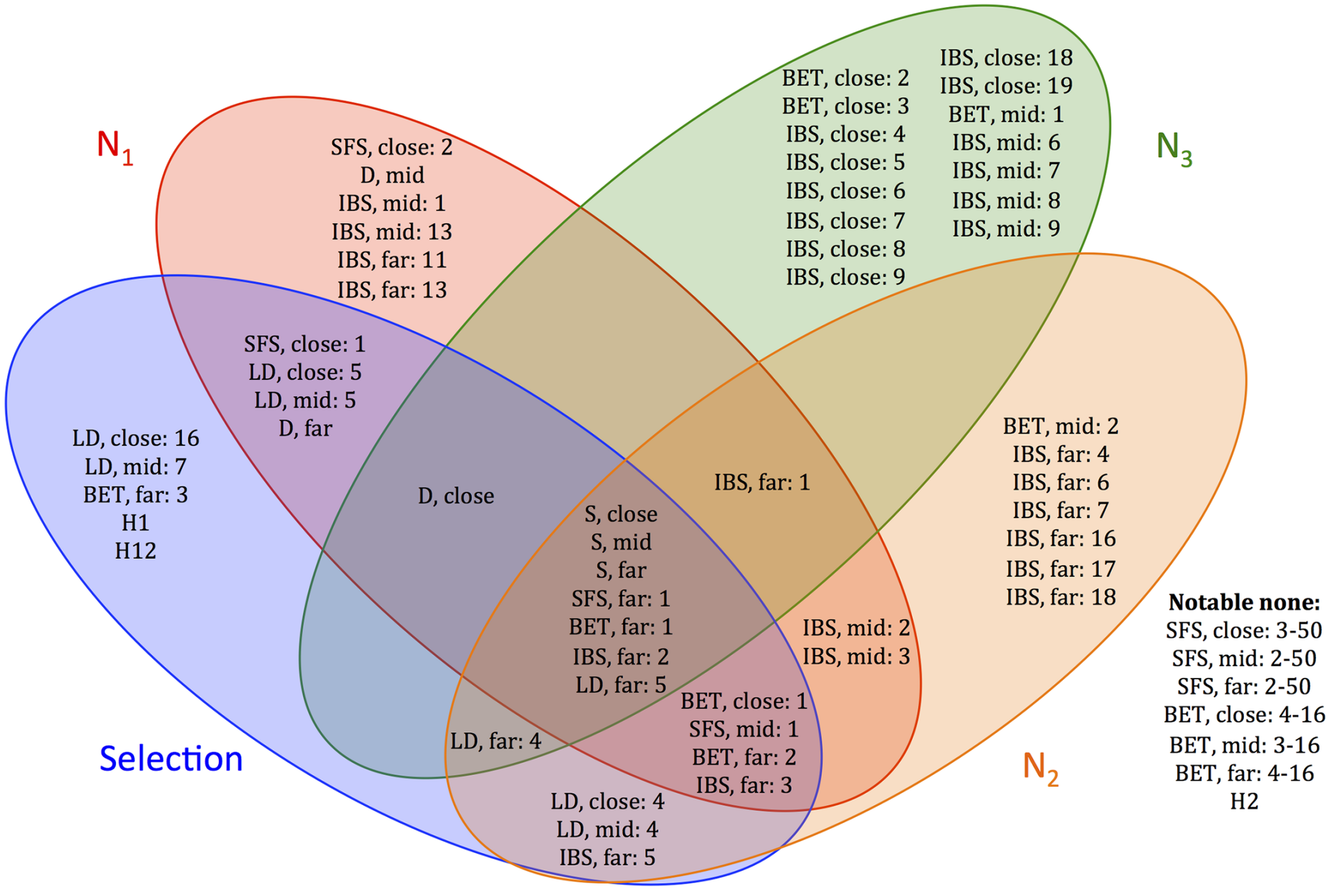
Deep Learning outperforms ABC

Dataset	Method	N_1 error	N_2 error	N_3 error
Full summary statistics	ABCtoolbox	0.062	0.043	0.218
	Deep learning	0.044	0.028	0.221
Filtered summary statistics	ABCtoolbox	0.161	0.035	0.311
	Deep learning	0.065	0.055	0.319

doi:10.1371/journal.pcbi.1004845.t008

Sheehan et al. 2016

Deep Learning: informative statistics



Sheehan et al. 2016