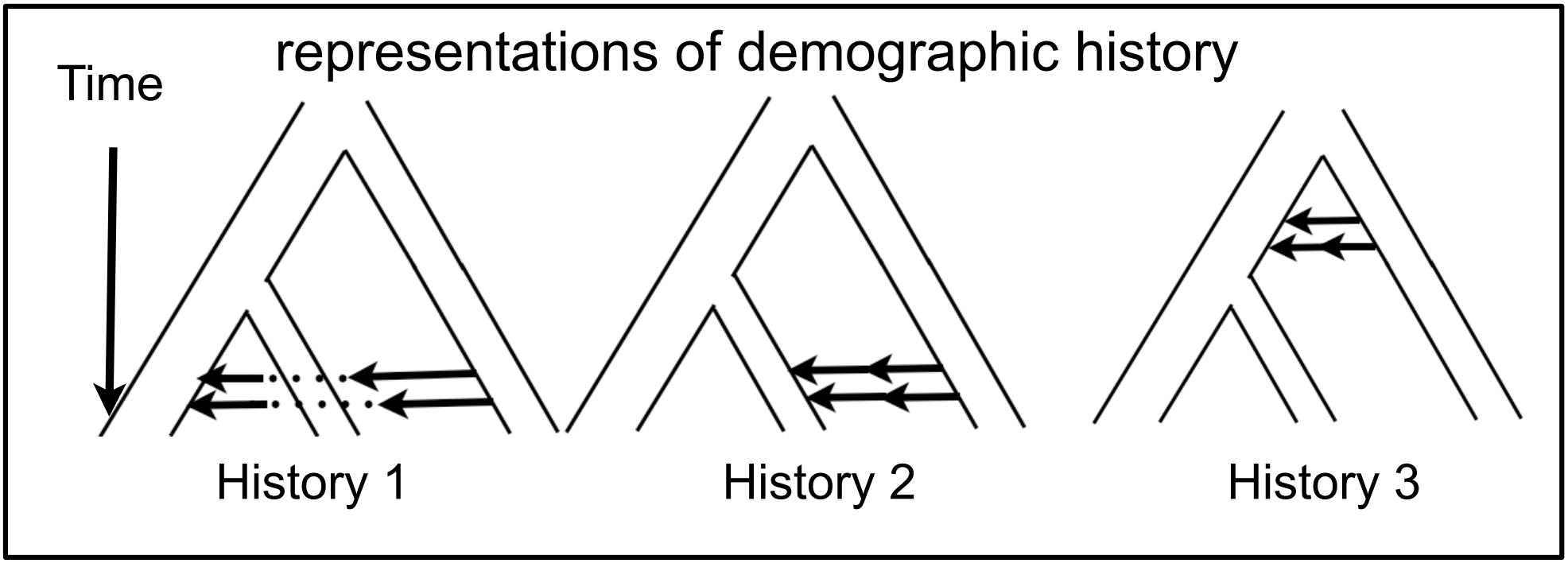
Lecture 7:

Population History and demographic Inference 2

outline

- Summary Statistics, Parameters and Models
- Coalescent Simulations
- Gene Trees vs Species Trees
- Felsenstein Equation
- MCMC, HMM and *maybe* more ABC methods

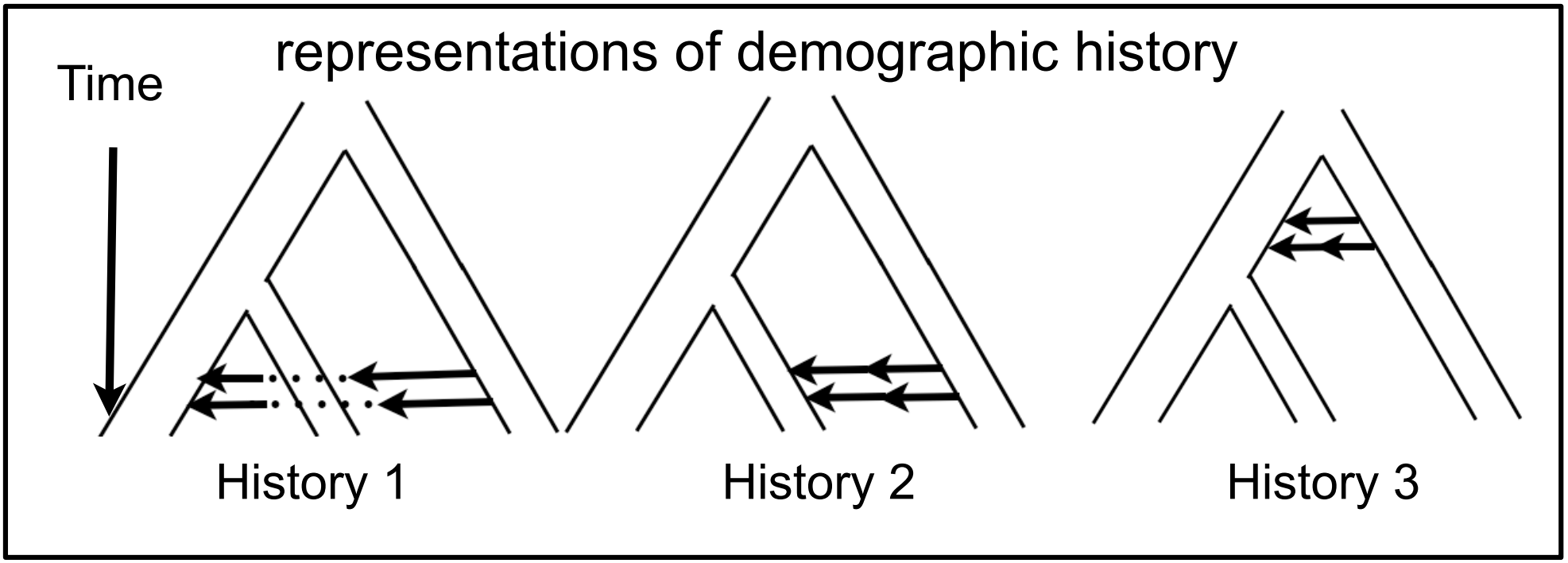
Population History and demographic Inference



Coalescent Theory allows for statistical inference

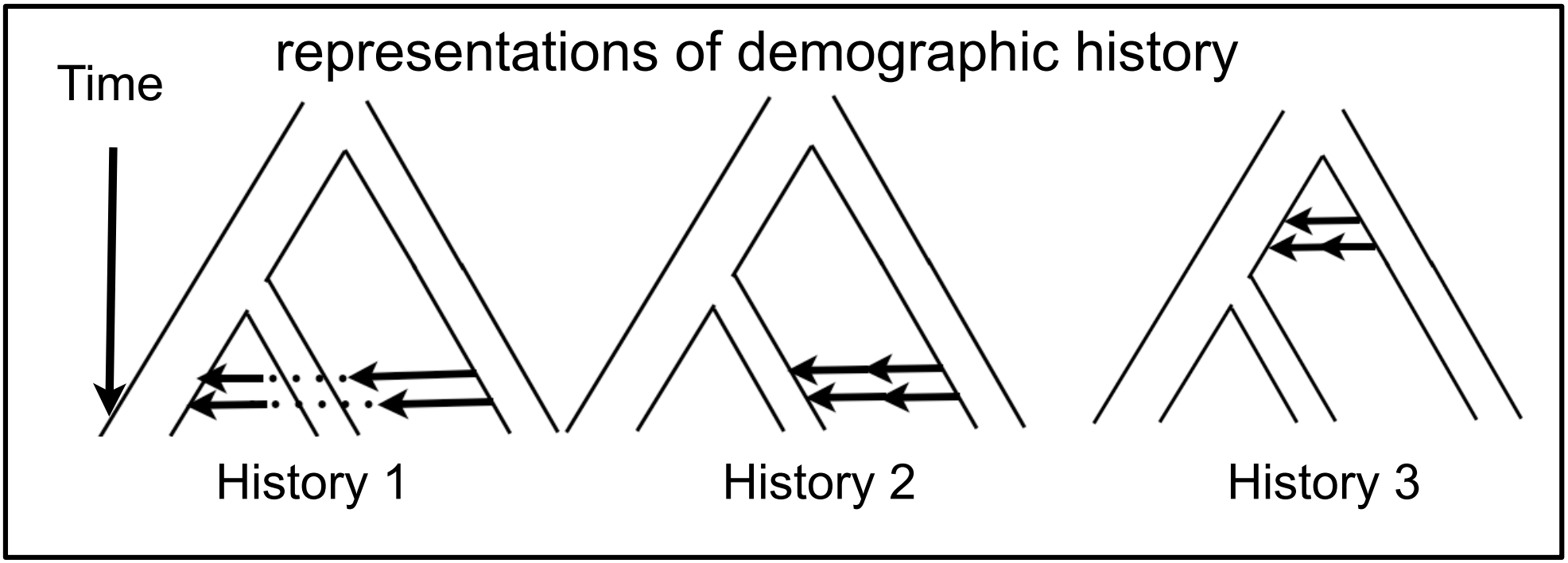
Testing historical demographic models

Population History and demographic Inference



Population Genetic **data**

Population History and demographic Inference

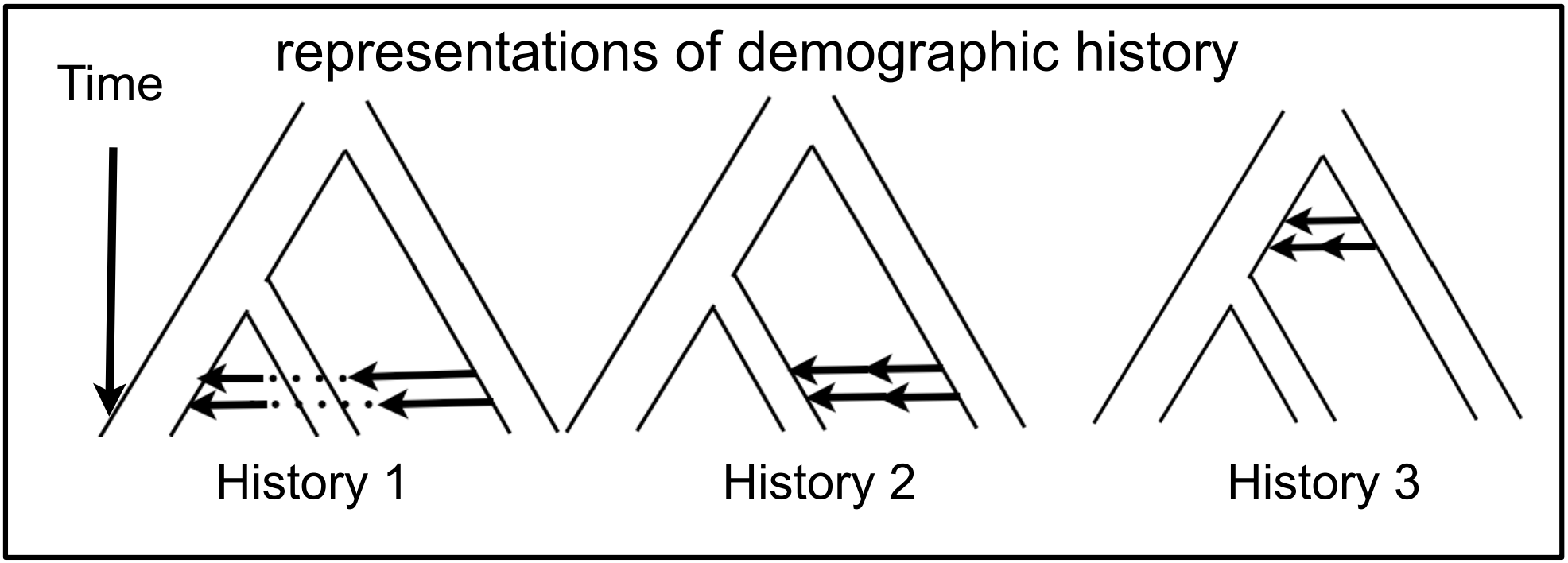


Population Genetic **data**

demographic models

parameters given model (s)

Population History and demographic Inference



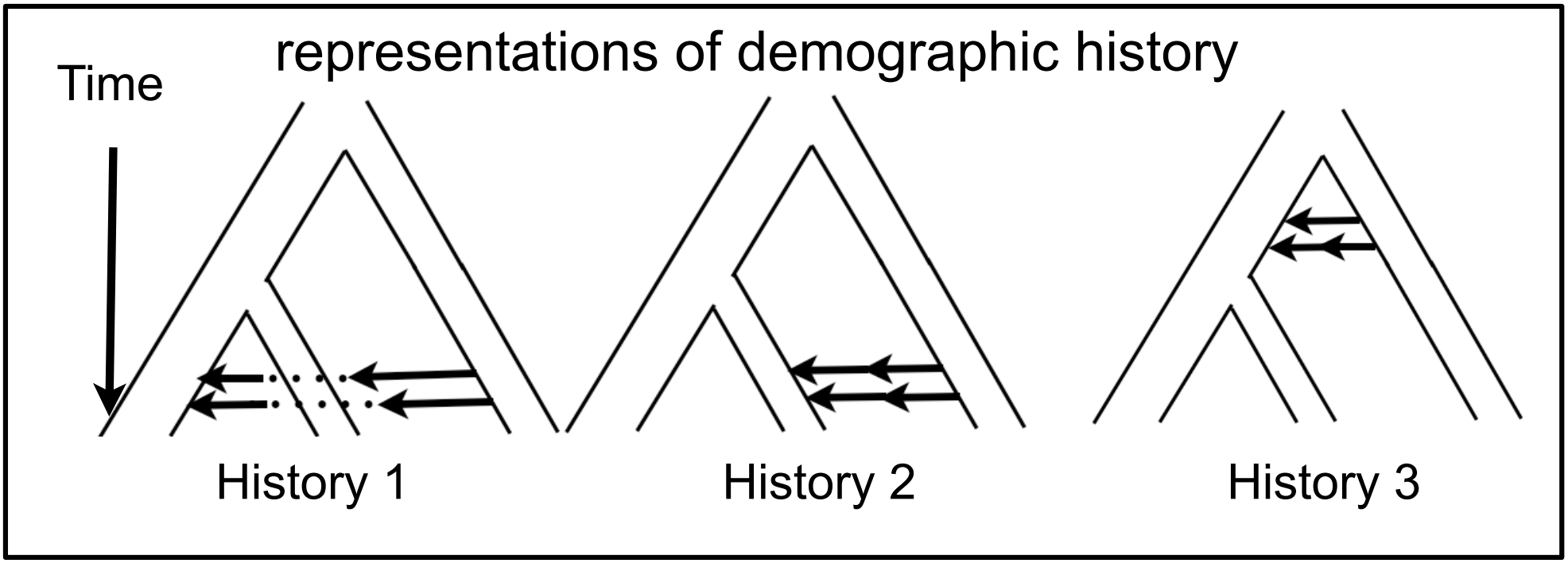
Population Genetic **data**

generate

demographic models

parameters given model (s)

Population History and demographic Inference



Population Genetic **data**

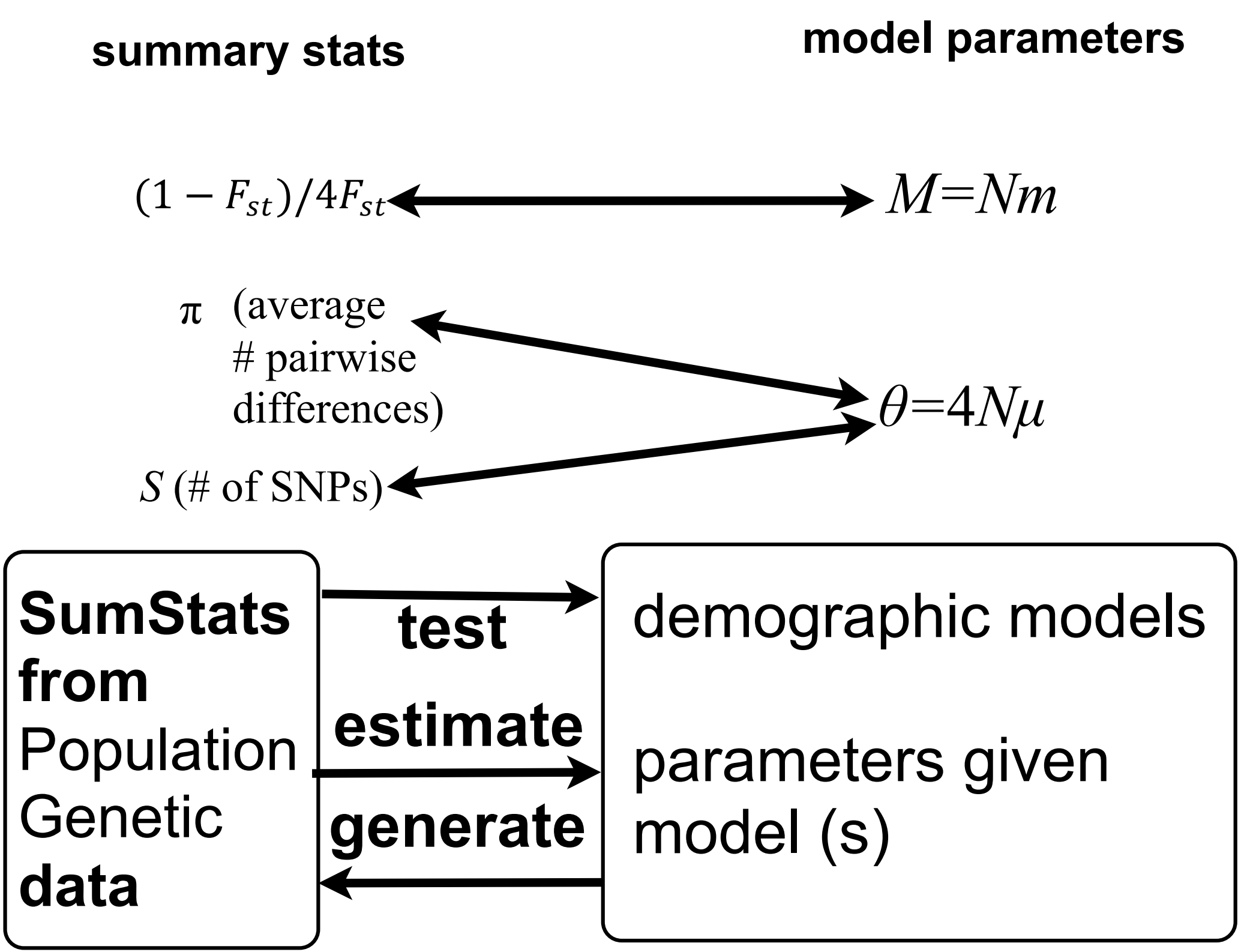
test

estimate

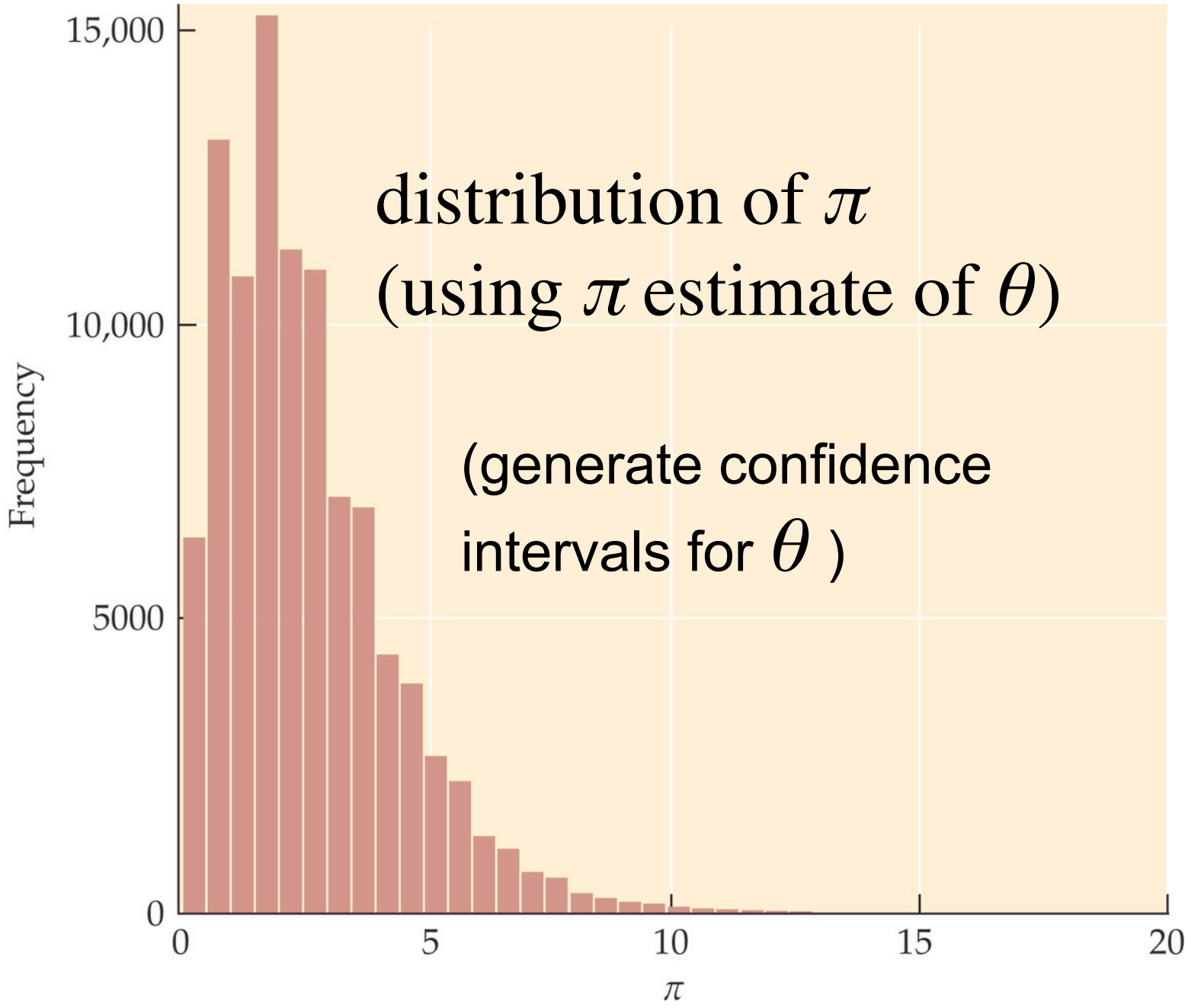
generate

demographic models

parameters given model (s)


summary stats
model parameters
$(1 - F_{st})/4F_{st}$
$M=Nm$
π (average # pairwise differences)
$\theta=4N\mu$
S (# of SNPs)
SumStats from Population Genetic data
test
estimate
generate
demographic models
parameters given model (s)

Simulating π in 100,000 coalescence simulations under the standard coalescence model with infinite sites mutation and θ = 2.6



***INTRODUCTION TO POPULATION GENETICS*, Figure 5.3**

The Tajima's Estimator

To estimate θ, we can use the assumption of an infinite sites model and the expected number of mutations separating two individuals. An **estimate** is an educated guess of the true value of a parameter based on information obtained from data. In our case, the parameter is θ and the data are the DNA sequences shown above. The data can be summarized in different ways. A popular way of summarizing DNA sequence data is in terms of the **average number of pairwise differences**, or π. The value of π is obtained by calculating the number of sites in which each pair of sequences differ, and then taking the average among all pairs of sequences. We can write this as

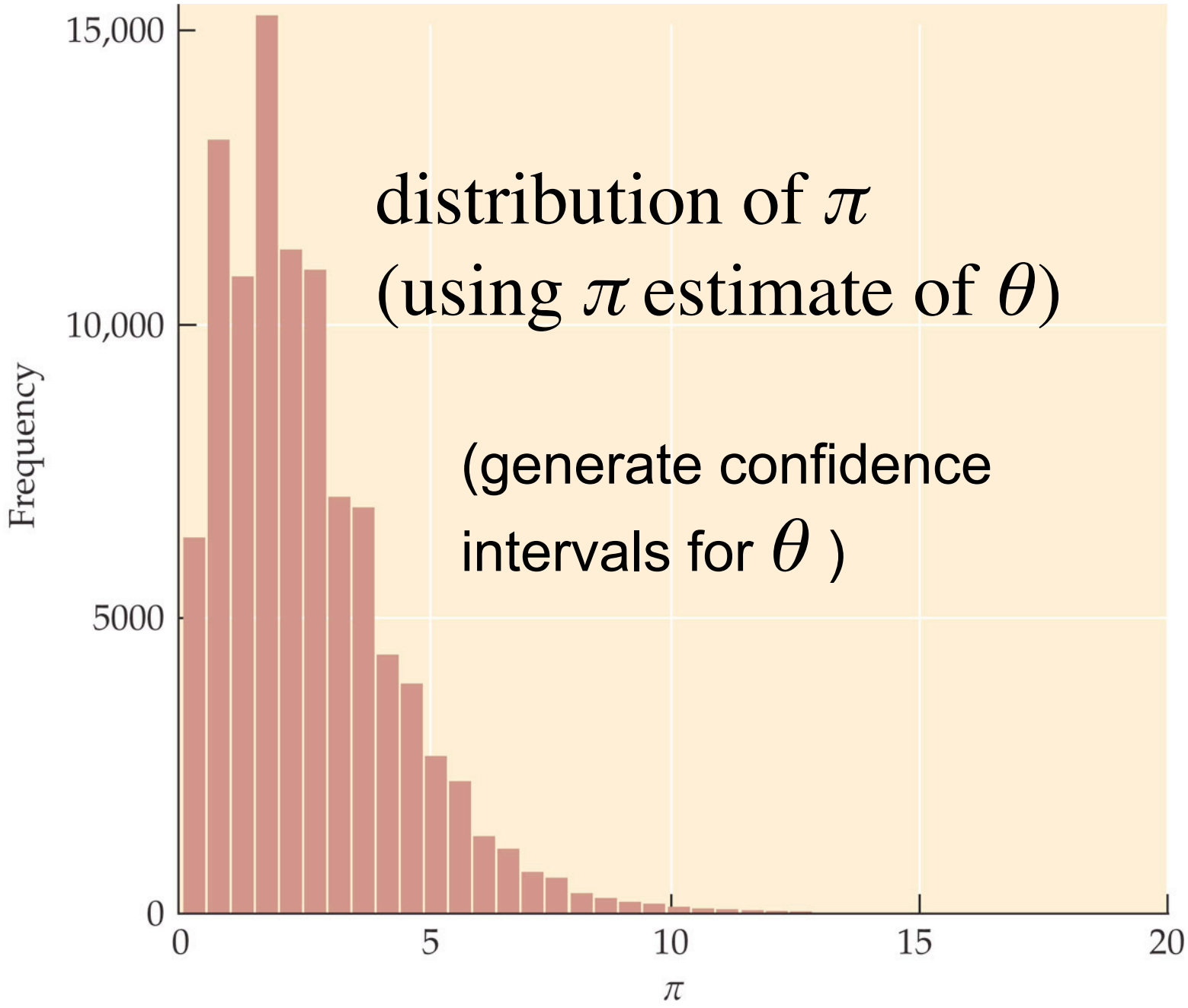
$$\pi = \frac{\sum_{i<j} d_{ij}}{n(n-1)/2} \qquad (3.5)$$

Nielsen and Slatkin 2013

We now have a convenient statistical method for estimating θ. A quantity for estimating parameters from the data is called an **estimator**. The estimator of θ based on π is sometimes called **Tajima's estimator** after the Japanese population geneticist F. Tajima. We denote an estimator of a quantity such as θ, by putting a circumflex, or "hat" on θ. We could, for example, write $\hat{\theta}_T = \pi$. The subscript T indicates that this is the particular estimator of $\hat{\theta}_T$ named after F. Tajima. It is important to distinguish the parameter θ, with a true value we may never know, from the estimator $\hat{\theta}_T$, which takes on a different value for each data set we consider.

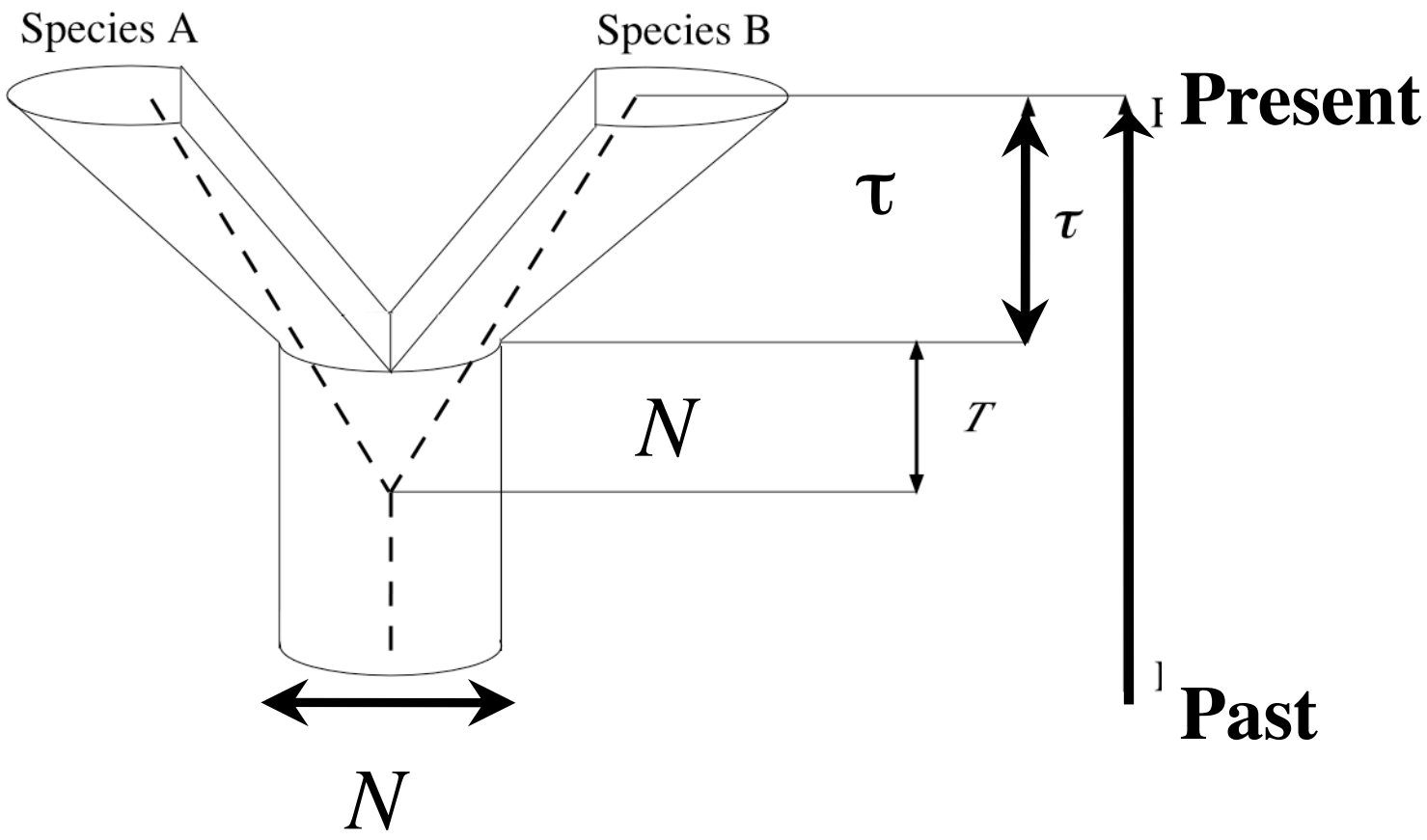
Nielsen and Slatkin 2013

Simulating π in 100,000 coalescence simulations under the standard coalescence model with infinite sites mutation and θ = 2.6

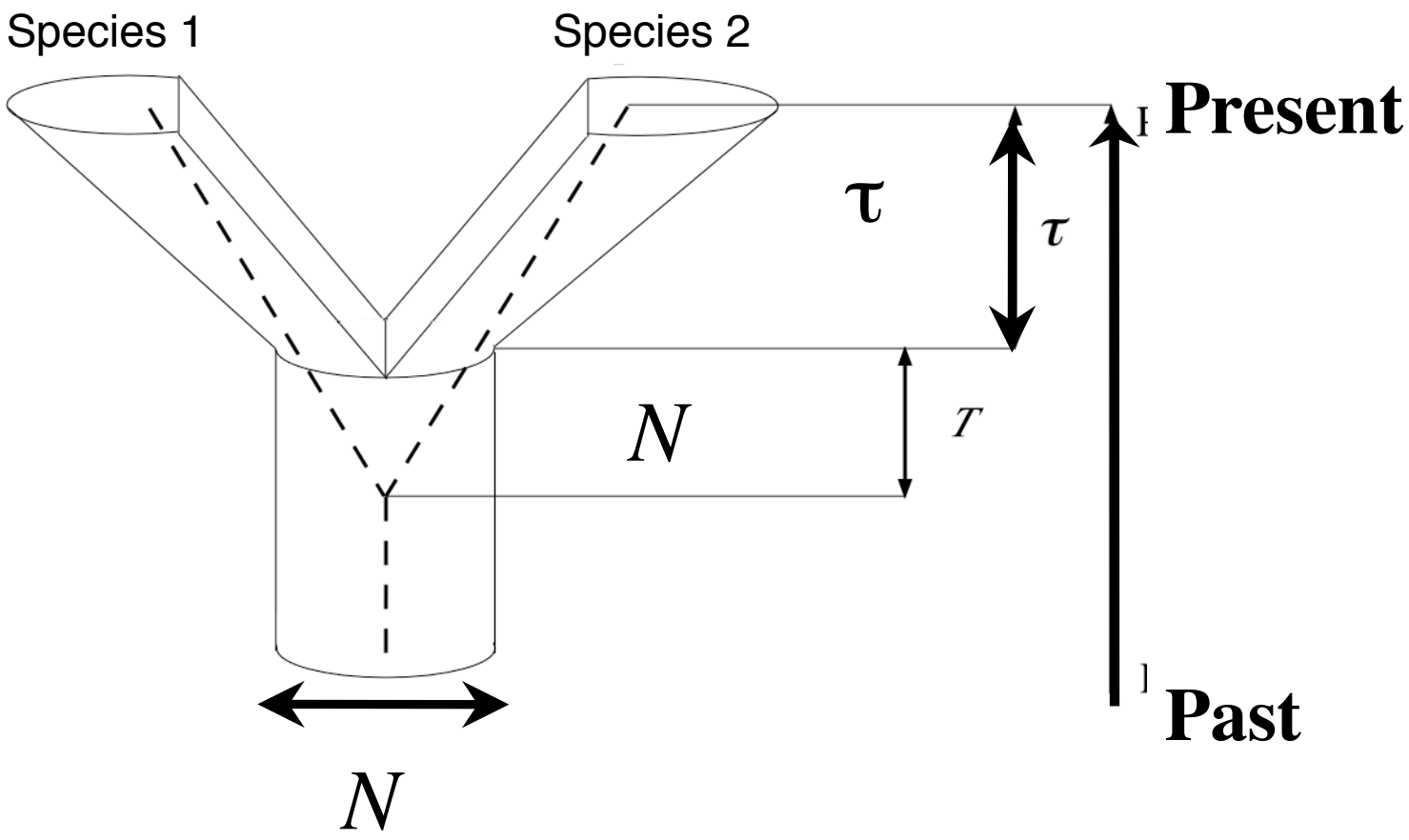


***INTRODUCTION TO POPULATION GENETICS*, Figure 5.3**

Genetic Divergence = $2\tau\mu + 2N\mu$



$$\pi_{net} = \pi_b - (\pi_1 + \pi_2)/2 \quad = 2\tau\mu + 2N\mu$$



Simulation Assignment: comparing single and double population models

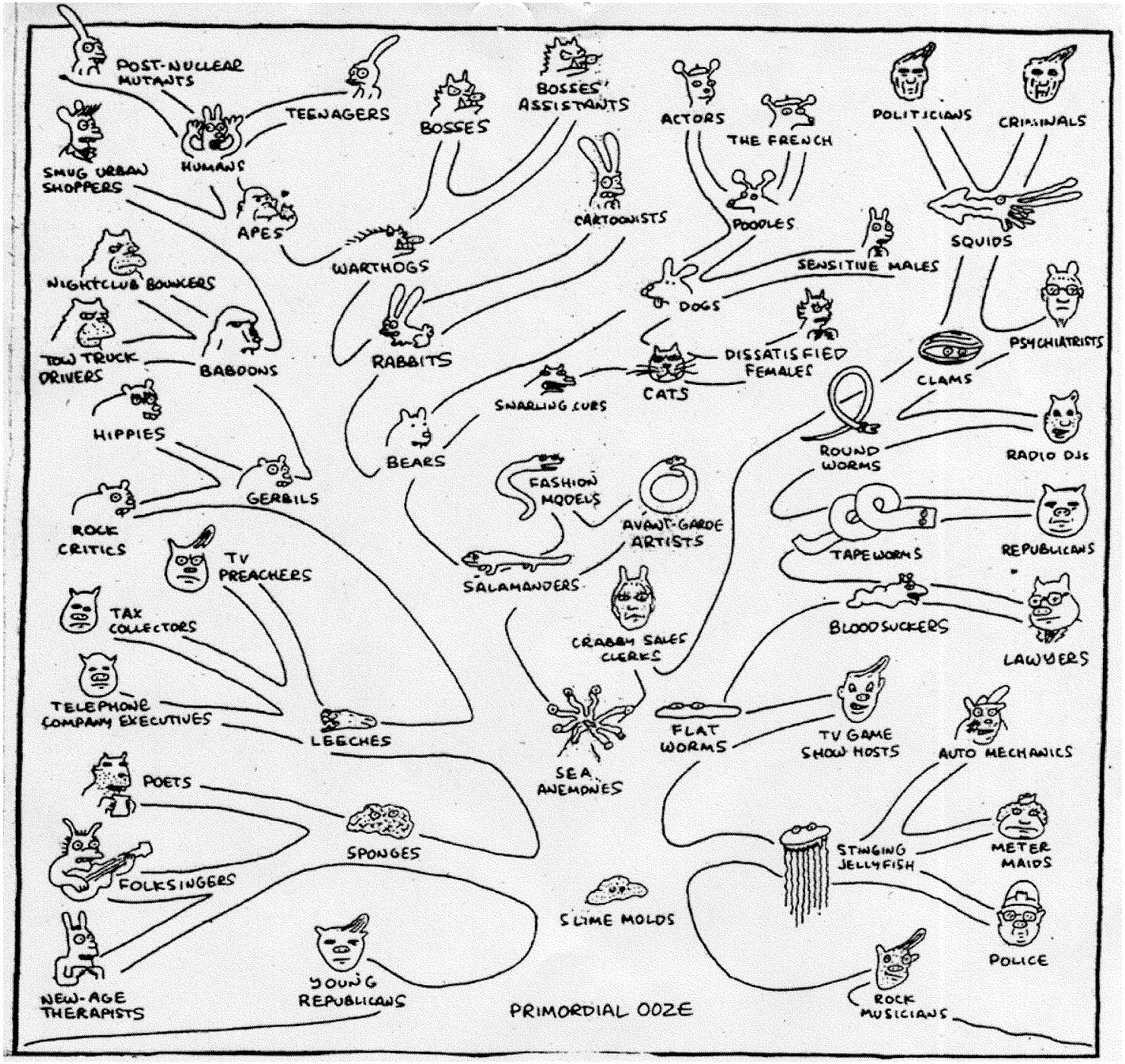
Use msPrime, SLiM or PipeMaster

1 locus, 1000 base pairs
μ=1e-7 per base pair

Simulate a 2 population model
$\tau = 5{,}000$ generations
$N_1 = 10{,}000, N_2 = 10{,}000, N_A = 10{,}000$
$n_1 = 10, n_2 = 10$

1. Simulate once and calculate an observed π (ignoring which of the 2 populations the 20 individuals come from).
2. Simulate a single population model with that observed π value 1,000 times (pretend you know the mutation rate exactly) and plot the distribution of simulated π values
3. indicate the 'true' $4N\mu$ on the plot. How off is it?
4. Simulate the 2 pop model 1,000 times and choose N values that you think will give you a π distribution that matches the one from the single pop model
5. plot results of 3 and 5

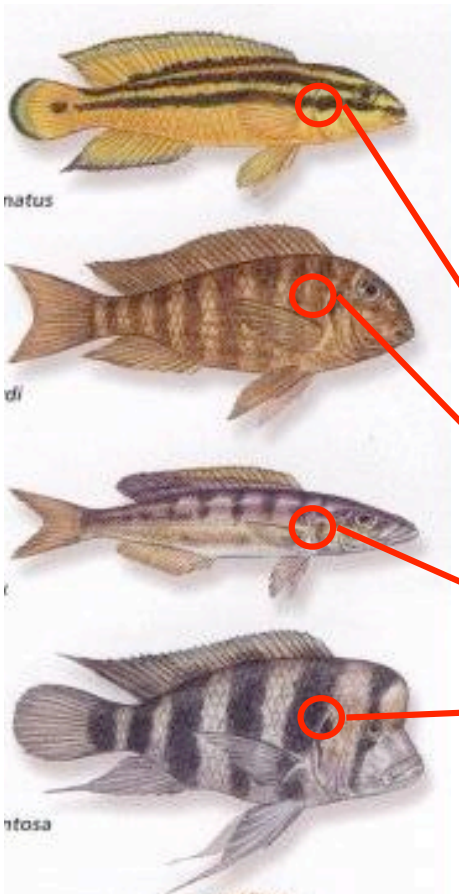
Using all the data - estimating a gene tree

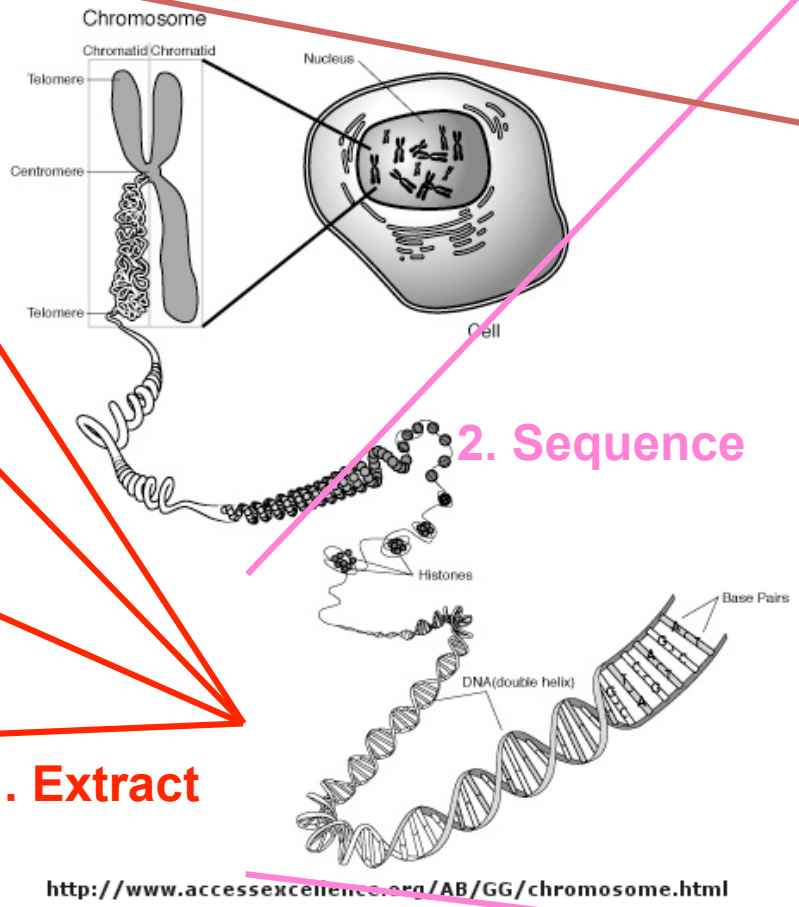


Constructing Phylogenetic Trees using Molecular characters

```
Outgroup     AAGCTTCATAGGAGCAACCATTCTAATAATAAGCCTCATAAAGCC
Species A    AAGCTTCACCGGCGCAGTTATCCTCATAATATGCCTCATAATGCC
Species B    GTGCTTCACCGACGCAGTTGTCCTCATAATGTGCCTCACTATGCC
Species C    GTGCTTCACCGACGCAGTTGCCCTCATGATGAGCCTCACTATGCA
```

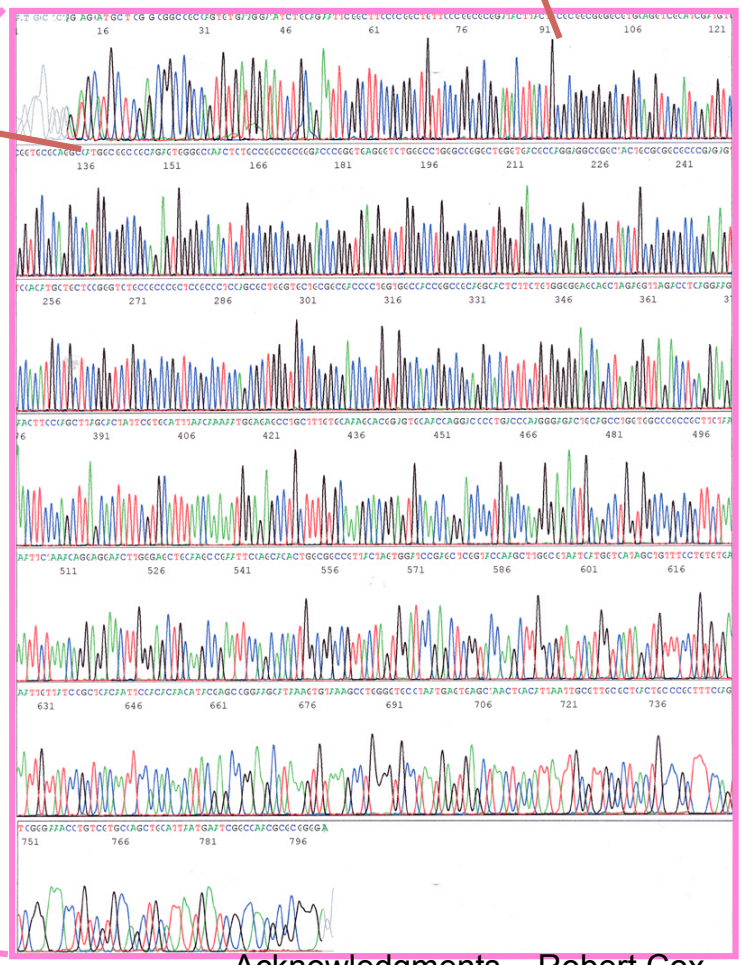
3. Align





2. Sequence

1. Extract



Acknowledgments – Robert Cox

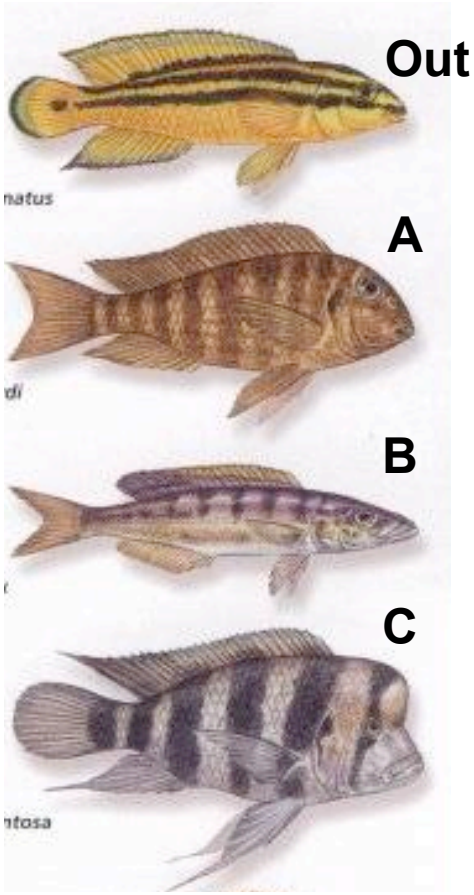
Molecular characters

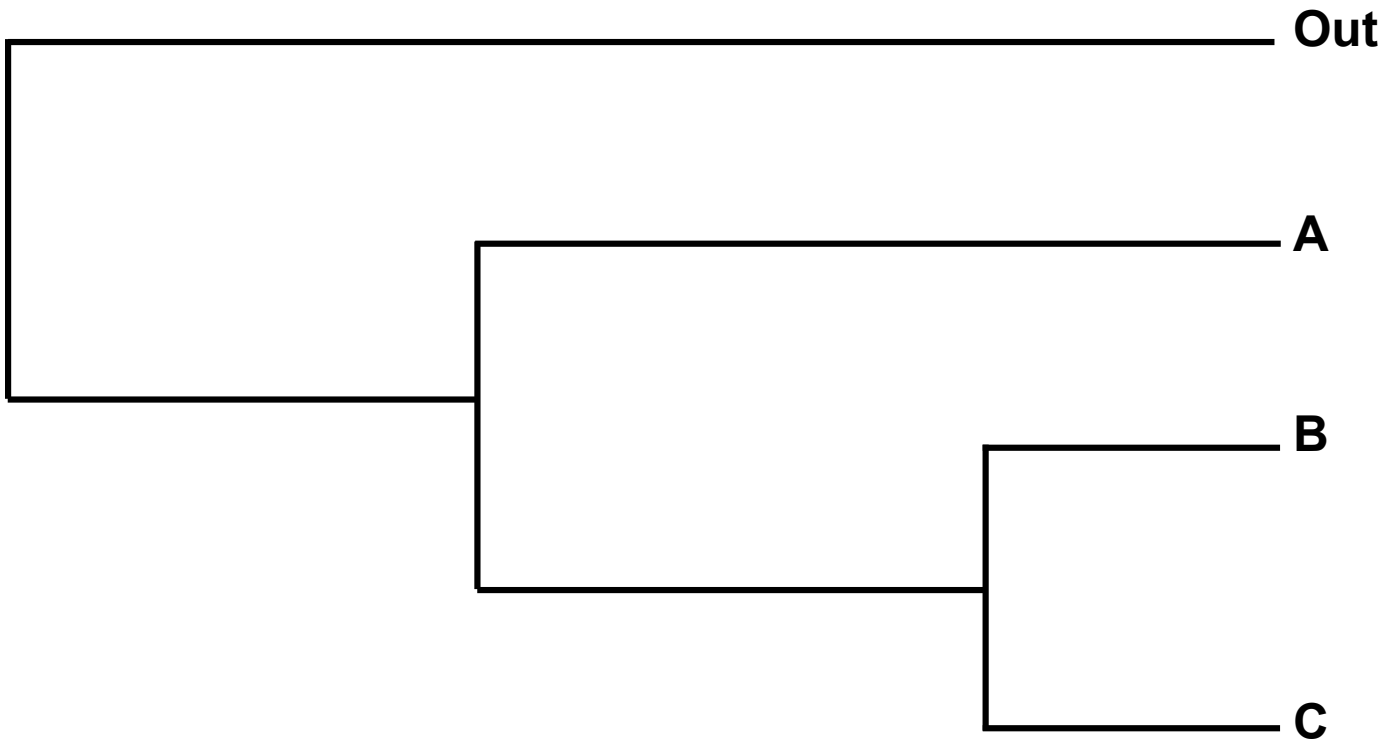
```
Outgroup
Species A
Species B
Species C
```

```
AAGCTTCATA
GAGCTTCACA
GTGCTTCACG
GTGCTTCACG
```

Invariable sites

These are not useful phylogenetic characters



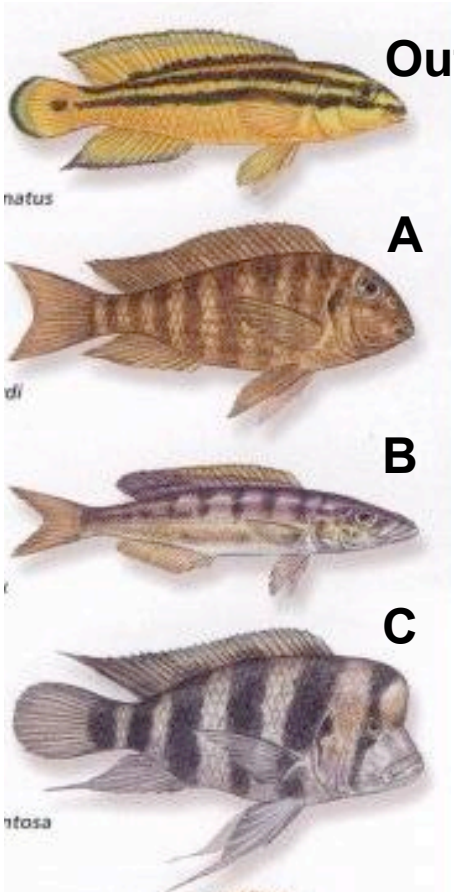


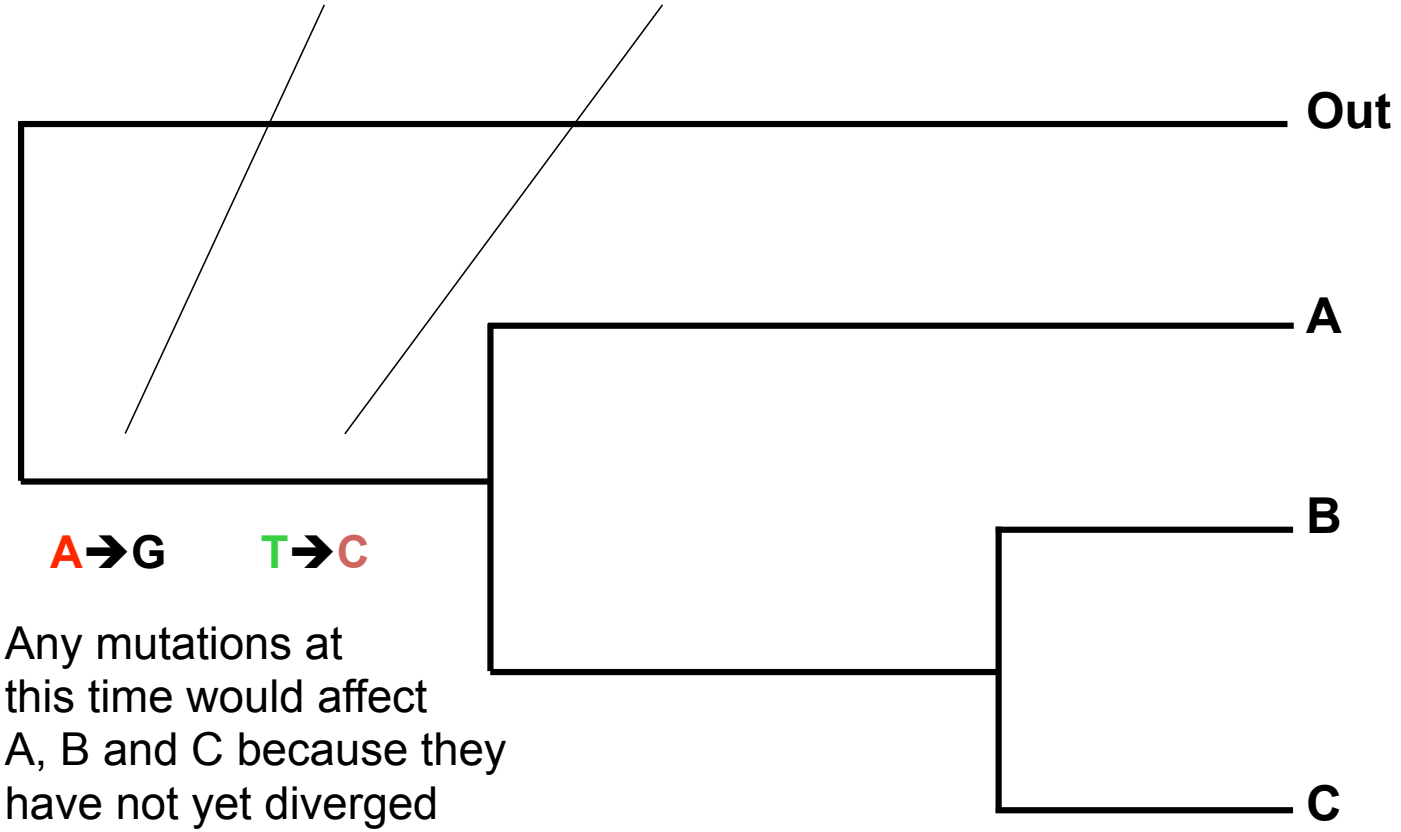
Acknowledgments – Robert Cox

Molecular characters

```
Outgroup    AAGCTTCATA
Species A   GAGCTTCACA
Species B   GTGCTTCACG
Species C   GTGCTTCACG
```

Synapomorphies supporting A+B+C



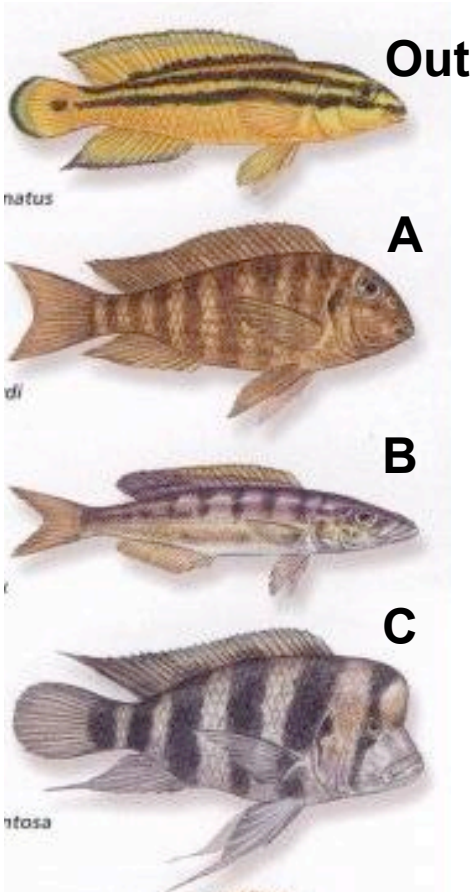


Any mutations at this time would affect A, B and C because they have not yet diverged

Acknowledgments – Robert Cox

Molecular characters

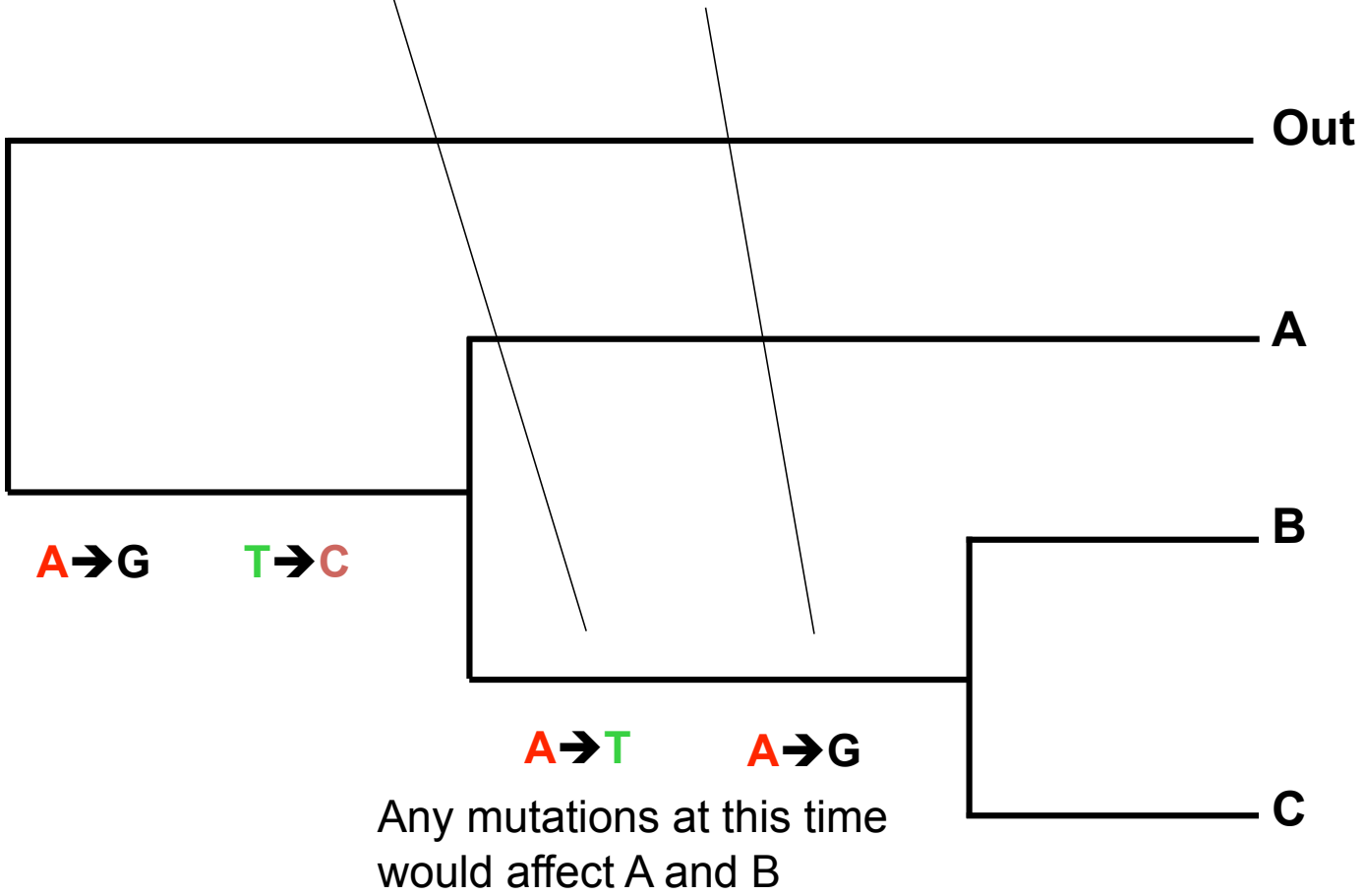
```
Outgroup
Species A
Species B
Species C
```




```
AAGCTTCATA
GAGCTTCACA
GTGCTTCACG
GTGCCTCACG
```

Synapomorphies supporting A+B+C

Synapomorphies supporting B+C



Any mutations at this time would affect A and B

Acknowledgments – Robert Cox

Molecular characters

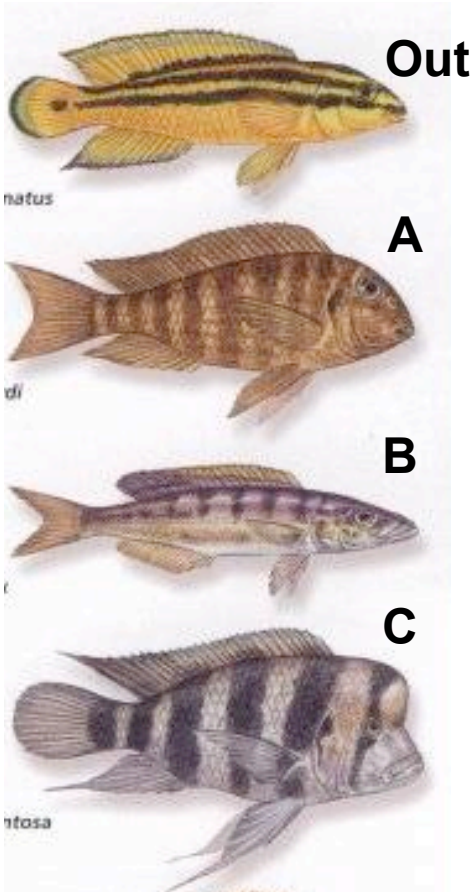
Outgroup
Species A
Species B
Species C

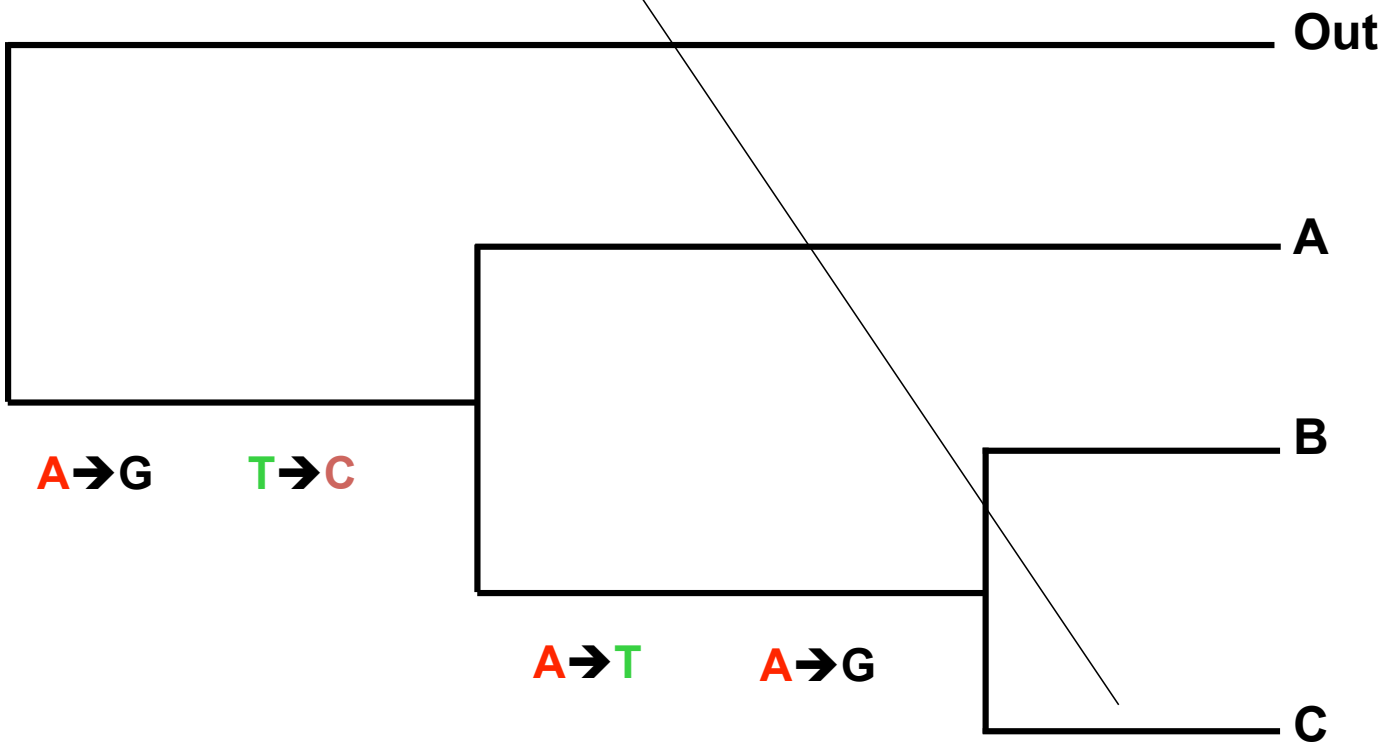
AAGCTTCATA
GAGCTTCACA
GTGCTTCACG
GTGCCTCACG

Synapomorphies supporting A+B+C

Synapomorphies supporting B+C

Apomorphy for C





Any mutations at this time would only affect C T→C

Acknowledgments – Robert Cox

Methods in Phylogenetic Reconstruction

Distance

Maximum Parsimony

Maximum Likelihood/Bayesian

* All algorithms are implemented using available software, eg. PAUP, PHYLIP, McClade, Mr. Bayes etc.

Comparison of Methods (the old school circa 1997)

Distance	**Maximum parsimony**	**Maximum likelihood**
Uses only pairwise distances	Uses only shared derived characters, infinite sites model	Uses all data
Minimizes distance between nearest neighbors	Minimizes total number of changes	Maximizes tree likelihood given specific parameter values
Very fast	Slow	***Very*** slow
Easily trapped in local optima	only uses infinite sites model	can use any evolution model (and use likelihood criteria to choose)
Good for generating tentative tree	Best option when homoplasy rare	optimal method if model is correct (but can be difficult in practice for large data sets)

$$L = Pr(data \mid \Theta)$$

Θ = model, tree & parameters

parameters = e.g. rates, times etc
model = infinite sites
tree = among all possible trees

Likelihood (*L*) is the probability of the data (alignment), given a tree (with topology *and* branch lengths specified) and a probabilistic model of evolution.

$$L = Pr(data \mid \Theta)$$

Θ = model, tree & parameters

parameters = e.g. rates
model = infinite sites
tree = among all possible trees

Likelihood (*L*) is the probability of the data (alignment), given a tree (with topology *and* branch lengths specified) and a probabilistic model of evolution.

Assumptions (the fine print):

- The tree is correct
- The probability that a position has a certain state at time 1 depends only on the state at time 0; knowing that it had some state prior to time 0 is irrelevant
- (this is called a *Markov process)*
- Data (individual sites) are independent
- A uniform evolutionary process operated across the entire tree (why might this be false? endosymbiosis? loss of function?),
- i.e., the process of evolution is a *homogeneous Markov process.*

Probability and prediction

Know parameters -> Predict Outcome (predict data)

Likelihood

Observe Data -> Estimate parameters

Likelihood and coin tosses

n = 100 (# of coin tosses)

h = 56 (# of heads)

$$Pr(data \mid \Theta) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

This starts the count of number of ways event can occur.

This is the probability of success for x trials.

This ends the count of number of ways event can occur.

This deletes duplications.

This is the probability of failure for the x trials.

If $p = 0.5$

$$\Pr(data|p = 0.50) = \frac{100!}{56!44!} 0.5^{56} 0.5^{44} = 0.0389$$

$$P(\text{heads}) = \frac{1}{2} = 0.5$$

Likelihood and coin tosses

n = 100 (# of coin tosses)

h = 56 (# of heads)

$$Pr(data \mid \Theta) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

This starts the count of number of ways event can occur.

This is the probability of success for x trials.

This ends the count of number of ways event can occur.

This deletes duplications.

This is the probability of failure for the x trials.

If $p = 0.5$

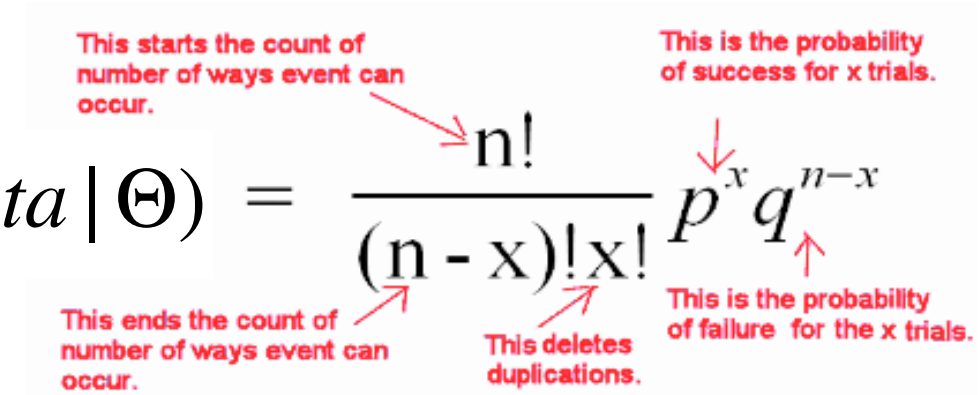
$$\Pr(data|p = 0.50) = \frac{100!}{56!44!} 0.5^{56} 0.5^{44} = 0.0389$$

But if $p = 0.52$

$$\Pr(data|p = 0.52) =$$ ie , bad old coin

n = 100 (# of coin tosses)

h = 56 (# of heads)

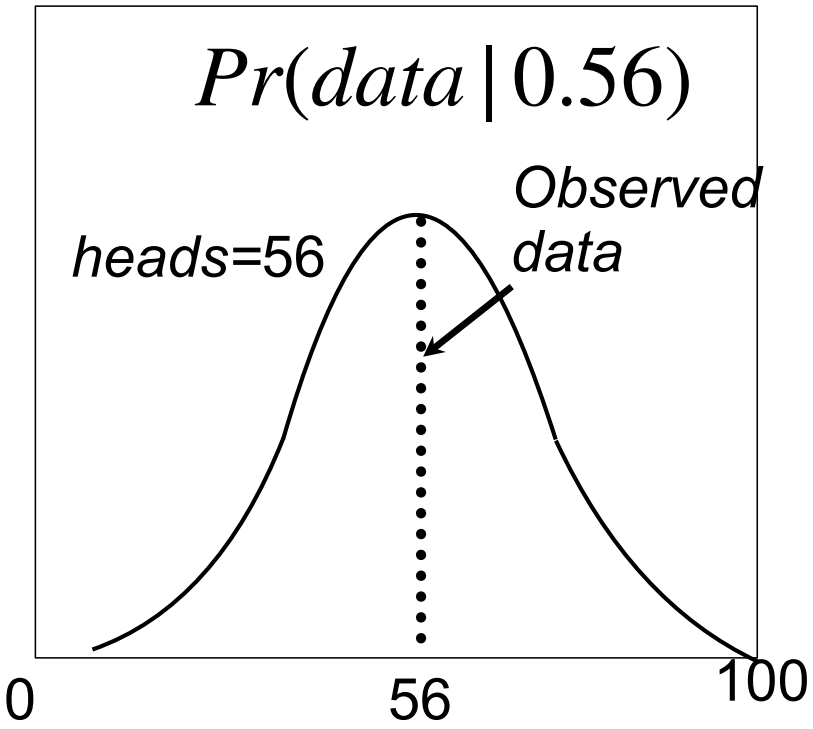


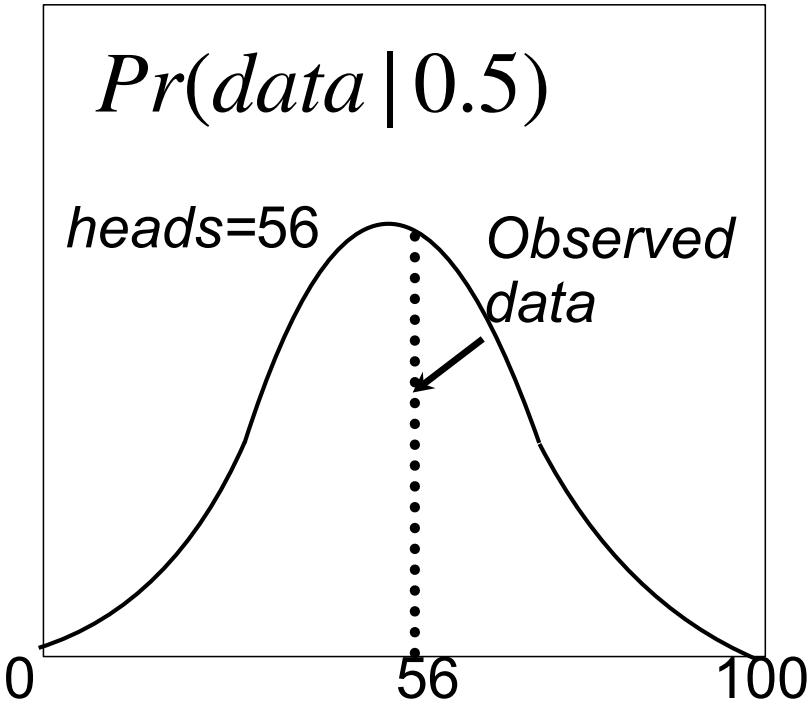
If $p = 0.5$

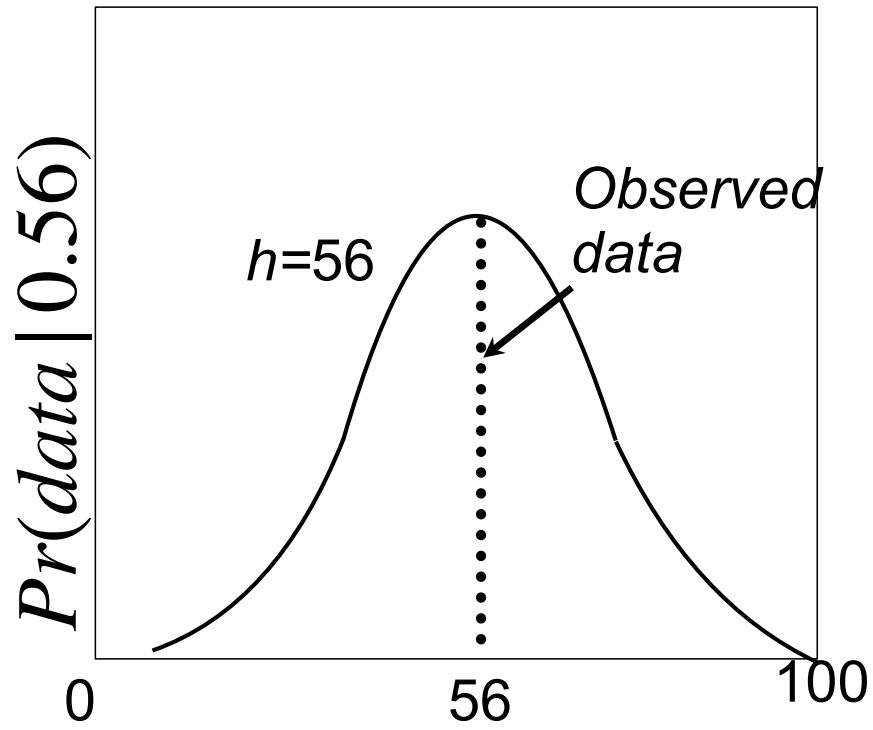
$$\Pr(data|p = 0.50) = \frac{100!}{56!44!}0.5^{56}0.5^{44} = 0.0389$$

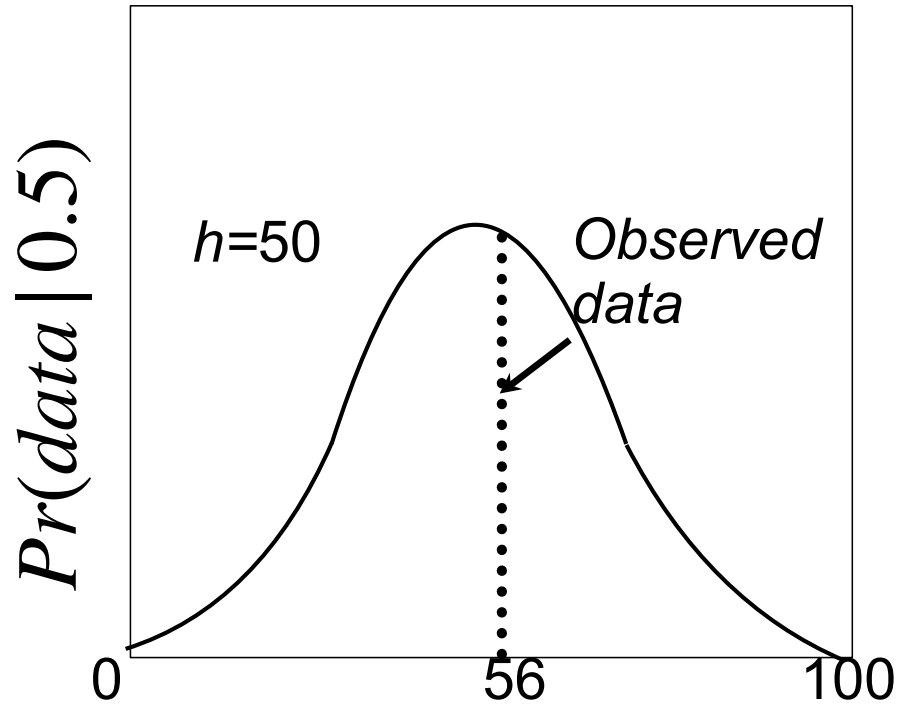
But if $p = 0.52$

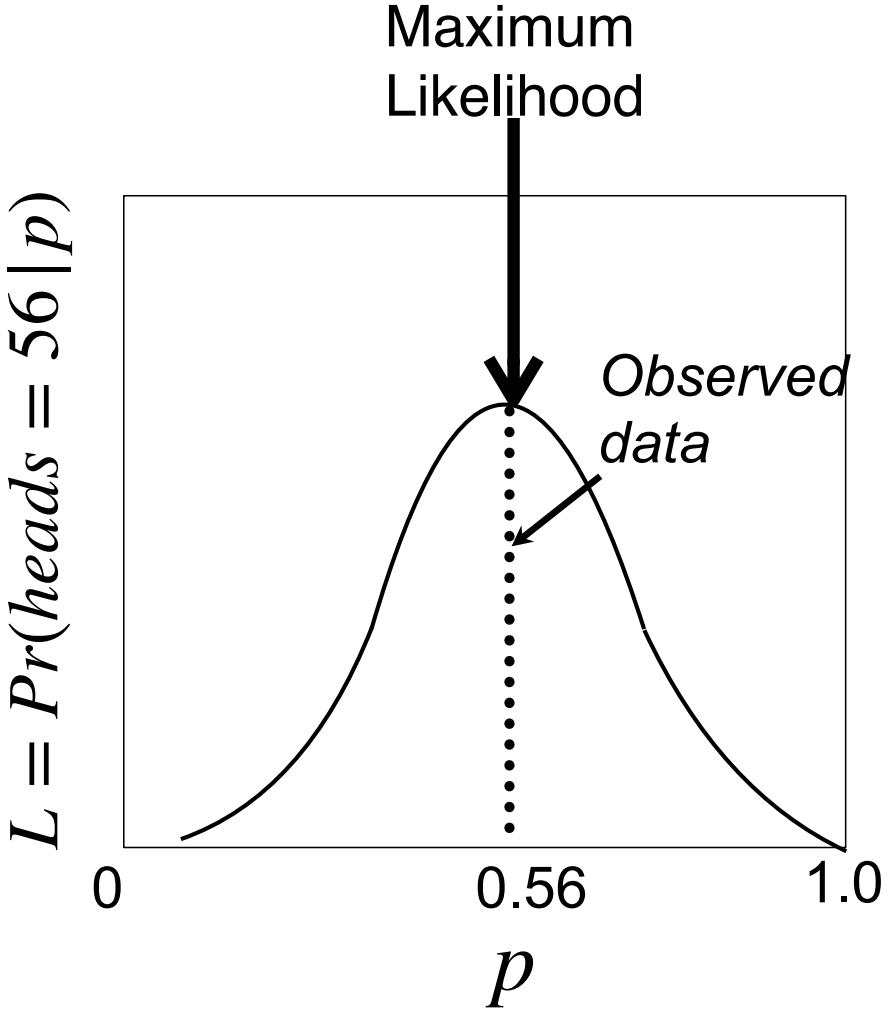
$$\Pr(data|p = 0.52) = \frac{100!}{56!44!}0.52^{56}0.52^{44} = 0.0581$$


$Pr(data|0.56)$
Observed data
heads=56
0
56
100


$Pr(data|0.5)$
heads=56
Observed data
0
56
100


$Pr(data\,|\,0.56)$
h=56
Observed data
0
56
100


$Pr(data\,|\,0.5)$
h=50
Observed data
0
56
100

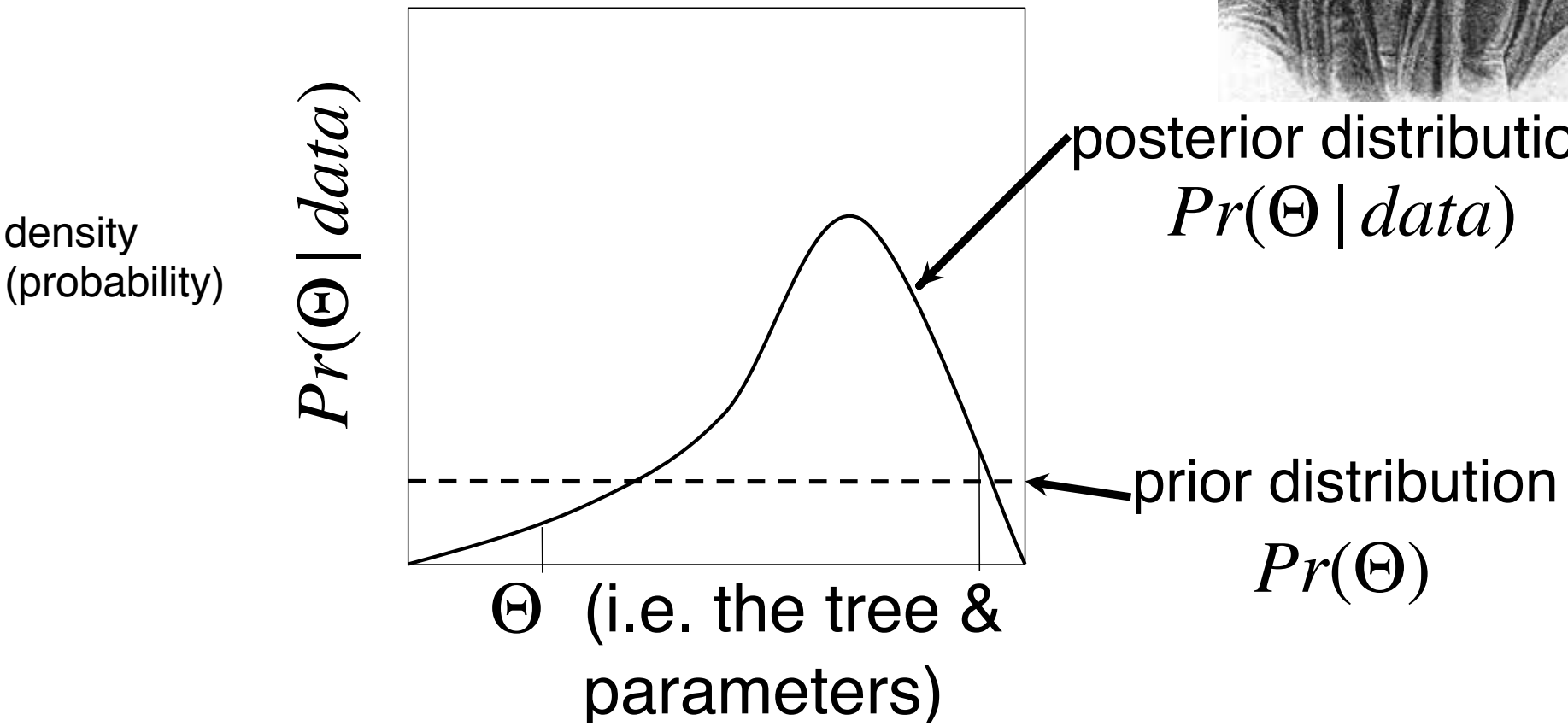

Maximum Likelihood
$L = Pr(heads = 56\,|\,p)$
Observed data
0
0.56
1.0
p

Bayesian

prior distribution for parameters, $Pr(\Theta)$

we can obtain posterior distribution $Pr(\Theta \mid data)$

$$Pr(\Theta \mid data) = \frac{Pr(data \mid \Theta) Pr(\Theta)}{Pr(data)}$$



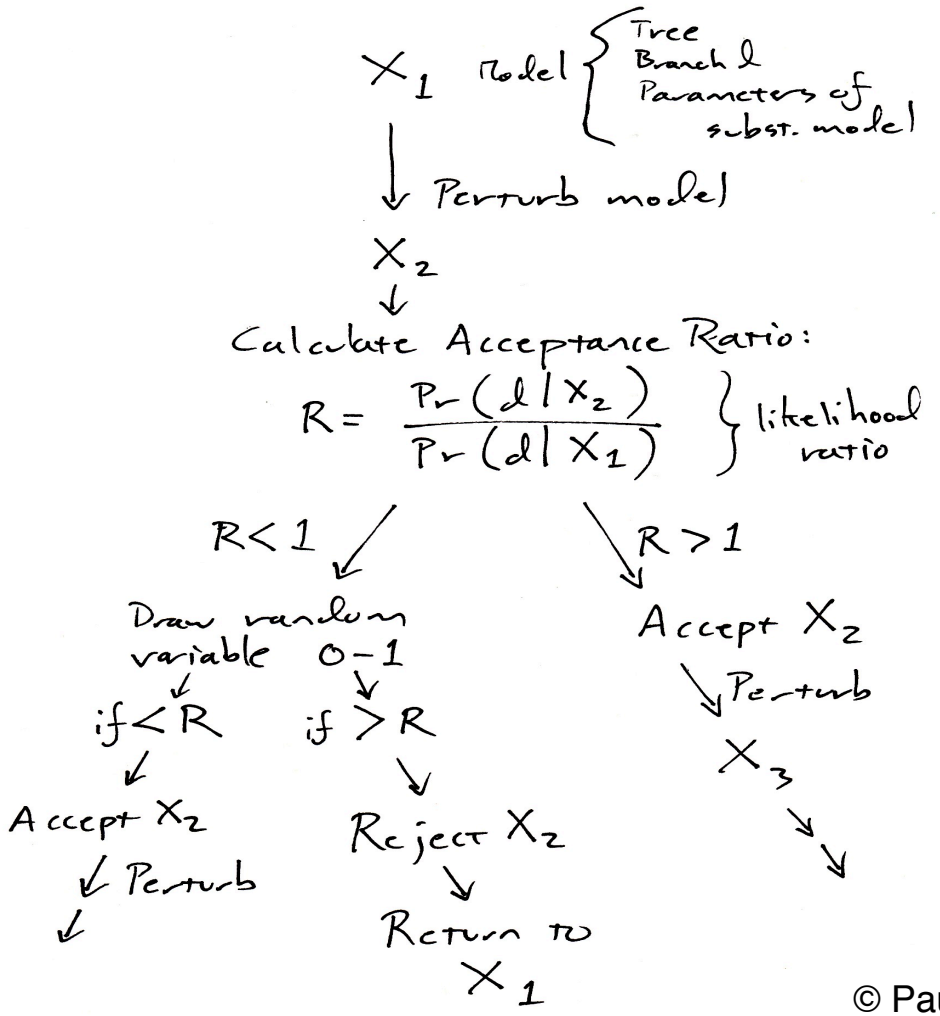
Calculating $Pr(data \mid \Theta)$ can be difficult

Can only do it in closed form for the simplest models

Complex models can in principle be solved via integration, but this is often intractable

must resort to numerical or simulation-based methods
(MCMC, ABC, supervised machine learning etc)

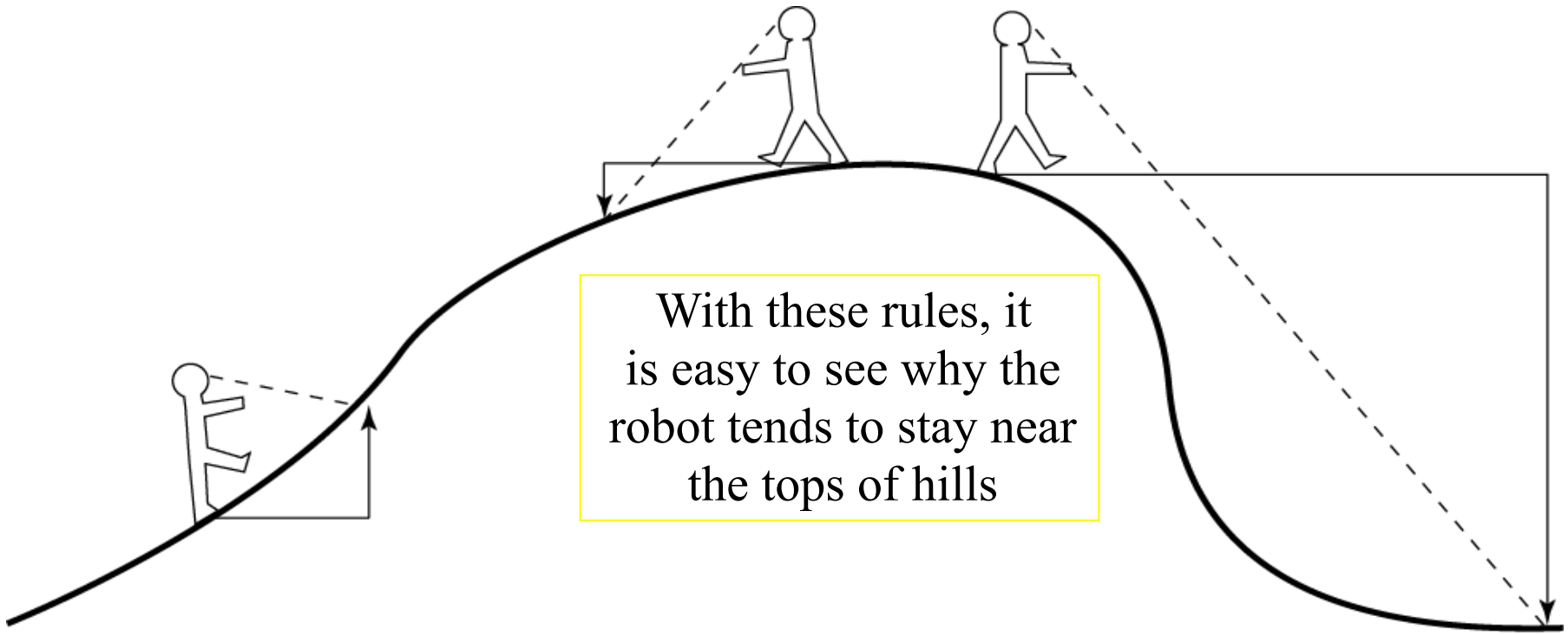
Markov Chain Monte Carlo (MCMC)



MCMC robot's rules

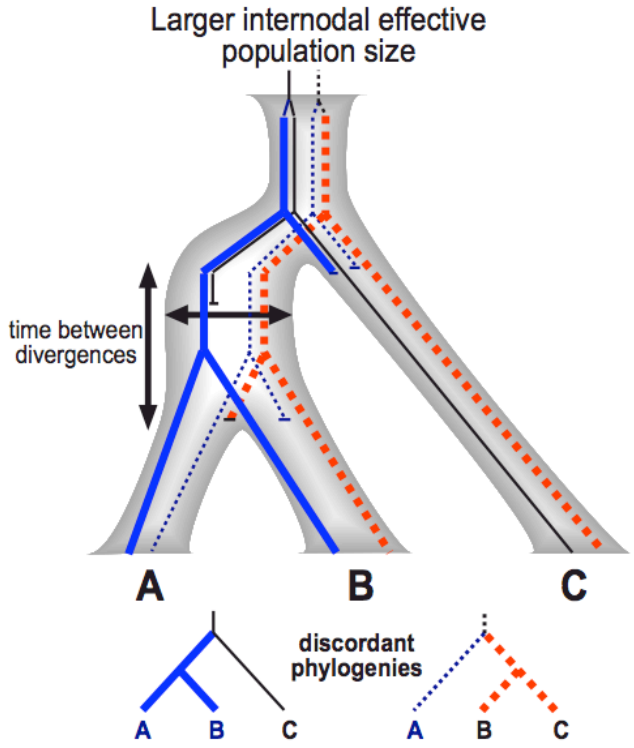
Slightly downhill steps are usually accepted

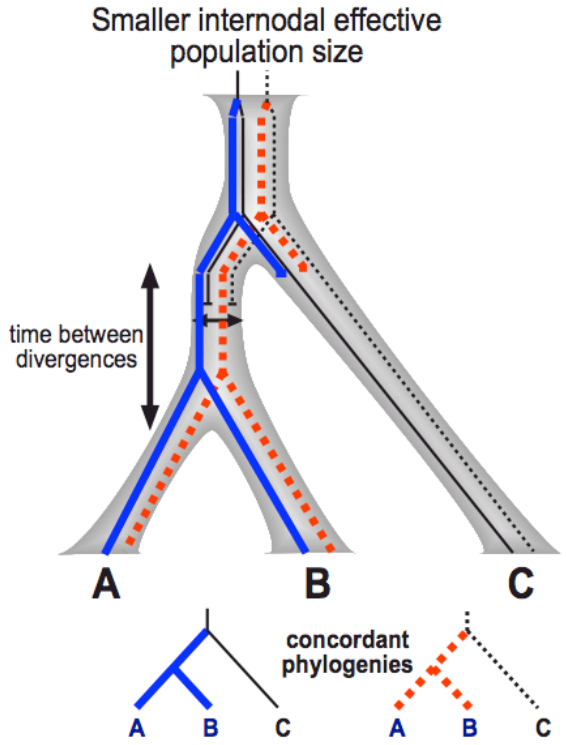
Drastic "off the cliff" downhill steps are almost never accepted



With these rules, it is easy to see why the robot tends to stay near the tops of hills

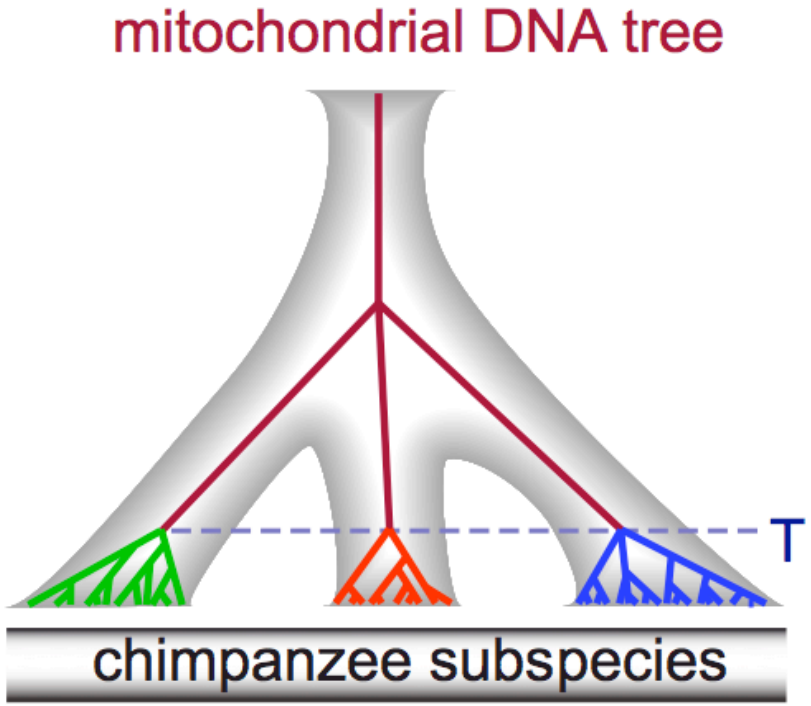
Uphill steps are always accepted

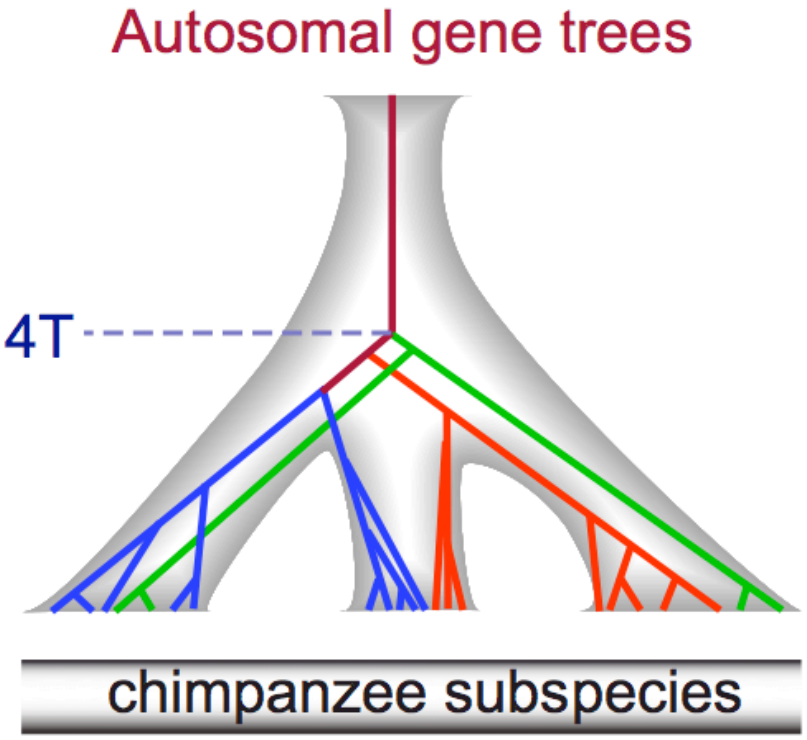




Often the gene tree Does not equal the "species" tree

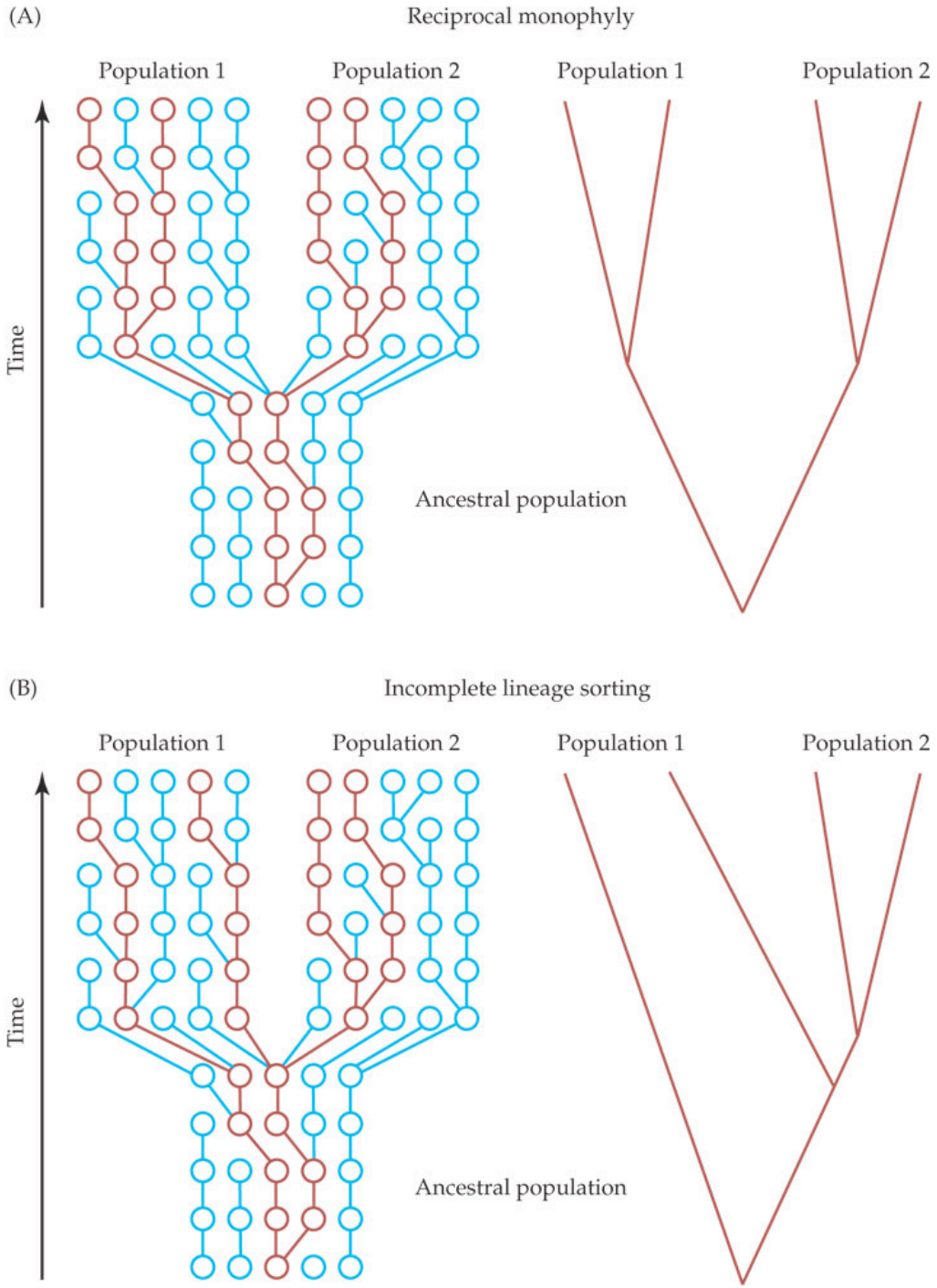
At any one particular gene, time to Co-ancestry can be long





Time to Co-ancestry and gene tree/species tree Congruence depends on Ne (effective population sizes)

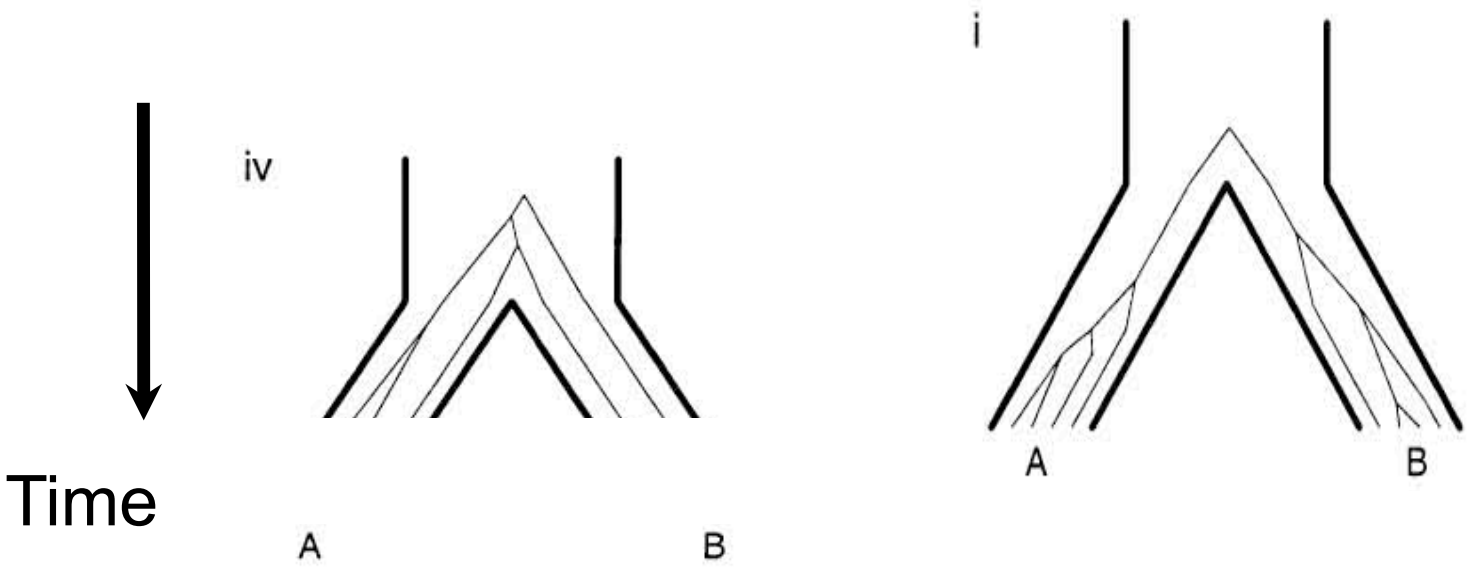
Autosomes have 4x size as Y and mtDNA



***INTRODUCTION TO POPULATION GENETICS*, Figure 5.5**

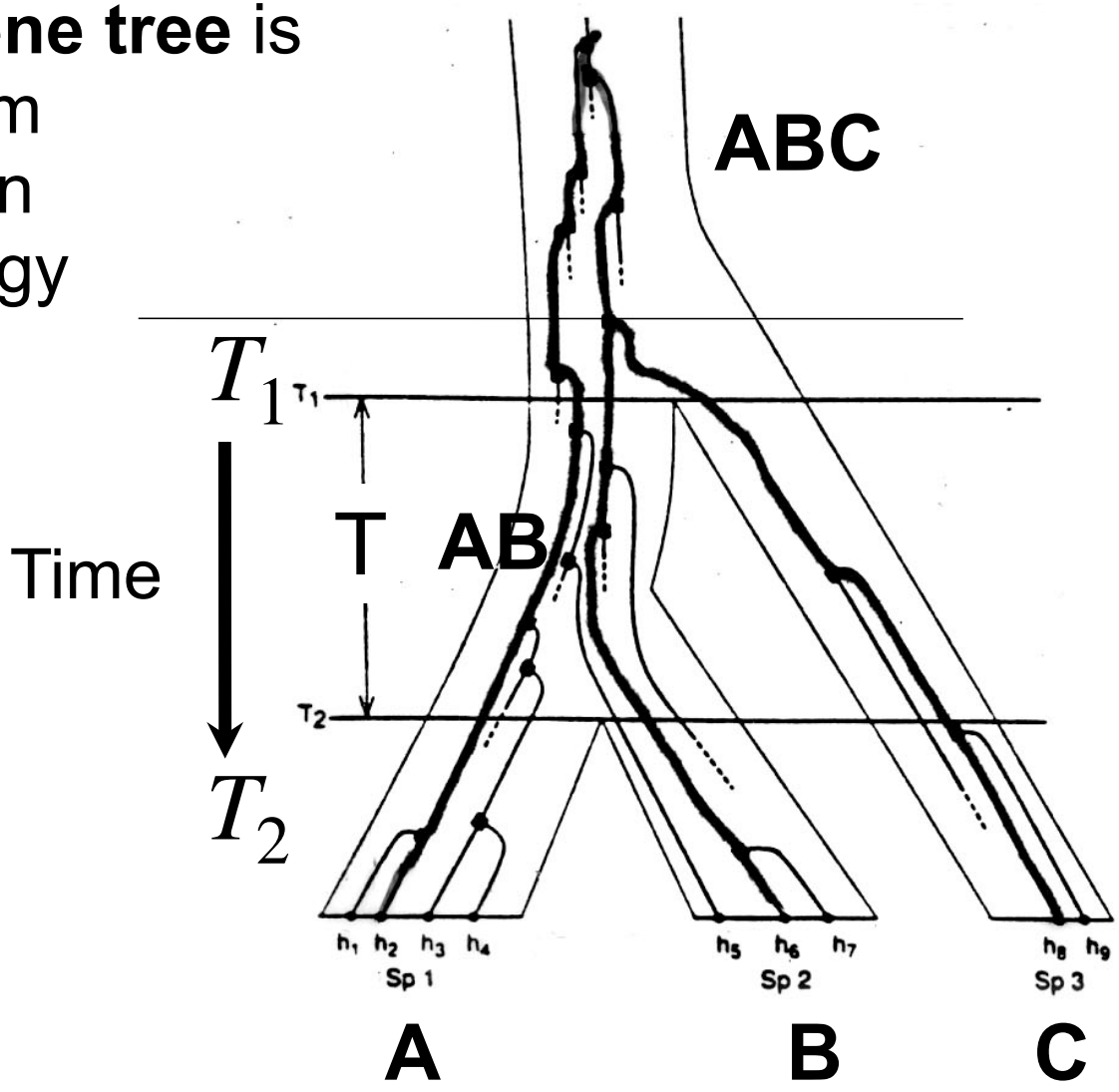
Incomplete
Lineage **S**orting
ILS

2 taxon species tree

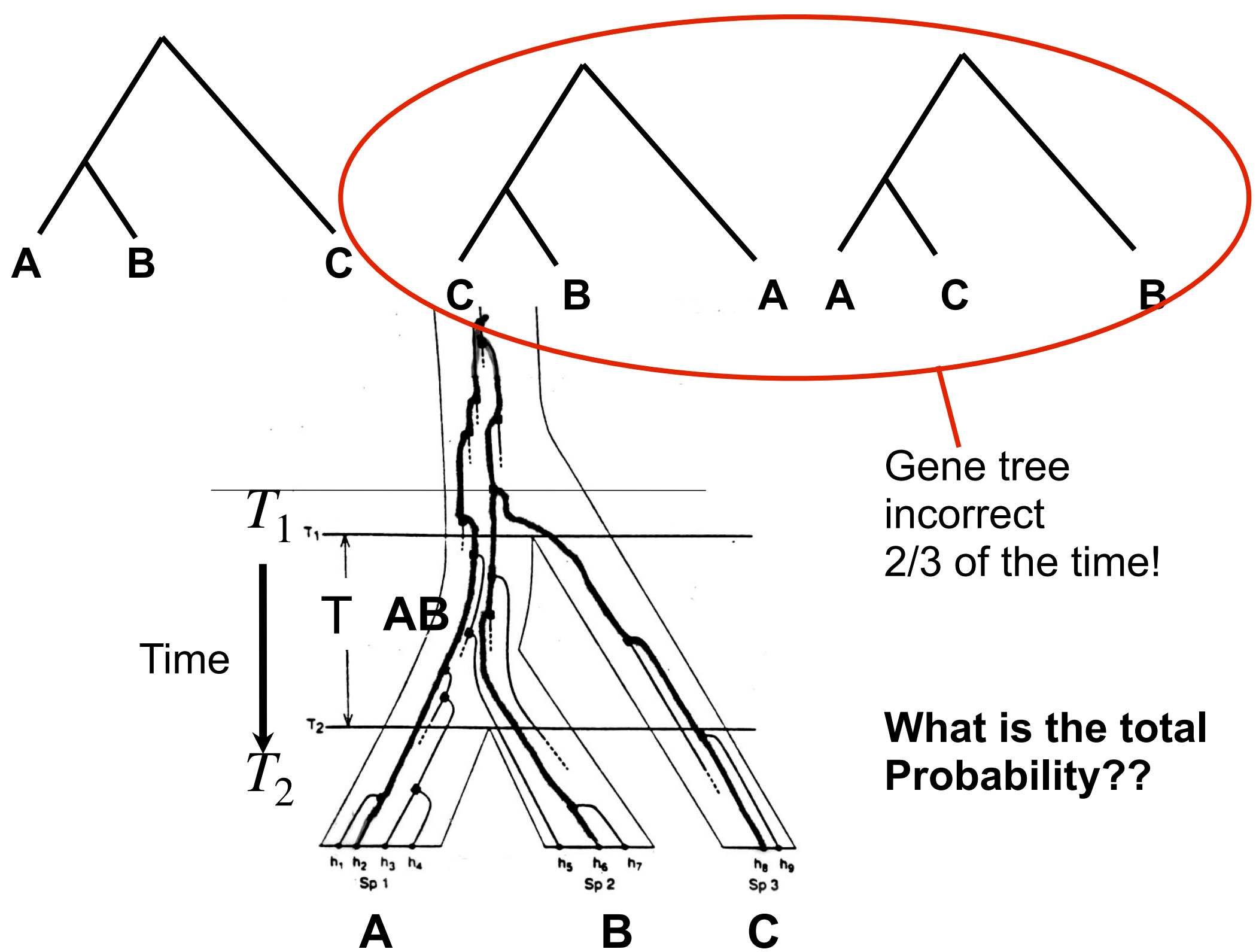


Pr(Gene tree = Species tree)
Increases with time

If no coalescent in ancestral species AB, (T generations) the **gene tree** is random 3 taxon topology

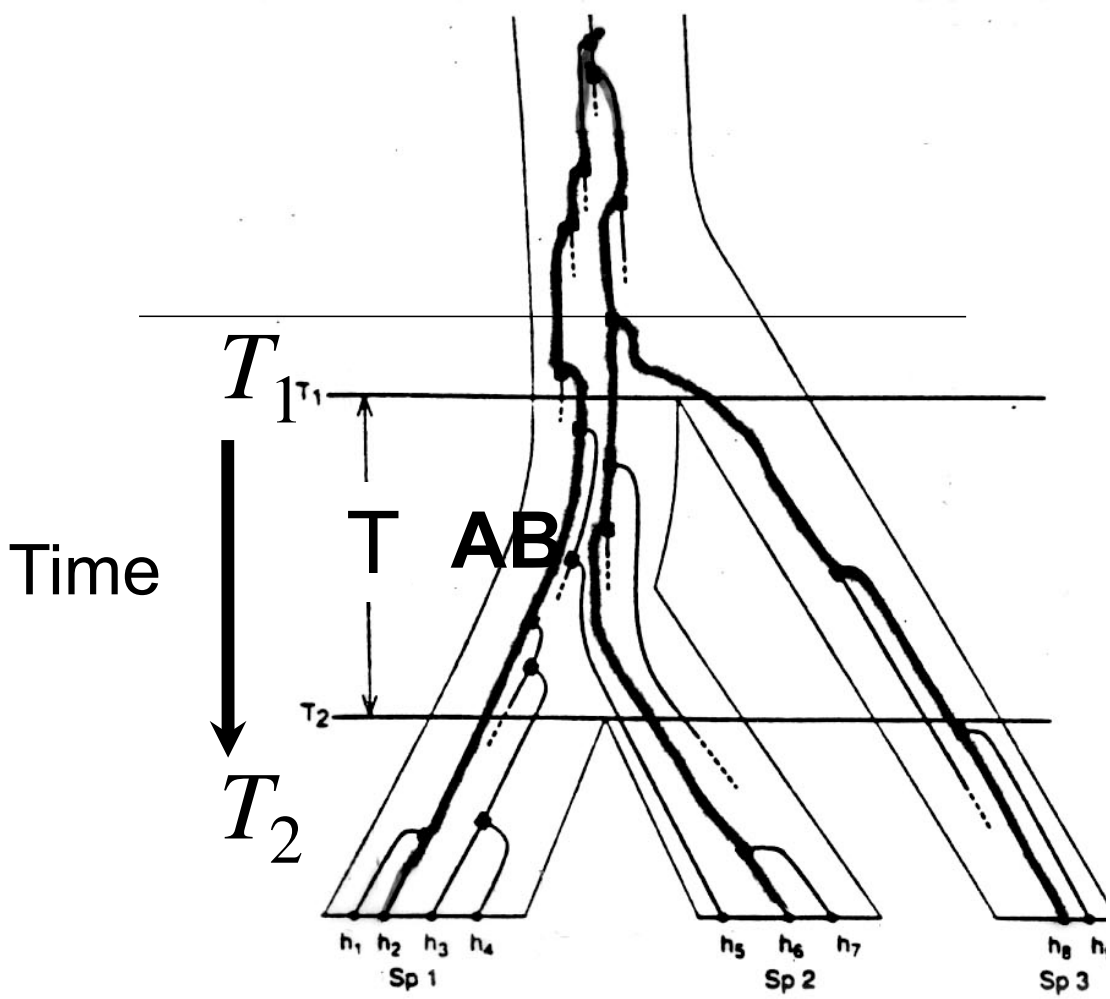


and Gene Tree incorrect 2/3 of the time!


A
B
C
C
B
A
A
C
B
Gene tree
incorrect
2/3 of the time!
T_1
T_1
T
AB
Time
T_2
T_2
h_1 h_2 h_3 h_4
Sp 1
h_5 h_6 h_7
Sp 2
h_8 h_9
Sp 3
A
B
C
What is the total
Probability??

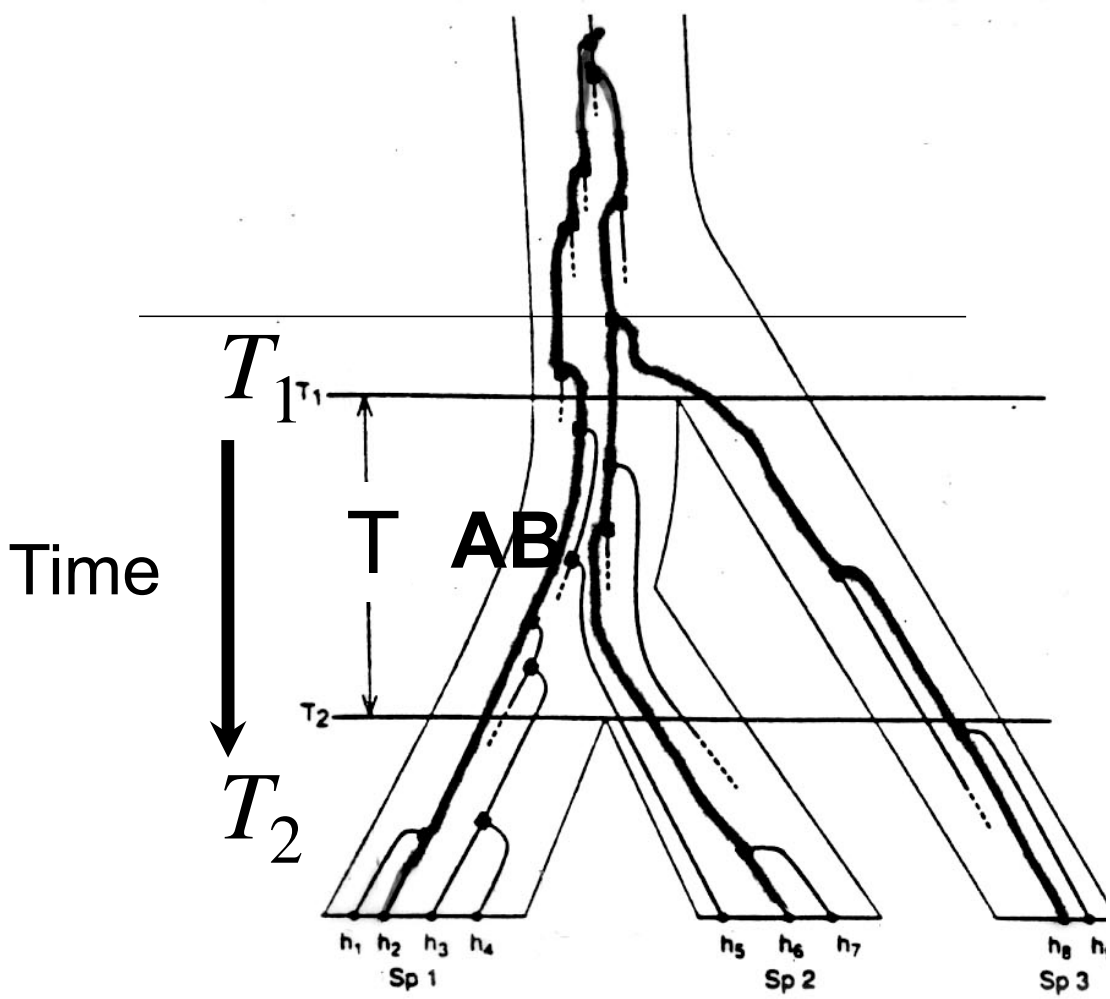
Because probability of no coalescent events in t generations is

$$Pr(k_t = k_0) = [1 - \frac{k_0(k_0 - 1)}{4N}]^t$$



Because probability of no coalescent events in t generations is

$$Pr(k_t = k_0) = [1 - \frac{k_0(k_0 - 1)}{4N}]^t$$



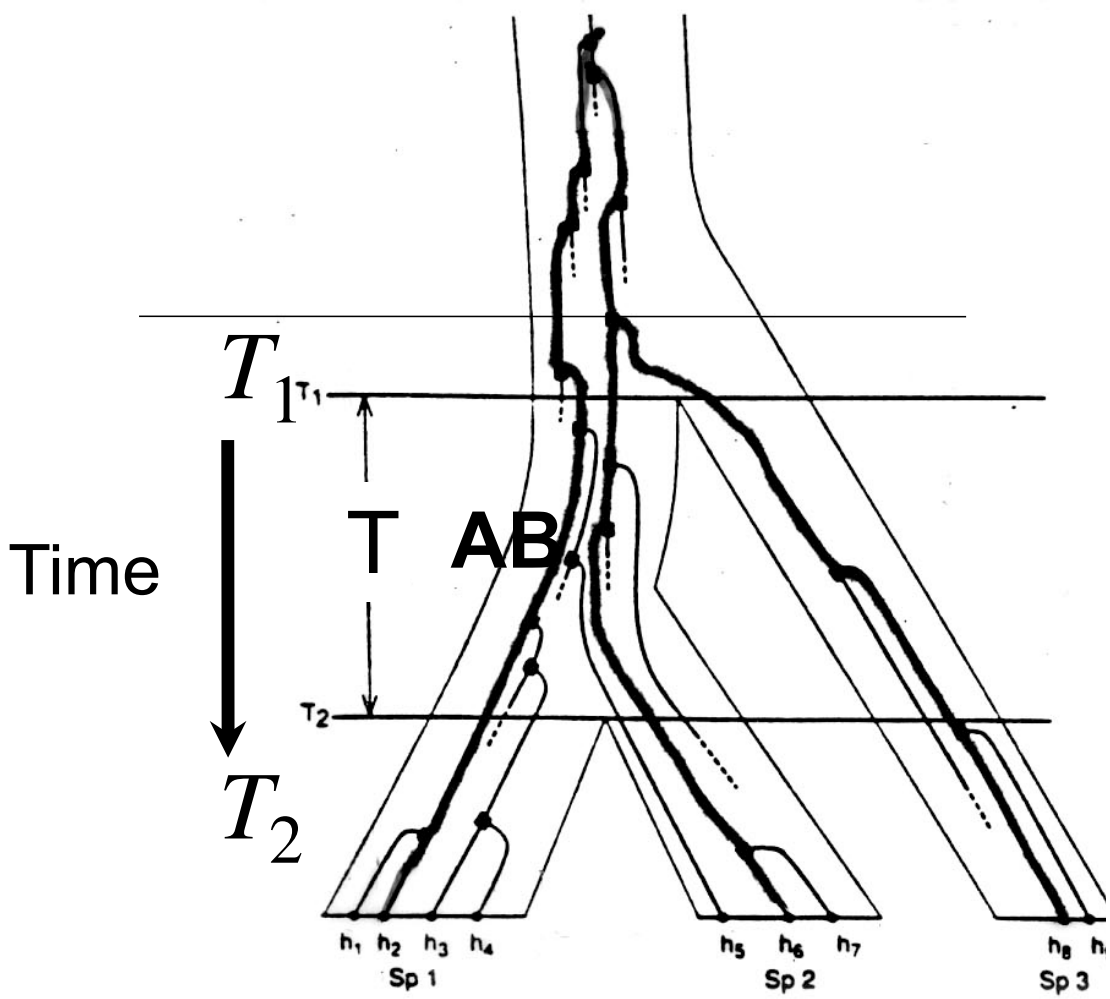
$$[1 - \frac{k_0(k_0 - 1)}{4N}]^t$$

$$[1 - \frac{2(2 - 1)}{4N}]^t$$

$$[1 - \frac{1}{2N}]^t$$

Because probability of no coalescent events in t generations is

$$Pr(k_t = k_0) = [1 - \frac{k_0(k_0 - 1)}{4N}]^t$$

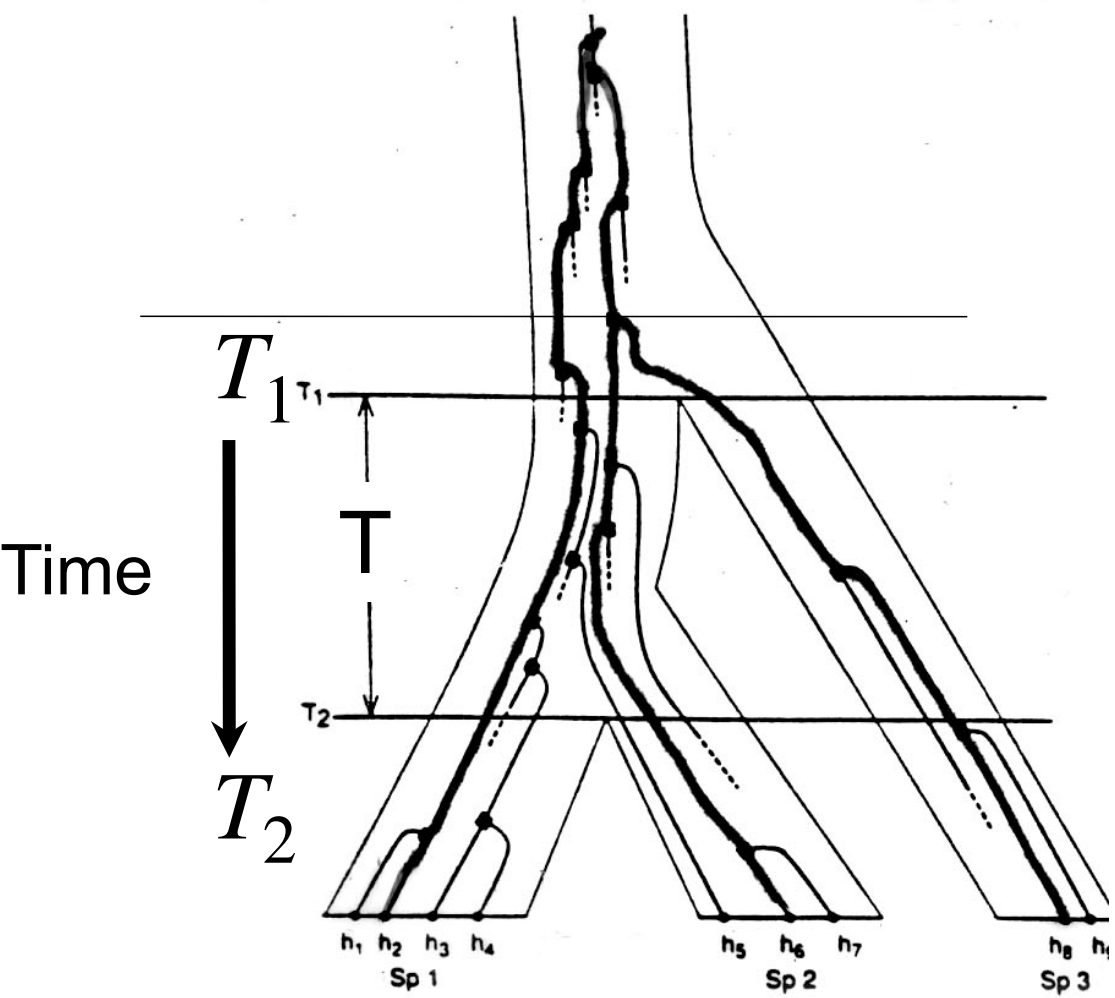


$$[1 - \frac{k_0(k_0 - 1)}{4N}]^t$$

$$[1 - \frac{2(2 - 1)}{4N}]^t$$

$$[1 - \frac{1}{2N}]^t$$

$$e^{-t/2N}$$



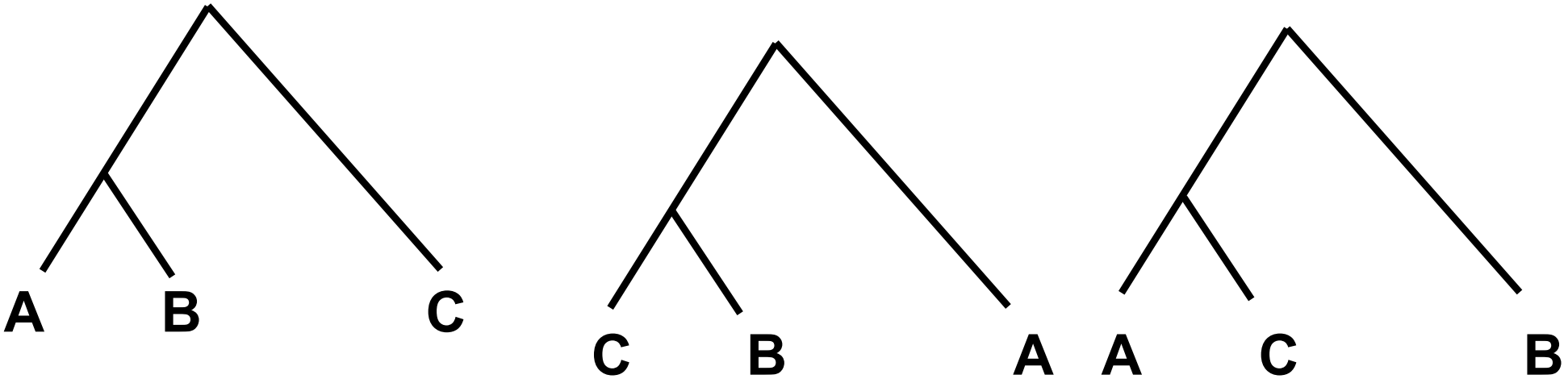
So, …. probability of NO coal events between T_1 and T_2 (T)= $e^{-t/2N}$

AND

Probability of a random 3-taxon topology = 2/3

Therefore probability of Gene tree ≠ Species tree =

$$\frac{2}{3}e^{-t/2N}$$



congruent gene tree

incongruent gene tree

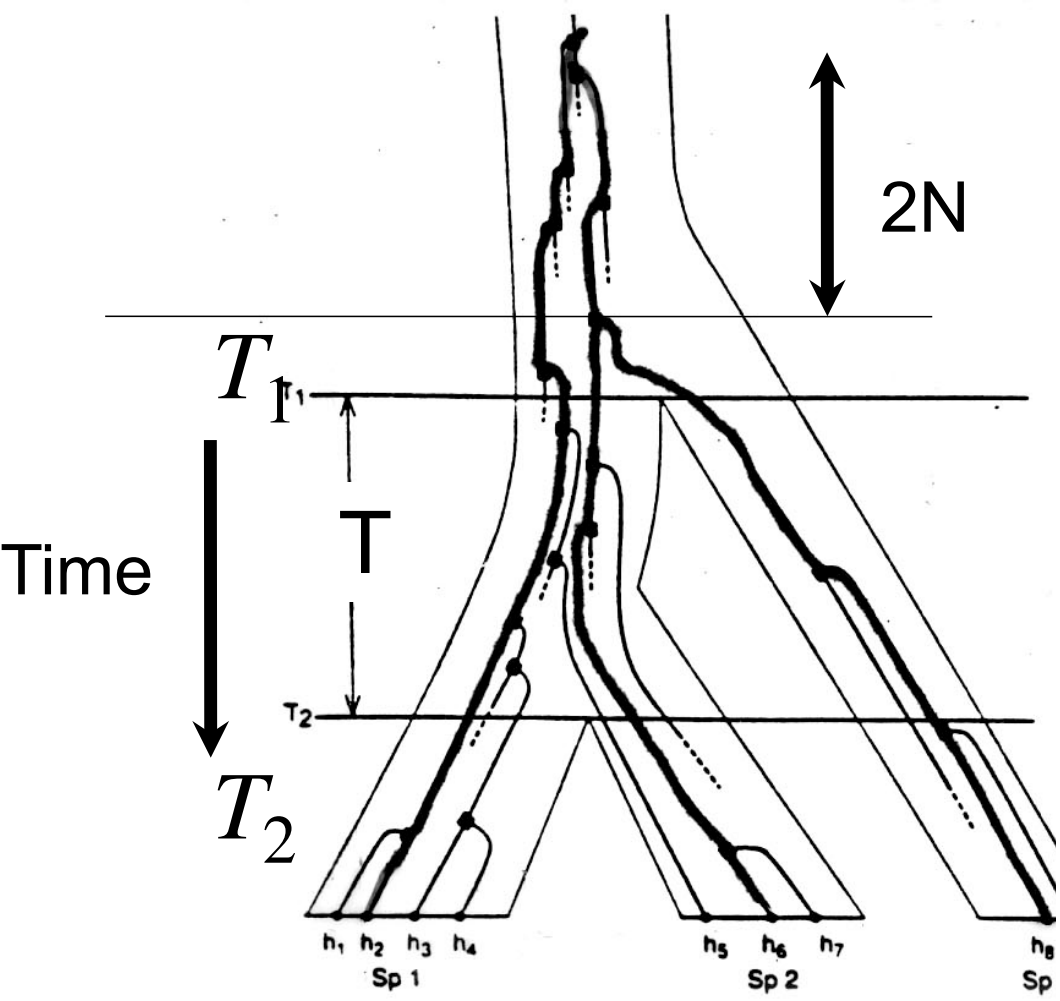
incongruent gene tree

$$Pr(incongruence) = \frac{2}{3}e^{-t/2N}$$

Nei 1987. *Molecular Evolutionary Genetics*
Wu 1991. Genetics 127:429-435
Hudson 1992. Genetics 131:509-512
Yang 2002. Genetics 162:1811-1823

Rannala B, Yang Z. 2003. Genetics 164, 1645-1656.

$$P_{incongruence} = 2/3 * e^{-t/2N}$$

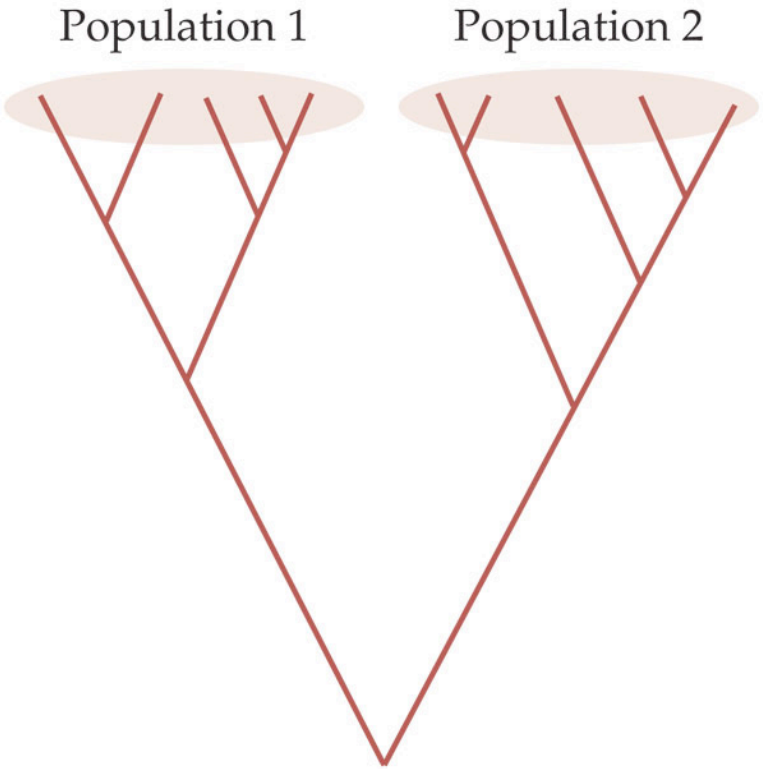


1. Pr(A&B do not coalesce in interval **T**)

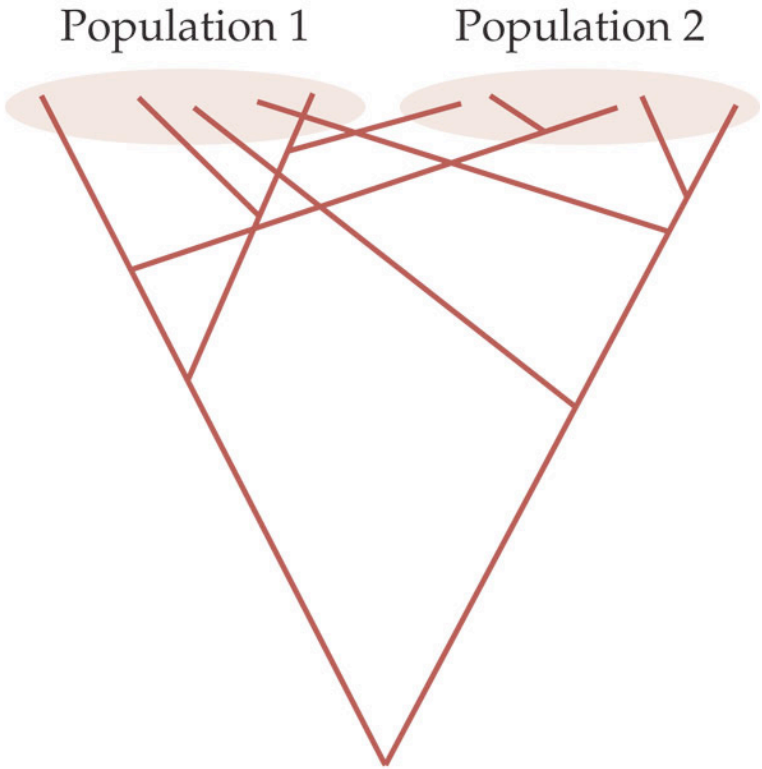
and

2. Pr(A&B **do not** coalesce before A&C **or** B&C)

(A) Old divergence, little gene flow

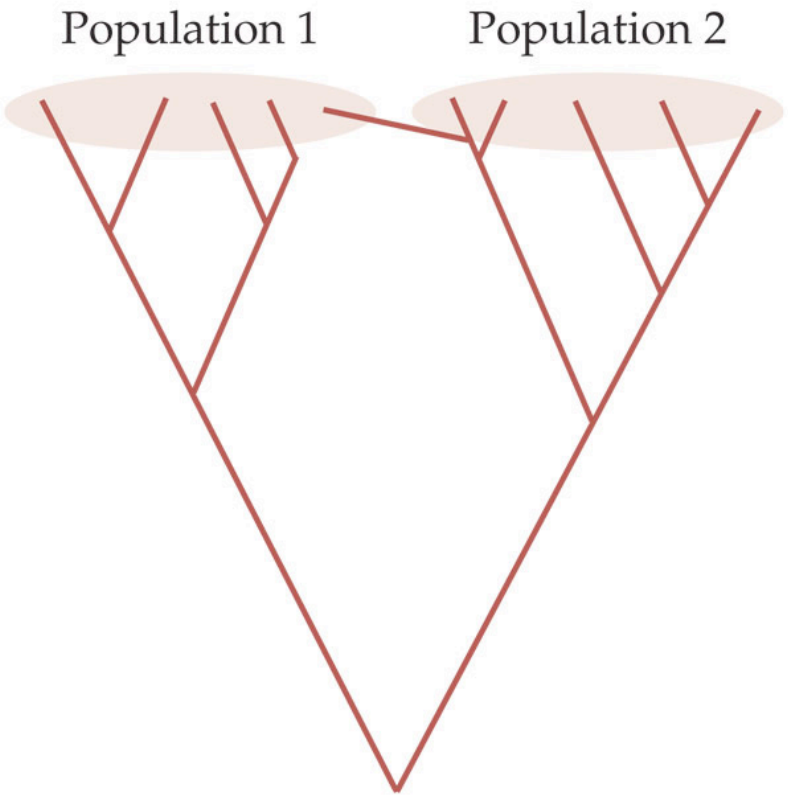


(B) Strong gene-flow, panmixia, very recent divergence

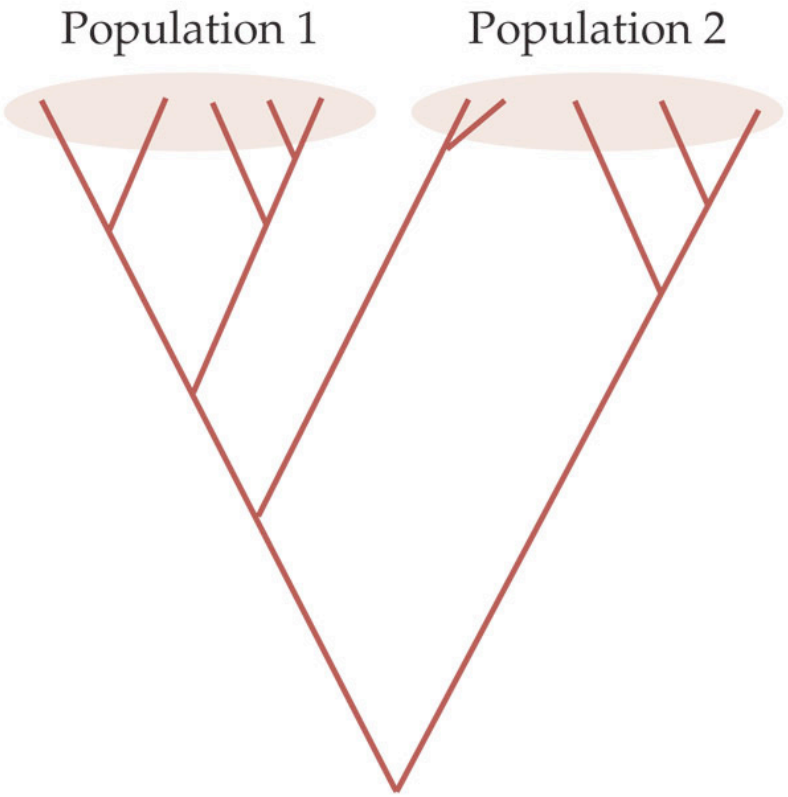


***INTRODUCTION TO POPULATION GENETICS*, Figure 5.8 (Part 1)**

(C) Old divergence, recent gene-flow

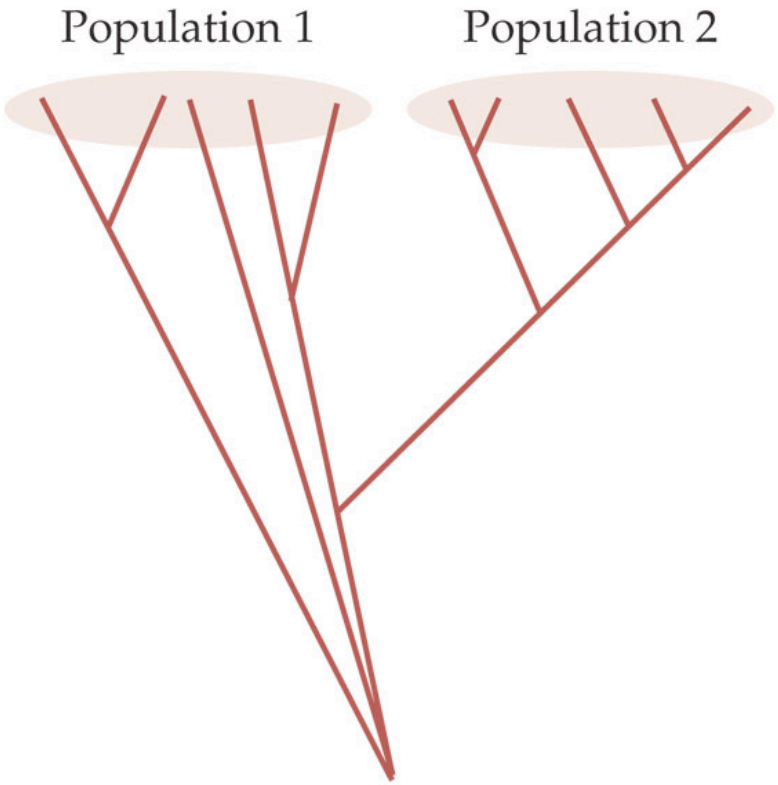


(D) Ongoing gene-flow, old divergence or recent divergence

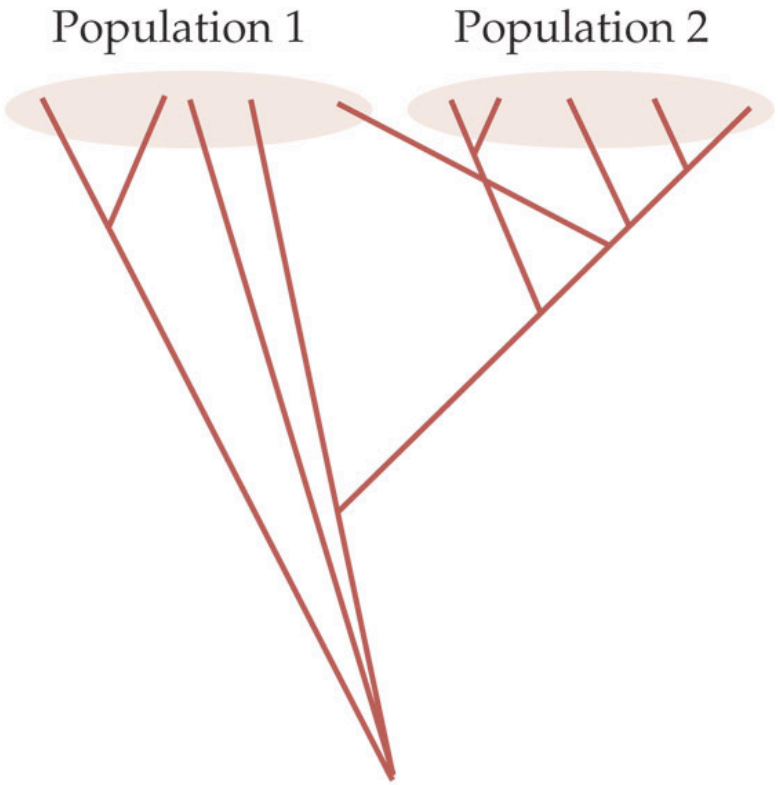


***INTRODUCTION TO POPULATION GENETICS*, Figure 5.8 (Part 2)**

(E) Old divergence or ongoing gene-flow, low N_e in population 2

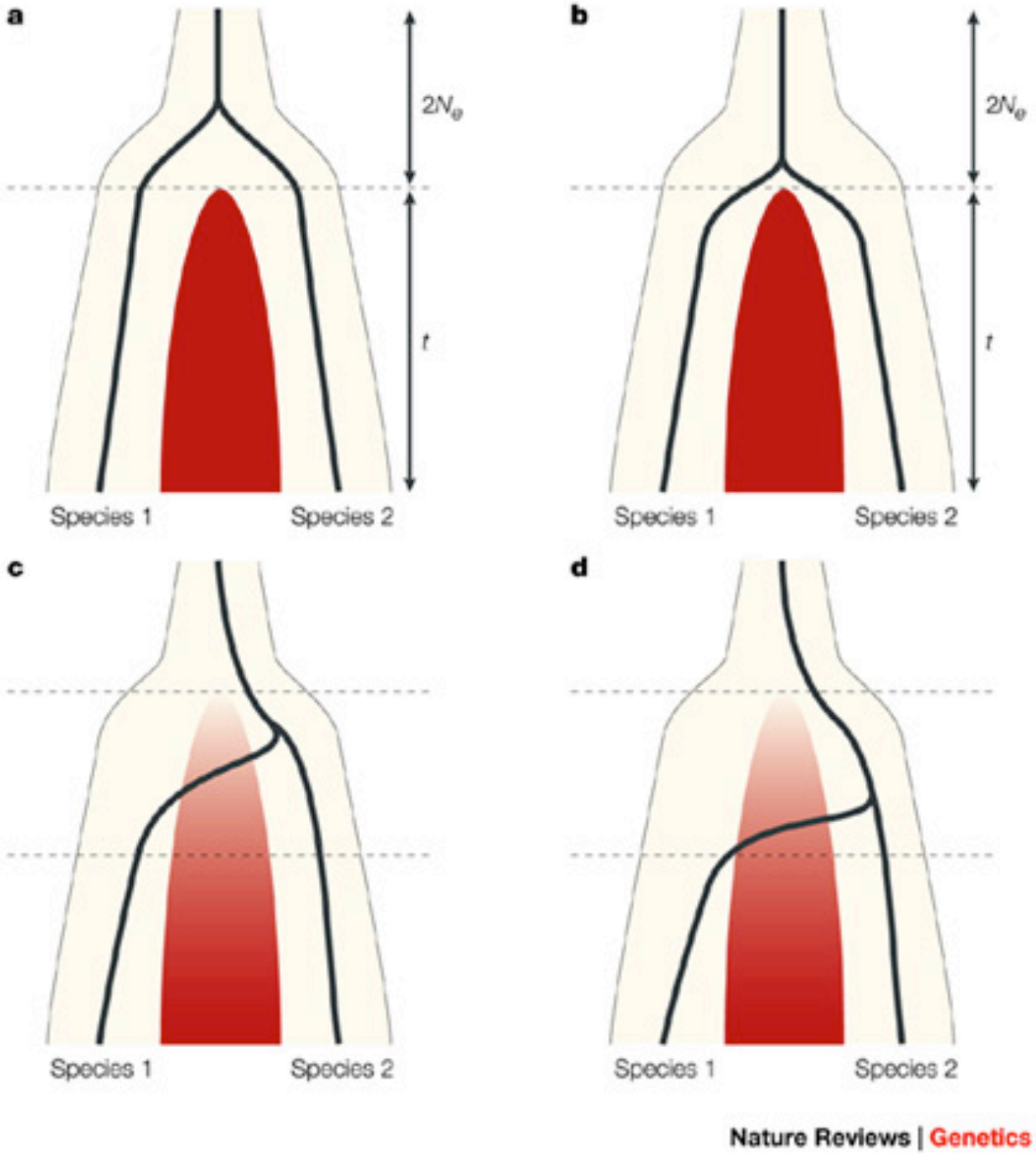


(F) Recent divergence or ongoing gene-flow, low N_e in population 2

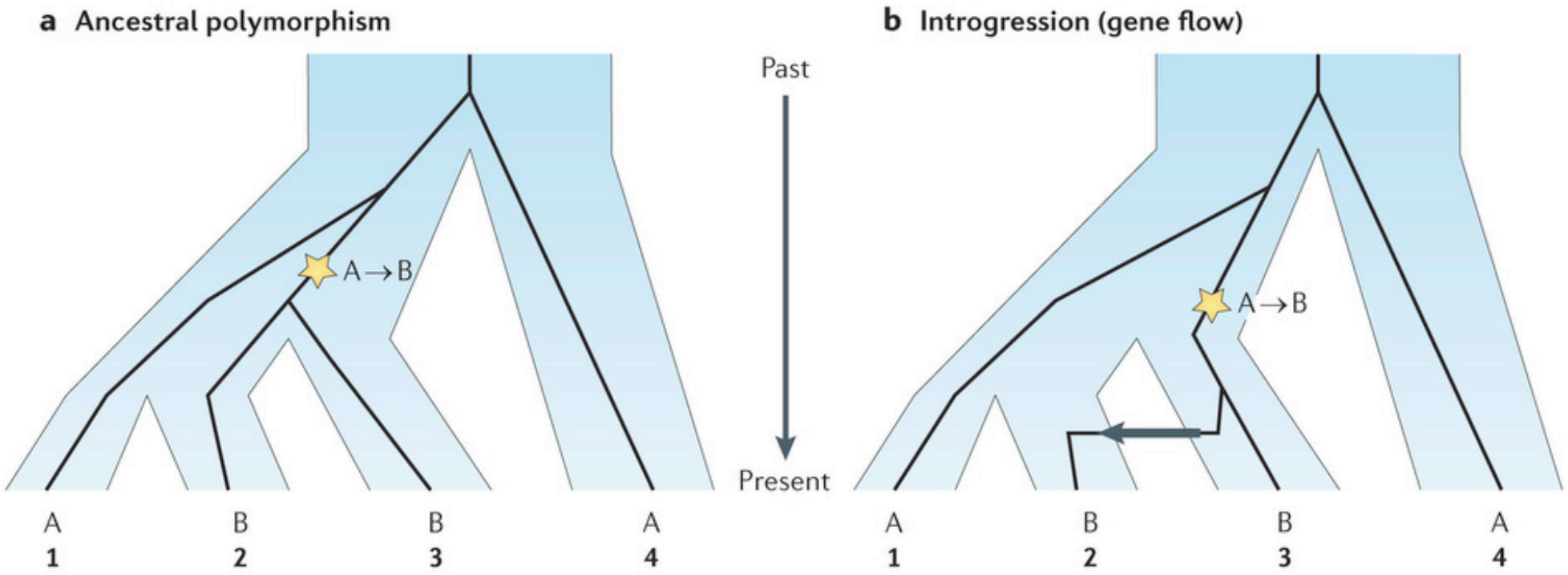


***INTRODUCTION TO POPULATION GENETICS,* Figure 5.8 (Part 3)**

Speciation w/ gene flow

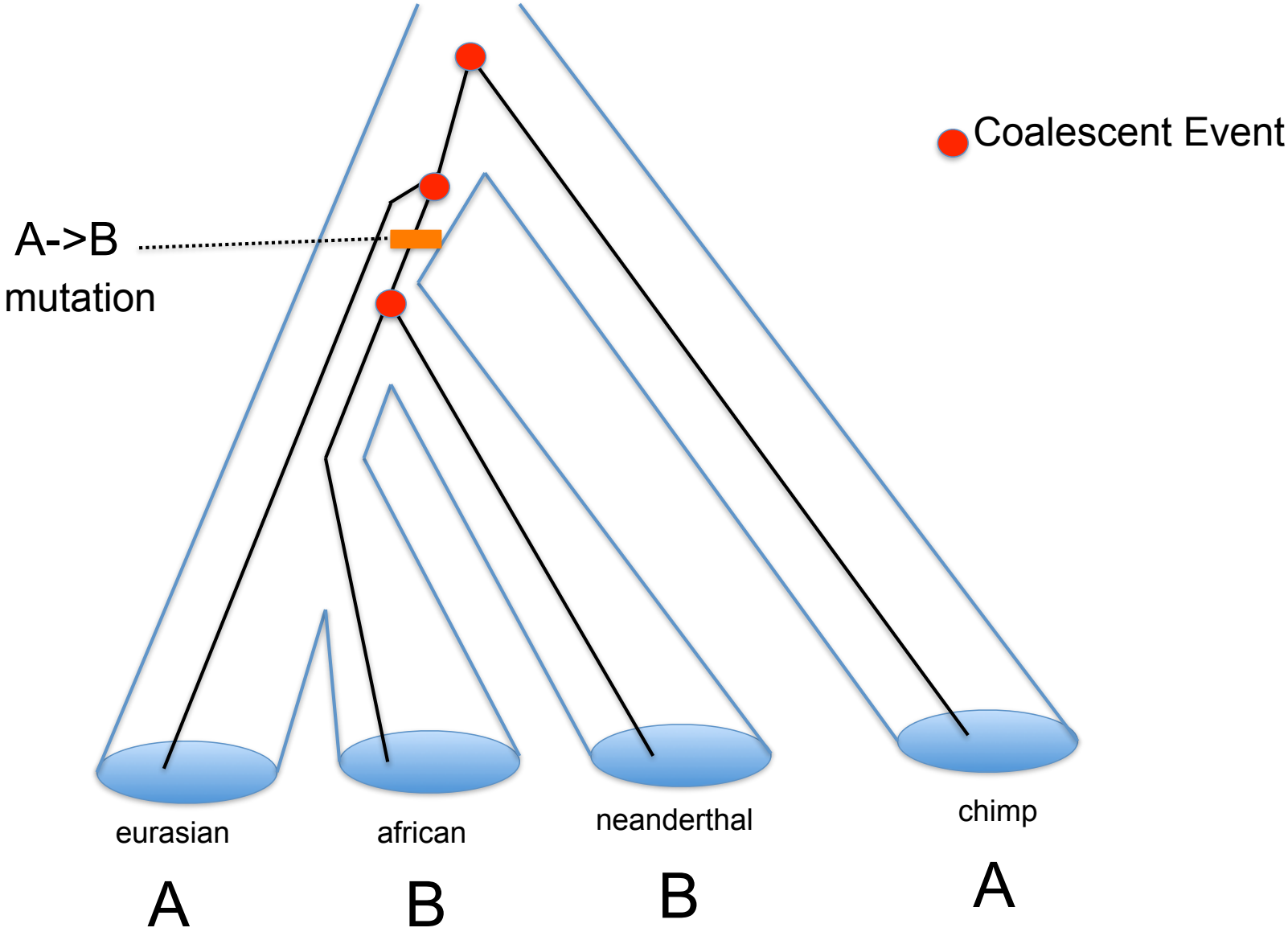


ABBA/BABA test

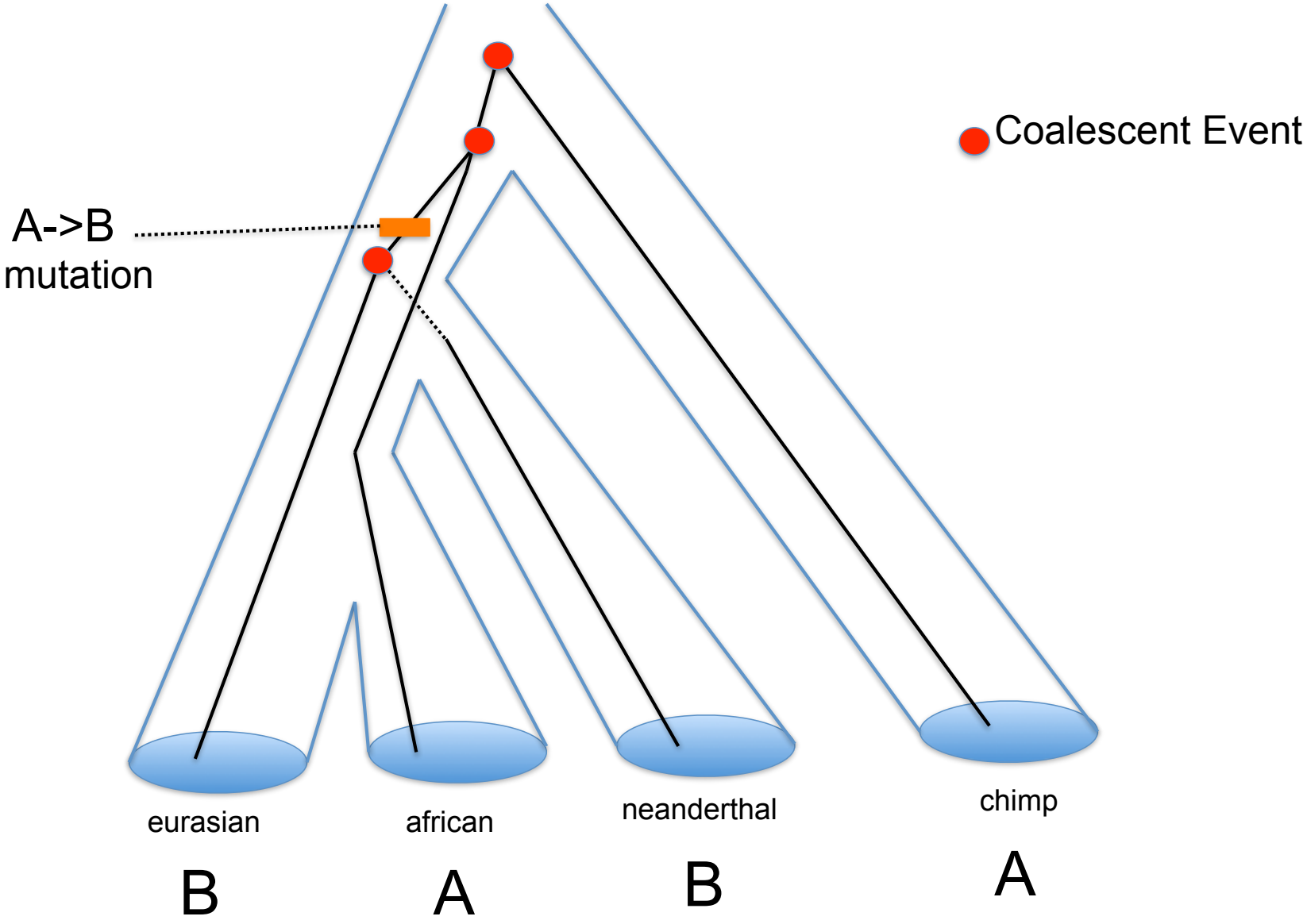


Did Humans – Neanderthals mix?

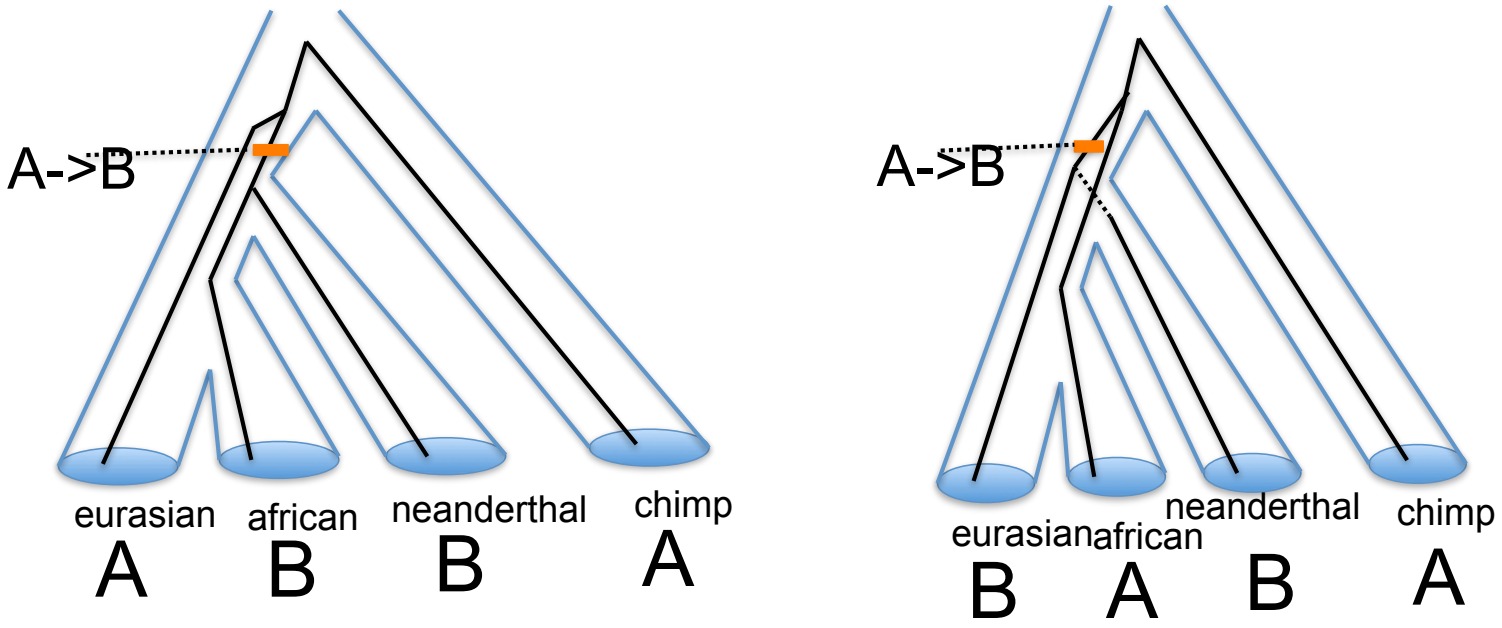
focus on mutations along the internal branch



Did Humans – Neanderthals mix?



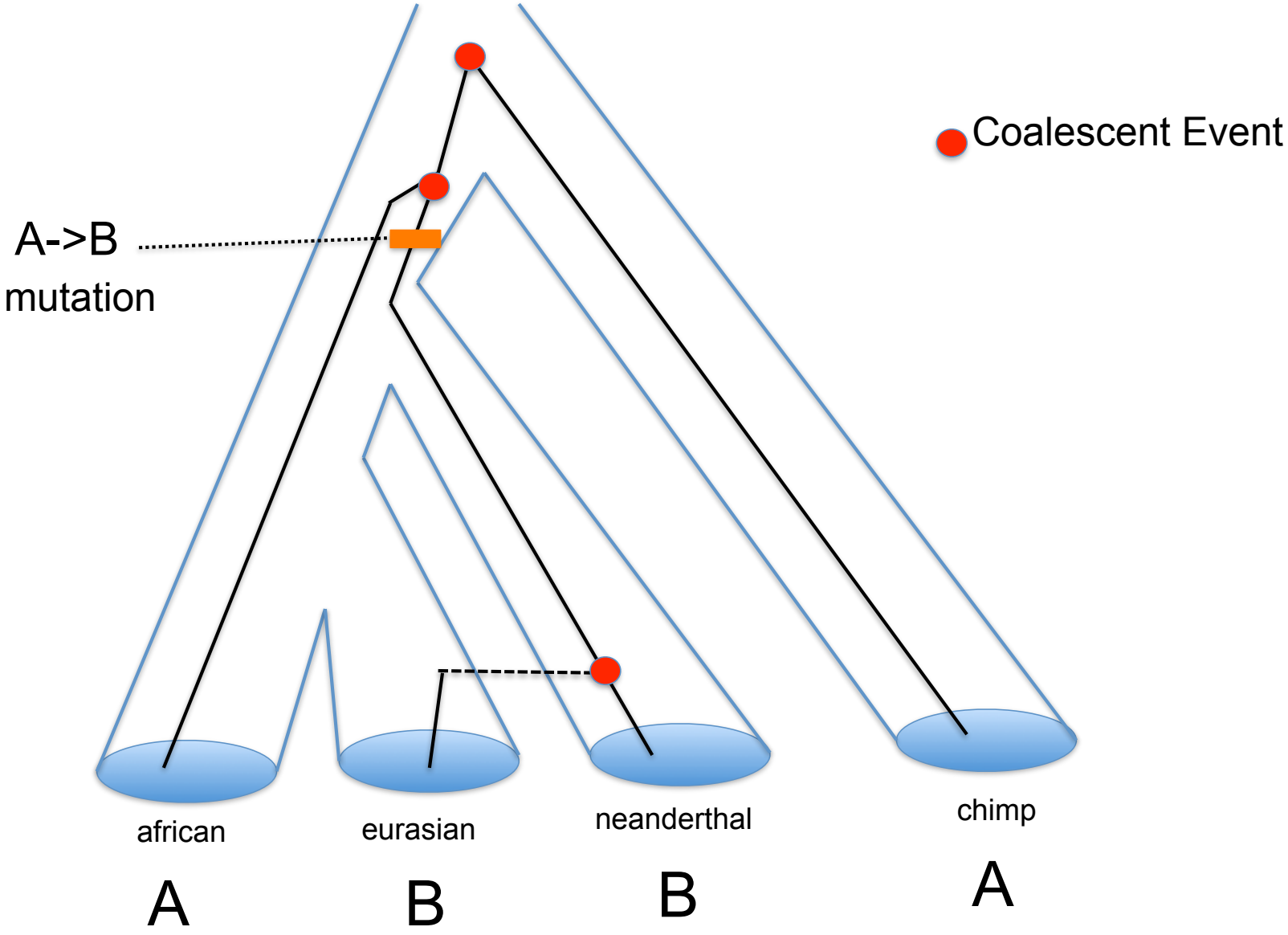
Given no admixture, should be 50% - 50%



ABBA and BABA equally likely if it is Ancestral polymorphism

50%/50% expectation under ancestral polymorphism
Out of 176,000 SNPs, one compares the observed proportion to the binomial expectation

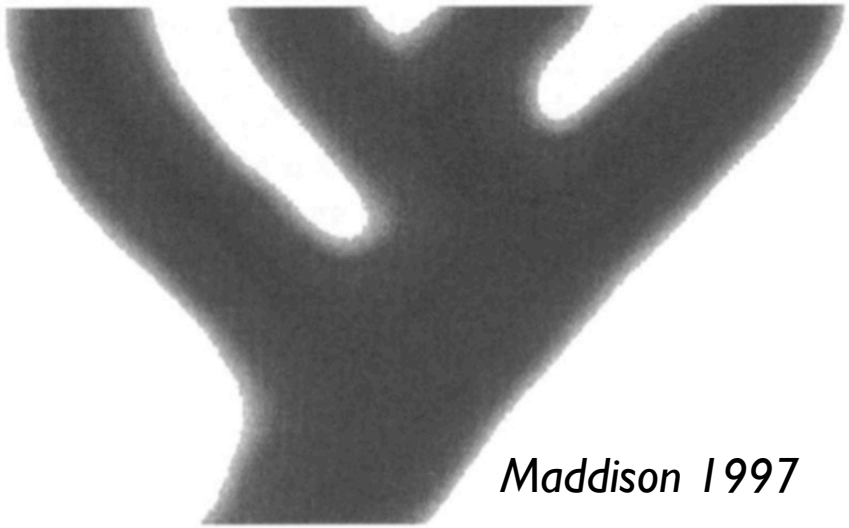
Did Humans – Neanderthals mix?



Bottom line - inference of population history from only 1 or a few loci or 1 sumstat is hazardous

approach	**problems**
Summary statistic/ simulation	lose a lot of information Can not distinguish histories based on one sumstat
using the estimated gene tree	Can not interpret the history based on one or a few unlinked gene trees statistical uncertainty in gene tree estimation step (inference only as good as the tree is)

Bottom line - inference of population history from only 1 or a few loci or 1 sumstat is hazardous



Maddison 1997

FIGURE 9. Phylogeny as a cloud of gene histories. Phylogeny is more like a statistical distribution than a simple tree of discrete thin branches. It has a central tendency, but it also has a variance because of the diversity of gene trees. Gene trees that disagree with the central tendency are not wrong; rather, they are part of the diffuse pattern that is the genetic history.

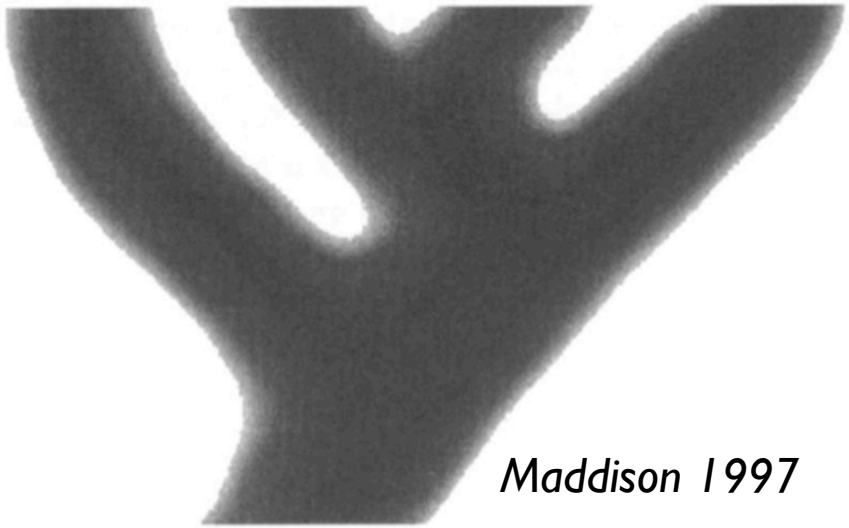
make likelihood function,

$$L = Pr(data \mid \Theta)$$

where Θ = population history parameters

Treat the population history like you would a phylogenetic tree in a likelihood frame work

account for coalescent stochasticity



Maddison 1997

FIGURE 9. Phylogeny as a cloud of gene histories. Phylogeny is more like a statistical distribution than a simple tree of discrete thin branches. It has a central tendency, but it also has a variance because of the diversity of gene trees. Gene trees that disagree with the central tendency are not wrong; rather, they are part of the diffuse pattern that is the genetic history.

$$\prod_{\text{loci}} \sum_{\substack{\text{possible} \\ \text{gene trees}}} [P(\text{sequences}|\text{gene tree}) \cdot P(\text{gene tree}|\text{species tree})]$$

Maddison 1997

$$\Pr(X|\Theta) = \int_G \Pr(X|G)p(G|\Theta)\,dG$$

G = coalescence gene trees

X = sequence alignments

Θ = species tree & parameters

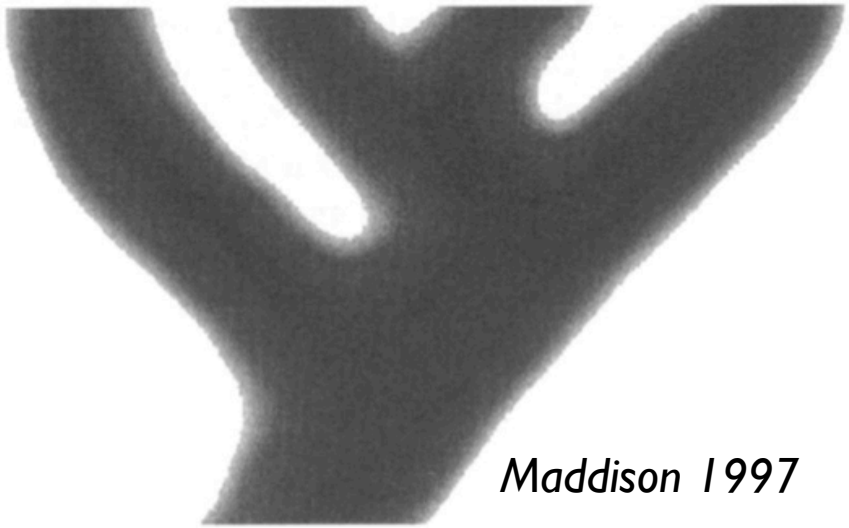
$$\Pr(X \mid \Theta) = \int_G \Pr(X \mid G) p(G \mid \Theta)\, dG$$

G = coalescent gene trees

considers all possible trees (weighted by their likelihood)

can be used to test hypotheses re: population history (ie Θ_1 vs Θ_2)

can be very hard to calculate for all possible G (often use MCMC)



Maddison 1997

FIGURE 9. Phylogeny as a cloud of gene histories. Phylogeny is more like a statistical distribution than a simple tree of discrete thin branches. It has a central tendency, but it also has a variance because of the diversity of gene trees. Gene trees that disagree with the central tendency are not wrong; rather, they are part of the diffuse pattern that is the genetic history.

sometimes even MCMC is too difficult

Too many parameters to calculate likelihood function

…… but we can simulate and approximate (ABC; approximate Bayesian computation or now, supervised machine learning)

simulate data
compress to vector of sumstats
parameters that produce data close to observed
approximates $\Pr(X \mid \Theta)$

To the limit of $D_{simulated} - D_{observed} = 0$

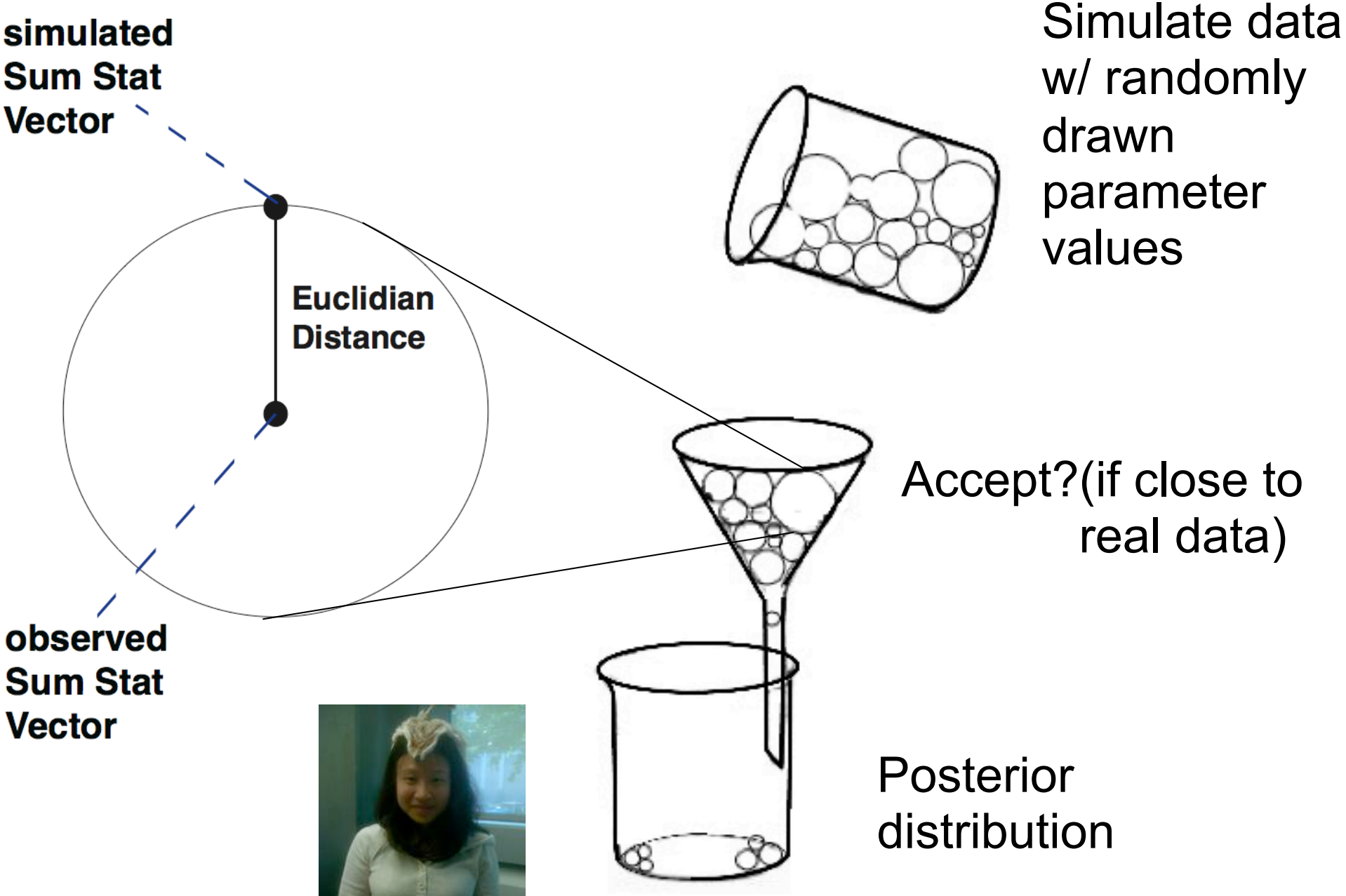
and

as # of simulations $\longrightarrow \infty$

the ABC posterior approaches true posterior

$\mathbf{P(\Phi^* | data)} \longrightarrow \mathbf{P(\Phi | data)}$

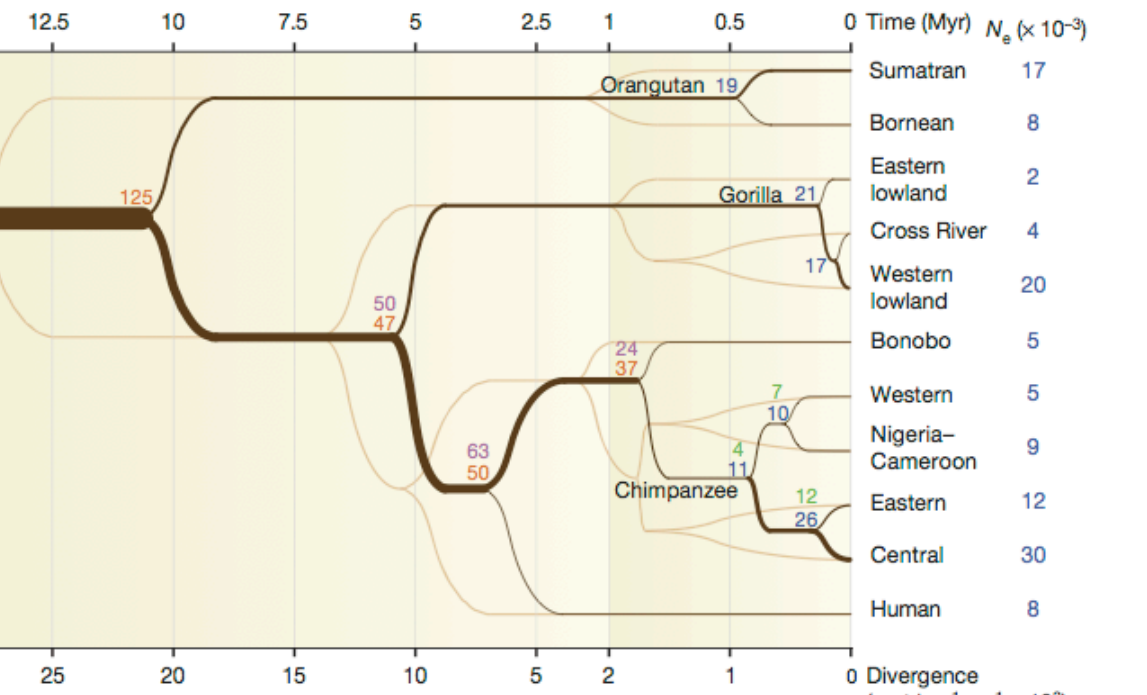
ABC (simulation based inference)

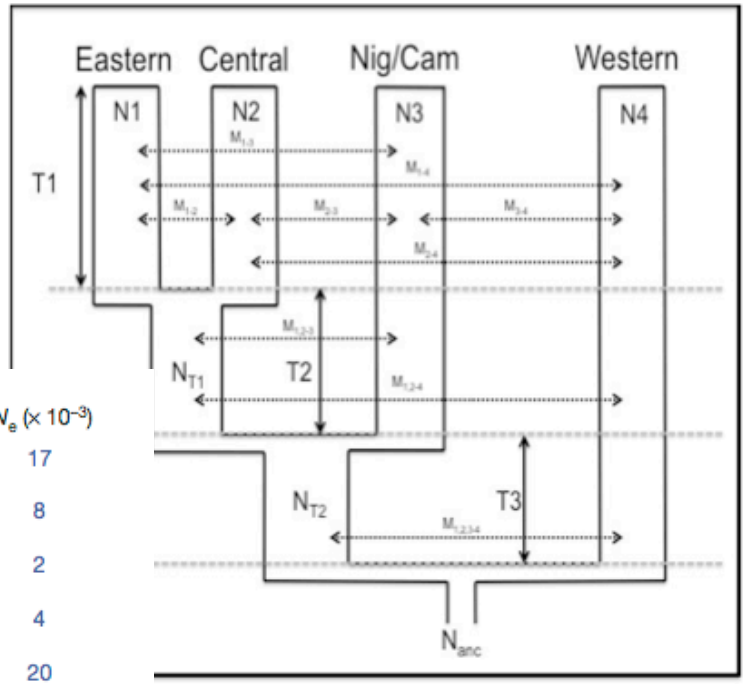


Wen Huang 2008

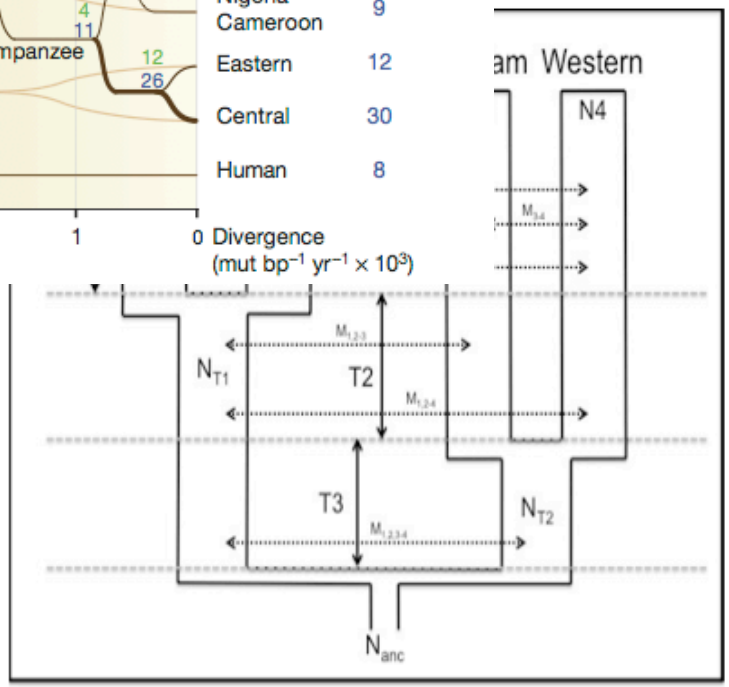
ABC

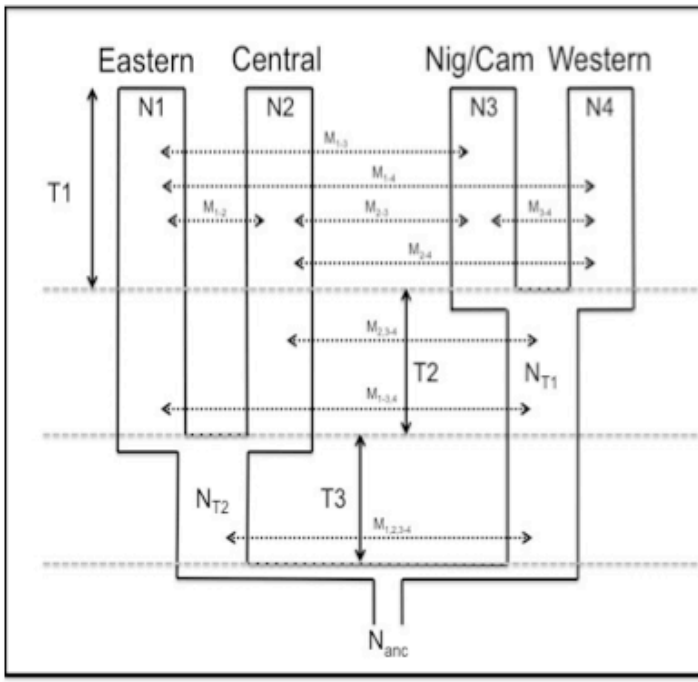
dealing with models that have many many parameters





lodel 3

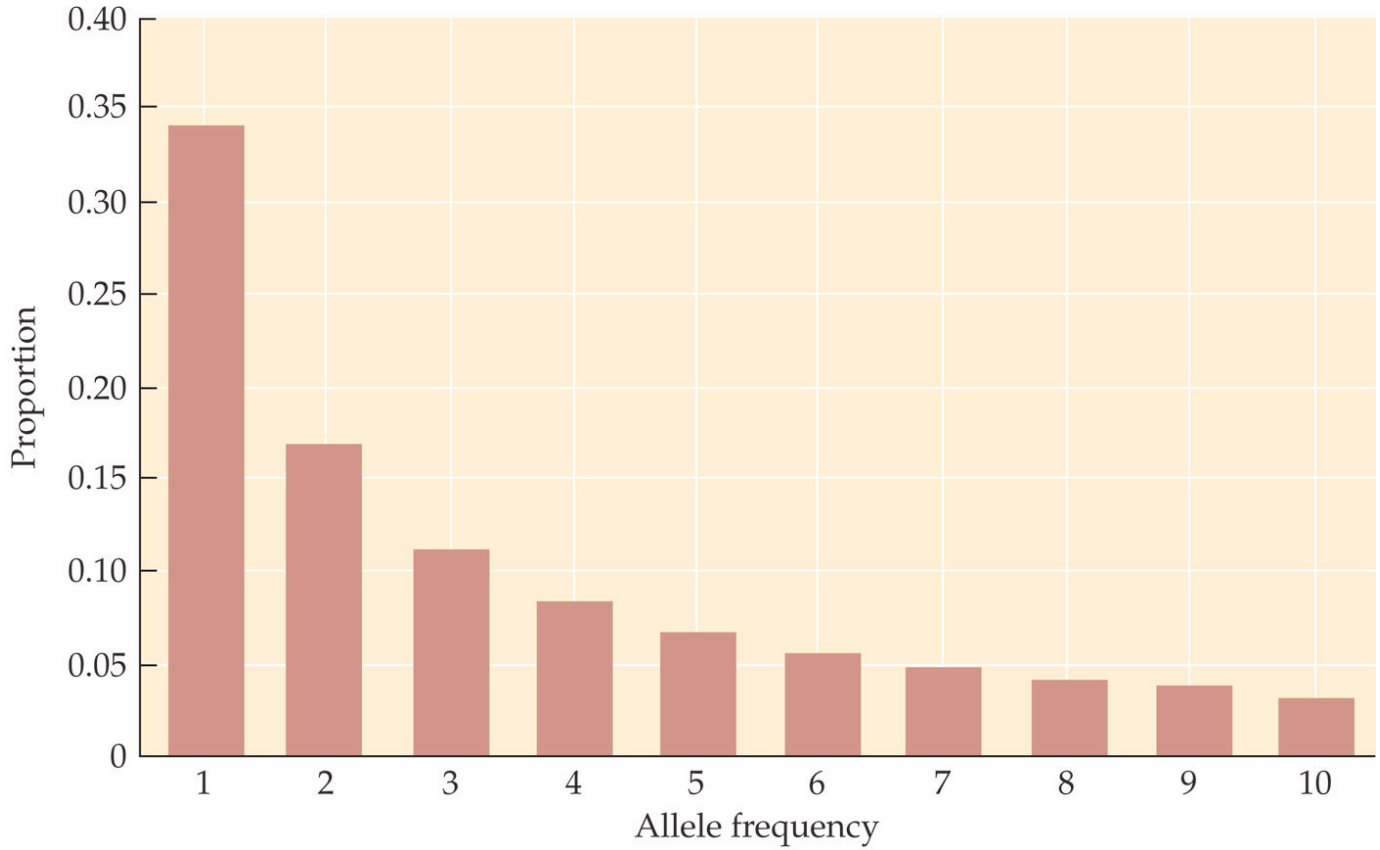




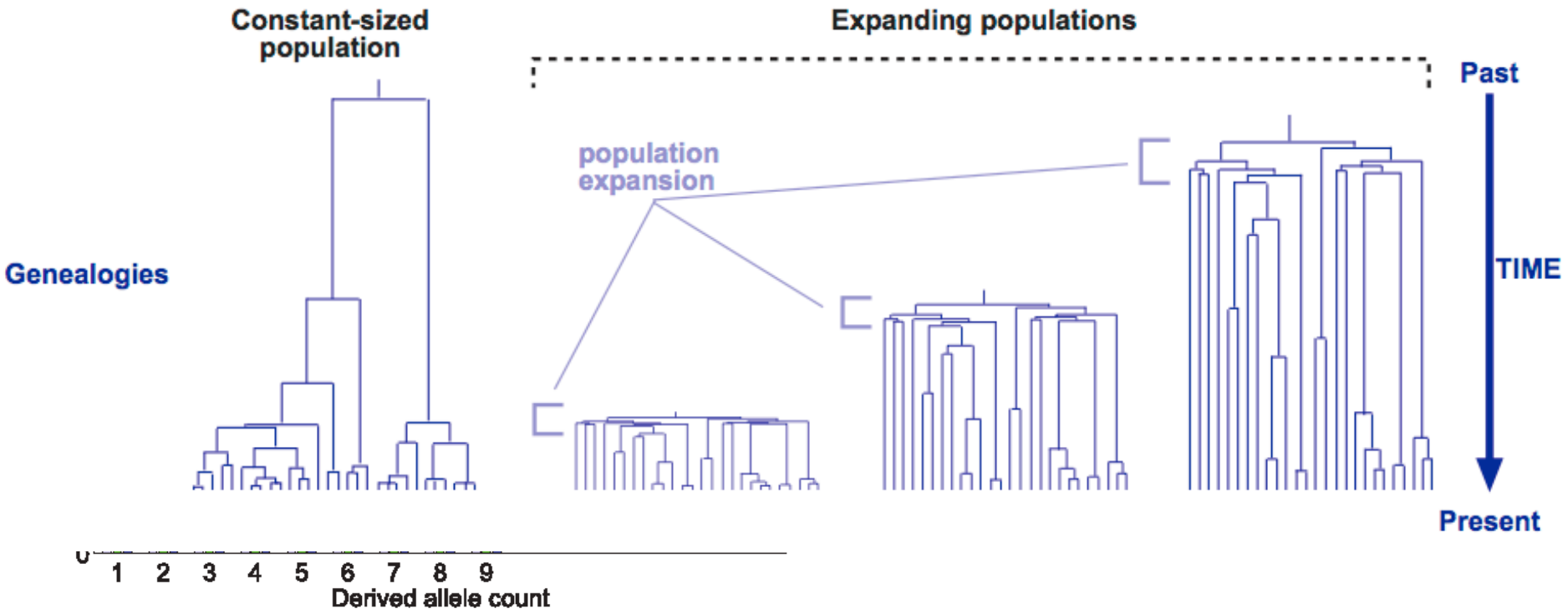
high-throughput data analysis

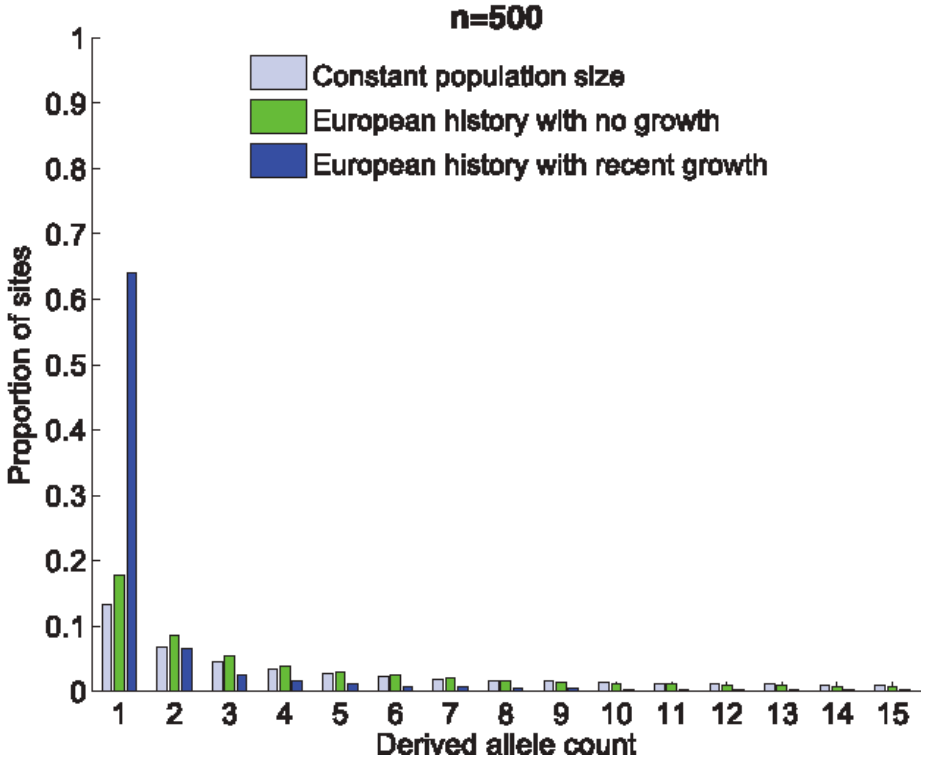
illumina SNP data (explicitly incorporating 10,000s of different coalescent trees becomes impractical)

SFS - Site Frequency Spectrum



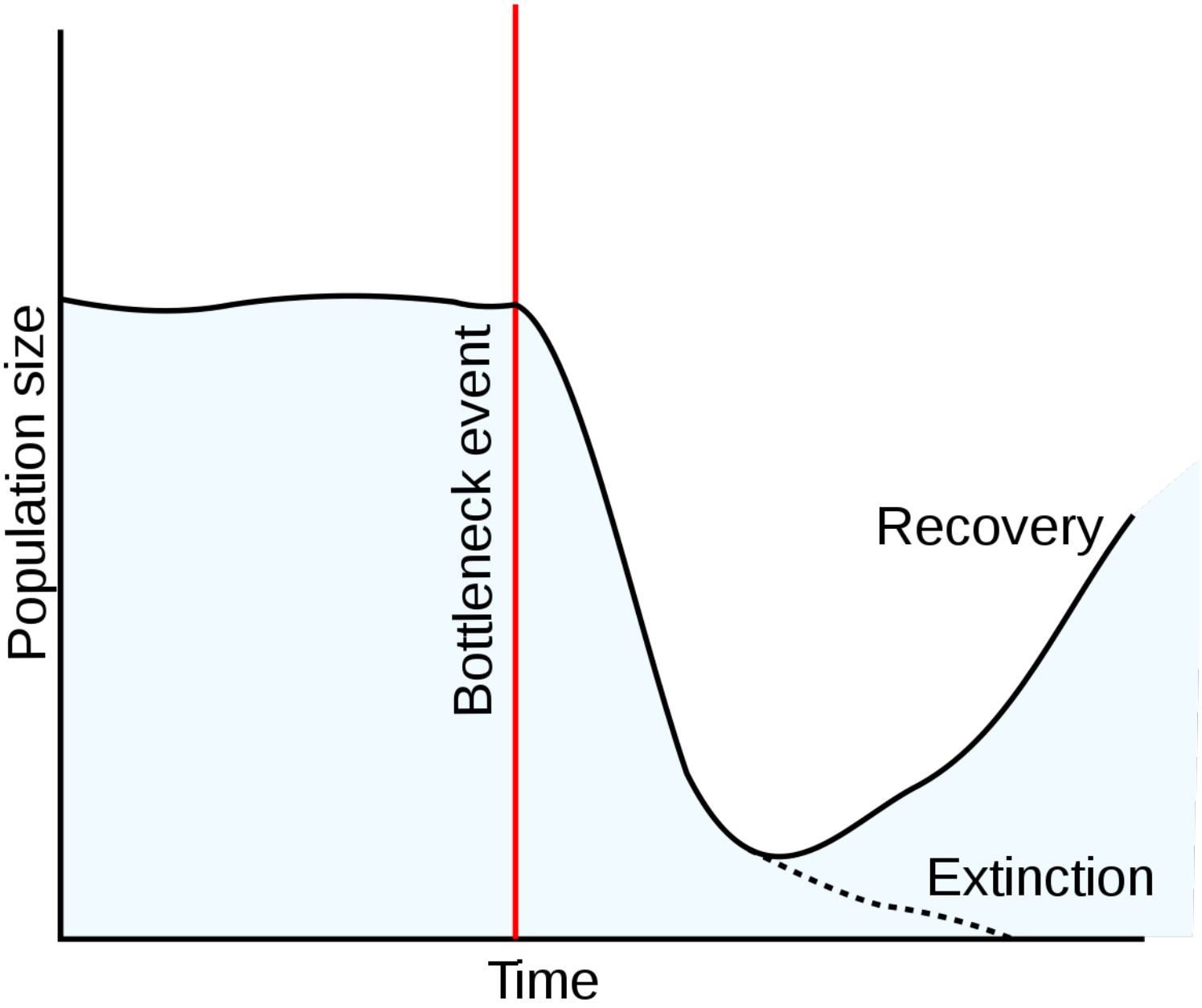
***INTRODUCTION TO POPULATION GENETICS*, Figure 3.9**



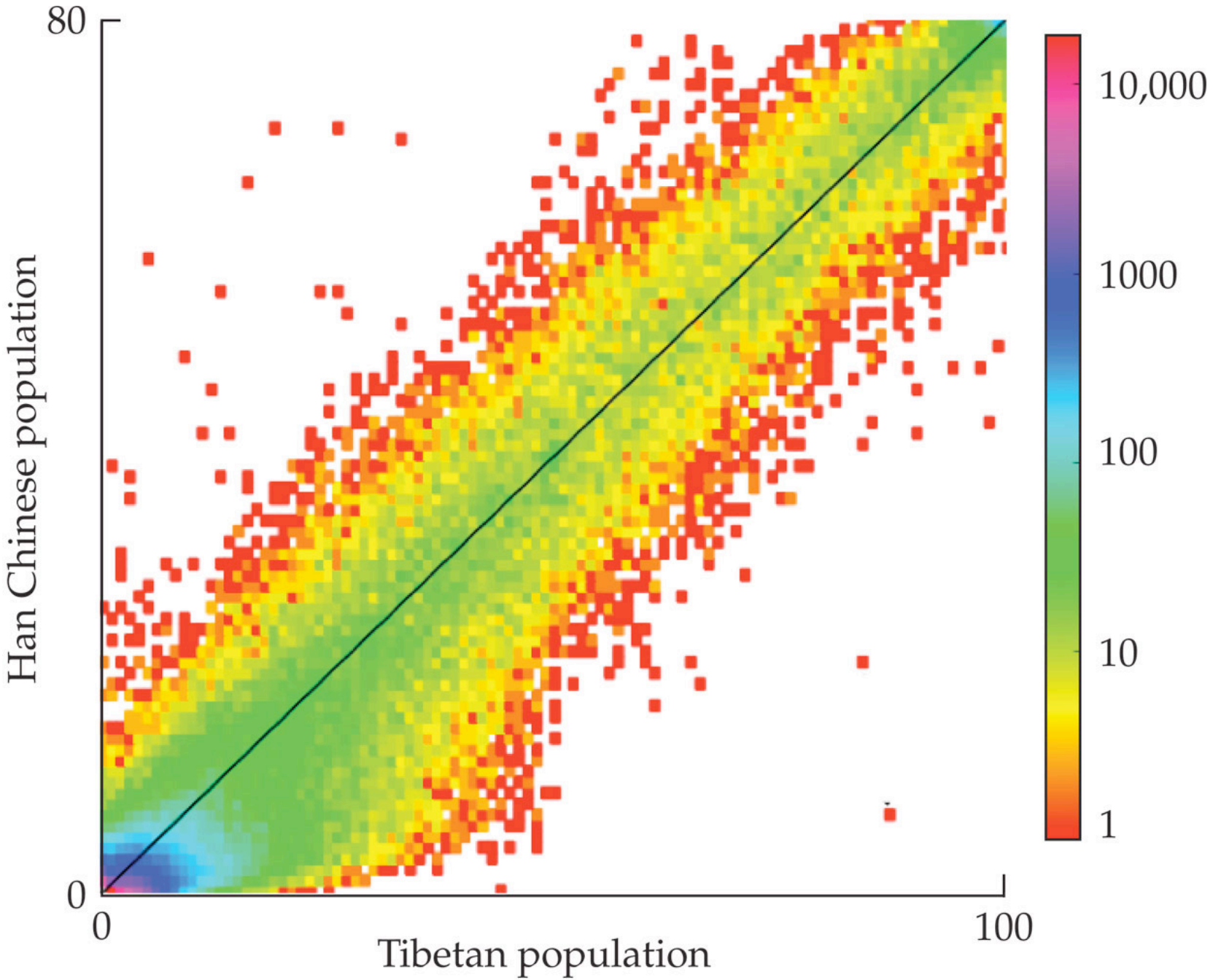


SFS - Site Frequency Spectrum

Bottleneck?, expansion? both ? (it depends)



Joint Site Frequency Spectrum

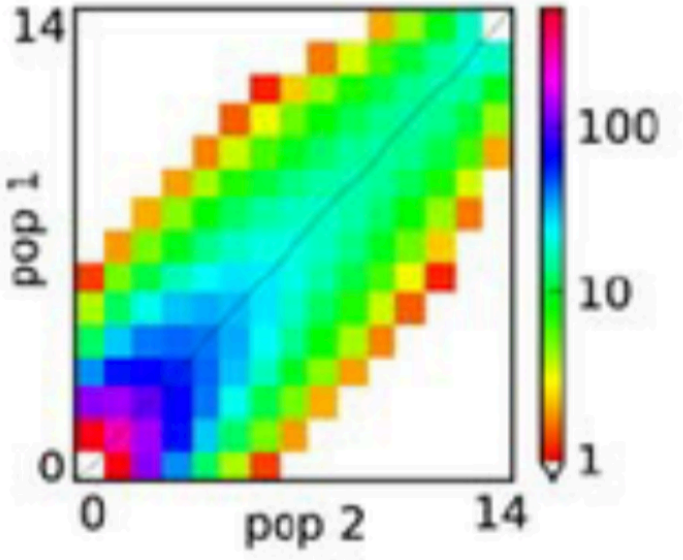


***INTRODUCTION TO POPULATION GENETICS,* Figure 5.14**

The joint site frequency spectrum of 2 populations

- If an allele has frequency f_1 in population 1 and frequency f_2 in population 2, its joint frequency distribution between the two populations is (f_1, f_2)

- The joint site frequency spectrum (of 2D SFS) is the 2-dimensional histogram of these joint site frequencies

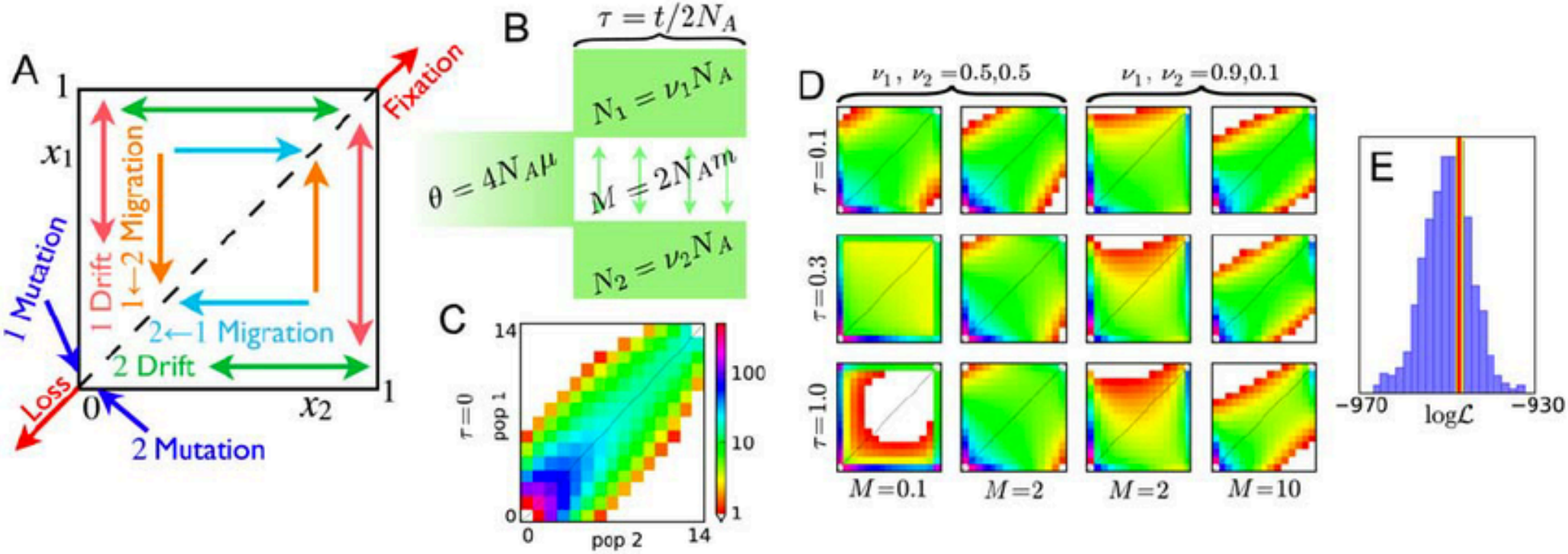


Does it look like these populations diverged very long ago?

Gutenkunst, et al., 2009

% K. Harris

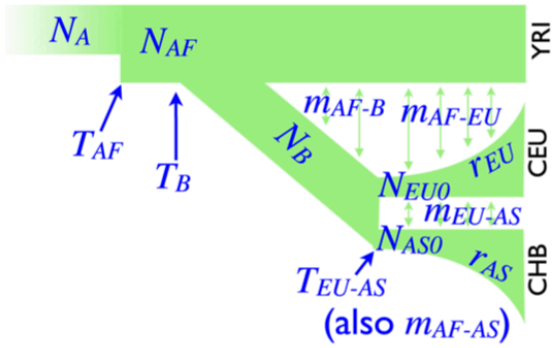
relationship
btw history and
shape of joint SFS

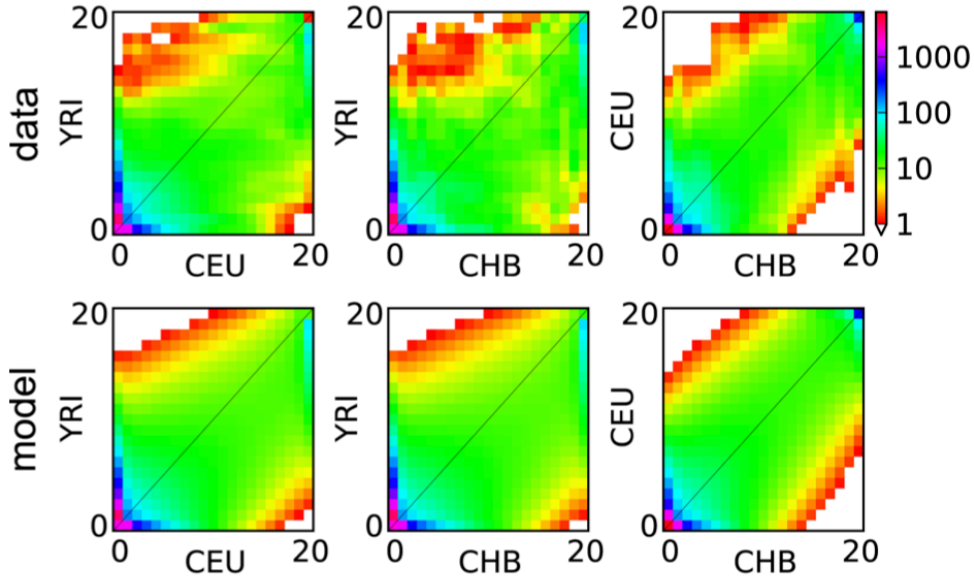


Since small populations diverge faster than large ones, must infer the divergence time jointly with the effective population sizes of the ancestral and present day populations

Gutenkunst et al 2009

An estimate of the joint demographic history of Europeans, Africans, and East Asians





Gutenkunst, et al. 2009

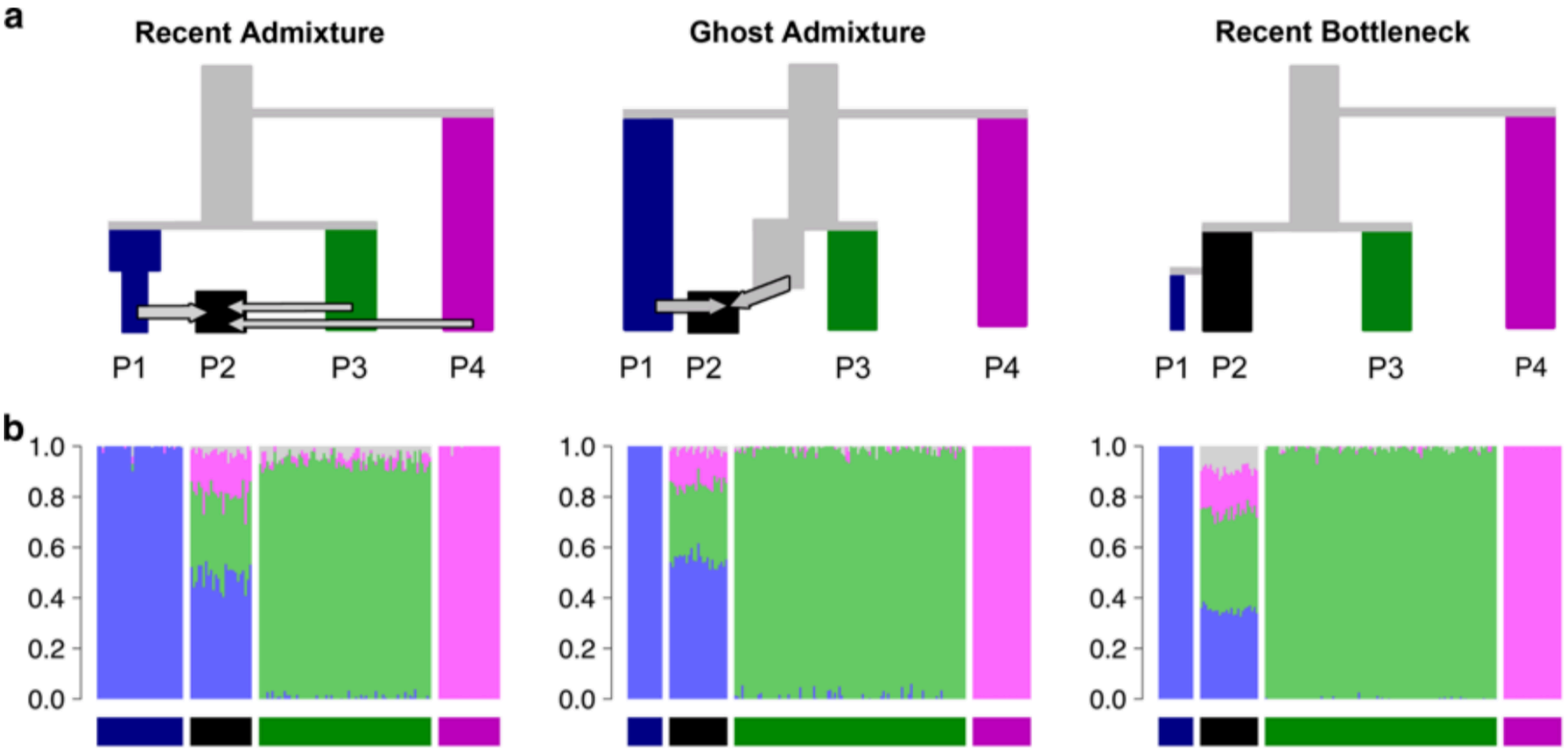
Programs like "dadi" calculate an approximate likelihood of the observed site frequency spectrum given a demographic model and optimize the parameters of the model

YRI = Yorubans from Ibadan, Nigeria
CEU = Individuals of European Descent from Utah
CHB = Han Chinese from Beijing

Why do you think a sample of Yorubans from Ibadan, Nigeria are being used as a proxy for all Africans?

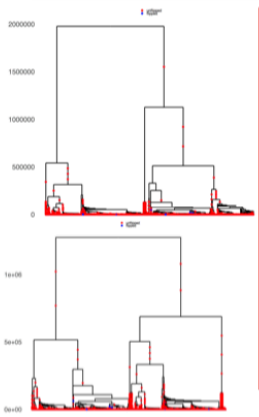
% K. Harris

Remember, STRUCTURE is agnostic to specific demographic histories. (i.e. you should use explicit demographic models for inference after assignment)



Lawson et al 2018

weird circularity “work-flow”



A hidden distribution of coalescent trees for a sample of individuals across many loci

Random mutation

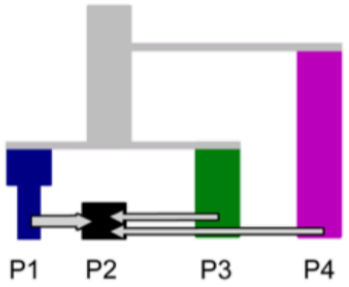
Mutations in observed DNA sequences give us some information about this distribution of coalescent trees

Random Wright-Fisher reproduction

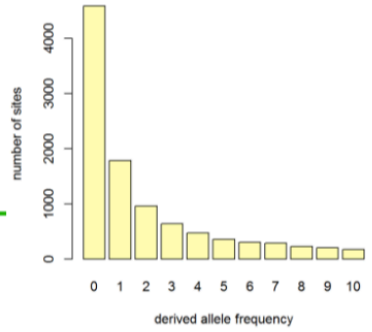
Mutation counting/ data compression

Demographic model
(Including population splits, migration, population bottlenecks population expansions

A summary statistic like the site frequency spectrum (SFS) is like a genomic TLDR, compressing genomic data in a way that preserves information about the underlying coalescent trees



Inference of the model that is most likely to have produced the summarized data

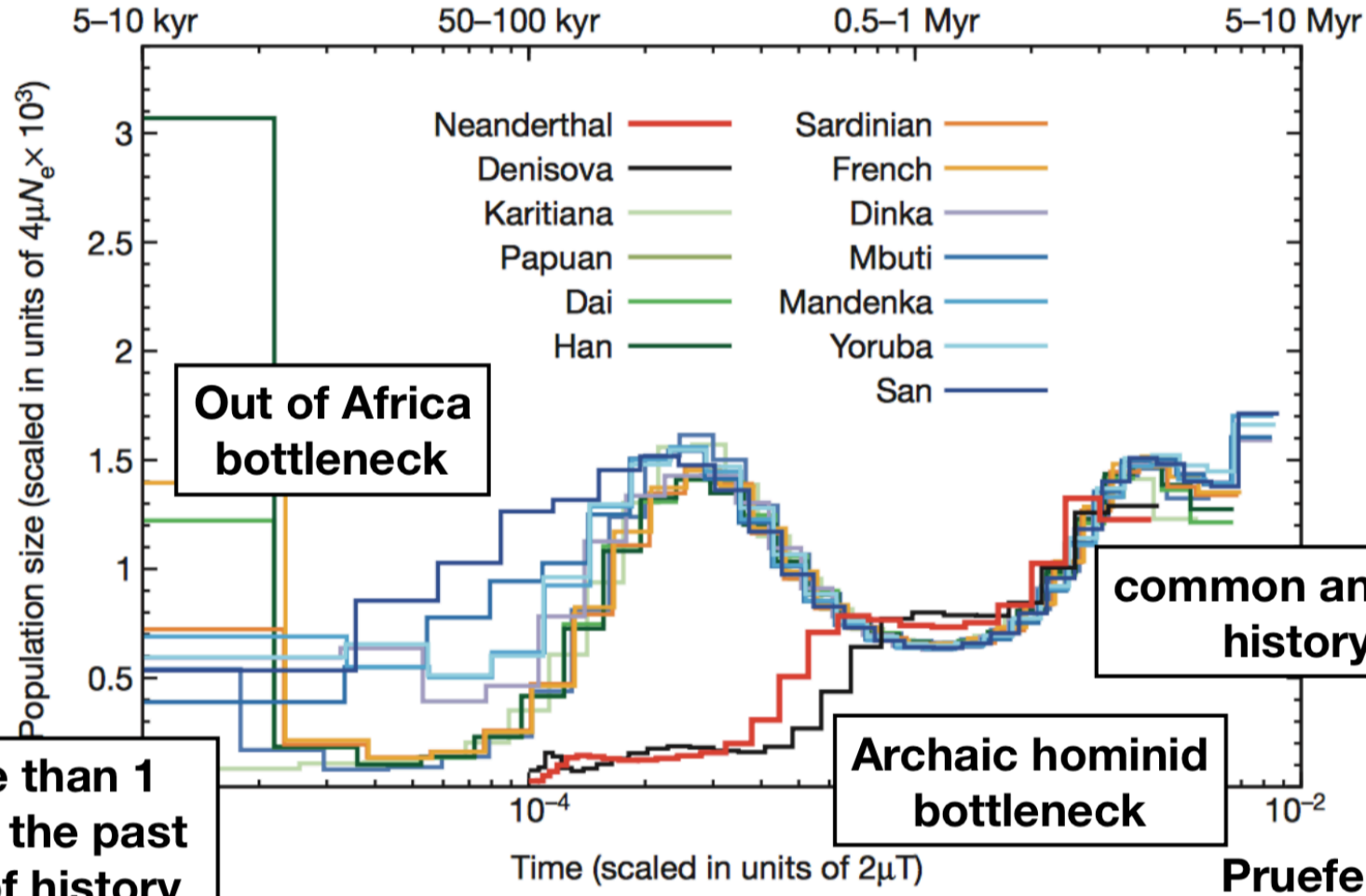


Inference of human population history from individual whole-genome sequences

Heng Li & Richard Durbin

% K. Harris

Nature **475**, 493–496 (28 July 2011) | Download Citation



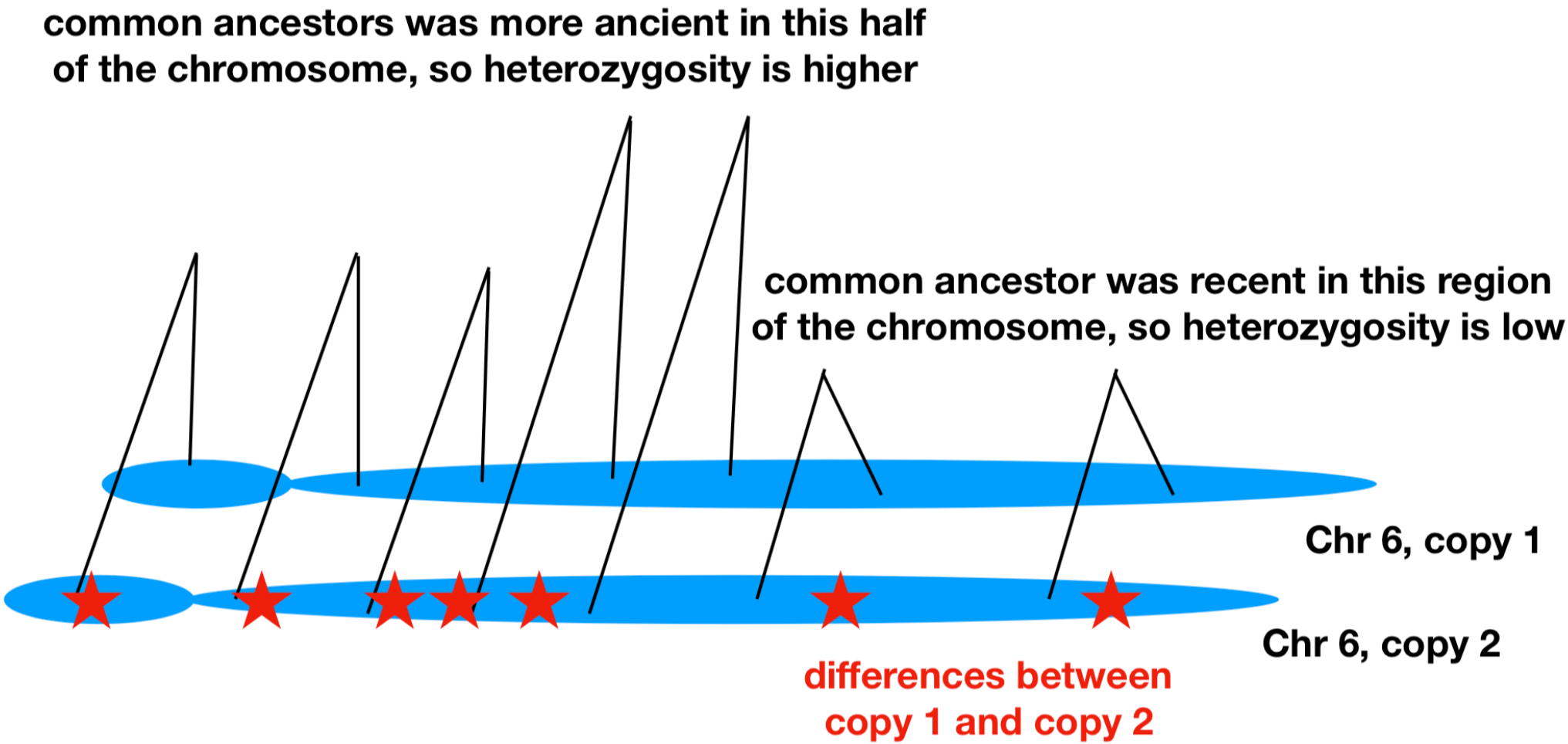
need more than 1 genome for the past 20k years of history

Pruefer, et al. 2014

The PSMC works by leveraging the variation of TMRCA along 1 diploid genome

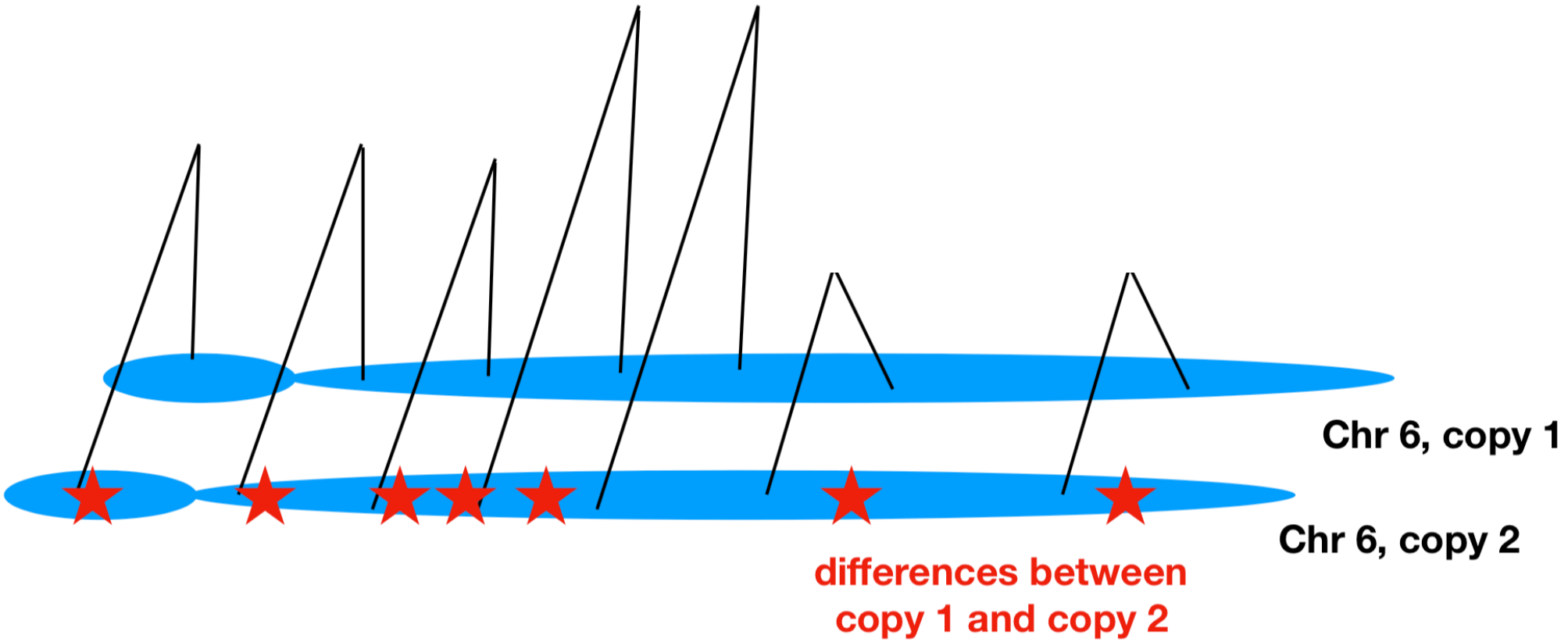
- The genealogical history of 1 genetic locus is just one of many histories that are possible given the population history
- Think about siblings: 1/4 of their genome coalesces 1 generation back, but the rest does not
- To accurately infer demographic history from genetic data, including changes in effective population size over time, information from many independent genetic loci is needed

Variation of coalescent history along the genome



% K. Harris

HMM, a hidden Markov model run along the chromosome to infer the time periods in which each locus coalesces (the “hidden” states)



% K. Harris

Inference of human population history from one genome

- Data from many individuals is required to estimate demographic history from the site frequency spectrum

% K. Harris

Inference of human population history from one genome

- Data from many individuals is required to estimate demographic history from the site frequency spectrum

- But heterozygosity in 1 genome is enough to estimate the average effective population size, and it's actually also enough to infer changes in N_e over time!

% K. Harris

Inference of human population history from one genome

- Data from many individuals is required to estimate demographic history from the site frequency spectrum

- But heterozygosity in 1 genome is enough to estimate the average effective population size, and it's actually also enough to infer changes in N_e over time!

- Recall: if we go back in time to a distant past generation, everyone alive then either has no present-day descendants or is an ancestor of everyone alive today

% K. Harris

Inference of human population history from one genome

- Data from many individuals is required to estimate demographic history from the site frequency spectrum
- But heterozygosity in 1 genome is enough to estimate the average effective population size, and it's actually also enough to infer changes in N_e over time!
- Recall: if we go back in time to a distant past generation, everyone alive then either has no present-day descendants or is an ancestor of everyone alive today
- DNA from a single individual is enough to estimate the effective size of this distant past population

% K. Harris

Demographic change interacts with selection in multiple ways

- When the population size grows, genetic drift slows down and selection becomes more efficient

% K. Harris

Demographic change interacts with selection in multiple ways

- When the population size grows, genetic drift slows down and selection becomes more efficient

- A population bottleneck speeds up genetic drift and makes natural selection get less efficient

% K. Harris

Demographic change interacts with selection in multiple ways

- When the population size grows, genetic drift slows down and selection becomes more efficient
- A population bottleneck speeds up genetic drift and makes natural selection get less efficient
- Population size changes shift the site frequency spectrum away from the shape SFS(n)~1/n that is expected under neutrality

% K. Harris

Demographic change interacts with selection in multiple ways

- When the population size grows, genetic drift slows down and selection becomes more efficient
- A population bottleneck speeds up genetic drift and makes natural selection get less efficient
- Population size changes shift the site frequency spectrum away from the shape SFS(n)~1/n that is expected under neutrality
- Natural selection also shifts the SFS away from the equilibrium 1/n shape

% K. Harris

Demographic change interacts with selection in multiple ways

- When the population size grows, genetic drift slows down and selection becomes more efficient
- A population bottleneck speeds up genetic drift and makes natural selection get less efficient
- Population size changes shift the site frequency spectrum away from the shape SFS(n)~1/n that is expected under neutrality
- Natural selection also shifts the SFS away from the equilibrium 1/n shape
- The SFS shape can be used to test for selection, but only after correcting for nonequilibrium demography

% K. Harris

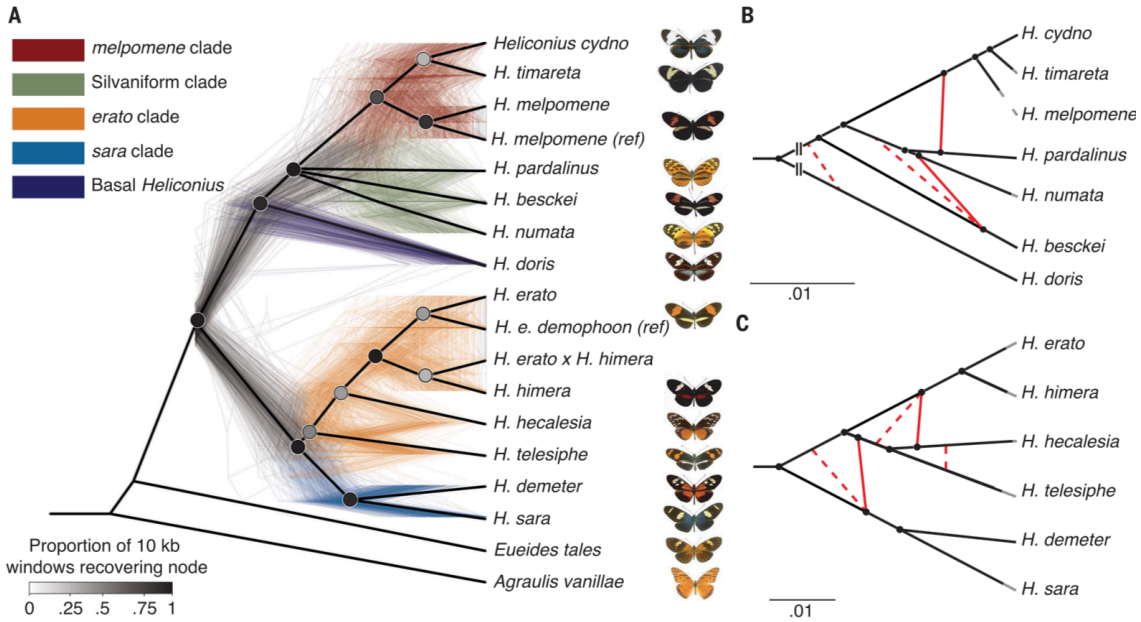
Next Week:

1. Simulation Assignment
2. Discuss paper

BUTTERFLY GENOMICS

Genomic architecture and introgression shape a butterfly radiation

Nathaniel B. Edelman[1]*, Paul B. Frandsen[2,3], Michael Miyagi[1], Bernardo Clavijo[4], John Davey[5,20], Rebecca B. Dikow[3], Gonzalo García-Accinelli[4], Steven M. Van Belleghem[6], Nick Patterson[7,8], Daniel E. Neafsey[8,9], Richard Challis[10], Sujai Kumar[11], Gilson R. P. Moreira[12], Camilo Salazar[13], Mathieu Chouteau[14], Brian A. Counterman[15], Riccardo Papa[6,16], Mark Blaxter[10], Robert D. Reed[17], Kanchon K. Dasmahapatra[5], Marcus Kronforst[18], Mathieu Joron[19], Chris D. Jiggins[20], W. Owen McMillan[21], Federica Di Palma[4], Andrew J. Blumberg[22], John Wakeley[1], David Jaffe[8,23], James Mallet[1]*



Edelman *et al.*, *Science* **366**, 594–599 (2019)

Simulation Assignment: comparing single and double population models

Use msPrime, SLiM or PipeMaster

1 locus, 1000 base pairs
μ=1e-7 per base pair

Simulate a 2 population model
$\tau = 5{,}000$ generations
$N_1 = 10{,}000, N_2 = 10{,}000, N_A = 10{,}000$
$n_1 = 10, n_2 = 10$
no migration

1. Simulate once and calculate an observed π (ignoring which of the 2 populations the 20 individuals come from).
2. Simulate a single population model with that observed π value 1,000 times (pretend you know the mutation rate exactly) and plot the distribution of simulated π values
3. indicate the 'true' $4N\mu$ (π) on the plot. How off is it?
4. Simulate the 2 pop model 1,000 times and choose N values that you think will give you a π distribution that matches the one from the single pop model.
5. plot results of 3 and 4