

## Guest lectures

---

- Mary Kuhner – research faculty in Genome Sciences
- mkkuhner@uw.edu, office number 3-8751 (email preferred)
- Office hours?
- I teach this course in even-numbered years
- I will give 3 lectures on coalescent theory and 1 on chromosome evolution

# Coalescent theory

---

- Basic idea—backwards in time genetics
- Gray whale example
- Deriving the coalescent from the Wright-Fisher model
- Uses and extensions of the coalescent:
  - Population size (genetic drift)
  - Population growth/shrinkage
  - Gene flow (migration)
  - Divergence of populations
  - Recombination
  - Selection
- Red drum example

## Basic idea of coalescent theory

---

- All term long you've looked at problems posed like this:
  - Given a starting situation–
  - And a given set of evolutionary forces–
  - What are the likely outcomes?
- This is a “forward time” approach to population genetics

## Basic idea of coalescent theory

---

- Suppose that I have data from an existing population and want to know how it got that way
- I don't know the starting situation
- Running forward from hypothetical starting situations may or may not ever give me my current data
- I would be better off with a “backward time” approach that started from what I know (current data)

## Basic idea of coalescent theory

---

- Consider the ancestry of my current gene copies backward in time:
  - Which ones are more or less closely related?
  - How big are the chunks that have the same common ancestor?
  - How long ago are their common ancestors?
- The answers to these questions contain a surprising amount of information about past evolutionary forces:
  - Population size (drift)
  - Gene flow
  - Selection
  - Recombination

## **What was the long-term population size of gray whales?**



Alter, Rynes and Palumbi (2007) DNA evidence for historic population size and past ecosystem impacts of gray whales. PNAS 104: 15162-15167.

## **What was the long-term population size of gray whales?**

---

- How many gray whales pre-whaling?
- Whaling ship records not conclusive
- Recent slowing of the observed growth rate may suggest recovery
- Molecular data an alternative source of information

## **What was the long-term population size of gray whales?**

---

- 10 loci:
  - 7 autosomal
  - 2 X-linked
  - 1 mtDNA
- Complex mutational model with rate variation among loci
- Complex population model with subdivision and copy number
- Complex demographic model relating  $N_{census}$  to  $N_e$

## What was the long-term population size of gray whales?

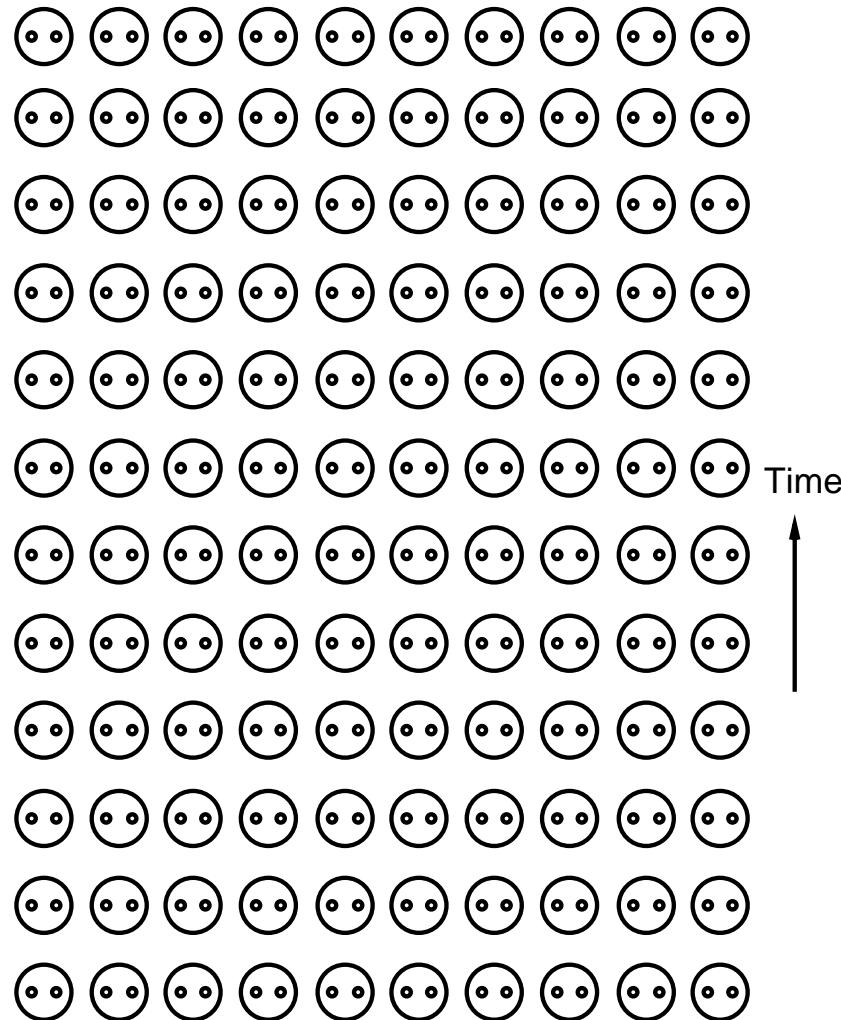
---

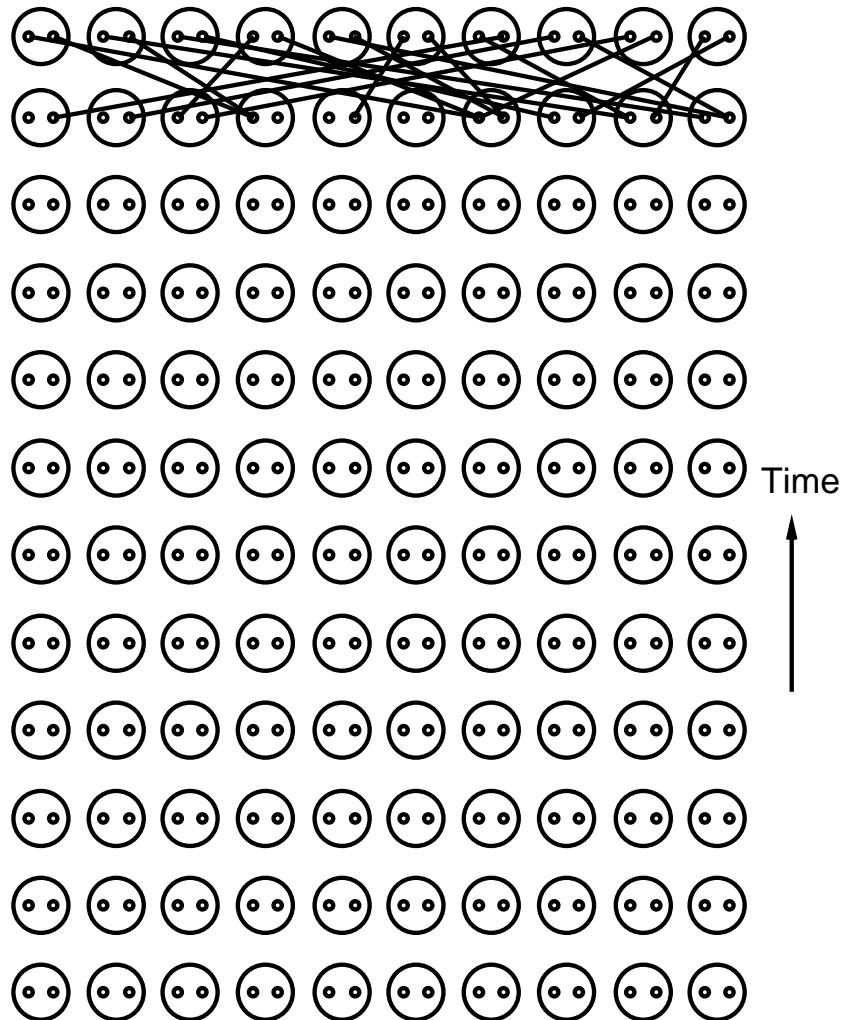
	Locus	n	Estimated N
Aut	ACTA	72	162,625
	BTN	72	76,369
	CP	76	77,319
	ESO	72	272,320
	FGG	72	180,730
	LACTAL	72	44,410
	WT1	80	51,972
X	G6PD	30	2,769
	PLP	52	92,655
mtDNA	Cytb	42	107,778
	All data		96,400 (78,500-117,700)
	Current census		18,000-29,000
	Previous models		19,480-35,430

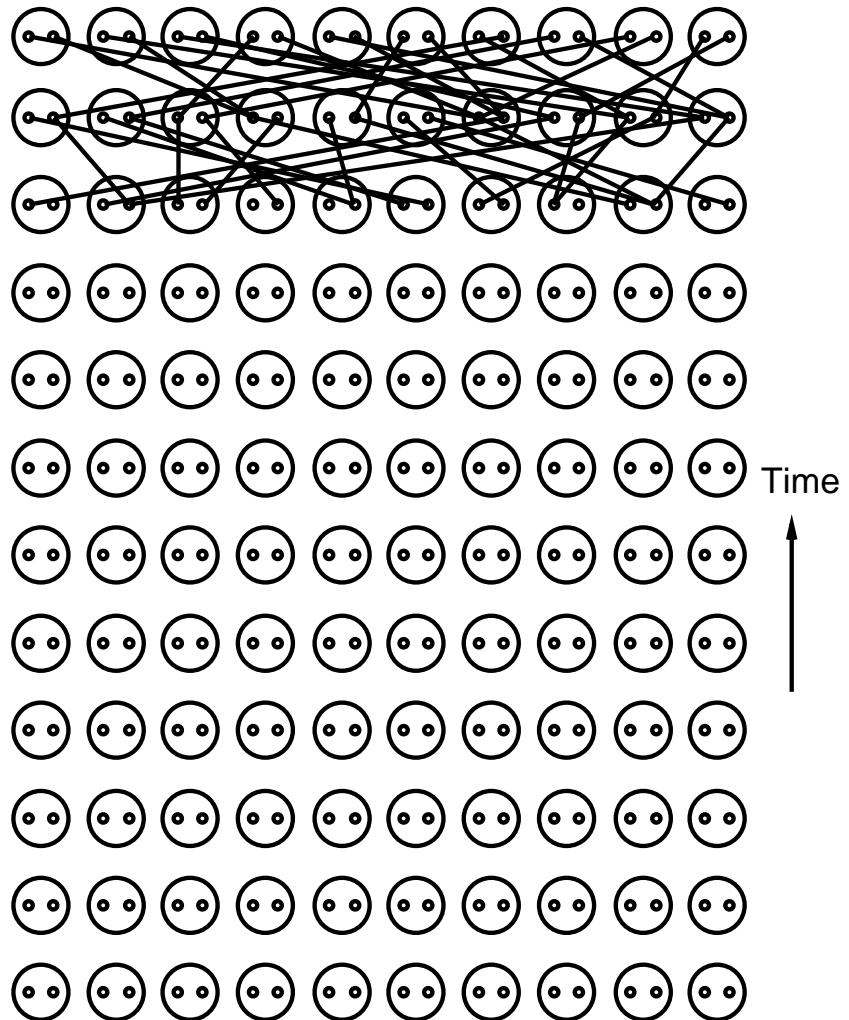
## **What was the long-term population size of gray whales?**

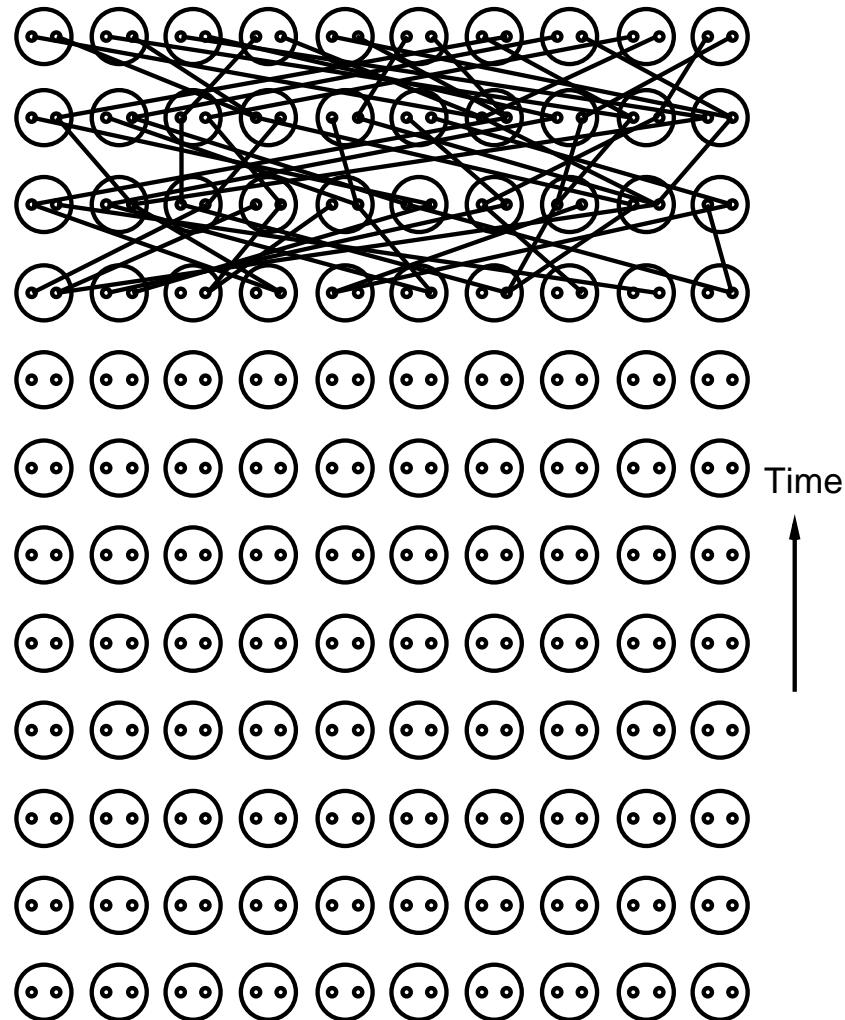
---

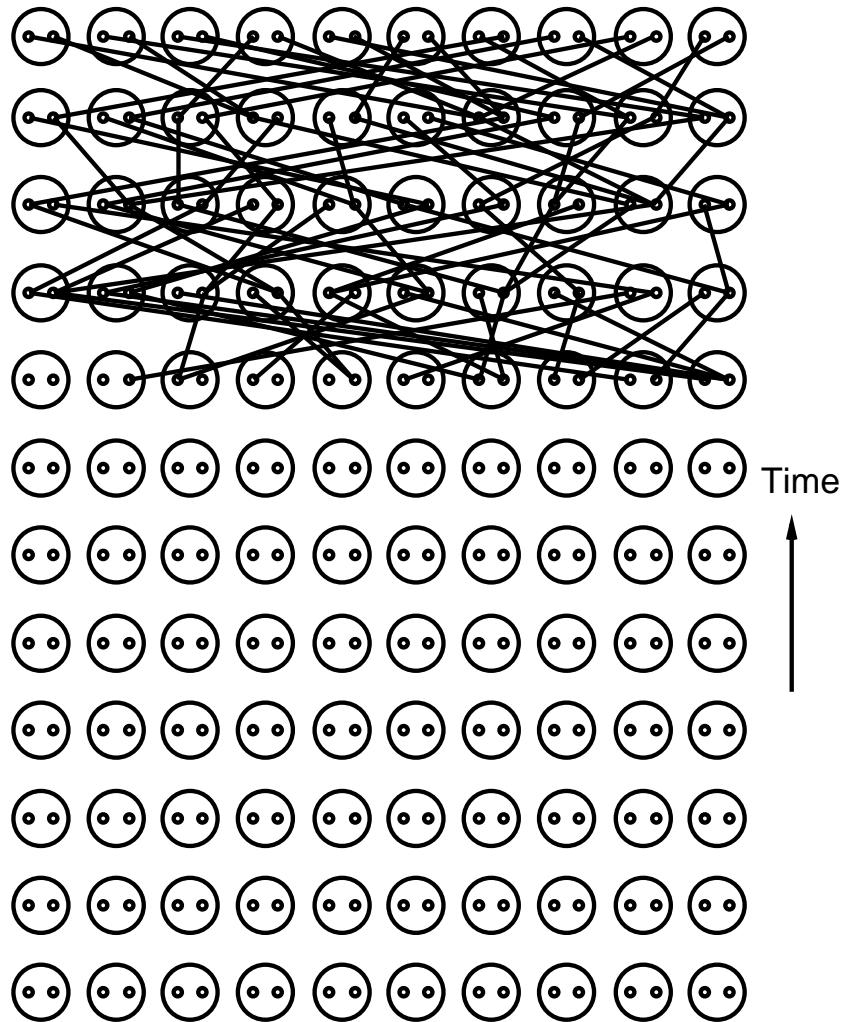
- Important conservation implications
- Effect on ecosystem significant:
  - Resuspension of up to 700 million cubic meters sediment
  - (12 Yukon Rivers worth)
  - Food for 1 million sea birds
- If accepted, result suggests halving gray whale kill rate
- Broadly similar results for minke, humpback, and fin whales

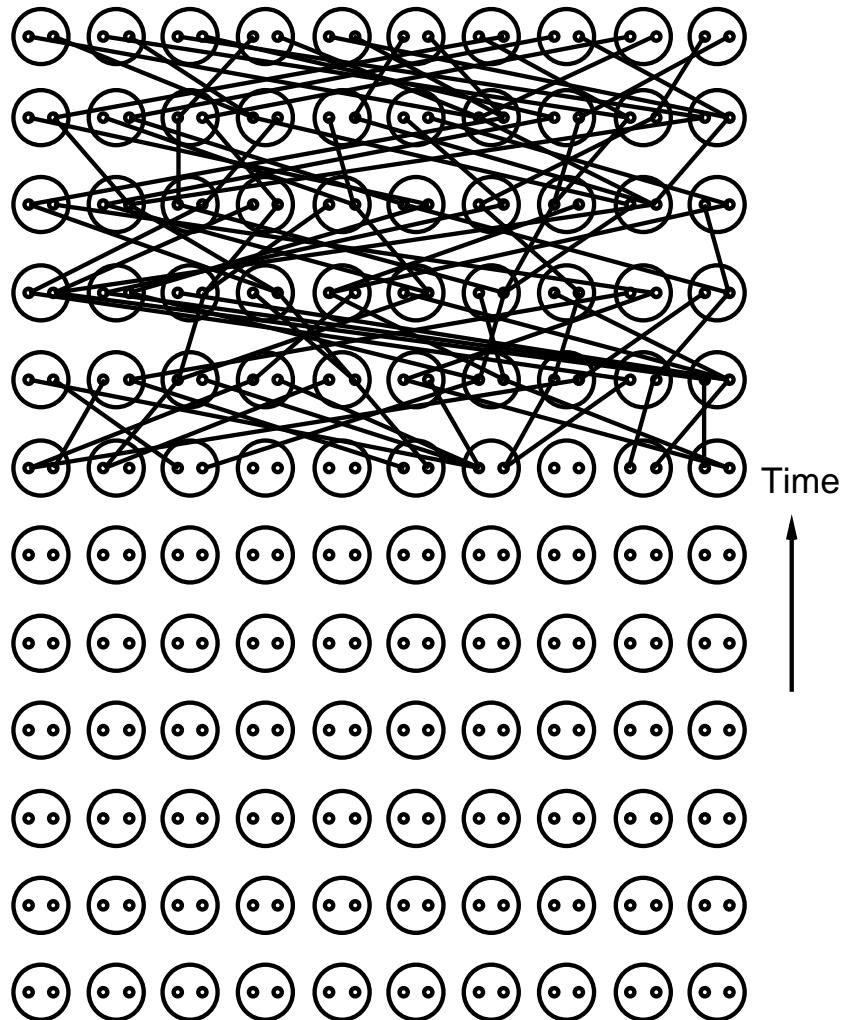


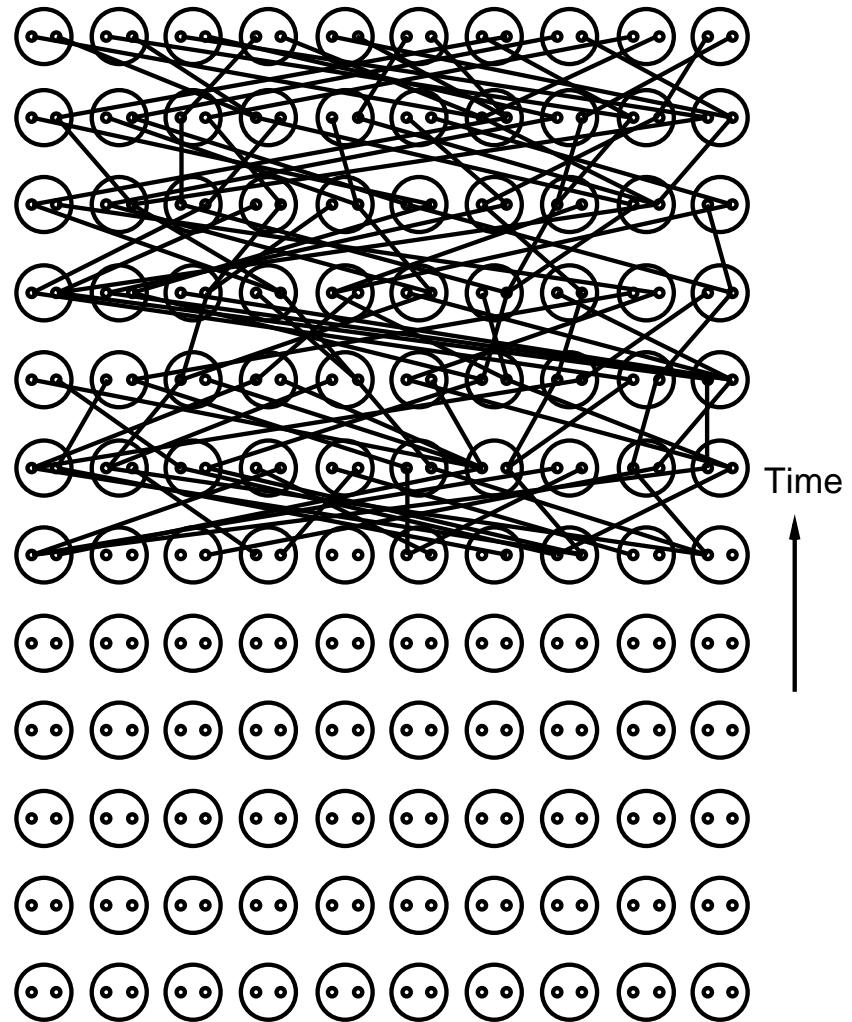


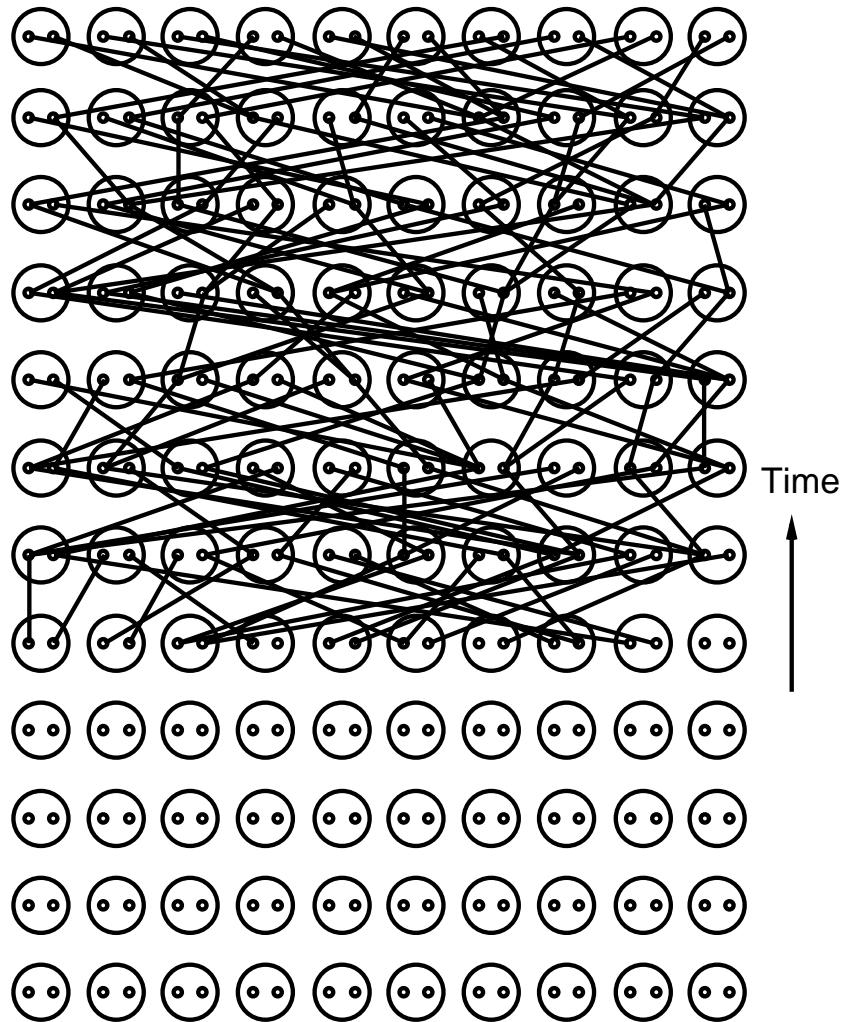


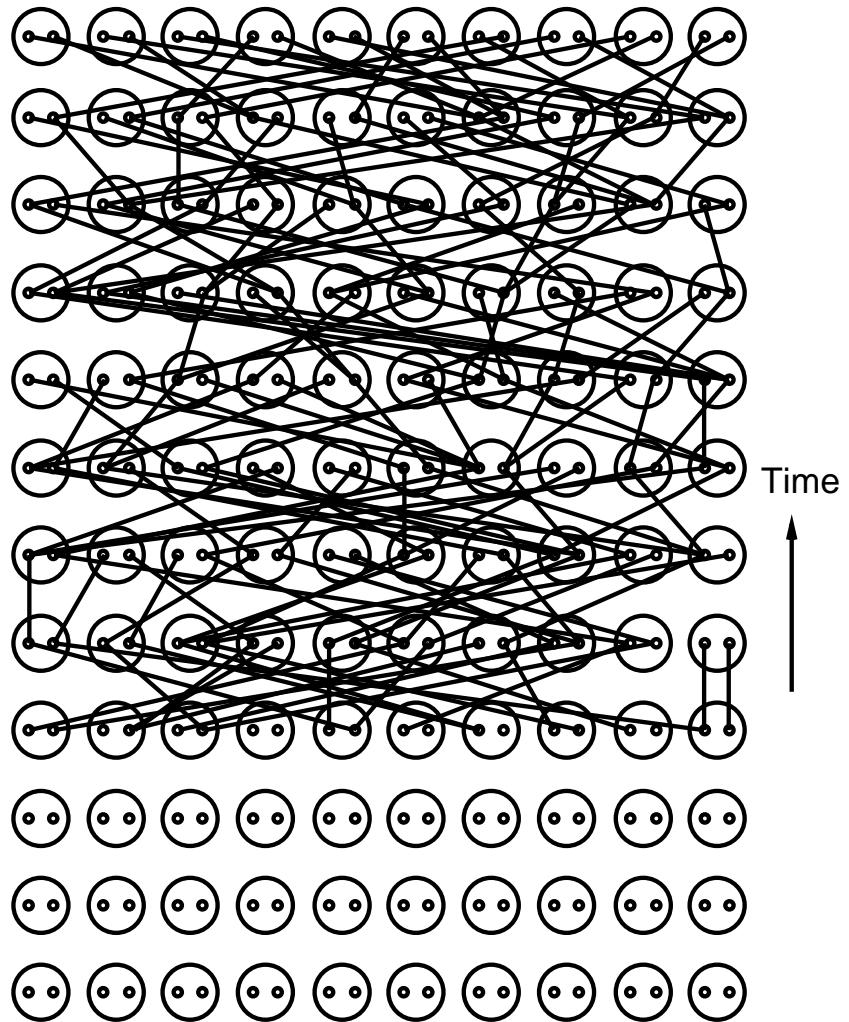


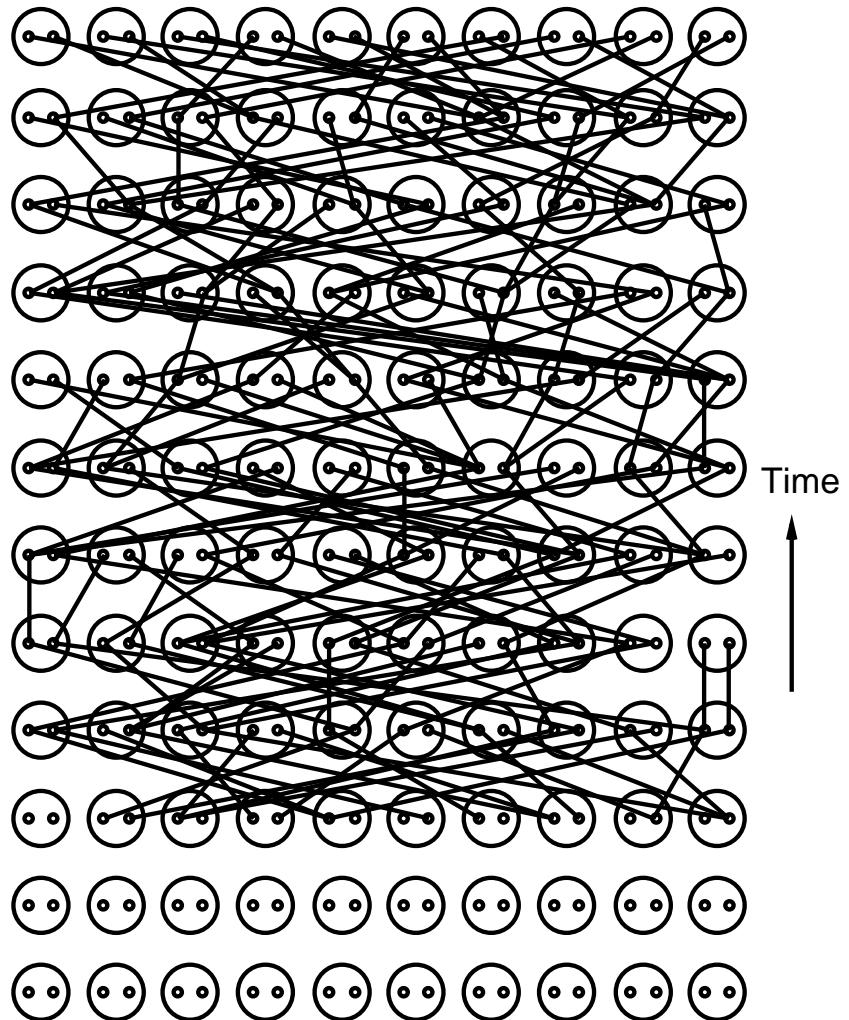


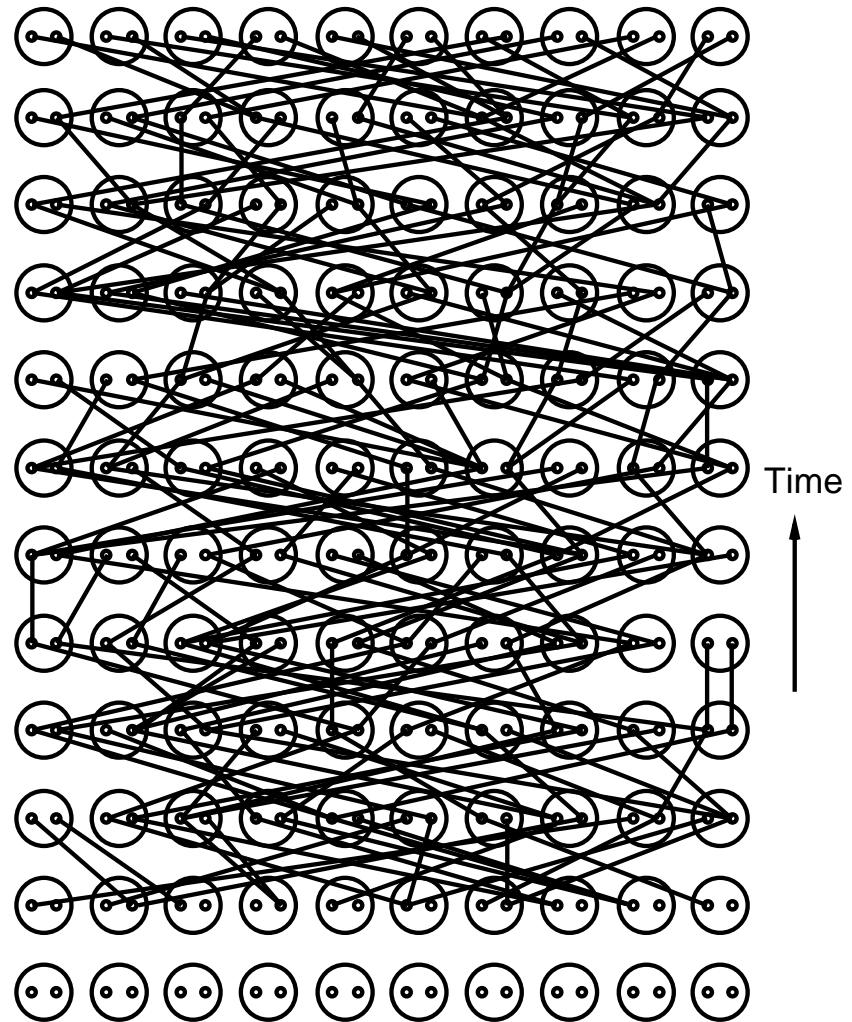


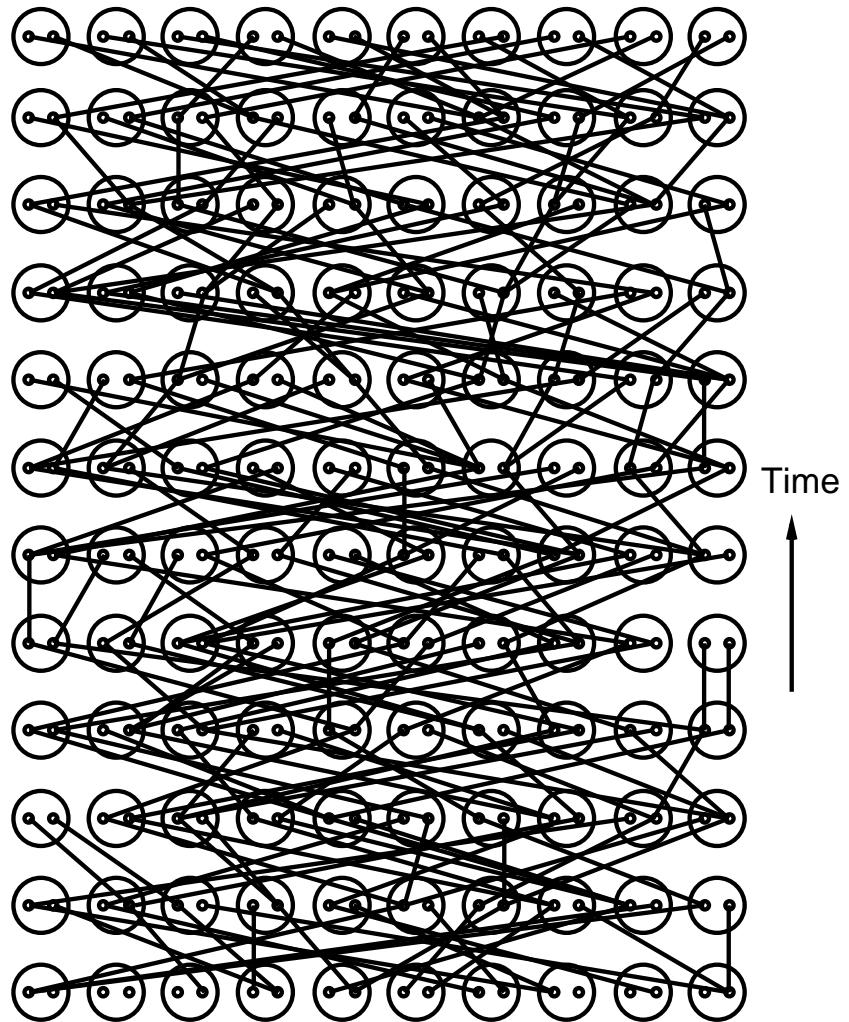


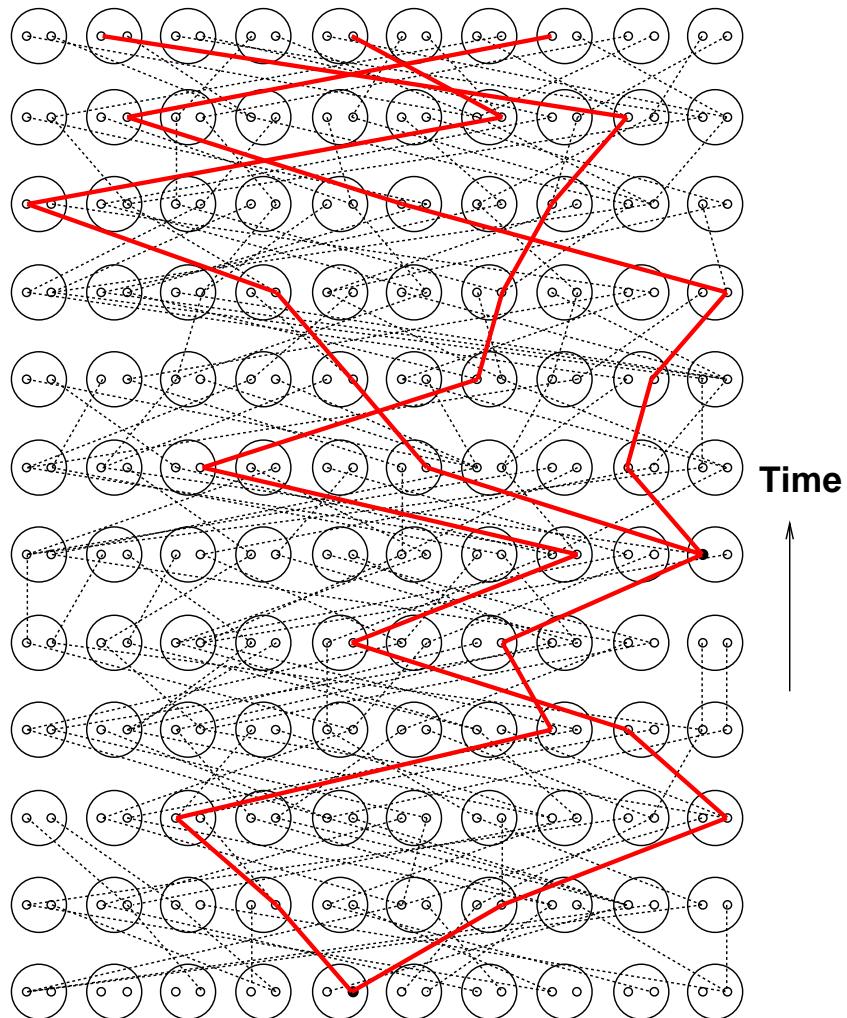


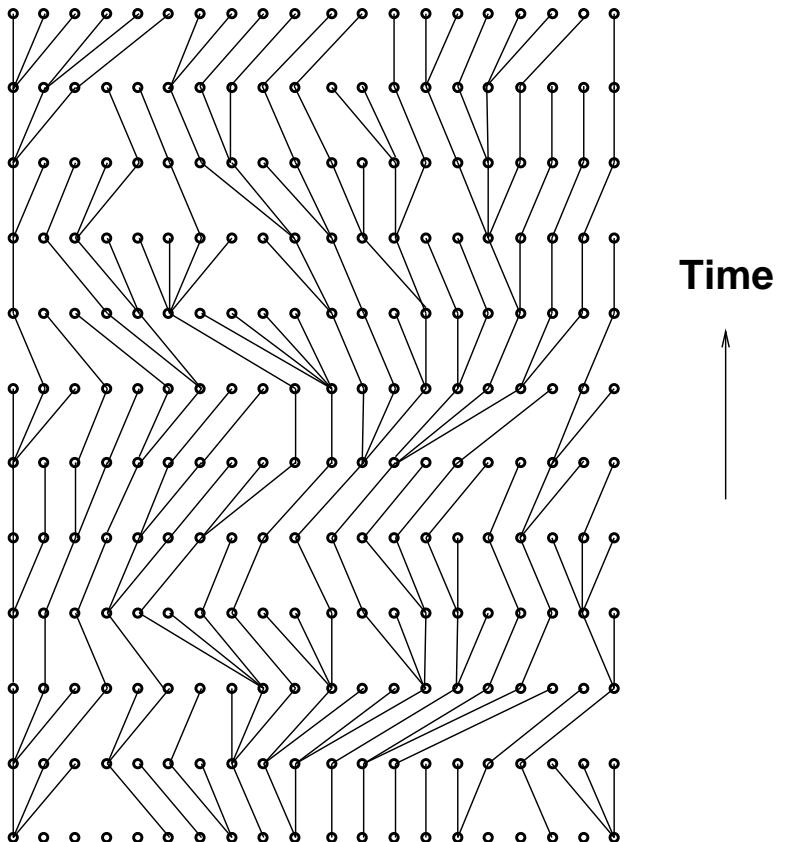


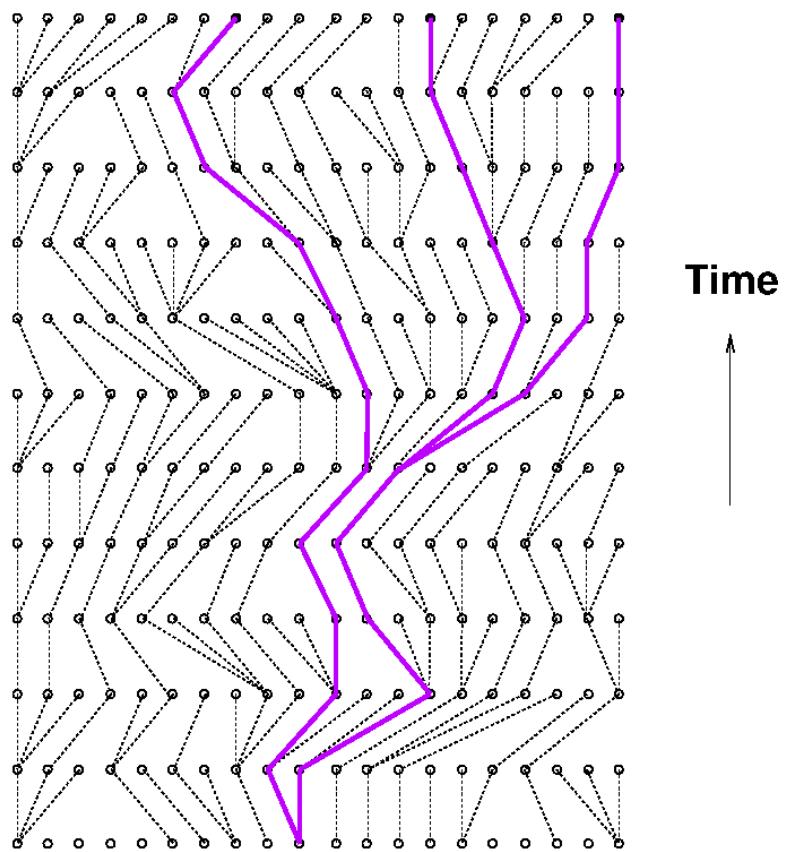




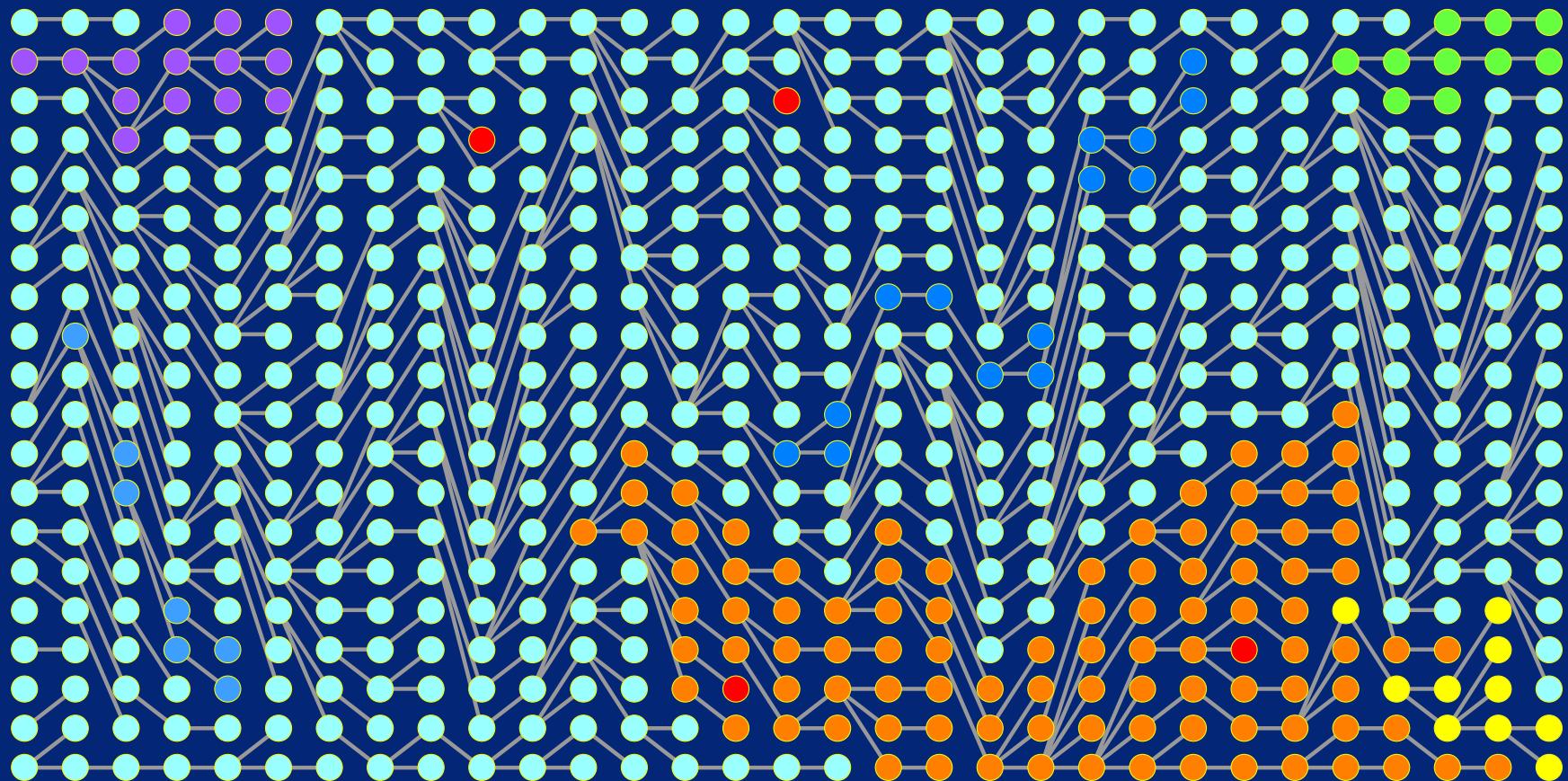








## Wright-Fisher population model



## **Wright-Fisher population model**

---

- Population size  $N$  is constant through time.
- Each individual gets replaced every generation.
- Next generation is drawn randomly from a large gamete pool.
- Only genetic drift affects the allele frequencies.

## Other population models

---

- Other population models can often be equated to Wright-Fisher
- The  $N$  parameter becomes the effective population size  $N_e$
- For example, cyclic populations have an  $N_e$  that is the harmonic mean of the various sizes
- Honeybees have an  $N_e$  that is the number of breeding individuals, not the vastly larger number of total bees

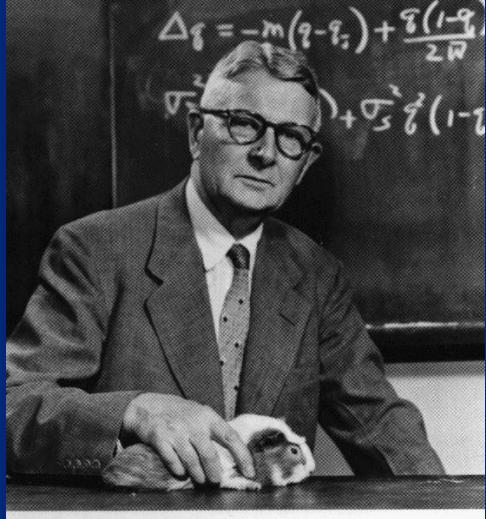
## From Wright-Fisher to the coalescent

---

- Key to coalescent theory: think BACKWARDS
- Each individual “chooses” its parent randomly from the previous generation
- When two lines come from the same parent, they “coalesce”
- As you look further back, eventually all lines must coalesce
- If there is no recombination, this is a within-population tree (genealogy)

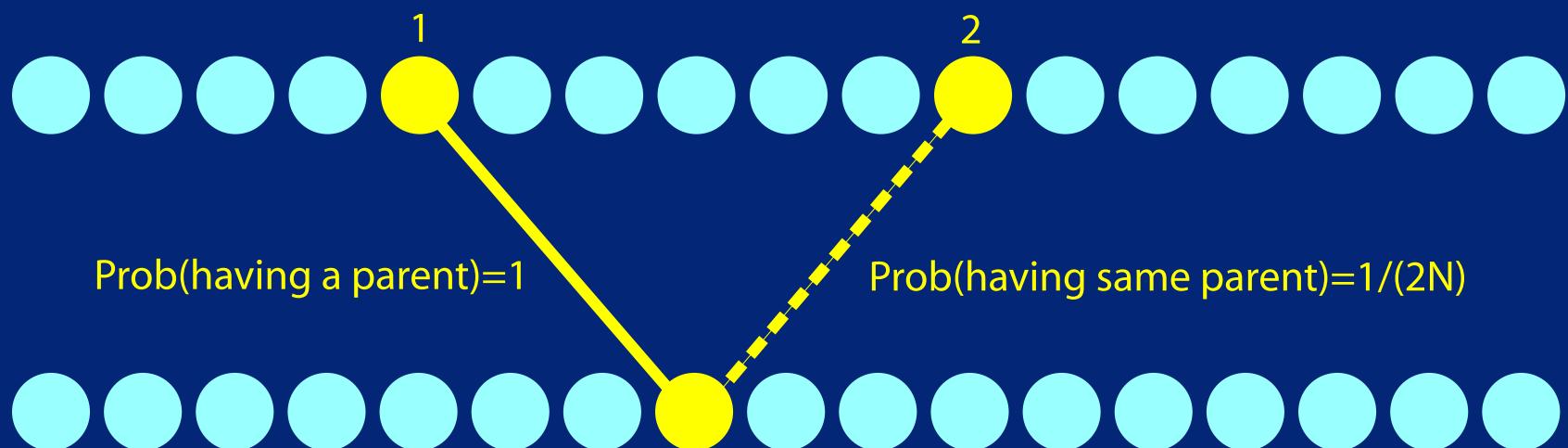
# The Coalescent

---



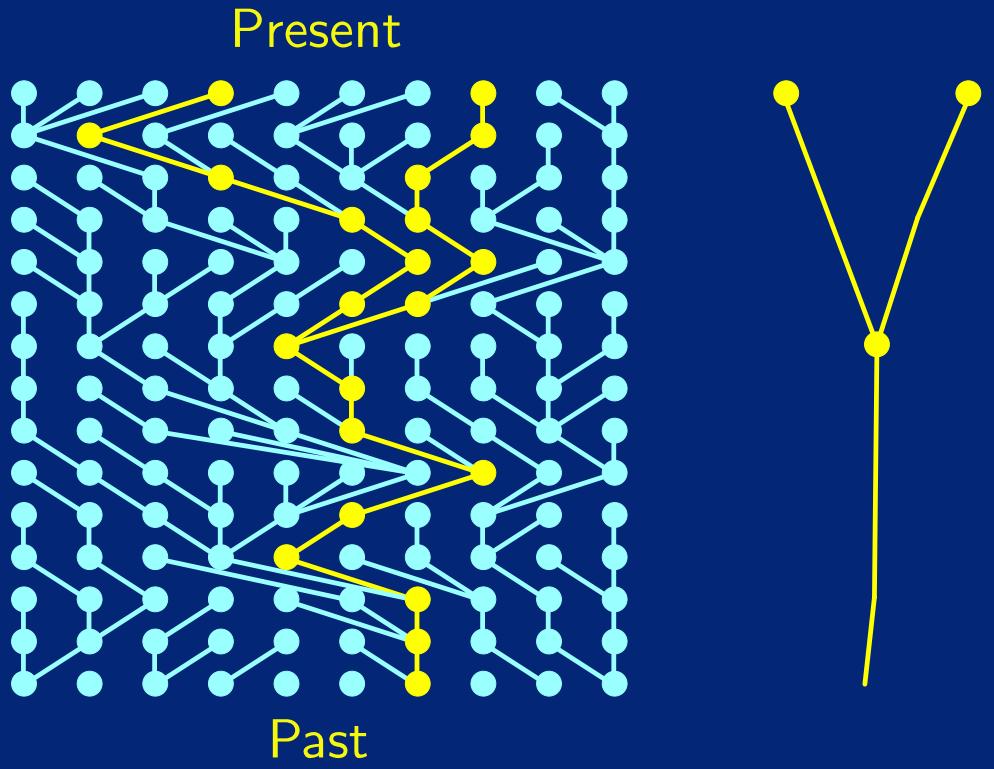
Sewall Wright showed that the probability that 2 gene copies come from the same gene copy in the preceding generation is

$$\text{Prob (two genes share a parent)} = \frac{1}{2N}$$



# The Coalescent

---



In every generation, there is a chance of  $1/2N$  to coalesce. Following the sampled lineages through generations backwards in time we realize that it follows a geometric distribution with

$$\mathbb{E}(u) = 2N \quad [\text{the expectation of the time of coalescence } u \text{ of two tips is } 2N]$$

# The Coalescent

---

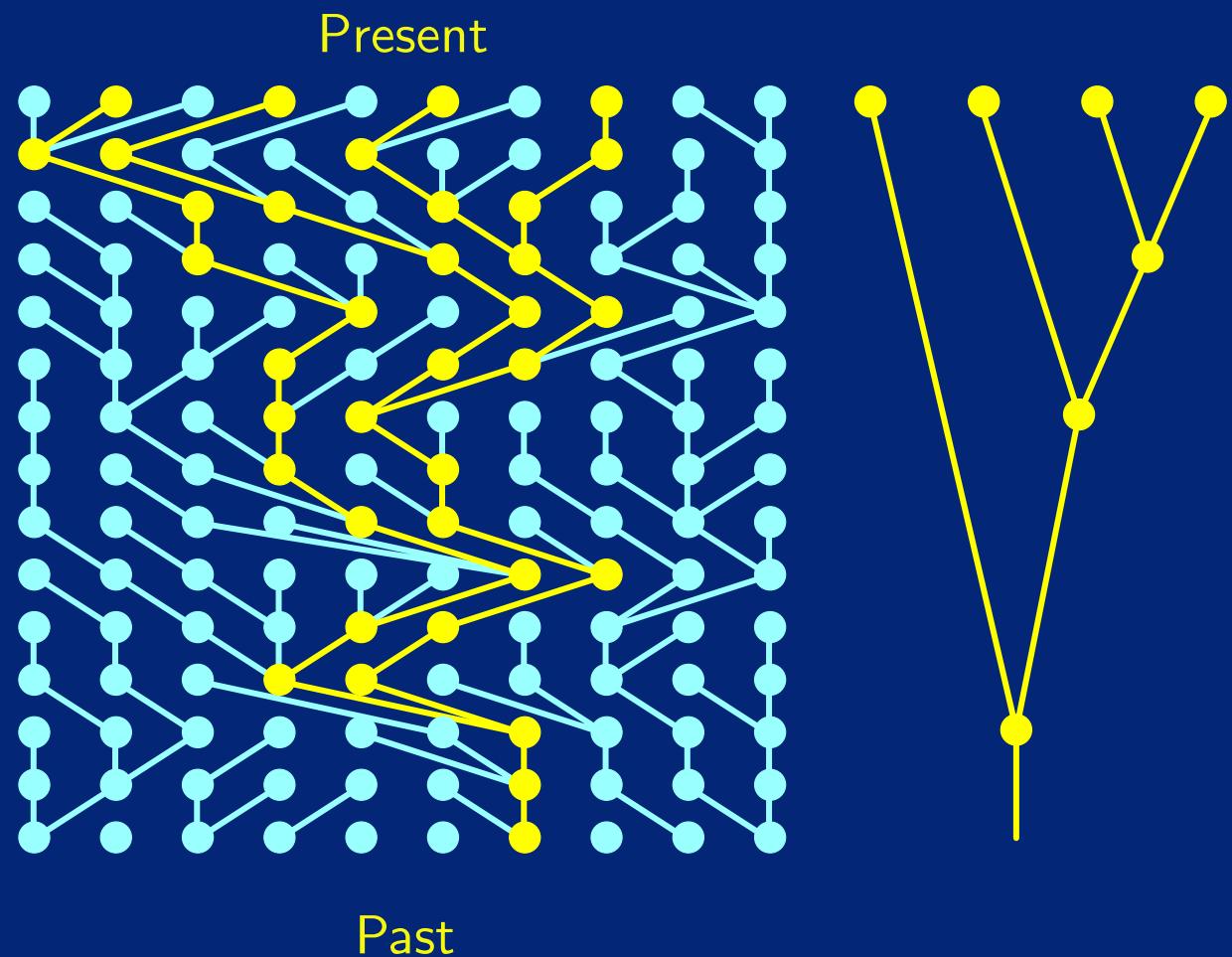


JFC Kingman generalized this for  $k$  gene copies.

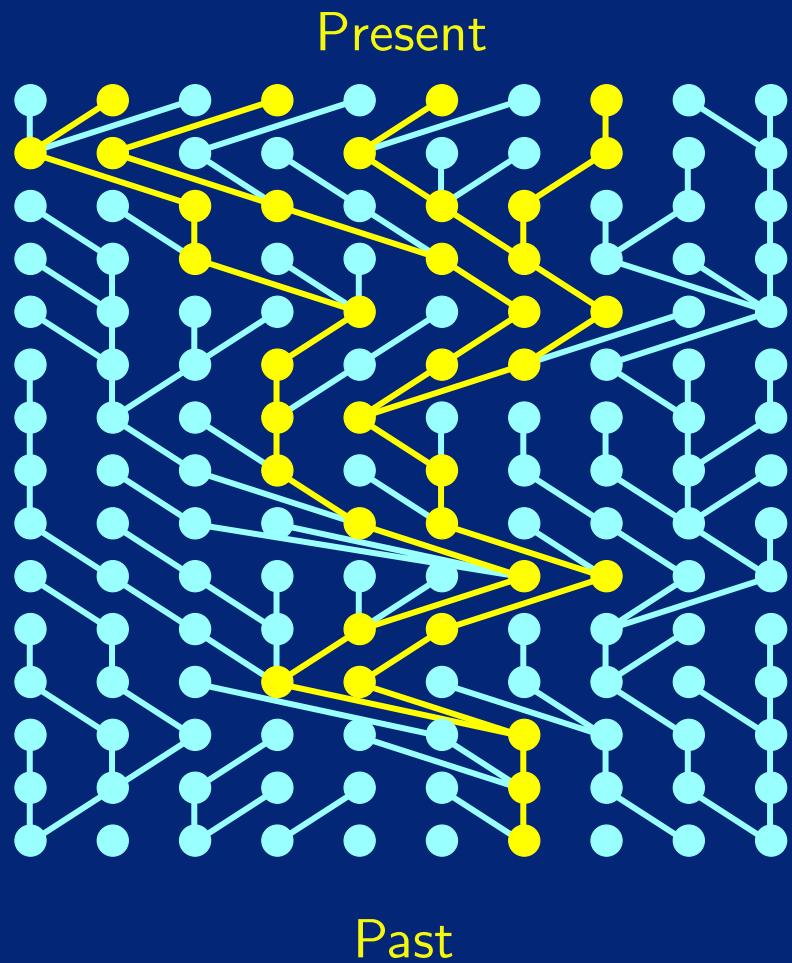
$$\text{Prob } (k \text{ copies are reduced to } k-1 \text{ copies}) = \frac{k(k-1)}{4N}$$

## Kingman's $n$ -coalescent

---

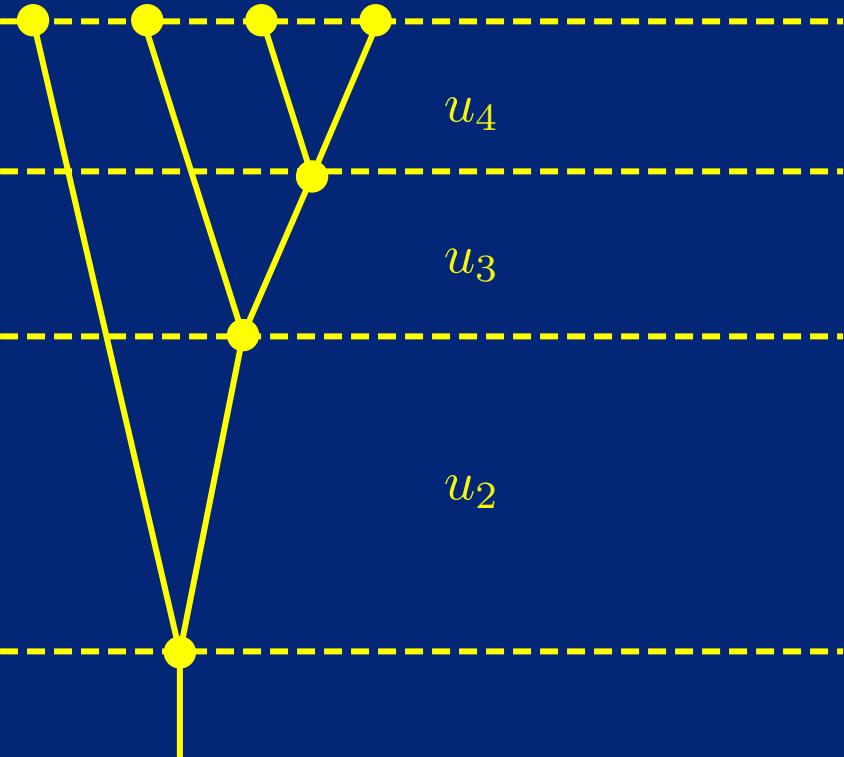


## Kingman's $n$ -coalescent



The expectation for the time interval  $u_k$  is

$$\mathbb{E}(u_k) = \frac{4N}{k(k-1)}$$

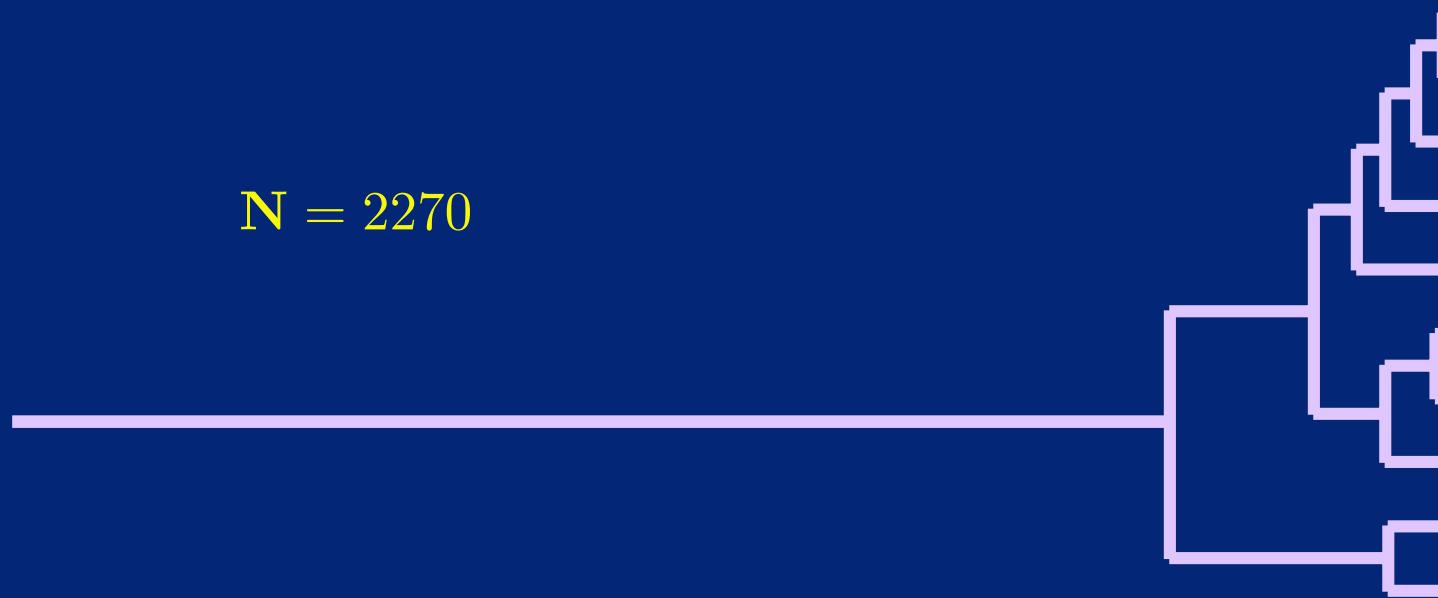


$$p(G|N) = \prod_i \exp(-u_i \frac{k(k-1)}{4N}) \frac{1}{2^N}$$

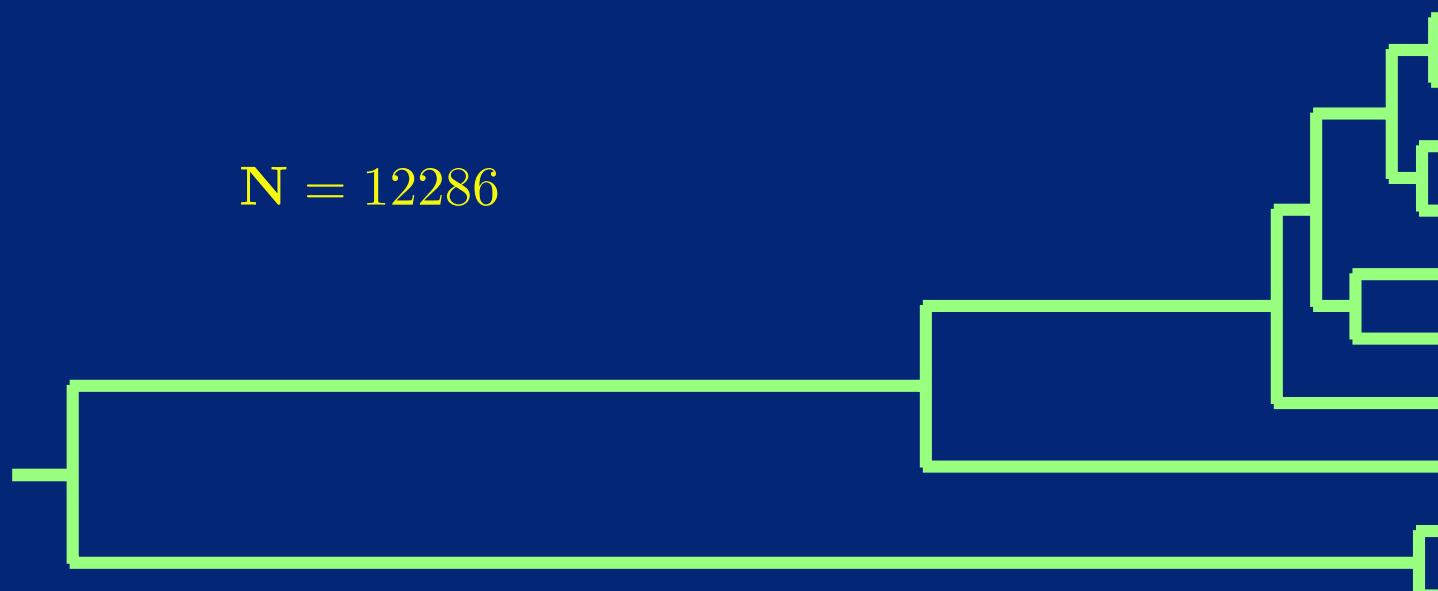
## Coalescence time depends on population size

---

$N = 2270$



$N = 12286$



## **This would be great if....**

---

- If we knew the tree, including its times, we would have a powerful estimator of population size
- Unfortunately this is difficult to infer
- Within-population variation is usually too low for really accurate phylogeny estimation
- We also have a problem with observing time directly

## The variable $\Theta$

---

- Goal: Estimate  $2N_e$ , the effective number of gene copies
- Problem: to estimate this, we need to know coalescence times
- We do not observe times (except in viruses and fossil DNA), only numbers of differences (mutations)

## The variable $\Theta$

---

- The number of differences is proportional to the product of mutation rate  $\mu$  and time
- We can only estimate the compound parameter  $4N_e\mu$  also called  $\Theta$
- One factor of 2 comes from each individual having two gene copies (so the number of gene copies is  $2N$ )
- The other comes from mutations accumulating on both branches of the tree, so in 1 unit of time we accumulate 2 units of mutations

## The variable $\Theta$

---

- It is disappointing not to get  $N_e$  directly
- If we can measure  $\mu$  experimentally we can convert  $\Theta$  to  $N_e$
- Even if we can't,  $\Theta$  is interesting:
  - Comparing populations with similar mutation rate
  - Expected “carrying capacity” of genetic diversity

## The coalescent without recombination: “Mitochondrial Eve”

---

- Cann, Stoneking and Wilson analyzed 149 human mtDNA sequences
- mtDNA (mitochondrial DNA) is:
  - Inherited only from the mother
  - Essentially haploid: the child receives only 1 genotype from the mother
  - Higher mutation rate than the nuclear genome
- Question: what is the population size of an mtDNA sequence relative to a nuclear sequence?

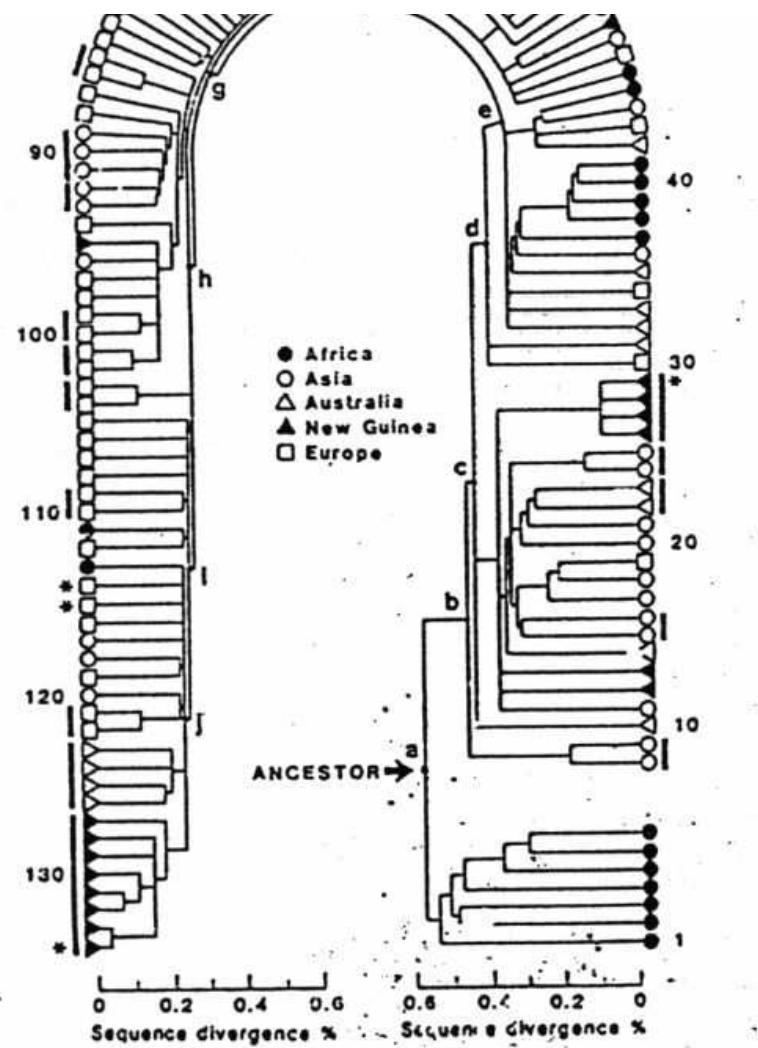


Fig. 3 a, Genealogical tree for 134 types of human mtDNA (133 restriction sites used. The tree accounts for the site differences ob

## The coalescent without recombination: “Mitochondrial Eve”

---

Three observations from Cann et al.

- All of their mtDNAs traced to a single ancestor
- That ancestor was dated to approx. 200,000 years ago
- She apparently lived in Africa

Let's take them in turn.

## All human mtDNA has a common ancestor

---

- The popular press found this the most exciting result
- It's actually a logical necessity
  - Start with some finite number of matrilines
  - Every generation some fail to reproduce
  - Given enough generations, only one will remain

## “Eve” dates to roughly 200,000 years ago

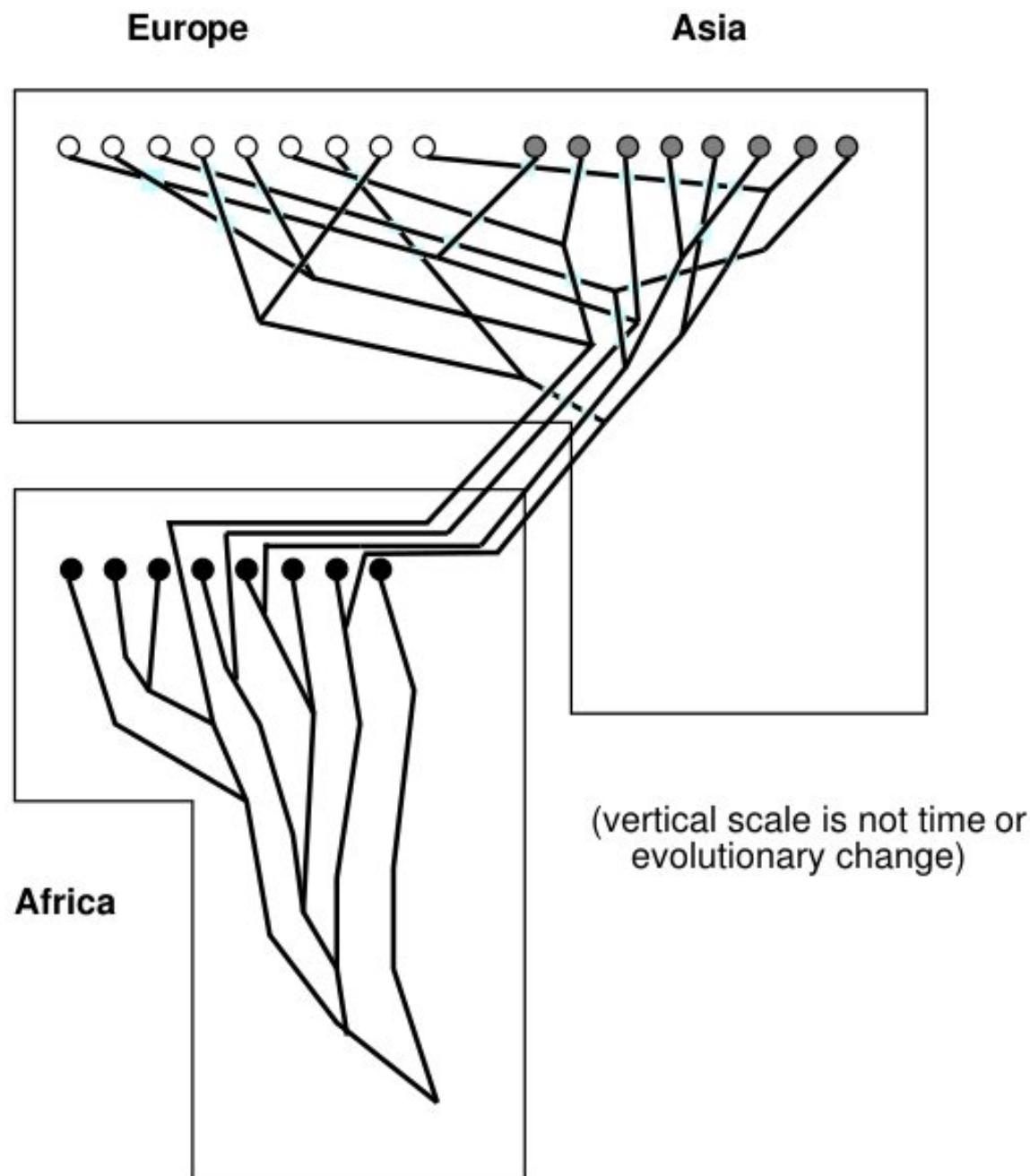
---

- We can't do a straight population size calculation because the population size is not constant
- But if it were:
  - Expected time to Eve is  $2N_f$  generations
  - (Why not  $4N_e$ ?)
  - 25 years/generation suggests  $N_f$  around 4000
- That doesn't sound like the widespread Eurasian *Homo erectus* population

## **The oldest splits lead to African lineages**

---

- Two hypotheses for modern human origins:
  - Widespread Eurasian and African populations evolved into modern humans (“multiregional hypothesis”)
  - Modern humans evolved in Africa and displaced previous Eurasian populations (“out-of-Africa hypothesis”)



## **Testing these hypotheses**

---

- When Eve lived there were tool-using hominoids all over Eurasia
- Their mtDNA does not seem to have survived
- Descendants of Eve have gotten all over the globe, at the expense of matrilines elsewhere
- This supports (doesn't prove) out-of-Africa hypothesis
- It at least proves that no population has been completely isolated for more than 200,000 years

## **Three reality checks about Eve**

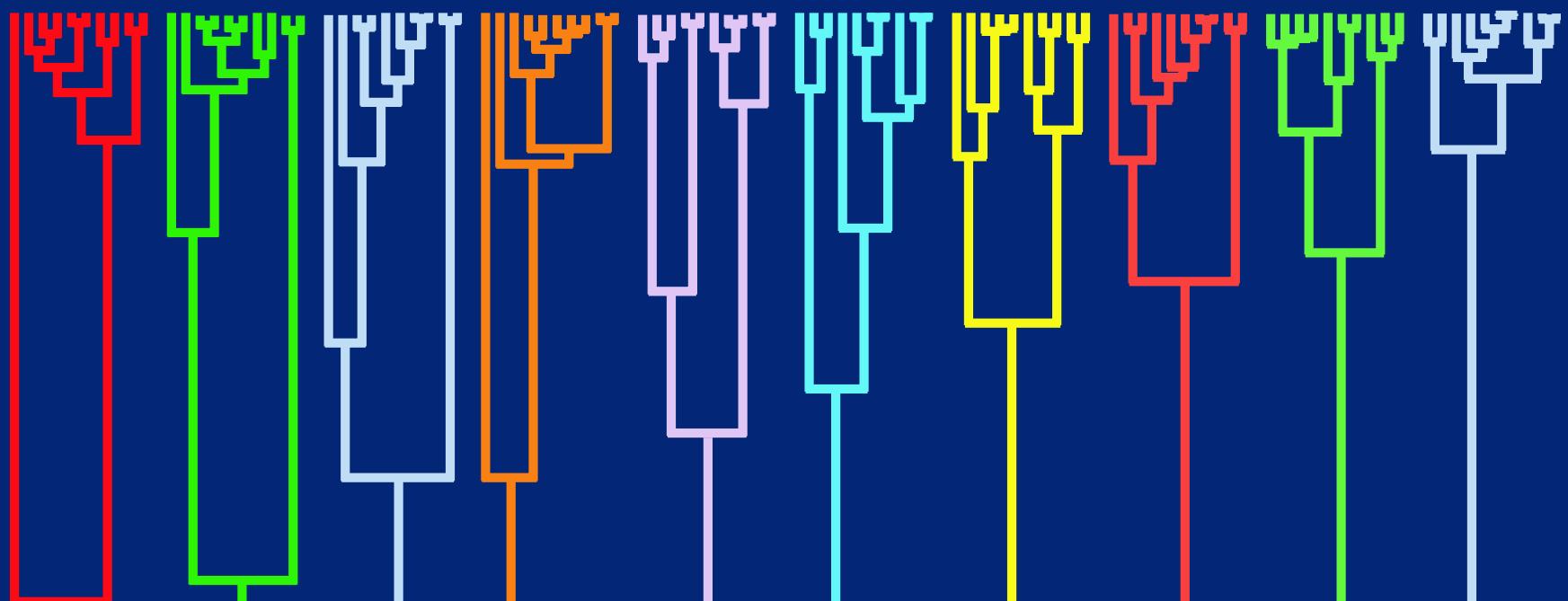
---

- Was Eve the only female in her generation?
- Is our nuclear genome likely to be descended from Eve?
- Could part of our nuclear genome come from those widespread Eurasian populations even if mtDNA did not?

## Variability of the coalescent

---

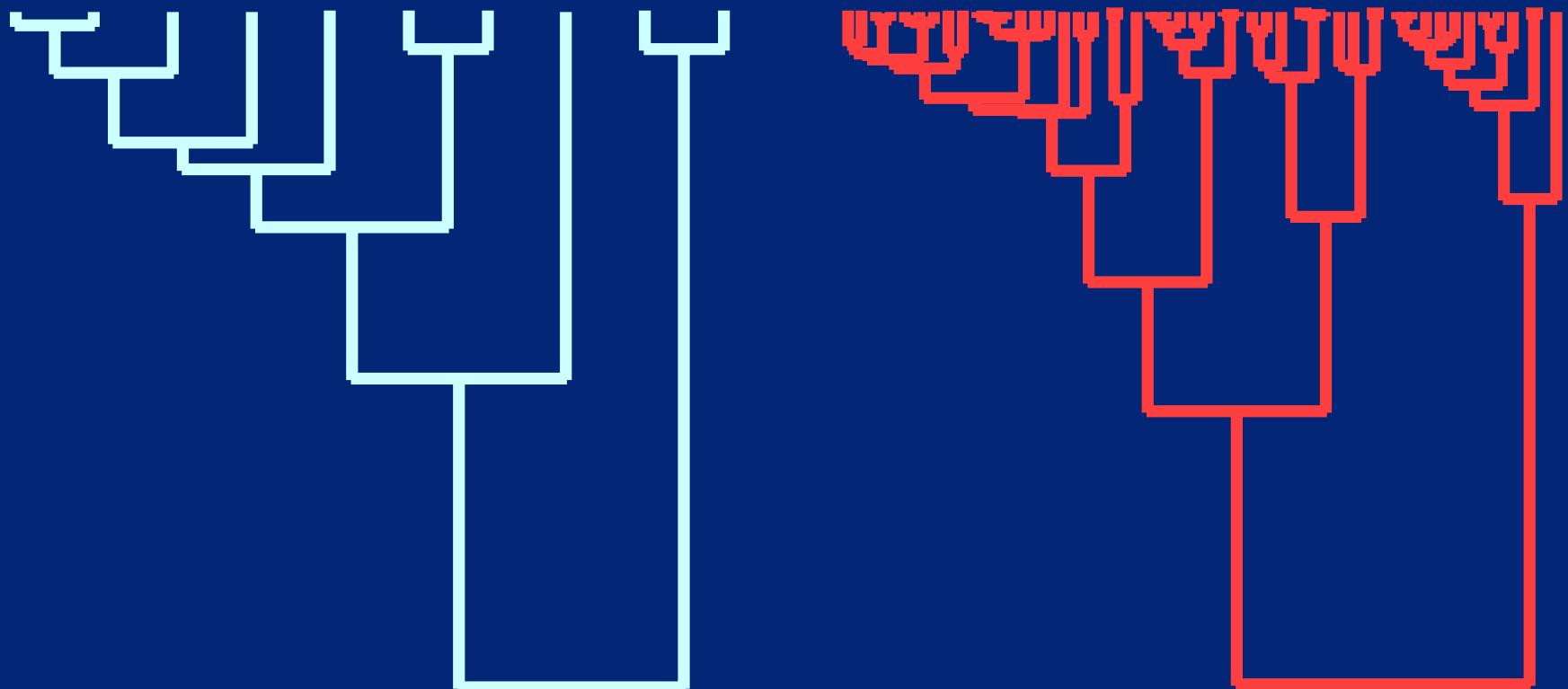
A single gene can give a misleading answer, because the coalescent is a highly variable distribution.



10 coalescent trees generated with the same population size,  $N = 10,000$

**Does sampling more individuals help? (No)**

---



## The origins of the rest of the genome

---

- Neanderthal and Denisovan genomes sequenced
- Several labs have searched human genomes for sequences which:
  - Have a rare haplotype very different from the common ones
  - That rare haplotype is also in Neanderthals or Denisovans
- They found a lot of hits:
  - Europeans and Asians appear to have some Neanderthal haplotypes
  - Asians appear to have some Denisovan haplotypes
  - Africans show few hits for either one
- Estimates of 4-5% admixture

## Why the different results?

---

- No Neanderthal or Denisovan mtDNA – initial conclusion was no admixture
- Reasonable amount of apparent Neanderthal and Denisovan nuclear DNA
- Could be:
  - Chance? (With 5% admixture, no surprise mtDNA doesn't show it)
  - Lower effective population size for mtDNA? (Lineages are lost faster)
  - Different fertility for male and female hybrids? (Common in other animals)
  - Negative selection for foreign mtDNA?
  - Positive selection for foreign nuclear DNA?

## **Approaches to using the coalescent**

---

- Summary-statistics approaches
- Many-tree approaches

## Summary-statistic approaches

---

- Summary statistics look at the bulk properties of coalescent trees.
- They often require a simplified model of mutation.
- Watterson developed an estimator of  $\Theta$  based on counting the number of variable sites
- We know how many variable sites to expect for various values of  $\Theta$ , sequence length and number of sequences
- This approach discards much of the information in the data

## Many-tree approaches

---

- Many groups, including mine, develop computer algorithms to estimate  $\Theta$  by considering many possible trees
- There are too many trees to consider all possibilities
- We write sampling algorithms which visit mainly the most likely trees
- While developed independently, this is similar to the Bayesian phylogenetic algorithm

# What do I mean by too many trees?

# Tips Topologies

3 3

4 18

5 180

6 2700

7 56700

8 1587600

9 57153600

10 2571912000

15 6958057668962400000

20 564480989588730591336960000000

30 43684666131030695124646801986207638914406400000000000000

$$50 \quad 3.28632 \times 10^{112}$$

100  $1.37416 \times 10^{284}$

## **Uses and extension of the coalescent**

---

- Genetic drift/population size estimation
- Population growth/shrinkage over time
- Migration between populations
- Recombination
- Divergence of populations
- Selection

## **Genetic drift (*Theta*)**

---

- With one time point, we estimate  $\Theta = 4N_e\mu$  in diploids
- The number estimated is  $2N_e\mu$  in haploids or  $N_e\mu$  in mtDNA
- Two ways to separate  $N_e$  and  $\mu$ :
  - Dated historical data (ancient DNA, etc.)
  - External estimate of mutation rate
- For most organisms,  $N_e$  is less than  $N$
- Demographic models can help resolve this

## Variable population size

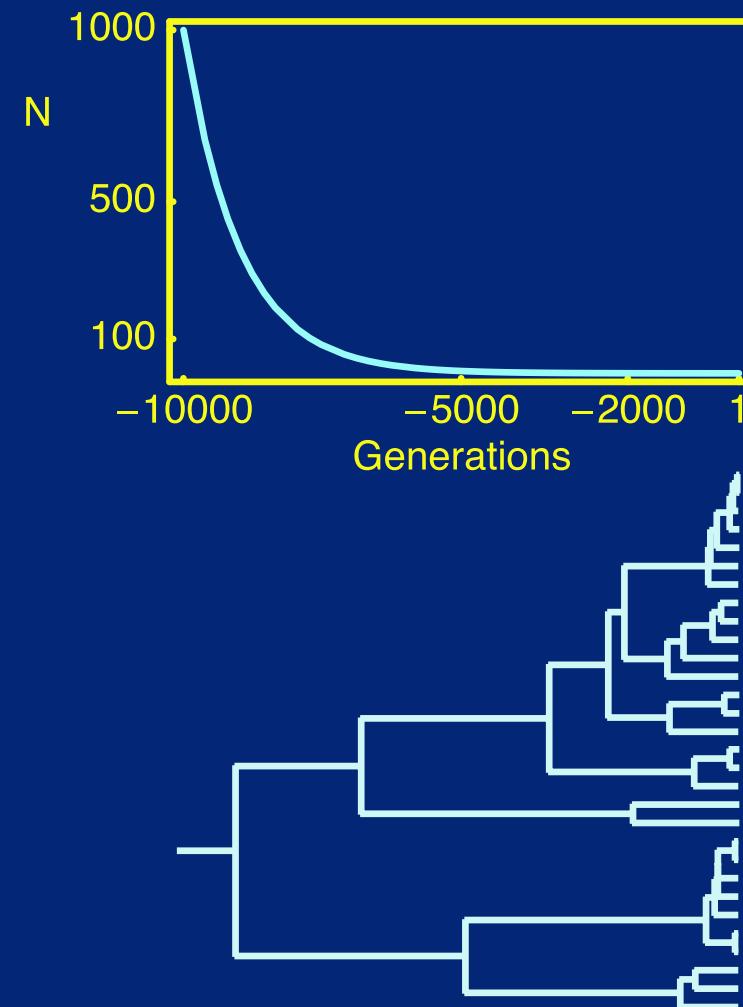
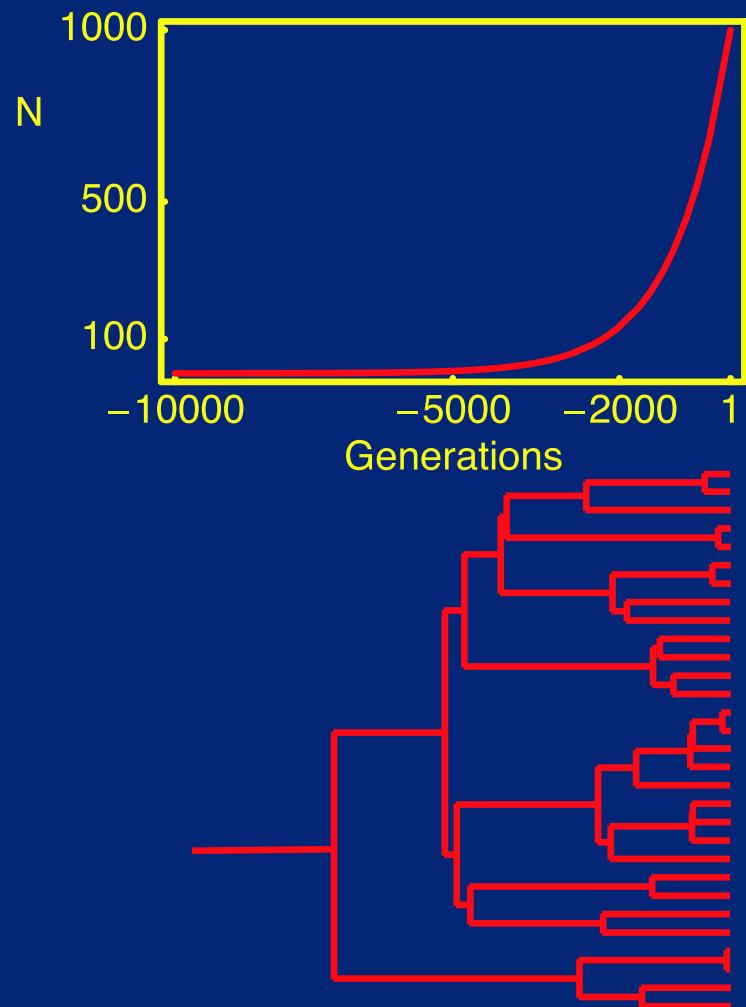
---

- In a small population lineages coalesce quickly
- In a large population lineages coalesce slowly

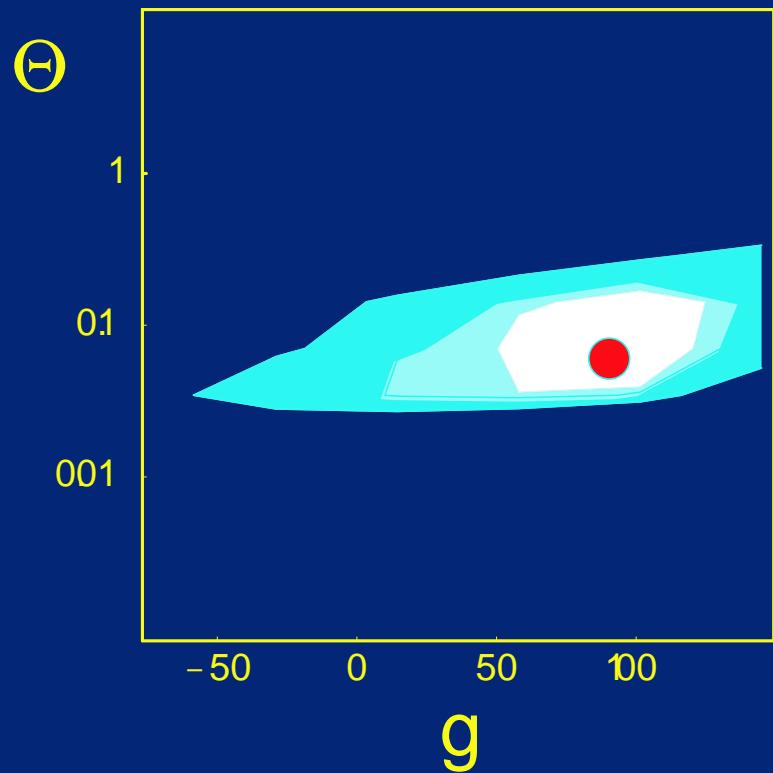
This leaves a signature in the data. We can exploit this and estimate the population growth rate  $g$  jointly with the current population size  $\Theta$ .

## Exponential population size expansion or shrinkage

---



## Grow a frog



Mutation Rate

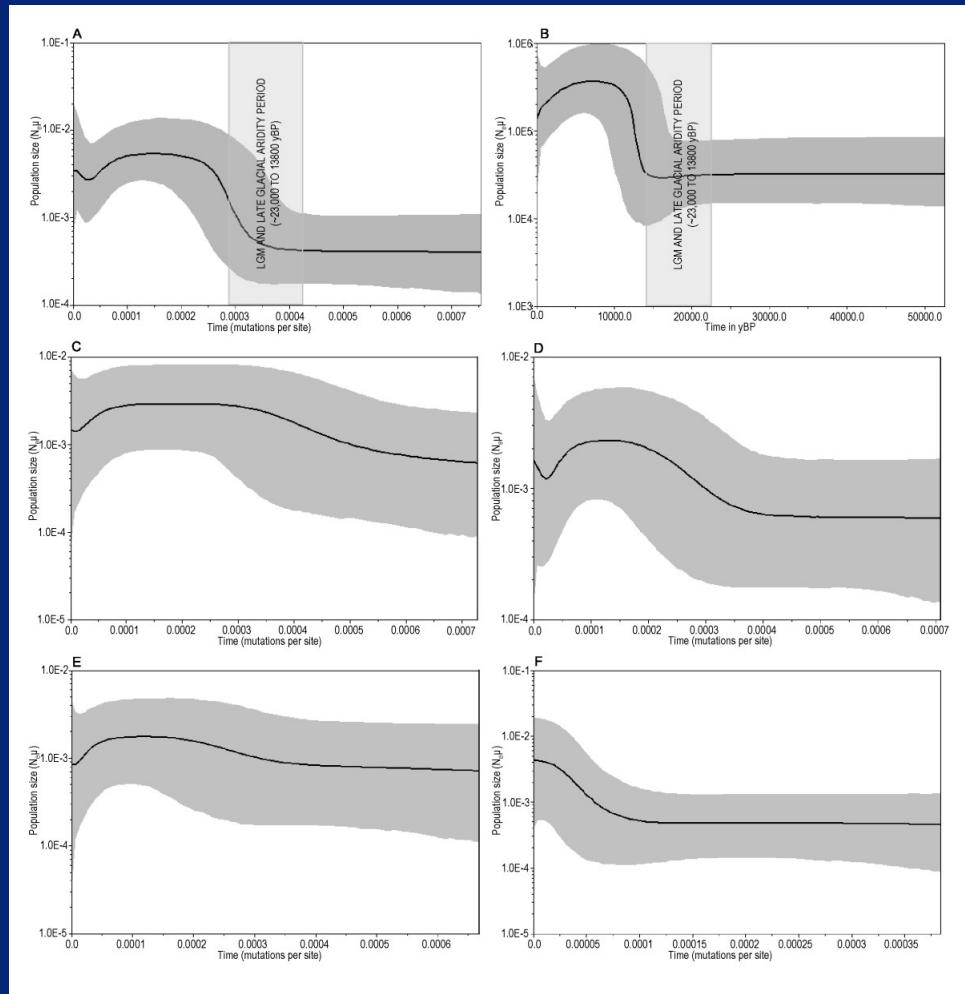
Population sizes

-10000 generations

Present

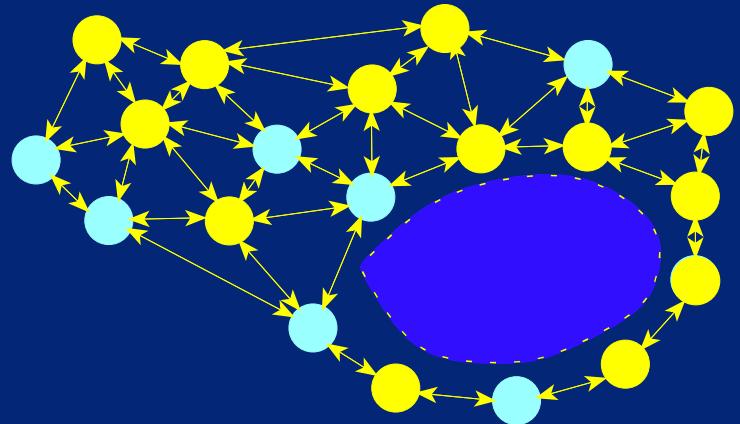
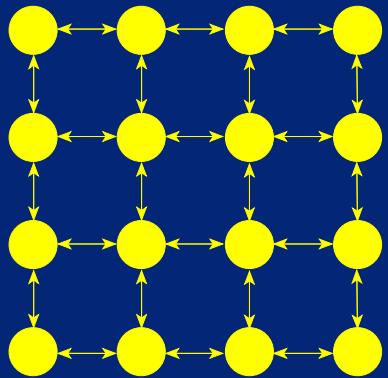
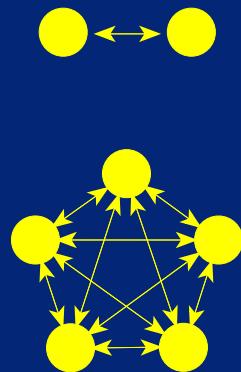
$10^{-8}$	8,300,000	8,360,000
$10^{-7}$	780,000	836,000
$10^{-6}$	40,500	83,600

# Bayesian skyline plots



## Gene flow

---

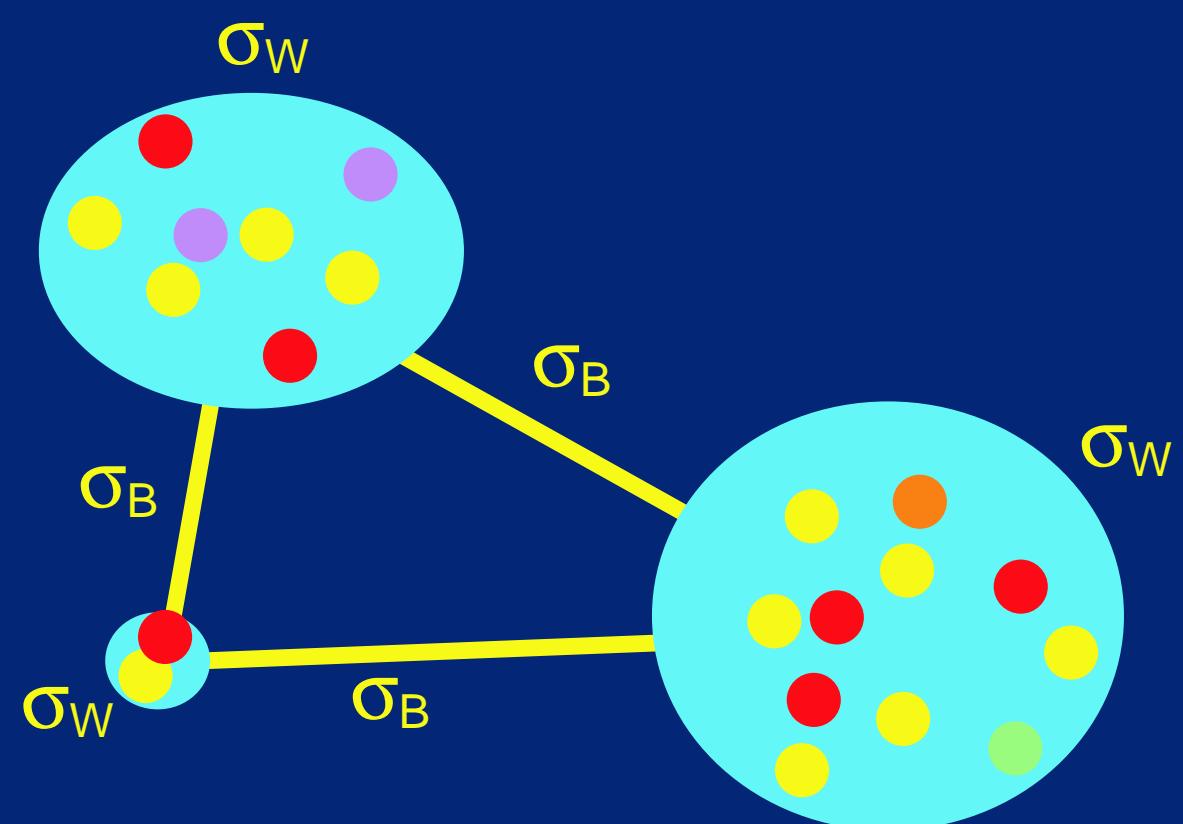


$$p(G|\Theta, M) = \prod_{u_j} \left( \prod_i^{pop.} g(\Theta_i, M_{.i}) \right) \begin{cases} \frac{2}{\Theta} \\ M_{ji} \end{cases}$$

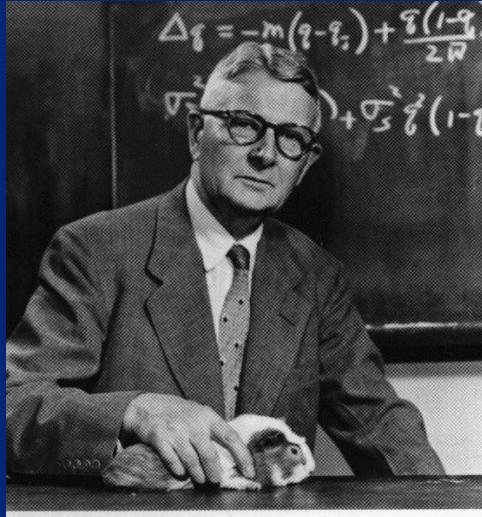
if event is a coalescence,  
if event is a migration from  $j$  to  $i$ .

## Gene flow: What researchers used (and still use)

$F_{ST}$



## What researchers used (and still use)



Sewall Wright showed that

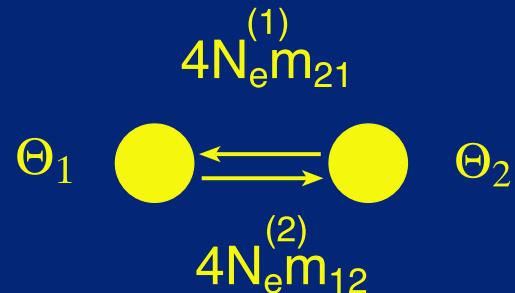
$$F_{ST} = \frac{1}{1 + 4Nm}$$

and that it assumes

- migration into all subpopulation is the same
- population size of each island is the same

## Simulated data and Wright's formula

---



True values	Estimated values
$0.01 \quad \text{↔} \quad 0.01$ 1. 1.	$1.14 \pm 0.77$
$0.01 \quad \xleftarrow{\text{10.}} \quad 0.01$ 1. 1.	$7.80 \pm 22.20$
$0.05 \quad \xleftarrow{\text{10.}} \quad 0.005$ 1. 1.	$11.46 \pm 18.54$

## Maximum Likelihood method to estimate gene flow parameters

---

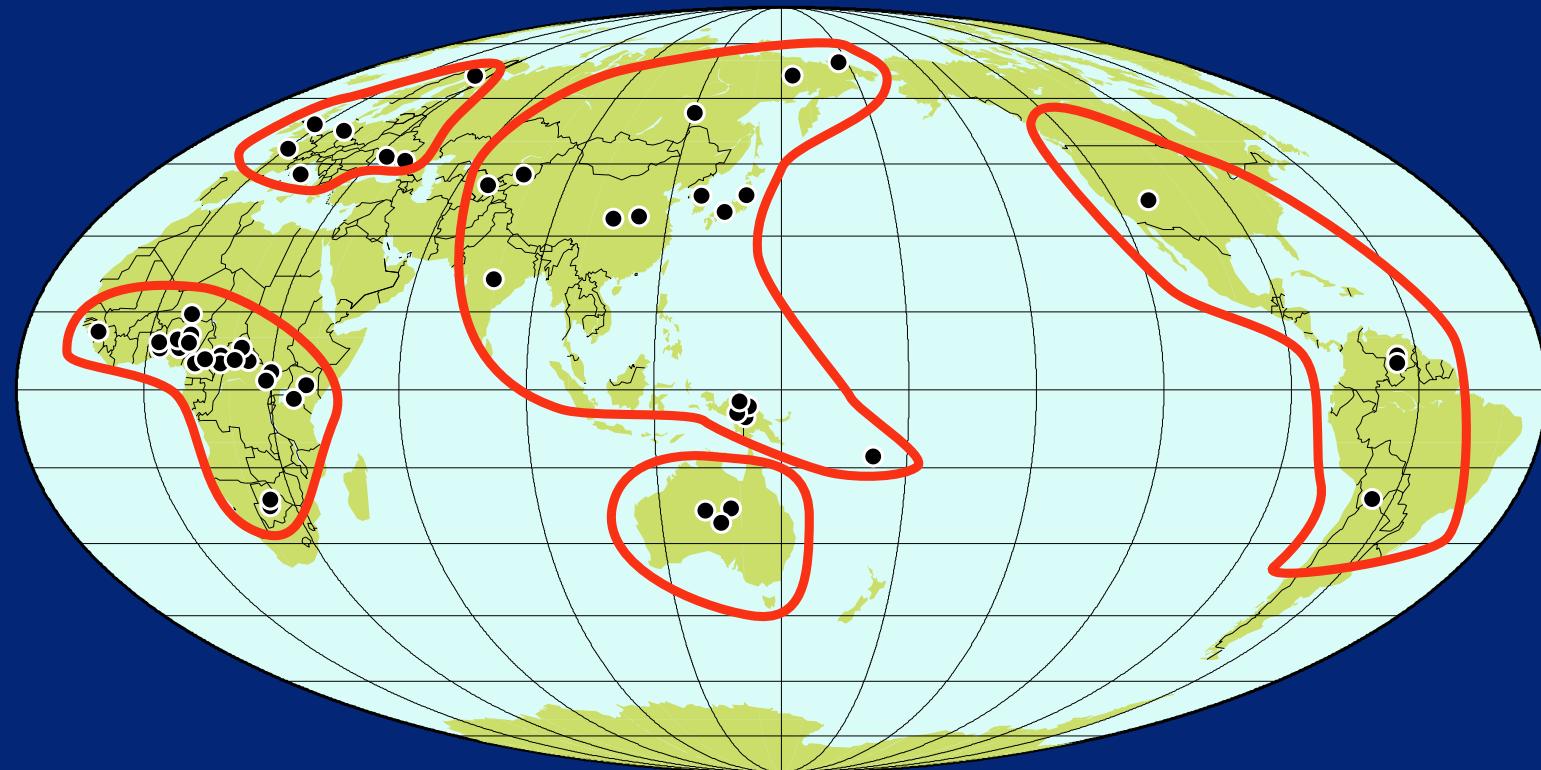
(Beerli and Felsenstein 1999)

100 two-locus datasets with 25 sampled individuals for each of 2 populations and 500 base pairs (bp) per locus.

	Population 1		Population 2	
	$\Theta$	$4N_e^{(1)}m_1$	$\Theta$	$4N_e^{(2)}m_2$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

## Complete mtDNA from 5 human “populations”

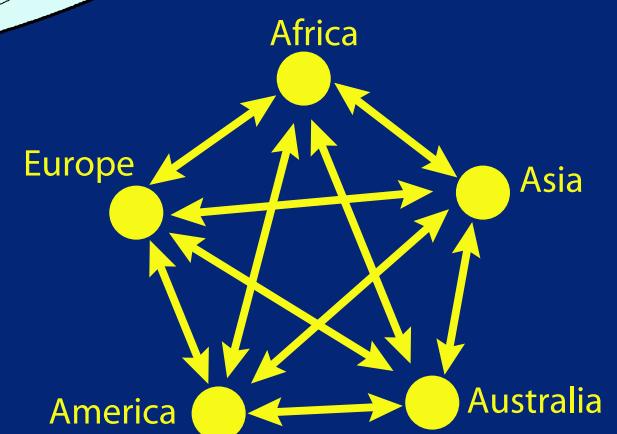
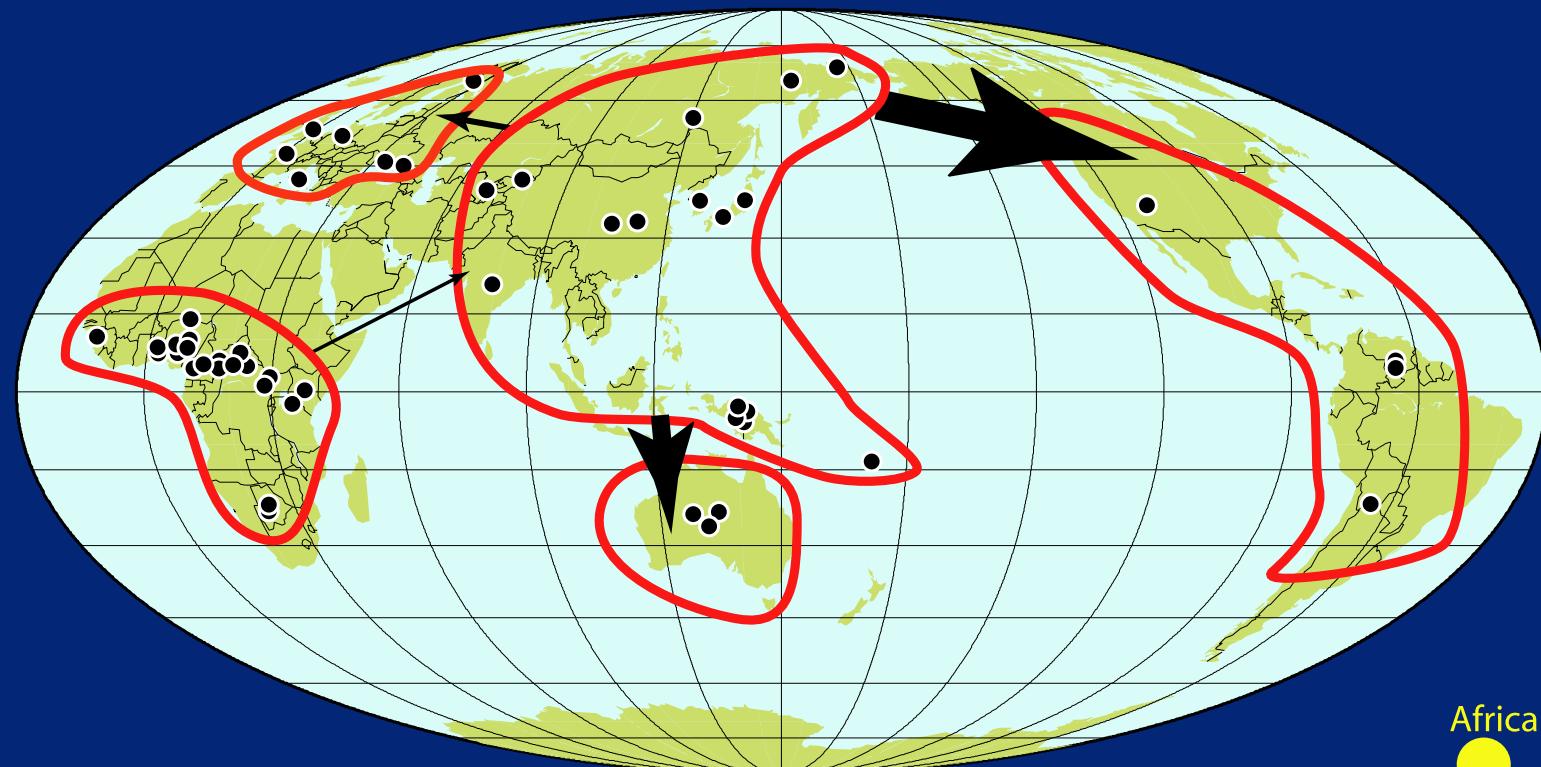
---



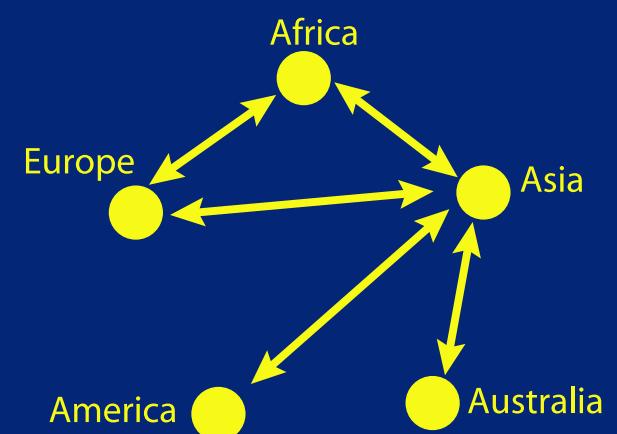
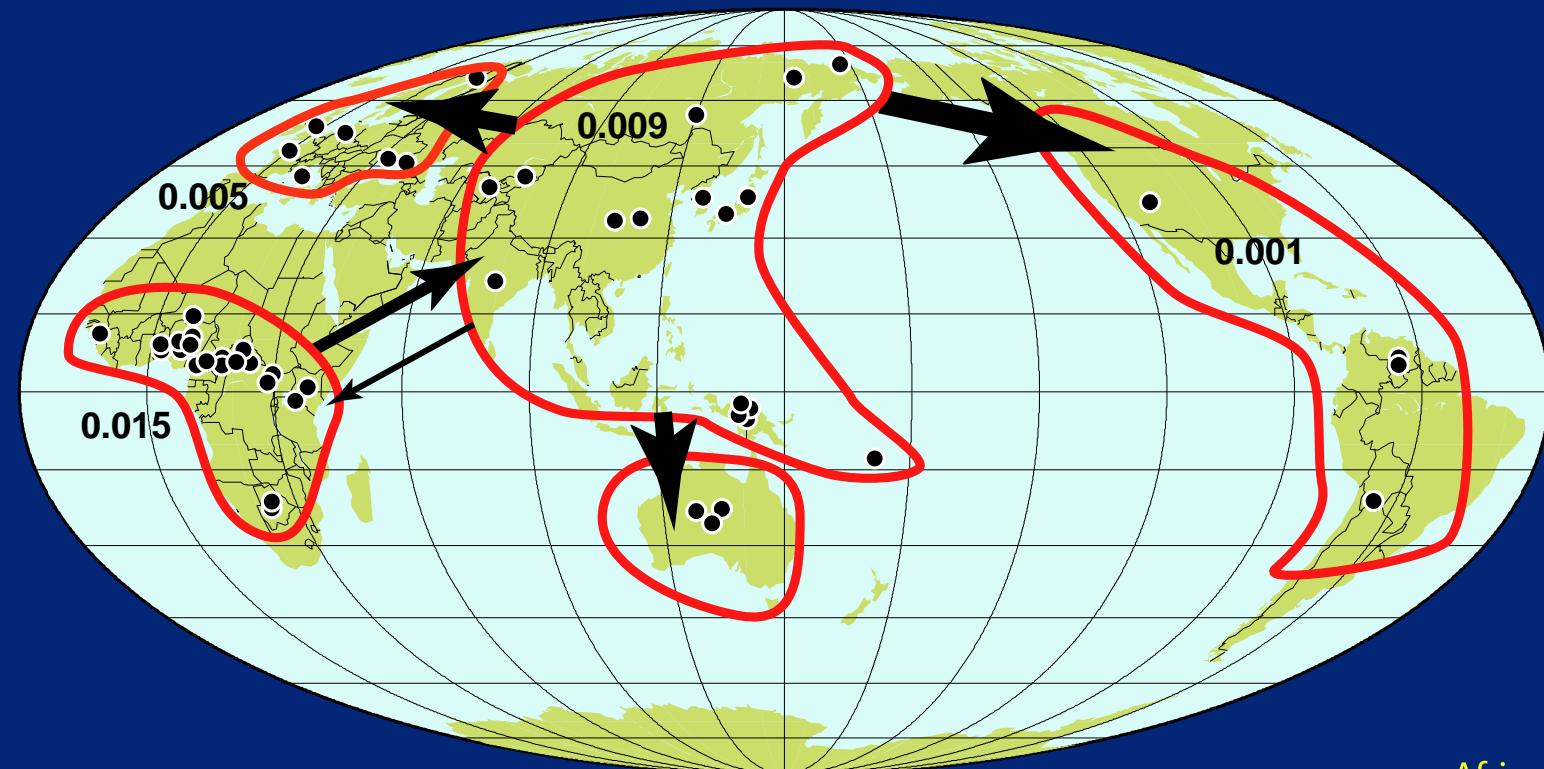
A total of 53 complete mtDNA sequences ( $\sim 16$  kb):  
Africa: 22, Asia: 17, Australia: 3, America: 4, Europe: 7.  
Assumed mutation model: F84+ $\Gamma$

## Full model: 5 population sizes + 20 migration rates

---



## Restricted model: only migration into neighbors allowed



## The coalescent in nuclear genes

---

- mtDNA has a single coalescent tree
- The nuclear genome has multiple trees due to recombination and assortment
- Say that my 17q comes from my maternal grandfather and my 17p comes from my maternal grandmother
- They will have totally different coalescent ancestries



## The coalescent in nuclear genes

---

- “Haplotype blocks” have the same common ancestor
- Length of blocks depends on:
  - Frequency of recombination (breaks up blocks)
  - Rate of coalescences (determines how distant common ancestors are)
  - Coalescence rate depends on population size—large populations have shorter blocks of linkage disequilibrium
  - Presence of hotspots
- The HapMap project tries to capitalize on this block structure

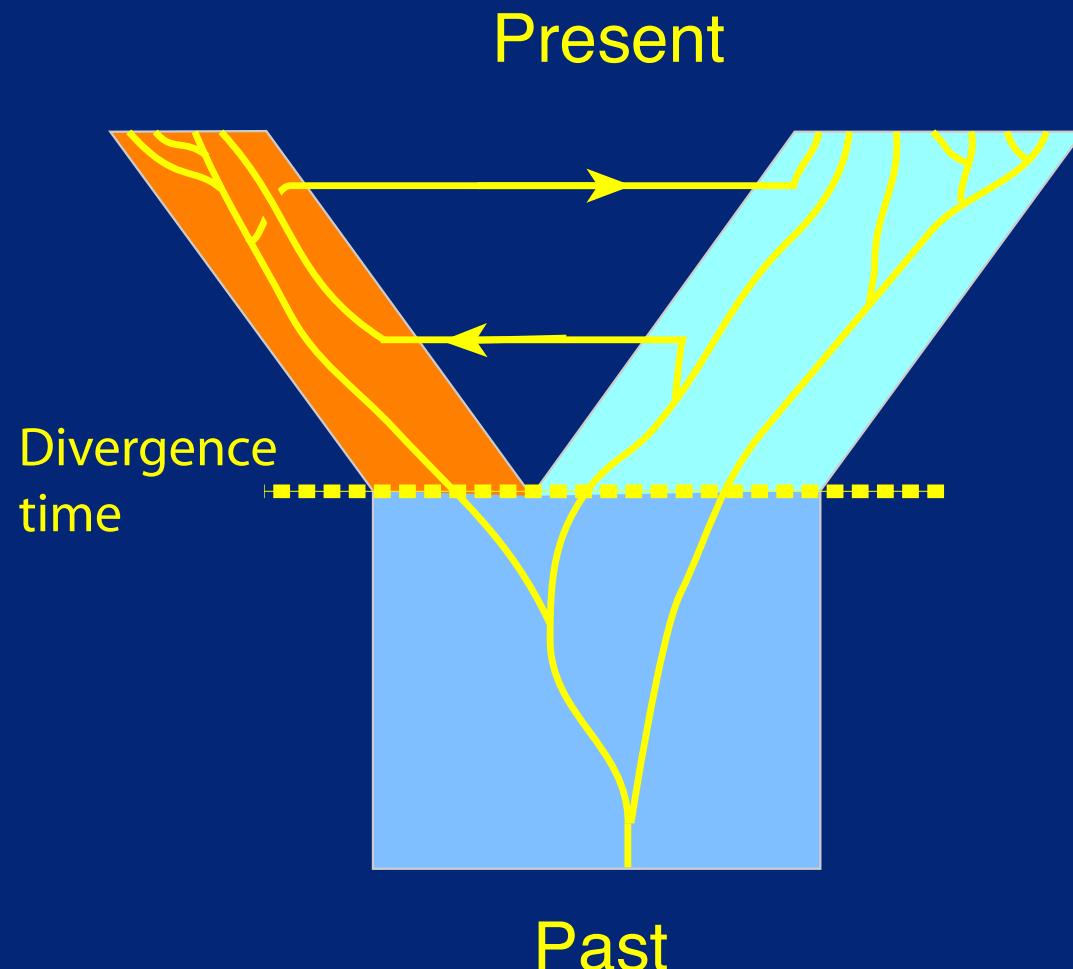
## Ancestral recombination graph (ARG)

---

- The history of the nuclear genome is a tangled graph
- I wrote a program (LAMARC) to search over such graphs using MCMC
- Much remains to be done in this field
- Many current approaches use very short sequences to try to avoid recombination
- More information is available from a full analysis, but it's hard

## Estimation of divergence time

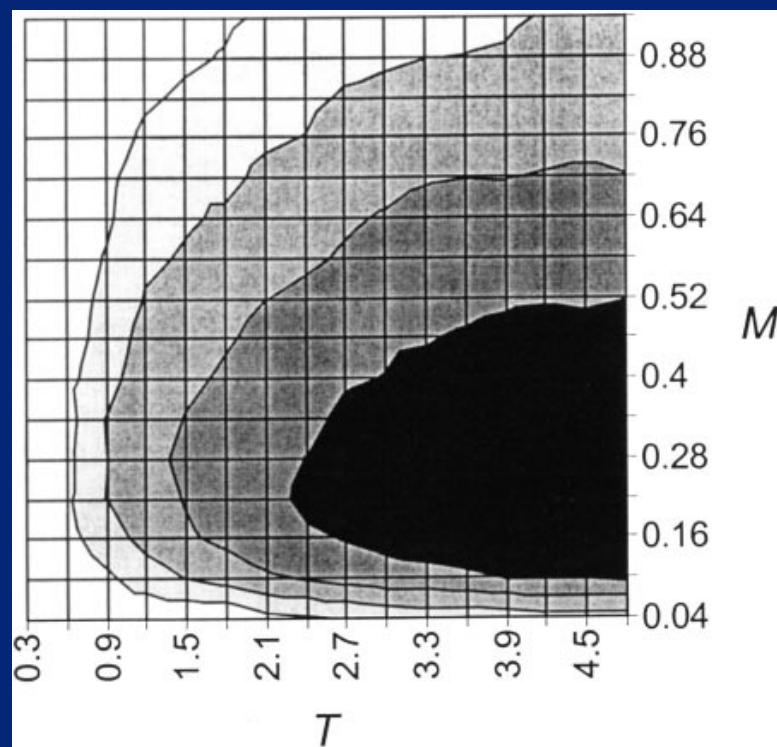
Wakeley and Nielsen (2001)



## Estimation of divergence time

---

Wakeley and Nielsen (2001) Figure 7. The joint integrated likelihood surface for  $T$  and  $M$  estimated from the data by Ortí et al. (1994). Darker values indicate higher likelihood.



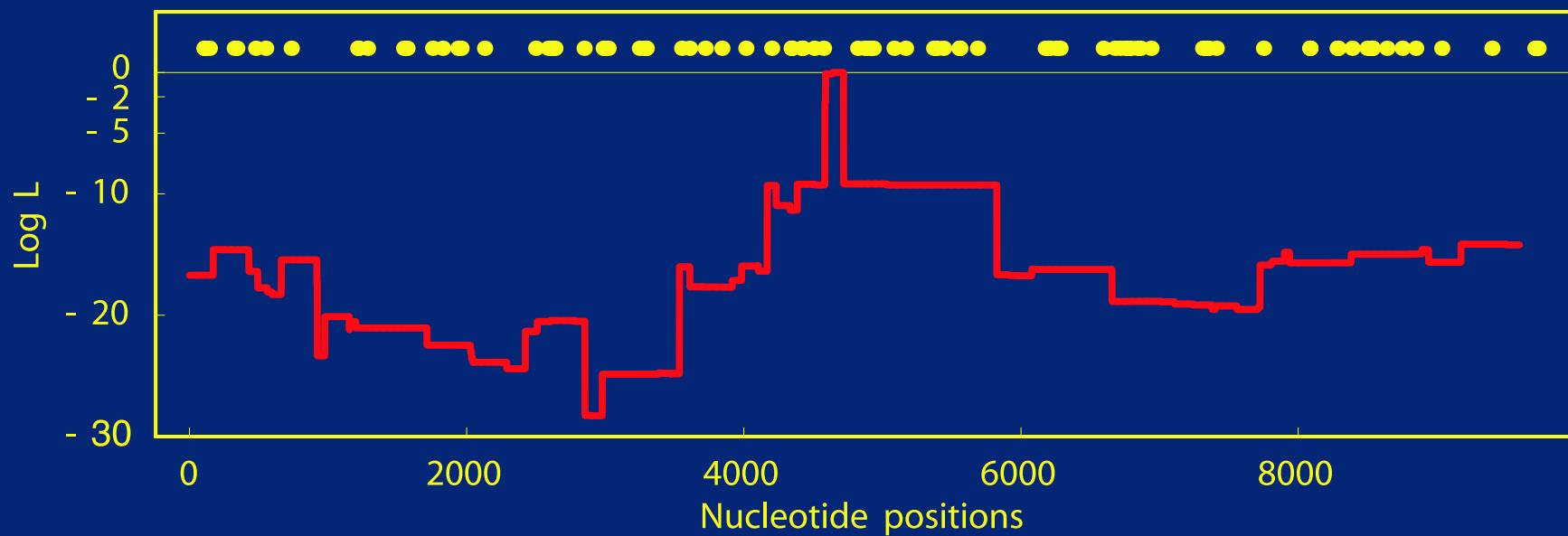
## Multiple time points

---

- Ancient DNA or historical samples of fast-evolving organisms
- Points must be:
  - Dated
  - Far enough apart for measurable evolution
- Advantages:
  - Separation of  $\Theta$  into  $N_e$  and  $\mu$
  - Much better resolution of growth rates

## Linkage disequilibrium mapping

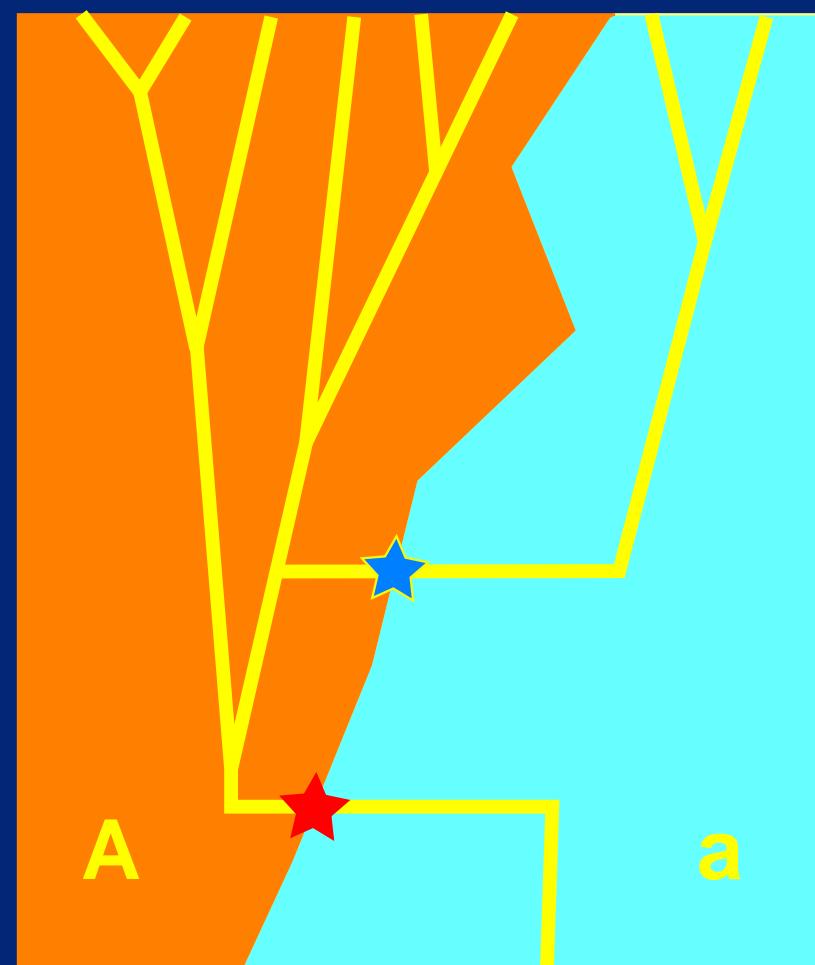
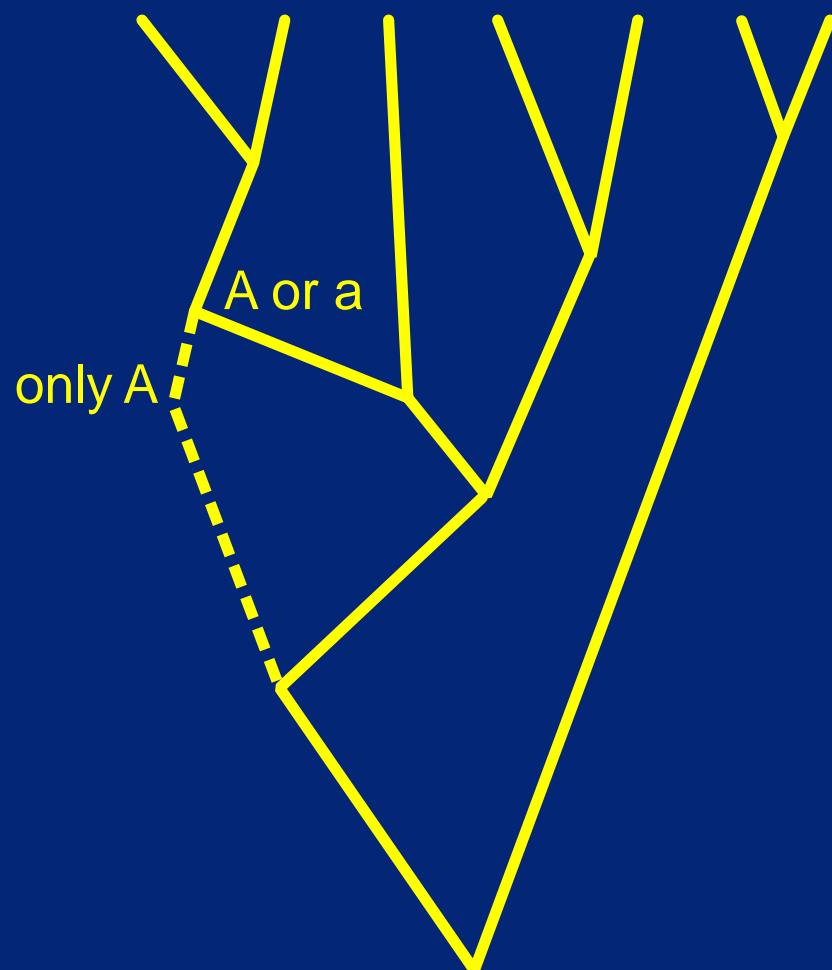
With a disease mutation model we can use the recombination estimator to post-analyze the sampled genealogies that were used to estimate  $r$  and find the location of the disease mutation on the DNA.



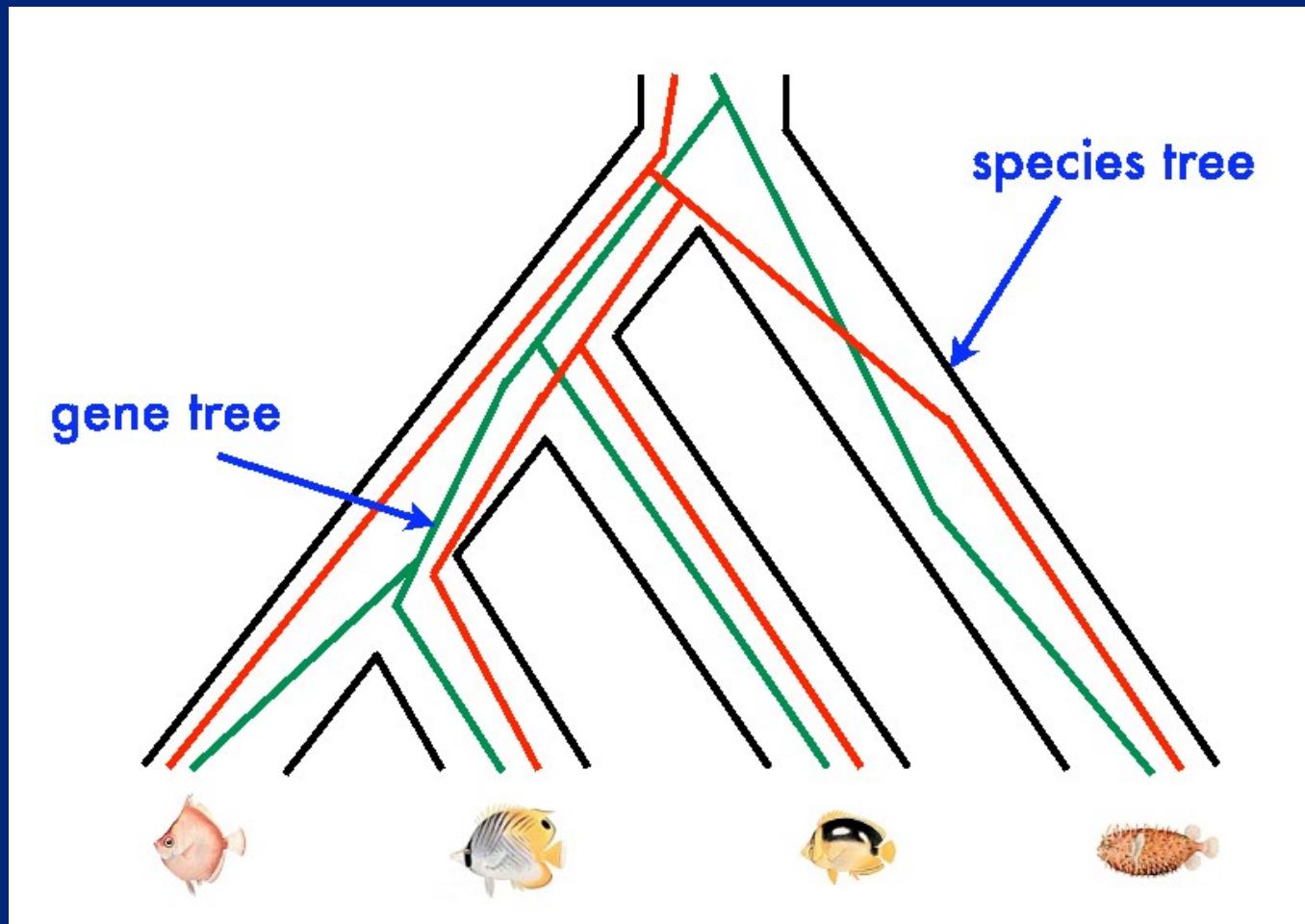
## Selection coefficient estimation

---

Krone and Neuhauser (1999), Felsenstein (unpubl)



## Coalescence at the boundary of speciation



## Coalescence at the boundary of speciation

---

- When we draw species phylogenies we hope the gene tree reflects the species tree
- For a given gene this may not be true:
  - More than one lineage survives from the ancestral species
  - They have a random coalescent back there
  - Their coalescence pattern won't reflect the speciation pattern as they coalesce before the species formed

## Coalescence at the boundary of speciation

---

- This is called “ancestral polymorphism” or “incomplete lineage sorting”
- It is mainly a problem for recent species
- Modern species tree methods need to take it into account
- Two major approaches:
  - Concatenate the genes and hope that the true signal wins out
  - Make a separate tree for each gene and try to reconcile them

## **What is the effective population size of red drum?**

---

Red drum, *Sciaenops ocellatus*, are large fish found in the Gulf of Mexico.



Turner, Wares, and Gold

Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish

Genetics 162:1329-1339 (2002)

## **What is the effective population size of red drum?**

---

- Census population size: 3,400,000
- Effective population size: ?
- Data set:
  - 8 microsatellite loci
  - 7 populations
  - 20 individuals per population

# What is the effective population size of red drum?

---

Three approaches:

1. Allele frequency fluctuation from year to year

- Measures current population size
- May be sensitive to short-term fluctuations

2. Coalescent estimate from *Migrate*

- Measures long-term harmonic mean of population size
- May reflect past bottlenecks or other long-term effects

3. Demographic models

- Attempt to infer genetic size from census size
- Vulnerable to errors in demographic model
- Not well established for long-lived species with high reproductive variability

## **Population model used for Migrate**

---

- Multiple populations along Gulf coast
- Migration allowed only between adjacent populations
- Allowing for population structure should improve estimates of population size



## **What is the effective population size of red drum?**

---

Estimates:

Census size ( $N$ ): 3,400,000

Allele frequency method ( $N_e$ ): 3,516 (1,785-18,148)

Coalescent method ( $N_e$ ): 1,853 (317-7,226)

The demographic model can be made consistent with these only by assuming enormous variance in reproductive success among individuals.

## **What is the effective population size of red drum?**

---

- Allele frequency estimators measure current size
- Coalescent estimators measure long-term size
- Conclusion: population size and structure have been stable

## **What is the effective population size of red drum?**

---

- Effective population size at least 1000 times smaller than census
- This result was highly surprising
- Red drum has the genetic liabilities of a rare species
- Turner et al. hypothesize an “estuary lottery”
- Unless the eggs are in exactly the right place, they all die