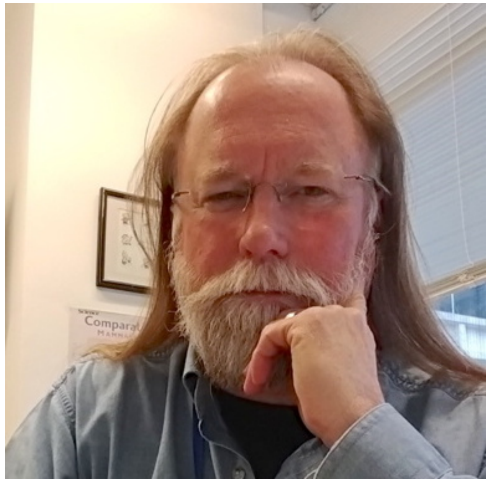
# AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

## PRINCIPLES OF COALESCENT THEORY AND POPULATIONS

# Today's Instructor



**Dr. Kurt Wollenberg**,
Ph.D. in Genetics

Ongoing Computational Biology projects:

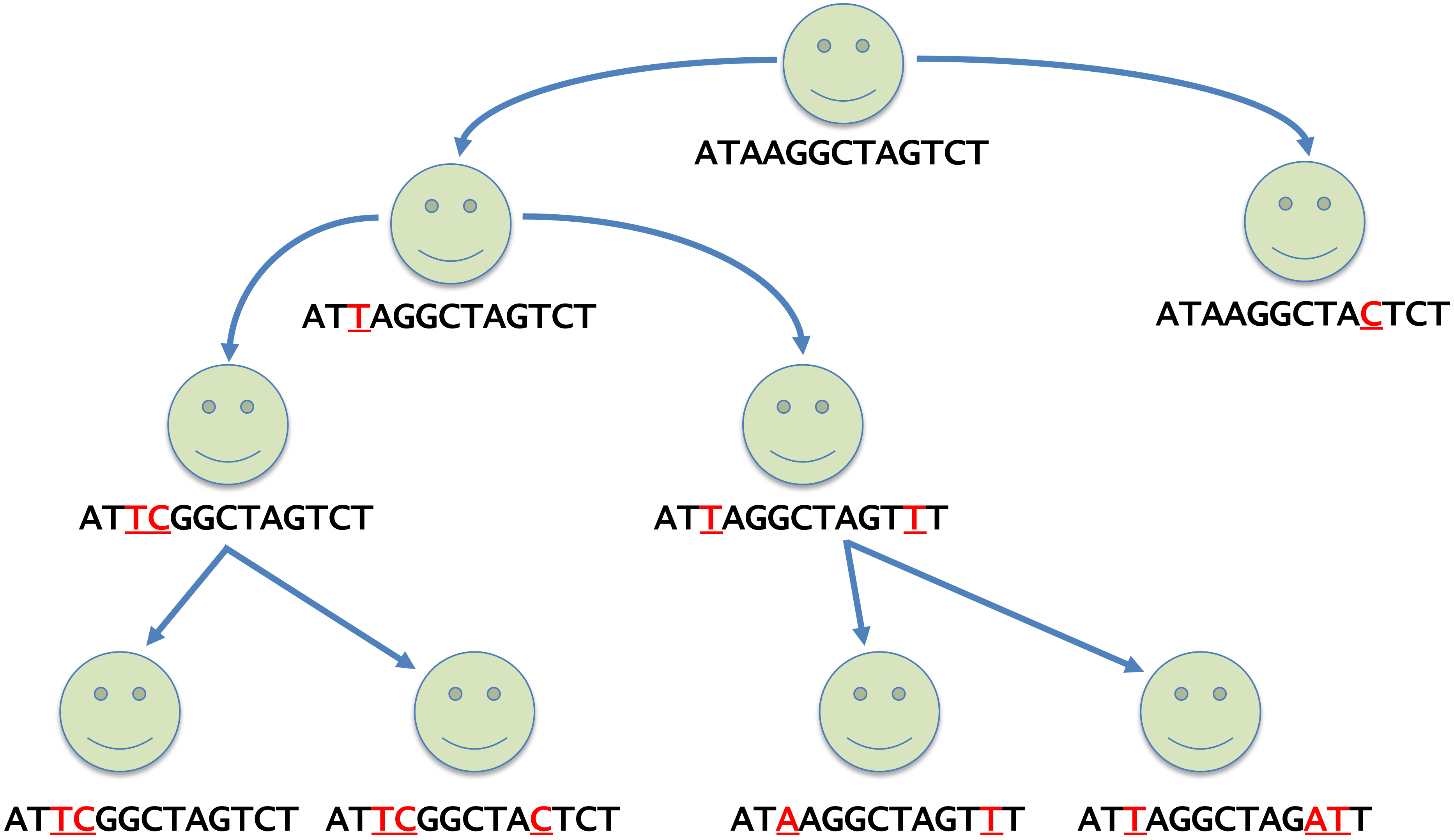
- Hepatitis B molecular evolution
- CLAG protein family evolution

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor: kurt.wollenberg@nih.gov

# Class Materials

- Directory on Uganda ACE server:
  - File directory: user@kla-ac-bio-03:/home/bcbb_teaching_files
  - Large data files
- NIAID github repository:
  - [https://github.com/niaid/Principles-of-Sequence-Analysis-and-Phylogenetics](https://github.com/niaid/Principles-of-Sequence-Analysis-and-Phylogenetics)
  - Code
  - Data files
  - Copies of lecture slides

# EVOLUTIONARY DIVERSITY



ATAAGGCTAGTCT

ATTAGGCTAGTCT

ATAAGGCTACTCT

ATTCGGCTAGTCT

ATTAGGCTAGTTT

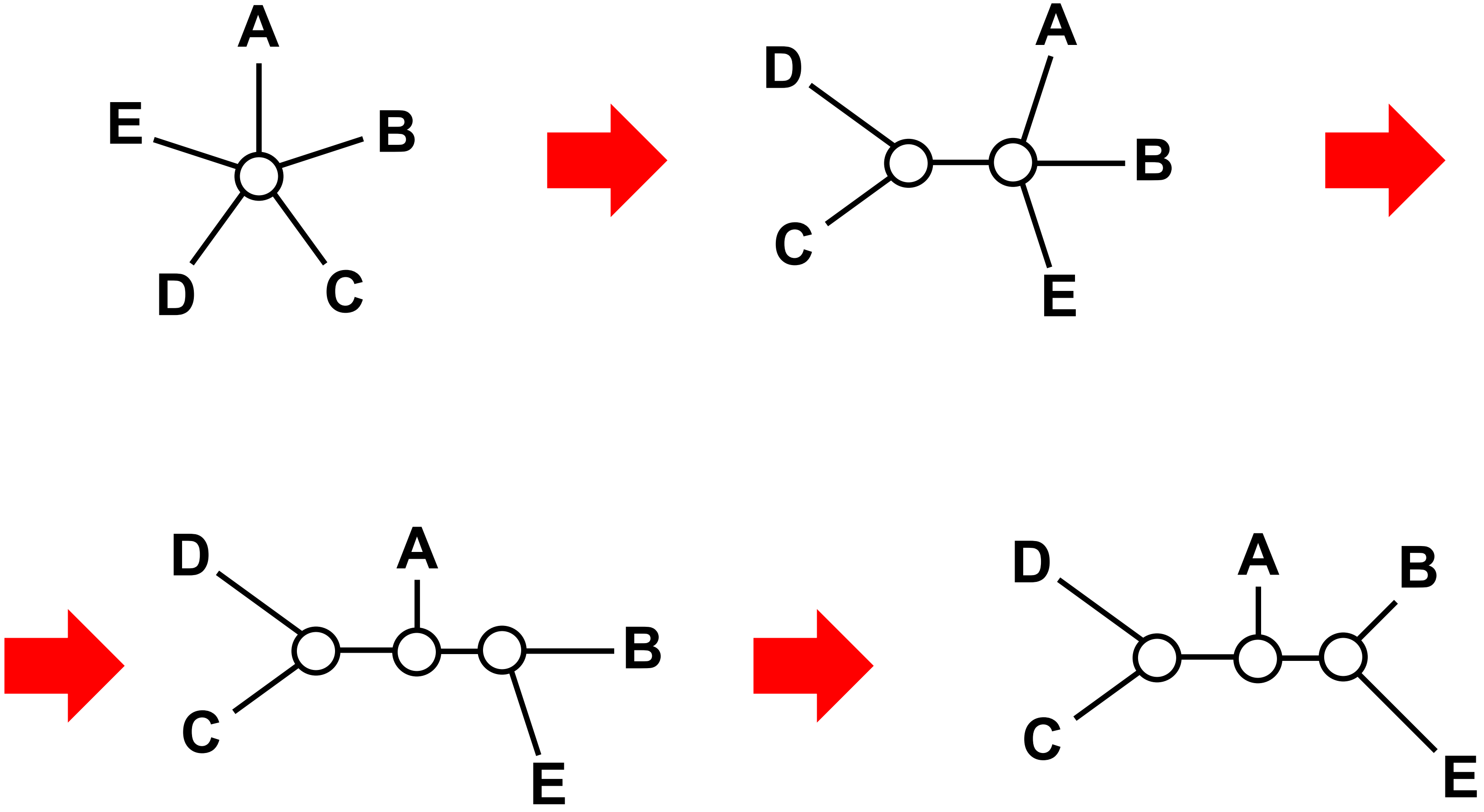
ATTCGGCTAGTCT

ATTCGGCTACTCT

ATAAGGCTAGTTT

ATTAGGCTAGATT

# EVOLUTIONARY DIVERSITY



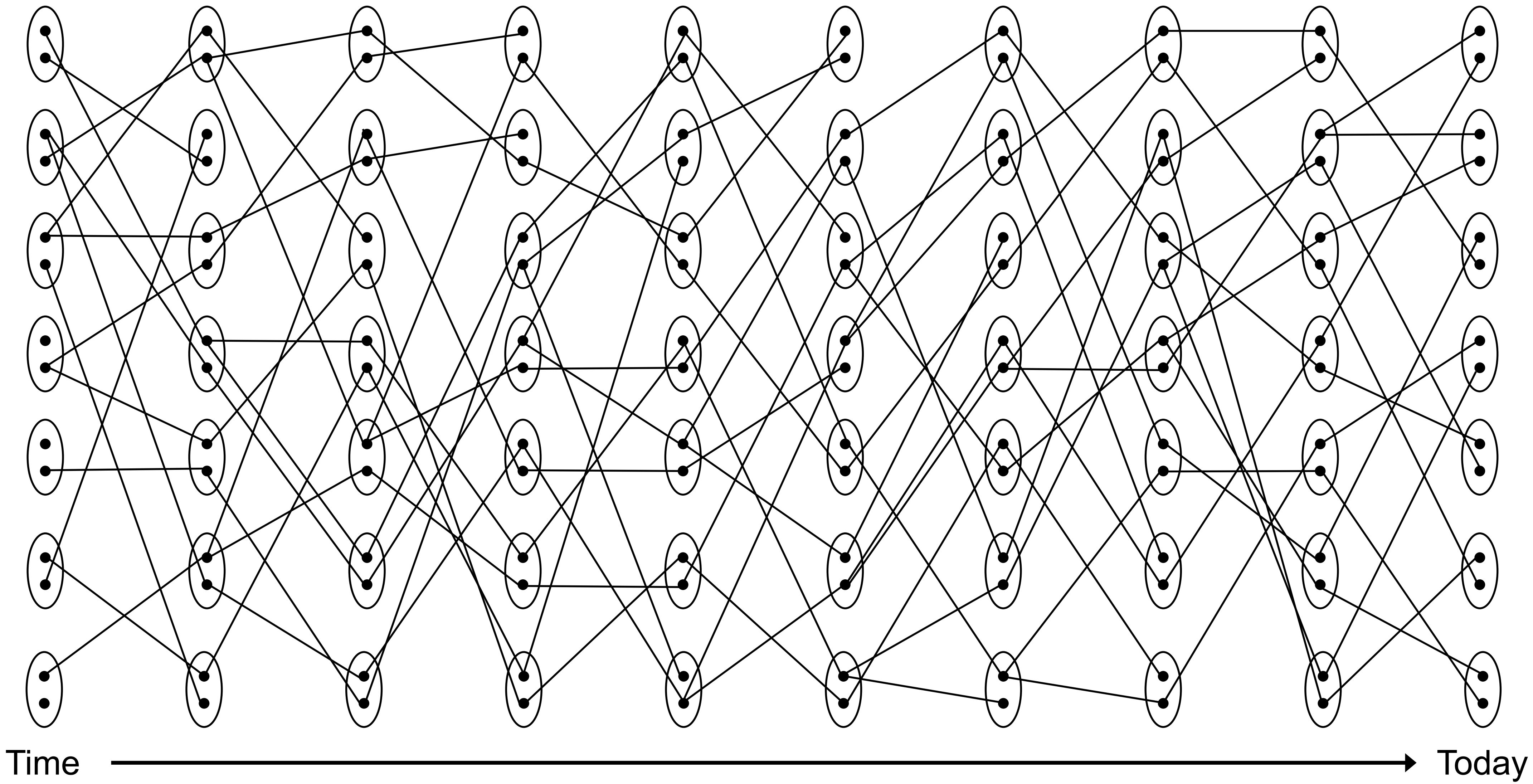
# EVOLUTIONARY DIVERSITY

Since most of our data are biological sequences, are there different ways to approach the data than from an organismal perspective?

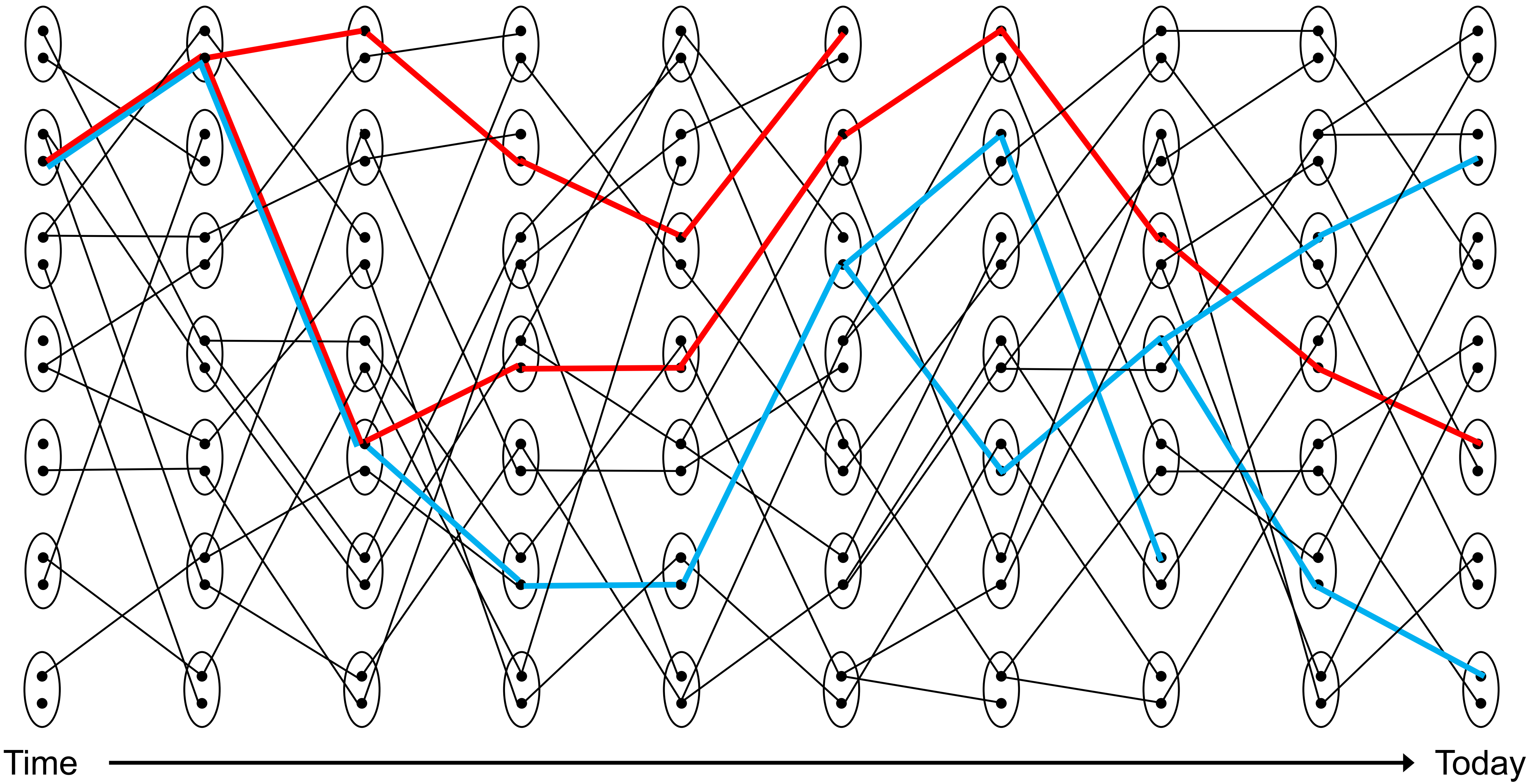
Gene genealogies

The relationships among members of a set of nonrecombining genetic elements not subject to selection on genotype.

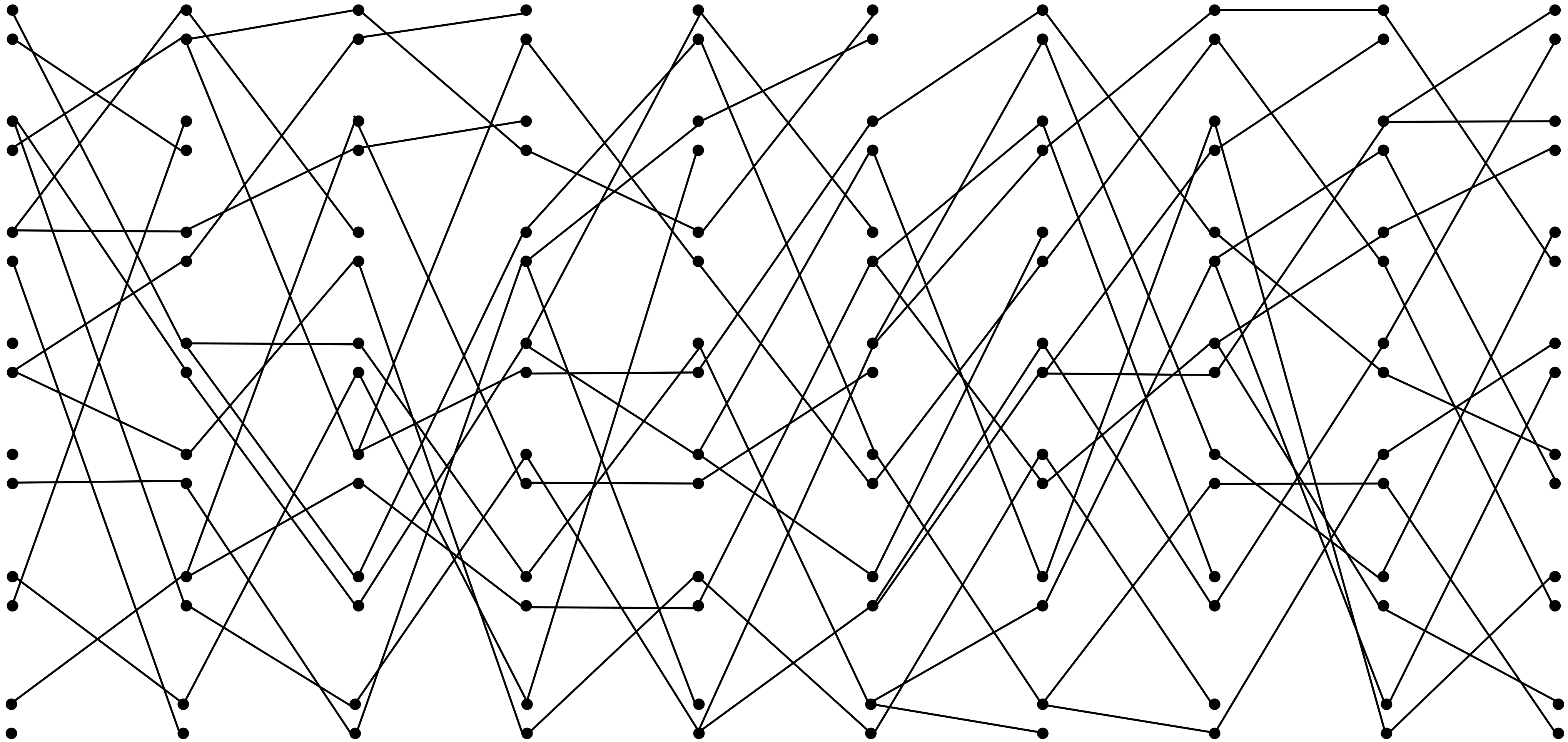
# GENE GENEALOGIES



# GENE GENEALOGIES

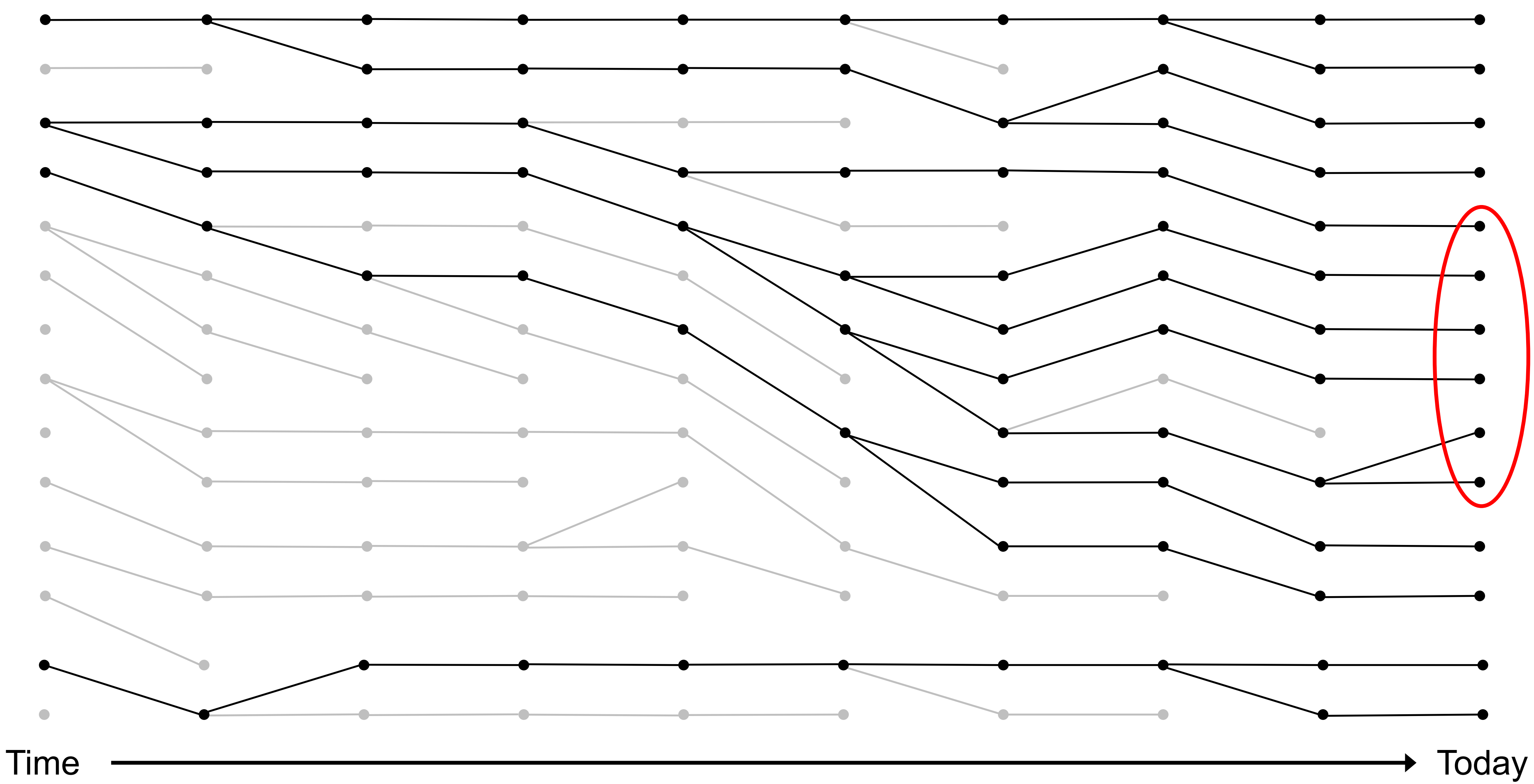


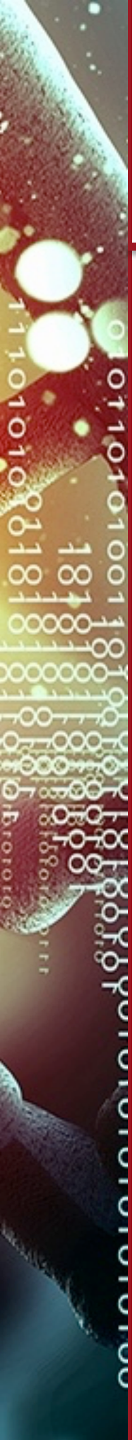
# GENE GENEALOGIES



Time → Today

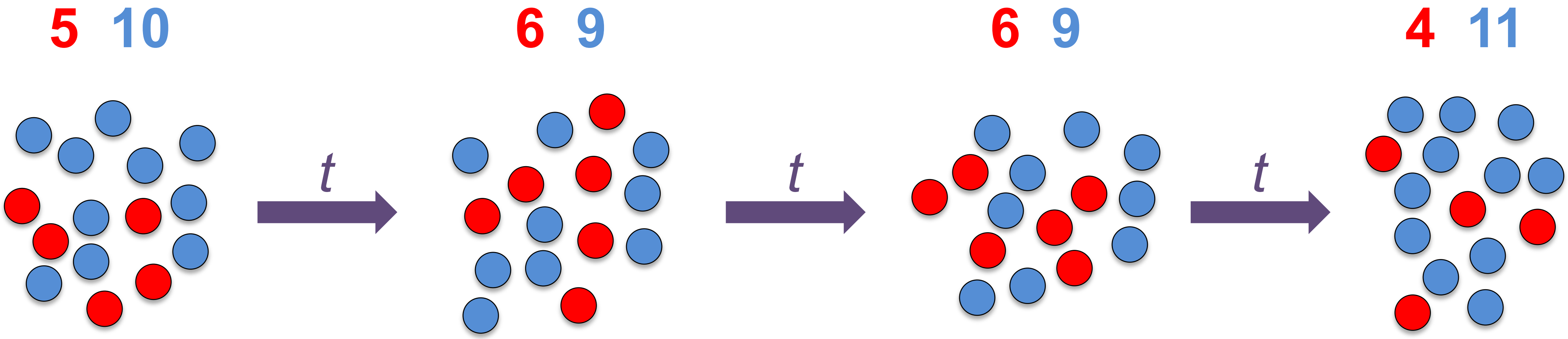
# GENE GENEALOGIES





# GENETIC DRIFT

Change in allele frequency over time



Why?

Directed causes: Changes the mean

**Stochastic causes: Changes the variance**

# GENETIC DRIFT

- Populations have a limited size.
- Not every member of a population will reproduce.
- Not every gamete produced by a reproducing individual will form an offspring
- This is more of an issue for sexual organisms, less for clonal organisms.
- This affects allele frequencies in the next generation.
- This is Genetic Drift.
- Genetic Drift – stochastic sampling of gametes for the generation of offspring.
- This ignores stochastic variation in survival or mating.

# GENETIC DRIFT

- Genetic drift is more pronounced the smaller the population.
- Small populations typically are the result of
  - Bottlenecks – an external event reduces the population to a small size
  - Founder events – a small subset of the original population becomes isolated
  - Skewed breeding structures – Only a few individuals of one gender will breed

# GENETIC DRIFT

Consider an ideal population of **diploid** organisms with **non-overlapping** generations of constant size. All individuals have the same fitness (**no selection**).

- Population size is *N*, therefore there are 2*N* genes at each locus.
- Consider one locus with only two alleles, $A_1$ and $A_2$.
- Allele frequencies are *p* for $A_1$ and *q* (= 1-*p*) for $A_2$.
- In each generation 2*N* gametes are sampled from an infinite gamete pool.

The probability of sampling *i* genes of type $A_1$ is

$$P_i = \frac{(2N)!}{i!\,(2N-i)!} p^i q^{2N-i}$$

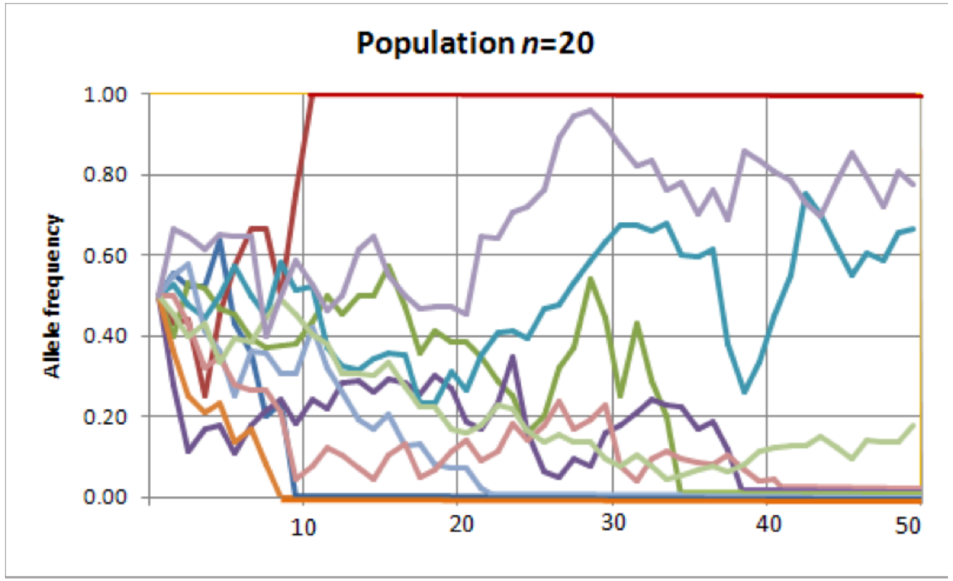
# GENETIC DRIFT

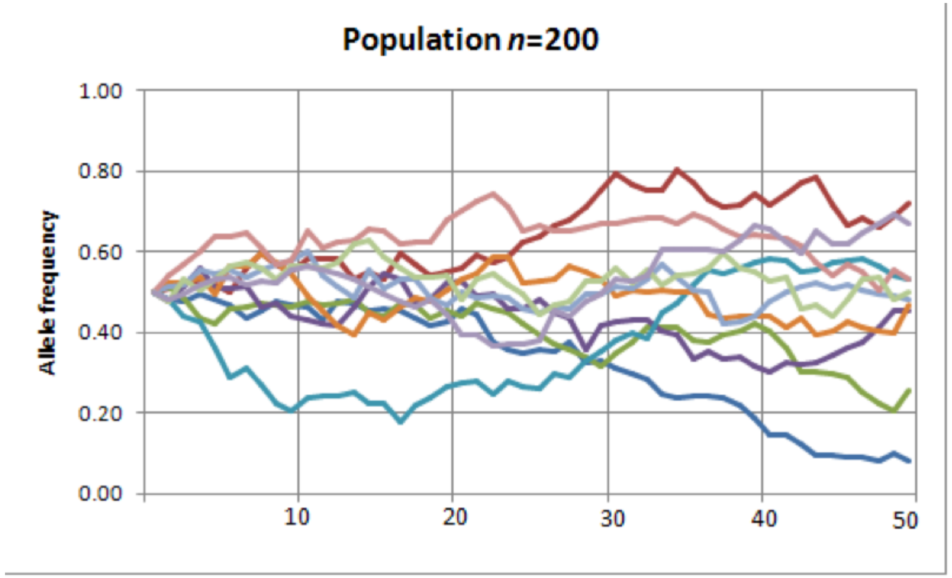
- The expected frequency of $p_i$ in each generation is $E(p_i) = p_0$
- The variance of $p_i$ in each generation is

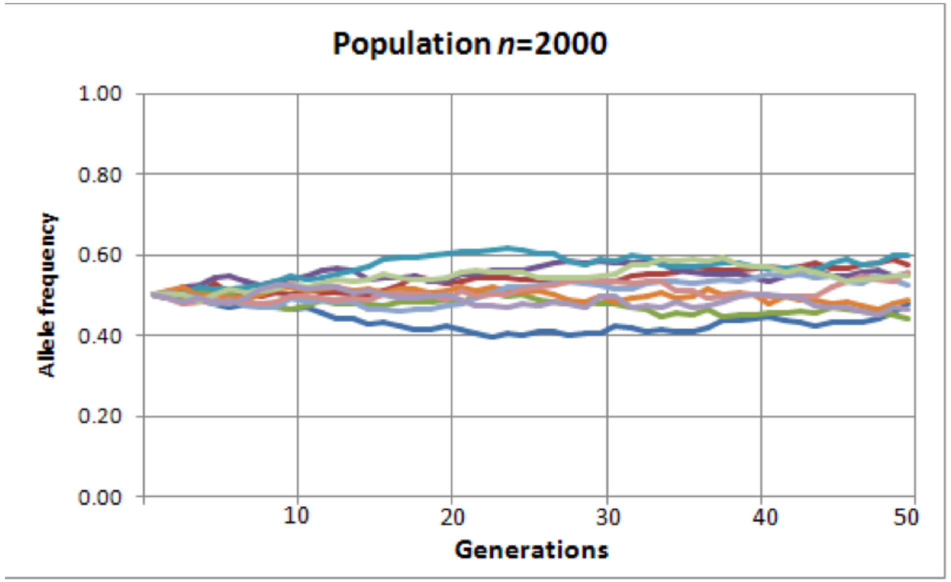
$$V(p_i) = \ p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^t\right] \approx p_0(1-p_0)(1-e^{-t/(2N)})$$

- Small populations increase the variance of allele frequencies from generation to generation.
- This increases the probability that an allele will be lost or become fixed in the next generation.

# GENETIC DRIFT

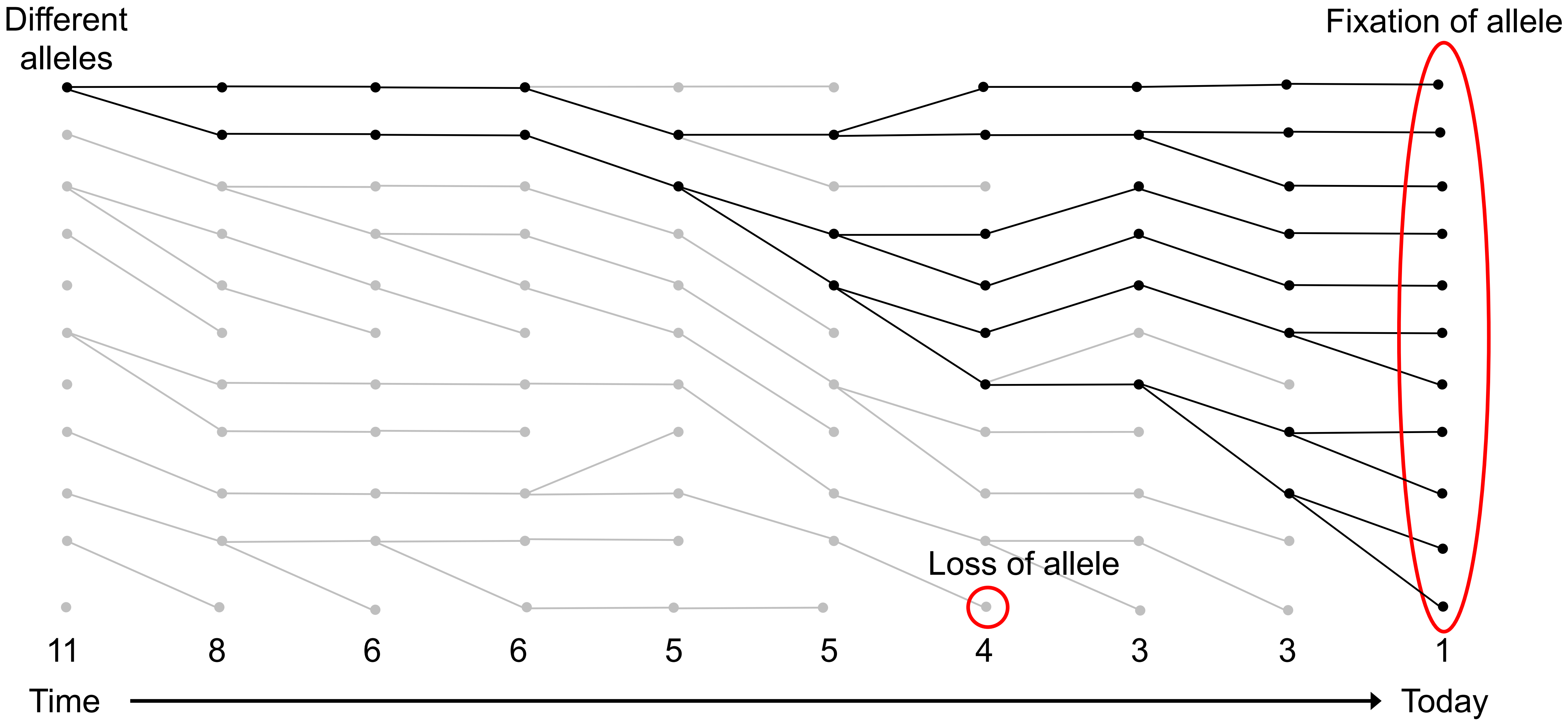

Population *n*=20


Population *n*=200


Population *n*=2000

# COALESCENT THEORY

## Genetic Drift



# POPULATION GENETICS

## Mutation – Drift Balance

So how will the process of mutation (adding new alleles to the population) affect how genetic drift (limiting allele sampling due to population size effects) drives alleles to fixation or loss?

# POPULATION GENETICS

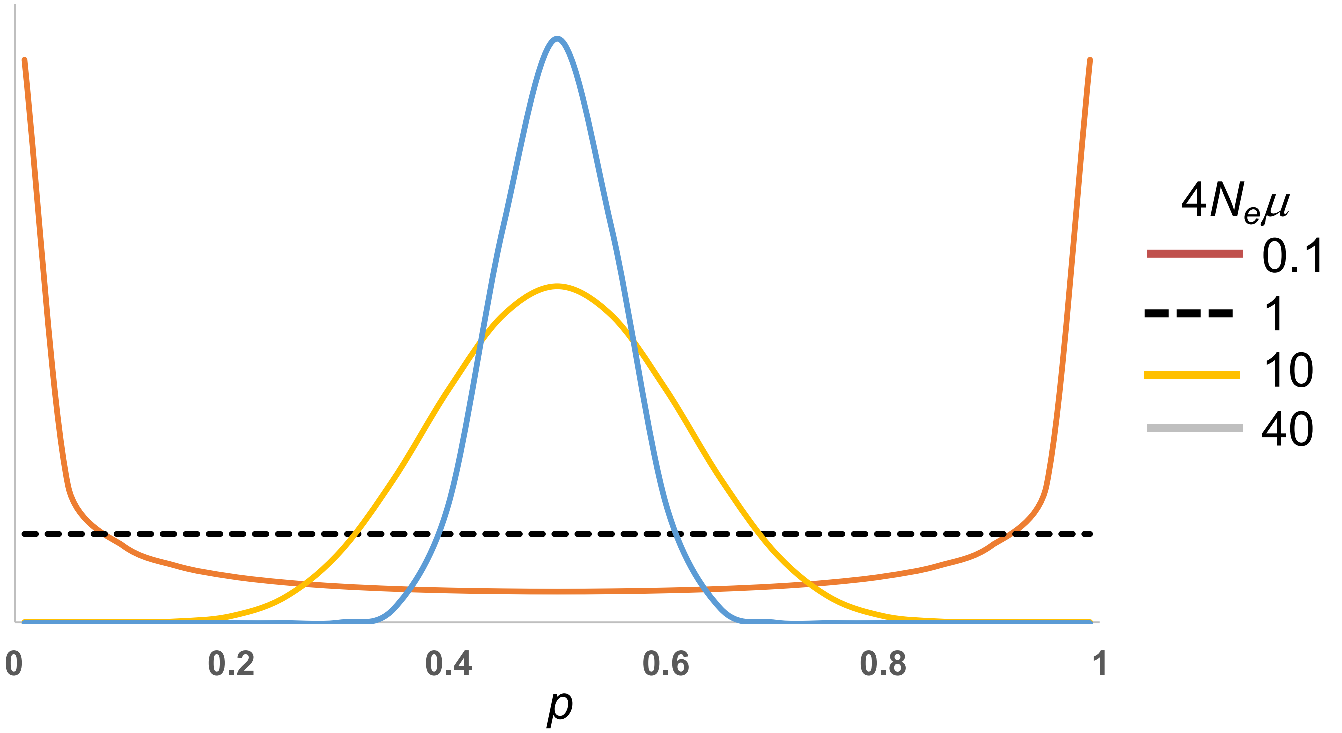
## Mutation – Drift Balance

- Mutation (Substitution) – generating new alleles, increasing population variation
  - A new allele occurs at a rate $\mu$ and has initial frequency 1/2*N*
- Drift - driving alleles to fixation or loss, reducing population variation
  - In a finite population the probability of fixation of a new allele is 1/2*N*
- Equilibrium heterozygosity: the probability that two randomly drawn alleles are different, averaged over all loci
- This is the probability that a mutation has occurred in one of the lineages (2$\mu$) divided by the probability of mutation or coalescence (1/2*N*) in these lineages
- Equilibrium heterozygosity: $H_e = \frac{2\mu}{2\mu + \frac{1}{2N}} = \frac{4N\mu}{4N\mu + 1}$

# POPULATION GENETICS

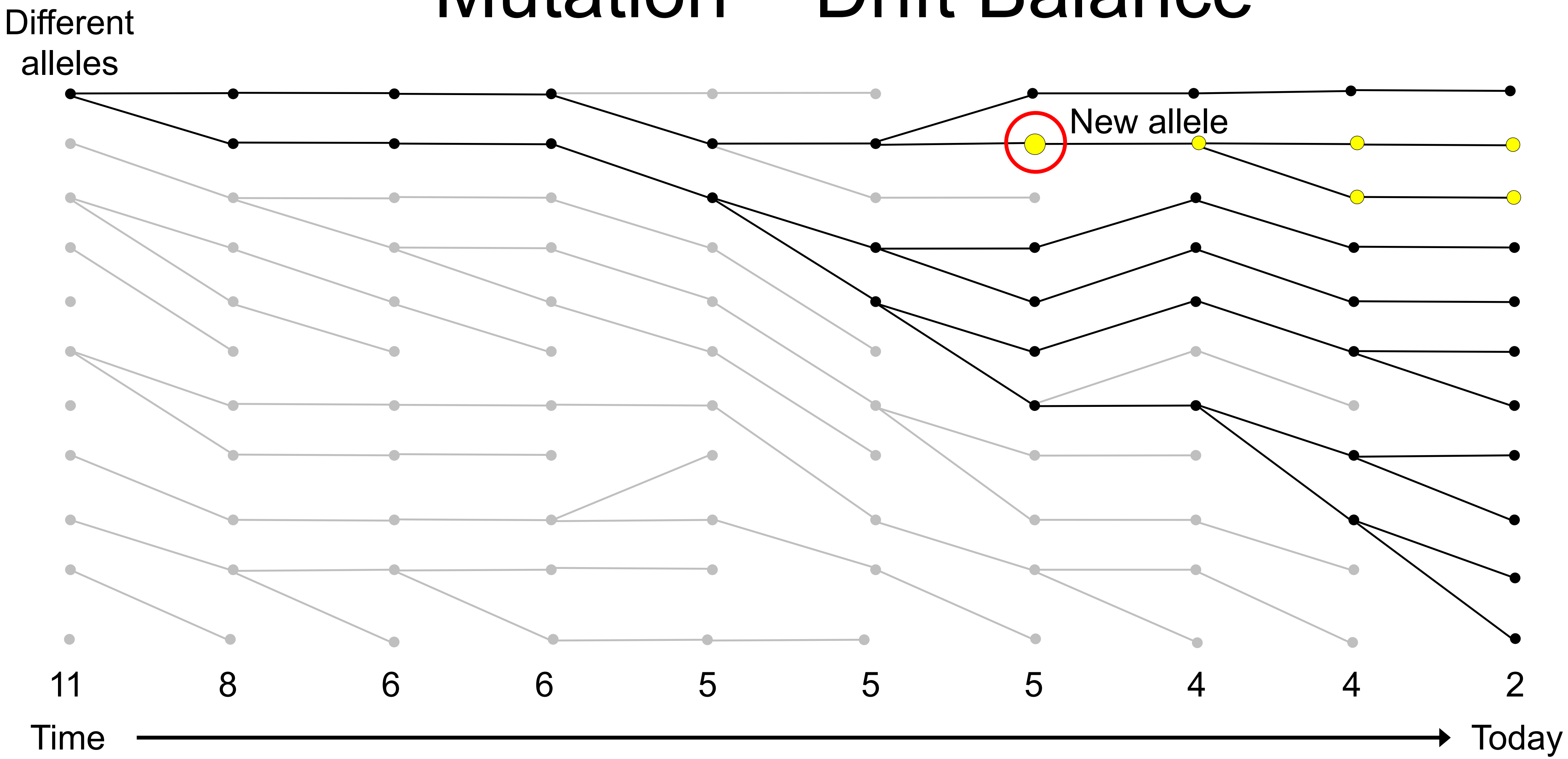
## Mutation – Drift Balance

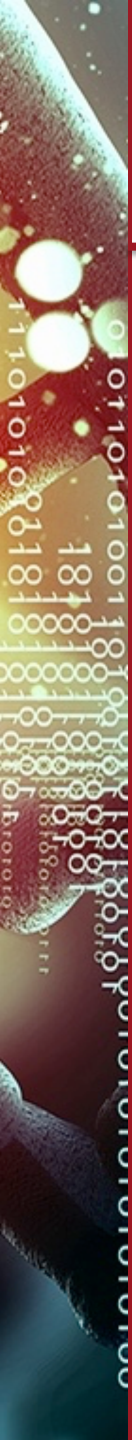
- Rather than equilibrium heterozygosity (which is expected to vary over time due to drift), what does the distribution of allele frequencies look like?



# POPULATION GENETICS

## Mutation – Drift Balance





# POPULATION GENETICS

## The Infinite Alleles Model

- If the process of mutation occurs randomly across the genome and the genome is very large it can be assumed that each mutation occurs at a different nucleotide
- This is the Infinite Sites Model of substitution
- Main impact of this model: there are no multiple substitutions
- The model of balance between substitution and drift is based on an infinite alleles model of substitution
- If the time to a common ancestor is small with respect to the product of the substitution rate and genome size this can be a reasonable model.

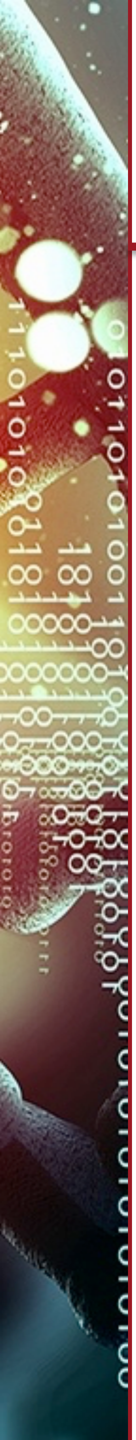
# POPULATION GENETICS

The Wright-Fisher population model

- *N* haploid individuals
- Random mating (panmictic) – all individuals have an equal probability of producing offspring
- Discrete (non-overlapping) generations
- No selection
- No migration in or out of the population

# COALESCENT MODELS

- Gene genealogy: the relationships among members of a set of nonrecombining genetic elements not subject to selection on genotype.
- Kingman (1982): a diploid population consists of 2N gene copies each with distinct genealogical histories.
- A coalescent approach examines data at the present time and models its behavior in the past.
- At generation *t* the state of the data in the previous generation (*t*+1) only depends on the states at generation *t*. This is, by definition, a Markov process.



# COALESCENT MODELS

- A coalescent event: two lineages in generation *t* have a common ancestor (coalesce) in generation *t*+1.
- For a diploid population with 2N alleles present in each generation the probability that two alleles in generation *t* have the same ancestor in generation *t*+1 is $P_C = \frac{1}{2N}$
- The probability these two alleles do not coalesce is $P_{NC} = 1 - \frac{1}{2N}$
- The probability that two alleles have not coalesced over *t* generations and then do at generation *t*+1 is $P_{C,t+1} = (1 - \frac{1}{2N})^t \frac{1}{2N}$

# COALESCENT MODELS

- The probability that two alleles have not coalesced over *t* generations and then do at generation *t*+1 is $P_{C,t+1} = (1 - \frac{1}{2N})^{t}\frac{1}{2N}$
- For reasonably large values of 2N (>100) this can be approximated as $P_{C,t+1} = \frac{1}{2N}e^{-\frac{t}{2N}}$
- For large values of t this approximates an exponential distribution, giving E(Time to coalescence) = 2*N* generations and Var(T) = 4$N^2$

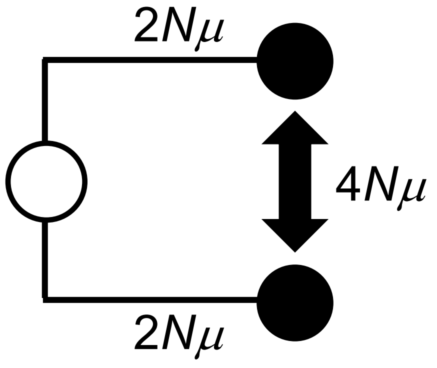
# COALESCENT MODELS

- For a sample *n* from a population of 2*N* genes there are *n*(*n*-1)/2 possible pairs with a common ancestor in the previous generation
- Each pair coalesces to an ancestor with probability 1/2*N*
- Now the probability of coalescing in generation *t*+1 is $P_{C,t+1} = (1 - \frac{n(n-1)}{4N})^t \frac{n(n-1)}{4N}$
- For reasonably large values of 2*N* this is approximately $P_{C,t+1} = \frac{n(n-1)}{4N} e^{-\frac{n(n-1)t}{4N}}$
- For large *t* the approximated exponential distribution
  - E(Time to coalescence) = 4*N*/*n*(*n*-1) generations
  - Var(T) = $16N^2/[n(n-1)]^2$

# COALESCENT MODELS

## The Coalescent (*n*-coalescent)

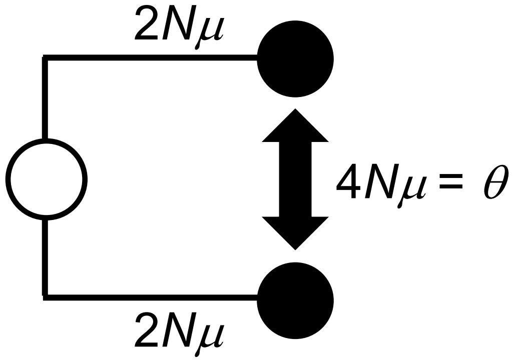
- For genetic data, coalescent intervals are typically expressed as the number of substitutions *b*, accumulating at rate $\mu$ over the 2*N* generations expected to the coalescence event



- Pr(coalescence) after *b* substitutions is $P_b = \frac{n(n-1)}{4N\mu} e^{-\frac{n(n-1)b}{4N\mu}}$

# COALESCENT MODELS

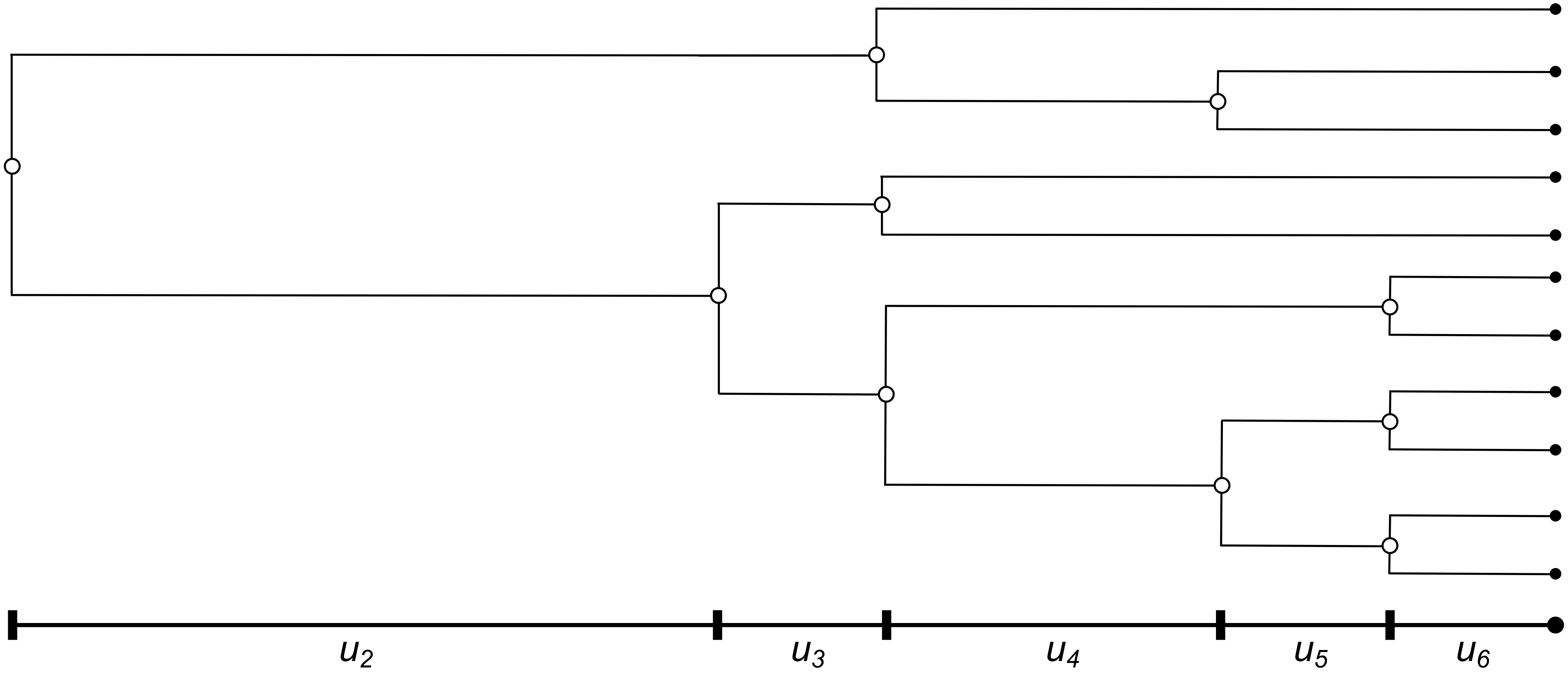
The Coalescent (*n*-coalescent)



- $4N\mu = \theta$, average number of mutational differences between any two randomly sampled sequences from a population with constant effective size

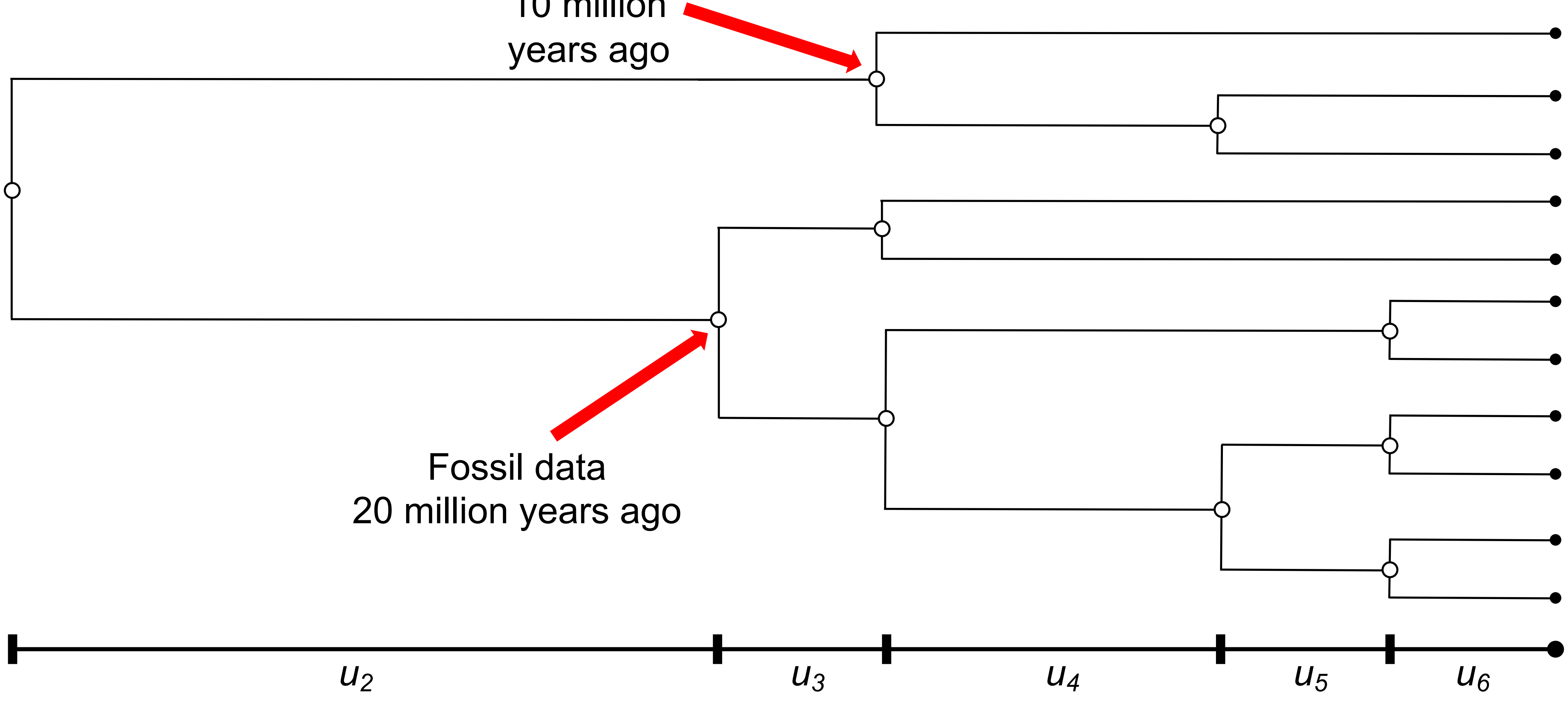
# COALESCENT THEORY

## Coalescent Intervals



# COALESCENT THEORY

## Estimating Time to Most Recent Common Ancestor

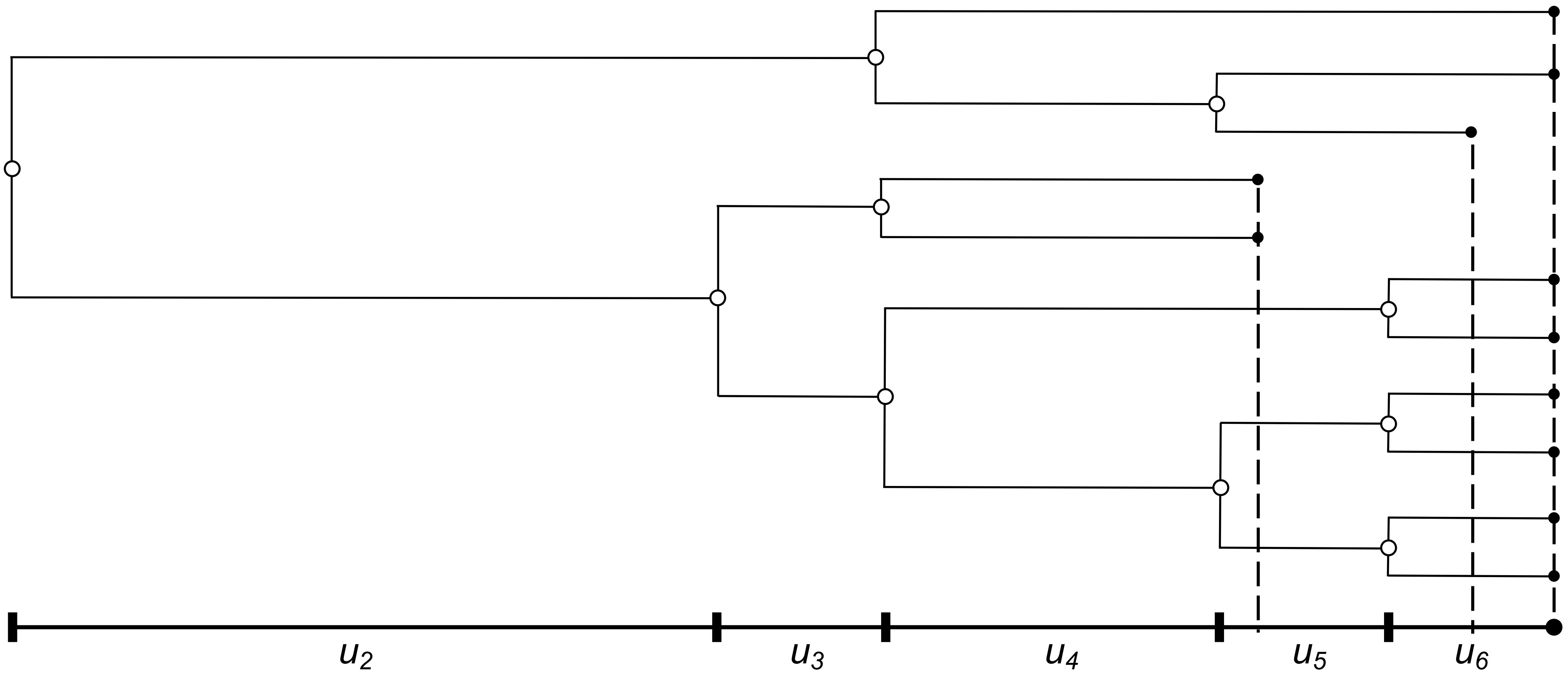


Fossil data
10 million
years ago

Fossil data
20 million years ago

# COALESCENT THEORY

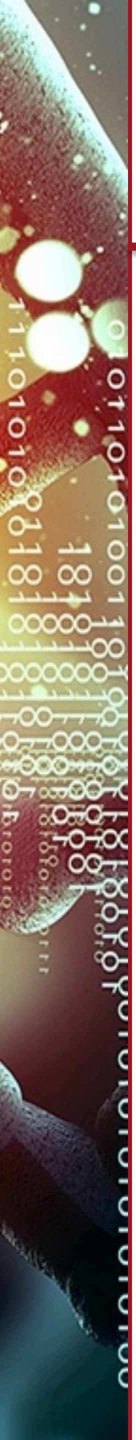
## Estimating Time to Most Recent Common Ancestor



# COALESCENT MODELS

## The Serial Coalescent (*s*-coalescent)

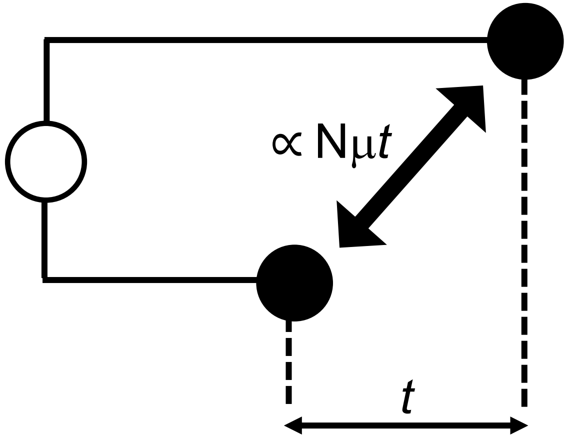
- With genetic data, coalescent intervals are typically expressed as the number of substitutions
- This means that time is scaled by substitution rate $\mu$
- For the standard coalescent, instead of N you must use $\theta = 2N\mu$ (haploid) or $4N\mu$ (diploid)
- Serial sampling (collecting sequences at different time points) allows you to separate time and substitution rate

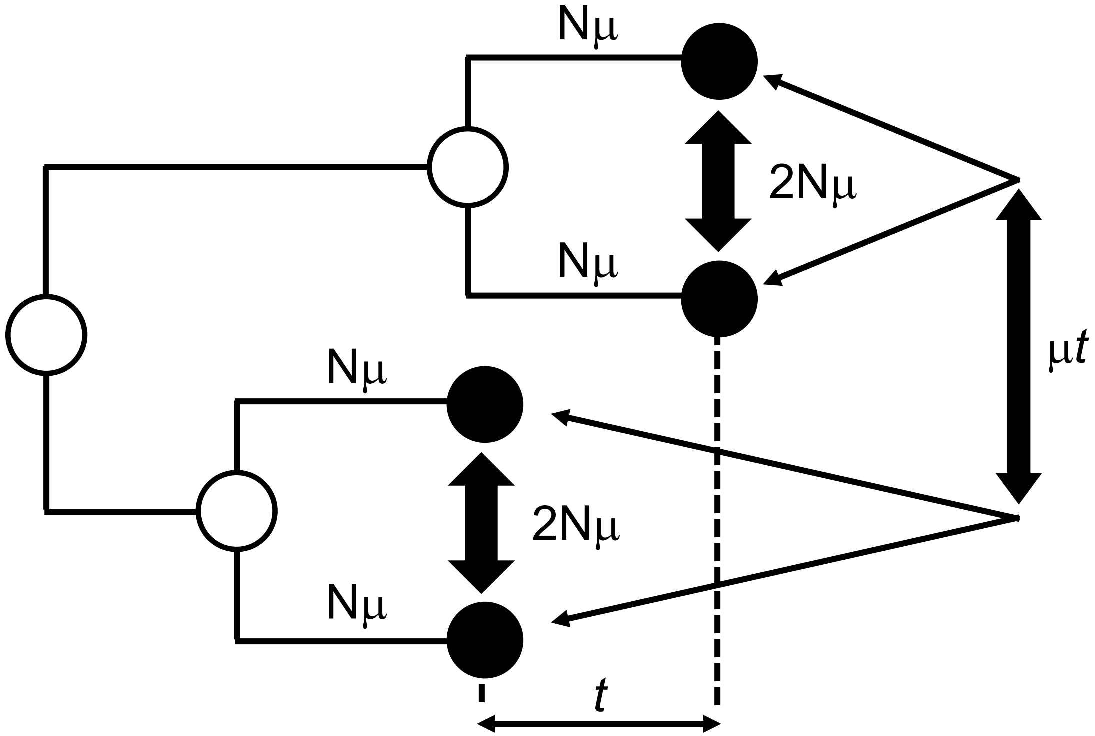


# COALESCENT MODELS

## The Serial Coalescent (*s*-coalescent)

- With serial sampled genetic data, coalescent intervals are now a function of time between samples and the substitution rate

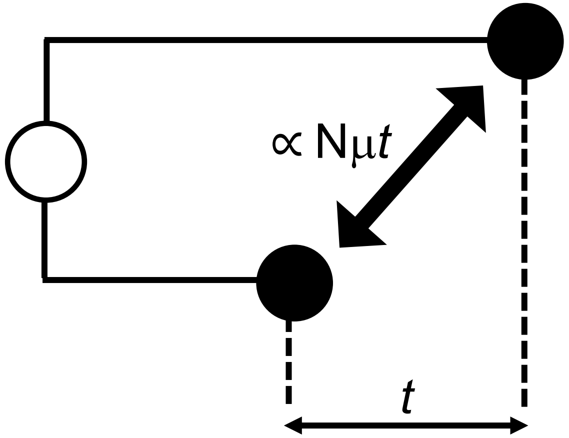


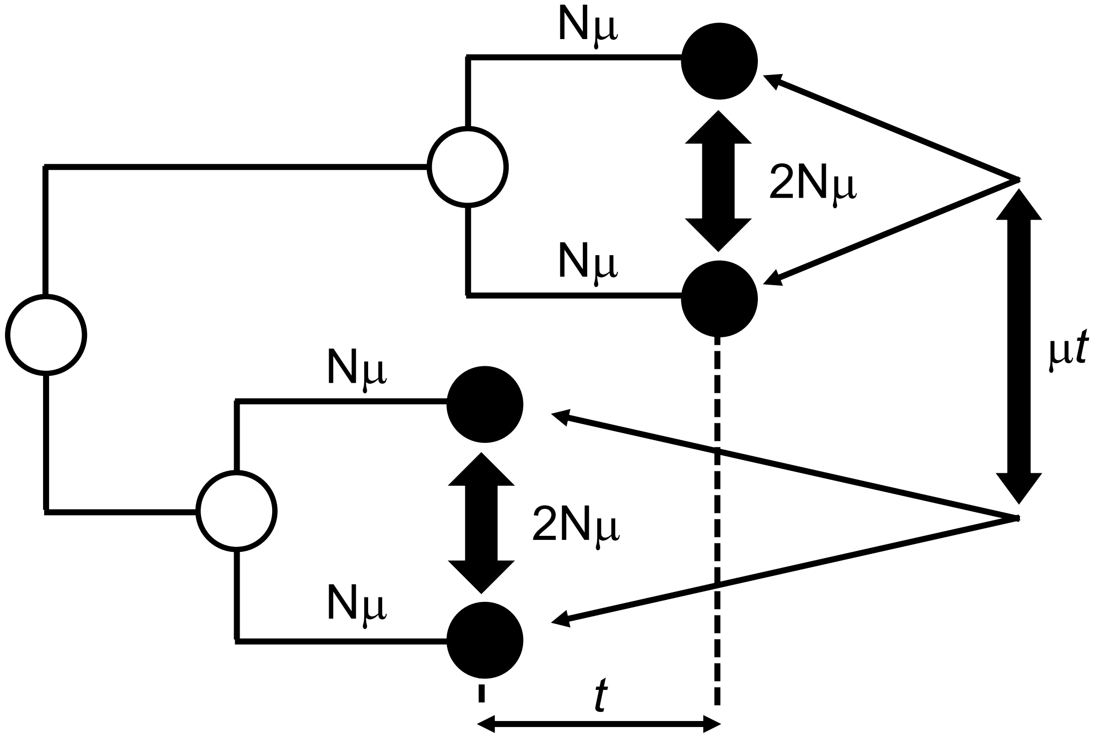


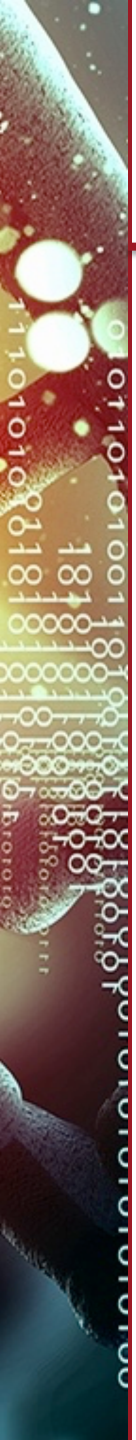
# COALESCENT MODELS

## The Serial Coalescent (*s*-coalescent)

- Having a calibration, either from external data or serial samples, allows one to estimate the expected time to the common ancestor (the coalescent event)



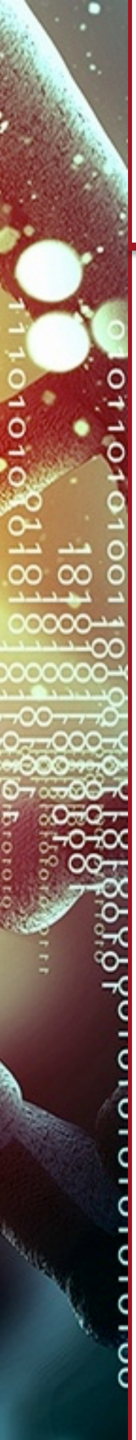




# DEMOGRAPHY AND PHYLOGENY

## Skyline and Skyride plots

- For coalescent interval lengths measured in substitutions per site $\gamma_i$ (where *i* is the number of lineages in the interval) the harmonic mean of the effective population size over the interval, $H_i$, is proportional to $\gamma_i \binom{i}{2} / \mu$
- A plot of $\gamma_i \binom{i}{2}$ over time is a skyline plot
- Scaling the magnitude of a skyline plot by substitution rate $\mu$ gives an estimate of $H_i$



# DEMOGRAPHY AND PHYLOGENY

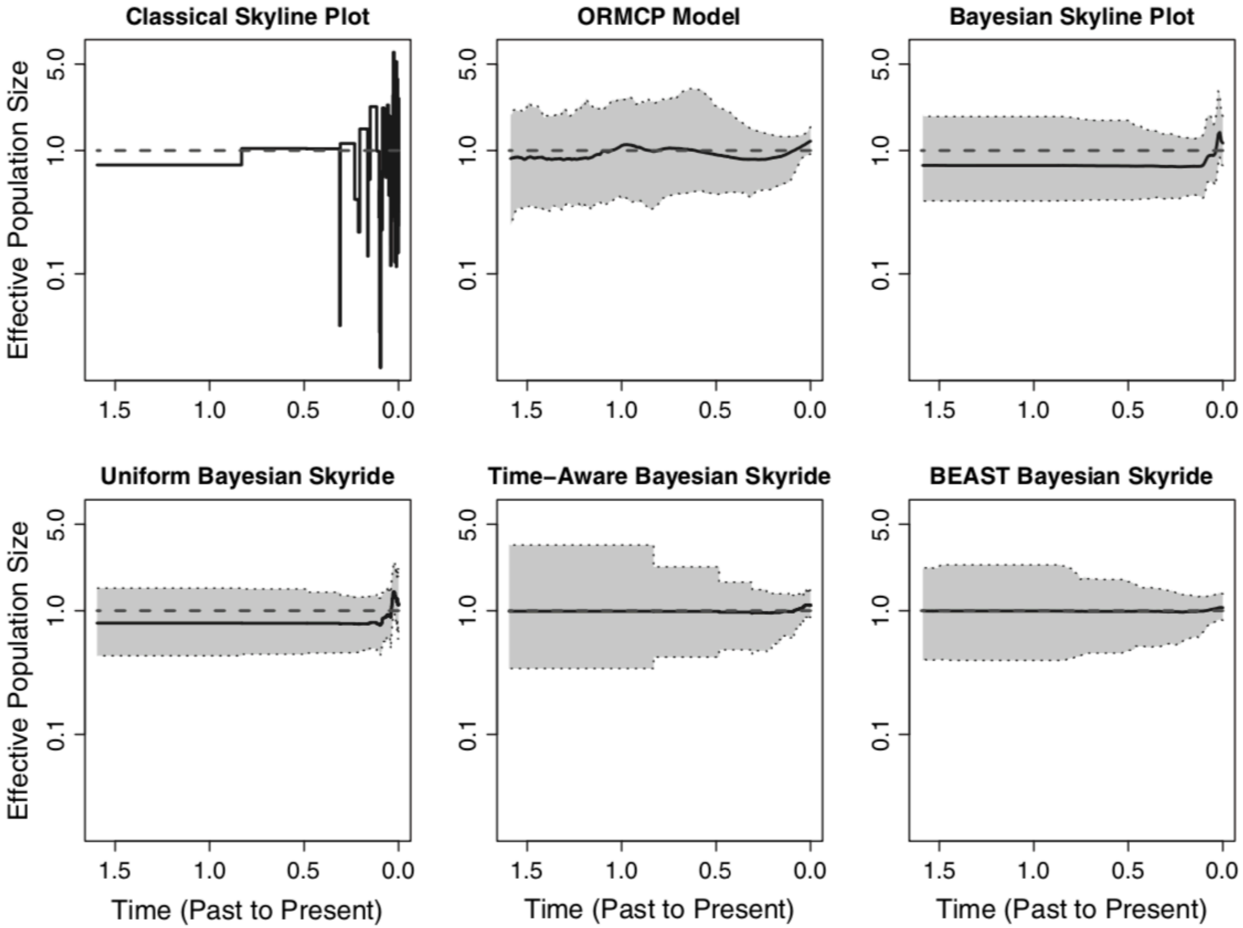
## Skyline and Skyride plots

- Implementations of skyline plots have used various methods to correct for extreme variance in estimates due to overfitting.
- These methods typically require the user specify the number of skyline intervals.
- Minin, Bloomquist, and Suchard (2008) published a method based on continuous changes of effective population size over time which they called the Skyride.
- The Skyride uses a Gaussian Markov random field smoothing function to reduce jumps in $N_e$ estimates and eliminate need for specification of population priors

# DEMOGRAPHY AND PHYLOGENY

## Skyline and Skyride plots

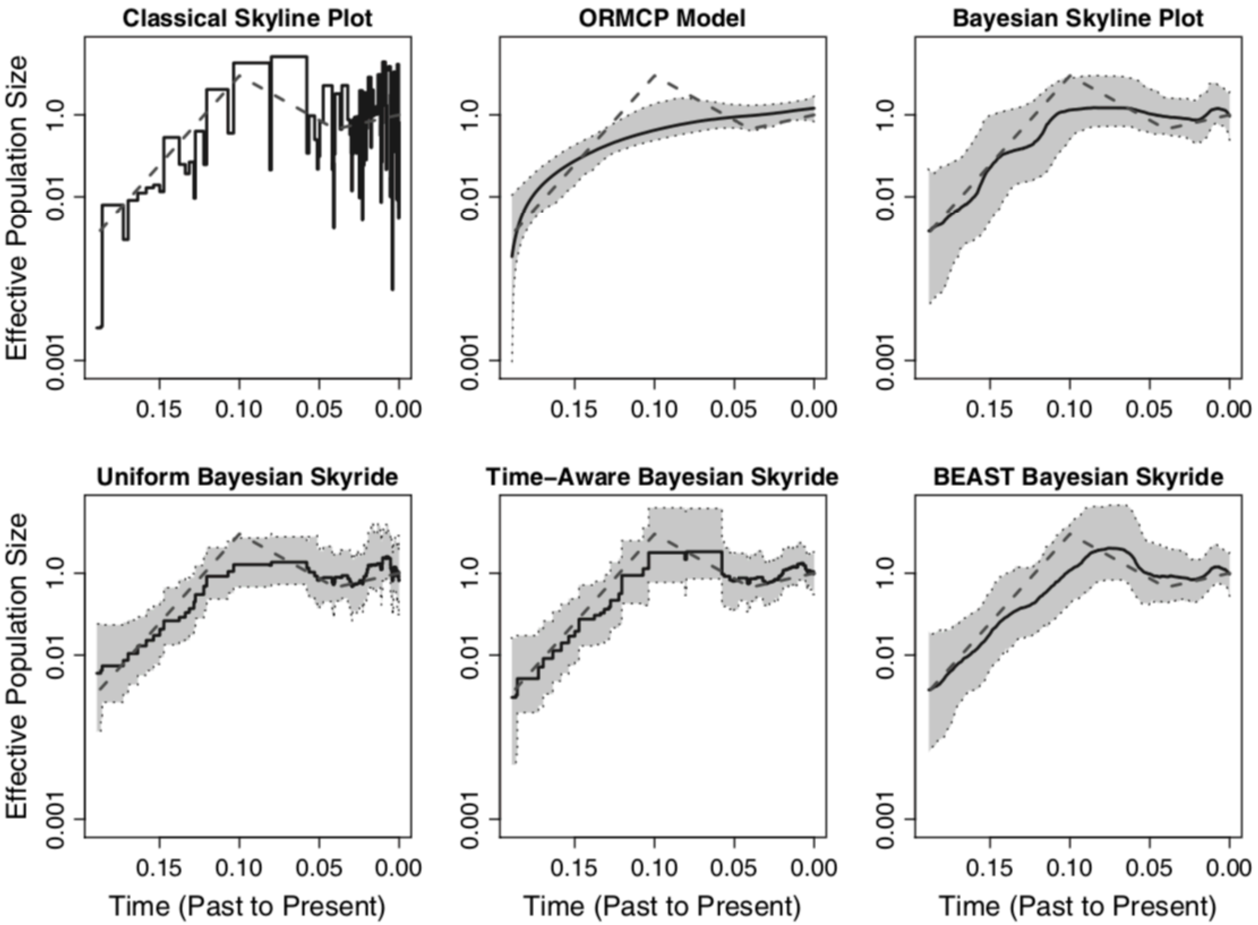
Minin, et al. 2008
Figure 2
Simulations for a constant population size.



# DEMOGRAPHY AND PHYLOGENY

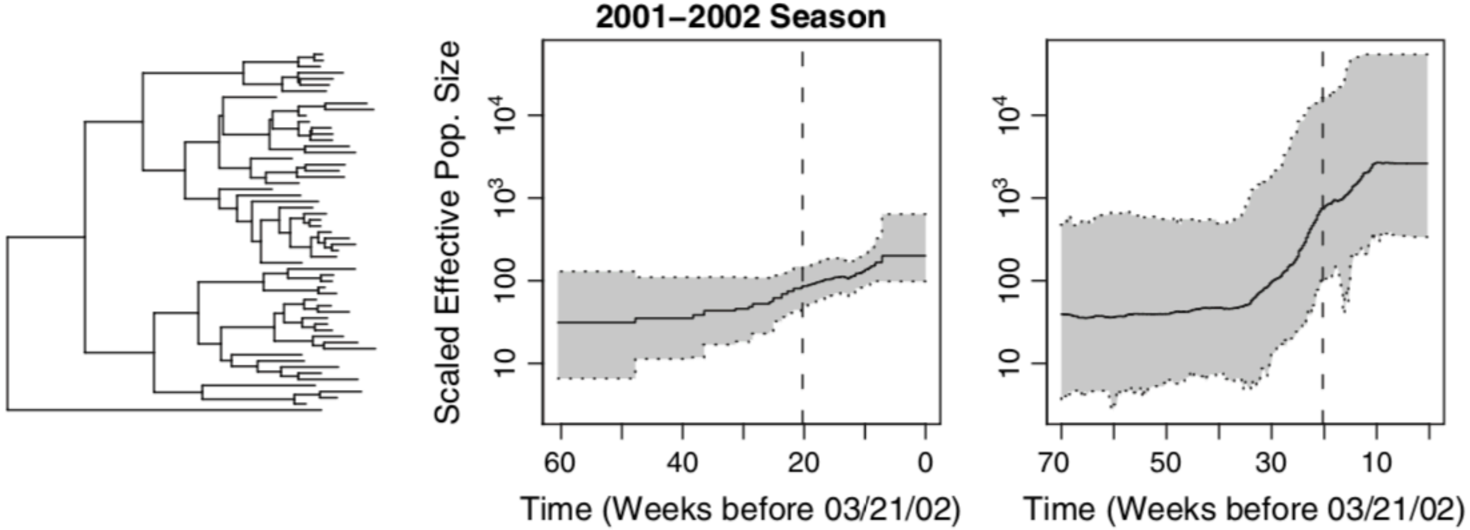
## Skyline and Skyride plots

Minin, et al. 2008
Figure 4
Simulations for a population with a bottleneck.



# DEMOGRAPHY AND PHYLOGENY

## Skyline and Skyride plots



Minin, et al. 2008 Figure 6. Intraseason dynamics of human influenza. Inferred genealogy, fixed-tree time-aware skyride, and BEAST skyride

# COALESCENT THEORY

Thank you