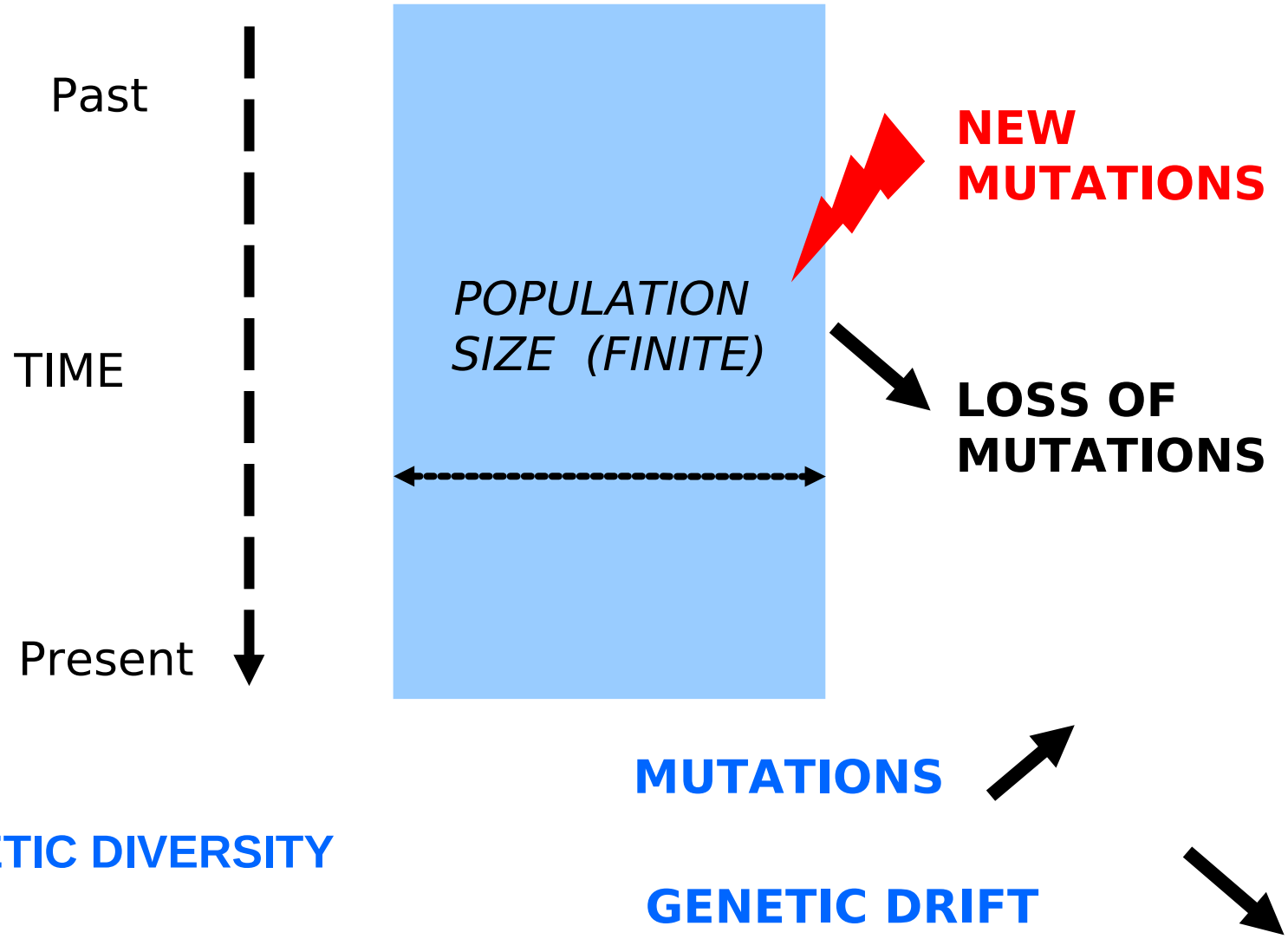
The basics of inference in population genetics: Wright Fisher and the coalescent

Oeiras, 13-17 May 2019

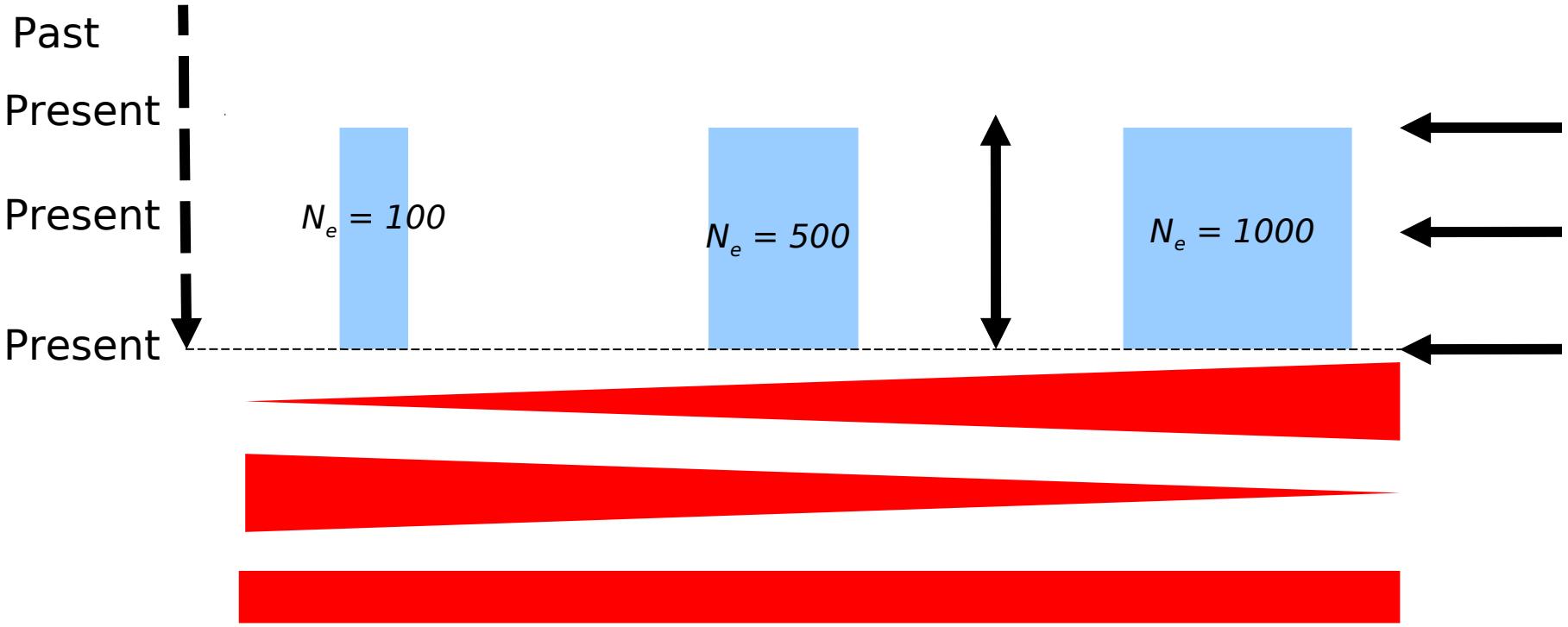
Mark Beaumont[1] / Lounès Chikhi[2,3]
Vitor Sousa [4] / Willy Rodriguez[5]/ Armando Arredondo[5]

[1] Department of Mathematics, Bristol University, UK
[2] Evolution et Diversité Biologique, CNRS, Toulouse, France
[3] Instituto Gulbenkian de Ciência, Oeiras, Portugal
[4] Faculdade da Universidade de Lisboal
[5] Institut de Mathématiques de Toulouse, Toulouse, France

GENETIC DIVERSITY



GENETIC DIVERSITY



GENETIC DIVERSITY

Different demographic histories can produce similar or counter-intuitive results

Towards a genealogical perspective on genetic data

- Genes (segments of chromosome) are transmitted from generation to generation by copying.
- In any generation some genes are copied a lot of times, others are never copied.
- A consequence of this is that any pair of genes share a common ancestor in the past.

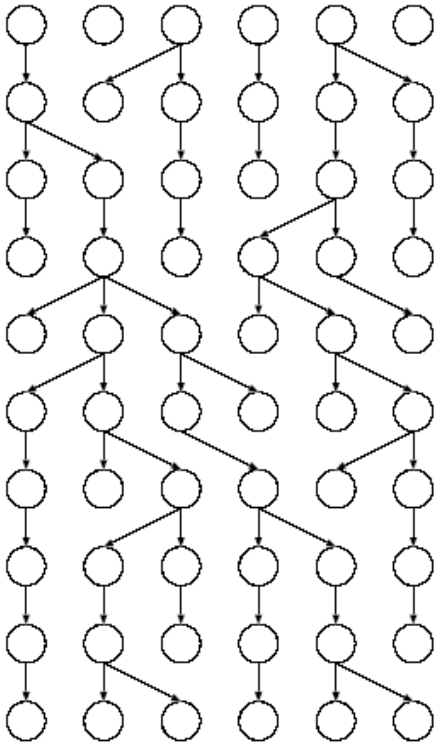
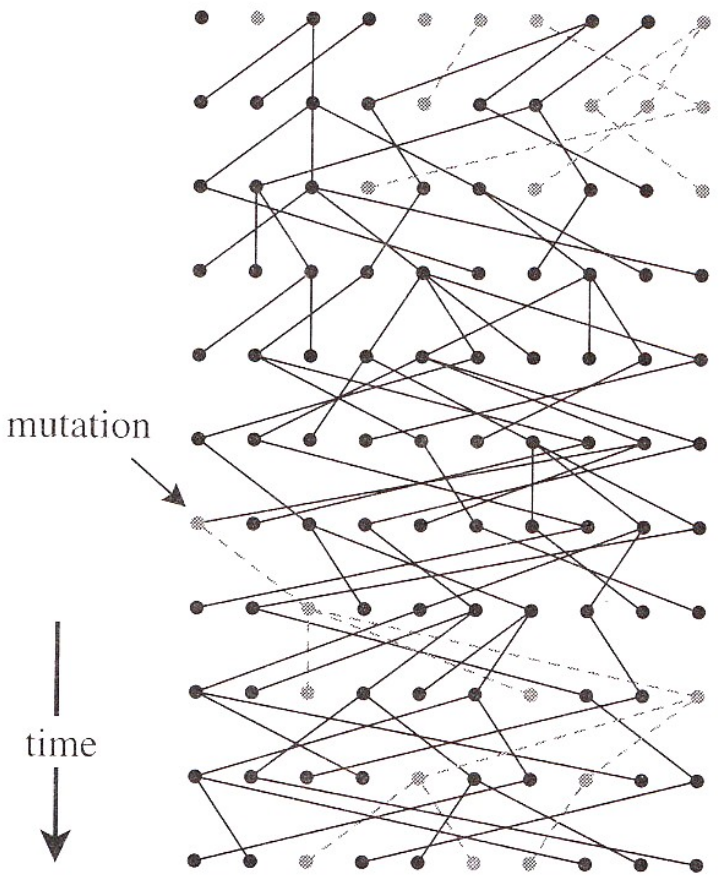


Figure 1: Forward process in the WFM.

- The copying process is error prone – mutations.
- Genes that share a recent common ancestor are unlikely to have a mutation in their ancestry.
- Genes with a distant common ancestor are more likely to have a mutation in their ancestry.
- Demography affects the copying process: in small populations genes have a greater chance of being lost than copied.



Wright-Fisher Model
for N = 6

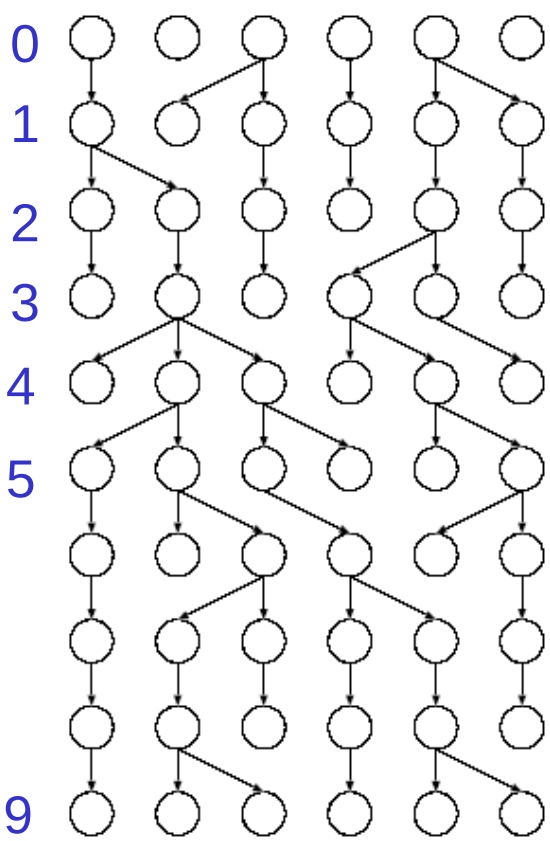


Figure 1: Forward process in the WFM.

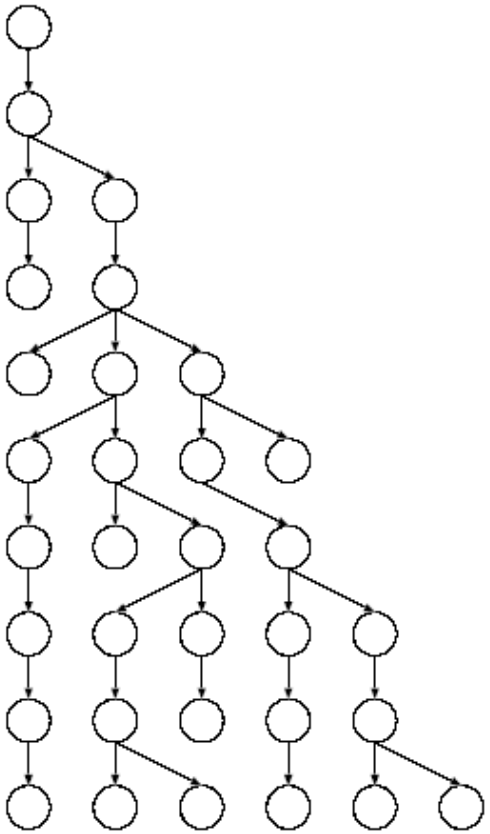


Figure 2: Pruned forward process in the WFM.

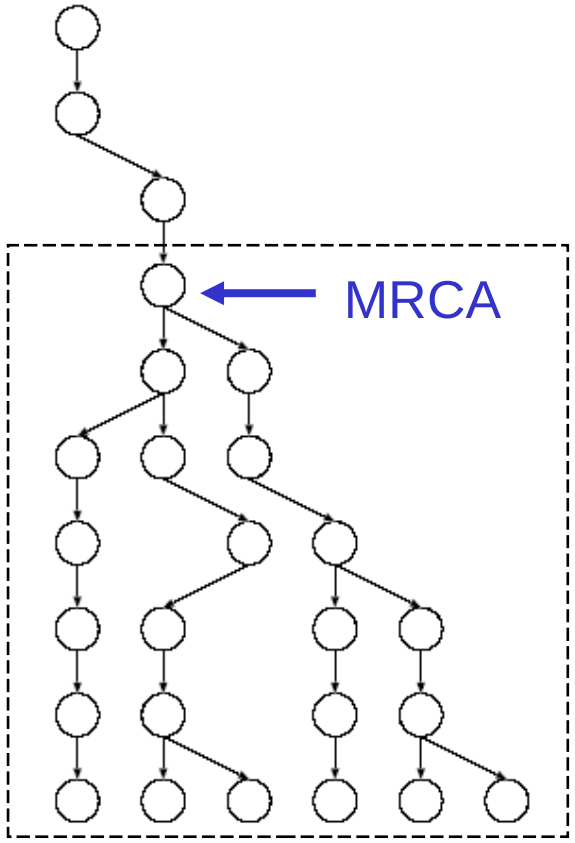


Figure 3: Relatives of alleles present in the final generation.

MRCA = Most Recent Common Ancestor

Wright-Fisher Model
for N = 6

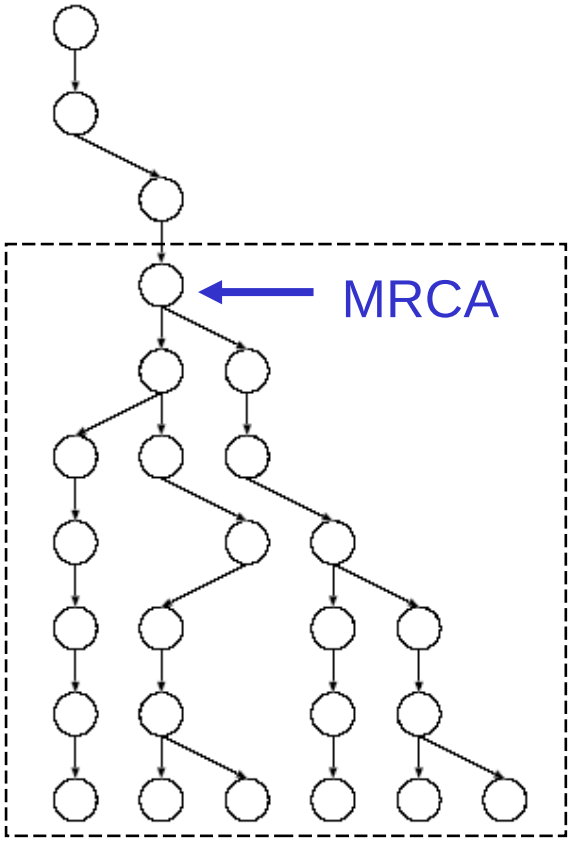


Figure 3: Relatives of alleles present in the final generation.

VISUAL BREAK !

http://willyrv.com/WFsimulation

You can change the number of genes and generations.
Not visually optimized:
use only n < 10 and T < 10

You can also use the R script:
plot_ms_trees_mig.R
(after R introduction)

MRCA = Most Recent Common Ancestor

Case of 2 gene copies (sample size of 2)

1 Probability of coalescence <u>1 generation</u> ago for <u>2 gene copies</u>:

We assume a population of N diploid individuals

Thus we have 2N haploid genomes

If we pick 2 genes at random:

- the probability that they are descended from the same copy is $1/2N$
- the probability that they come from different copies is $1-(1/2N)$.

Why is the first probability equal to 1/2N ?

Case of 2 gene copies (sample size of 2)

1 Probability of coalescence <u>1 generation</u> ago for <u>2 gene copies</u>:

We assume a population of N diploid individuals

Thus we have 2N haploid genomes

If we pick 2 genes at random:

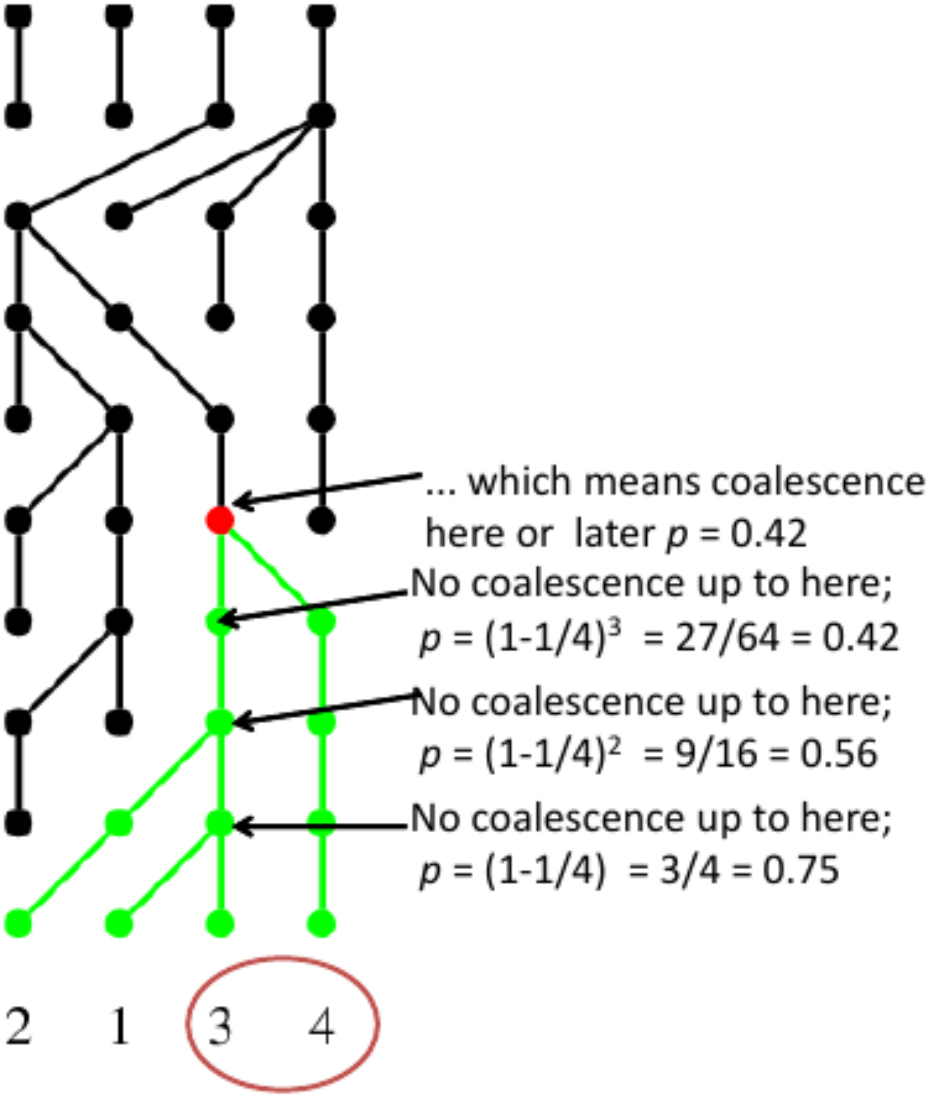
- the probability that they are descended from the same copy is $1/2N$
- the probability that they come from different copies is $1-(1/2N)$.

Why is the first probability equal to 1/2N ?
To understand this we consider that each gene 'chooses' its parent at random in the previous generation.

Take a pair of genes and let one choose its parent; then the probability the other chooses the same parent is 1/(2N).

Another way to look at it is to arrange all (2N)2 possible pairs of parents chosen by the two genes on a grid (matrix), of which 2N (those on the diagonal) will be identical, giving (2N)/(2N)2 = 1/(2N)

Case of 2 gene copies (sample size of 2)



Example:
2N = 4
s=2

Case of 2 gene copies (sample size of 2)

2 Probability of coalescence GREATER THAN T generations ago for 2 gene copies:

(We make the assumption that N is large and t is large)

It is the probability that they **do not coalesce** for *T* generations:

$$P(\text{coalescence time} > T) = [1-(1/2N)]^T$$

For large N, 1/2N is small and we can approximate:

$$[1-(1/2N)]^T \approx e^{-T/2N}$$

If we rescale time in units of 2N (continuous time): t= T/2N we have :

$$P(\text{coalescence time} > t) \approx e^{-t}$$

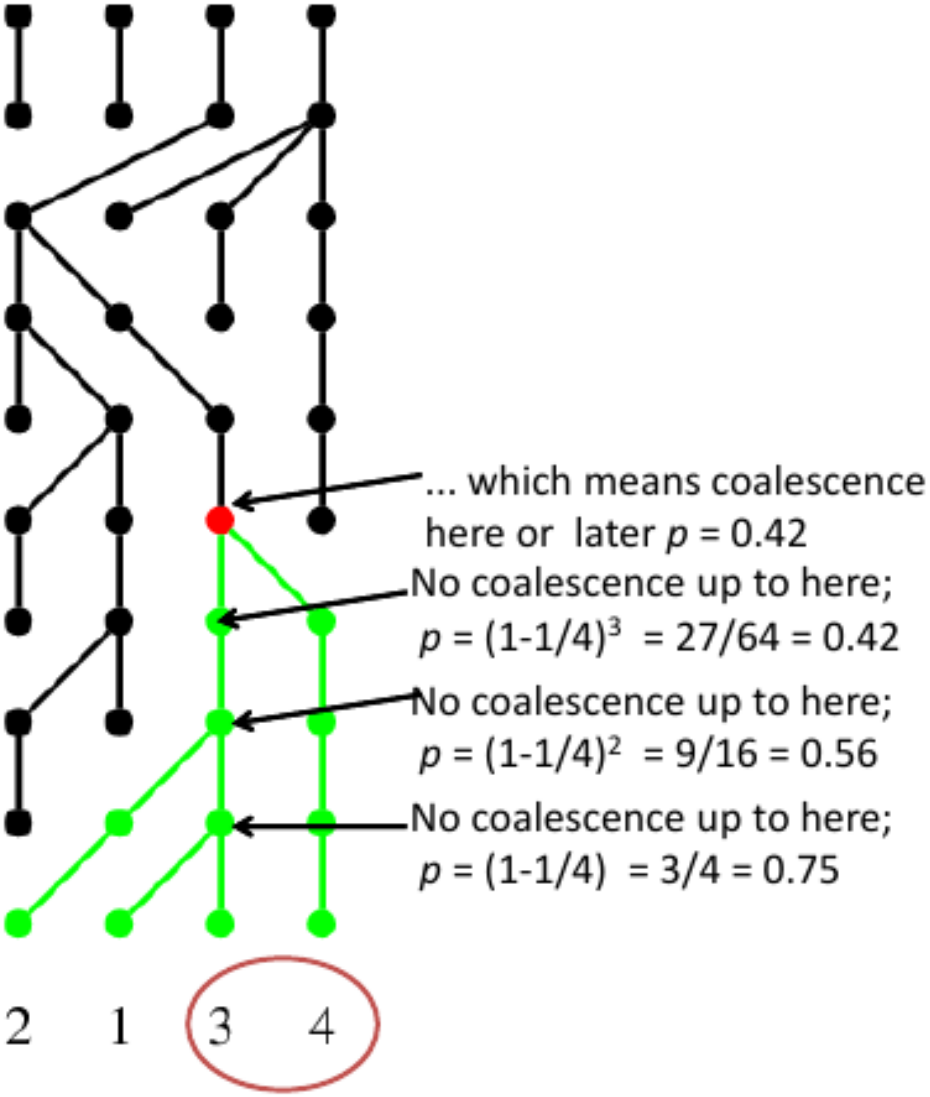
In other words: coalescence time follows an exponential distribution, defined as:

$f(t)=\lambda e^{-\lambda t}$ - where λ and $t > 0$

with an expectation of $1/\lambda$ and a variance of $1/\lambda^2$.

With coalescence rescaling we have $\lambda = 1$

Case of 2 gene copies (sample size of 2)



Case of k gene copies (sample size of k)

1.Probability that <u>any two</u> of the k sequences coalesce <u>1 generation ago</u> is

$(1/2N)\times k(k-1)/2 = k(k-1)/4N$, (ignoring multiple coalescence)

indeed, there are k(k-1)/2 possible pairs of sequences in k sequences.

Note that for k=2 you find the result from above.

2. Probability that <u>any two</u> of the k sequences coalesce <u>t+1 generations</u> ago is:

$$k(k-1)/4N\ e^{-tk(k-1)/4N}$$

Thus, following the same rational as above, we can show that the time T_k until the first coalescence follows an exponential distribution where $\lambda=k(k-1)/4N$.

Thus, the expectation of this time is

$E(T_k) = 4N/k(k-1)$.

Note that these results require k(k-1)/4N to be <<1 which is true if k<<N. This means that we assume that the sample size is small compared to the population size. So in theory the coalescent should not work for small populations (endangered species, bottlenecked pops). In practice, simulations show that the coalescent actually works very well even in such cases.

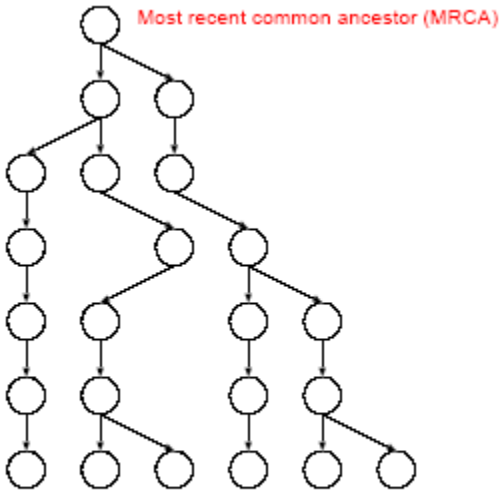


Figure 4: Ancestral tree of alleles present in the final generation.

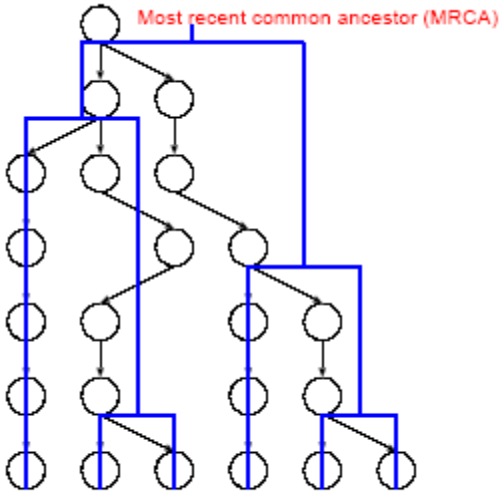
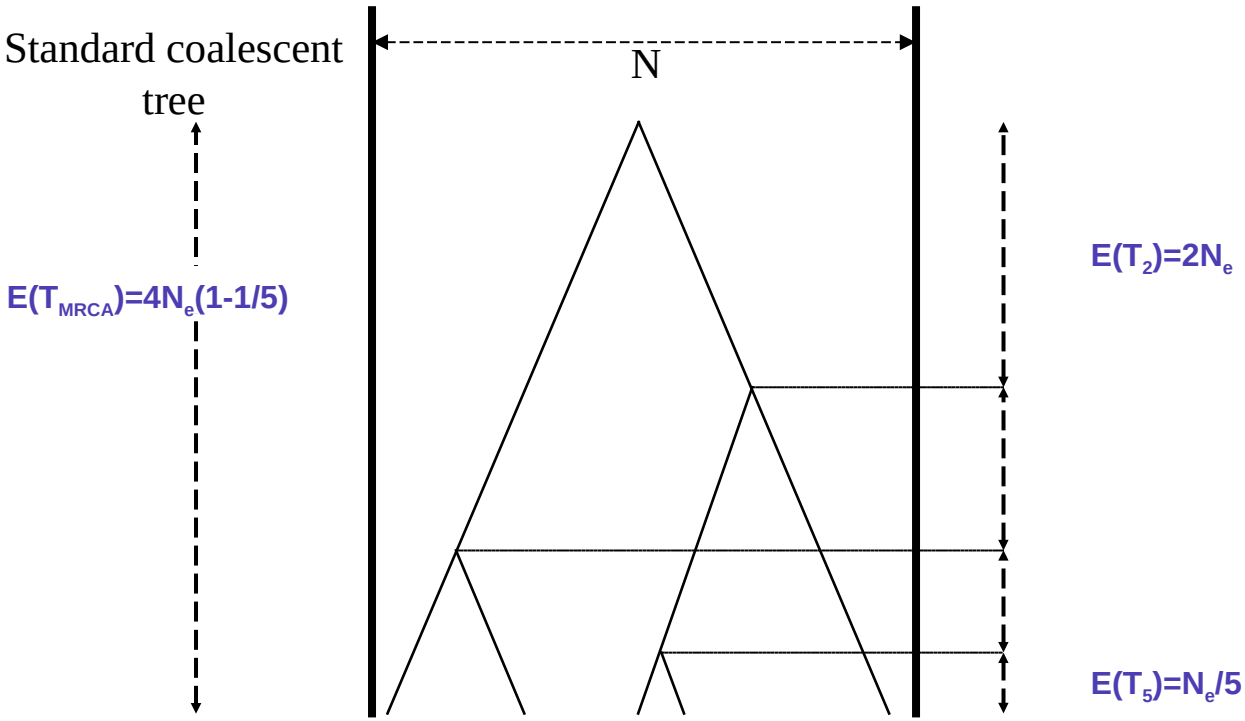
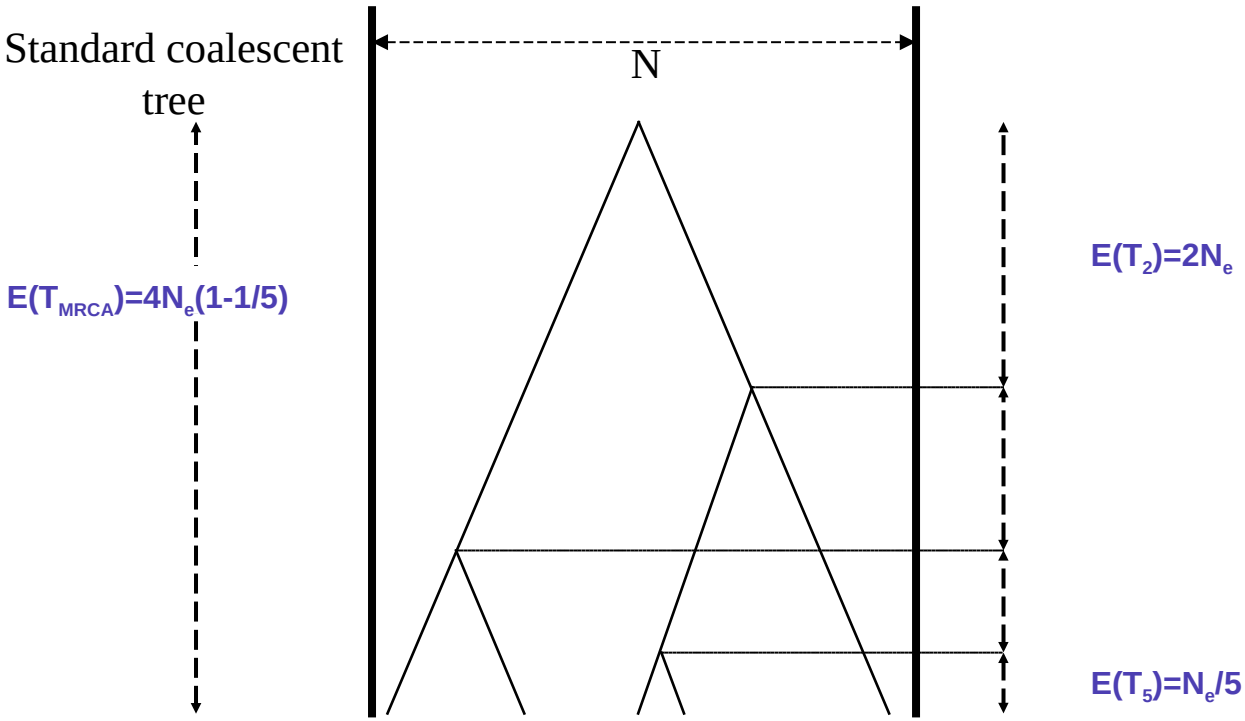


Figure 5: The genealogy of the population at time $t = 9$.

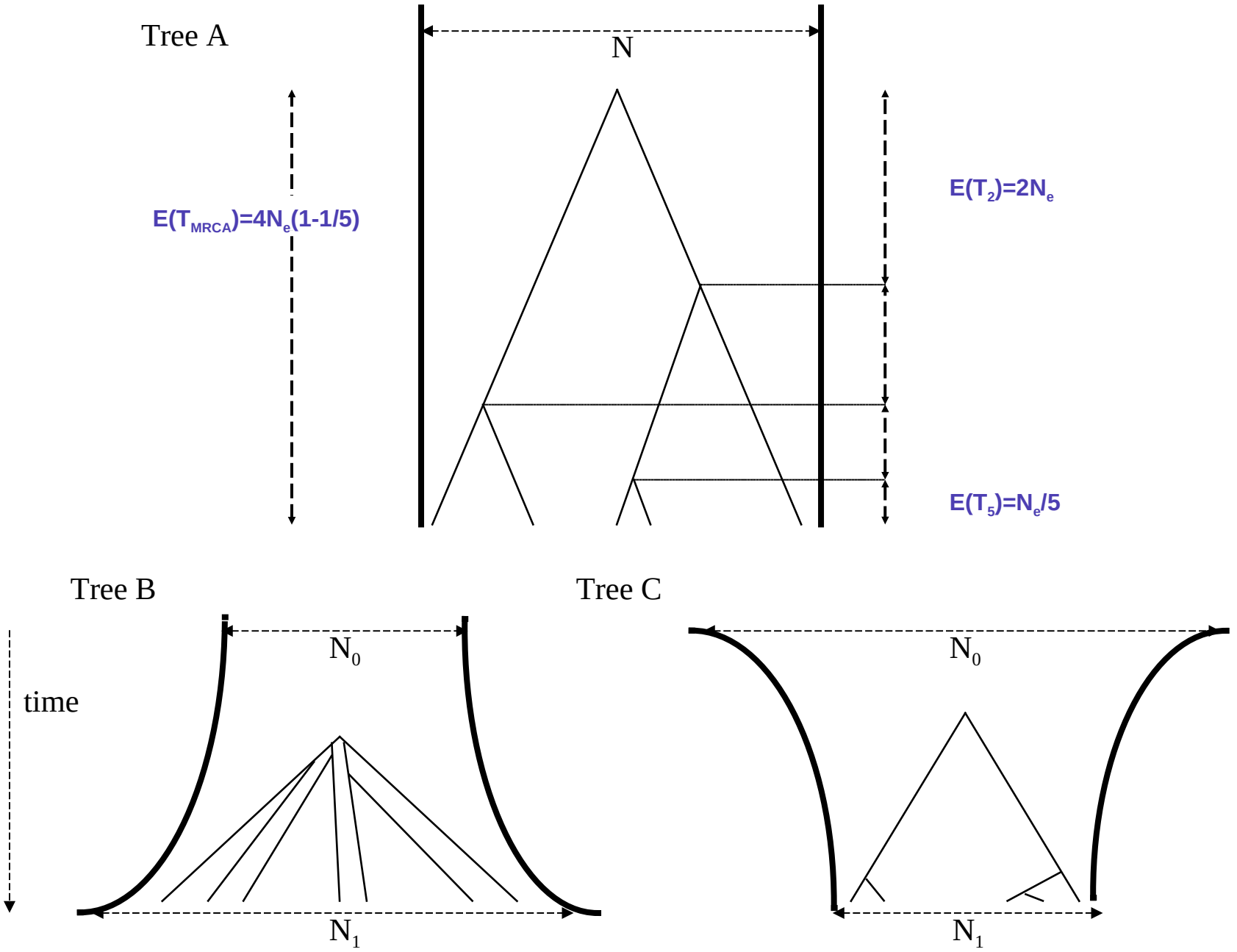
- Coalescent = probabilistic model
- Gives the distribution coalescence times (Wright-Fisher, Moran).
- Exponential law (T_k = time during which there are k lineages):
  - $E(T_k) = 4N/k(k-1)$
- $E(T_{MRCA}) = E(T_k) + E(T_{k-1}) + \ldots + E(T_2) = 4N(1- 1/k) \sim 4N$
- $E(T_2) = 4N(1- 1/2) = 2N$



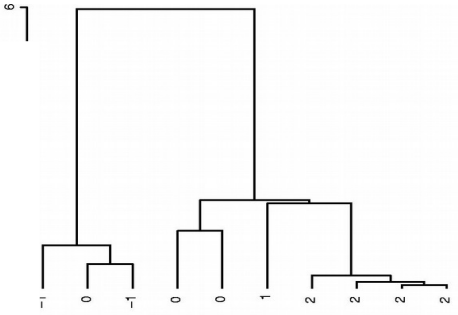
- $E(T_k) = 4N/k(k-1)$
- $E(T_{MRCA}) \sim 4N$
- $E(T_2) = 4N(1- 1/2) = 2N$

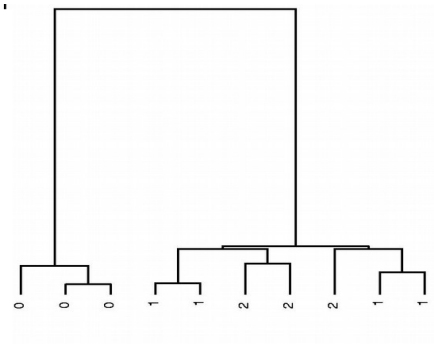


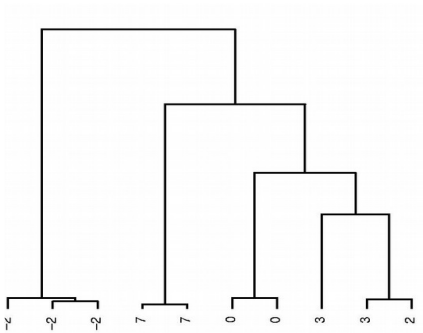
- $E(T_k) = 4N/k(k-1)$
- WHAT DO YOU EXPECT TO HAPPEN IF POPULATION SIZE CHANGES?
- POPULATION GROWTH / CRASH ?

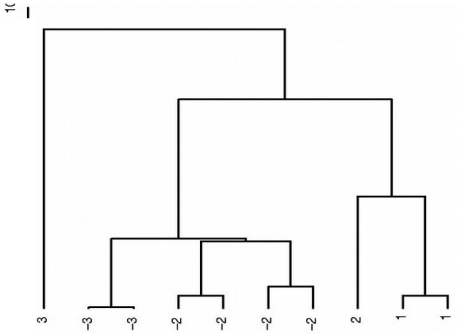

Tree A
N
$E(T_{MRCA})=4N_e(1-1/5)$
$E(T_2)=2N_e$
$E(T_5)=N_e/5$
Tree B
time
N_0
N_1
Tree C
N_0
N_1

Stable

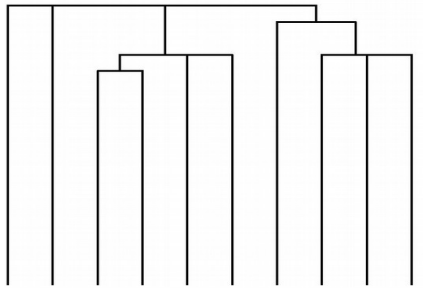




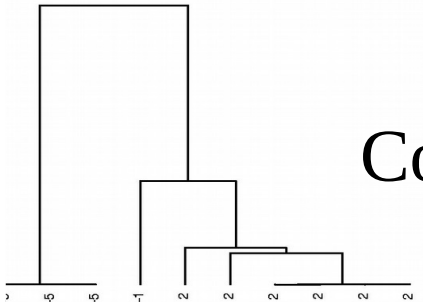


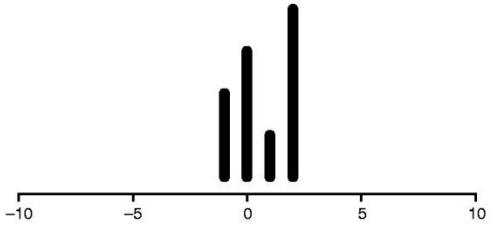


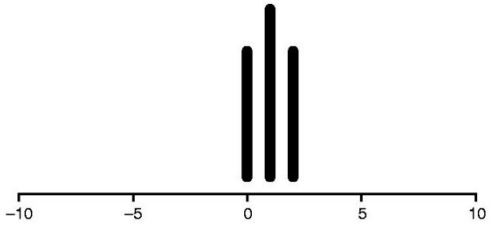
Growing



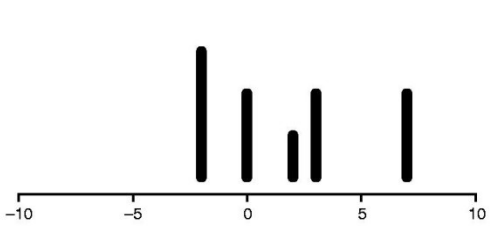
Contracting

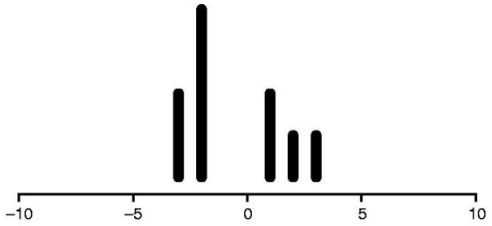




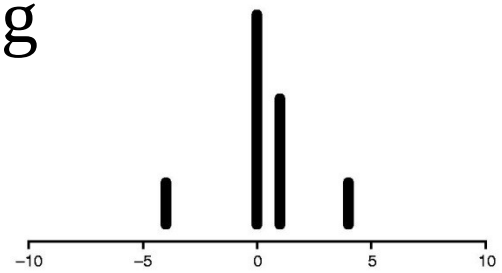


Stable

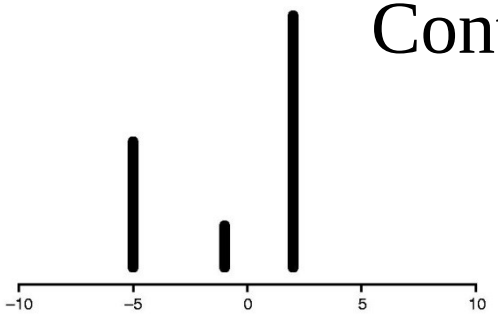


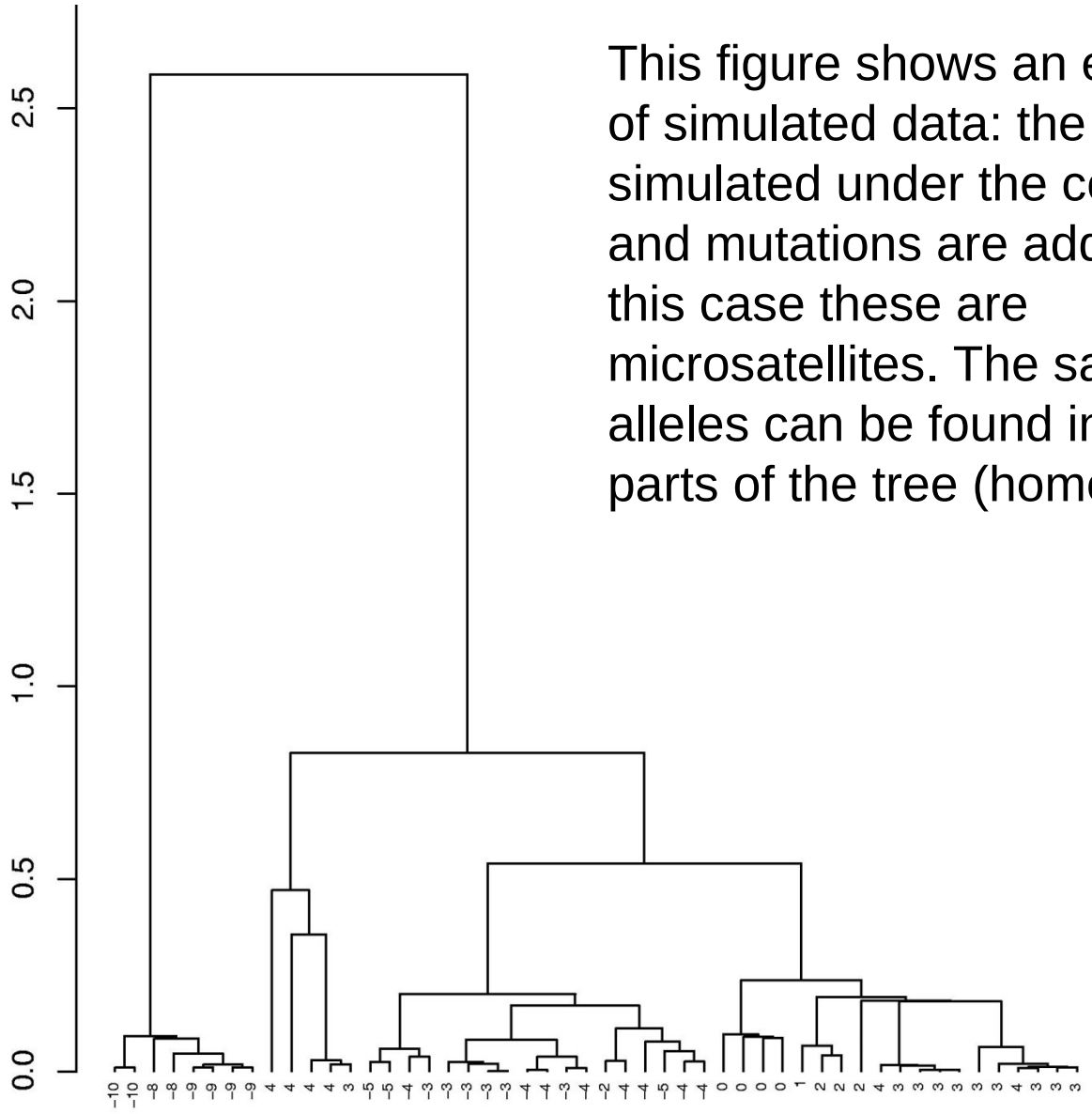


Growing



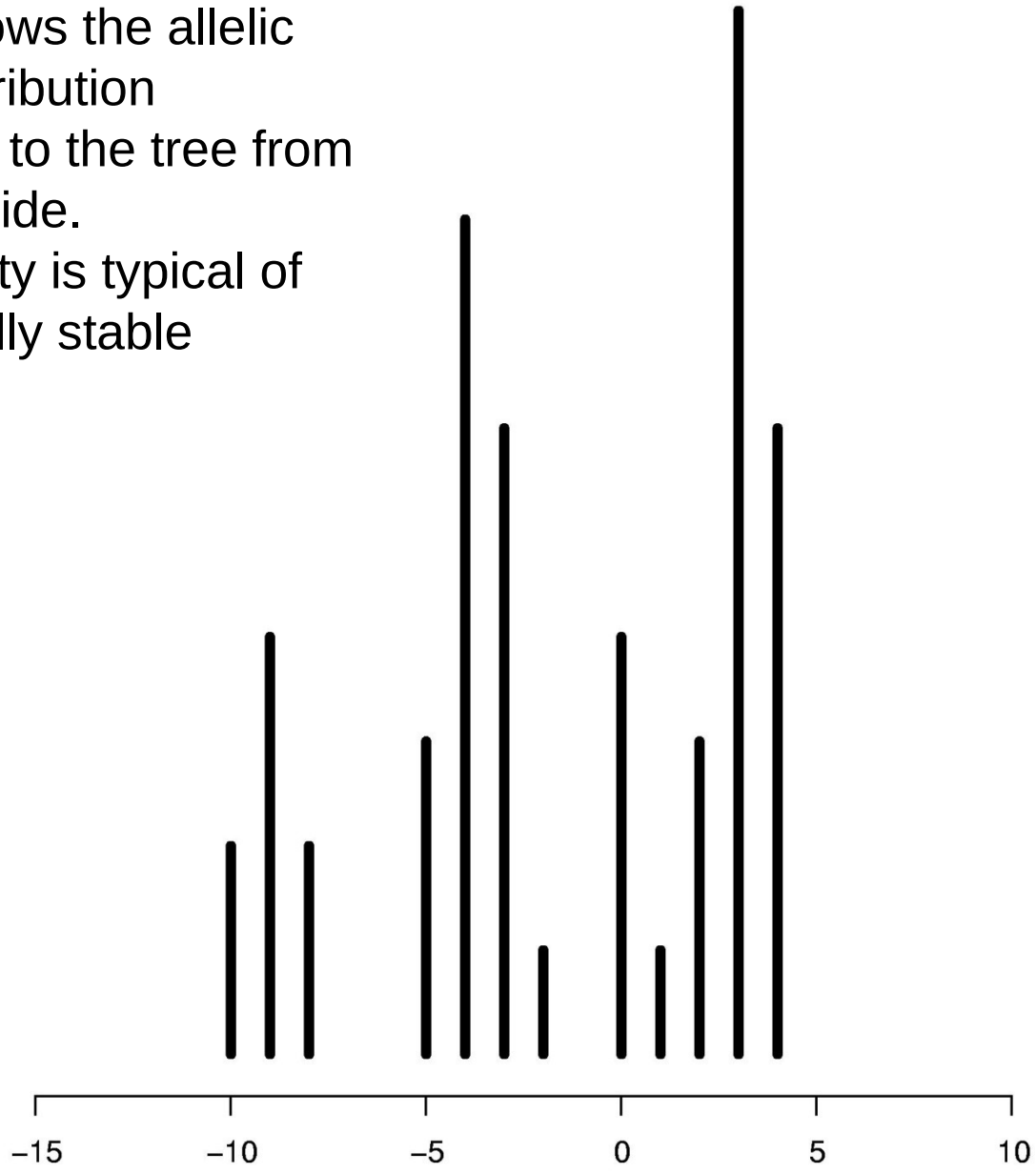
Contracting





This figure shows an example of simulated data: the tree is simulated under the coalescent and mutations are added. In this case these are microsatellites. The same alleles can be found in different parts of the tree (homoplasy).

This figure shows the allelic frequency distribution corresponding to the tree from the previous slide.
Its multimodality is typical of demographically stable populations.



NEXT STEPS

- Simulations with ms and SPAms
- Introduction to R
- Coalescent simulations and tree visualization with R

ANY QUESTIONS ?