ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

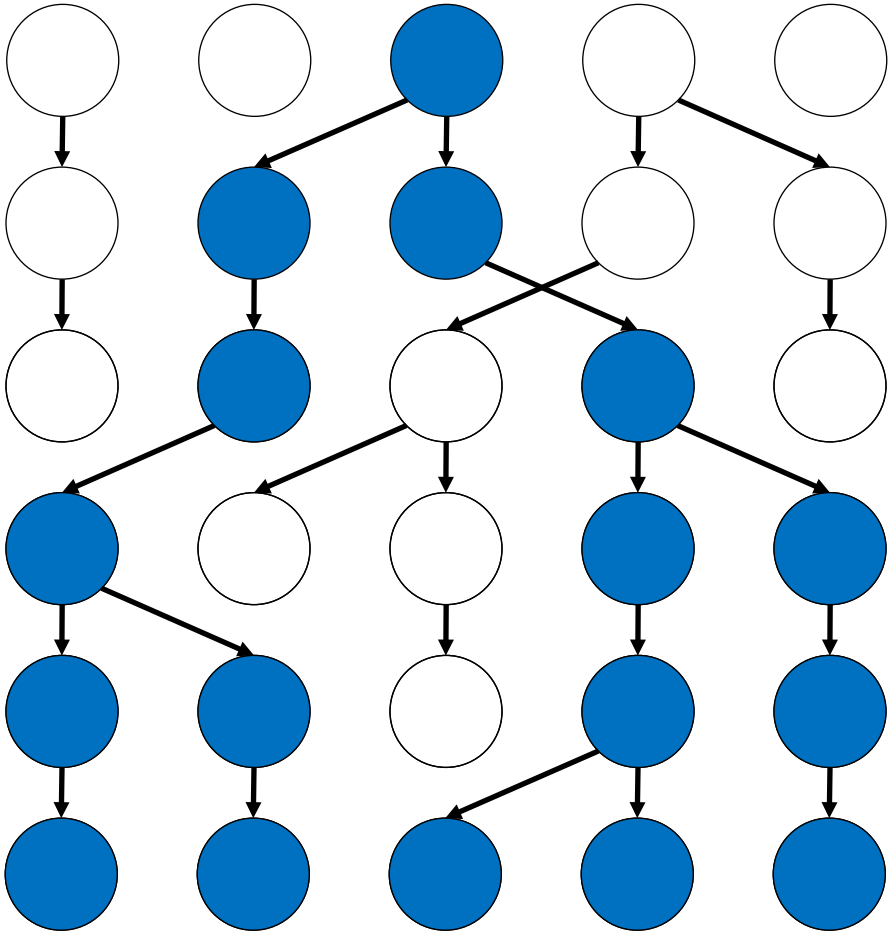
Coalescent theory

Niko Beerenwinkel

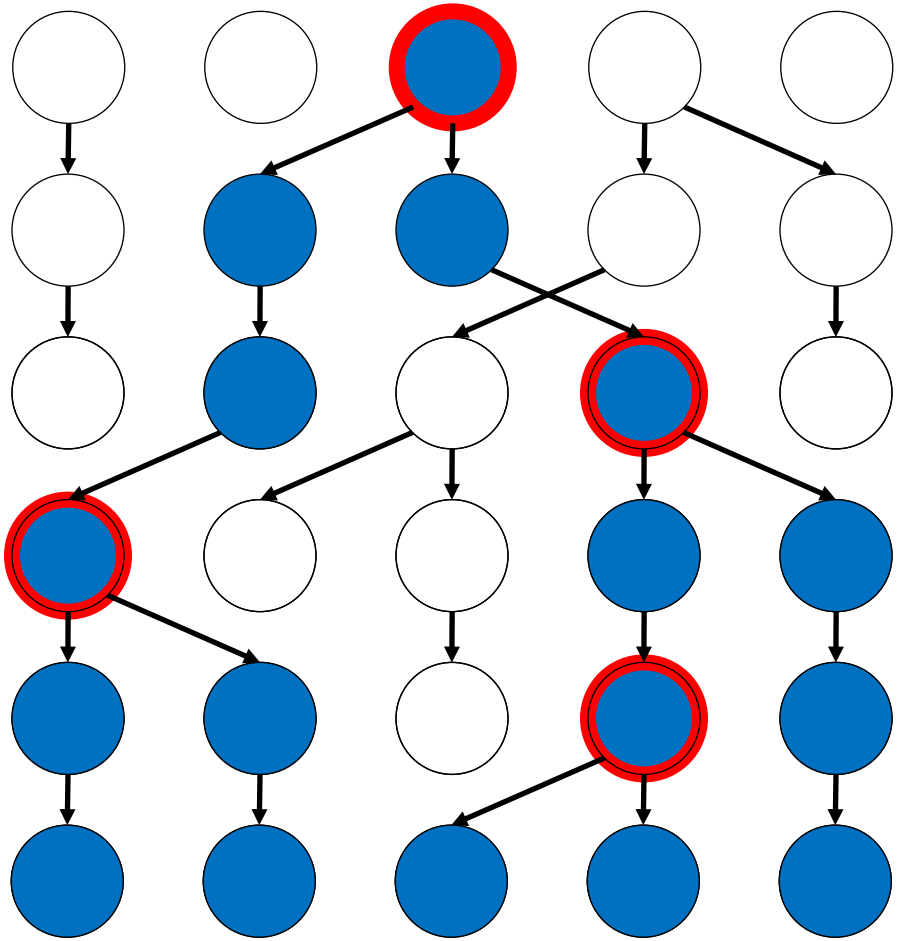
Outline

- The coalescent
- Coalescence time
- Detecting selection

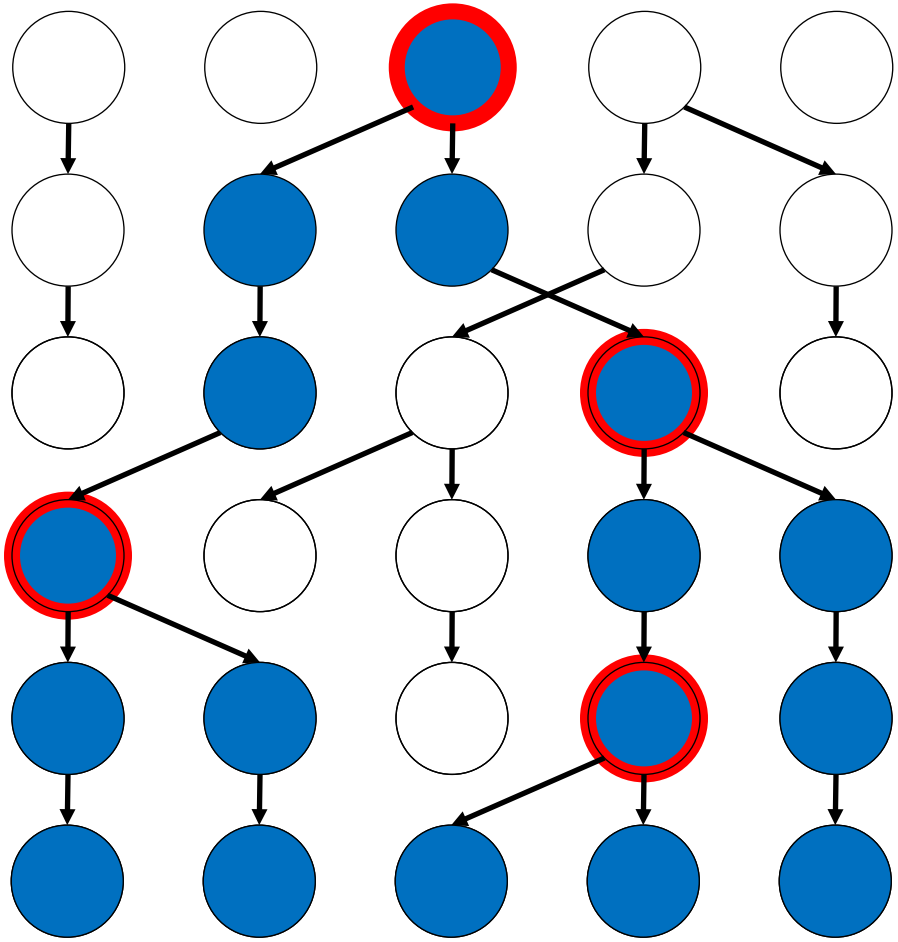
The Wright-Fisher process

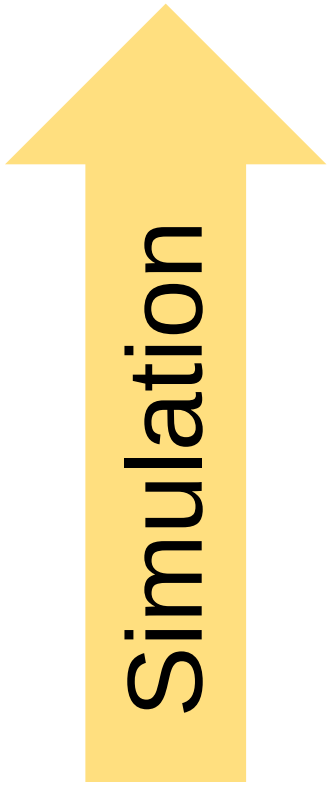


The Wright-Fisher process

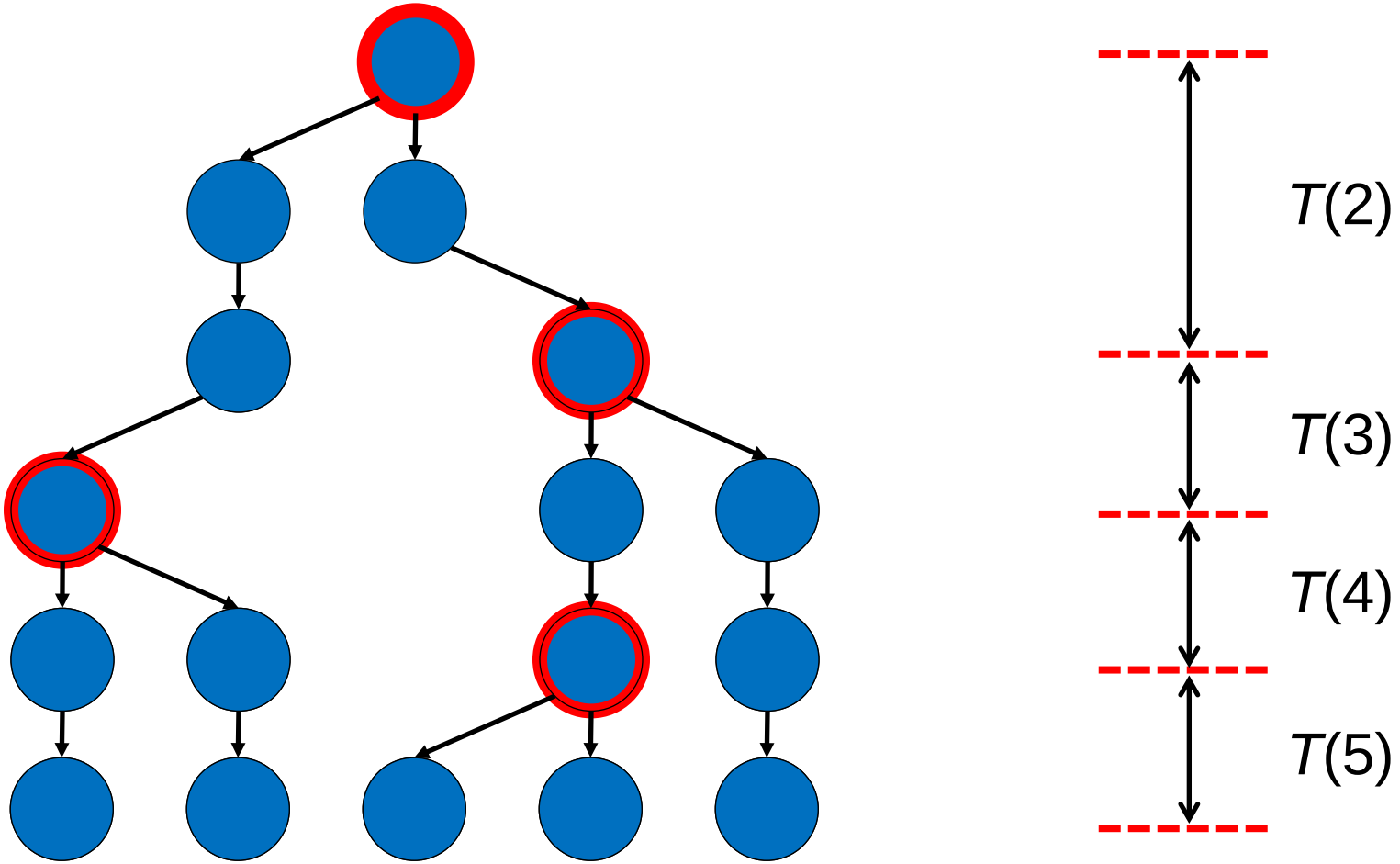


Coalescent events

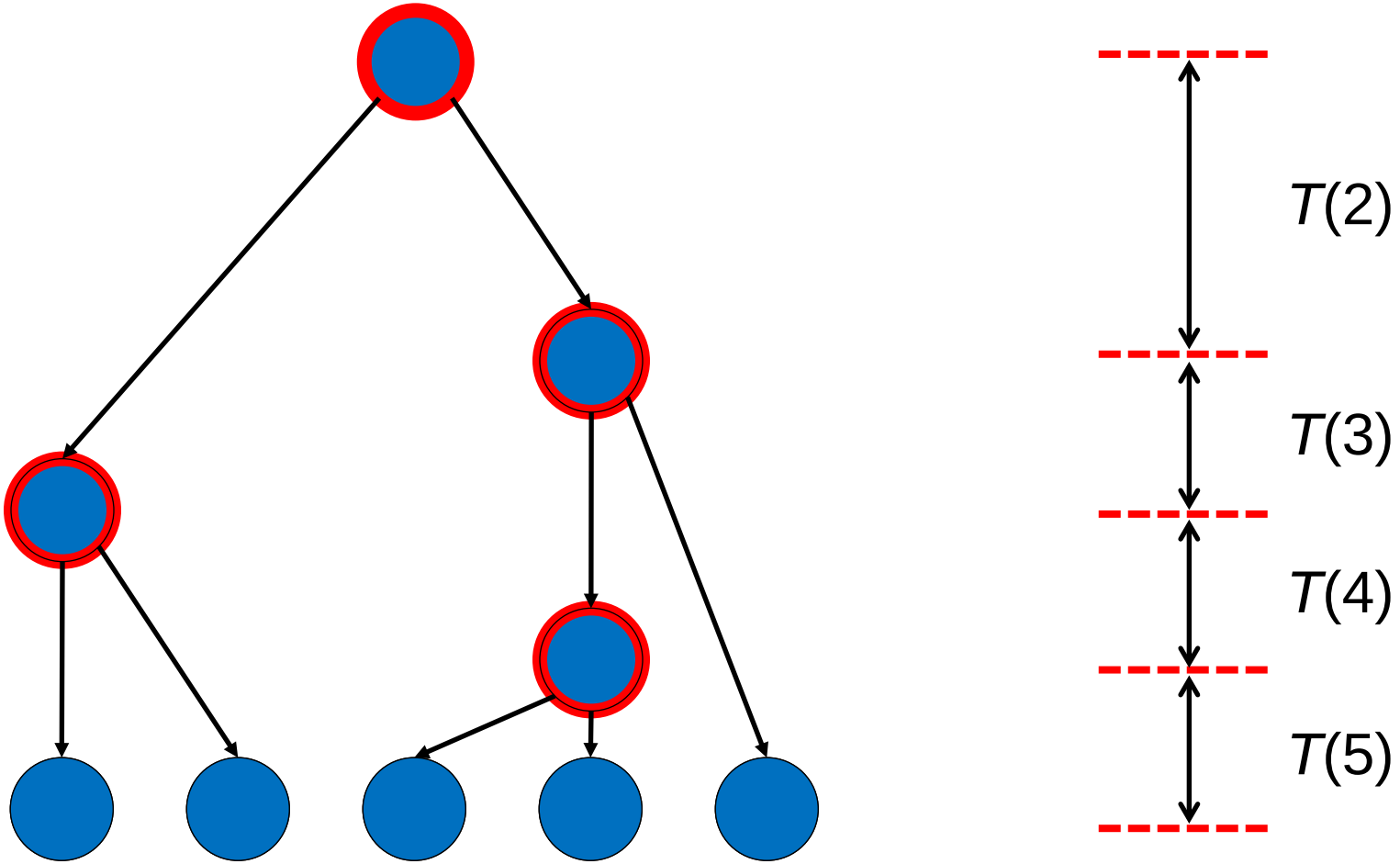




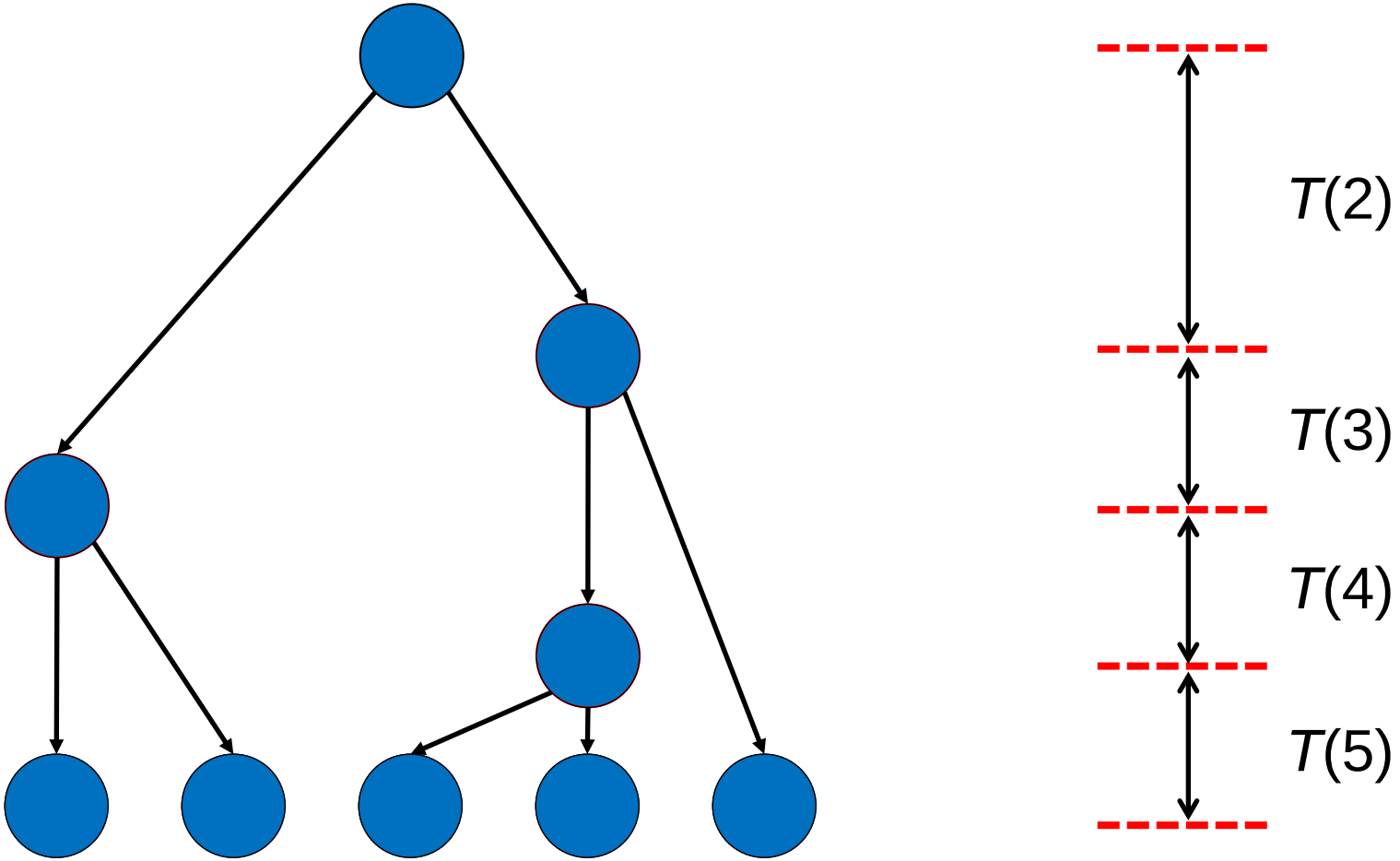
Coalescent times



Coalescent times

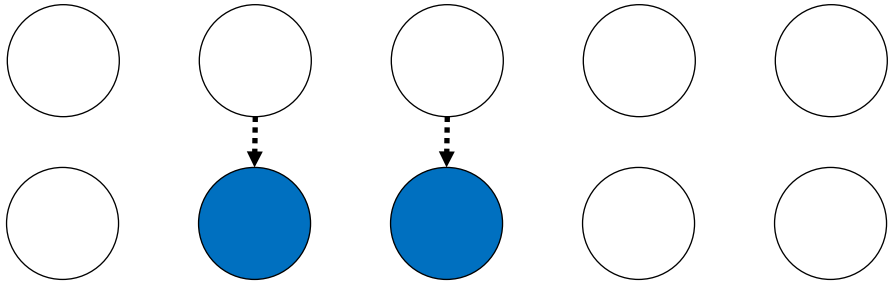


Coalescent times



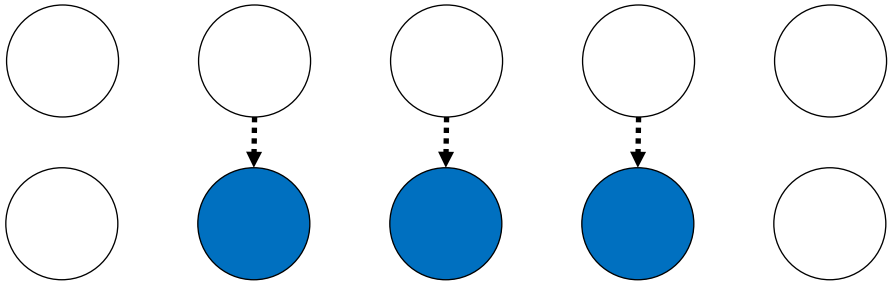
The probability that *j* genes have no common ancestor in the previous generation

$j = 2$



$$1 - \frac{1}{N}$$

$j = 3$



$$\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)$$

$$\prod_{i=1}^{j-1}\left(1 - \frac{i}{N}\right) = 1 - \binom{j}{2} N^{-1} + O(N^{-2})$$

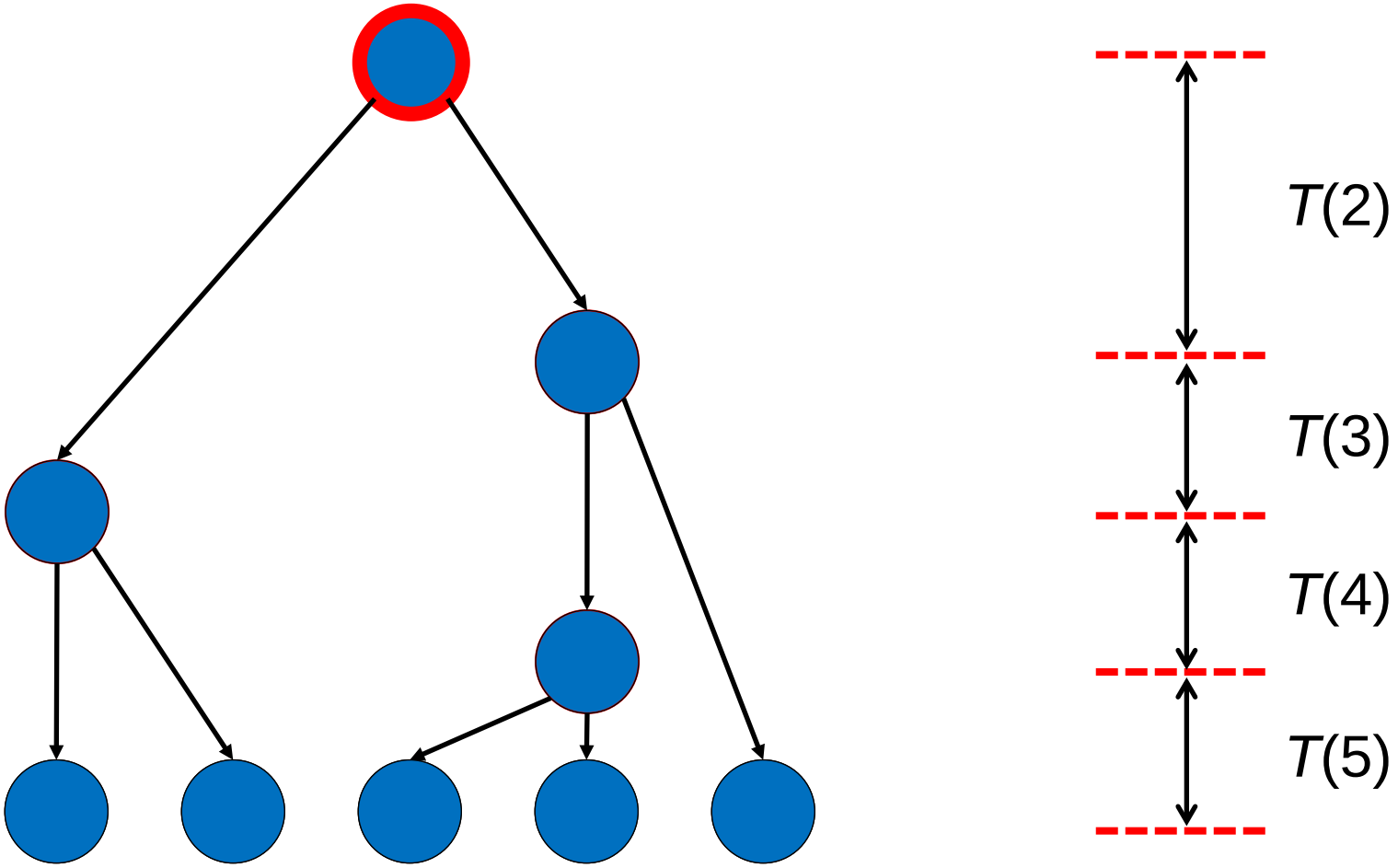
The coalescent

- We measure time in units of *N* generations.
- Let T(*j*) be the coalescence time between *j* and *j* − 1 genes:

$$P(T(j) > t) = \left[\prod_{i=1}^{j-1}\left(1 - \frac{i}{N}\right)\right]^{Nt} \longrightarrow \exp\left[-\binom{j}{2}t\right]$$

  as $N \to \infty$.
- Thus, in the diffusion limit, only pairwise coalescence events occur.
- The coalescence time is distributed exponentially with parameter (*j* choose 2) = [*j*(*j* − 1)]/2.
- This stochastic process is called the coalescent.

Most recent common ancestor (MRCA)



$$T_{\mathrm{MRCA}}(5) = T(2) + T(3) + T(4) + T(5)$$

Time to the MRCA: expectation

- For a sample of size *n*, the time to MRCA is

$$T_{\mathrm{MRCA}}(n) = \sum_{j=2}^{n} T(j)$$

- E[*T*(*j*)] = 1 / (*j* choose 2) = 2 / [*j*(*j* − 1)], hence:

$$\mathrm{E}[T_{\mathrm{MRCA}}(n)] = \sum_{j=2}^{n} \mathrm{E}[T(j)] = \sum_{j=2}^{n} \frac{2}{j(j-1)}$$

$$= 2 \sum_{j=2}^{n} \left(\frac{1}{j-1} - \frac{1}{j} \right) = 2 \left(1 - \frac{1}{n} \right)$$

- Note that $\mathrm{E}[T(2)] = 1$ and $\lim_{n \to \infty} \mathrm{E}[T_{\mathrm{MRCA}}(n)] = 2.$

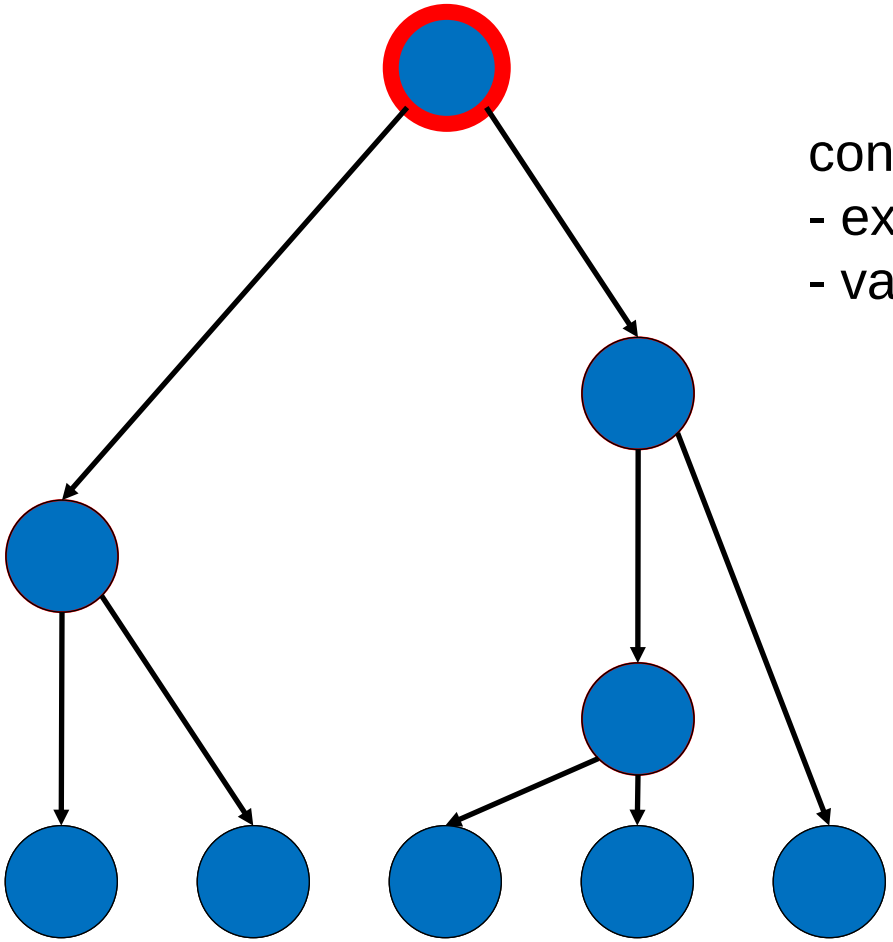
Time to the MRCA: variance

- T(*j*) are independent and var[*T*(*j*)] = 1 / (*j* choose 2)^2

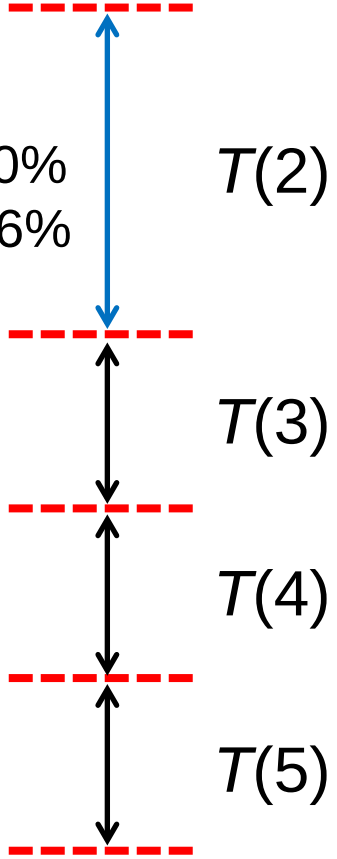
$$\begin{aligned}
\mathrm{var}[T_{\mathrm{MRCA}}(n)] &= \sum_{j=2}^{n} \mathrm{var}[T(j)] = \sum_{j=2}^{n} \left(\frac{2}{j(j-1)} \right)^2 \\
&= 4 \sum_{j=2}^{n} \left(\frac{1}{j-1} - \frac{1}{j} \right)^2 \\
&= 8 \sum_{j=1}^{n} \frac{1}{j^2} + \frac{4}{n^2} - 8 \left(1 - \frac{1}{n} \right) - 4
\end{aligned}$$

- $\mathrm{var}[T(2)] = 1, \quad \lim_{n \to \infty} \mathrm{var}[T_{\mathrm{MRCA}}(n)] = \frac{8\pi^2}{6} - 12 \approx 1.16$

$T_{MRCA}(n)$ is dominated by $T(2)$

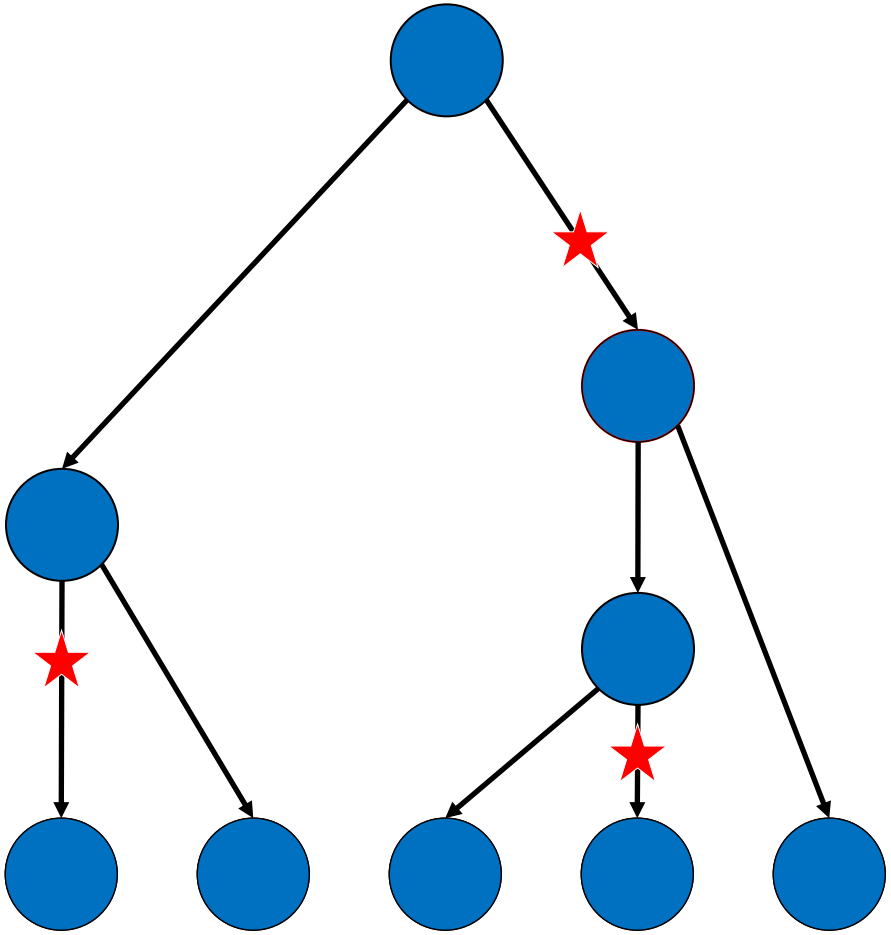


contribution to
- expectation: 50%
- variance: ~86%



Example: Four realizations for a sample of size $n = 10$

The mutation process is superimposed on the coalescent.



We assume a Poisson process that puts down mutations independently on all branches at rate $\theta/2$, where $\theta = 2\,N\,u$ is the scaled mutation rate.

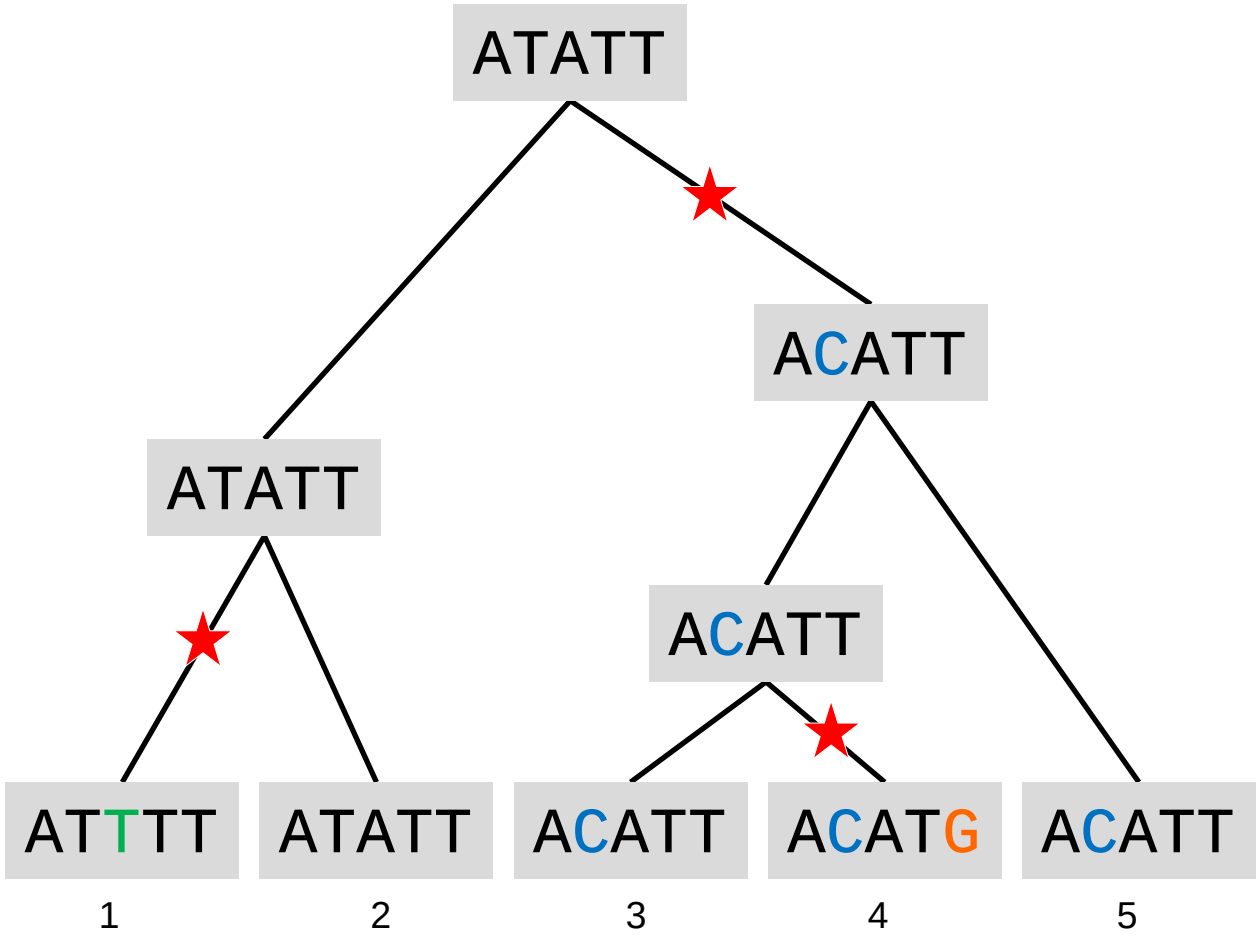
The infinite sites model

- Suppose observed individuals are identified genetically and the genomic region is long:

...ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACCAA...

- We assume an infinite number of sites (loci) and each mutation to affect a different nucleotide site.
- Thus, each mutation produces a new version of the gene and there is an infinite number of alleles.
- The infinite sites model is appropriate for long DNA sequences under neutral evolution.

Number of segregating sites, *S*



1 ATTTT
2 ATATT
3 ACATT
4 ACATG
5 ACATT

segregating sites

$$S = 3$$

Number of segregating sites, *S*

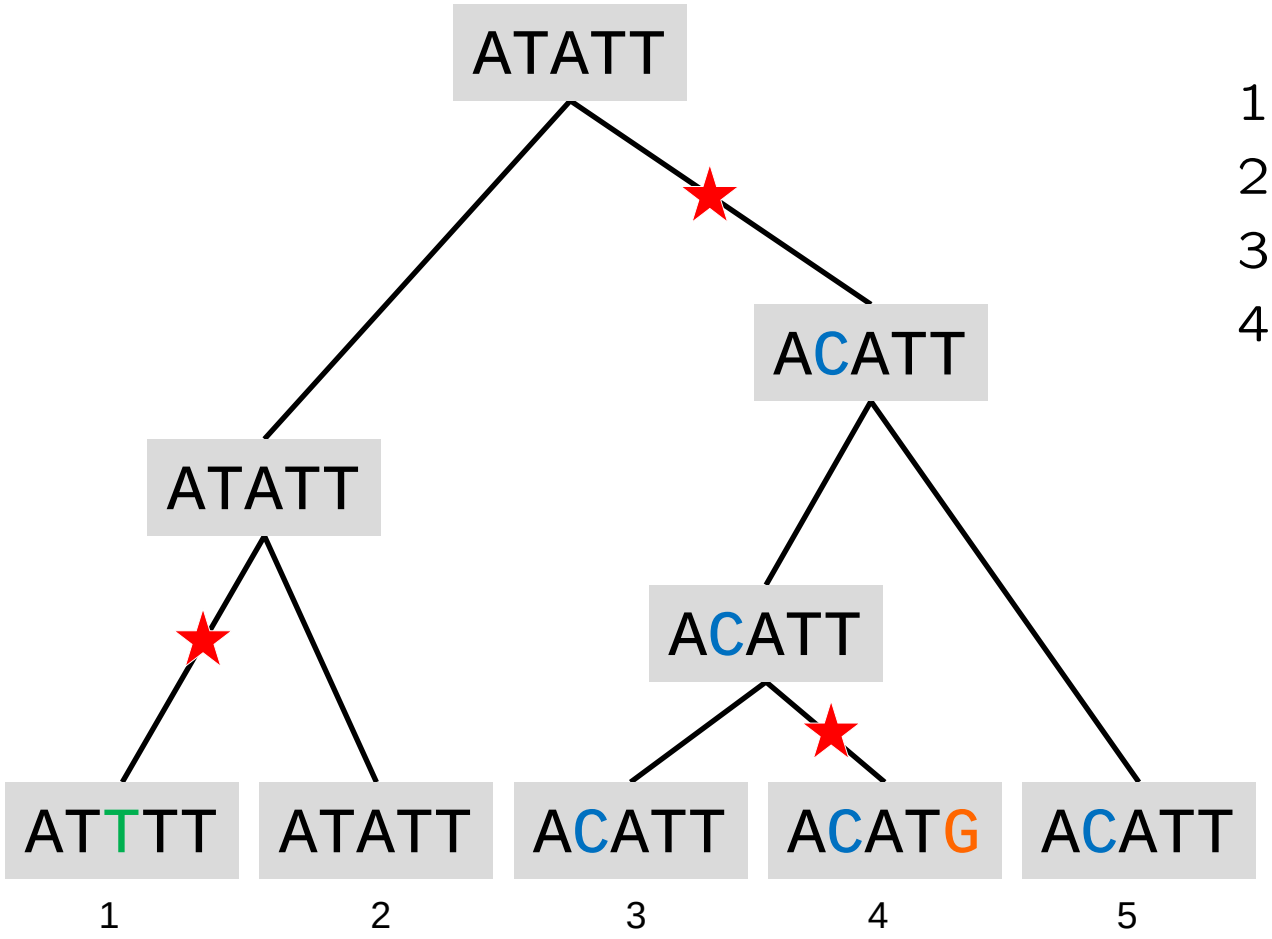
- Under the infinite sites model, *S* is equal to the total number of mutations of the genealogy.
- The total branch length is

$$T_{\text{tot}}(n) = \sum_{j=2}^{n} jT(j)$$

- Hence,

$$\mathrm{E}[S] = \frac{\theta}{2}\mathrm{E}[T_{\text{tot}}(n)] = \frac{\theta}{2}\sum_{j=2}^{n} j\frac{1}{\binom{j}{2}} = \theta\sum_{j=2}^{n}\frac{1}{j-1} = \theta\, c_n$$

Average pairwise nucleotide distance, *K*

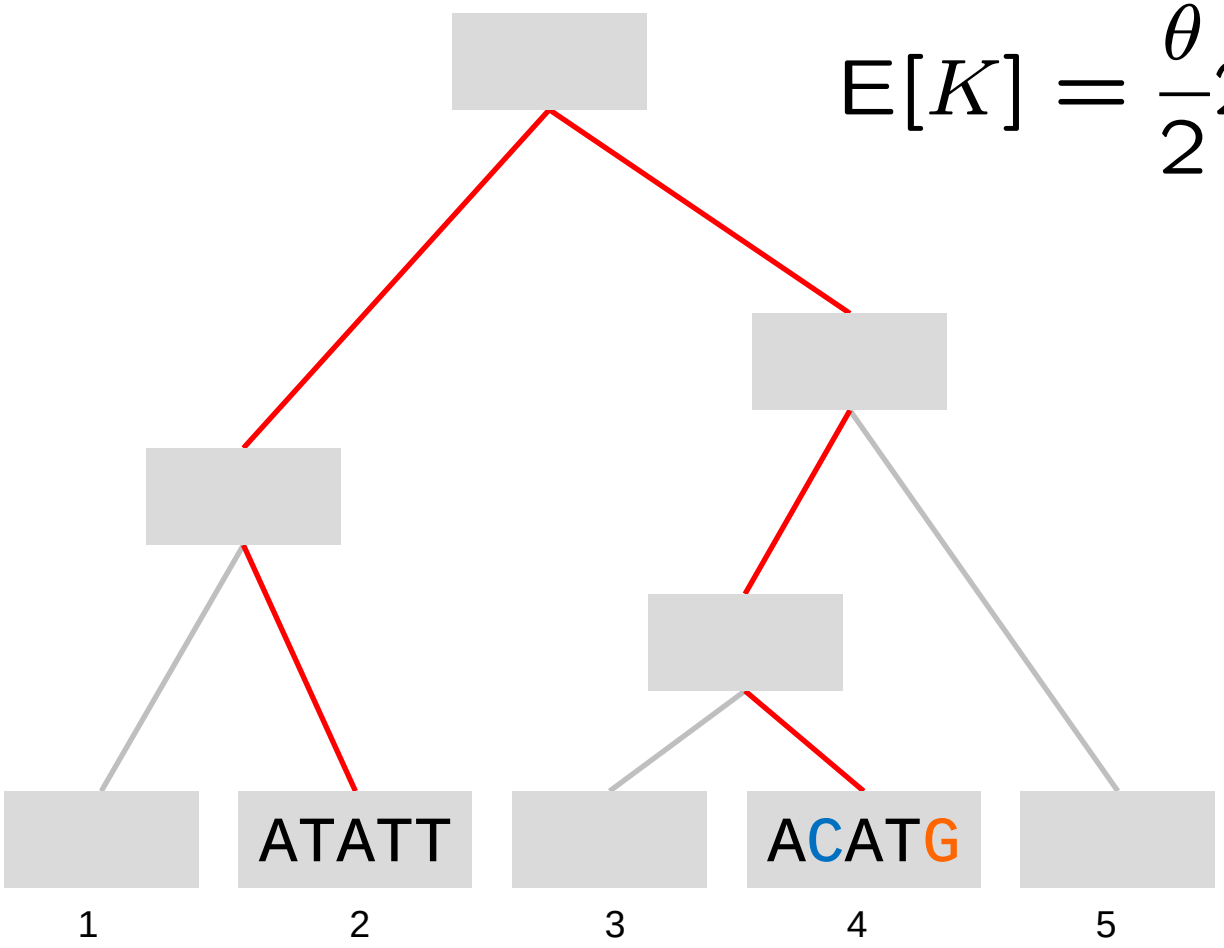


$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 2 & 3 & 2 \\ & 1 & 2 & 1 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix} \end{array}$$

$$K = 14/10$$

Average pairwise nucleotide distance, *K*

$$\mathrm{E}[K] = \frac{\theta}{2} 2\mathrm{E}[T(2)] = \theta$$



Detecting selection

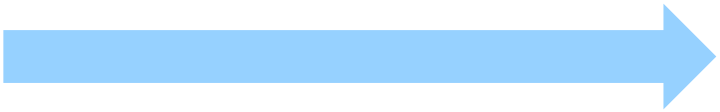
- Under the **neutral** infinite sites model, we have two different estimates of the mutation rate:

$$\mathrm{E}[K] = \theta = c_n^{-1}\mathrm{E}[S]$$

- Selection changes the allele frequencies in the population and affects these two estimates in different ways:
  - *S* ignores allele frequency changes, but is sensitive to low-frequency deleterious alleles.
  - *K* is strongly affected by allele frequencies, but largely insensitive to low-frequency deleterious alleles.

Example: All mutations are deleterious

1	ATTTT
2	ATATT
3	ACATT
4	ACATG
5	ACATT



1	ATTTT
2	ATATT
3	ACATT
4	ATATG
5	ATATT

$S = 3$
$K = 14/10$

$S = 3$
$K = 12/10$

Tajima's *D*

- The following test statistic is used for detecting selection (more precisely, deviation from neutrality):

$$D = \frac{\hat{K} - c_n^{-1}\hat{S}}{\sqrt{\hat{V}}} \quad \text{(Tajima's D)}$$

  where $\hat{K}$, $\hat{S}$, $\hat{V}$ are estimates of *K*, *S*, and the variance of $\hat{K} - c_n^{-1}\hat{S}$, respectively.
- The distribution of *D* under the null hypothesis of no selection is approximated by simulations of the coalescent.

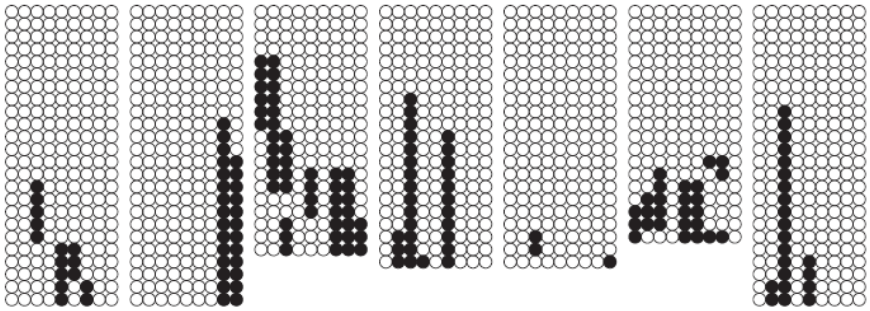
Inference under the coalescent

- Basic idea: The likelihood of the model parameters ϑ (e.g., mutation rate, population size, etc.) given observed data $\mathcal{D}$ (DNA sequences) is

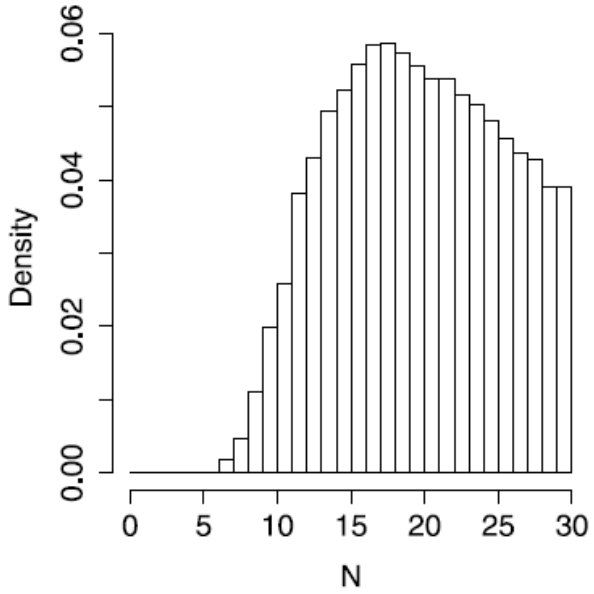
$$L(\vartheta) = P(\mathcal{D} \mid \vartheta) = \int \underbrace{P(\mathcal{D} \mid \mathcal{T}, \vartheta)}_{\text{statistical phylogenetic tree model}} \underbrace{P(\mathcal{T} \mid \vartheta)}_{\text{coalescent}} \, d\mathcal{T}$$

  where $\mathcal{T}$ runs over all phylogenetic tree models.
- Use MCMC to approximate integral
- Model parameters are estimated by maximum likelihood or Bayesian inference.

Example: The number of stem cells in a human colonic crypt estimated from methylation patterns of differentiated cells



7 crypts, 9 methylation sites



Summary

- The coalescent is a stochastic process that describes the random sampling of genealogies. It is based on the Wright-Fisher process.
- The coalescence time is distributed exponentially with parameter (*j* choose 2) in generation *j*.
- Tajima's *D* can detect selection by comparing, under the infinite sites model, two different estimates of the mutation rate, one based on the number of segregating sites, the other on pairwise distances.
- The coalescent can be used to infer population parameters from observed DNA sequence data.

References

- Rosenberg NA et al. (2002) Nat Rev Genet 3:380
- Neuhauser C (2007) Handbook of Statistical Genetics (D.J. Balding et al., editors), Chapter 22, pp. 755-780
- Nicolas P et al. (2007) PLoS Comput Biol 3(3):e28