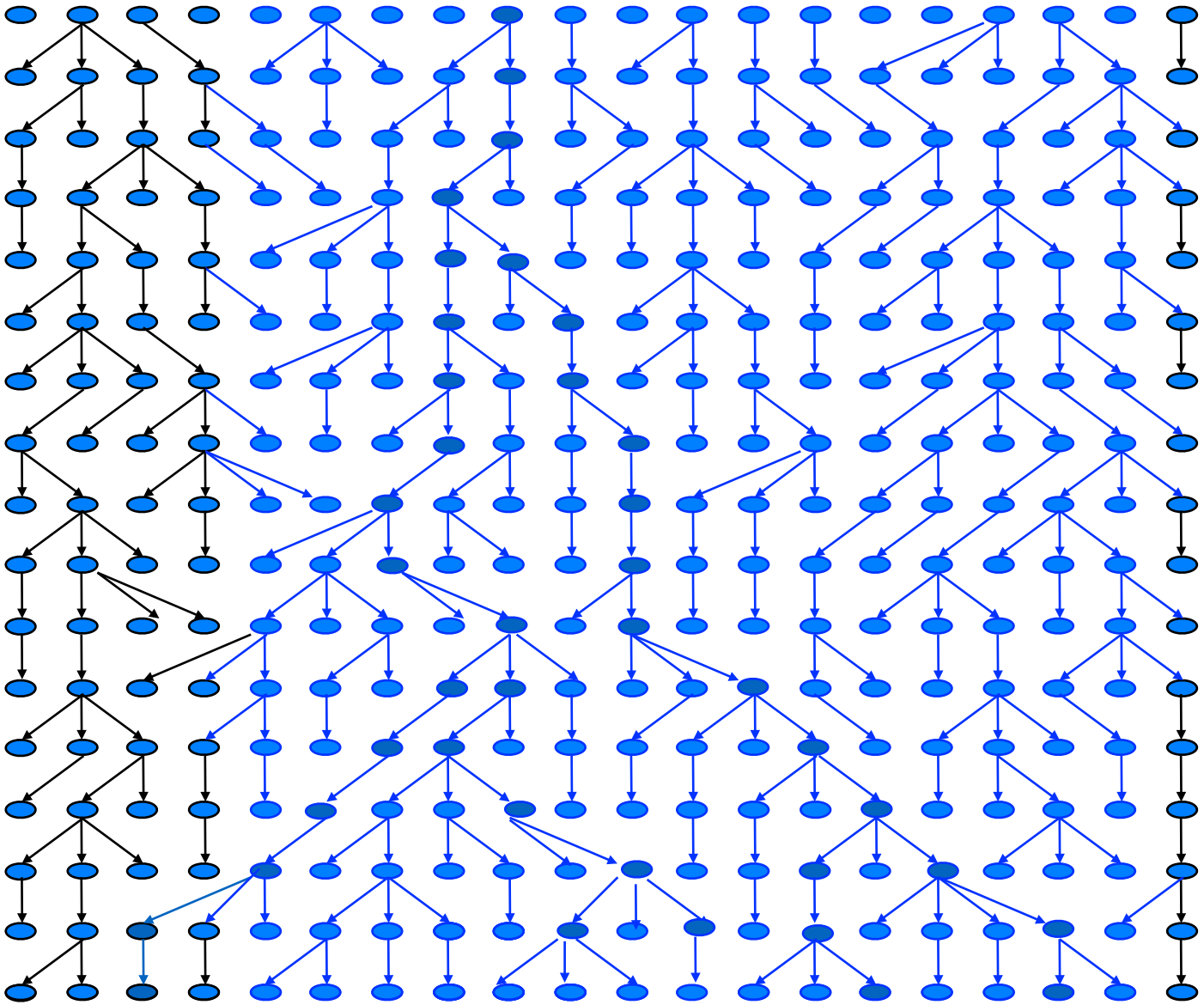
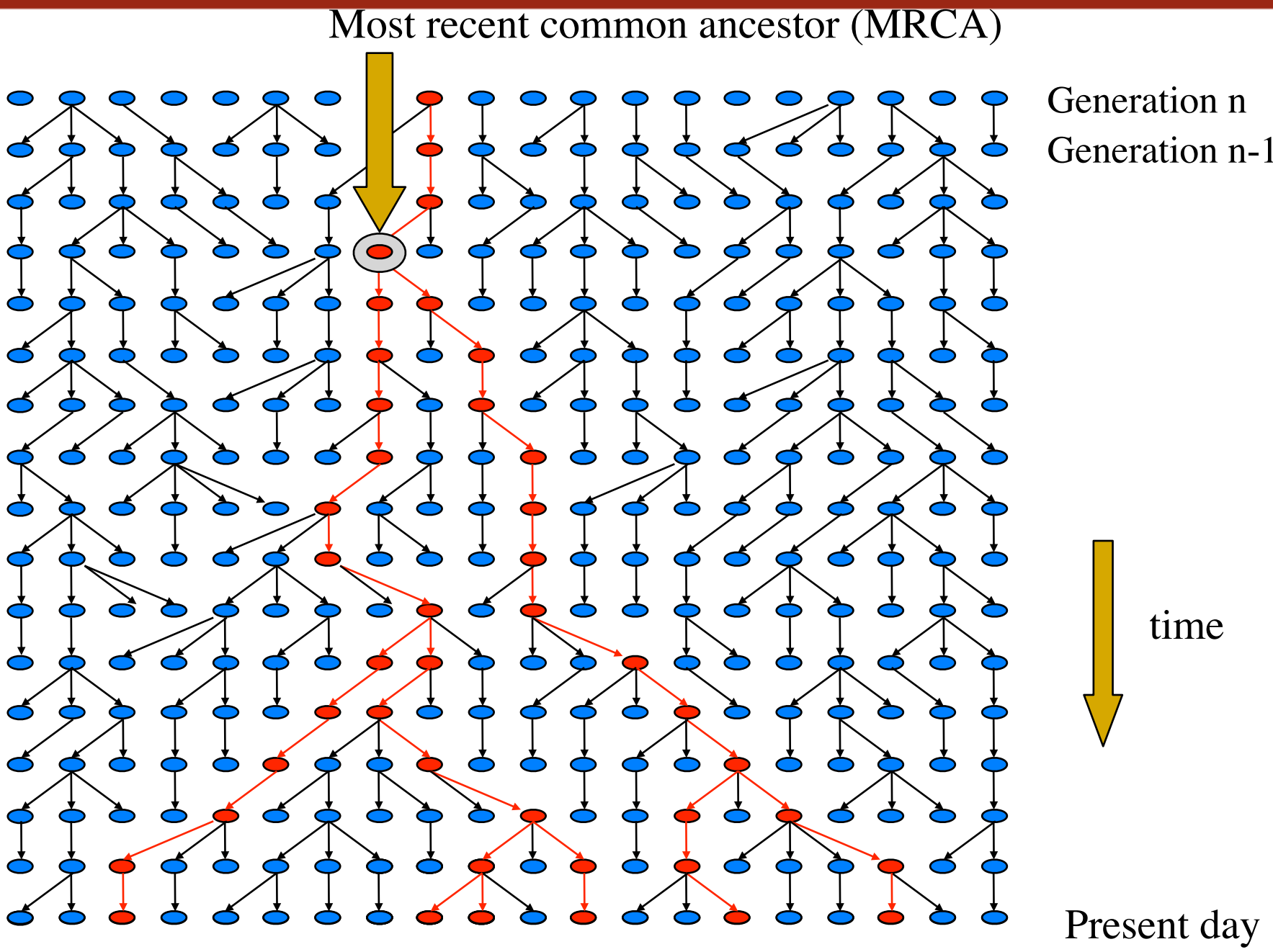
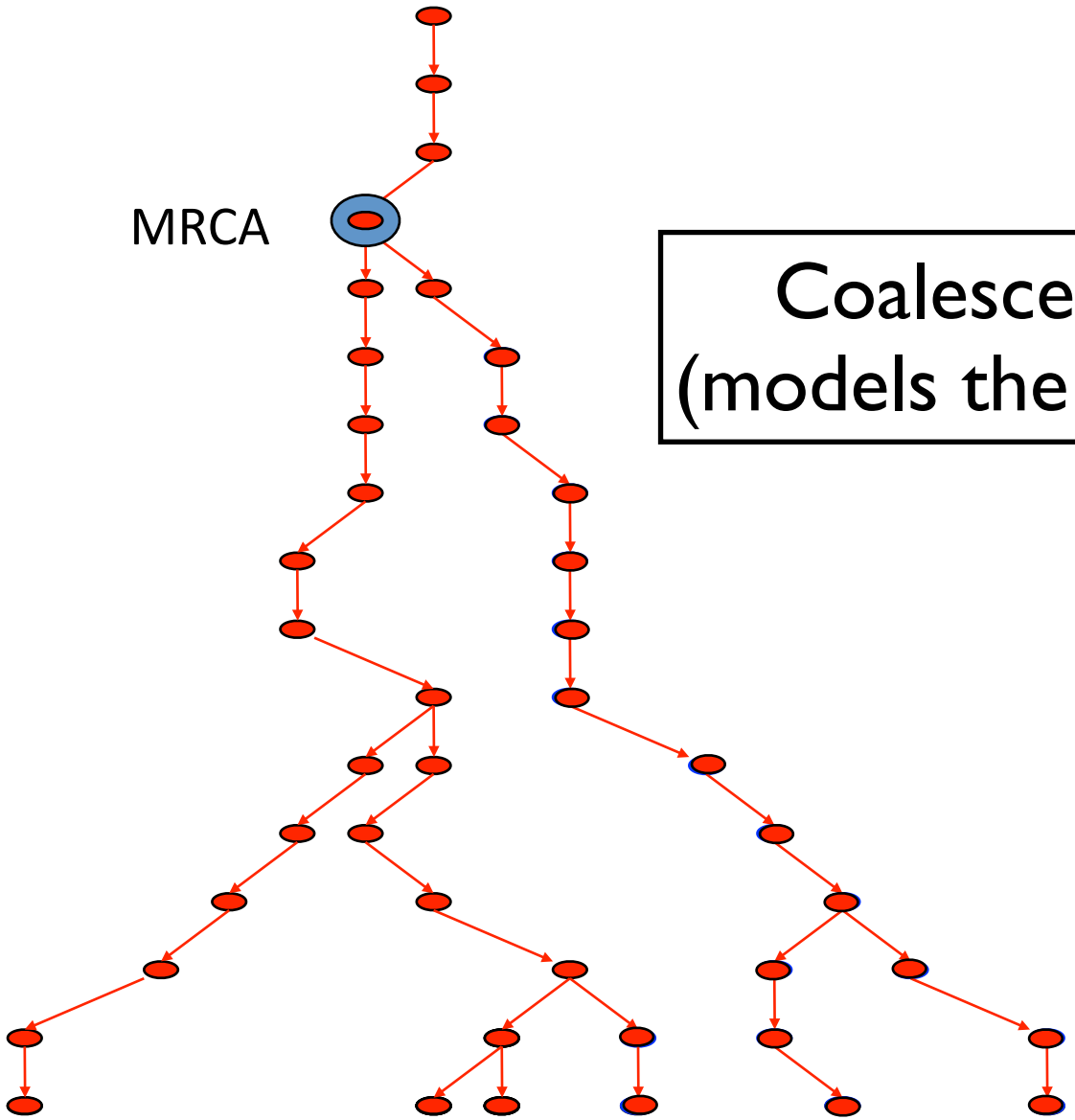
Population Genetics, Coalescent Trees and Urns

Caspar Friedrich - “Dolmen in the snow” (1807)

Schematic of evolving population




Most recent common ancestor (MRCA)
Generation n
Generation n-1
time
Present day



Coalescent tree
(models the ancestry)

time

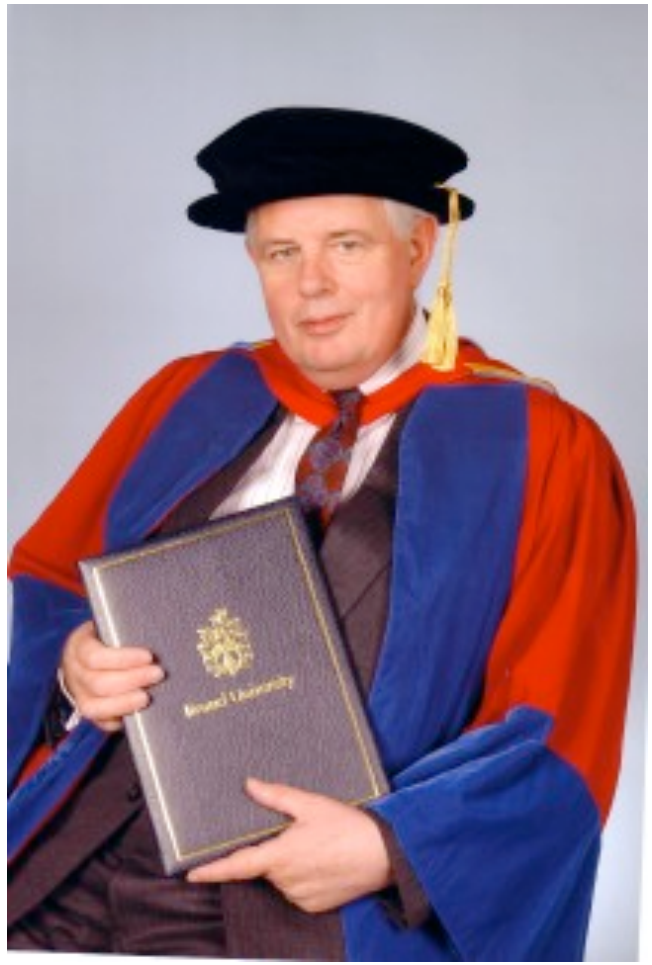
Present day

Ancestral methods

- The coalescent was introduced in Kingman (1982) as a mathematical description of the genealogy that underlies the evolution of a population.

- It has become a standard tool for the analysis of molecular population data.

- It allows for efficient modeling/analysis of random samples drawn from a population.



John Kingman, 1939-

Coalescent trees - algorithm

- Time runs in reverse! (We start at the bottom and work our way up.)
- Start with a sample of n individuals (set k=n)
- Time to next event ~exp(k(k-1)/2)
  - Choose two lines of ancestry, uniformly at random, to coalesce. (So keep a list of the all the lines of ancestry that still exist.)
  - Label the new node as n+(n-k+1) . Add it to the list of lines of ancestry that exist. Remove the two lines that coalesced.
  - Set k=k-1
- Keep going until you get down to one line (i.e., k=1)
- Forwards in time, this is equivalent to a model in which the offspring distribution is Multinomial(N,1/N,1/N,…,1/N)

Markov Chain

A *Markov chain* is a sequence of random variables X_1, X_2, X_3, ... with the Markov property, namely that, given the present state, the future and past states are independent (Wikipedia).

Generic example: a random walk: X(i+1)=X(i) +R, where R is some random variable.

Born: 14 June 1856 in Ryazan, Russia
Died: 20 July 1922 in Petrograd (now St Petersburg), Russia

http://www-history.mcs.st-and.ac.uk/Biographies/Markov.html

Theoretical results

- Let T_i be the time for the coalescence from i to i-1 lines.
- Let T be the overall height of the tree. Then:
-

$$
\begin{aligned}
E(T) &= E(\sum_{i=2}^{n} T_i) \\
&= \sum_{i=2}^{n} E(T_i) \\
&= \sum_{i=2}^{n} 2/i(i-1) \\
&= \frac{2}{n(n-1)} + \frac{2}{(n-1)(n-2)} + \cdots \frac{2}{3 \times 2} + \frac{2}{2 \times 1} \\
&= \cdots
\end{aligned}
$$

Theoretical results

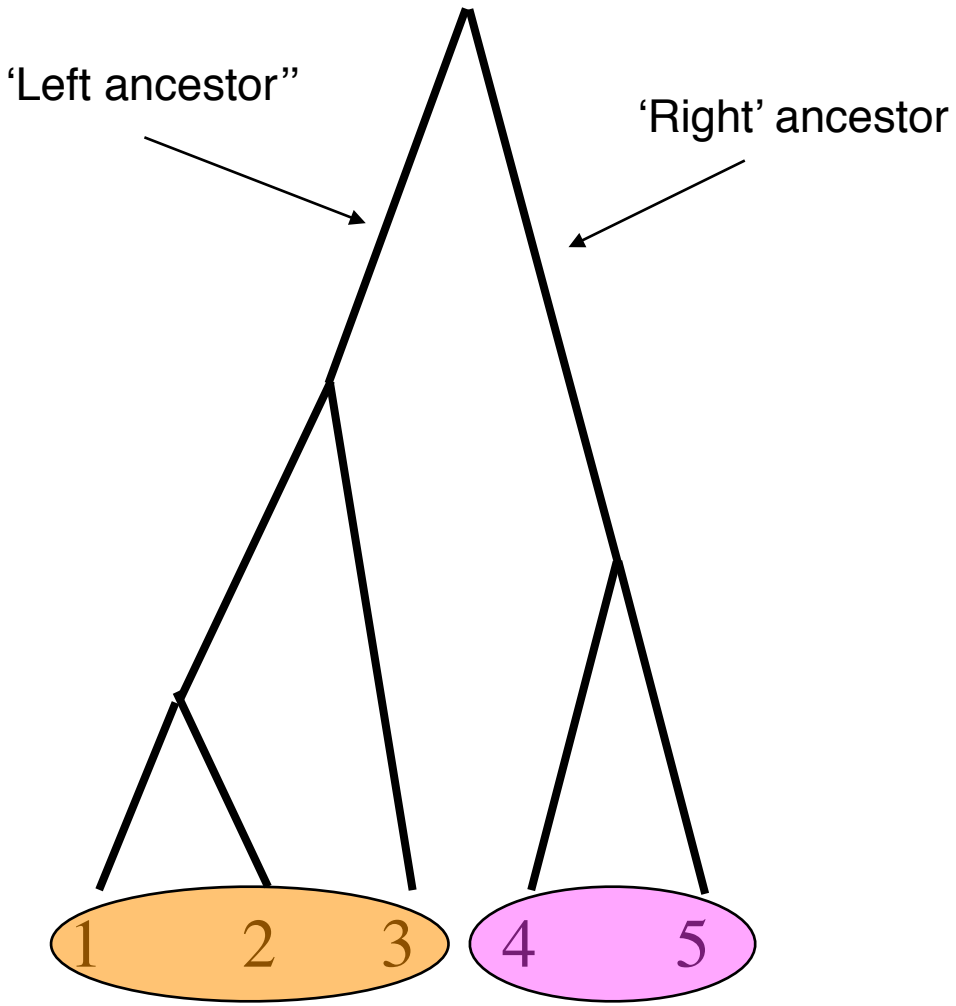
- Claim E(Tree height)=2(1-1/n)
- Can think of time here as being measured in units of N generations, where N is the population size (Kingman, 1982).
- Note that Expected tree height increases as sample size n increases.
- But also note that E(Tree height) never exceeds (or even reaches) 2 (or 2N generations).

Theoretical results

- Let L_i be the tree length for coalescence from i to i-1 lines.
- Let L be the overall height of the tree. Then:
-

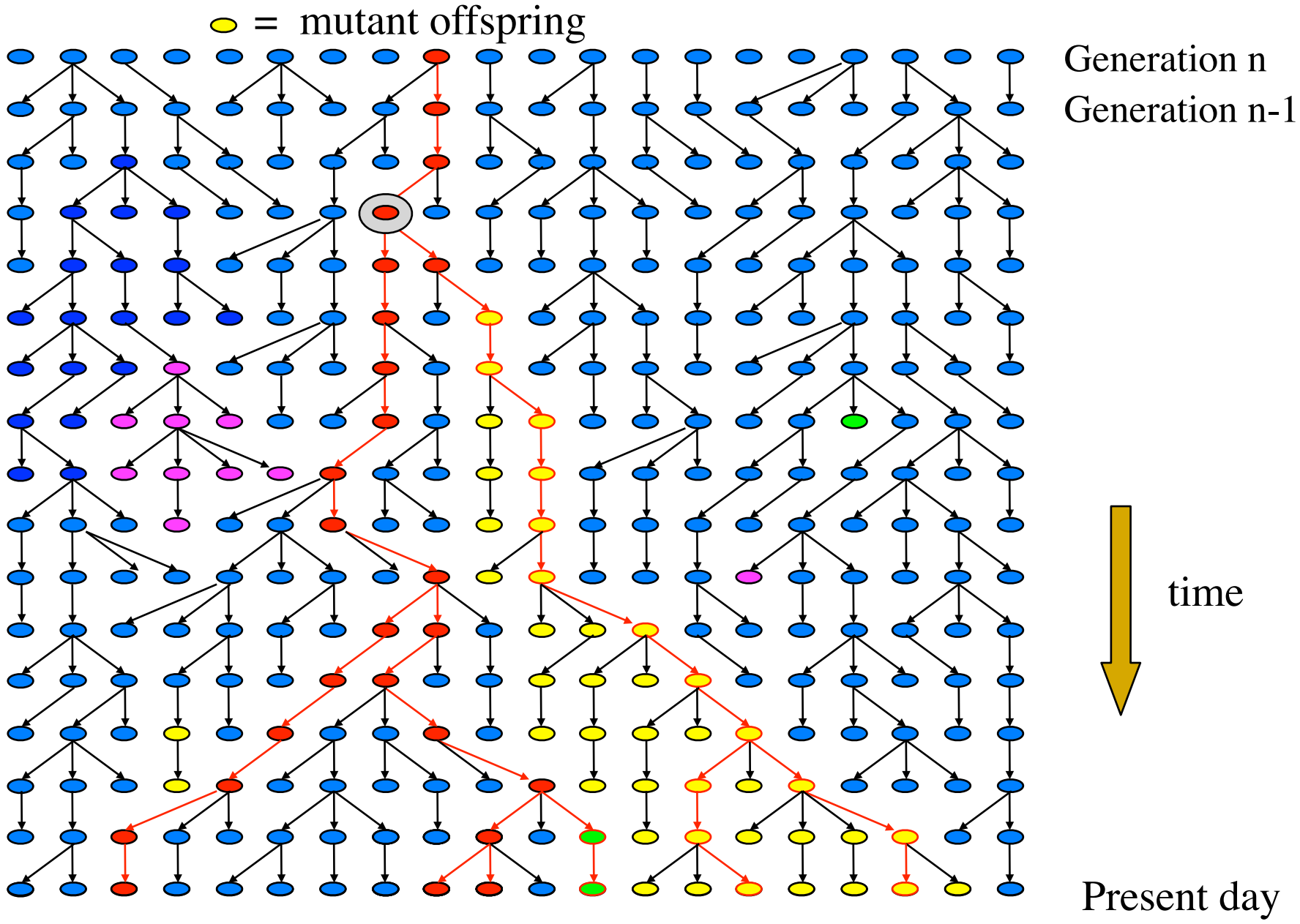
$$
\begin{aligned}
E(L) &= E(\sum_{i=2}^{n} L_i) = E(\sum_{i=2}^{n} iT_i) \\
&= \sum_{i=2}^{n} E(iT_i) \\
&= \sum_{i=2}^{n} 2i/i(i-1) \\
&= \frac{2}{(n-1)} + \frac{2}{(n-2)} + \cdots \frac{2}{2} + \frac{2}{1} \\
&= 2\sum_{i=2}^{n} \frac{1}{n-1} \\
&= 2\sum_{i=1}^{n-1} \frac{1}{n} \\
&\longrightarrow 2log(n-1), \quad \text{as } n \longrightarrow \infty
\end{aligned}
$$

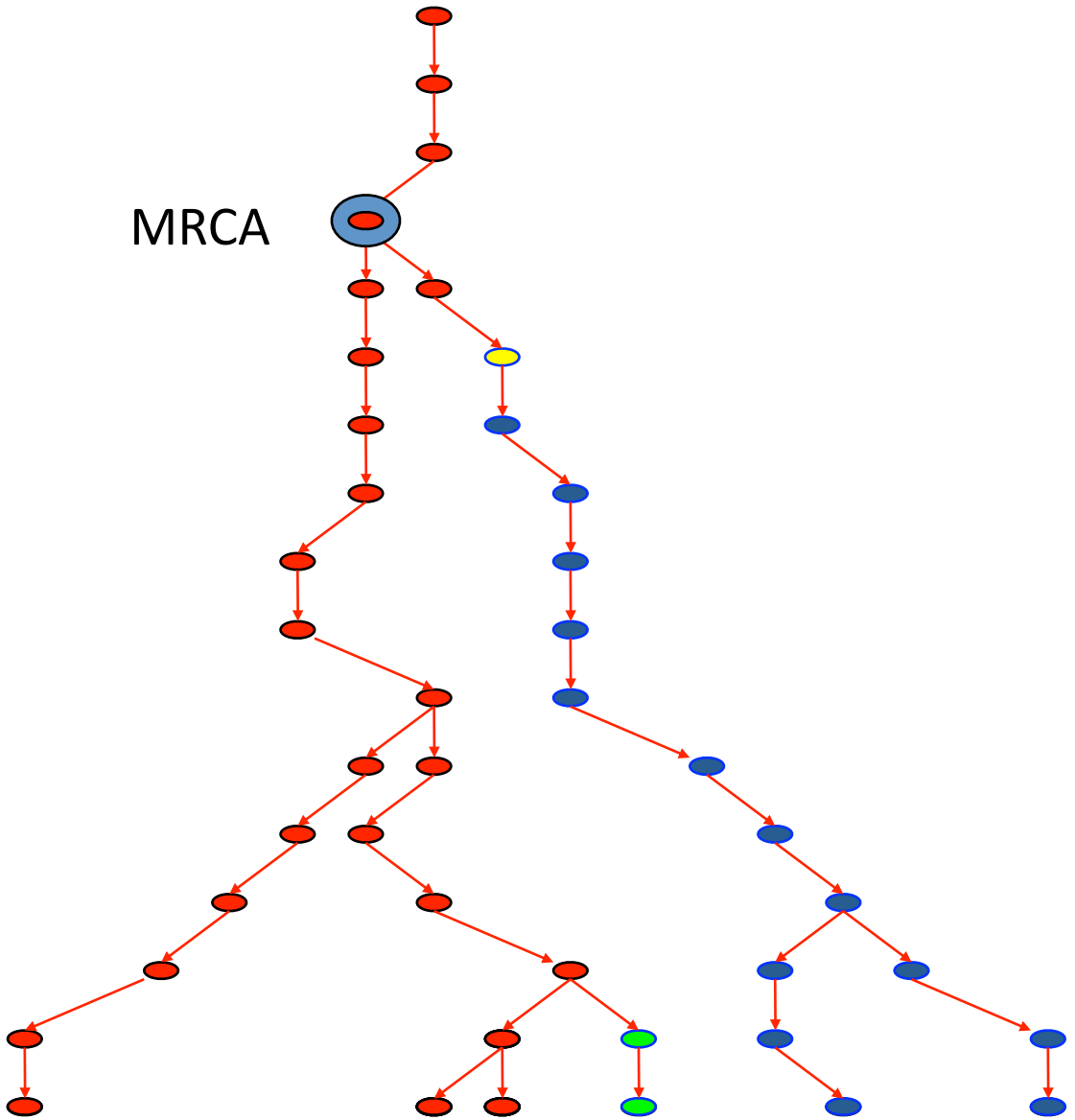
Optionmal Coalescent - Number of descendants of last two lines



Suppose we have a sample of size 50. What is the distribution of the number of descendants of the ‘left (or right) ancestor’?

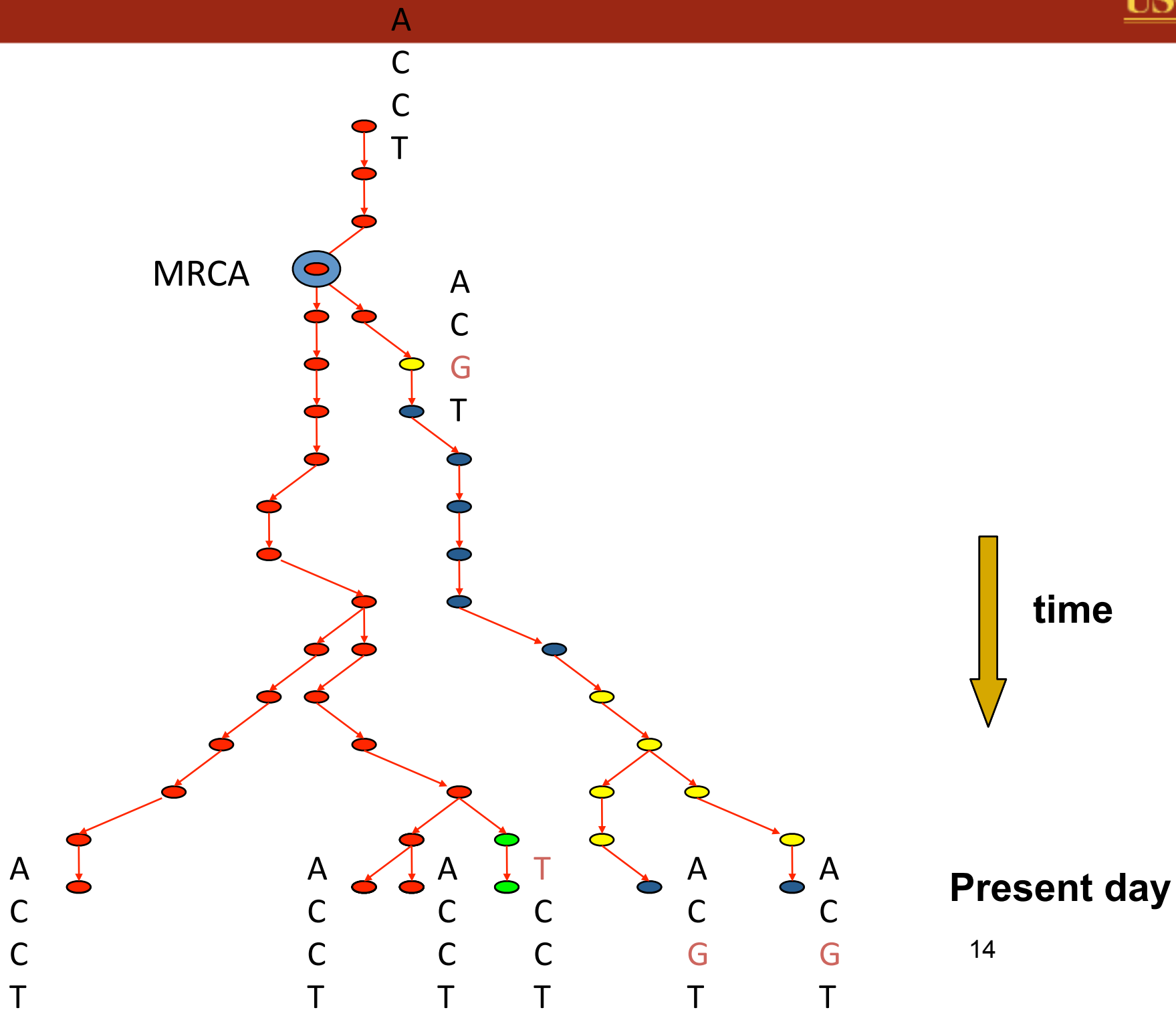
Now we add mutation





time

Present day


A
C
C
T
MRCA
A
C
G
T
time
A
C
T
T
A
C
C
T
A
C
C
T
T
C
C
T
A
C
G
T
A
C
G
T
Present day

- Mutations may now appear on lines of ancestry.
- Recall, N = population size
- Suppose we have k lines of ancestry, indexed 1,2,...,k

- In a discrete generation model:
  - P(lines i and j coalesce)=1/N
  - k lines: P(some pair coalesce)=k(k-1)/2N

  - P(line i mutates)=u
  - k lines: P(some line mutates)~ku
  - Define θ=2Nu: P(some line mutates)~kθ/2N

- P(coalesce) = k(k-1)/2N
- P(mutation) = kθ/2N

- The **Jump chain** (the chain observed only at times when the state changes) is as follows: one of two things will happen
  - P(coalesce) = [k(k-1)/2N] / [k(k-1)/2N + kθ/2N]
    = (k-1)/(k-1+θ)
  - P(mutation) = θ/(k-1+θ)
- Again, the jump chain is a **Markov chain**. ("What you do next depends only upon where you are, not how you got there.")

Random nice results:

$$E(\#\text{ mutations}) = \sum_{k=2}^{n} \mathrm{E}\left(\exp\binom{k}{2} k\theta/2\right)$$

$$= \sum_{k=2}^{n} \frac{\theta}{k-1} = \theta \sum_{k=1}^{n-1} \frac{1}{k}$$

- Watterson's estimator of mutation rate:

$$\hat{\theta} = S / \sum_{k=1}^{n-1} \frac{1}{k}$$

where S is the observed number of segregating sites (estimate)

A sample is enough

- P(sample of n has same MRCA as population of N) =

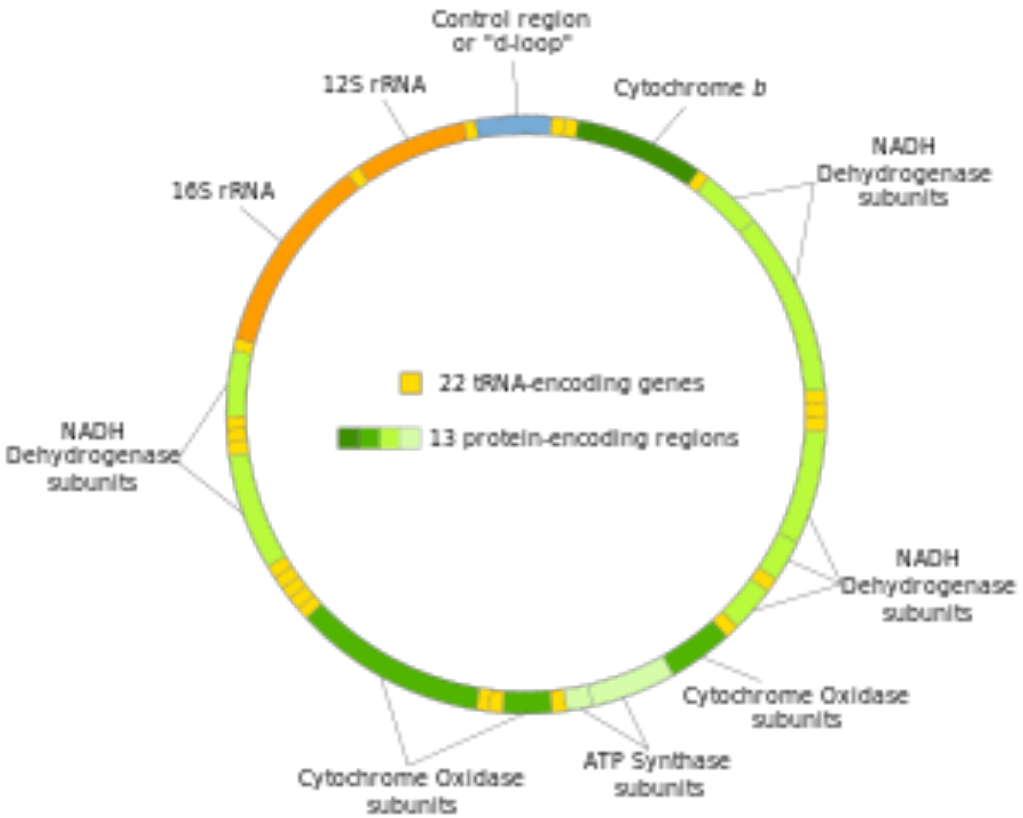
$$\frac{n-1}{n+1}\frac{N+1}{N-1}$$

(Saunders, I.W., Tavaré, S., Watterson, G.A.: On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. 16, 471–491 (1984))

Mutation models

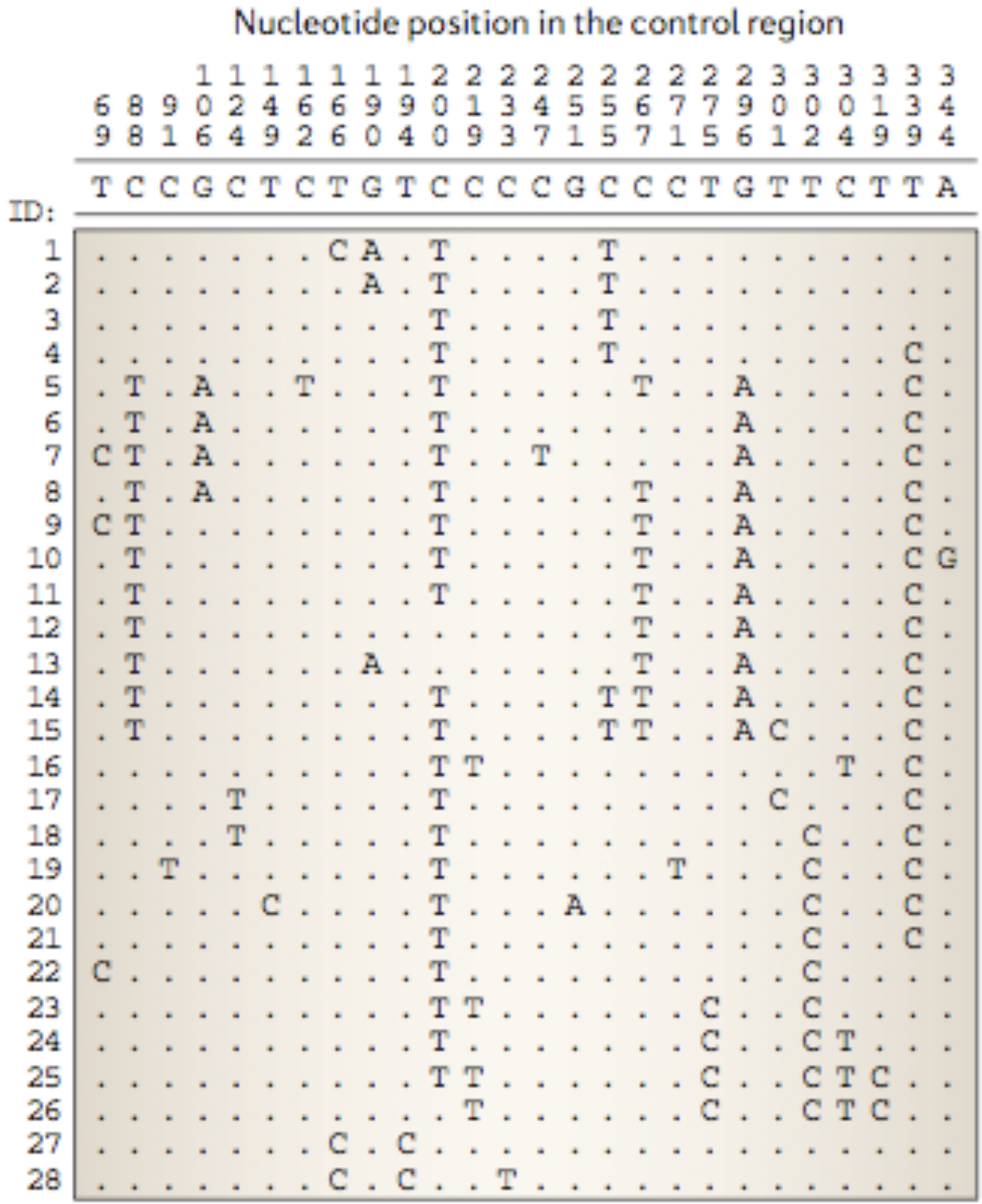
- There are many different mutation models:
  - **Infinite Alleles** - each mutation creates a previously unseen allele [type]
  - **Finite Sites** - there are a finite number of sites at which mutation can occur (e.g. mtDNA)
  - **Infinite Sites** - assume there are an infinite number of sites that could mutate. Thus, each mutation will occur at a unique site. (Sounds a bit like Infinite alleles, but actually carries more information about ancestry.).
  - What model for the human genome?

mtDNA



- Just 16000 base pairs
- Loop structure
- Codes for ~40 genes
- Energy factories
- Up to ~1000/cell
- Originally from bacteria engulfed by early eukaryote ancestors

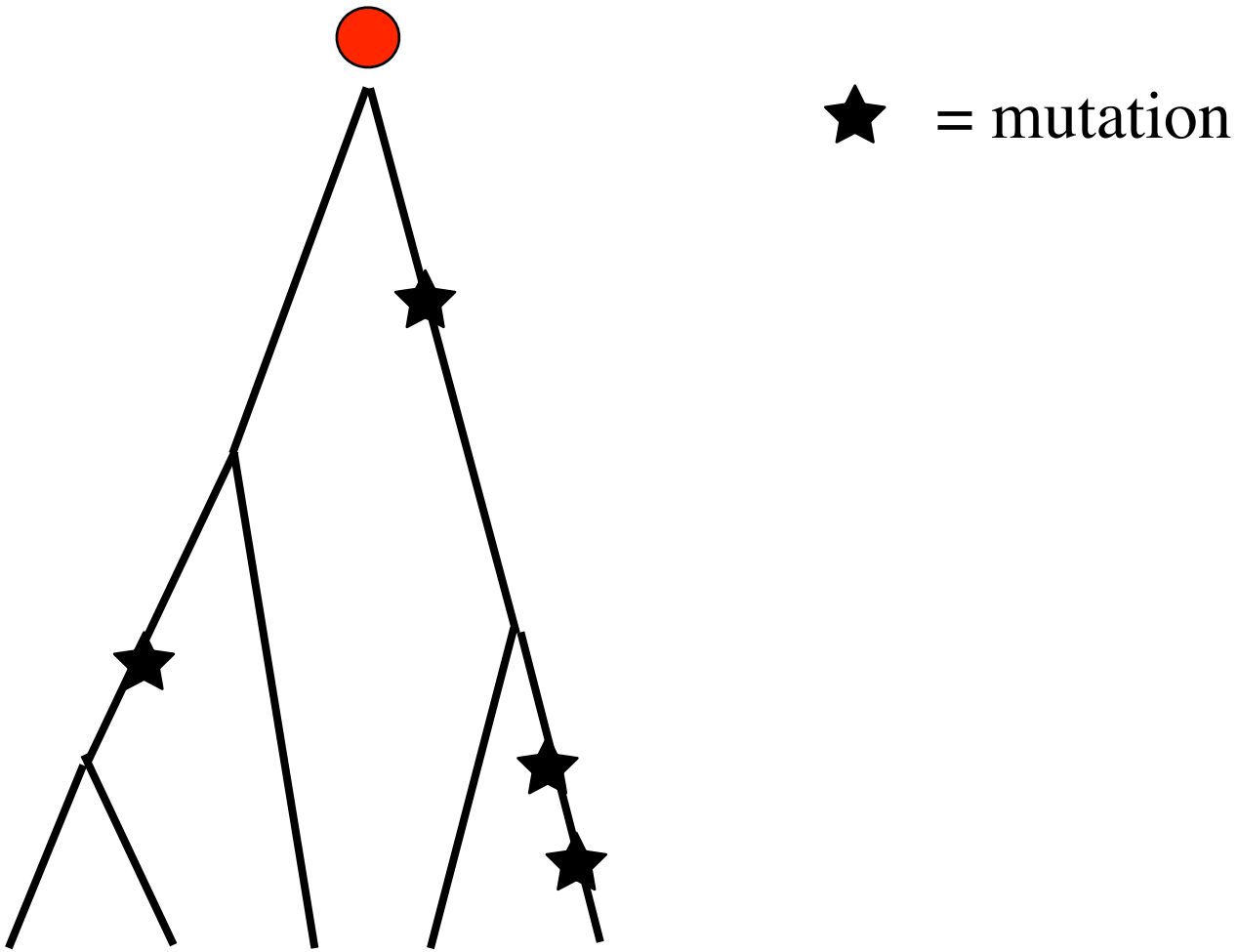
http://en.wikipedia.org/wiki/Mitochondrial_DNA


Nucleotide position in the control region
ID:
1 . . . . C A . T . . . . T . . . .
2 . . . . A . T . . . . T . . . .
3 . . . . T . . . . T . . . .
4 . . . . T . . . . T . . . . C . .
5 . T . A . . T . . . T . . . . . T . . A . . . . C . .
6 . T . A . . . . . . T . . . . A . . . . C . .
7 C T . A . . . . . . T . . T . . . . . A . . . . C . .
8 . T . A . . . . . . T . . . . . T . . A . . . . C . .
9 C T . . . . T . . . . . T . . A . . . . C . .
10 . T . . . . T . . . . . T . . A . . . . C G
11 . T . . . . T . . . . . T . . A . . . . C . .
12 . T . . . . T . . A . . . . C . .
13 . T . . . . . . A . . . . T . . A . . . . C . .
14 . T . . . . T . . . . T T . . A . . . . C . .
15 . T . . . . T . . . . T T . . A C . . . C . .
16 . . . . T T . . . . T . C . .
17 . . . . T . . . . . T . . . . C . . . C . .
18 . . . . T . . . . . T . . . . C . . C . .
19 . . T . . . . T . . . . T . . . C . . C . .
20 . . . . . C . . . . T . . . A . . . . . . C . . C . .
21 . . . . T . . . . C . . C . .
22 C . . . . T . . . . C . . . . .
23 . . . . T T . . . . . . C . . C . . . . .
24 . . . . T . . . . C . . C T . . . .
25 . . . . T T . . . . . . C . . C T C . . .
26 . . . . T . . . . . . C . . C T C . . .
27 . . . . C . C . . . .
28 . . . . C . C . . T . . . .

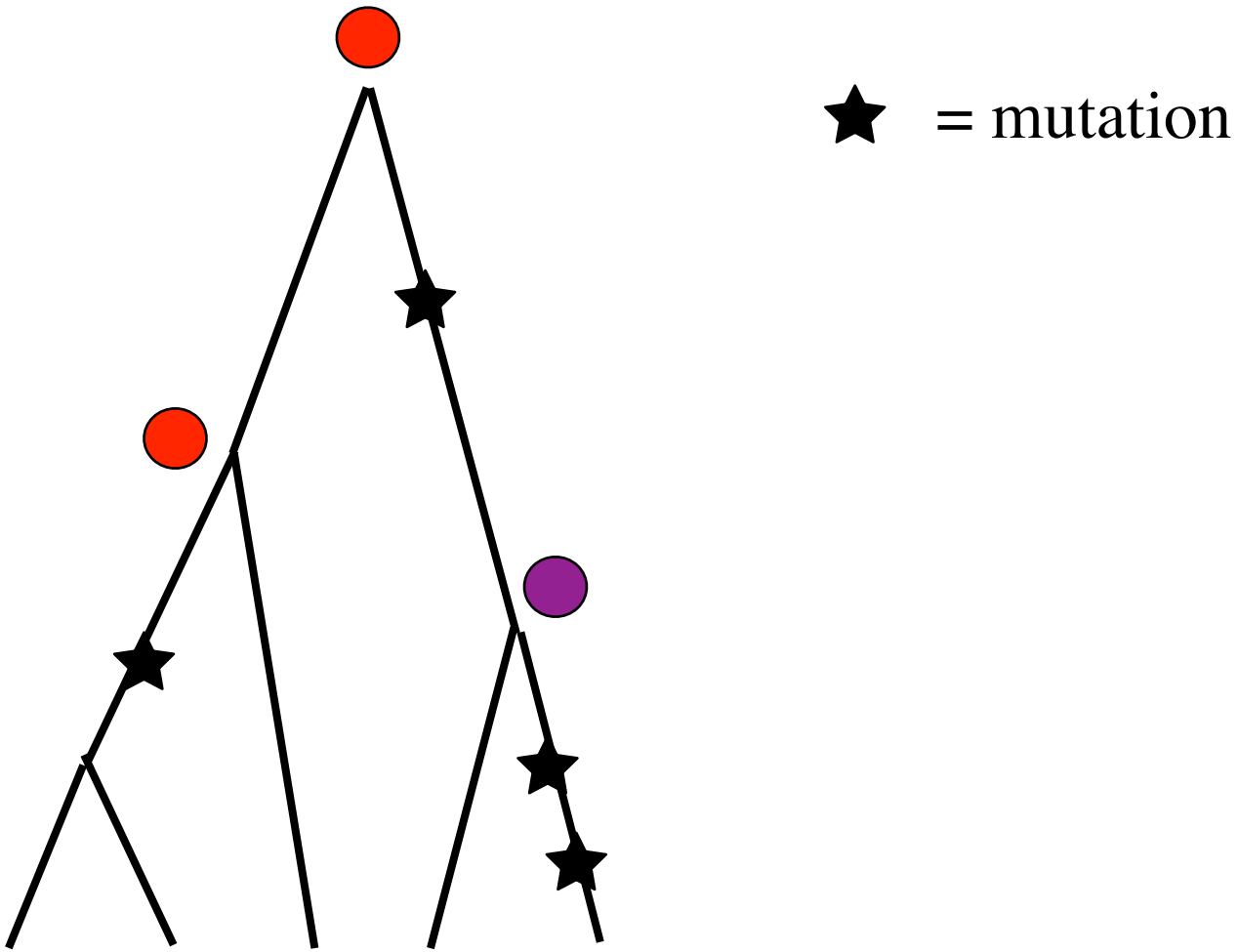
infinite alleles model

- Think of types as colors
  - The type of the bottom of a branch is the same as its type at the top if there are no mutations on the branch
  - Otherwise, it is of some unique, new type

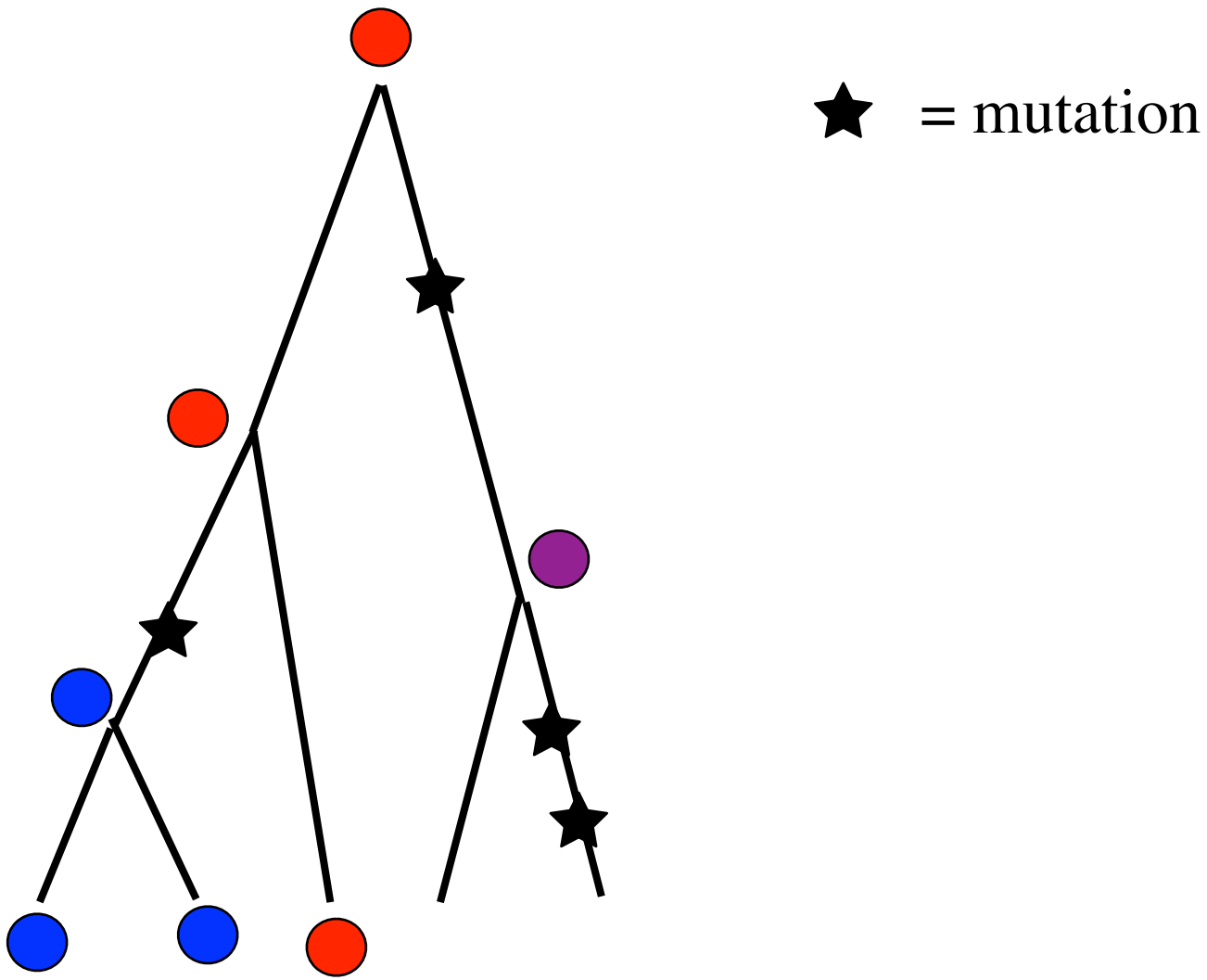
Illustration



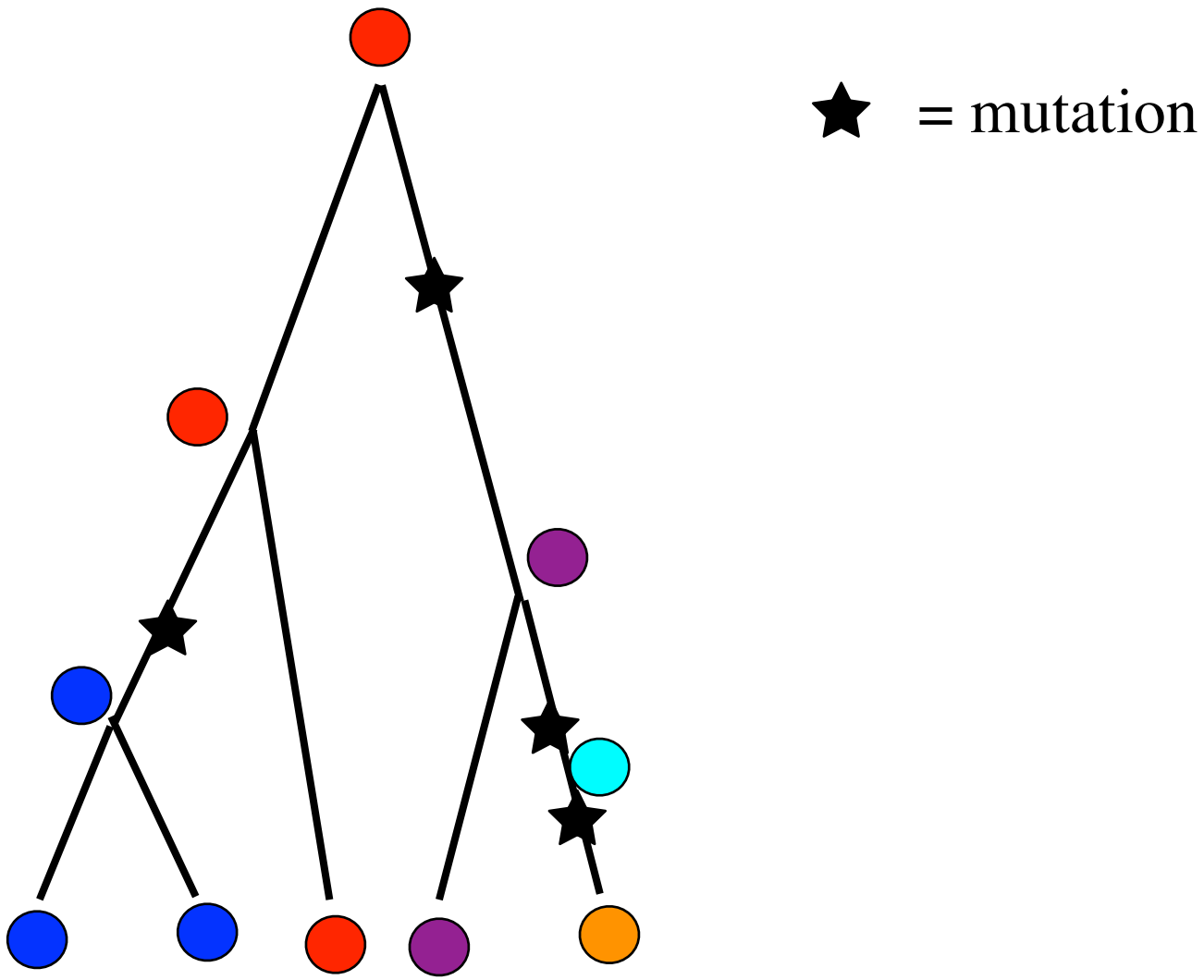
Illustration



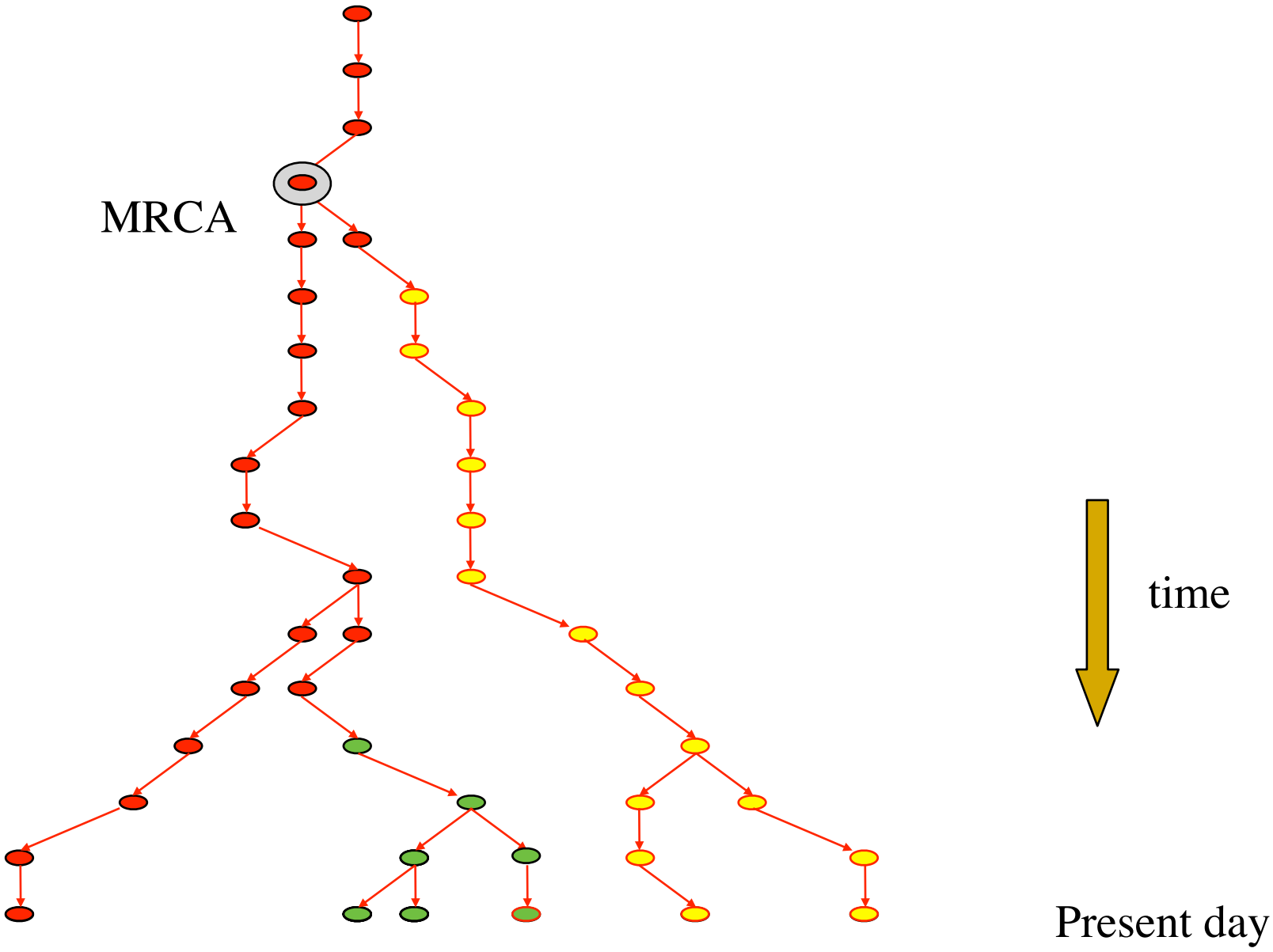
Illustration

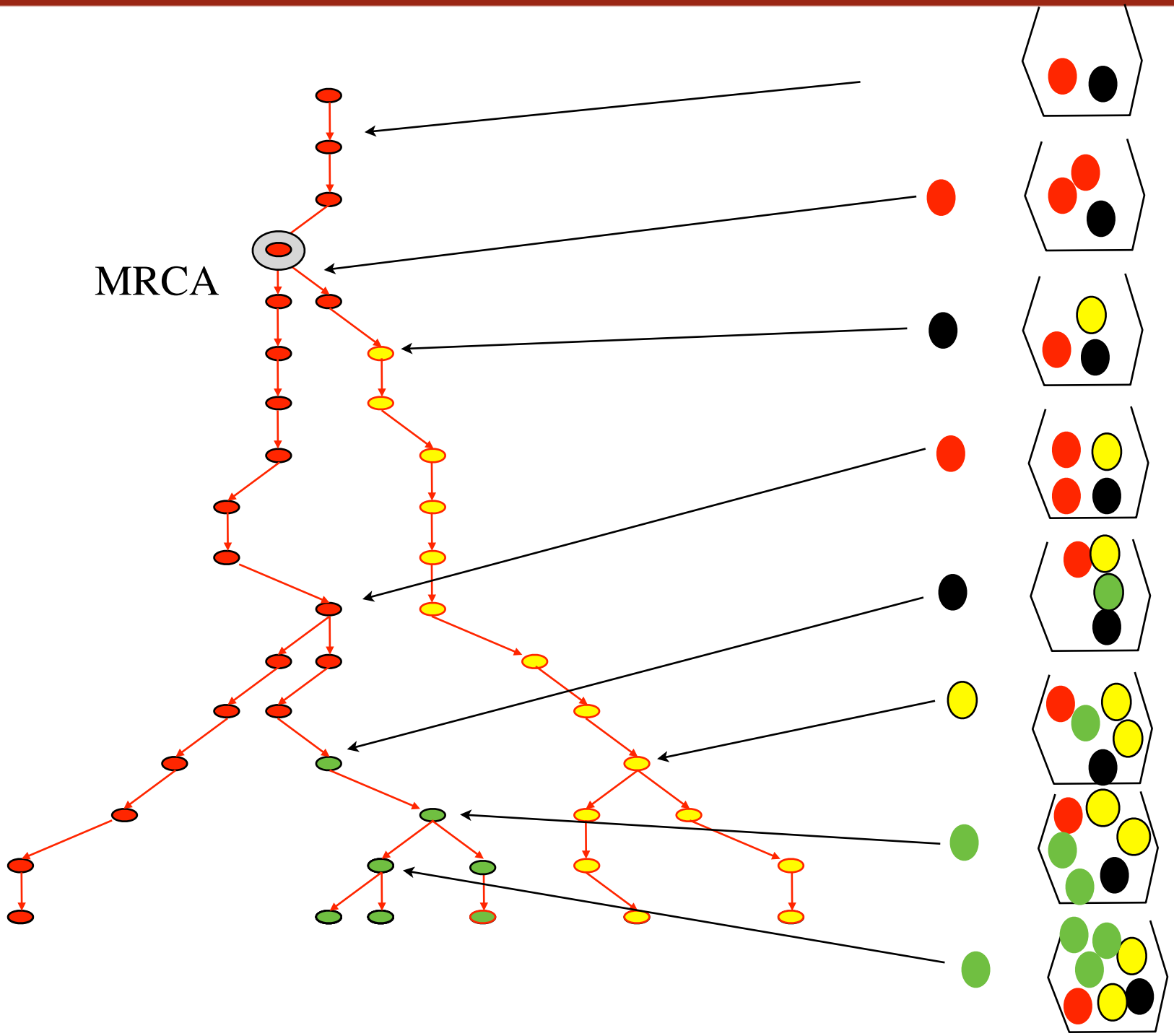


Illustration



Comparing the coalescent to the Urn model


MRCA
time
Present day


MRCA

Real-life is complicated….

The coalescent with recombination.

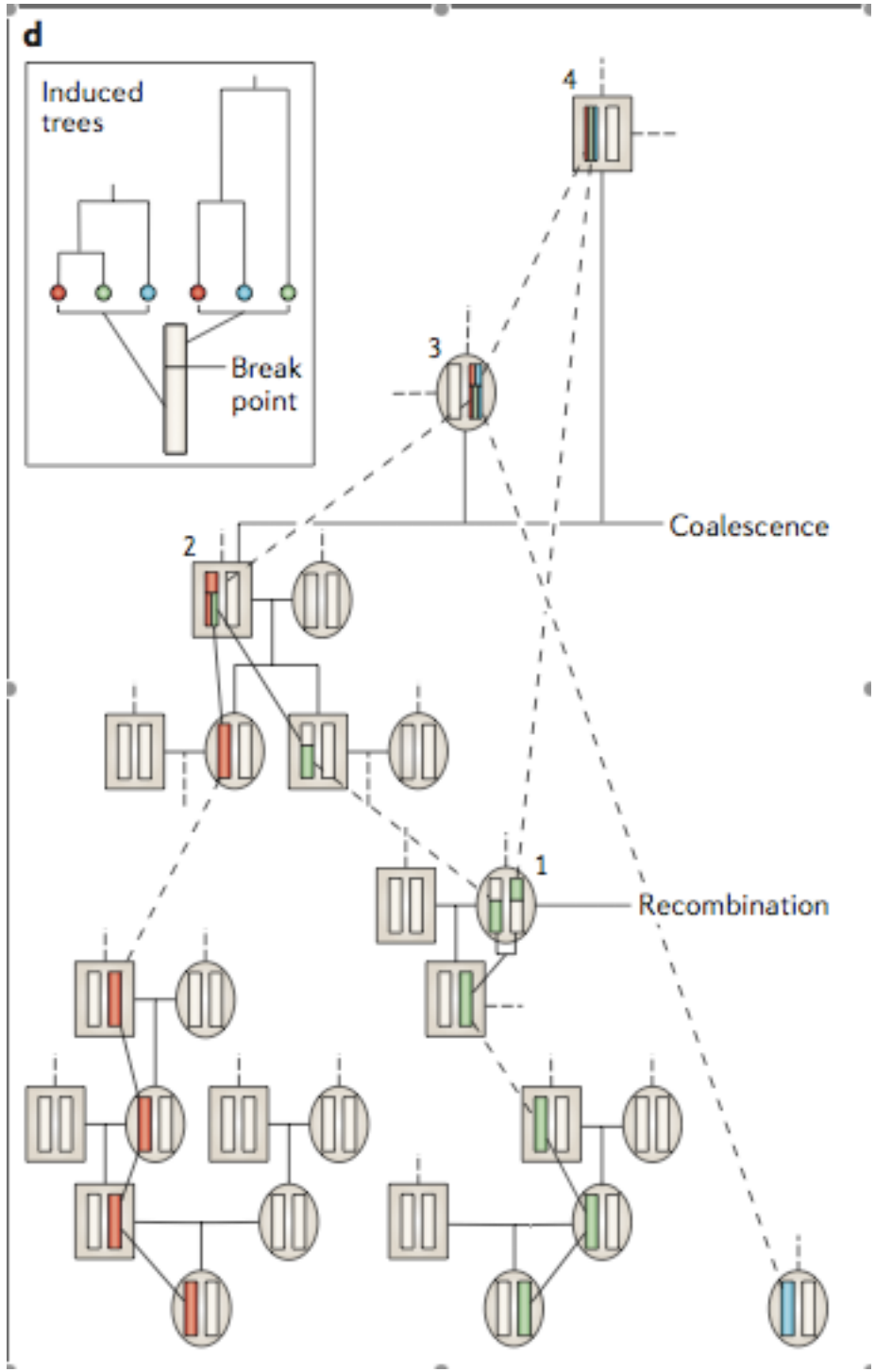
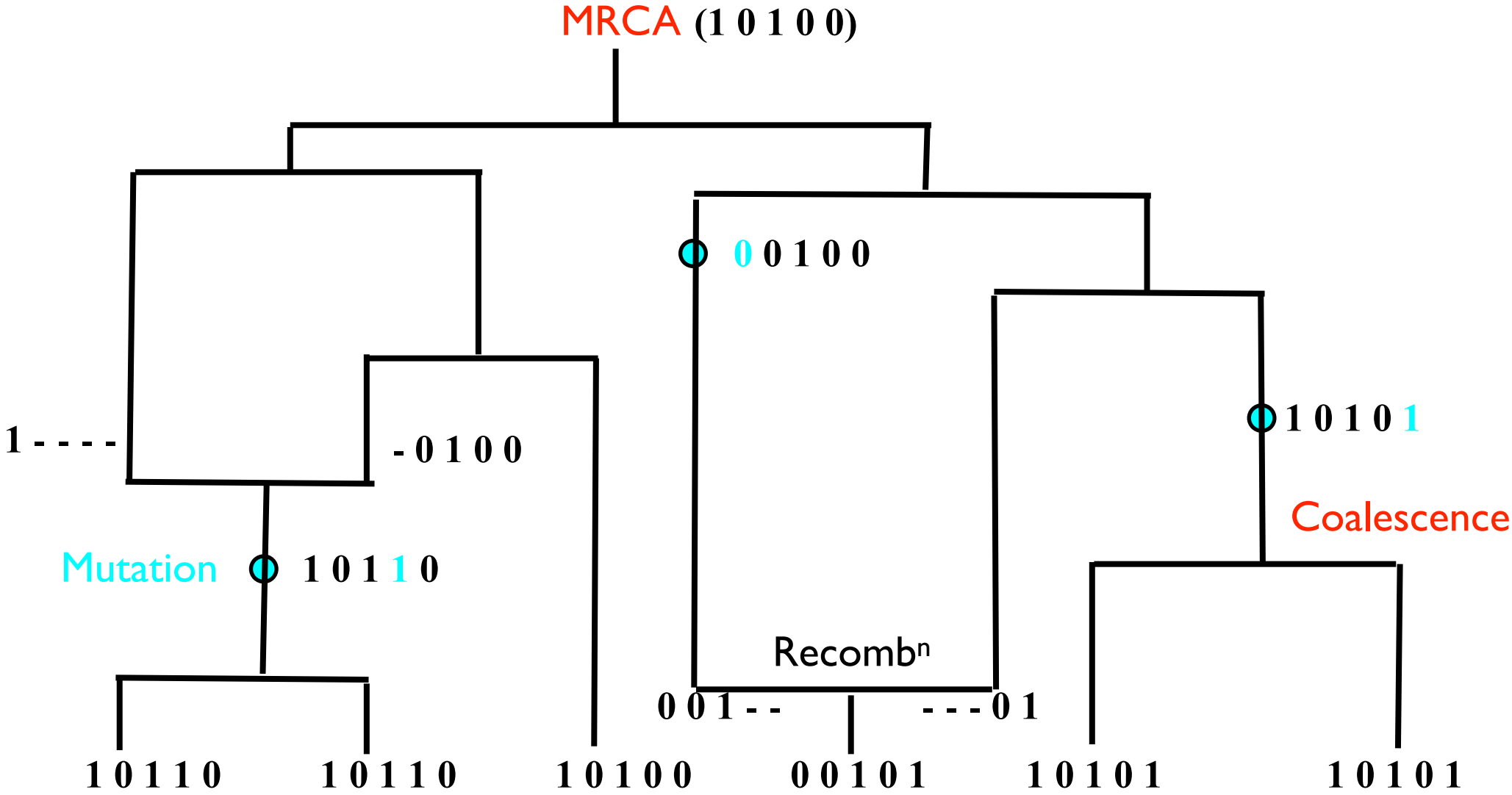


Figure 5: Representation of an ancestry for markers subject to recombination



We trace the ancestry of a sample of 6 marker sequences, until we reach the MRCA. Mutational events are marked in blue. (Markers not ancestral to the sample are marked '-') Not all mutations on this **graph** will appear in the final sample.

References

Coalescent Theory, M. Nordborg:
https://onlinelibrary.wiley.com/doi/abs/10.1002/0470022620.bbc21

Coalescent Theory: An Introduction, J. Wakeley (2009).
https://www.amazon.com/Coalescent-Theory-Introduction-John-Wakeley/dp/0974707759/ref=sr_1_1?crid=AK4KGA9B24JO&keywords=coalescent+theory+wakeley&qid=1582304288&sprefix=coalescent+theory%2Caps%2C207&sr=8-1

Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. P. Donnelly (1986).
https://www.sciencedirect.com/science/article/pii/0040580986900377

END