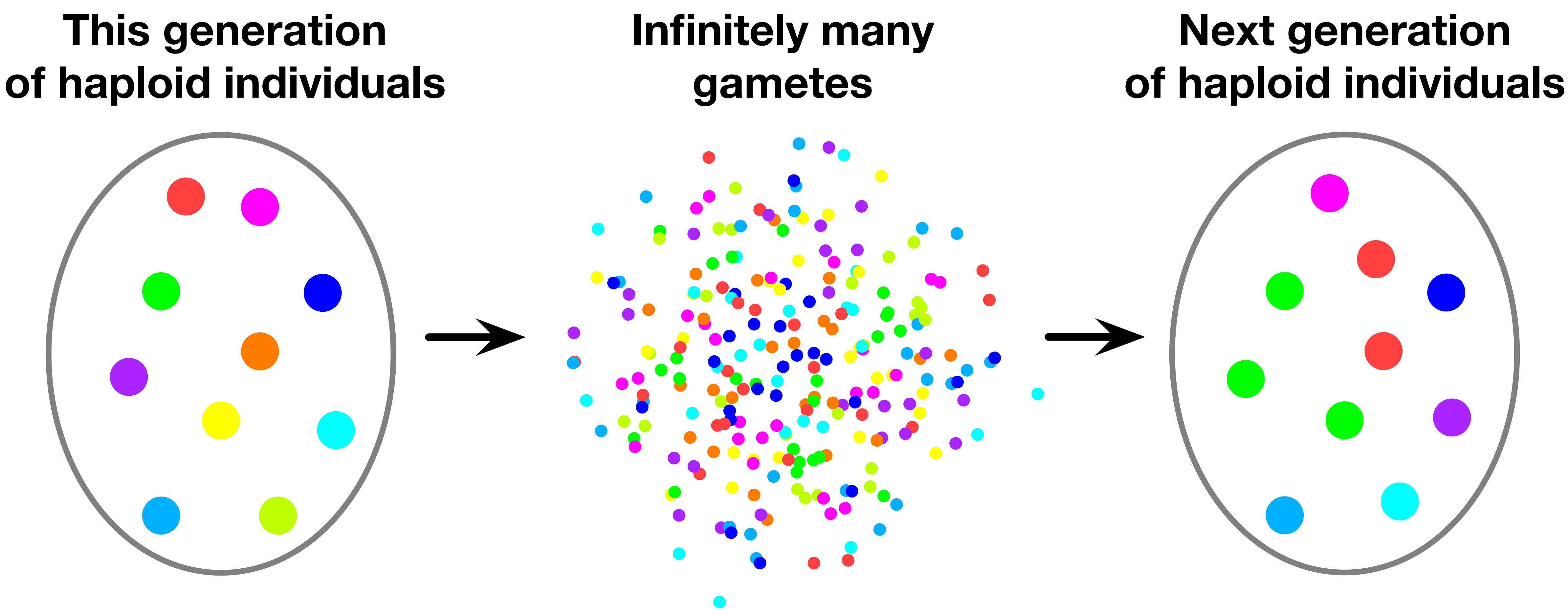
The coalescent

Week 2

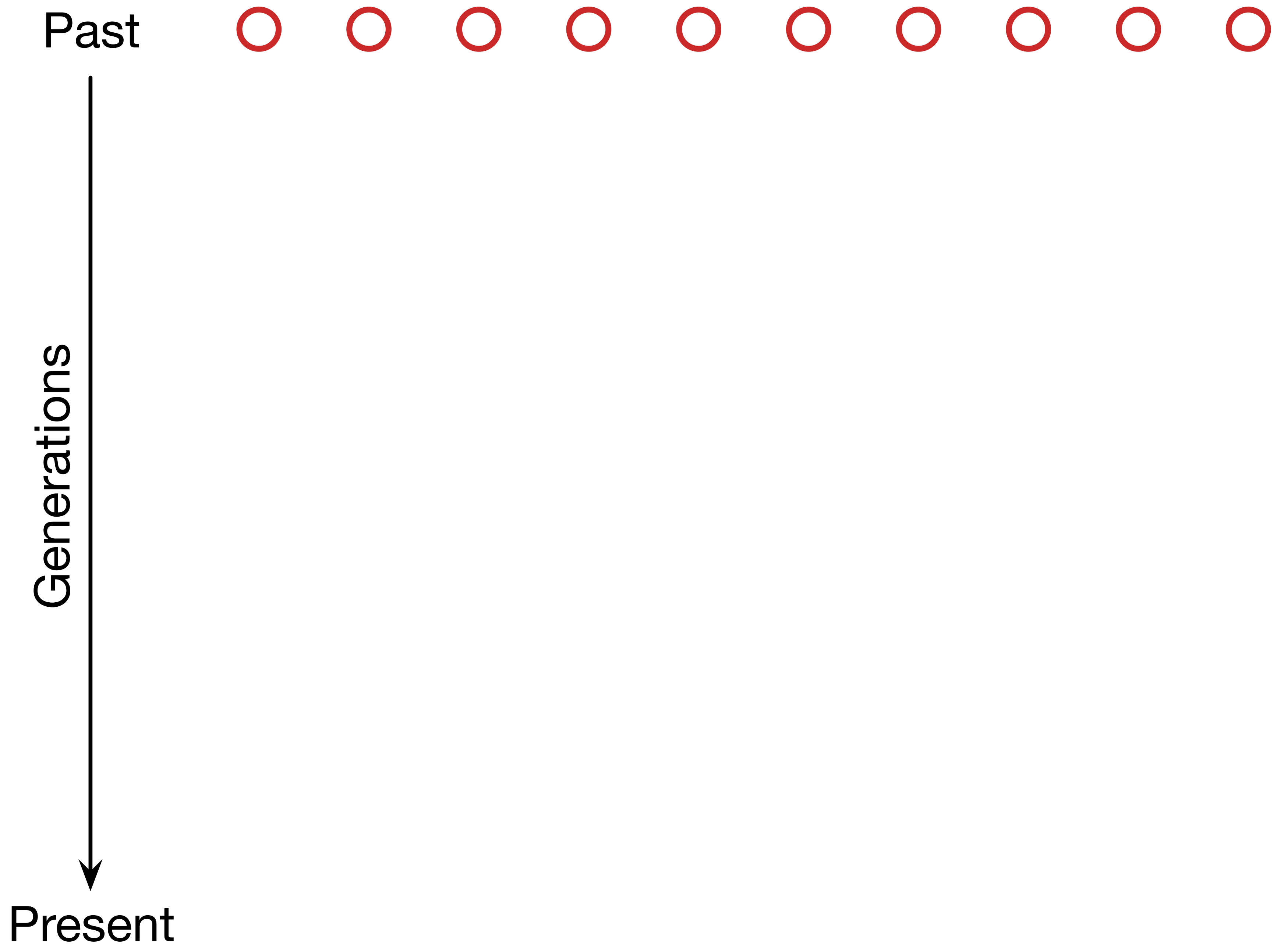
Kasper Munch

Popgen basics

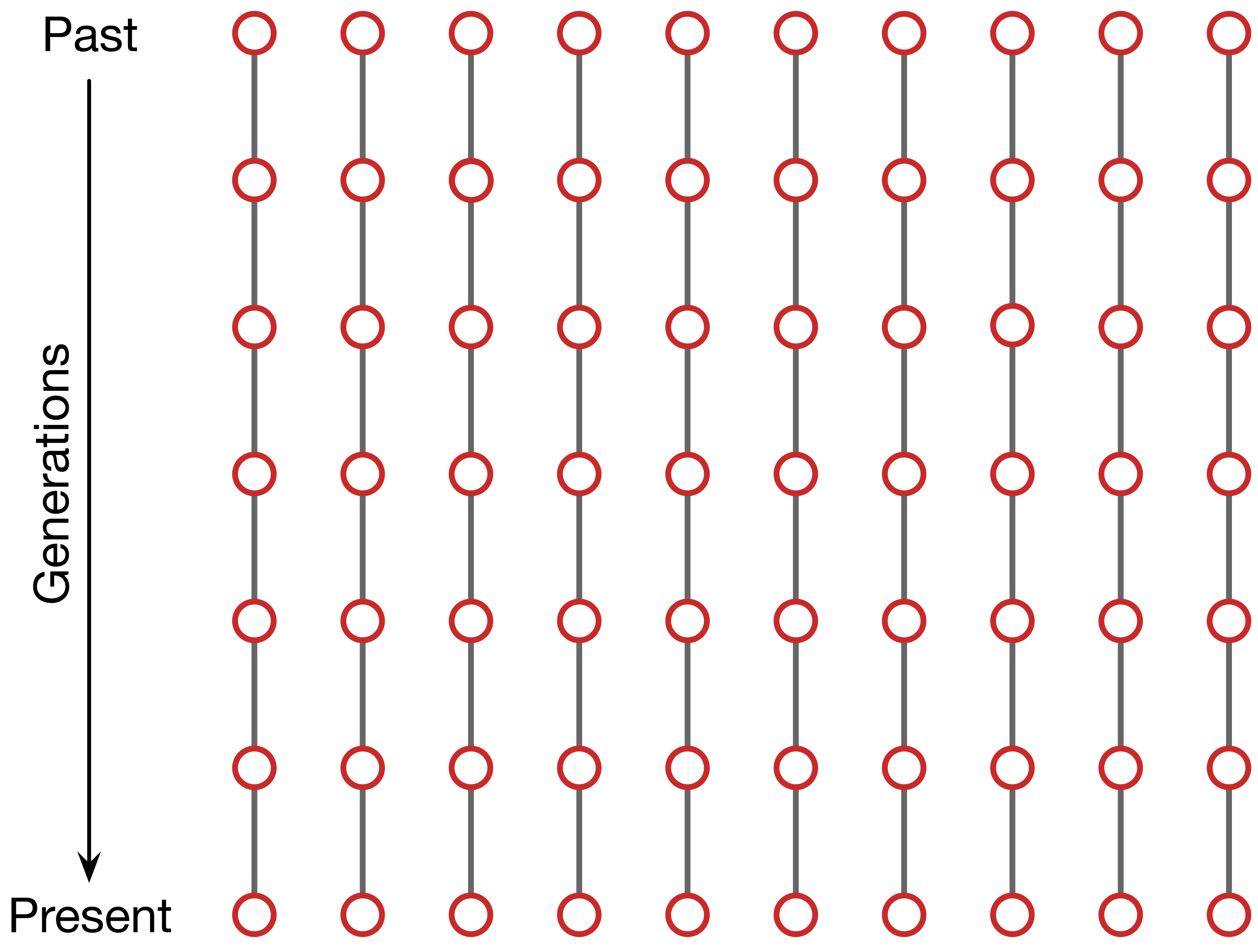
Wright-Fisher model



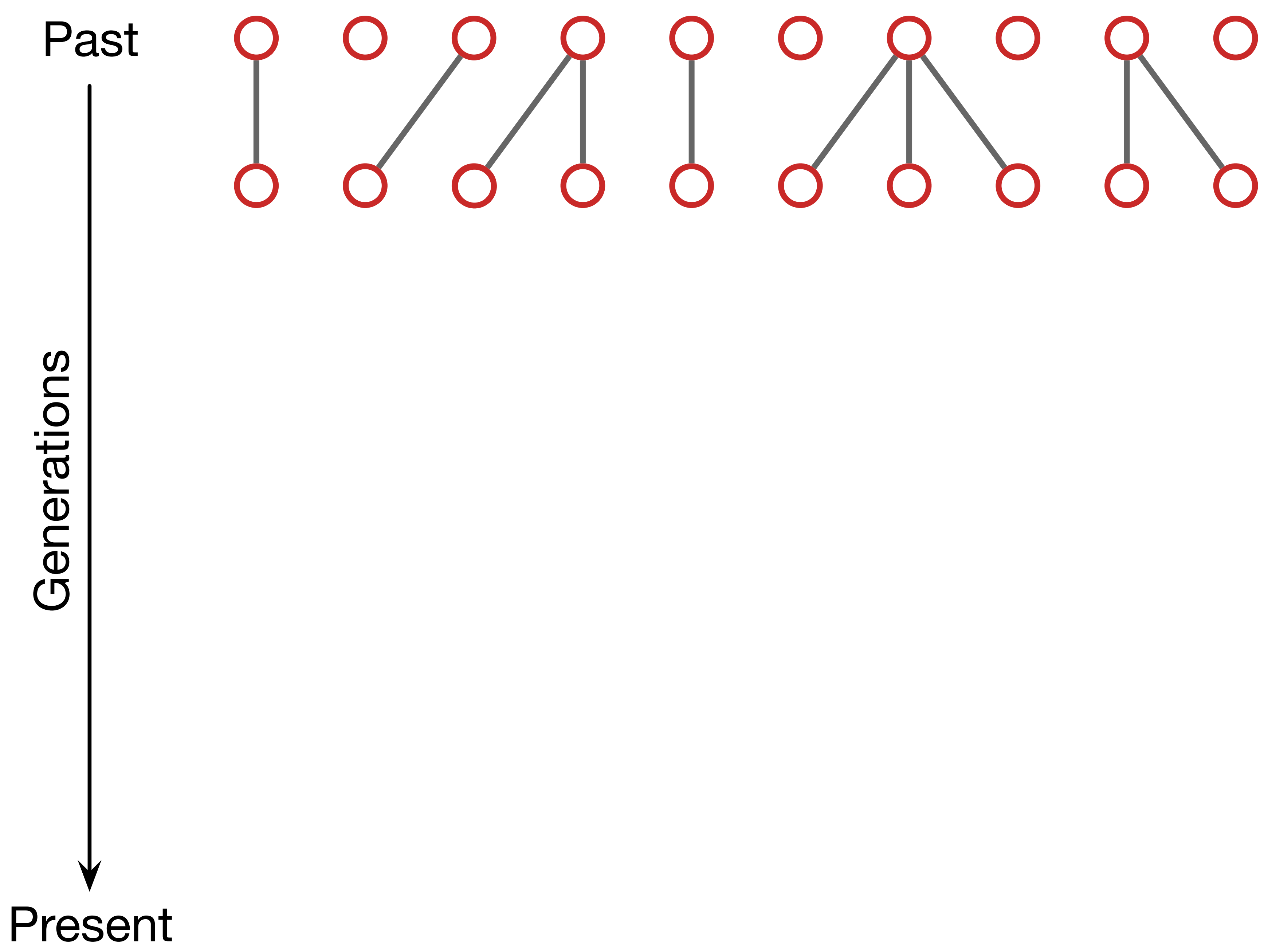
Propagation of genes



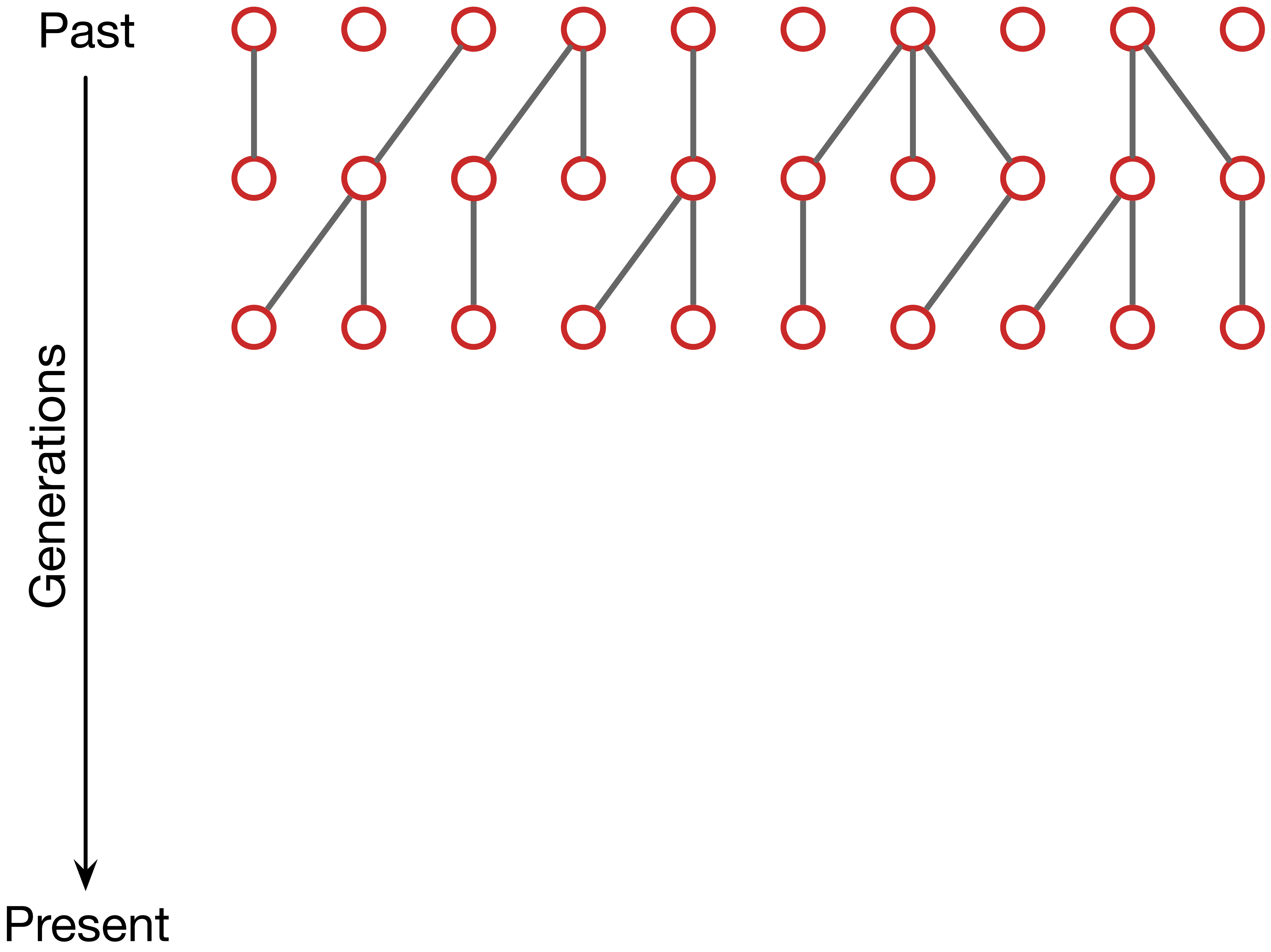
Propagation of genes



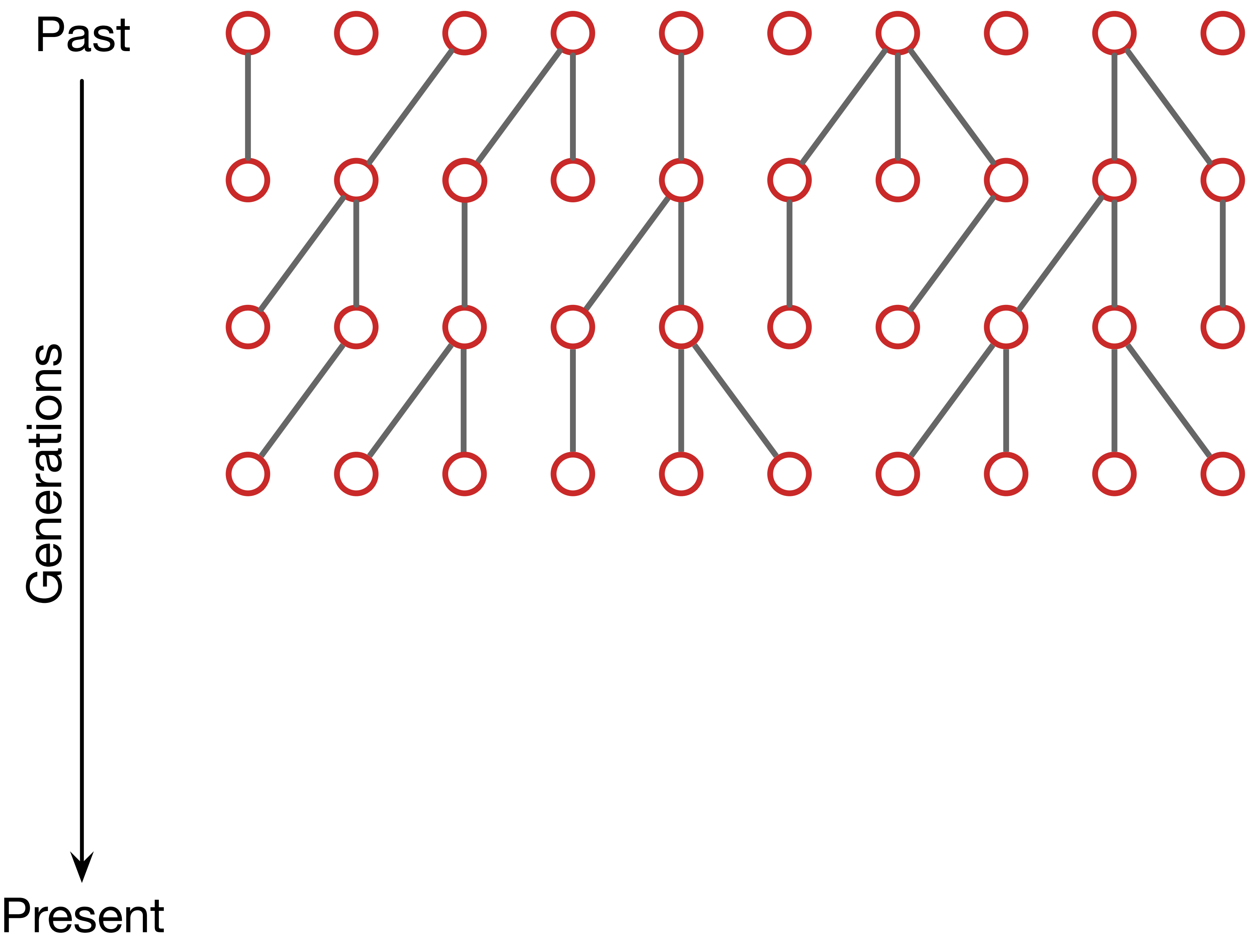
Propagation of genes



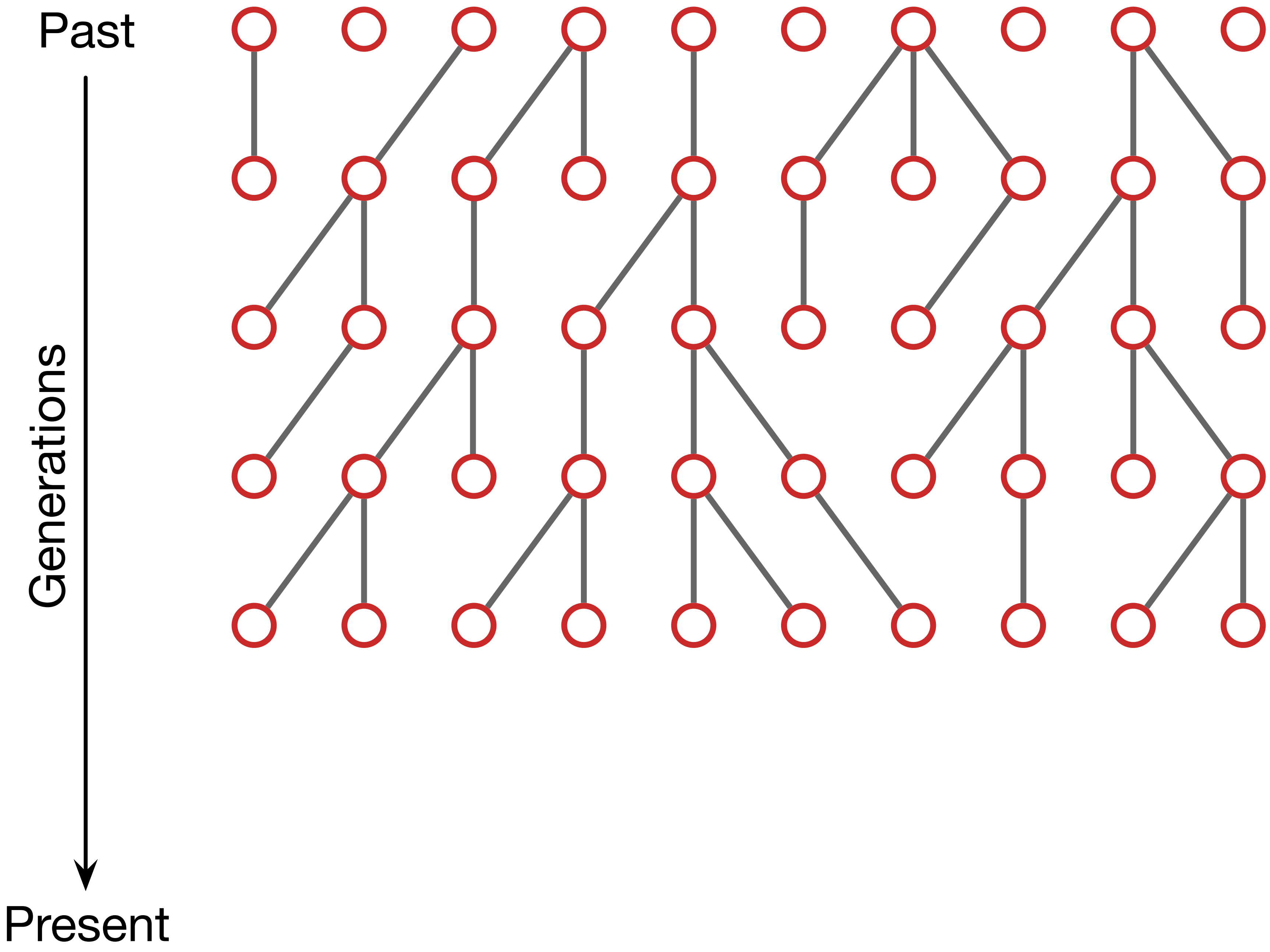
Propagation of genes



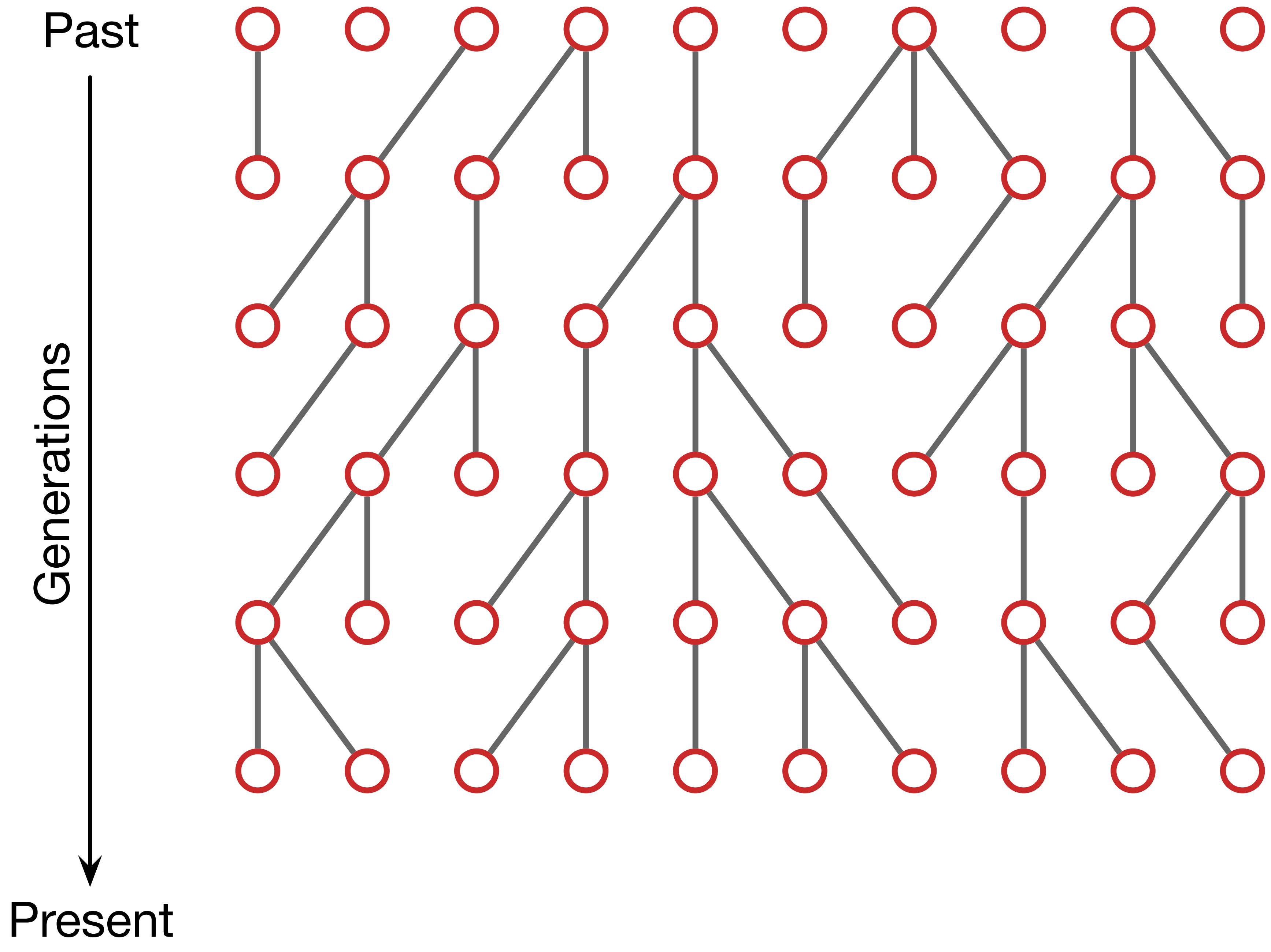
Propagation of genes



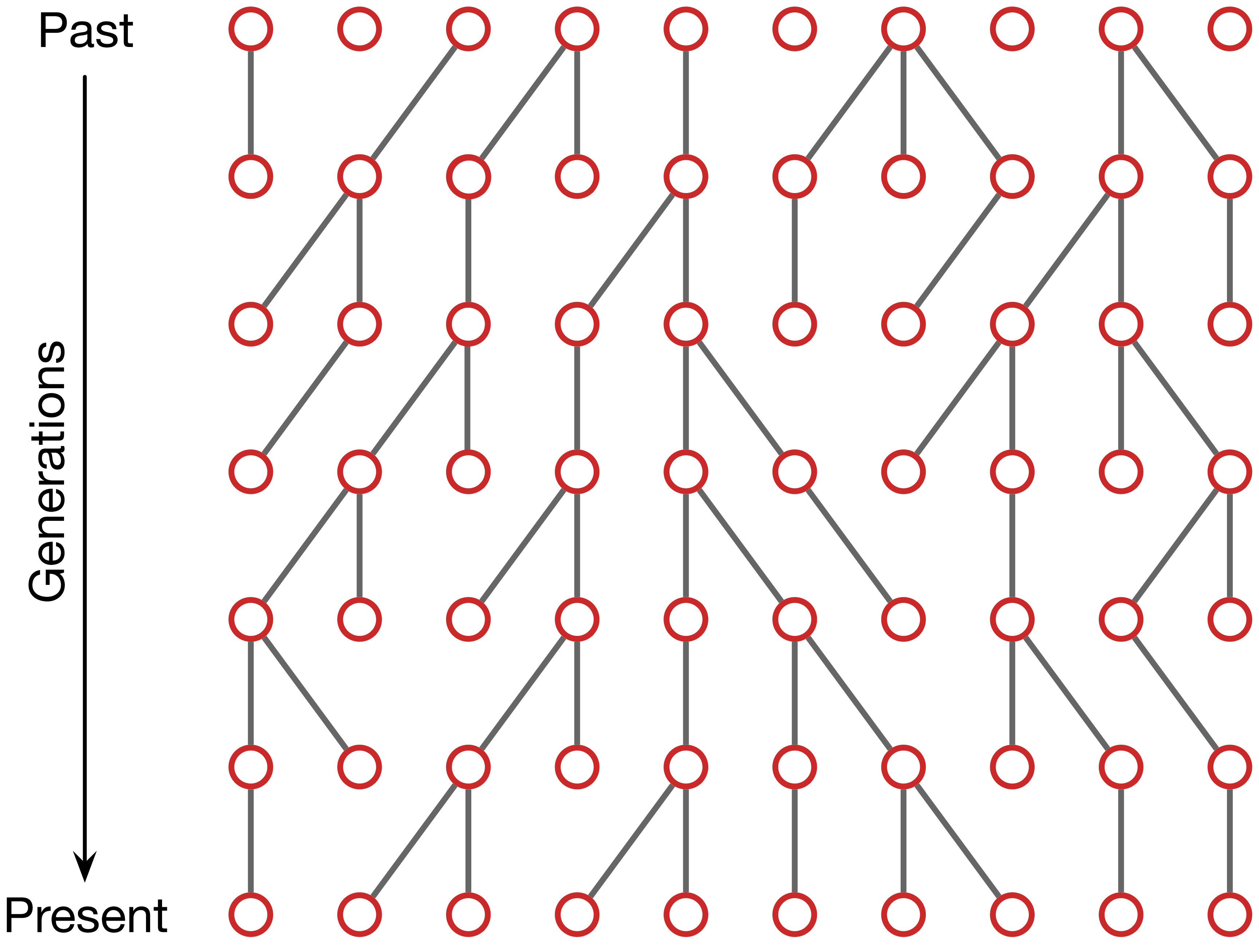
Propagation of genes



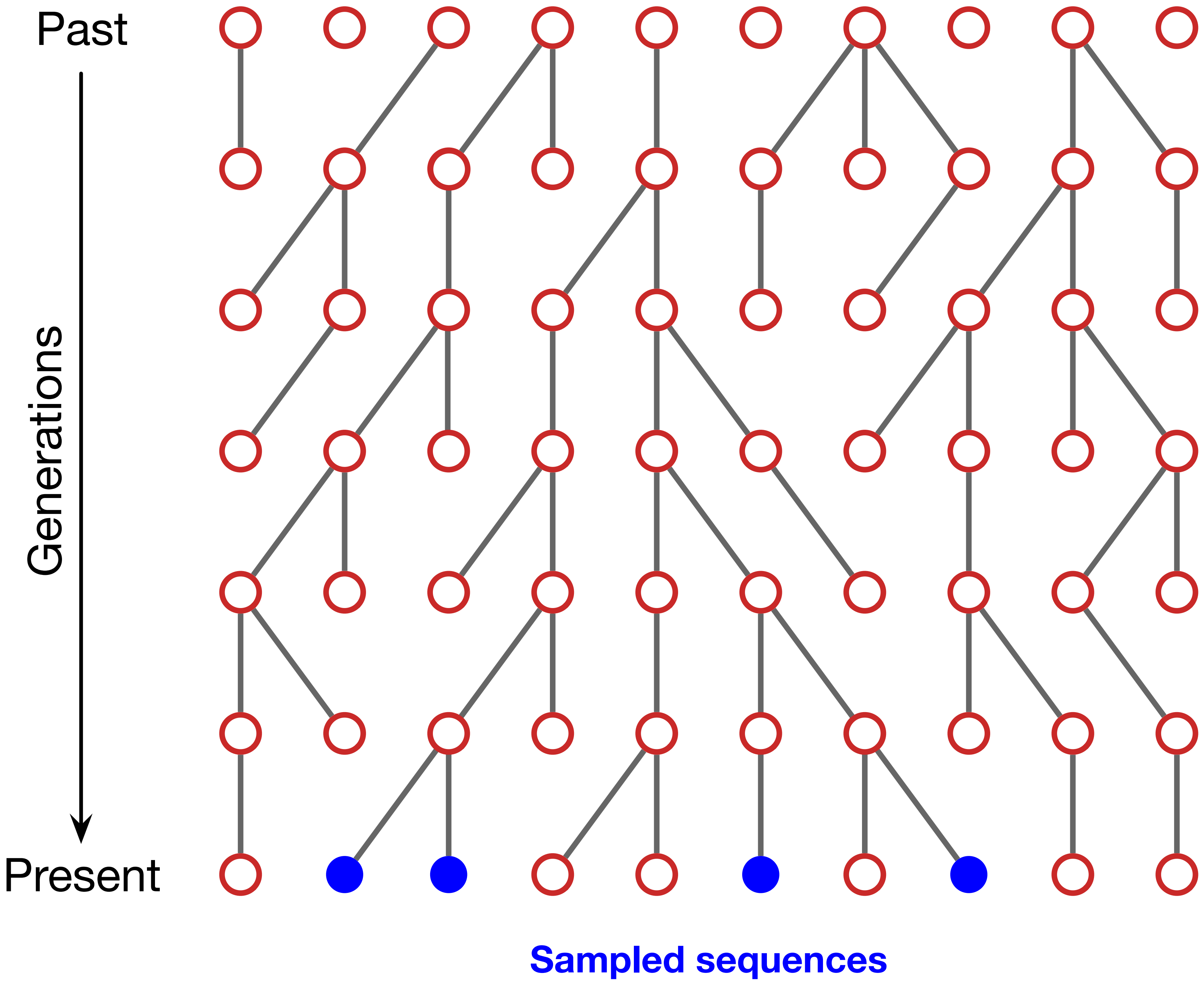
Propagation of genes



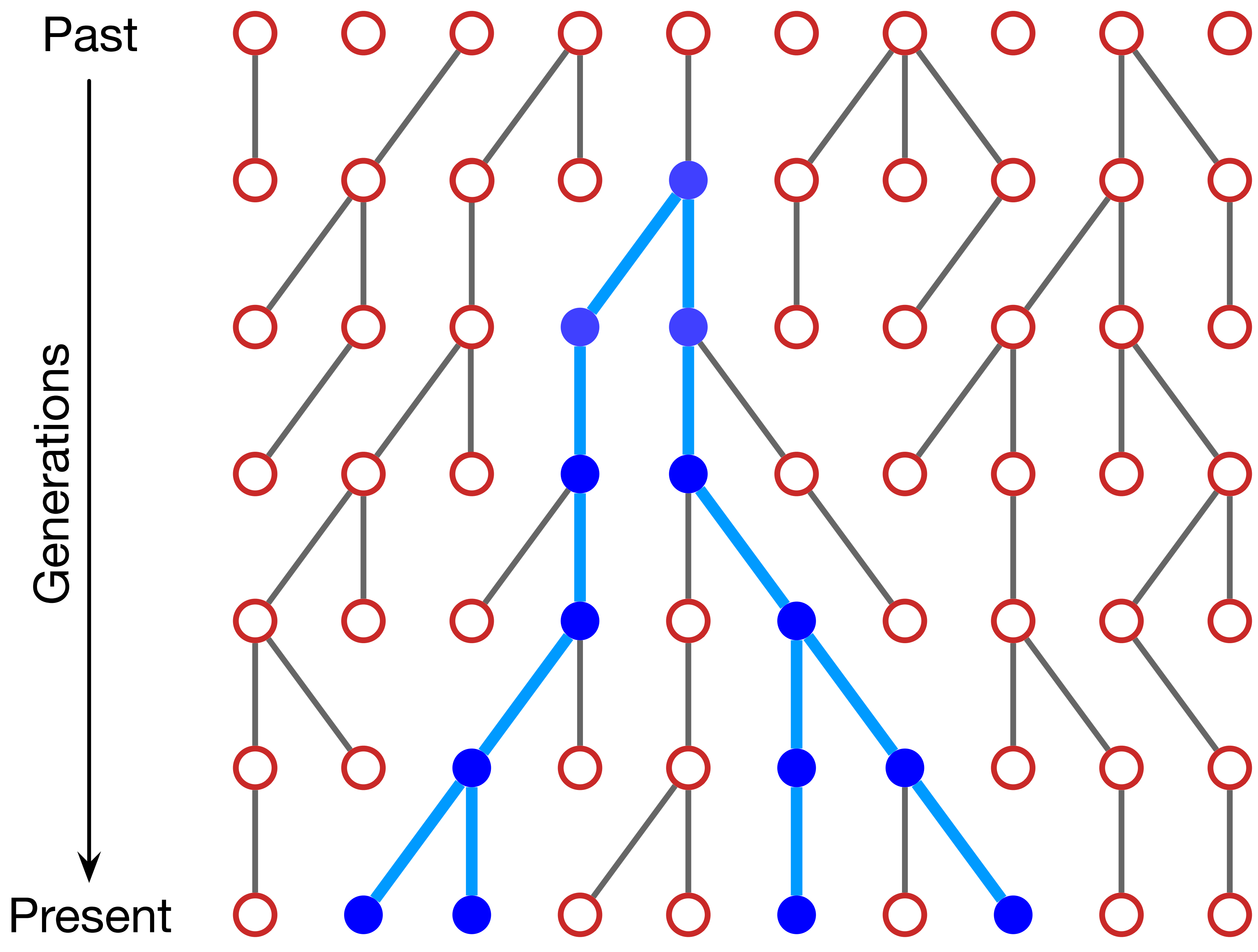
Propagation of genes



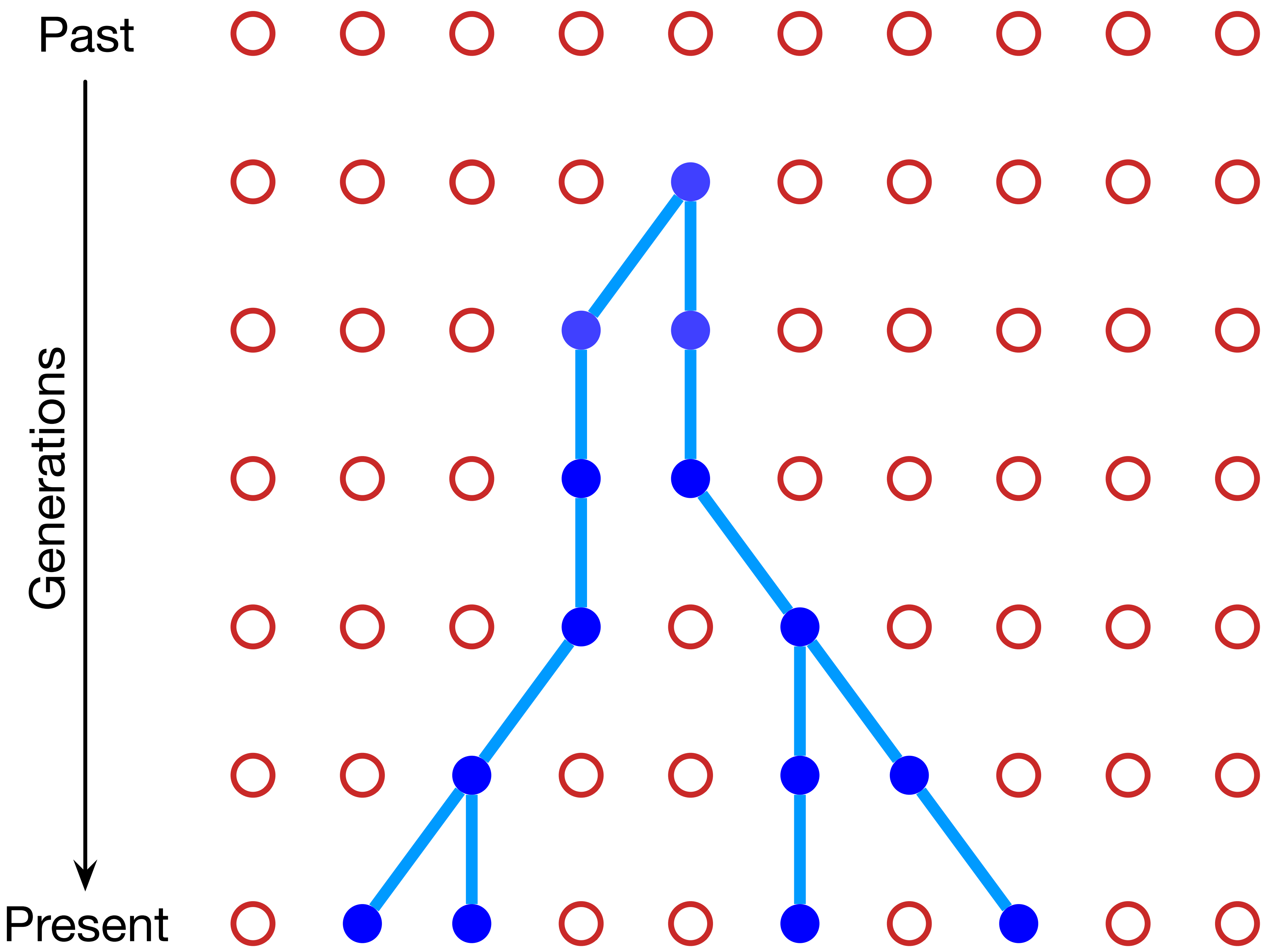
The coalescent



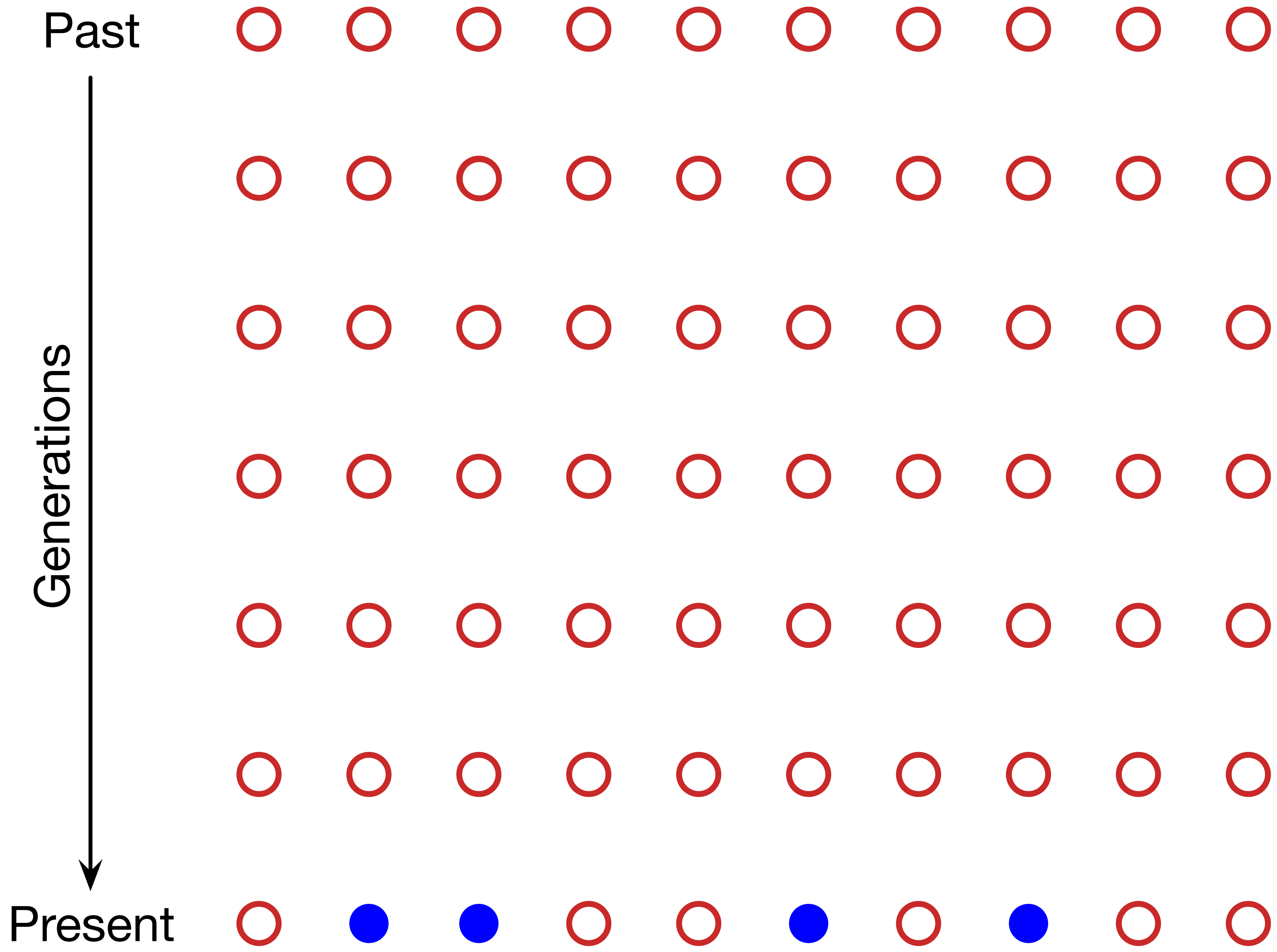
The coalescent



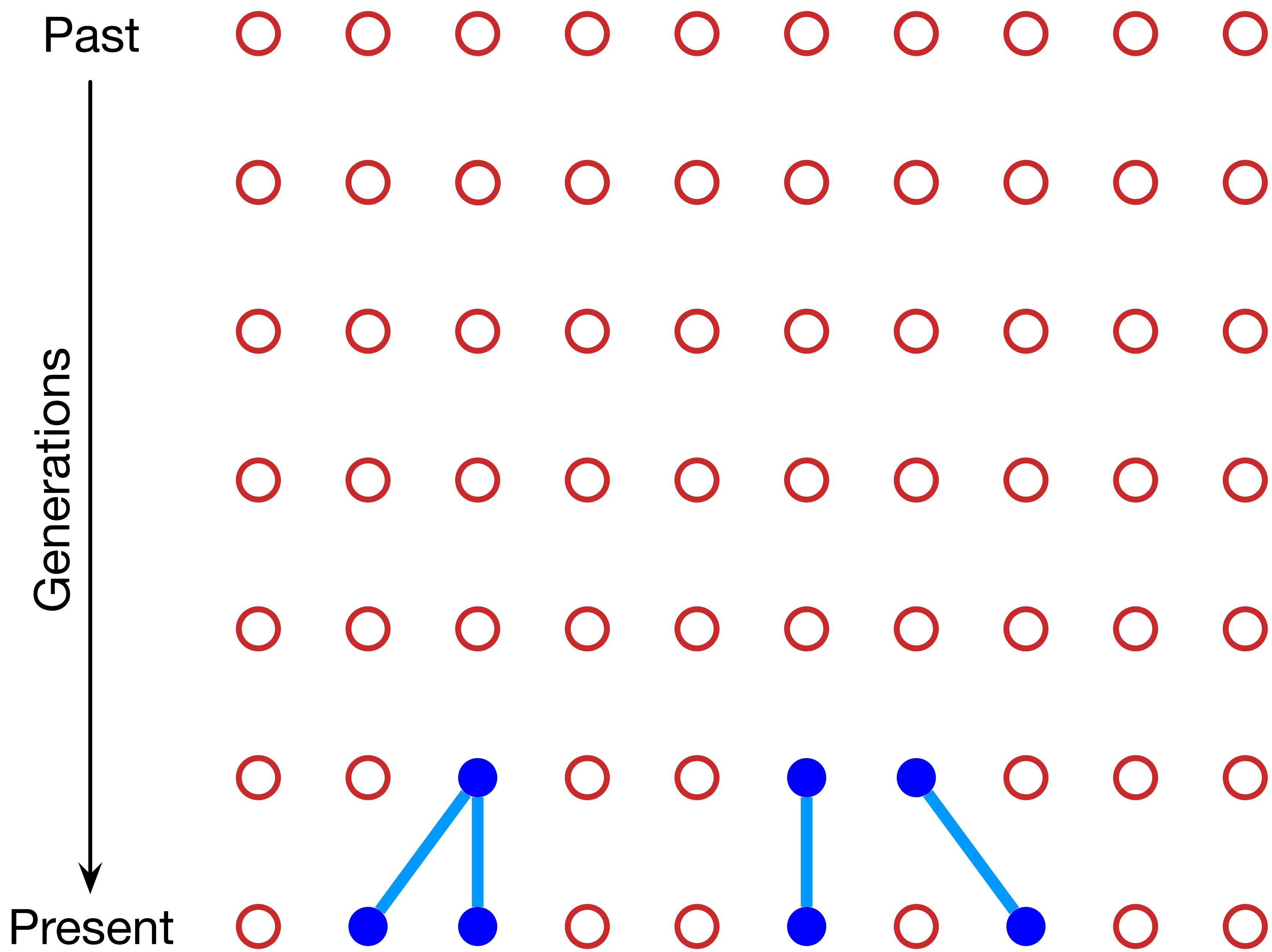
The coalescent



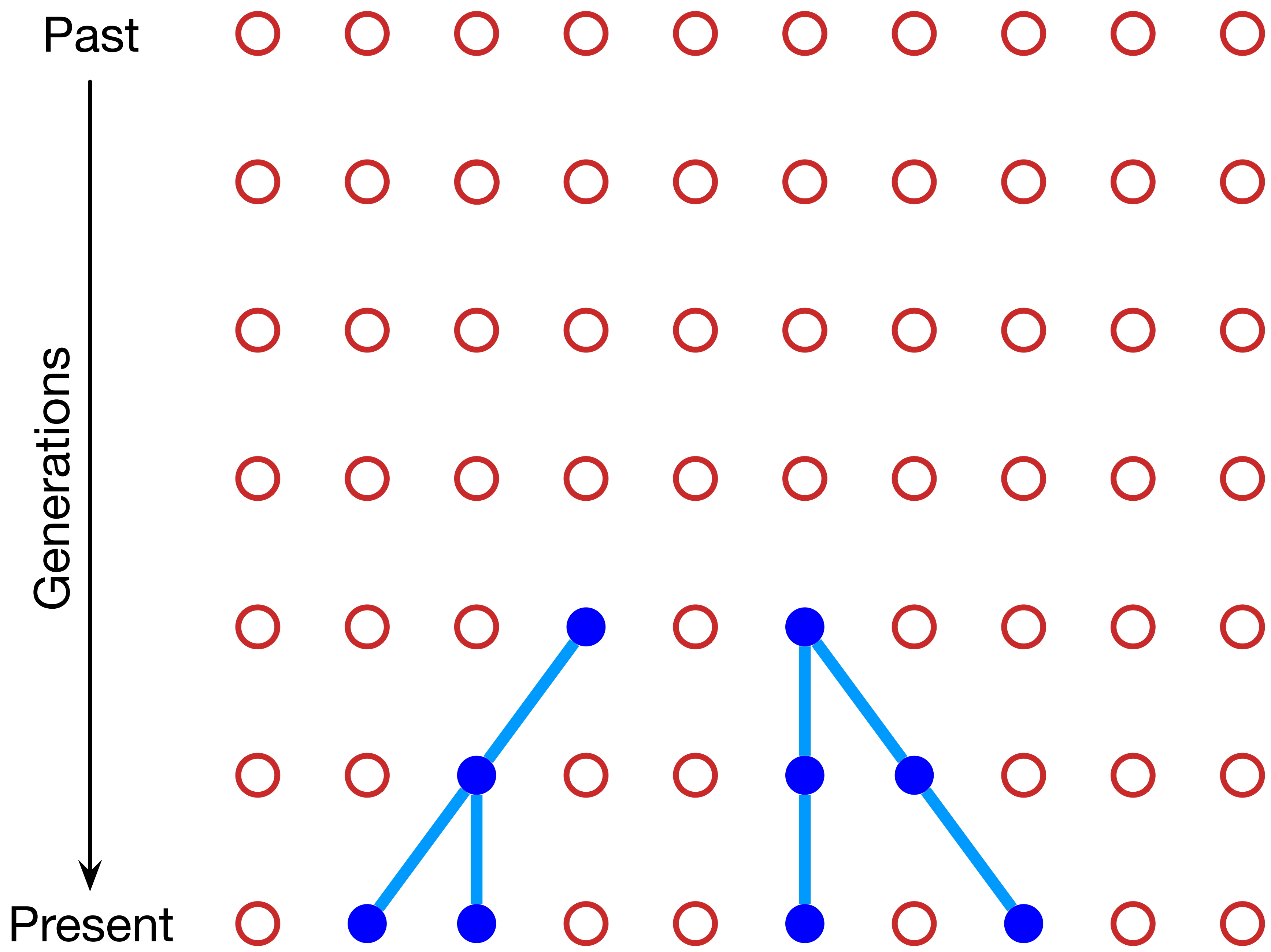
The coalescent



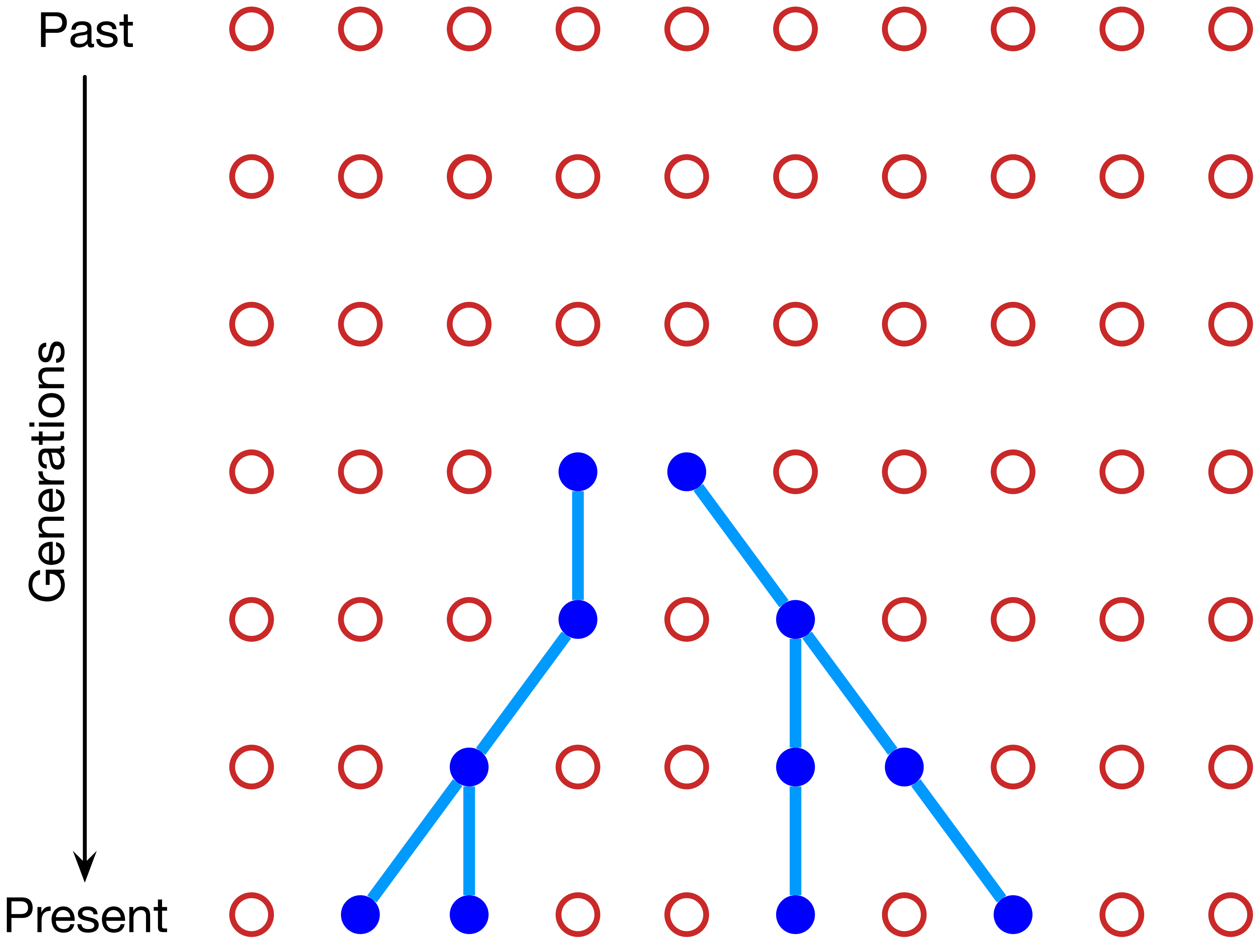
The coalescent



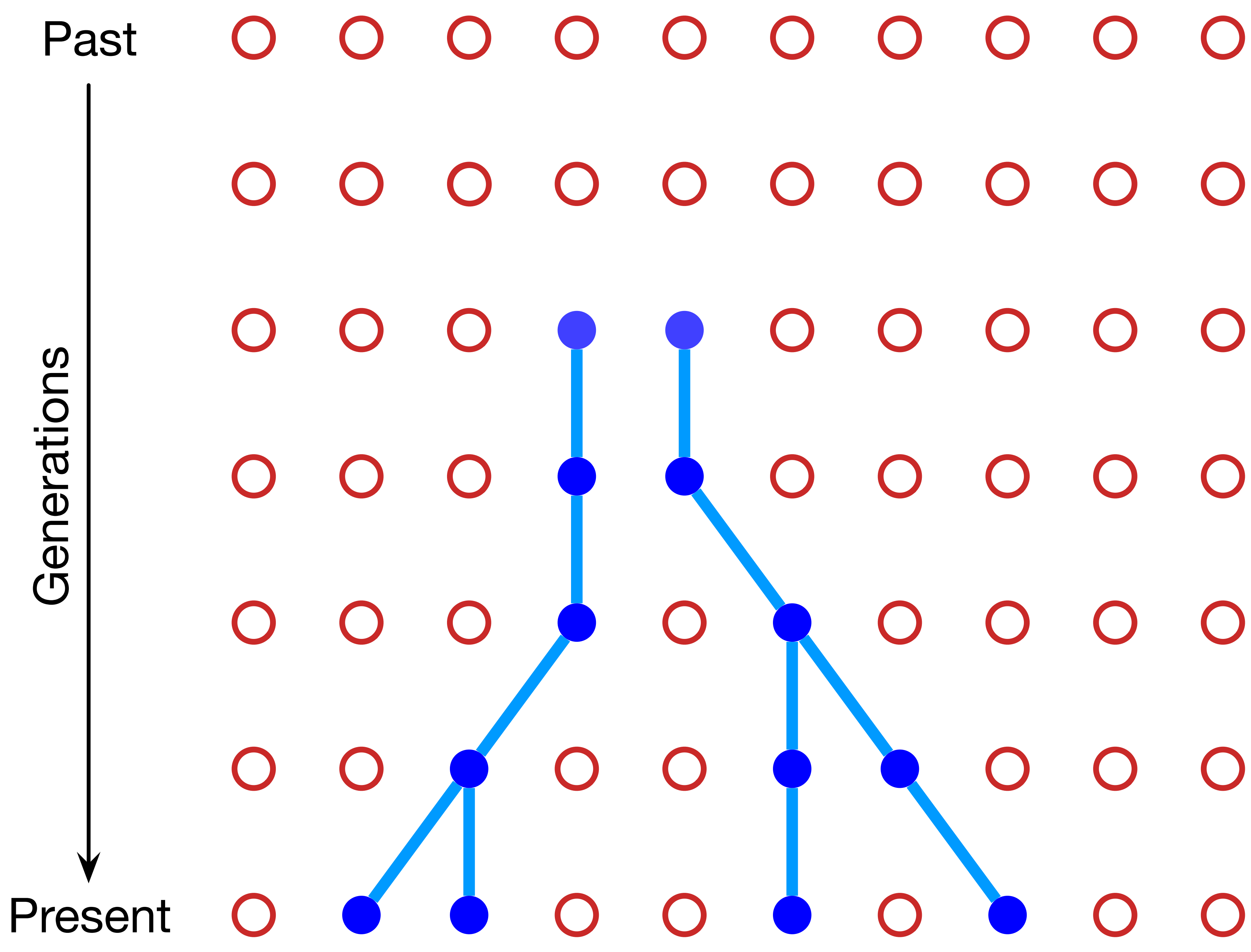
The coalescent



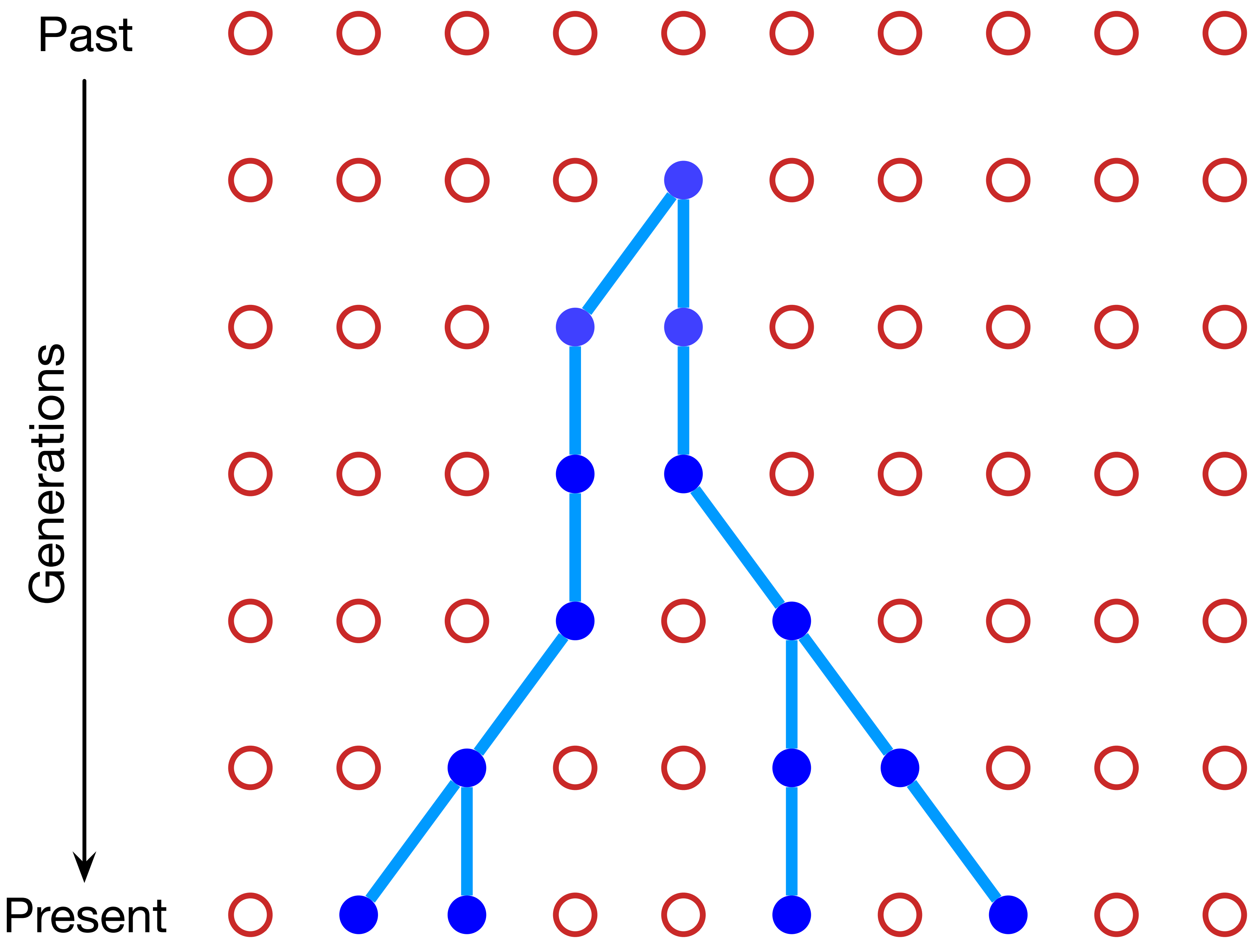
The coalescent



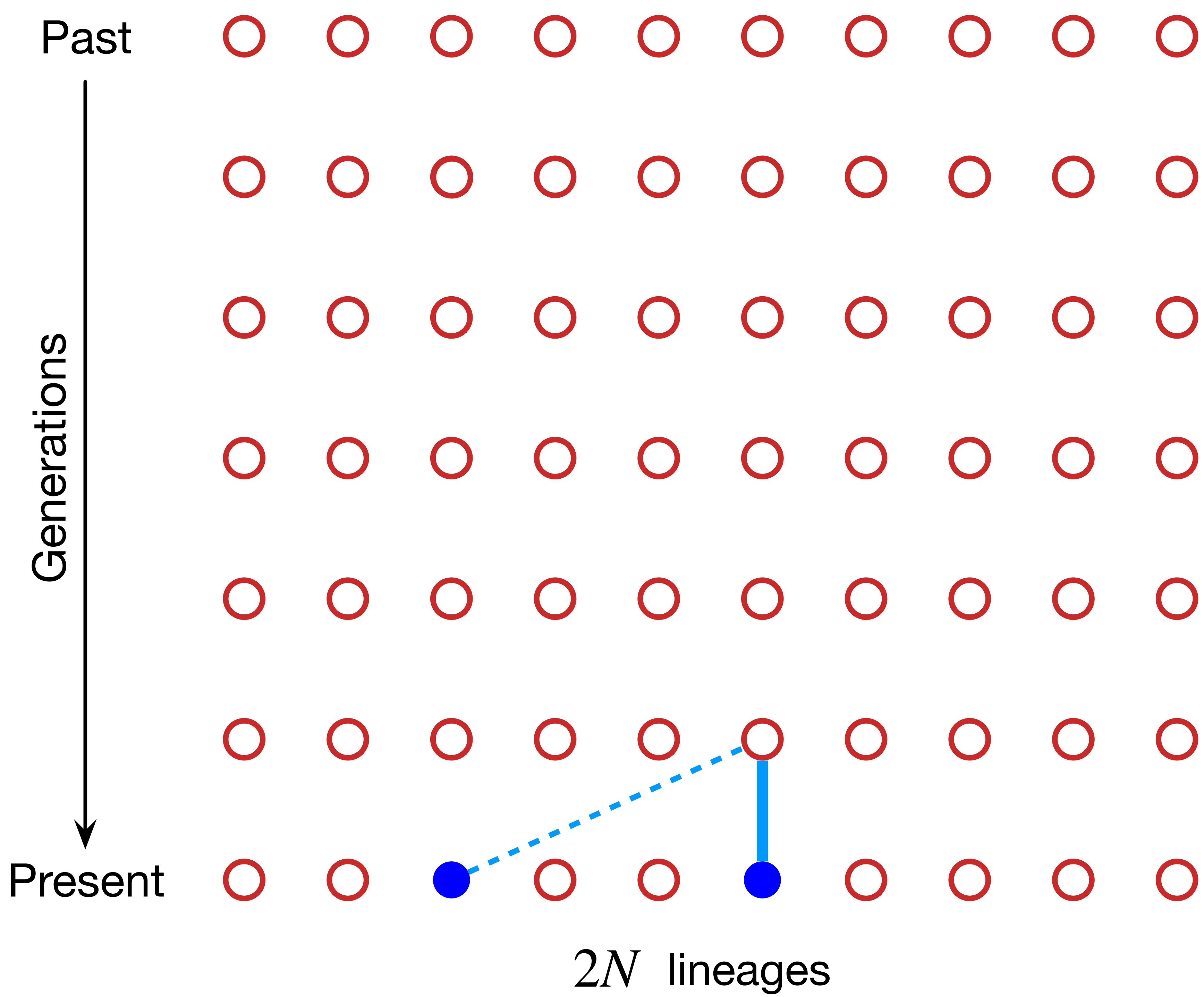
The coalescent



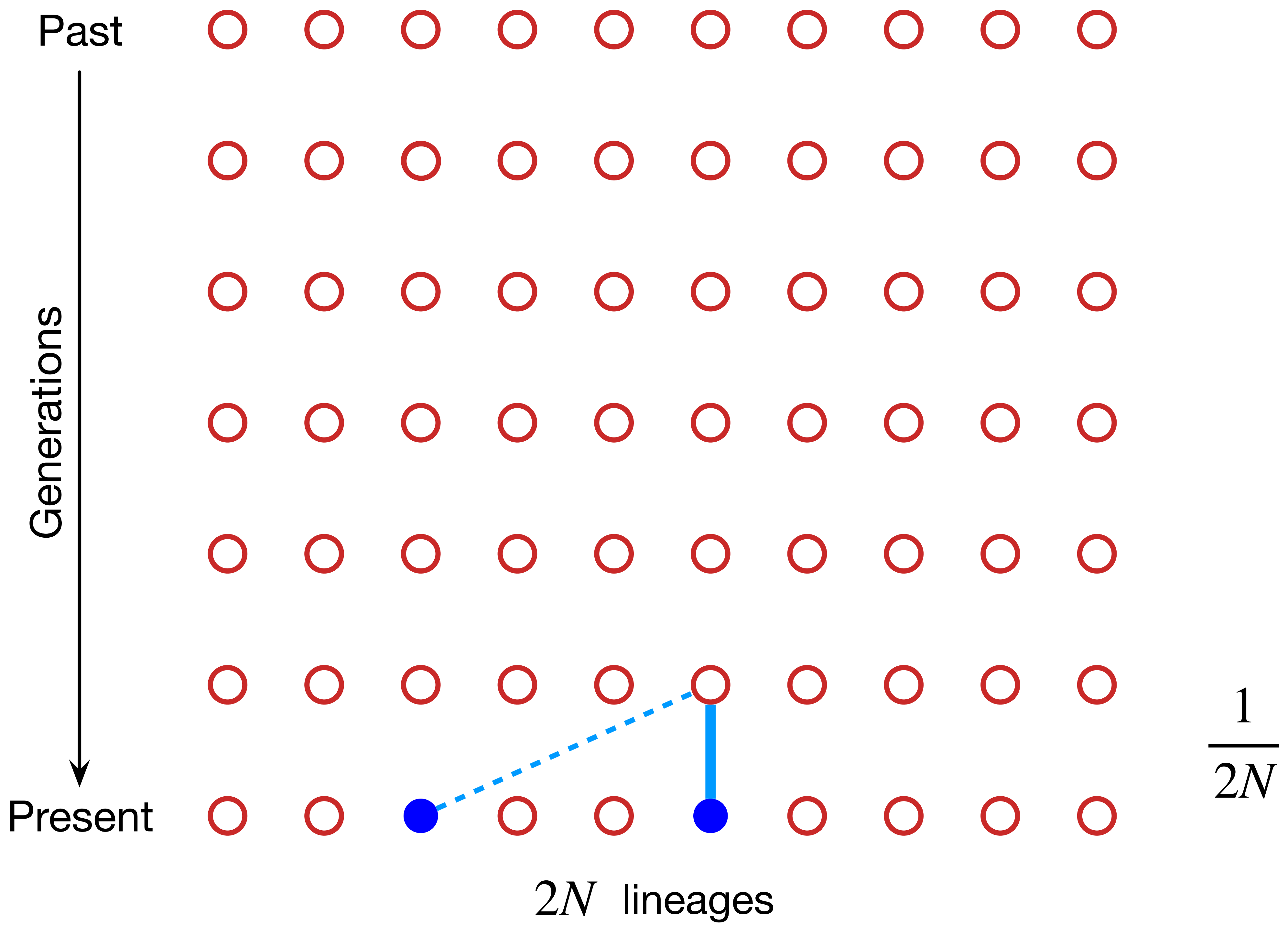
The coalescent



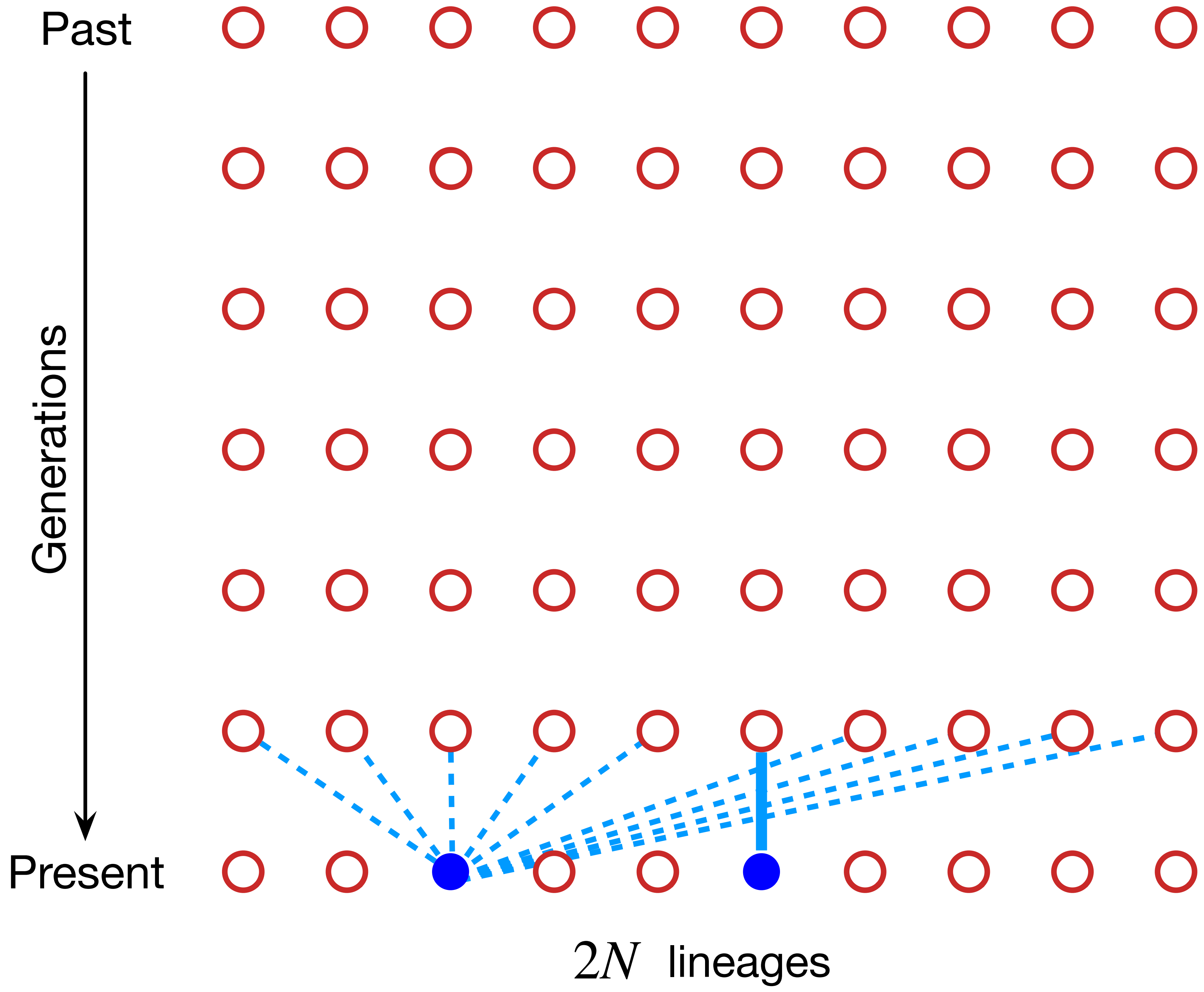
The coalescent



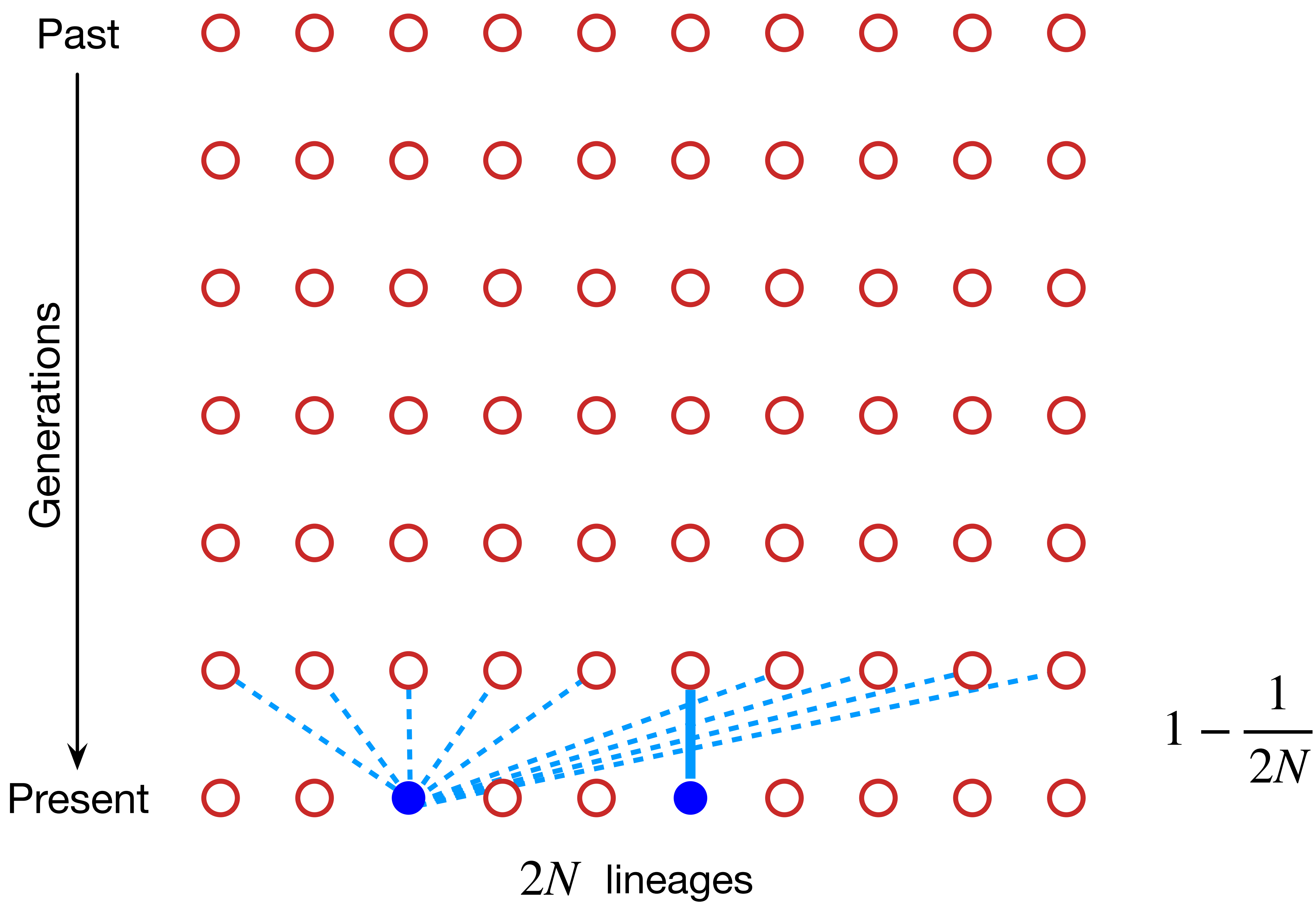
The coalescent



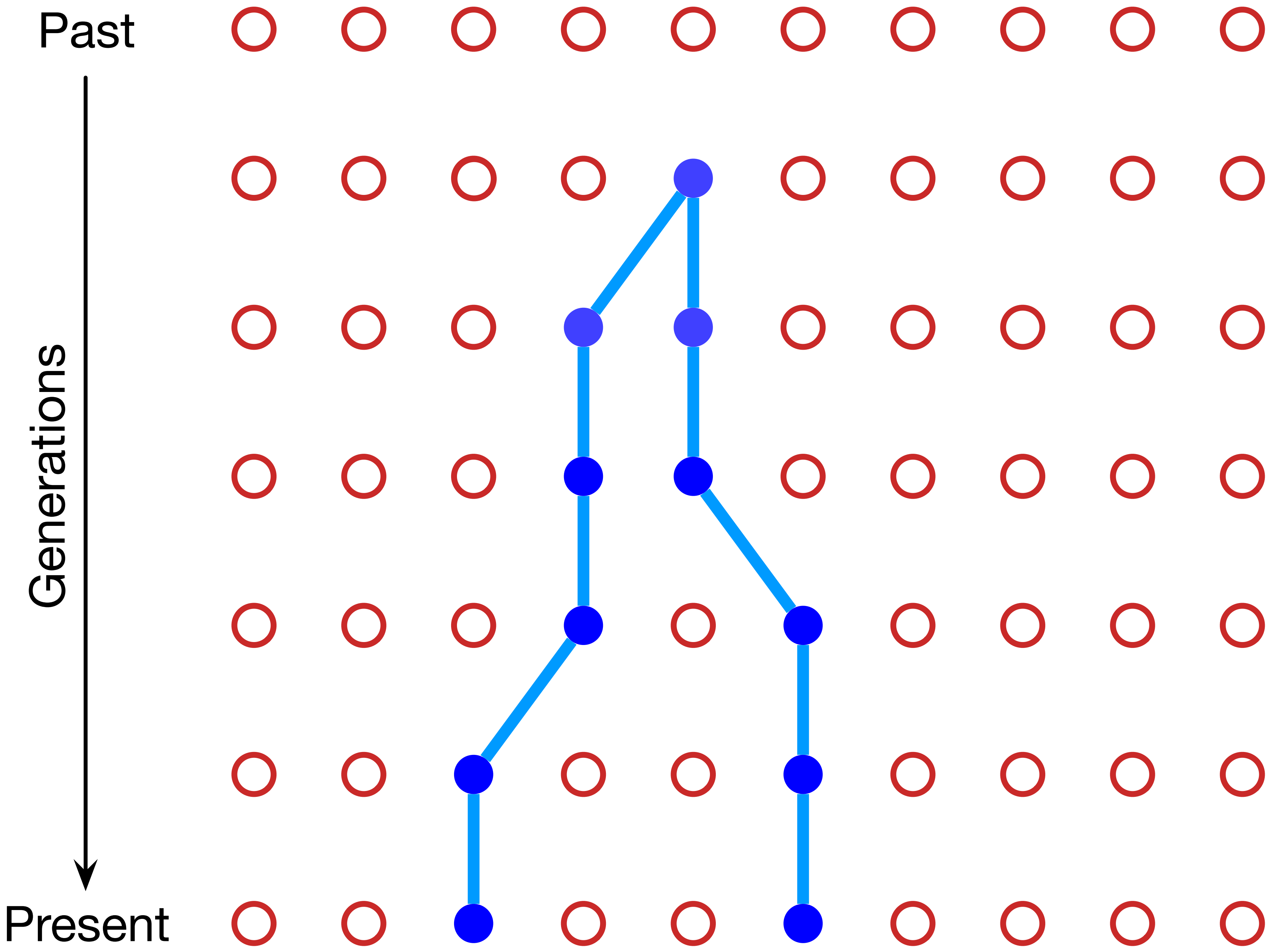
The coalescent



The coalescent



The coalescent



$$\frac{1}{2N}$$

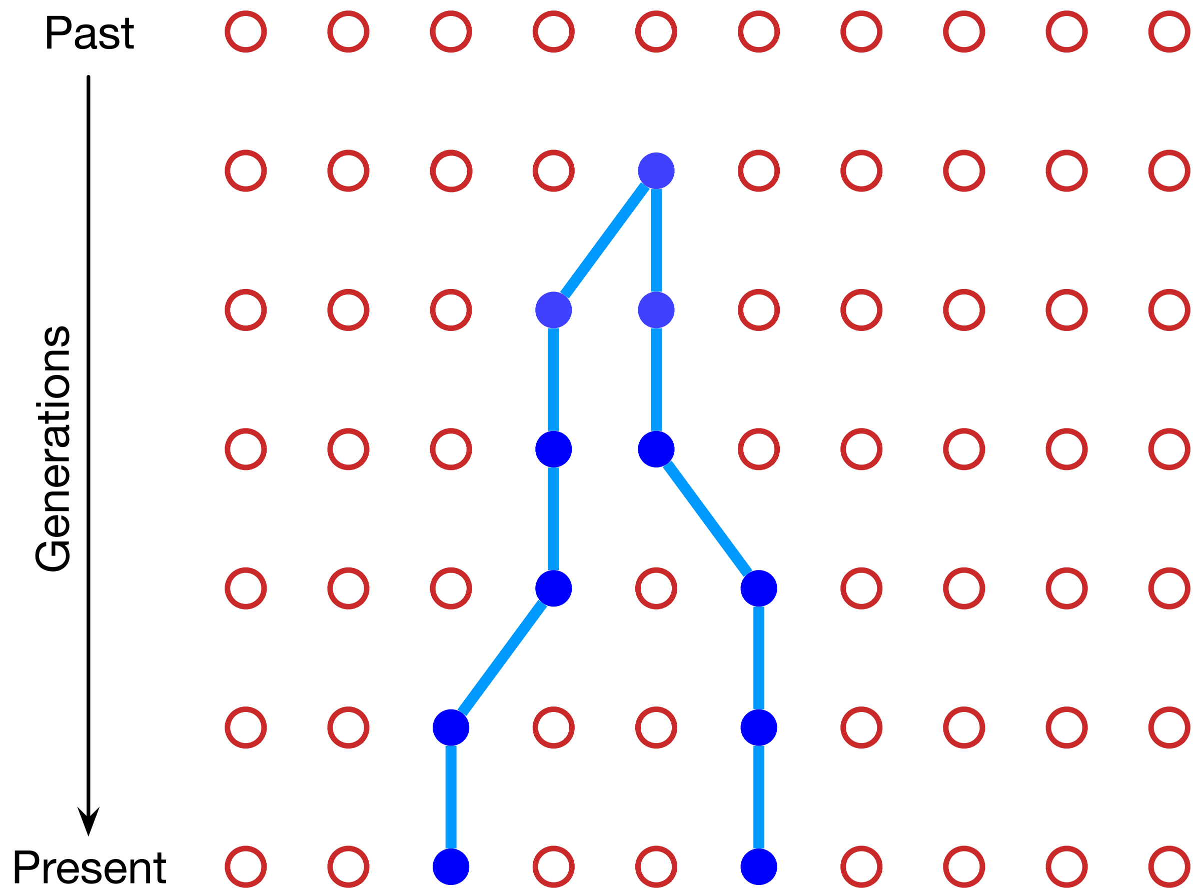
$$1 - \frac{1}{2N}$$

$$1 - \frac{1}{2N}$$

$$1 - \frac{1}{2N}$$

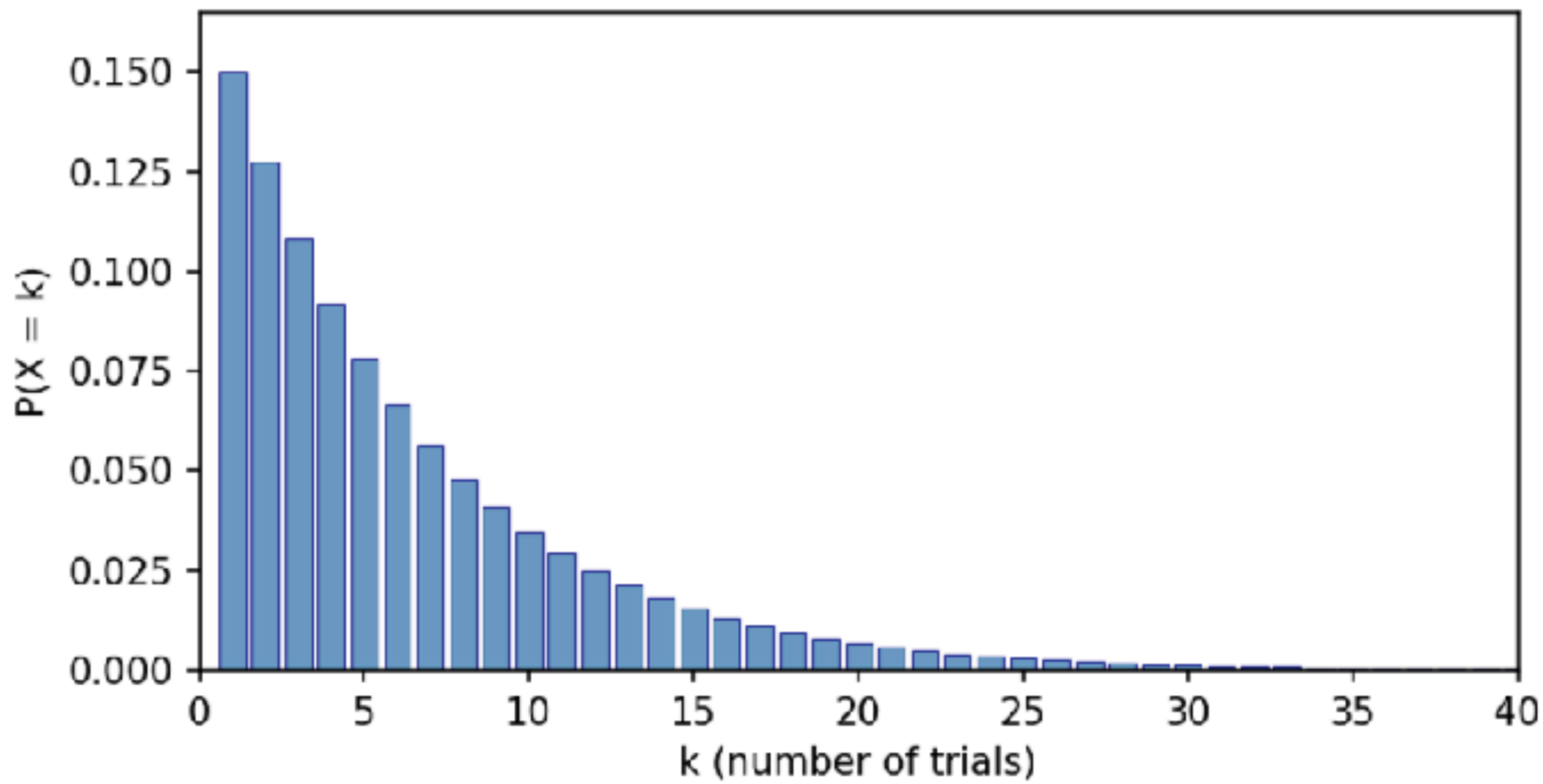
$$1 - \frac{1}{2N}$$

The coalescent



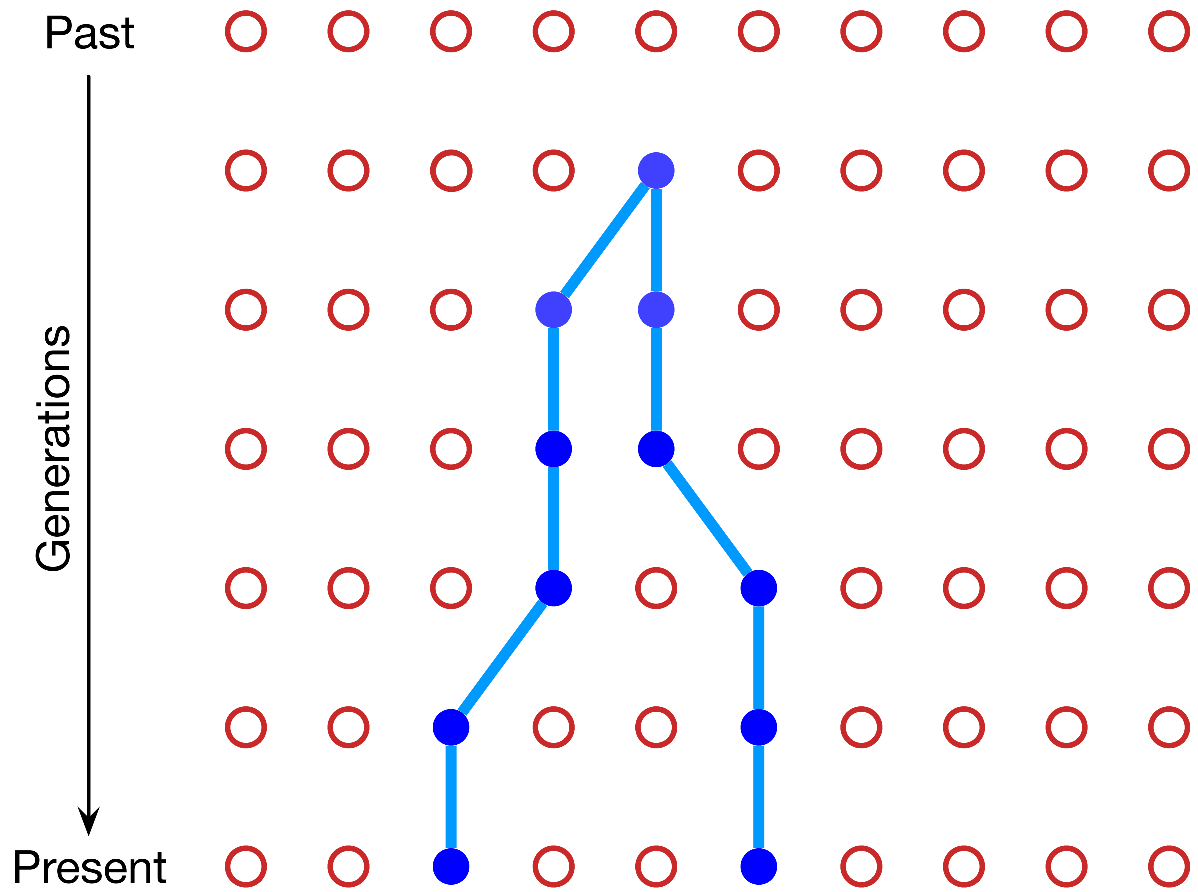
$$\mathbb{P}(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1}\left(\frac{1}{2N}\right)$$

The coalescent



$$\mathbb{P}(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right)$$

The coalescent



$$\mathbb{P}(k_{t+1} = 1 \mid k_t = 2) = \frac{1}{2N}$$

$$\mathbb{E}(T_2) = 2N$$

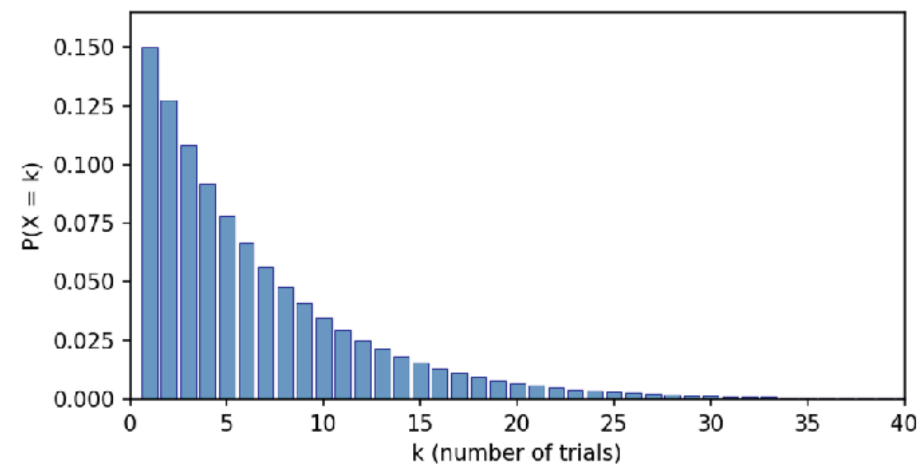
Coalescent

Continuous approximation

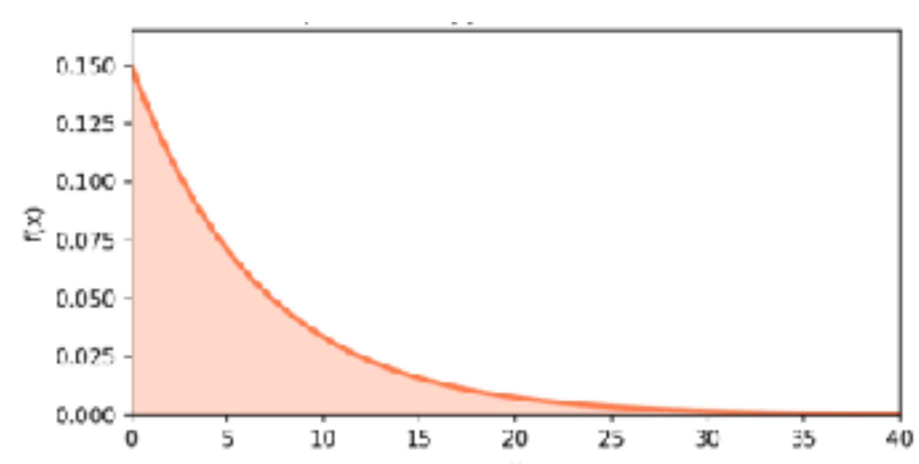
$$\mathbb{P}(T_2 > t) = 1 - \left(1 - \frac{1}{2N}\right)^t$$

Assuming N is very large

$$\mathbb{P}(T_2 > t) = 1 - e^{-t\frac{1}{2N}}$$



Geometric distribution (**discrete**)



Exponential distribution (**continuous**)

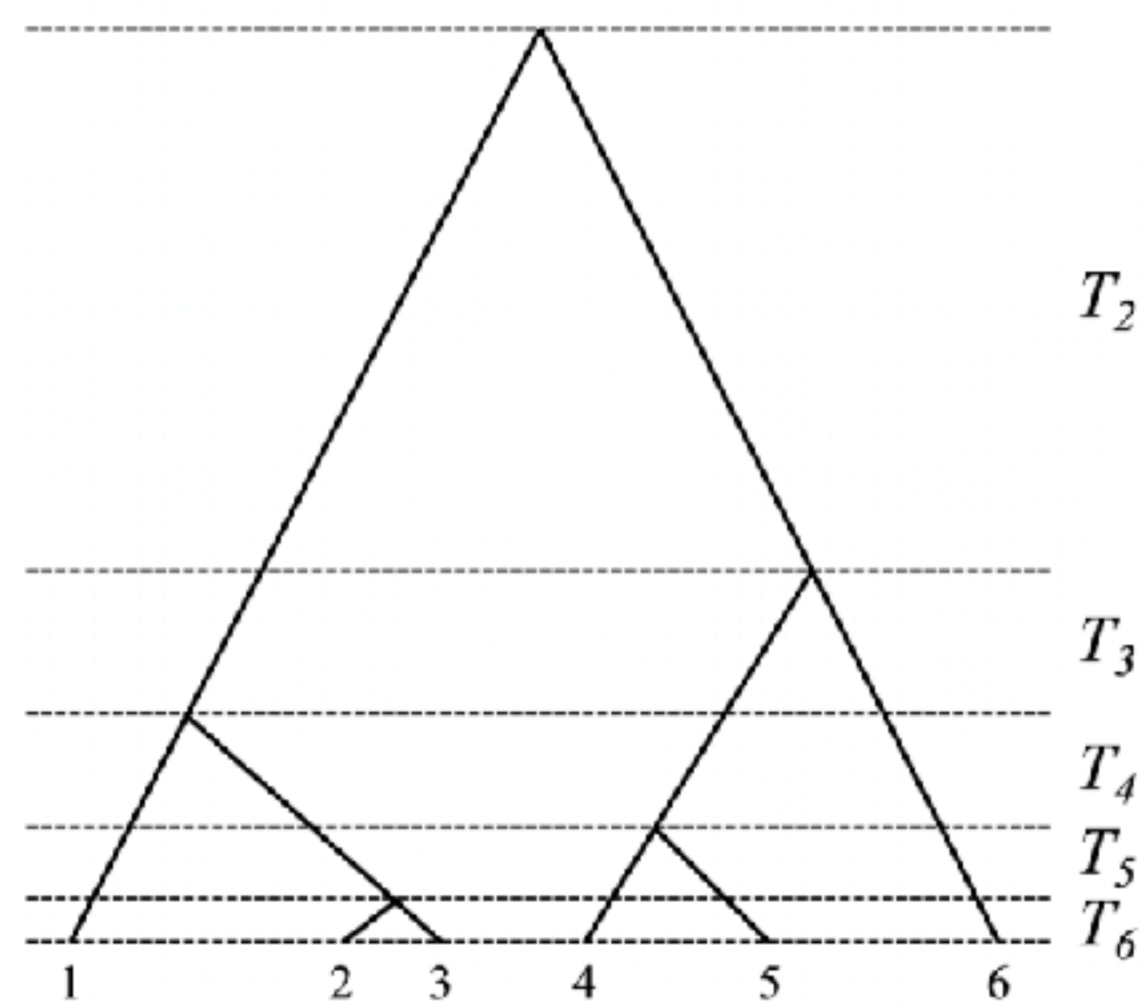
Coalescent

Next coalescence among k sequences

$$\mathbb{P}(T_k > t) = 1 - e^{-t\frac{\binom{k}{2}}{2N}}$$

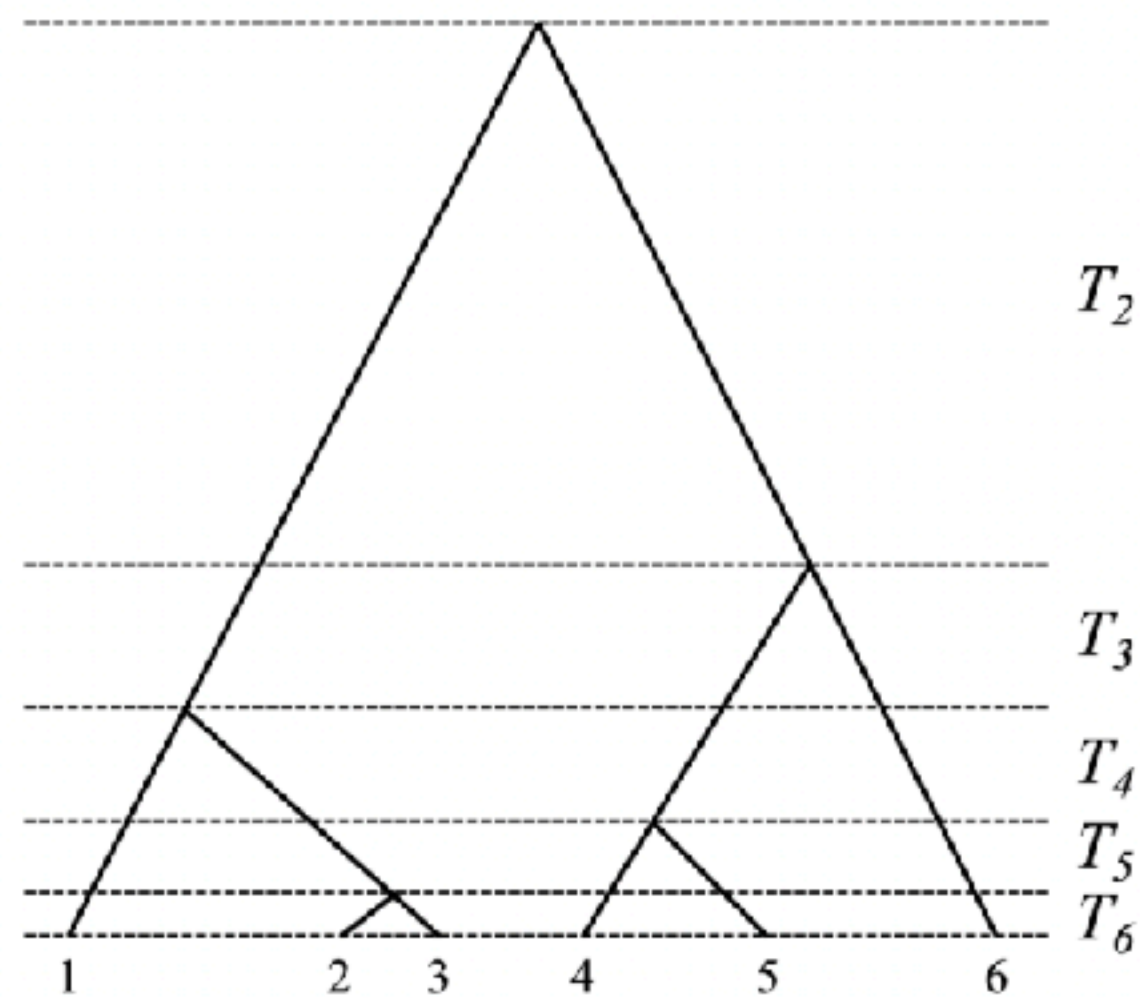
$$\mathbb{P}(T_k < t) = e^{-t\frac{\binom{k}{2}}{2N}}$$

$$\mathbb{E}(T_{k\to k-1}) = \frac{2N}{\binom{k}{2}}$$



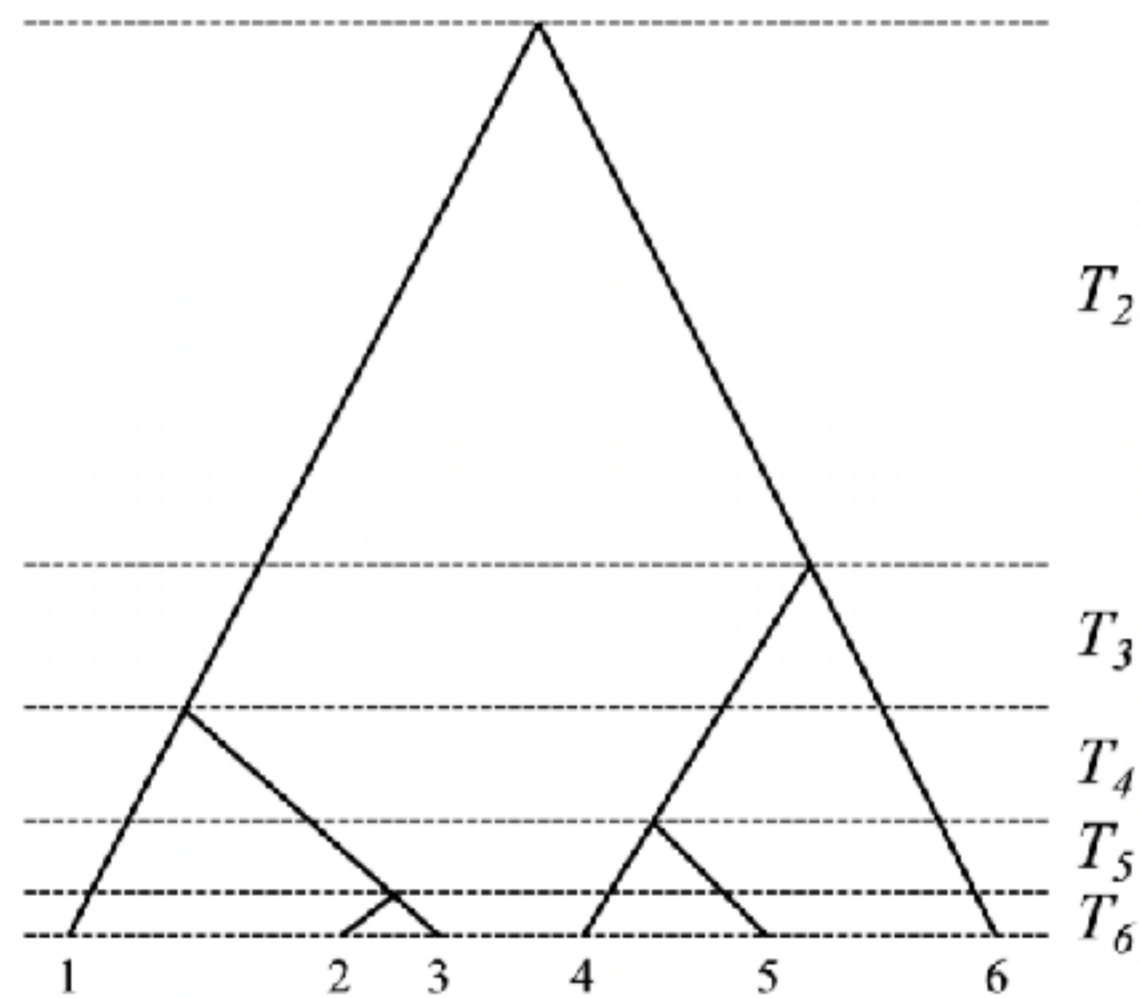
Total branch length

$$\mathbb{E}(T_{total}) = \sum_{k=2}^{n} k\, \mathbb{E}(T_{k\to k-1})$$



Total branch length

$$
\begin{aligned}
\mathbb{E}(T_{total}) &= \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1}) \\
&= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)}
\end{aligned}
$$

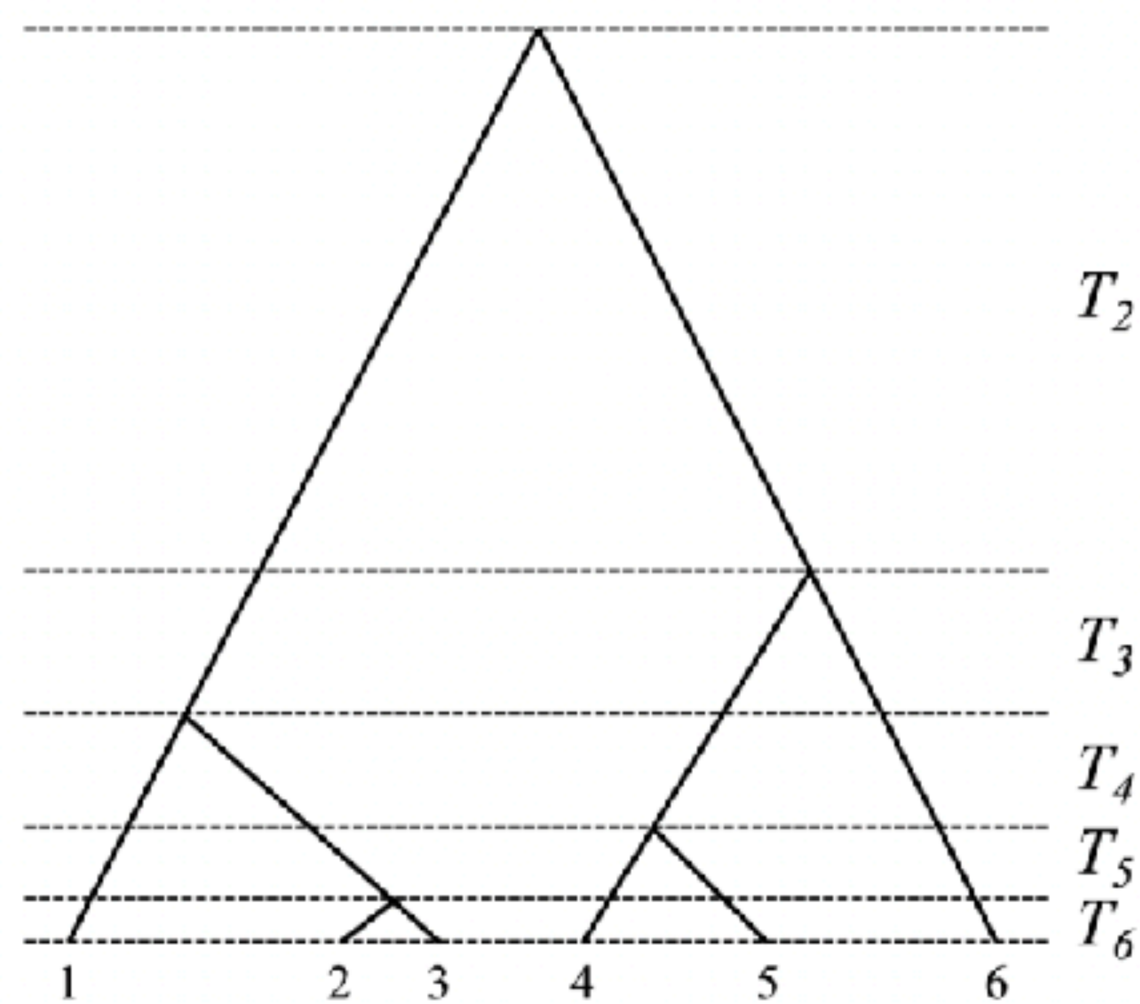


Total branch length

$$\mathbb{E}(T_{total}) = \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1})$$

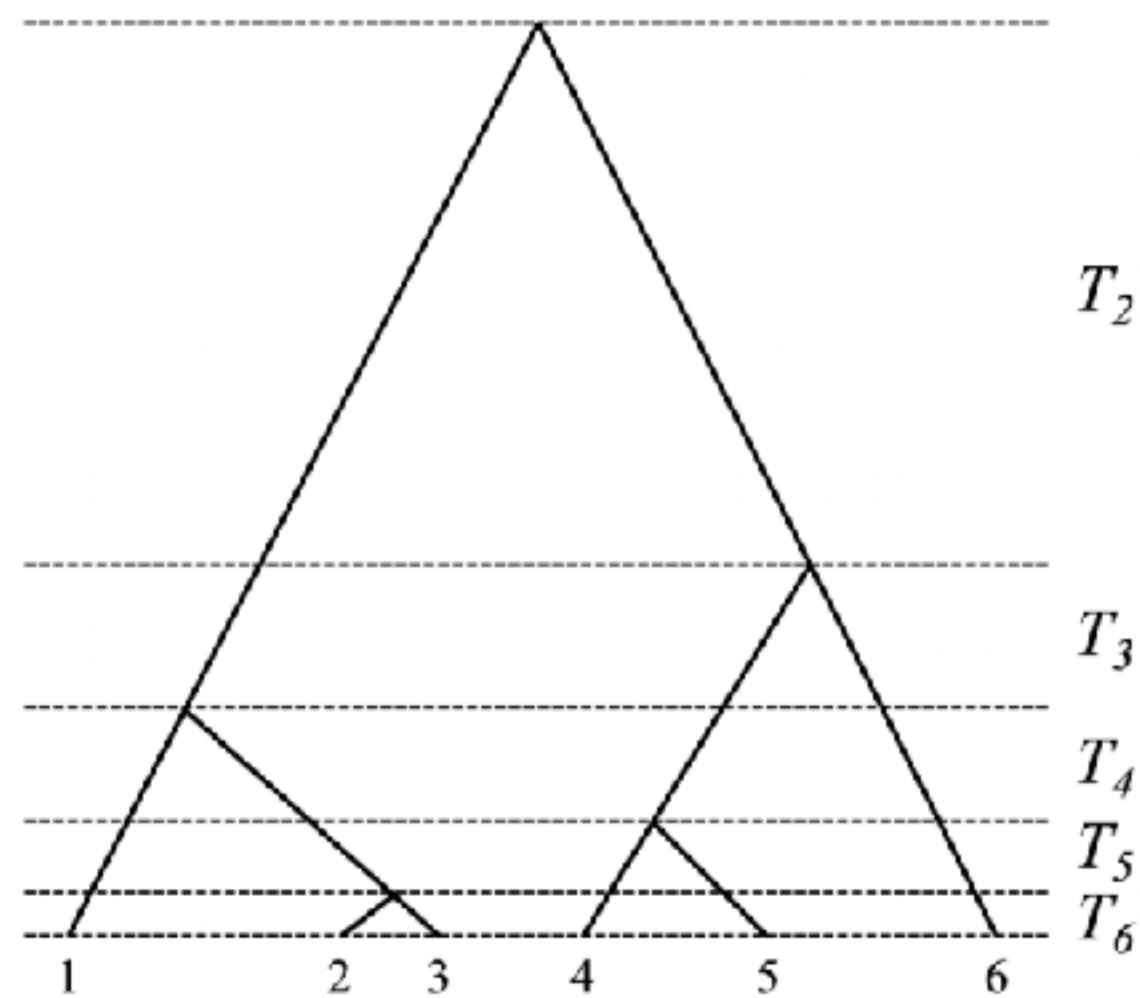
$$= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)}$$

$$= 4N \sum_{k=2}^{n} \frac{1}{(k-1)}$$

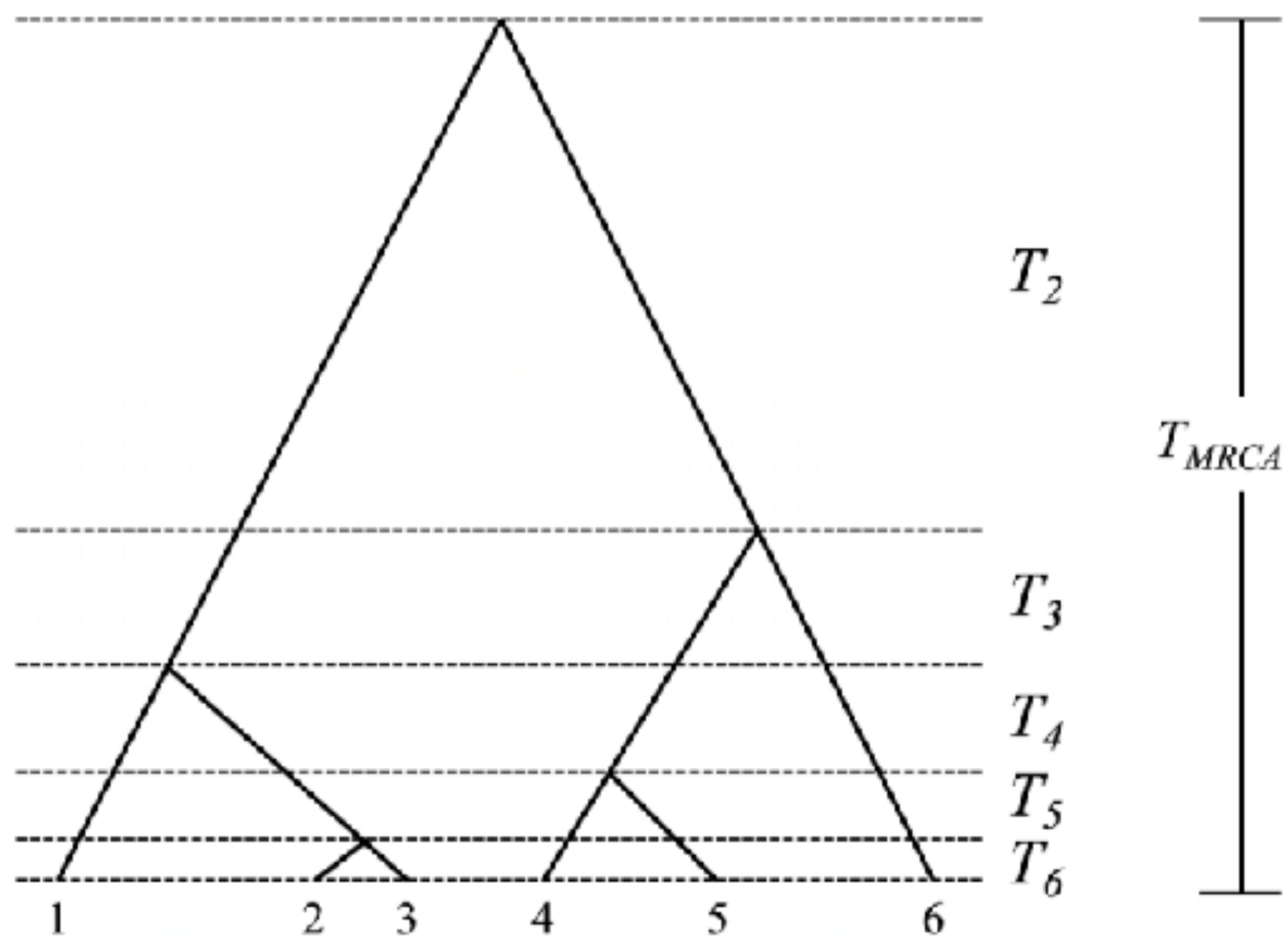


Total branch length

$$\begin{aligned}
\mathbb{E}(T_{total}) &= \sum_{k=2}^{n} k\, \mathbb{E}(T_{k\to k-1}) \\
&= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N \sum_{k=2}^{n} \frac{1}{(k-1)} \\
&= 4N \sum_{k=1}^{n-1} \frac{1}{k}
\end{aligned}$$



Total tree height - TMRCA



$$\mathbb{E}(T_{TMRCA}) = 4N\left(1 - \frac{1}{k}\right)$$

Total tree height - TMRCA

$$\mathbb{E}(T_{TMRCA}) = \frac{4N}{k(k-1)} + \frac{4N}{(k-1)(k-2)} + \frac{4N}{(k-2)(k-3)} + \cdots + \frac{4N}{2}$$

Total tree height - TMRCA

$$\mathbb{E}(T_{TMRCA}) = \frac{4N}{k(k-1)} + \frac{4N}{(k-1)(k-2)} + \frac{4N}{(k-2)(k-3)} + \cdots + \frac{4N}{2}$$
$$= 4N\left(\frac{1}{k(k-1)} + \frac{1}{(k-1)(k-2)} + \frac{1}{(k-2)(k-3)} + \cdots + \frac{1}{2}\right)$$

$$\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$$

Total tree height - TMRCA

$$
\begin{aligned}
\mathbb{E}(T_{TMRCA}) &= \frac{4N}{k(k-1)} + \frac{4N}{(k-1)(k-2)} + \frac{4N}{(k-2)(k-3)} + \cdots + \frac{4N}{2} \\
&= 4N\left(\frac{1}{k(k-1)} + \frac{1}{(k-1)(k-2)} + \frac{1}{(k-2)(k-3)} + \cdots + \frac{1}{2}\right) \\
&= 4N\left(\left(\frac{1}{k-1} - \frac{1}{k}\right) + \left(\frac{1}{k-2} - \frac{1}{k-1}\right) + \left(\frac{1}{k-3} - \frac{1}{k-2}\right) + \cdots + \left(\frac{1}{1} - \frac{1}{2}\right)\right)
\end{aligned}
$$

$$
\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}
$$

Total tree height - TMRCA

$$
\begin{aligned}
\mathbb{E}(T_{TMRCA}) &= \frac{4N}{k(k-1)} + \frac{4N}{(k-1)(k-2)} + \frac{4N}{(k-2)(k-3)} + \cdots + \frac{4N}{2} \\
&= 4N\left(\frac{1}{k(k-1)} + \frac{1}{(k-1)(k-2)} + \frac{1}{(k-2)(k-3)} + \cdots + \frac{1}{2}\right) \\
&= 4N\left(\left(\frac{1}{k-1} - \frac{1}{k}\right) + \left(\frac{1}{k-2} - \frac{1}{k-1}\right) + \left(\frac{1}{k-3} - \frac{1}{k-2}\right) + \cdots + \left(\frac{1}{1} - \frac{1}{2}\right)\right) \\
&= 4N\left(\cancel{\frac{1}{k-1}} - \frac{1}{k} + \cancel{\frac{1}{k-2}} - \cancel{\frac{1}{k-1}} + \cancel{\frac{1}{k-3}} - \cancel{\frac{1}{k-2}} + \cdots + \frac{1}{1} - \cancel{\frac{1}{2}}\right)
\end{aligned}
$$

$$\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$$

Total tree height - TMRCA

$$
\begin{aligned}
\mathbb{E}(T_{TMRCA}) &= \frac{4N}{k(k-1)} + \frac{4N}{(k-1)(k-2)} + \frac{4N}{(k-2)(k-3)} + \cdots + \frac{4N}{2} \\
&= 4N\left(\frac{1}{k(k-1)} + \frac{1}{(k-1)(k-2)} + \frac{1}{(k-2)(k-3)} + \cdots + \frac{1}{2}\right) \\
&= 4N\left(\left(\frac{1}{k-1} - \frac{1}{k}\right) + \left(\frac{1}{k-2} - \frac{1}{k-1}\right) + \left(\frac{1}{k-3} - \frac{1}{k-2}\right) + \cdots + \left(\frac{1}{1} - \frac{1}{2}\right)\right) \\
&= 4N\left(\cancel{\frac{1}{k-1}} - \frac{1}{k} + \cancel{\frac{1}{k-2}} - \cancel{\frac{1}{k-1}} + \cancel{\frac{1}{k-3}} - \cancel{\frac{1}{k-2}} + \cdots + \frac{1}{1} - \cancel{\frac{1}{2}}\right) \\
&= 4N\left(1 - \frac{1}{k}\right)
\end{aligned}
$$

$$
\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}
$$

SGDP paper

Week 1

Kasper Munch

Discussion

SGDP paper

ARTICLE

doi:10.1038/nature18964

The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

1. What is the title of the paper?

2. What does the abstract say are the most important findings or contributions?

3. Which figures support each of these findings?

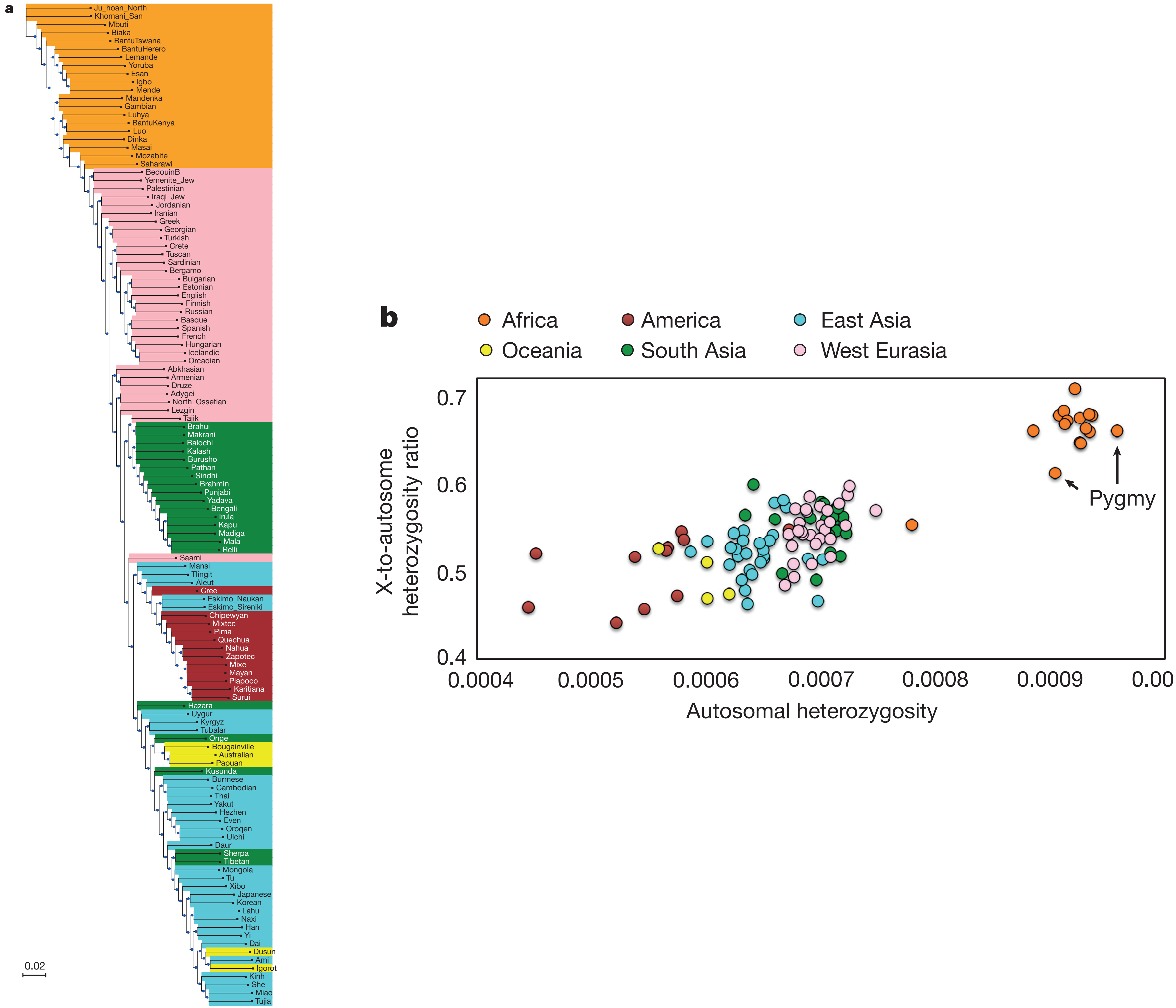
4. Pick one of these findings and discuss the associated text and figure with two fellow students.

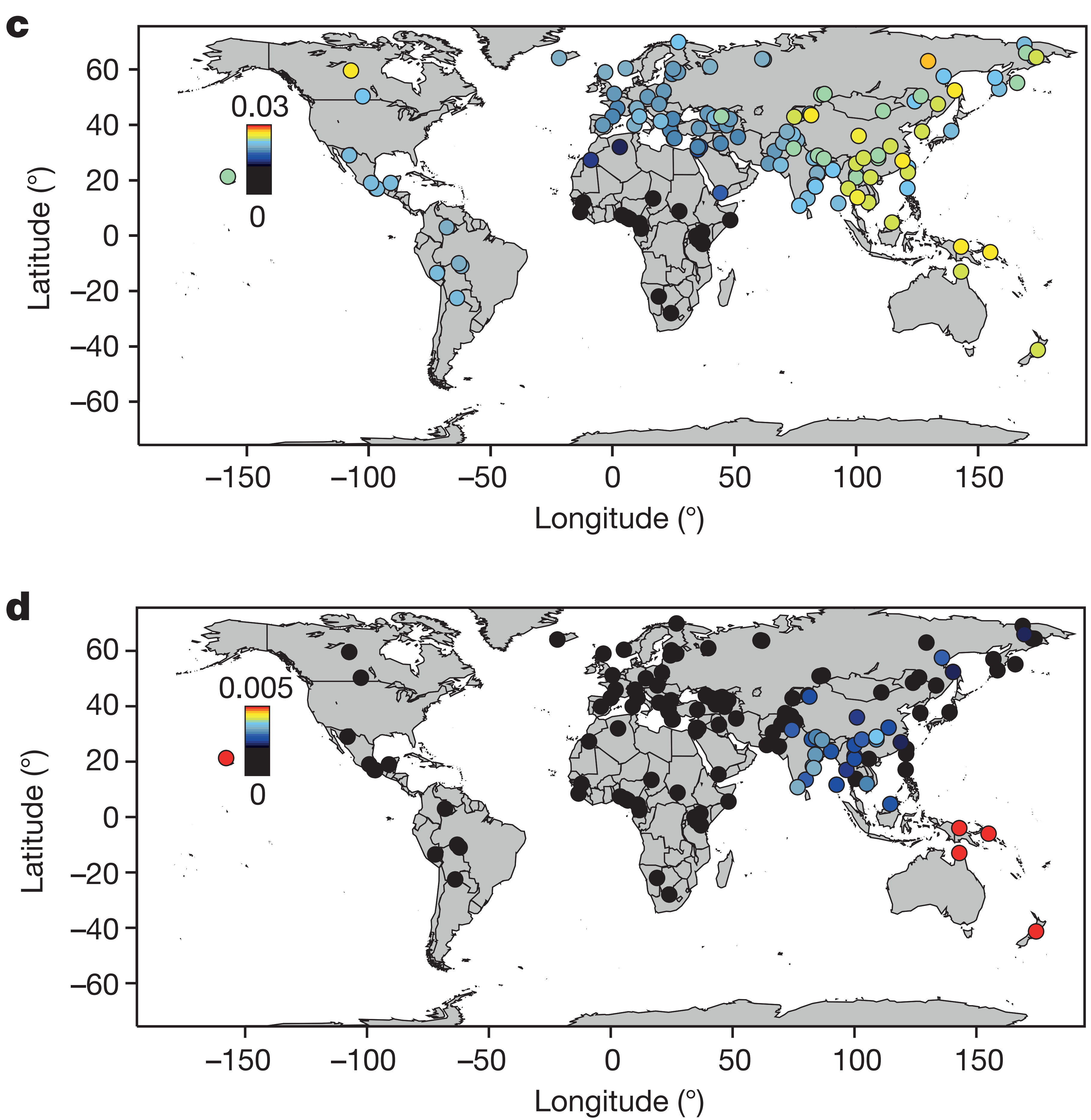
    1. What do they find?

    2. How do they find it?

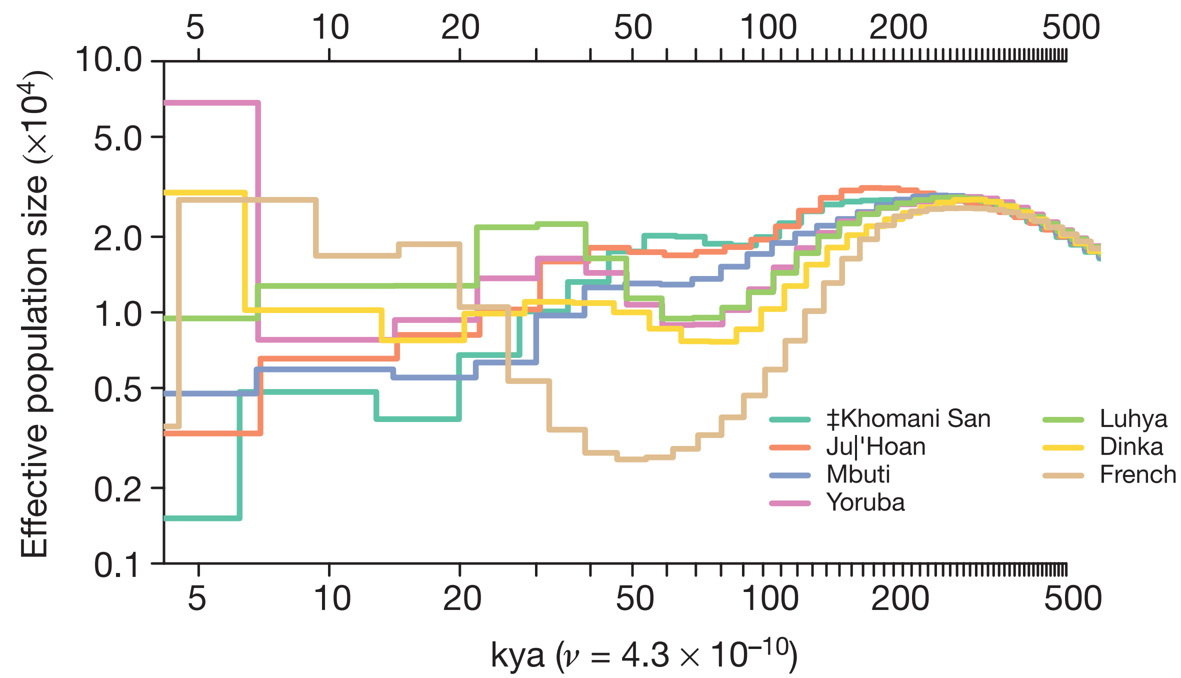
    3. How does the figure represent the result?

    4. Do you think it is interesting/relevant, and if so how/why?


a
Ju_hoan_North
Khomani_San
Mbuti
Biaka
BantuTswana
BantuHerero
Lemande
Yoruba
Esan
Igbo
Mende
Mandenka
Gambian
Luhya
BantuKenya
Luo
Dinka
Masai
Mozabite
Saharawi
BedouinB
Yemenite_Jew
Palestinian
Iraqi_Jew
Jordanian
Iranian
Greek
Georgian
Turkish
Crete
Tuscan
Sardinian
Bergamo
Bulgarian
Estonian
English
Finnish
Russian
Basque
Spanish
French
Hungarian
Icelandic
Orcadian
Abkhasian
Armenian
Druze
Adygei
North_Ossetian
Lezgin
Tajik
Brahui
Makrani
Balochi
Kalash
Burusho
Pathan
Sindhi
Brahmin
Punjabi
Yadava
Bengali
Irula
Kapu
Madiga
Mala
Relli
Saami
Mansi
Tlingit
Aleut
Cree
Eskimo_Naukan
Eskimo_Sireniki
Chipewyan
Mixtec
Pima
Quechua
Nahua
Zapotec
Mixe
Mayan
Piapoco
Karitiana
Surui
Hazara
Uygur
Kyrgyz
Tubalar
Onge
Bougainville
Australian
Papuan
Kusunda
Burmese
Cambodian
Thai
Yakut
Hezhen
Even
Oroqen
Ulchi
Daur
Sherpa
Tibetan
Mongola
Tu
Xibo
Japanese
Korean
Lahu
Naxi
Han
Yi
Dai
Dusun
Ami
Igorot
Kinh
She
Miao
Tujia
0.02
b
Africa
America
East Asia
Oceania
South Asia
West Eurasia
X-to-autosome heterozygosity ratio
0.7
0.6
0.5
0.4
Pygmy
0.0004
0.0005
0.0006
0.0007
0.0008
0.0009
0.001
Autosomal heterozygosity

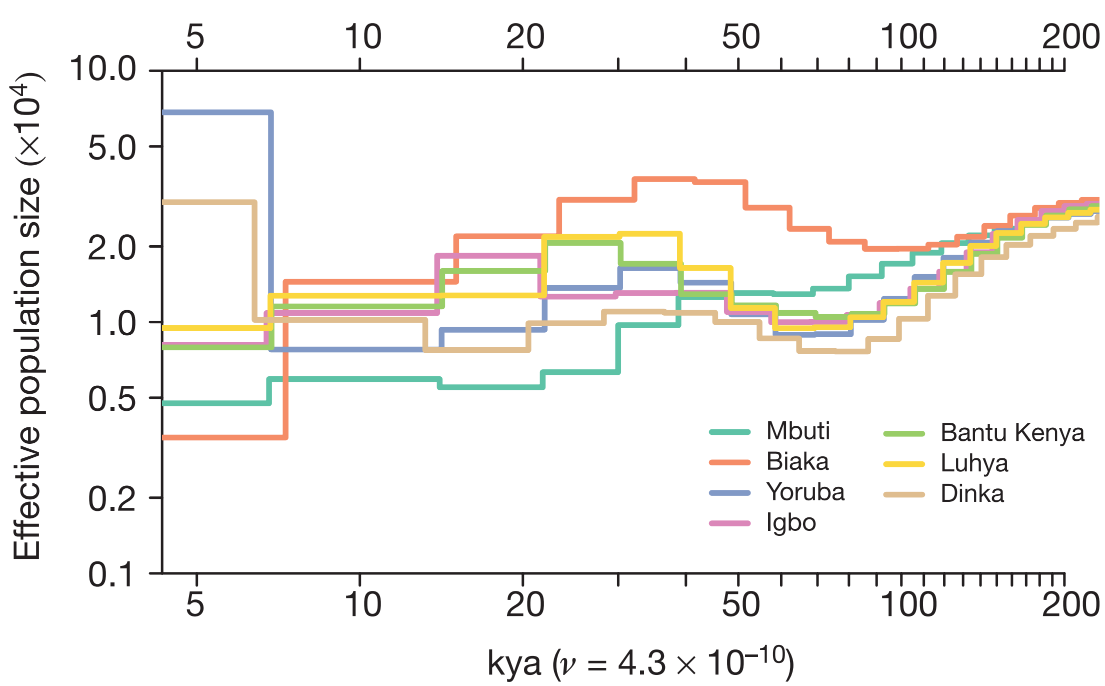

c
0.03
0
Latitude (°)
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
Longitude (°)
d
0.005
0
Latitude (°)
Longitude (°)

d



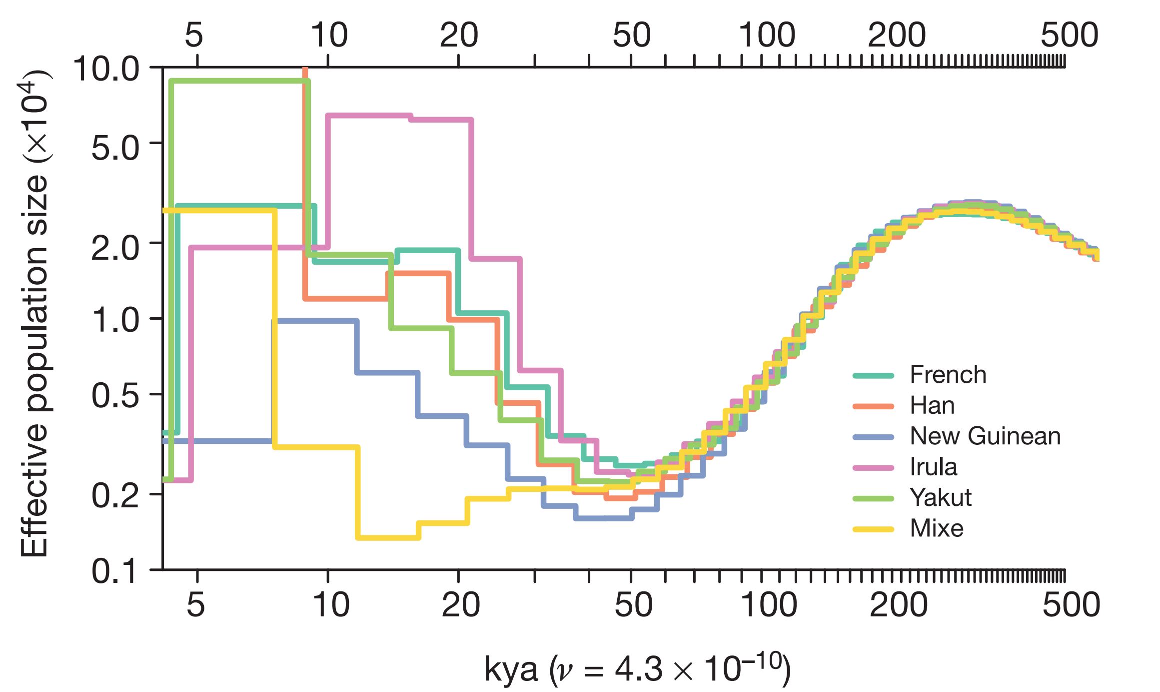
Effective population size (×10^4)
kya (ν = 4.3 × 10−10)
‡Khomani San
Ju|'Hoan
Mbuti
Yoruba
Luhya
Dinka
French

e

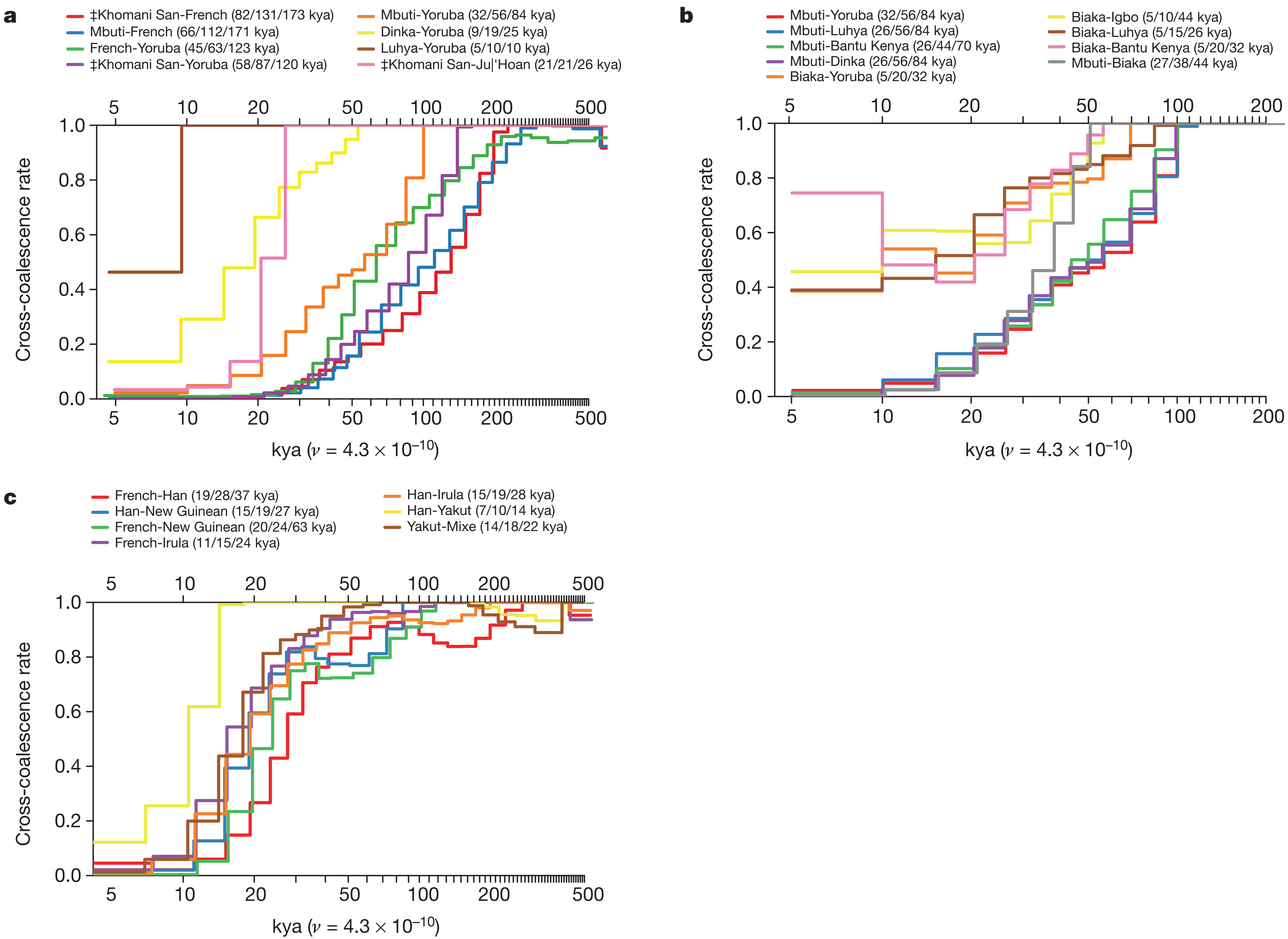


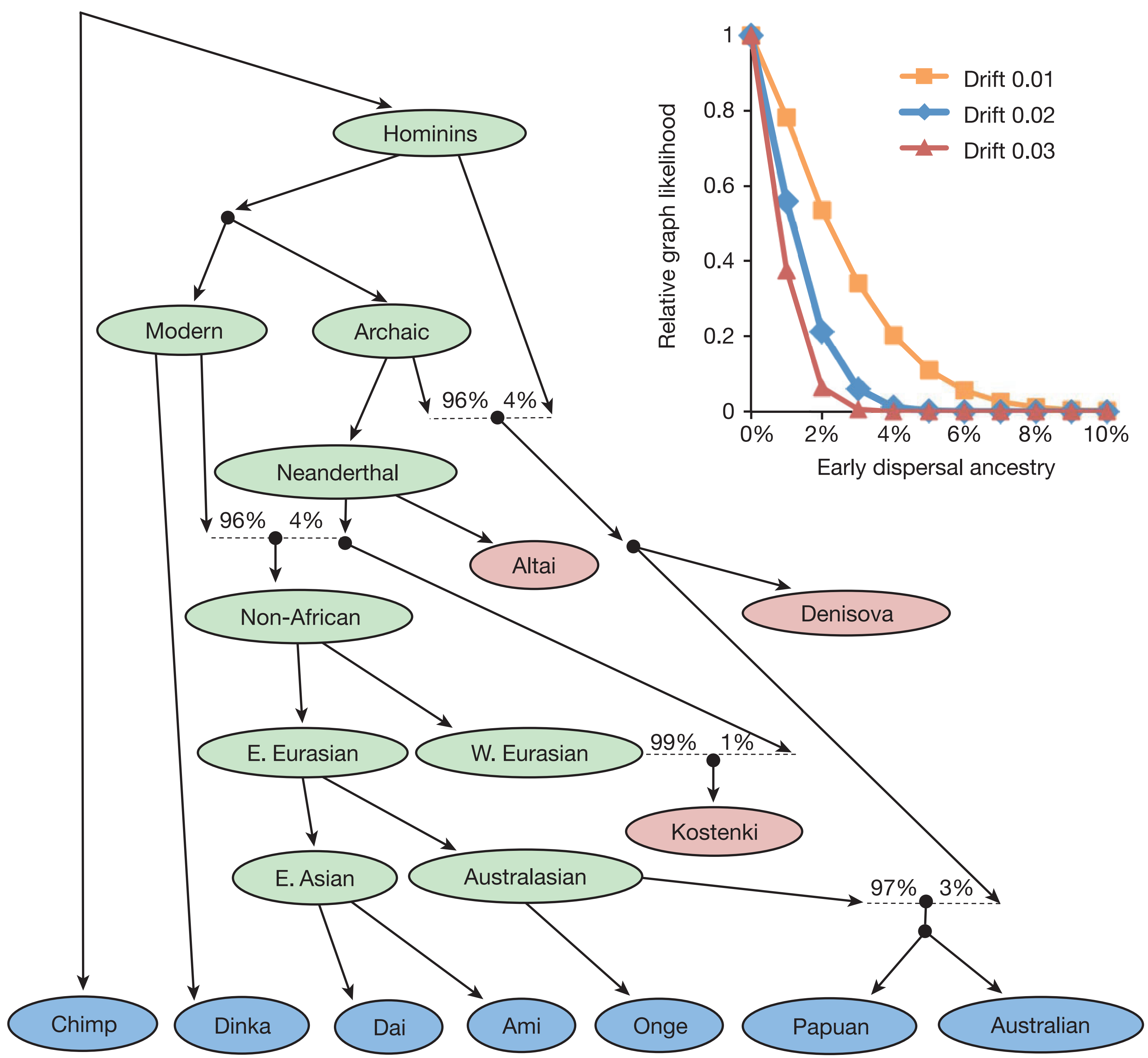
Effective population size (×10^4)
kya (ν = 4.3 × 10−10)
Mbuti
Biaka
Yoruba
Igbo
Bantu Kenya
Luhya
Dinka

f



Effective population size (×10^4)
kya (ν = 4.3 × 10−10)
French
Han
New Guinean
Irula
Yakut
Mixe


a
‡Khomani San-French (82/131/173 kya)
Mbuti-French (66/112/171 kya)
French-Yoruba (45/63/123 kya)
‡Khomani San-Yoruba (58/87/120 kya)
Mbuti-Yoruba (32/56/84 kya)
Dinka-Yoruba (9/19/25 kya)
Luhya-Yoruba (5/10/10 kya)
‡Khomani San-Ju|'Hoan (21/21/26 kya)
Cross-coalescence rate
kya ($\nu = 4.3 \times 10^{-10}$)
b
Mbuti-Yoruba (32/56/84 kya)
Mbuti-Luhya (26/56/84 kya)
Mbuti-Bantu Kenya (26/44/70 kya)
Mbuti-Dinka (26/56/84 kya)
Biaka-Yoruba (5/20/32 kya)
Biaka-Igbo (5/10/44 kya)
Biaka-Luhya (5/15/26 kya)
Biaka-Bantu Kenya (5/20/32 kya)
Mbuti-Biaka (27/38/44 kya)
Cross-coalescence rate
kya ($\nu = 4.3 \times 10^{-10}$)
c
French-Han (19/28/37 kya)
Han-New Guinean (15/19/27 kya)
French-New Guinean (20/24/63 kya)
French-Irula (11/15/24 kya)
Han-Irula (15/19/28 kya)
Han-Yakut (7/10/14 kya)
Yakut-Mixe (14/18/22 kya)
Cross-coalescence rate
kya ($\nu = 4.3 \times 10^{-10}$)



Hominins
Modern
Archaic
96%
4%
Neanderthal
96%
4%
Altai
Denisova
Non-African
E. Eurasian
W. Eurasian
99%
1%
Kostenki
E. Asian
Australasian
97%
3%
Chimp
Dinka
Dai
Ami
Onge
Papuan
Australian
Relative graph likelihood
Early dispersal ancestry
Drift 0.01
Drift 0.02
Drift 0.03
1
0.8
0.6
0.4
0.2
0
0%
2%
4%
6%
8%
10%

The coalescent with mutations

Week 2

Kasper Munch

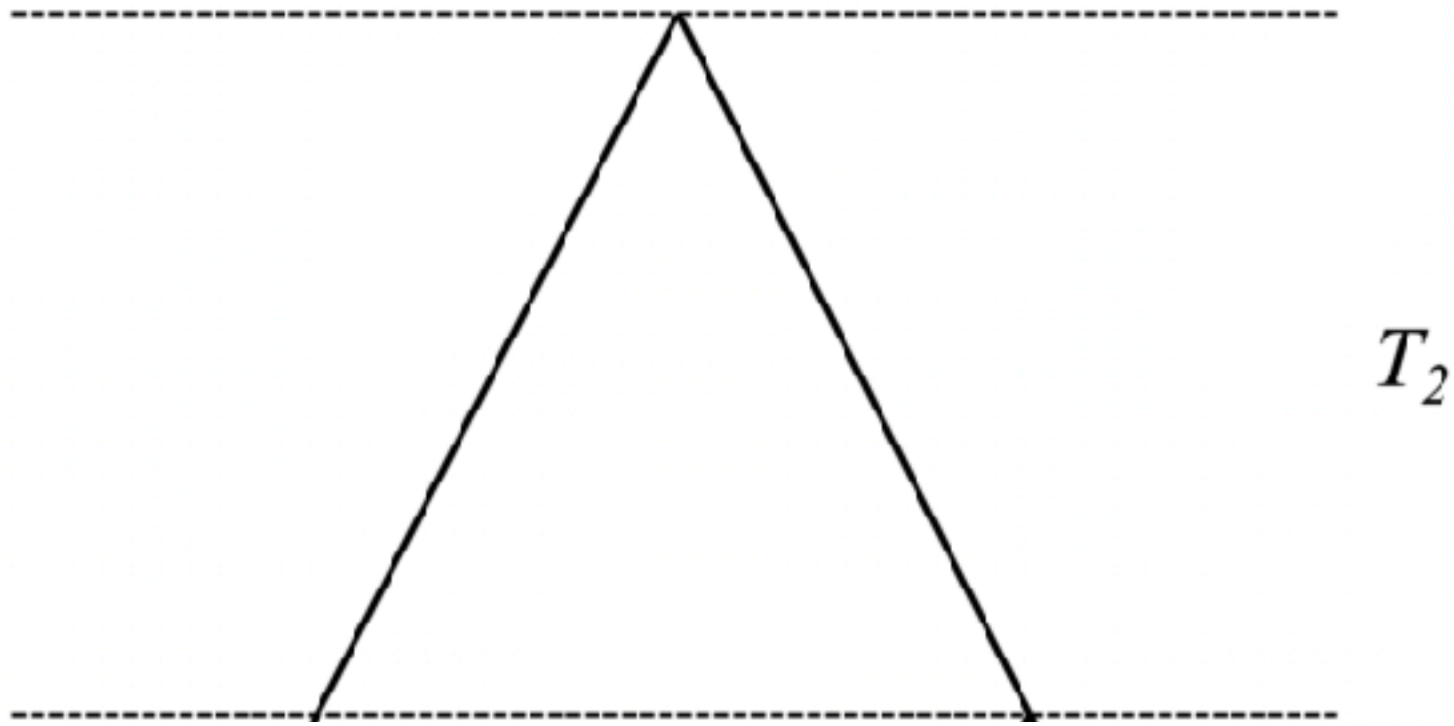
Popgen basics

Measuring diversity

- **Pairwise differences**
- **Segregating sites**

```
GCTCACCGGAATTATCCGATATGCTAGTA
GCTTACCGGAATTATGCGATATGCTTGTA
GCTCACCGGAATTATGCGATATGGTAGAA
GCTCACCGGAATTATGCGATATGGTAGAA
GCTCACCGGGATGATGCGATATGCTAGTA
GCTCACCGGAATTATGCGATATGCTAGAA
GCTTACCGGAATTATCCGATATGCTAGTA
GCTCACAGGGATTATGCGCTATGCTAGTA
GCTCACCGGAATTATGCGATATGGTAGAA
GCTCACCGGAATTATCCGATATGCTAGTA
```

Pairwise differences



$$\mathbb{E}(\pi) = 2\,\mathbb{E}(T_2)\,\mu = 4N\mu = \theta$$

Segregating sites

$$\mathbb{E}(T_{total}) = \sum_{k=2}^{n} k\,\mathbb{E}(T_{k \to k-1})$$

Segregating sites

$$\mathbb{E}(T_{total}) = \sum_{k=2}^{n} k\,\mathbb{E}(T_{k \to k-1})$$

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

Segregating sites

$$
\begin{aligned}
\mathbb{E}(T_{total}) &= \sum_{k=2}^{n} k \, \mathbb{E}(T_{k \to k-1}) \\
&= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N \sum_{k=2}^{n} \frac{1}{(k-1)}
\end{aligned}
$$

Segregating sites

$$
\begin{aligned}
\mathbb{E}(T_{total}) &= \sum_{k=2}^{n} k\, \mathbb{E}(T_{k\to k-1}) \\
&= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N \sum_{k=2}^{n} \frac{1}{(k-1)} \\
&= 4N \sum_{k=1}^{n-1} \frac{1}{k}
\end{aligned}
$$

Segregating sites

$$\mathbb{E}(T_{total}) = \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1})$$

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

$$= 4N \sum_{k=2}^{n} \frac{1}{(k-1)}$$

$$= 4N \sum_{k=1}^{n-1} \frac{1}{k}$$

$$\mathbb{E}(S) = \mu \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1})$$

Segregating sites

$$\begin{aligned}
\mathbb{E}(T_{total}) &= \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1}) \\
&= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N \sum_{k=2}^{n} \frac{1}{(k-1)} \\
&= 4N \sum_{k=1}^{n-1} \frac{1}{k}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(S) &= \mu \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1}) \\
&= \mu \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)}
\end{aligned}$$

Segregating sites

$$
\begin{aligned}
\mathbb{E}(T_{total}) &= \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1}) \\
&= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N \sum_{k=2}^{n} \frac{1}{(k-1)} \\
&= 4N \sum_{k=1}^{n-1} \frac{1}{k}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(S) &= \mu \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1}) \\
&= \mu \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N\mu \sum_{k=2}^{n} \frac{1}{(k-1)}
\end{aligned}
$$

Segregating sites

$$\begin{aligned}
\mathbb{E}(T_{total}) &= \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1}) \\
&= \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N \sum_{k=2}^{n} \frac{1}{(k-1)} \\
&= 4N \sum_{k=1}^{n-1} \frac{1}{k}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(S) &= \mu \sum_{k=2}^{n} k\,\mathbb{E}(T_{k\to k-1}) \\
&= \mu \sum_{k=2}^{n} k \frac{2N}{\left(\frac{k(k-1)}{2}\right)} \\
&= 4N\mu \sum_{k=2}^{n} \frac{1}{(k-1)} \\
&= 4N\mu \sum_{k=1}^{n-1} \frac{1}{k}
\end{aligned}$$

Segregating sites

$$\mathbb{E}(T_{total}) = \sum_{k=2}^{n} k\, \mathbb{E}(T_{k\to k-1})$$

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

$$= 4N \sum_{k=2}^{n} \frac{1}{(k-1)}$$

$$= 4N \sum_{k=1}^{n-1} \frac{1}{k}$$

$$\mathbb{E}(S) = \mu \sum_{k=2}^{n} k\, \mathbb{E}(T_{k\to k-1})$$

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

$$= 4N\mu \sum_{k=2}^{n} \frac{1}{(k-1)}$$

$$= 4N\mu \sum_{k=1}^{n-1} \frac{1}{k}$$

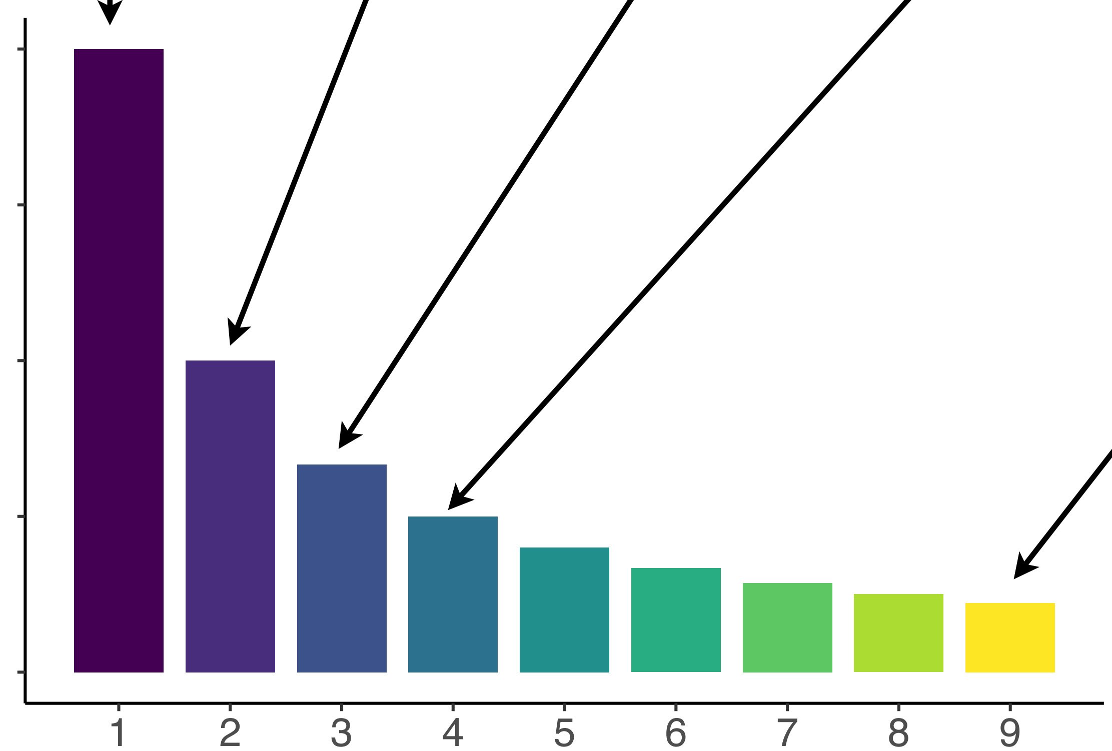
$$= \theta \sum_{k=1}^{n-1} \frac{1}{k}$$

Site Frequency Spectrum (SFS)

10 samples

$$\mathbb{E}(S) = 4N\mu \sum_{k=1}^{n-1} \frac{1}{k}$$

$$= 4N\mu\frac{1}{1} + 4N\mu\frac{1}{2} + 4N\mu\frac{1}{3} + 4N\mu\frac{1}{4} + \cdots + 4N\mu\frac{1}{n-1}$$



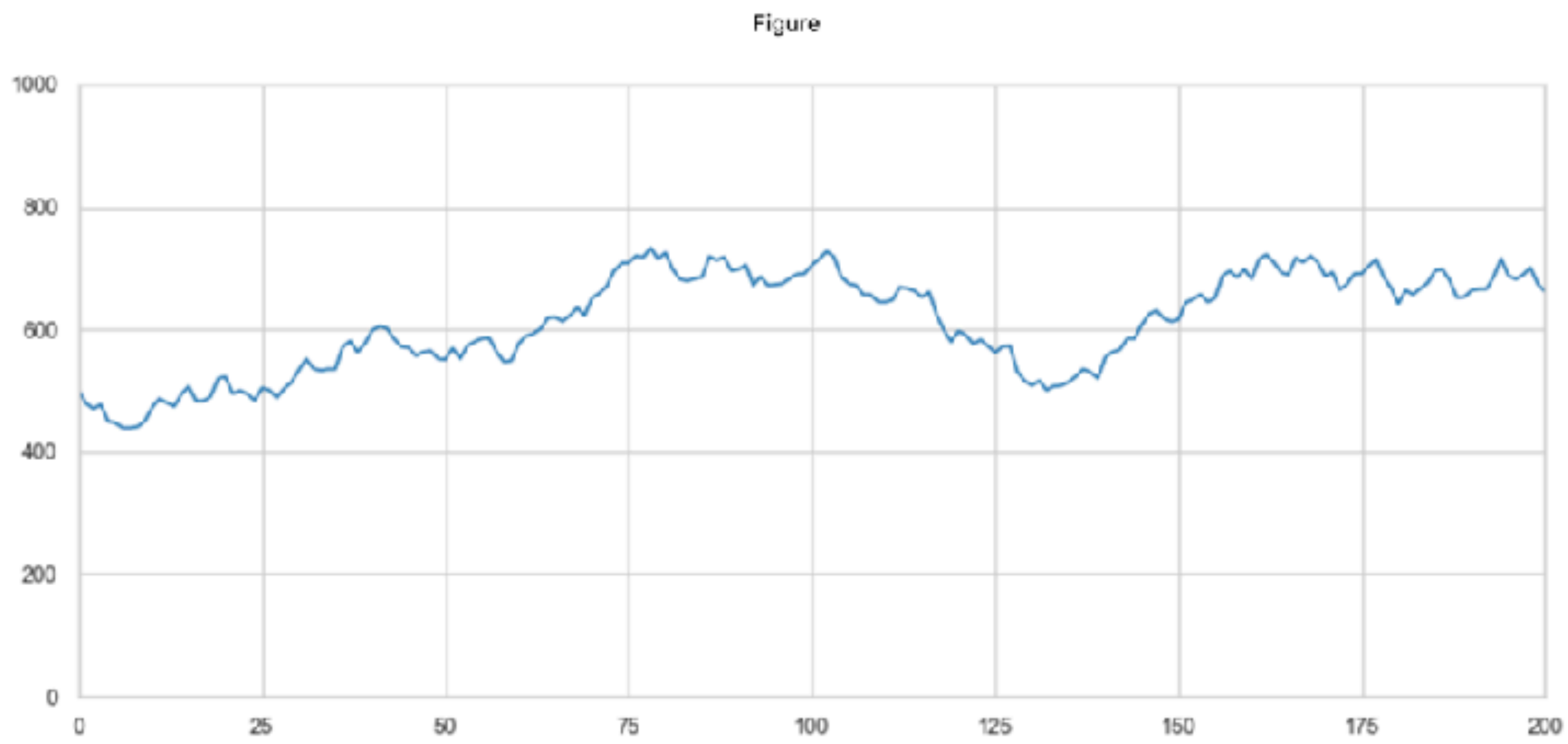
Genetic drift and the coalescent

Week 2

Kasper Munch

Genetic drift

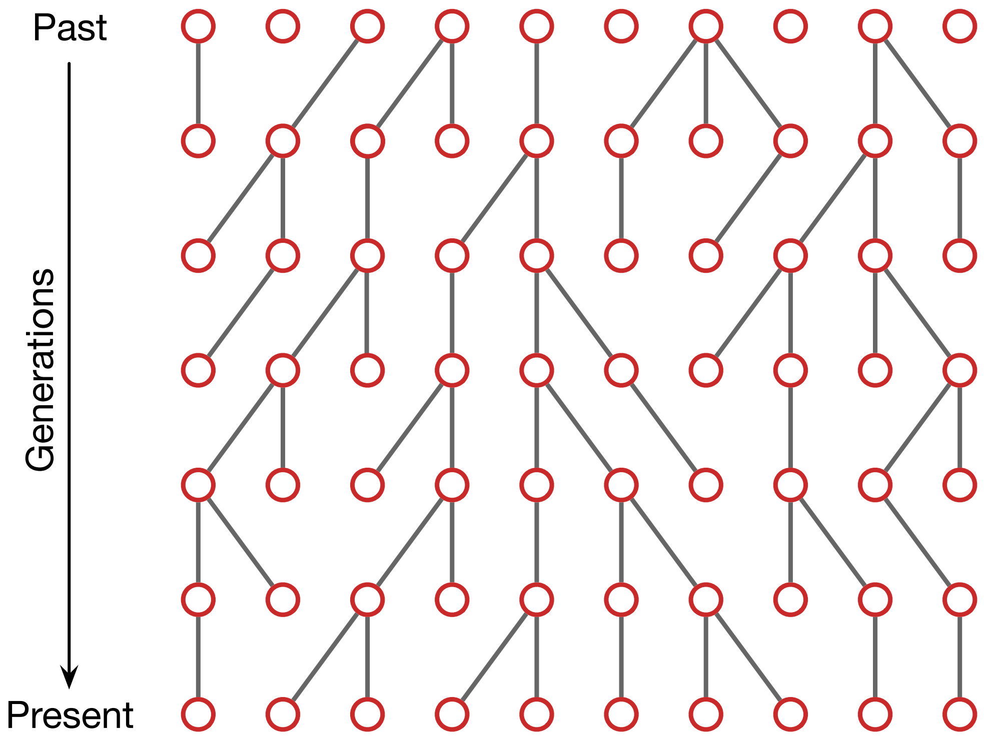
Allele frequency over time



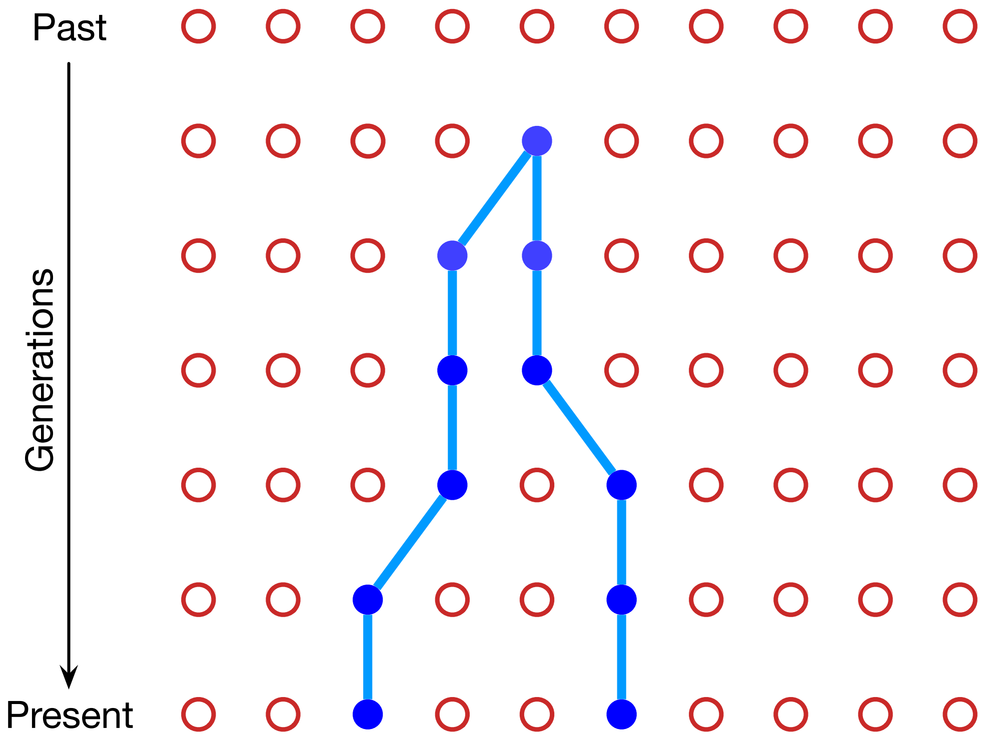
Jupyter Notebook

Effective population size

The N in a constant sized WF-population that would give the same …



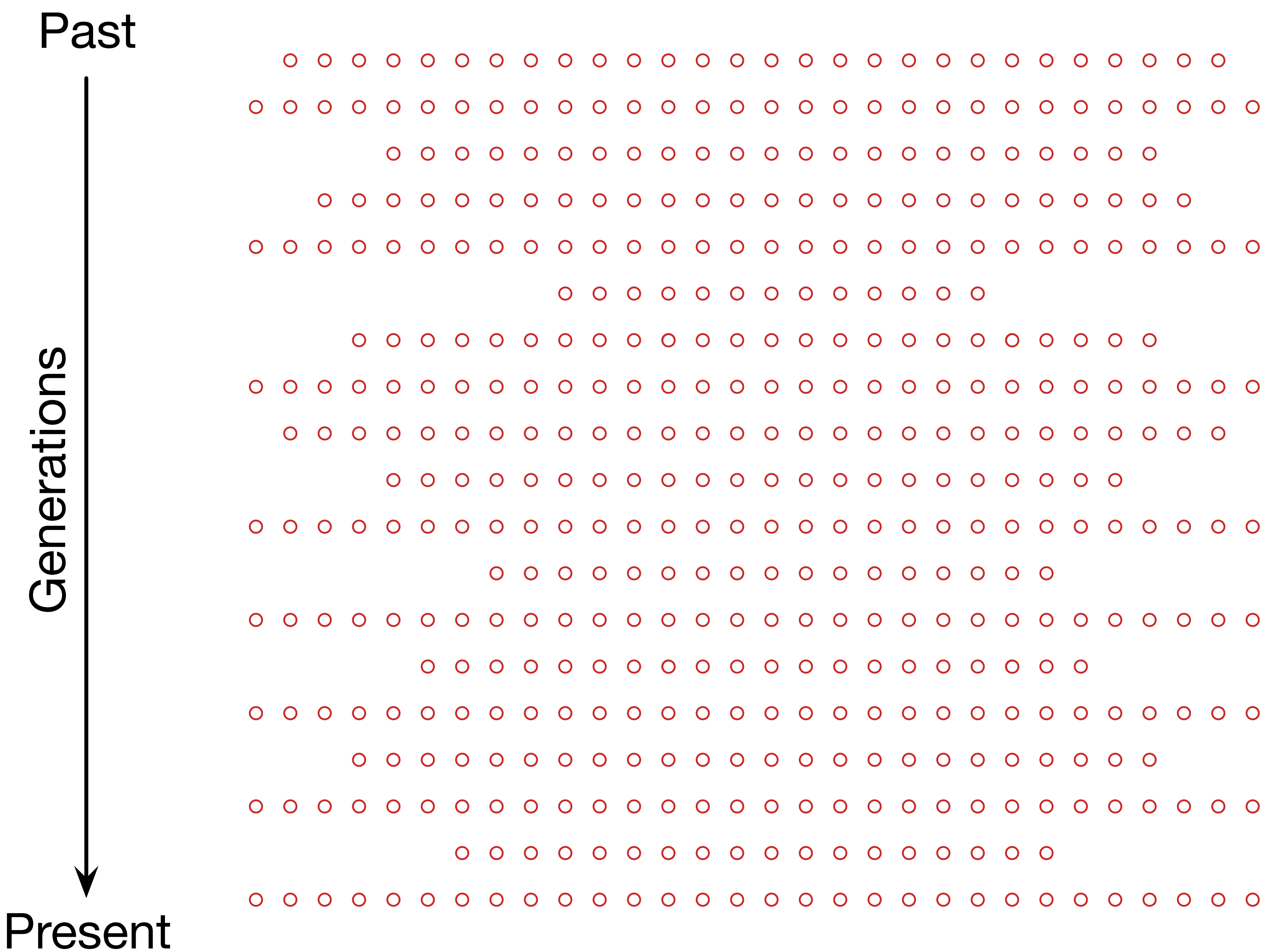
… development of variance in allele frequency



… coalescence rate

Population sizes are not constant

How do we find the average coalescence rate?



$$N_e = \frac{i}{\sum_{i=0}^{g} \frac{1}{N_i}}$$

Harmonic mean
of the effective
population sizes of
generations
(or even sized epoch)

Some individuals reproduce more

How do we find the average coalescence rate?

No matter how many males and females that reproduce - everyone has a father and a mother. So *each* lineage spend the same time in **males** and **females**.

$$\frac{1}{N_e}$$

Some individuals reproduce more

How do we find the average coalescence rate?

No matter how many males and females that reproduce - everyone has a father and a mother. So *each* lineage spend the same time in **males** and **females**.

Numbers of breeding **males** and **females**

$$\frac{1}{N_e} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_M} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_F}$$

Some individuals reproduce more

How do we find the average coalescence rate?

No matter how many males and females that reproduce - everyone has a father and a mother. So *each* lineage spend the same time in **males** and **females**.

Numbers of breeding **males** and **females**

$$\frac{1}{N_e} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_M} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_F}$$

$$= \frac{1}{4} \times \frac{1}{2N_M} + \frac{1}{4} \times \frac{1}{2N_F}$$

Some individuals reproduce more

How do we find the average coalescence rate?

No matter how many males and females that reproduce - everyone has a father and a mother. So *each* lineage spend the same time in **males** and **females**.

Numbers of breeding **males** and **females**

$$\frac{1}{N_e} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_M} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_F}$$

$$= \frac{1}{4} \times \frac{1}{2N_M} + \frac{1}{4} \times \frac{1}{2N_F}$$

$$= \frac{\frac{1}{2N_M} + \frac{1}{2N_F}}{4}$$

Some individuals reproduce more

How do we find the average coalescence rate?

No matter how many males and females that reproduce - everyone has a father and a mother. So *each* lineage spend the same time in **males** and **females**.

Numbers of breeding **males** and **females**

$$\frac{1}{N_e} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_M} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_F}$$

$$= \frac{1}{4} \times \frac{1}{2N_M} + \frac{1}{4} \times \frac{1}{2N_F}$$

$$= \frac{\frac{1}{2N_M} + \frac{1}{2N_F}}{4}$$

$$N_e = \frac{4}{\frac{1}{2N_M} + \frac{1}{2N_F}}$$

Some individuals reproduce more

How do we find the average coalescence rate?

No matter how many males and females that reproduce - everyone has a father and a mother. So *each* lineage spend the same time in **males** and **females**.

Numbers of breeding **males** and **females**

$$\frac{1}{N_e} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_M} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_F}$$

$$= \frac{1}{4} \times \frac{1}{2N_M} + \frac{1}{4} \times \frac{1}{2N_F}$$

$$= \frac{\frac{1}{2N_M} + \frac{1}{2N_F}}{4}$$

$$N_e = \frac{4}{\frac{1}{2N_M} + \frac{1}{2N_F}}$$

$$= \frac{2}{\frac{1}{N_M} + \frac{1}{N_F}}$$

Harmonic mean of the number of breeding males and females

Some individuals reproduce more

How do we find the average coalescence rate?

No matter how many males and females that reproduce - everyone has a father and a mother. So *each* lineage spend the same time in **males** and **females**.

Numbers of breeding **males** and **females**

$$\frac{1}{N_e} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_M} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2N_F}$$

$$= \frac{1}{4} \times \frac{1}{2N_M} + \frac{1}{4} \times \frac{1}{2N_F}$$

$$= \frac{\frac{1}{2N_M} + \frac{1}{2N_F}}{4}$$

$$N_e = \frac{4}{\frac{1}{2N_M} + \frac{1}{2N_F}}$$

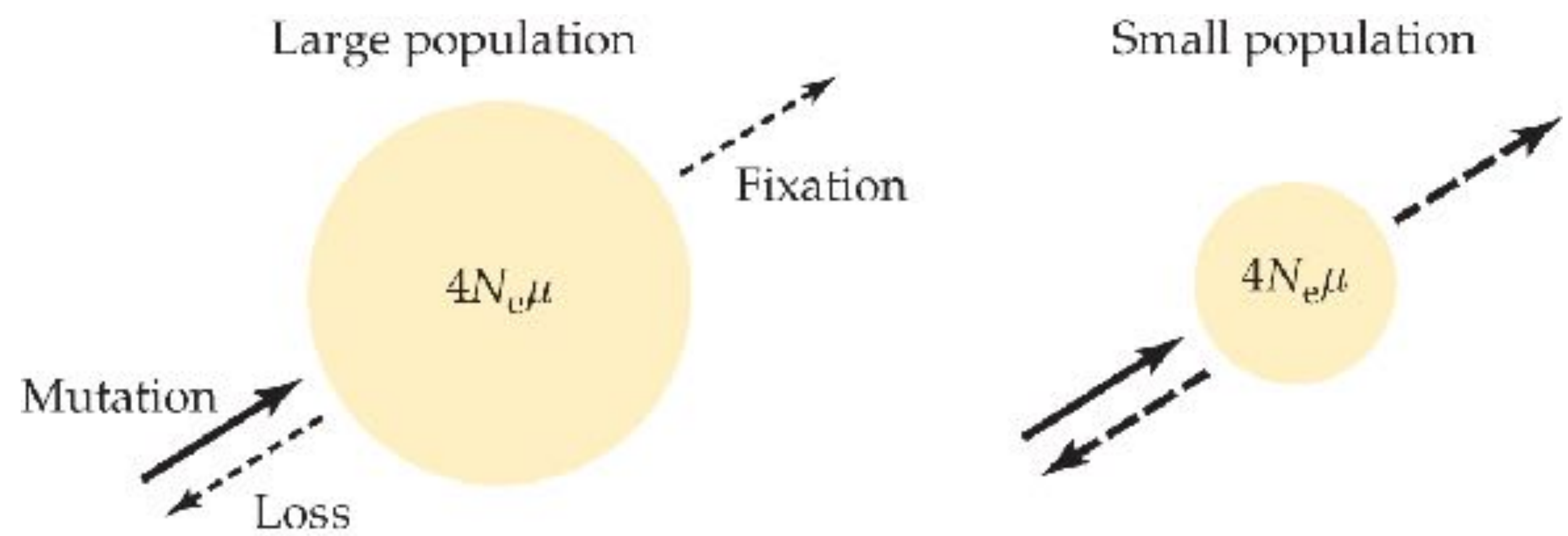
$$= \frac{2}{\frac{1}{N_M} + \frac{1}{N_F}}$$

Harmonic mean of the number of breeding males and females

$$= \frac{4N_M 4N_F}{N_M + N_F}$$

Coalescent

Drift - mutation equilibrium



Probability of homozygosity (IBD):

$$\frac{1/2N}{1/2N + 2\mu} = \frac{1}{1 + 4N\mu} = \frac{1}{1 + \theta}$$

Probability of heterozygosity:

$$\frac{2\mu}{1/2N + 2\mu} = \frac{4N\mu}{1 + 4N\mu} = \frac{\theta}{1 + \theta}$$

Discussion

Drift and Coalescent

- How is heterozygosity determined by genetic drift and mutation rate?
- What is the correspondence between genetic drift forward in time and the coalescent backwards in time?
- What determines the coalescence rate?
- What does the rate of coalescence in a time-interval tell us about the effective population size at that time?
- How does the total depth of the coalescent tree change as we add more samples?

Discussion

Effective population size

- How is Ne defined in terms of allele frequencies and coalescence rates?
- When is N and Ne the same?
- How do you compute the Ne for a population with fluctuating size, or for different numbers of (reproducing) males and females?
- Why do we use the harmonic mean for coalescence rate?
- Why is the Ne of the X chromosome 3/4 of the autosomal one?

Coalescent with recombination

Week 2

Kasper Munch

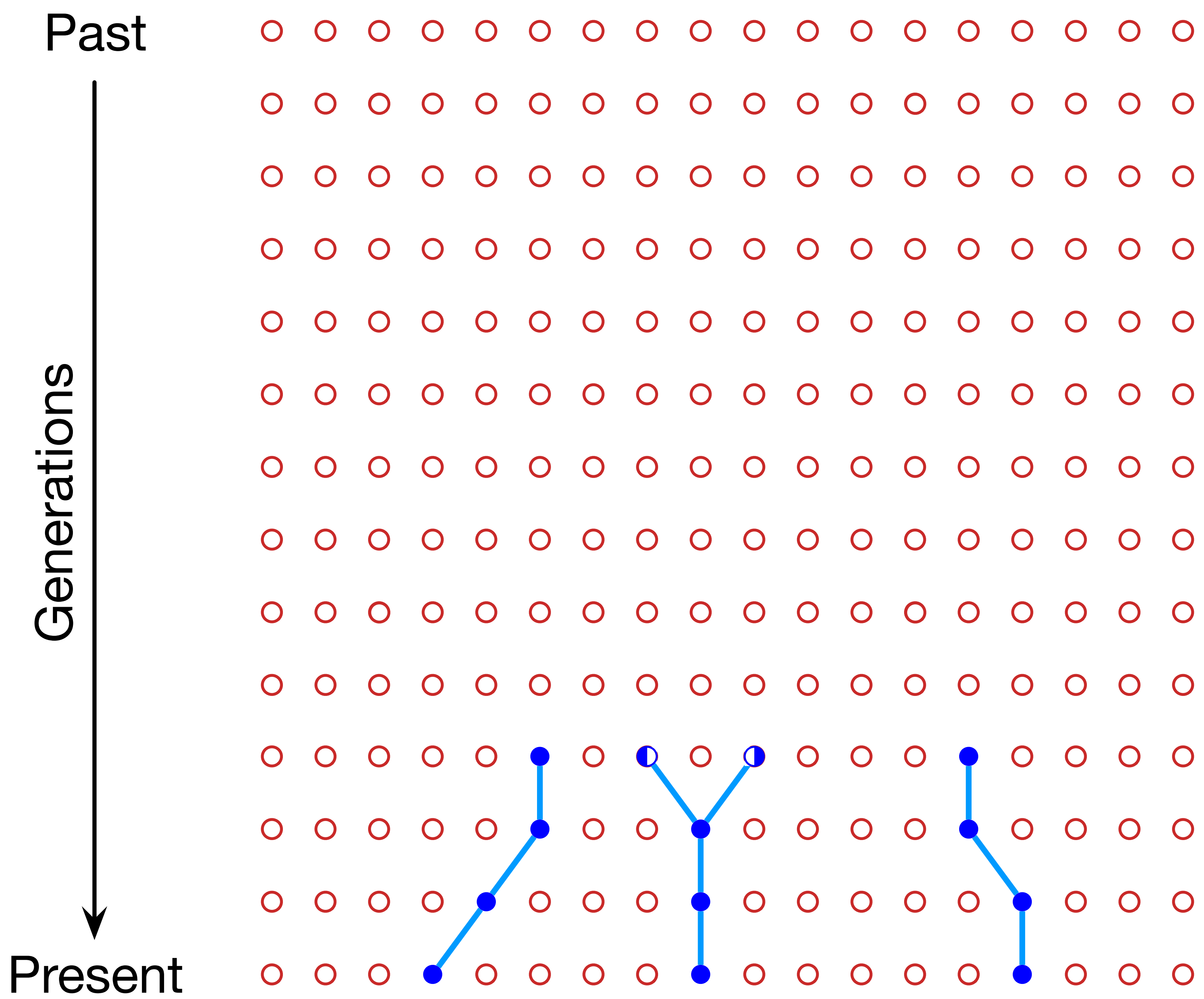
Coalescent with recombination

A recombination makes a germ cell with two parts, each inherited from different grandparents.

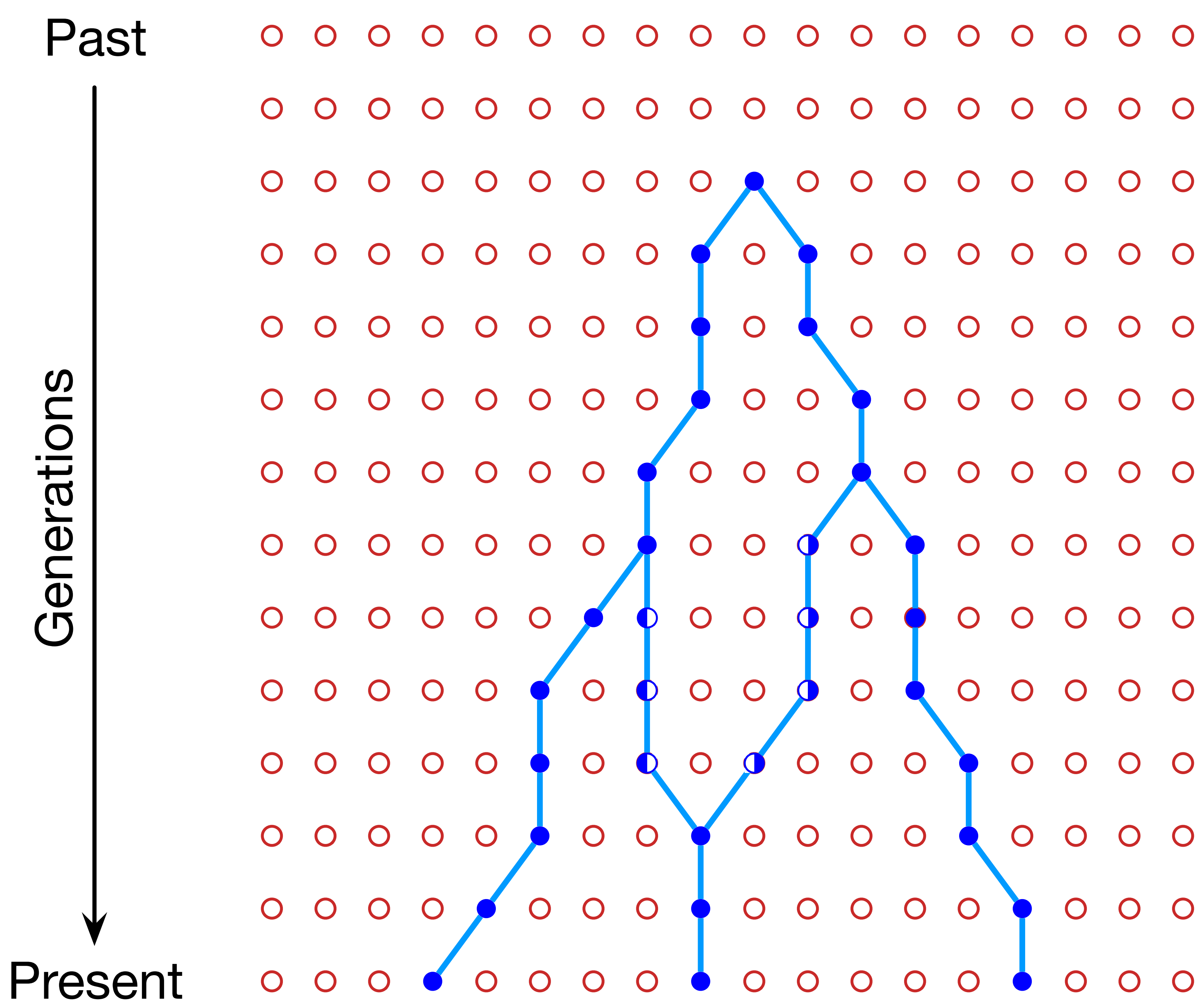
Each part is carried by different people in the previous generation.

Who are your ancestors?

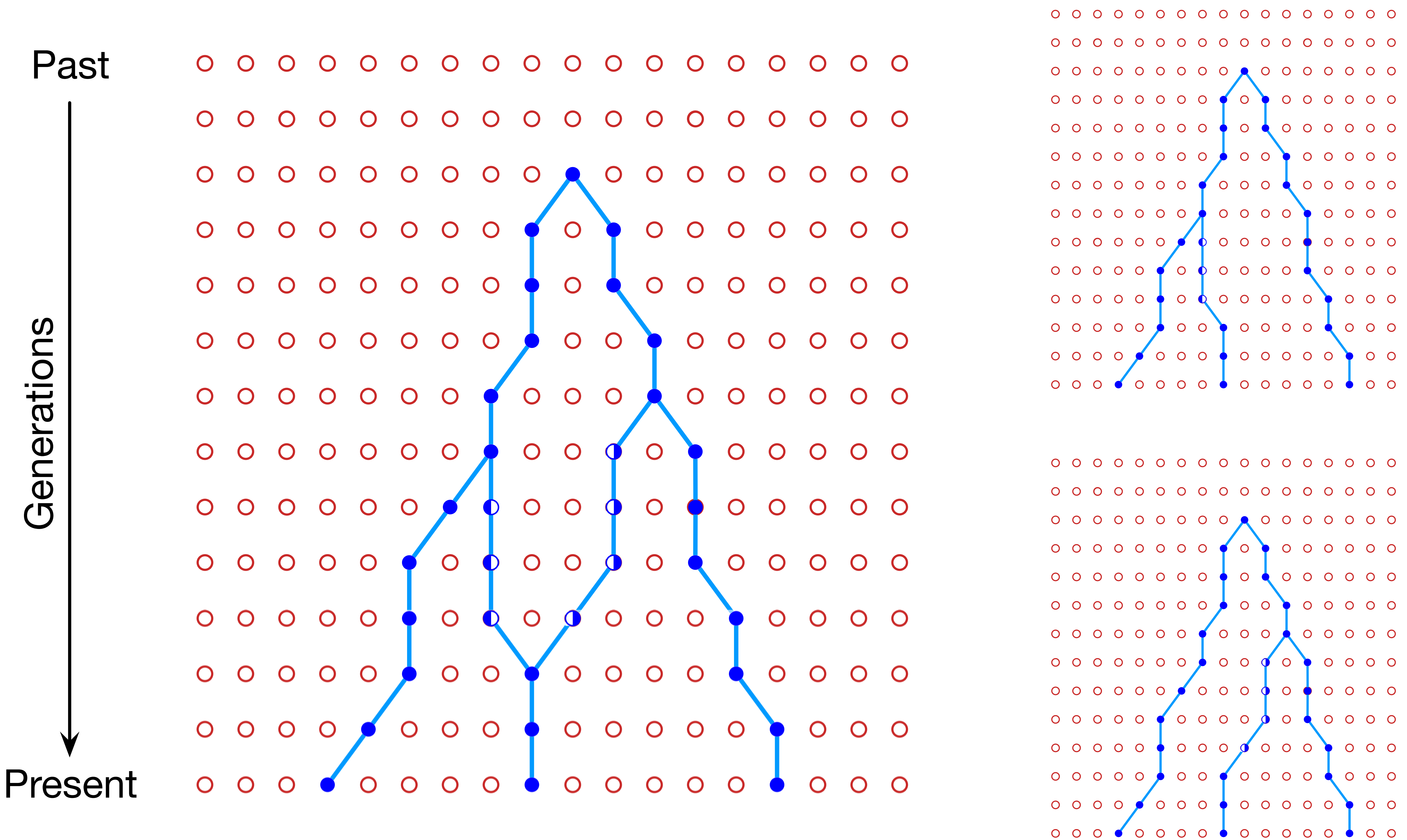
Coalescent and recombination



Coalescent and recombination



Coalescent and recombination



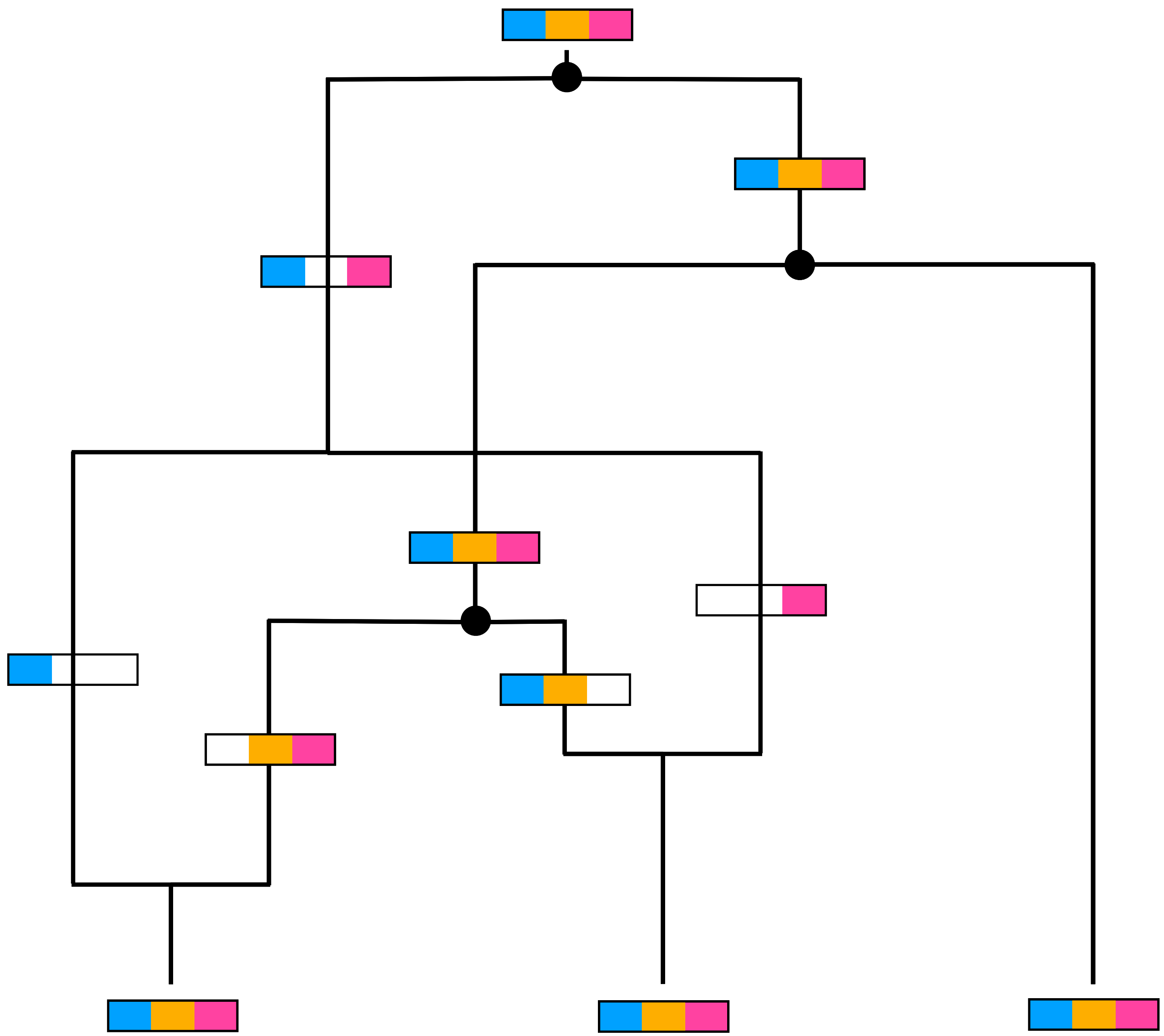
Ancestral recombination graph

Week 3

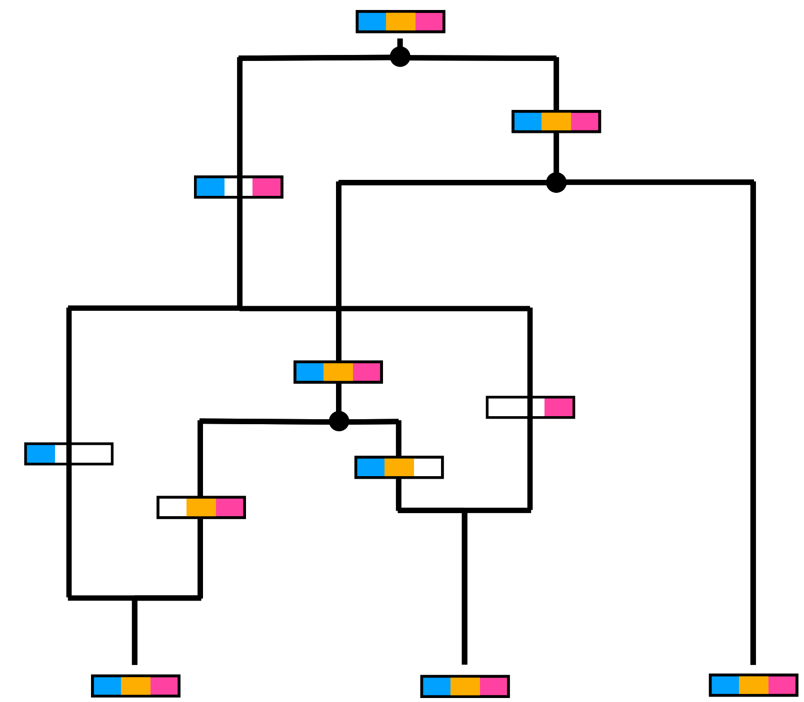
Kasper Munch

Ancestral recombination graph (ARG)

Ancestral Recombination Graph
(ARG)



Depicting ancestry on an ARG



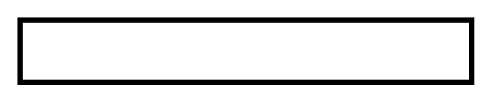
Ancestral sequence:

“A sequence segment in at least one *sampled* sequence carried by an *ancestor* sequence“

Examples:

Sequence ancestral to samples, with three sections separated by recombinations:

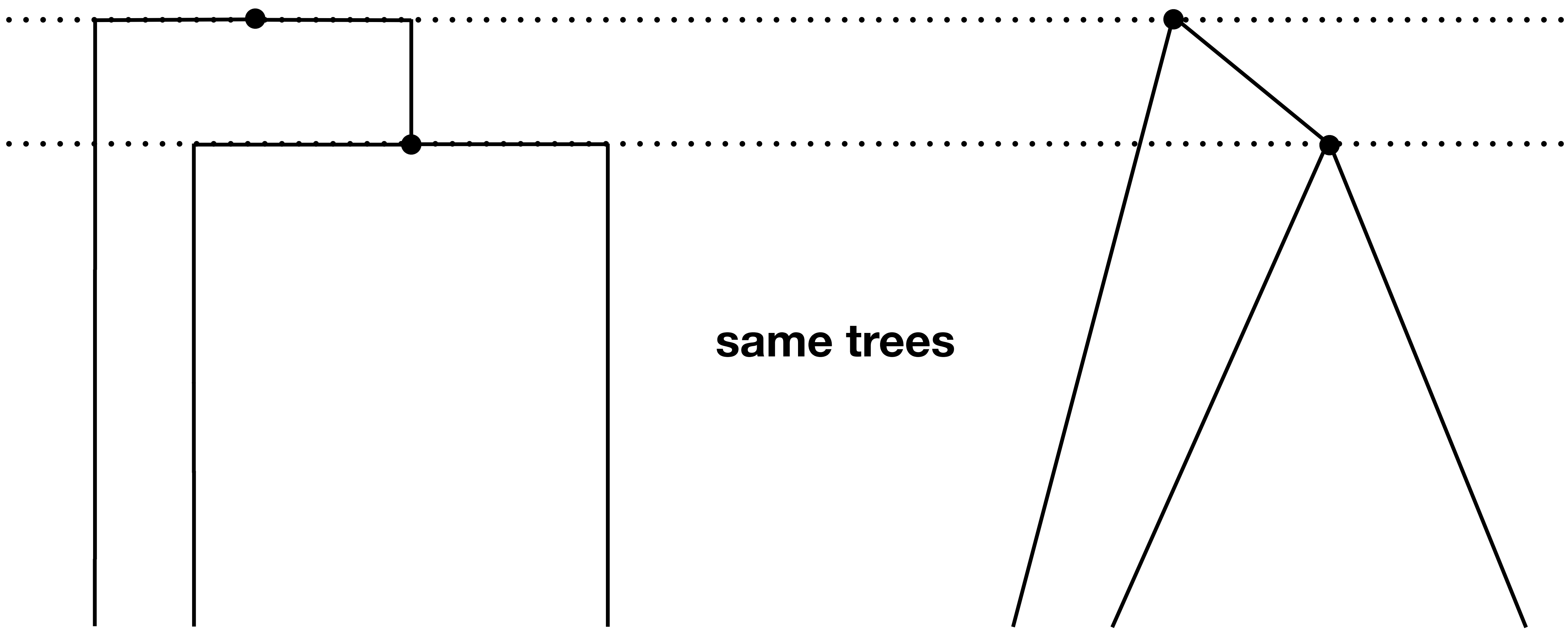
Sequence not ancestral to samples:



Sequence with two sections ancestral to samples and one part not ancestral:

Depicting trees

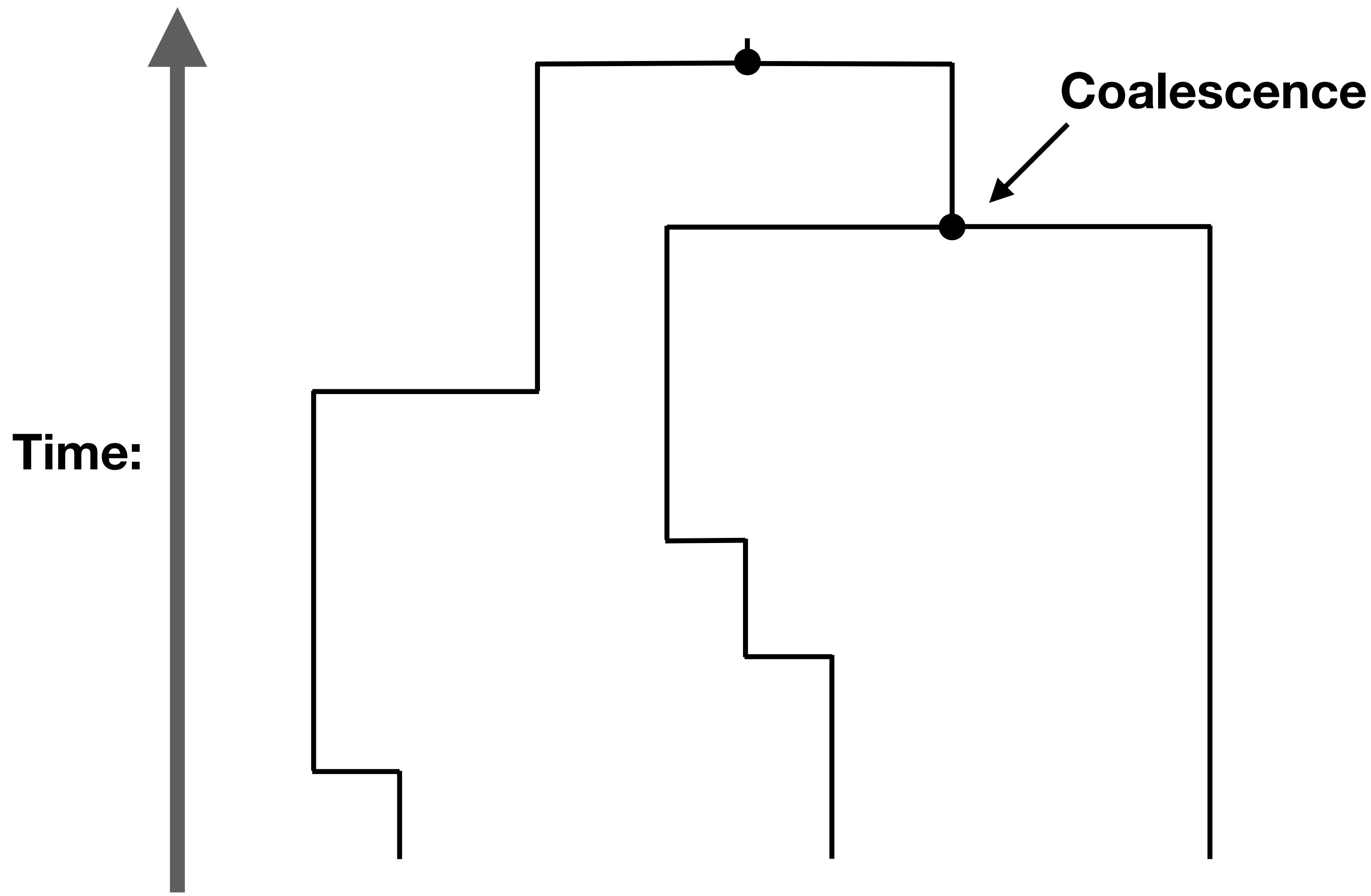
Dendrograms and classic trees



Only *vertical* distance represents time

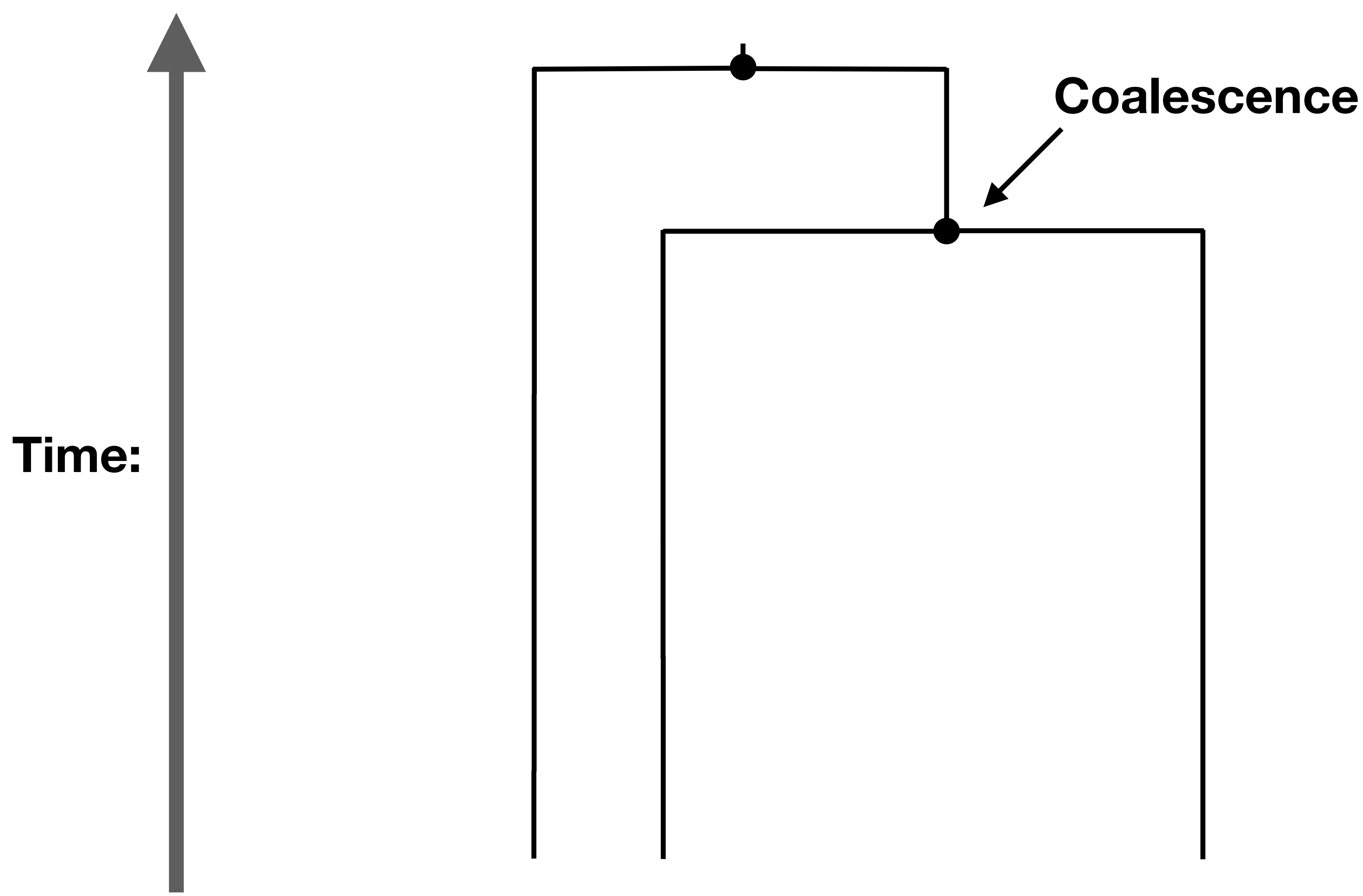
Depicting trees

Strange shapes



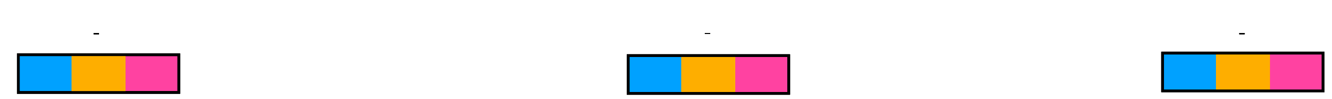
Depicting trees

Strange shapes



The ARG process

Tracing ancestry of the sampled sequences

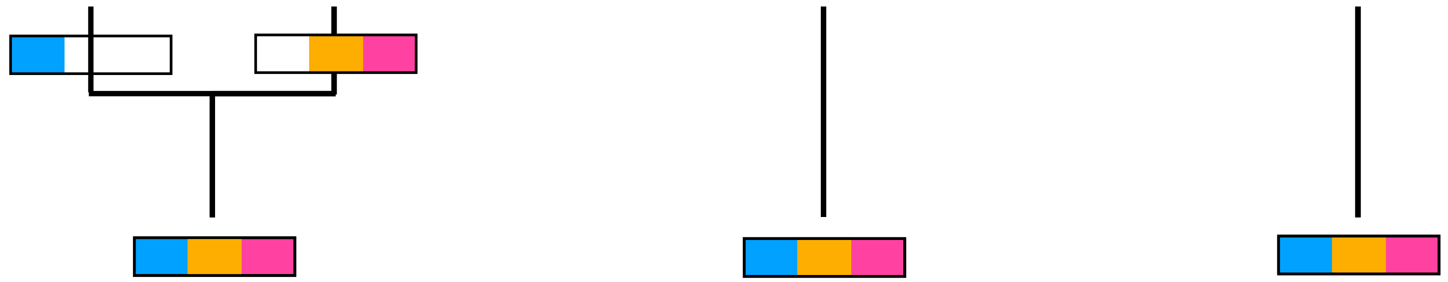


Sampled sequences

The ARG process

Tracing ancestry of the sampled sequences

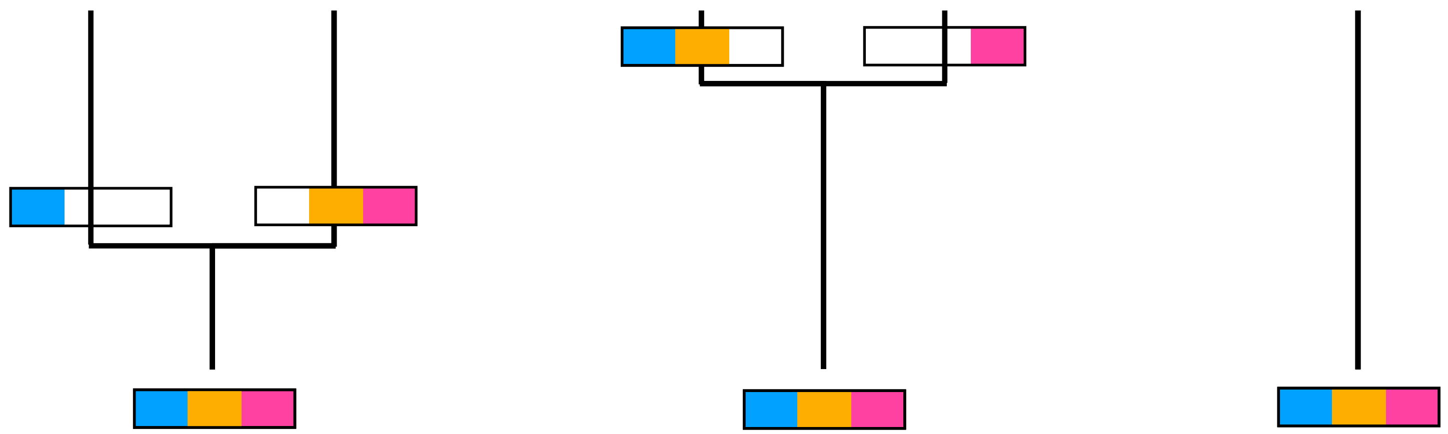
A recombination between
blue and yellow segment



The ARG process

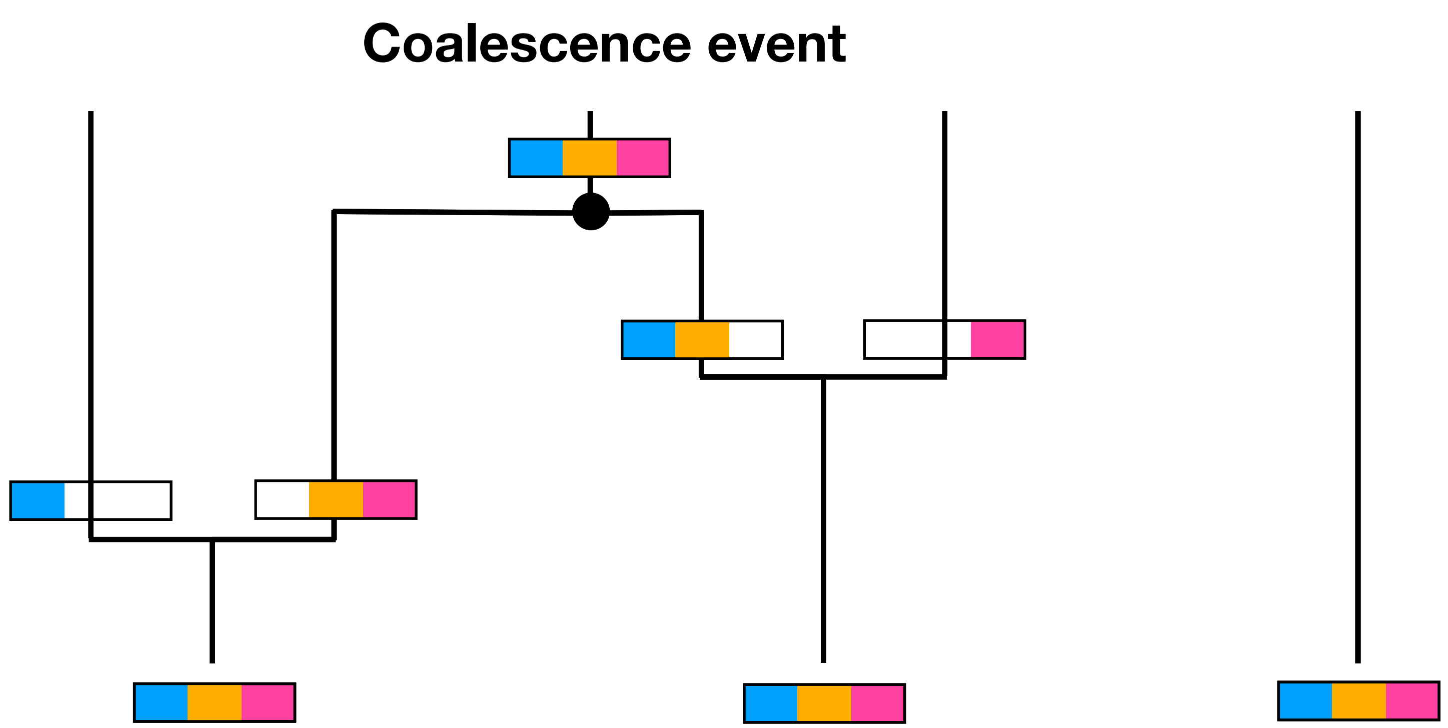
Tracing ancestry of the sampled sequences

A recombination between
yellow and pink segment



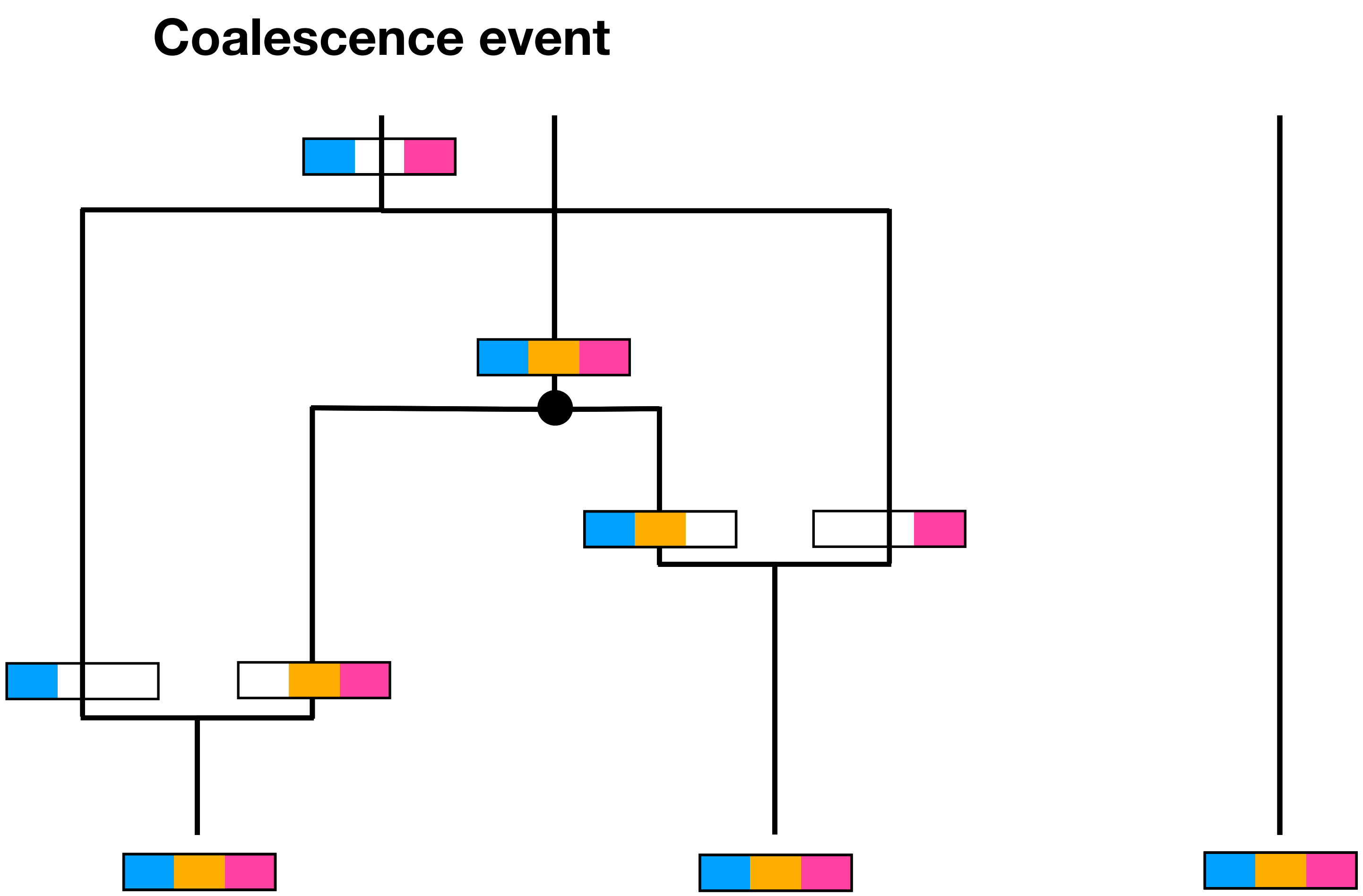
The ARG process

Tracing ancestry of the sampled sequences



The ARG process

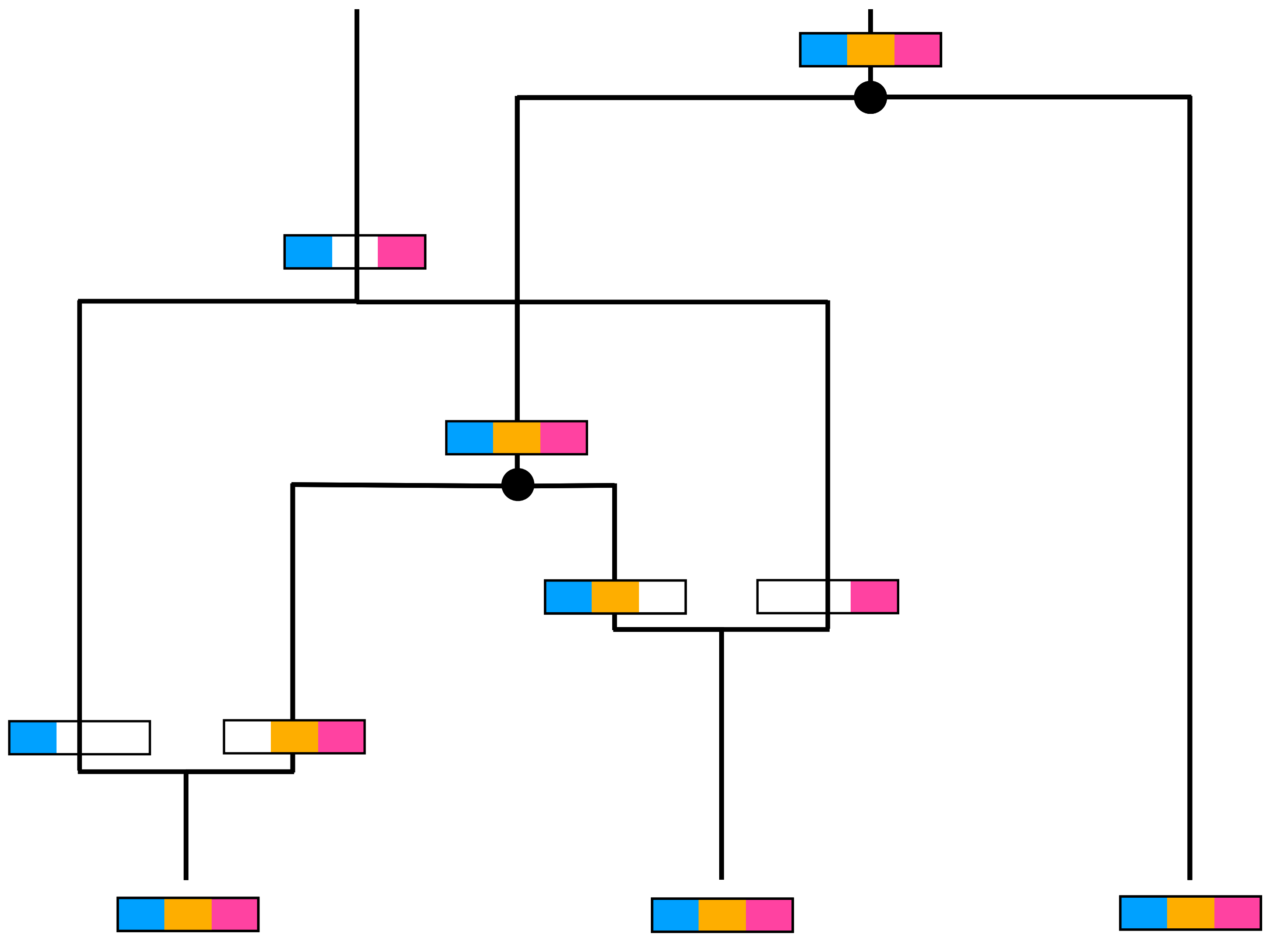
Tracing ancestry of the sampled sequences



The ARG process

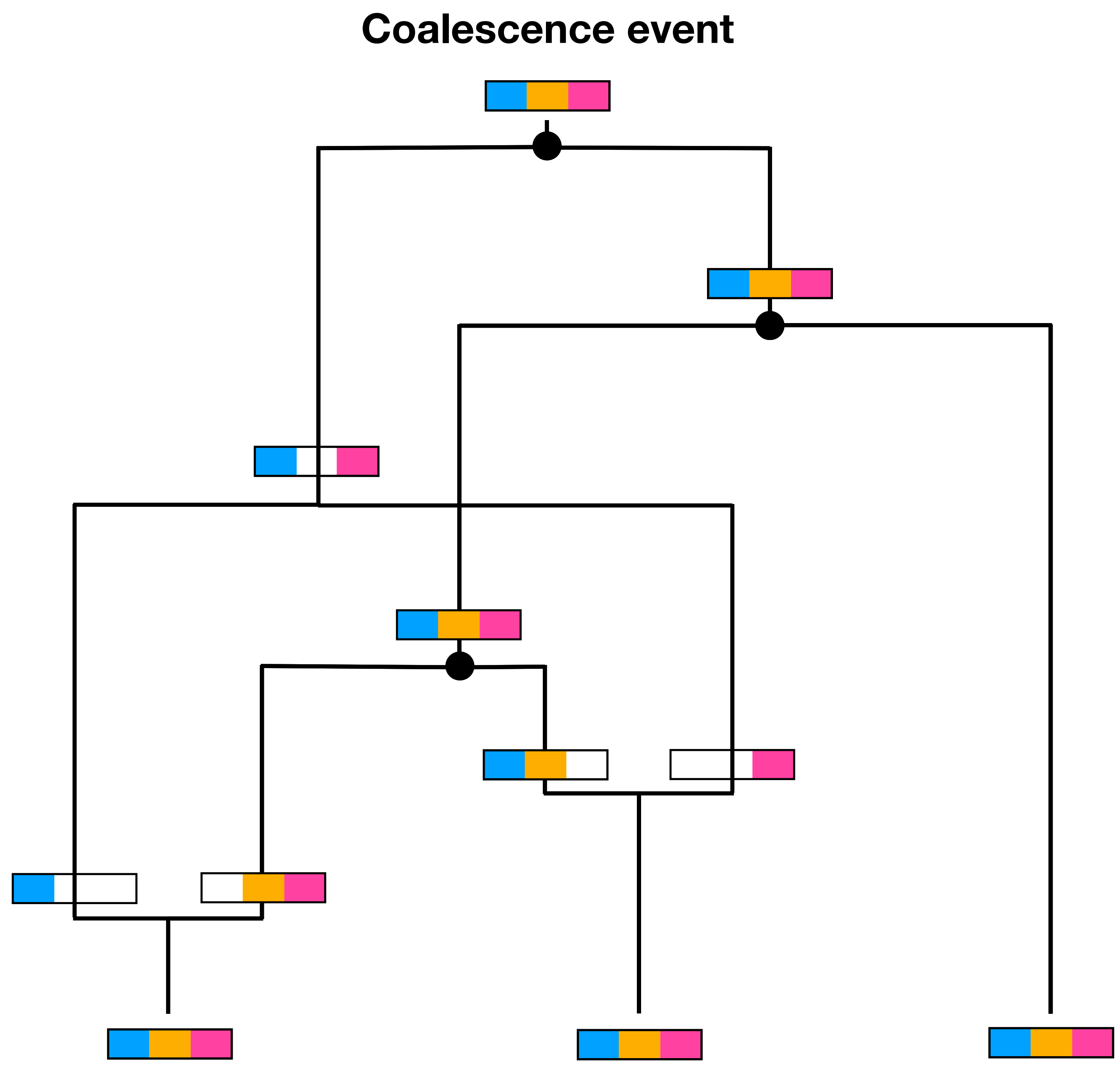
Tracing ancestry of the sampled sequences

Coalescence event



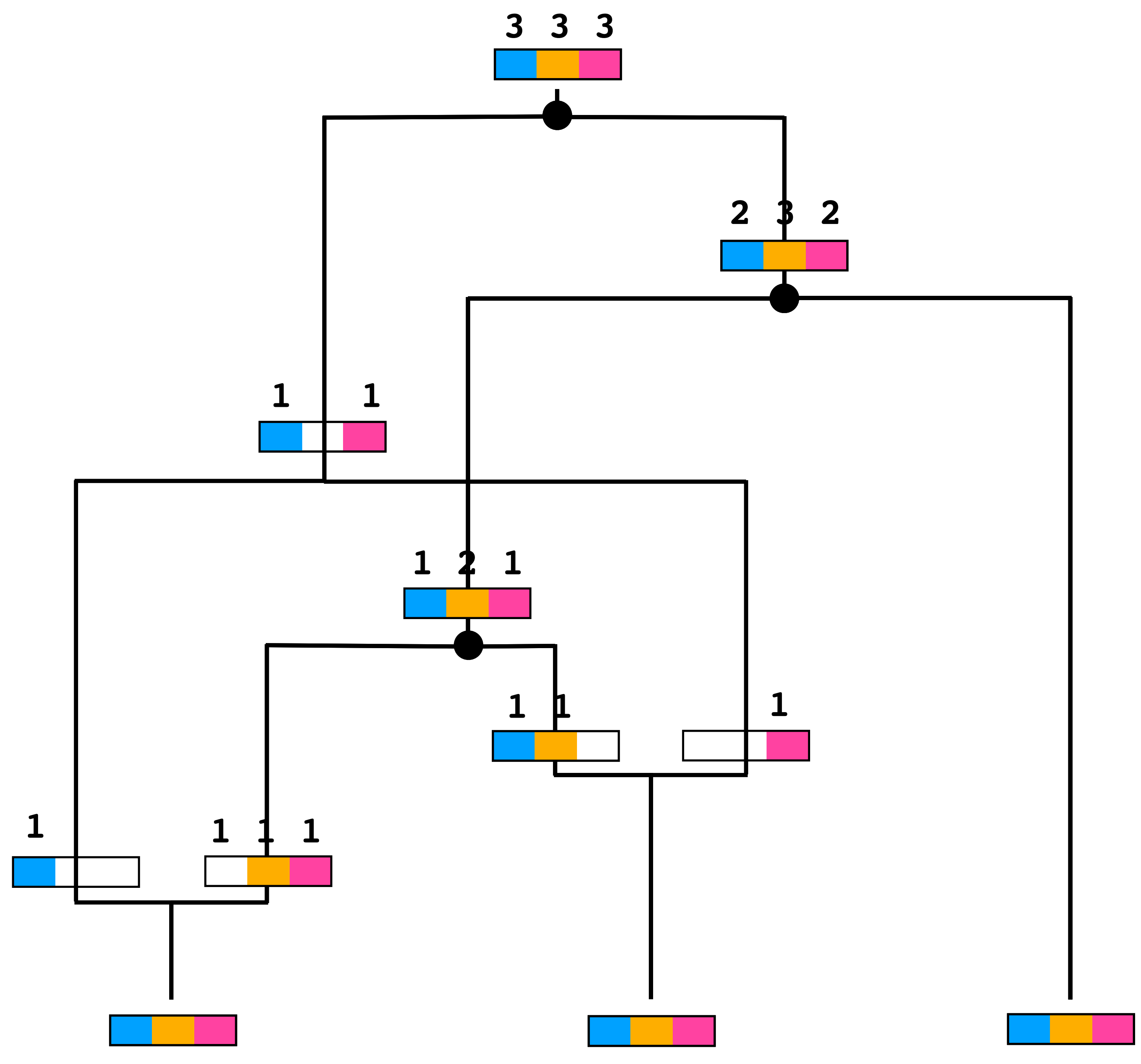
The ARG process

Tracing ancestry of the sampled sequences



The ARG process

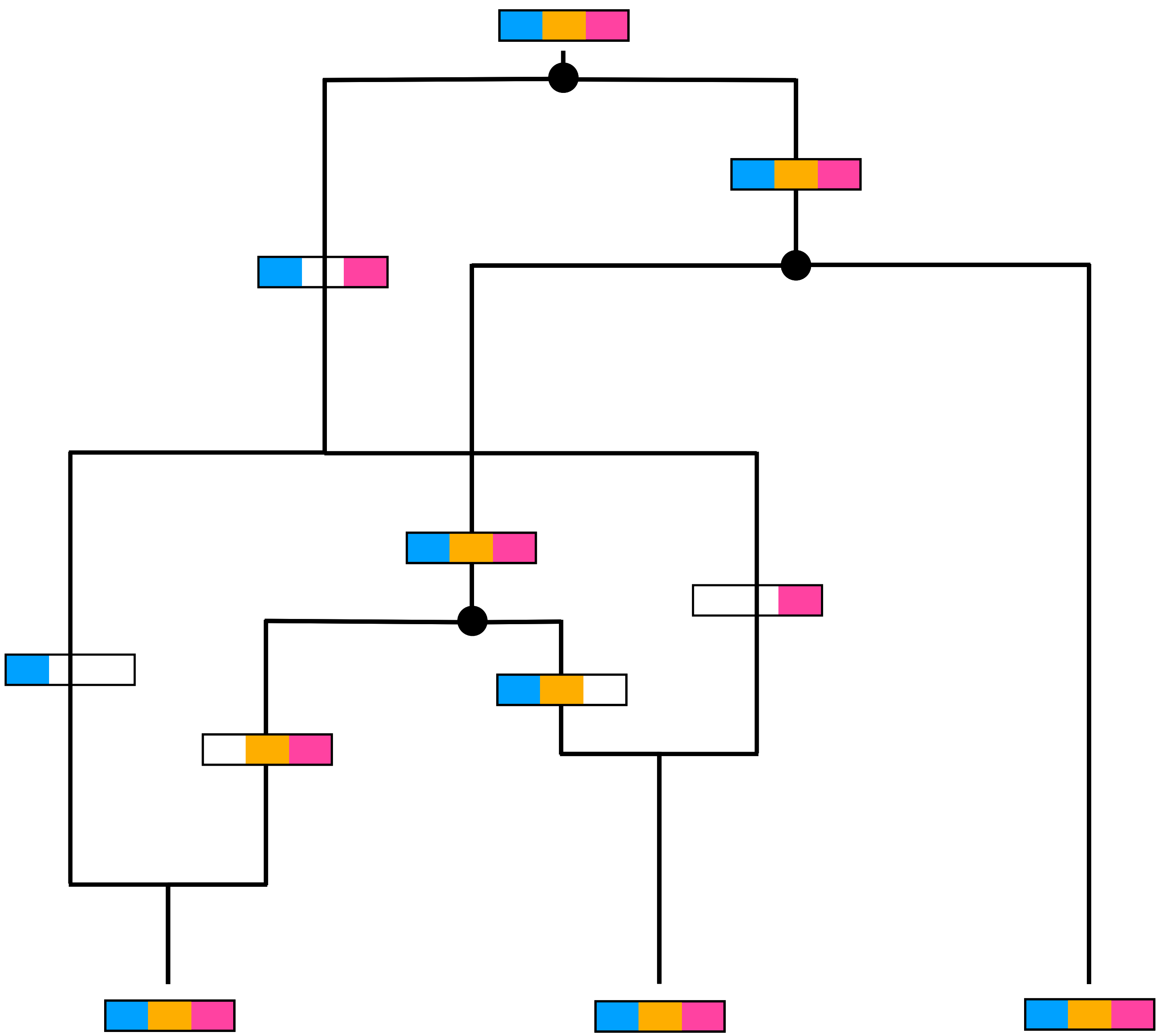
Nr. descendants of segments



Trees in the ARG

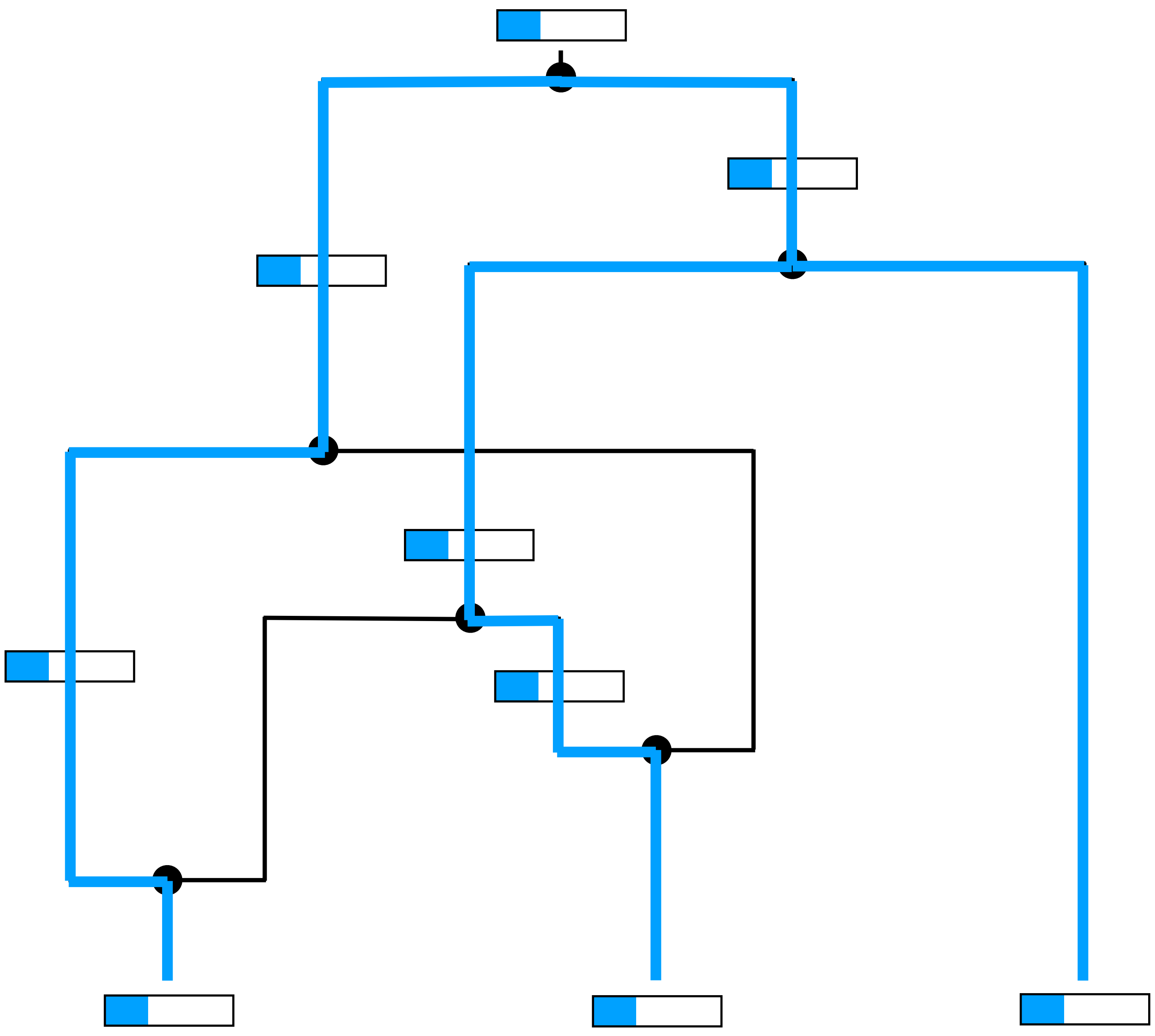
Trees in the ARG

The ancestry of each segment is a tree



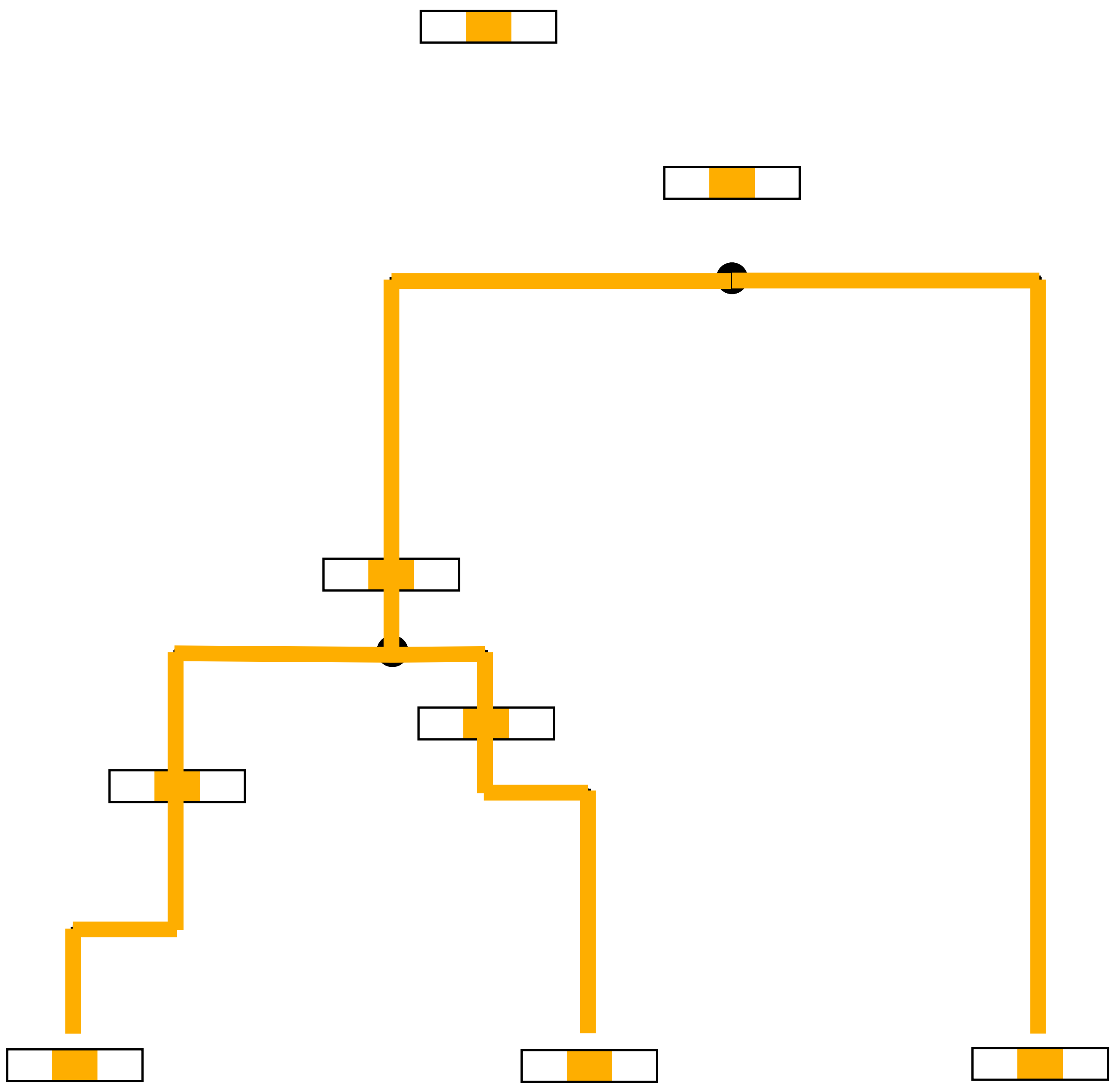
Trees in the ARG

The ancestry of each segment is a tree



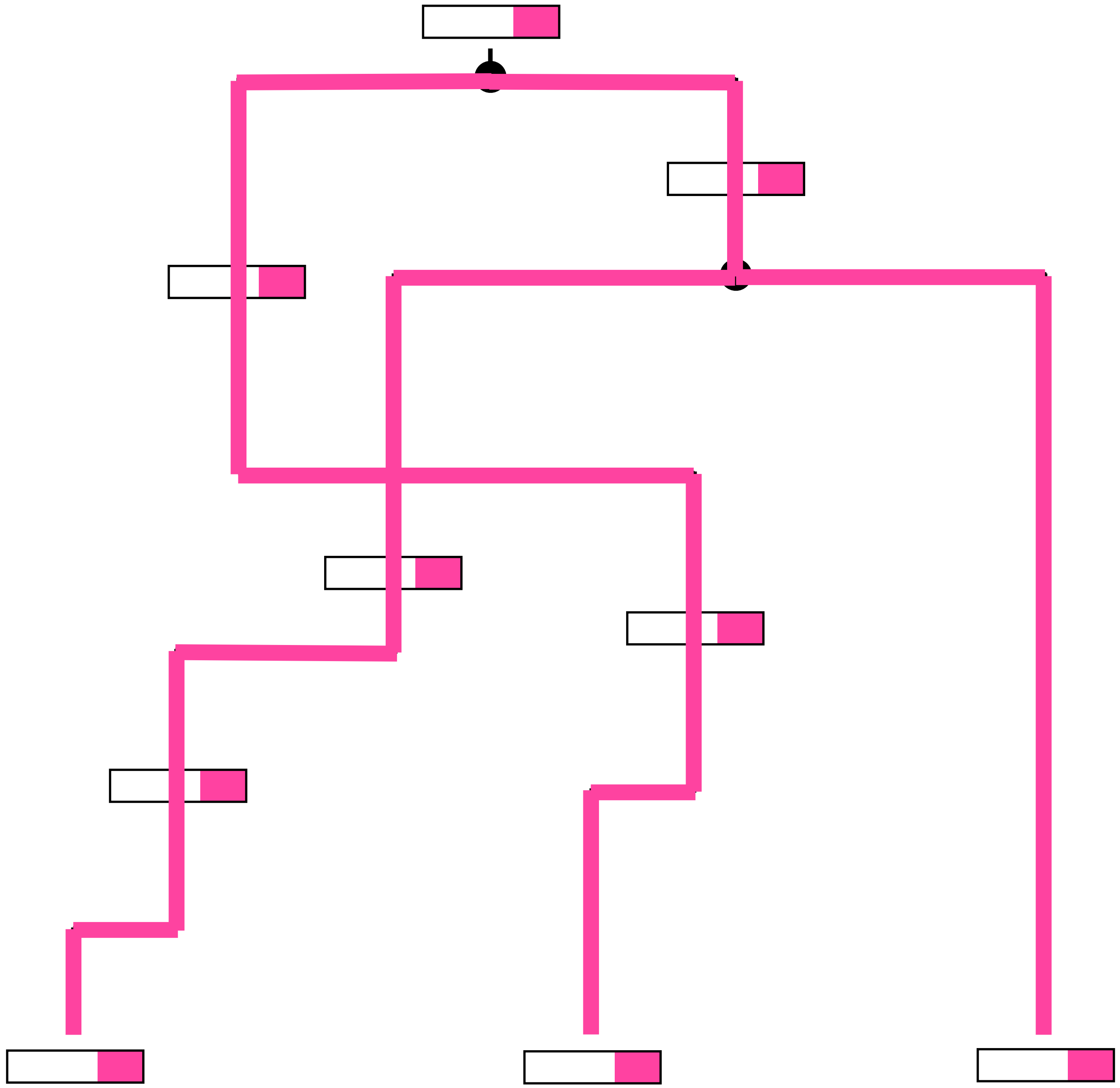
Trees in the ARG

The ancestry of each segment is a tree



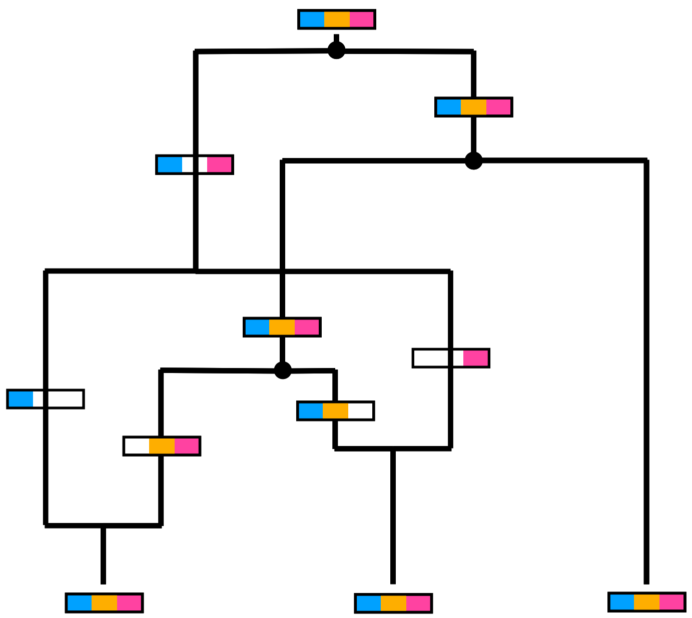
Trees in the ARG

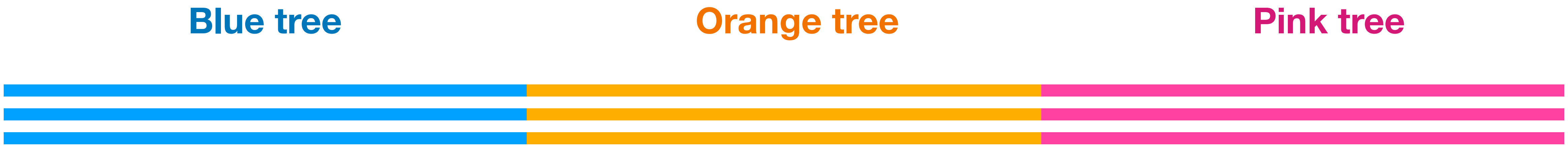
The ancestry of each segment is a tree



Trees in the ARG

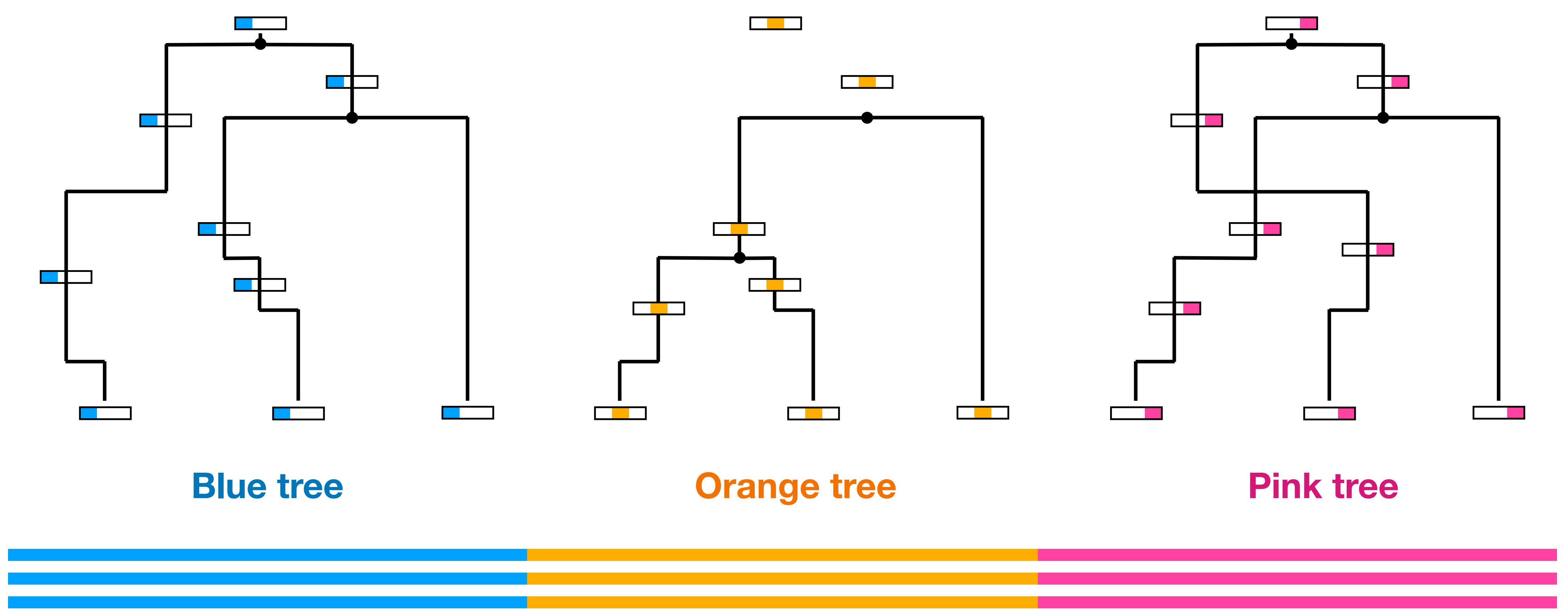
The ancestry of each segment is a tree





Trees in the ARG

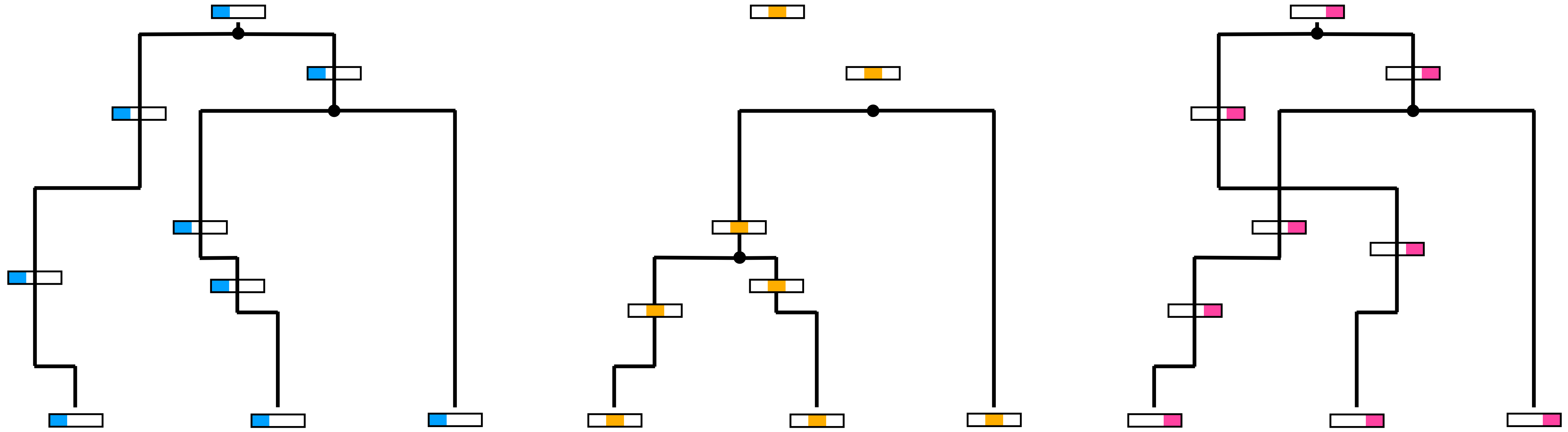
The ancestry of each segment is a tree



Trees in the ARG

Each tree is a coalescent tree

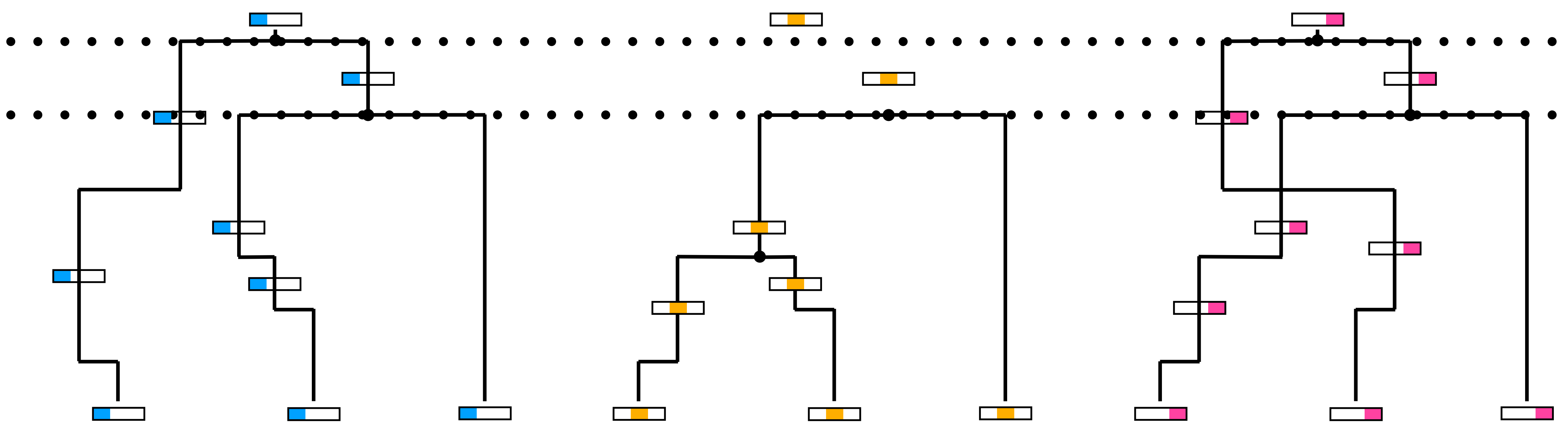
- Each tree obeys the standard coalescence process



Trees in the ARG

Each tree is a coalescent tree

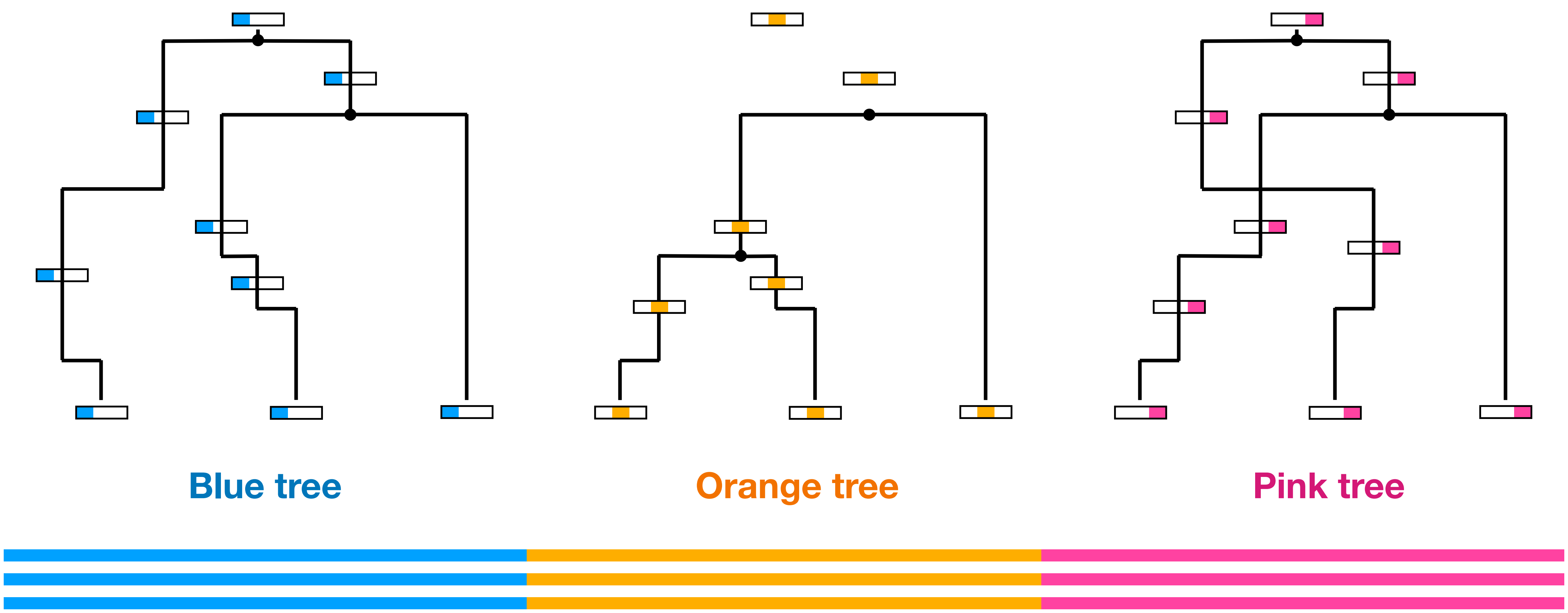
- Coalescence events can be shared by trees that are not neighbours. E.g., the last coalescence in the first and last tree:



Trees in the ARG

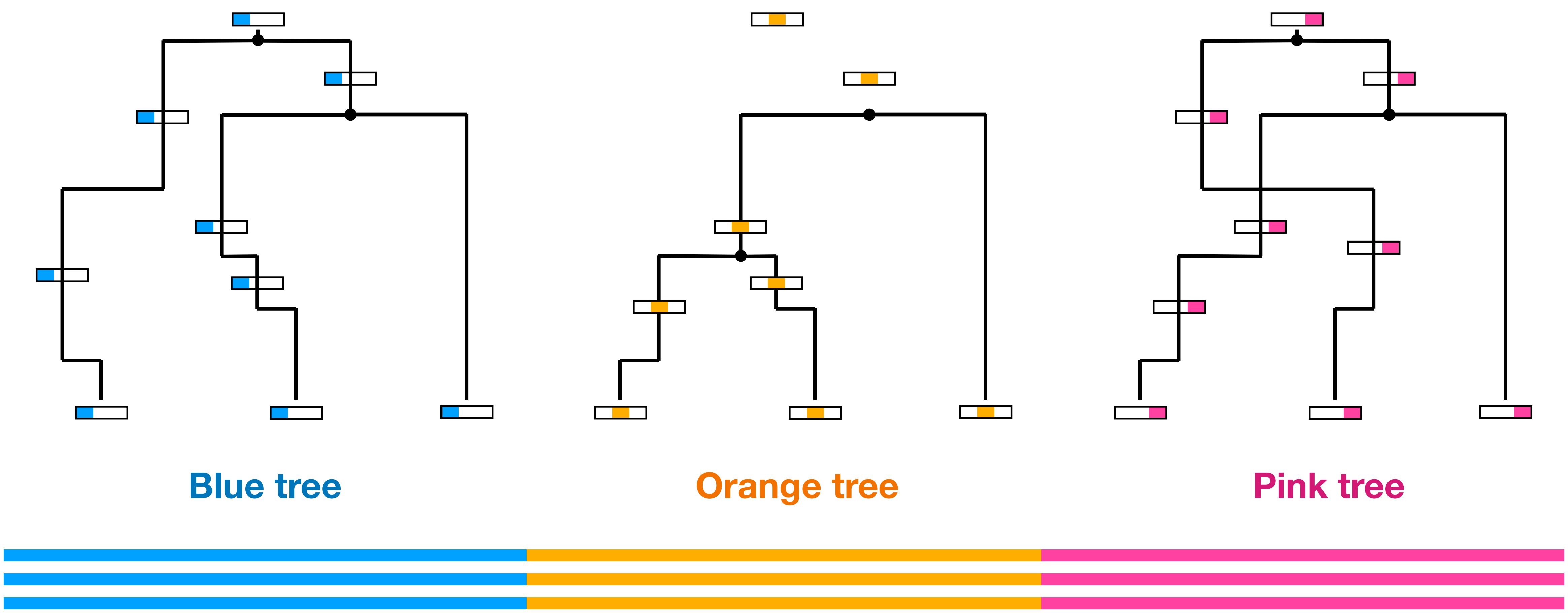
The ARG defines a collection of trees

- The ARG is made up of coalescence trees for consecutive segments along genome:

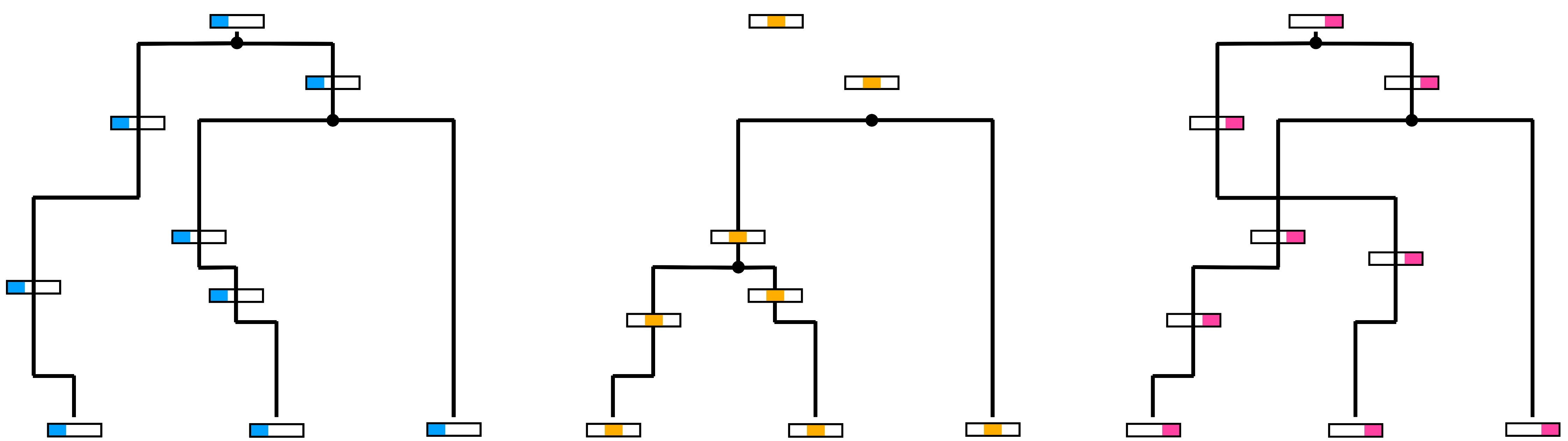


Trees in the ARG

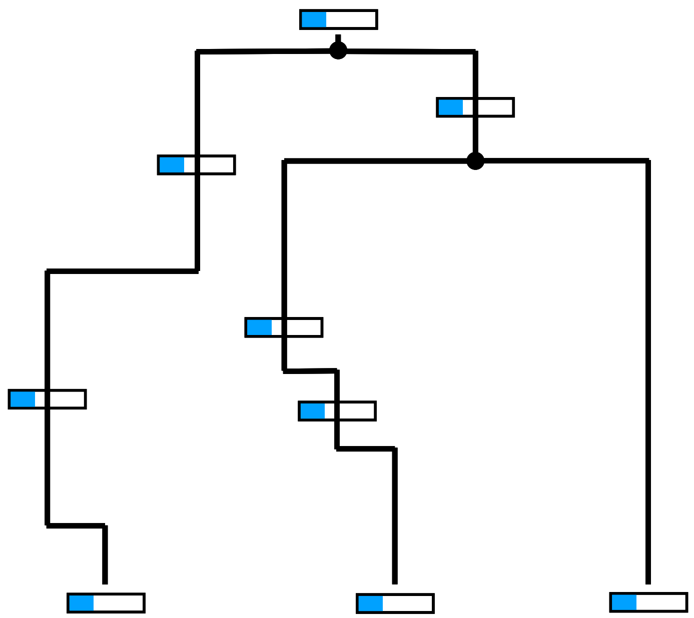
- Generating the ARG **backwards in time** produces individual coalescence trees for consecutive segments **along the genome**:

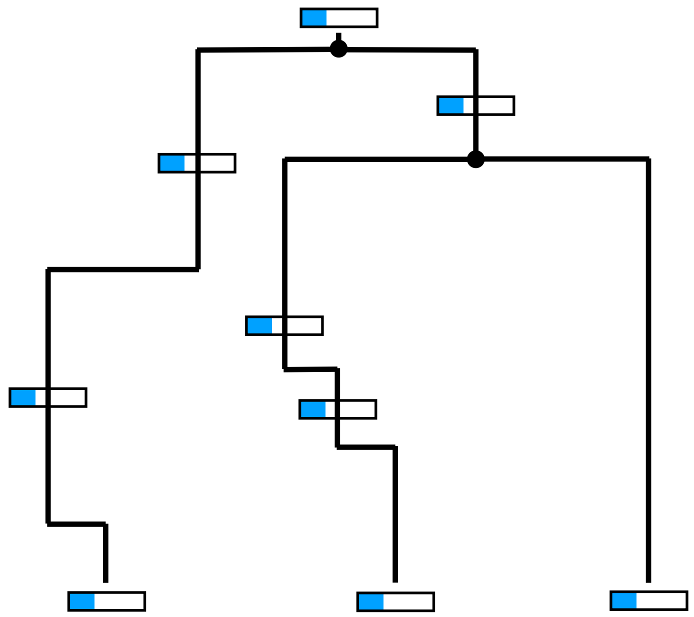


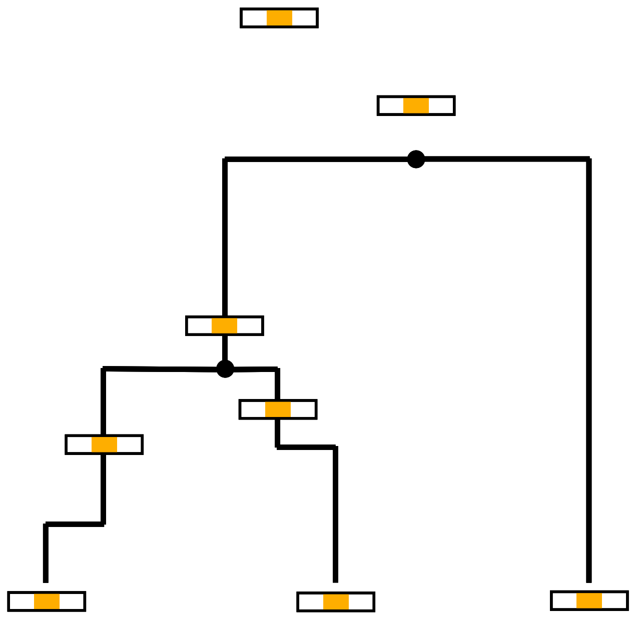
Adding trees, segment by segment

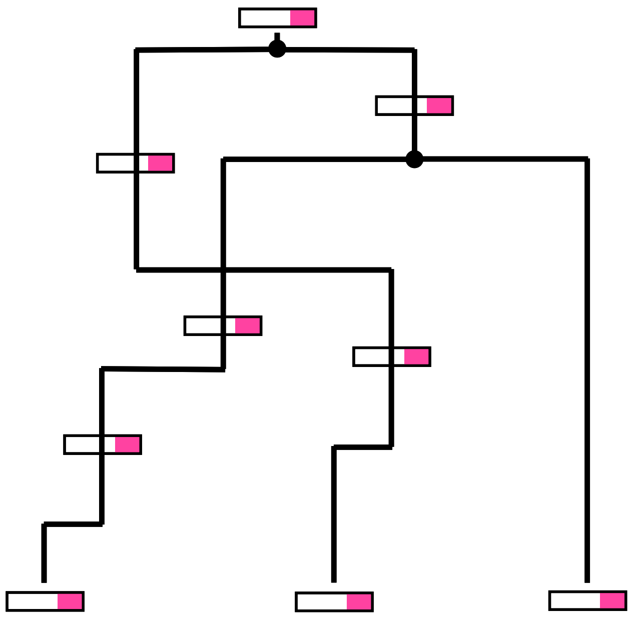


Adding trees, segment by segment

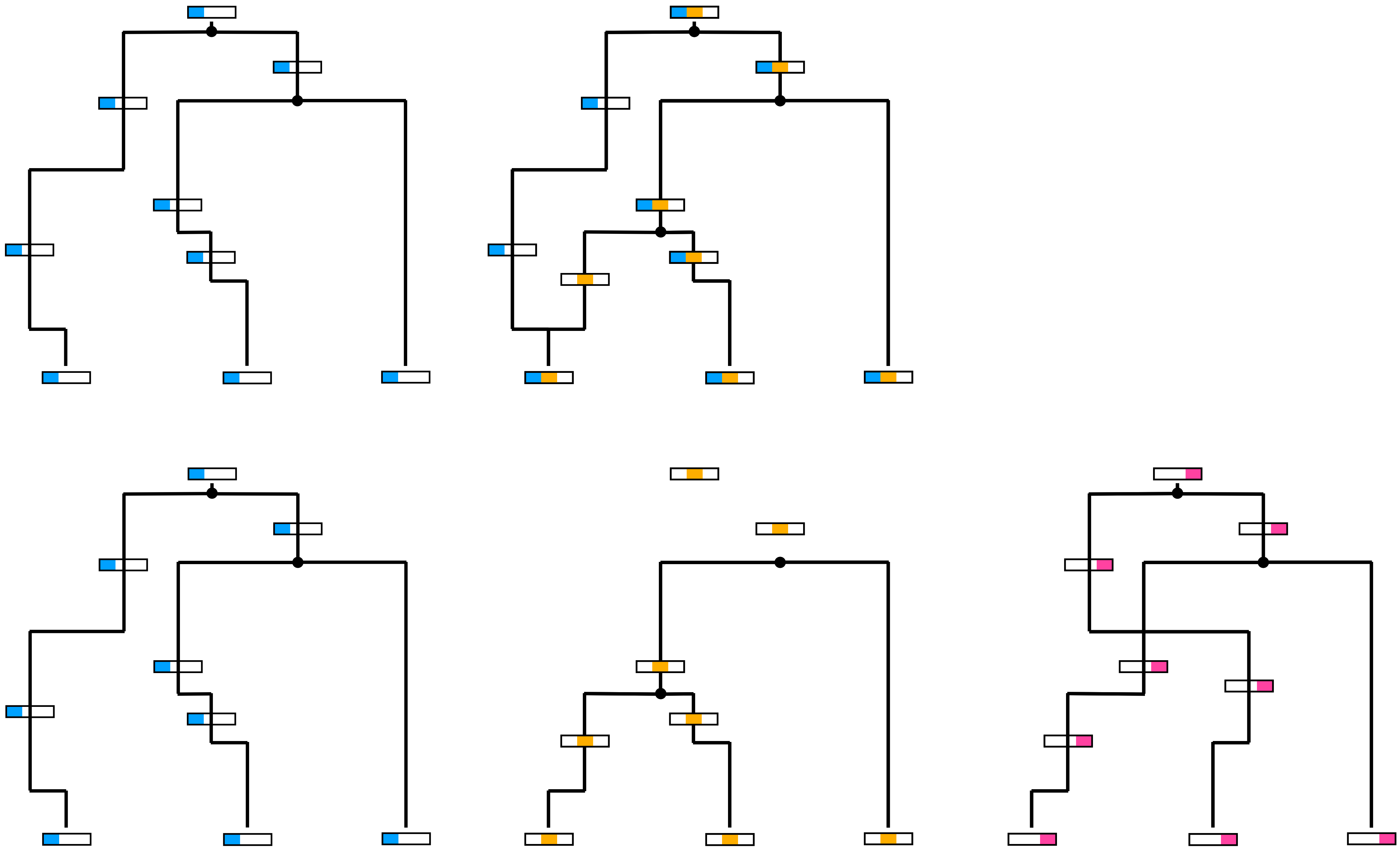




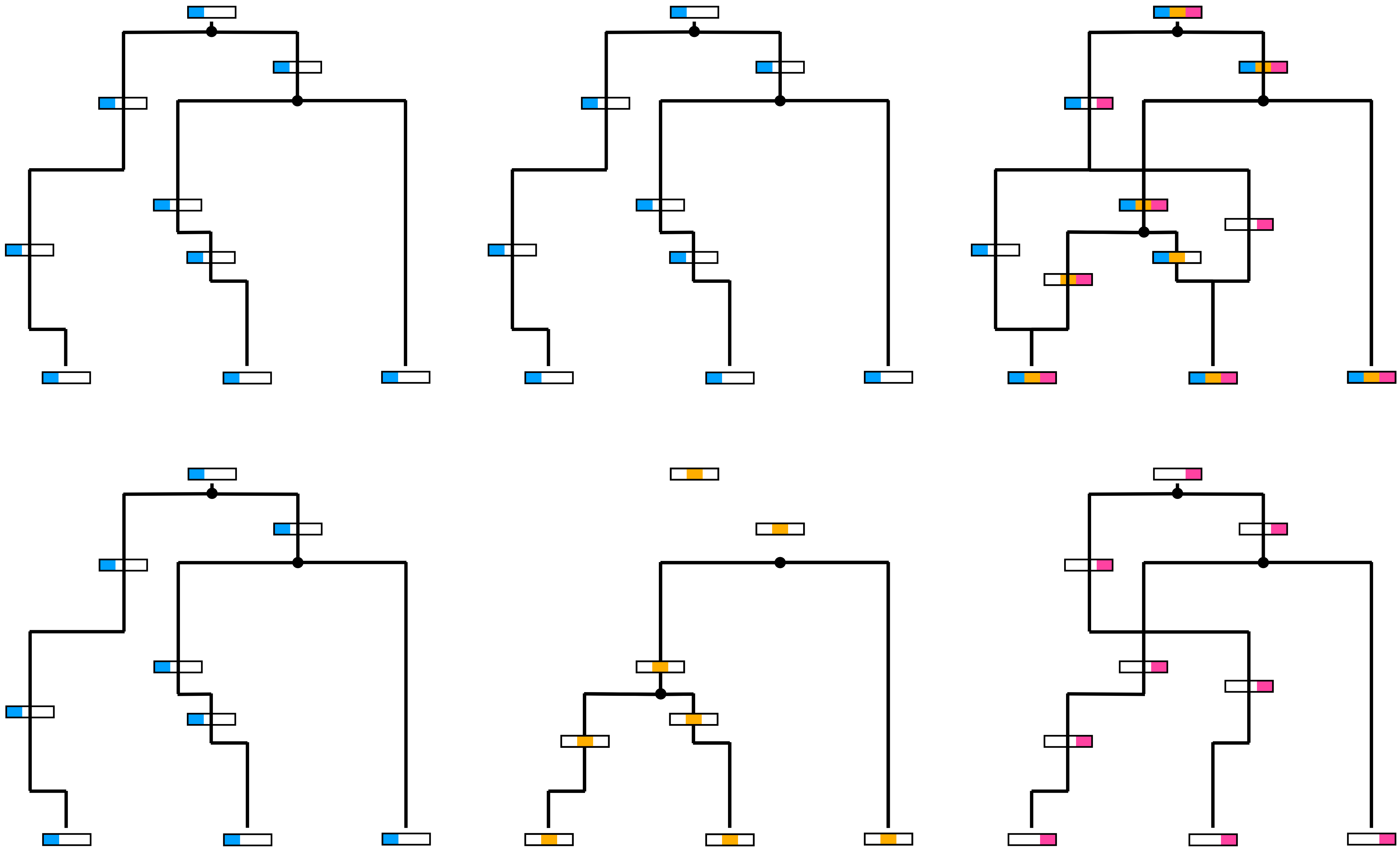




Adding trees, segment by segment



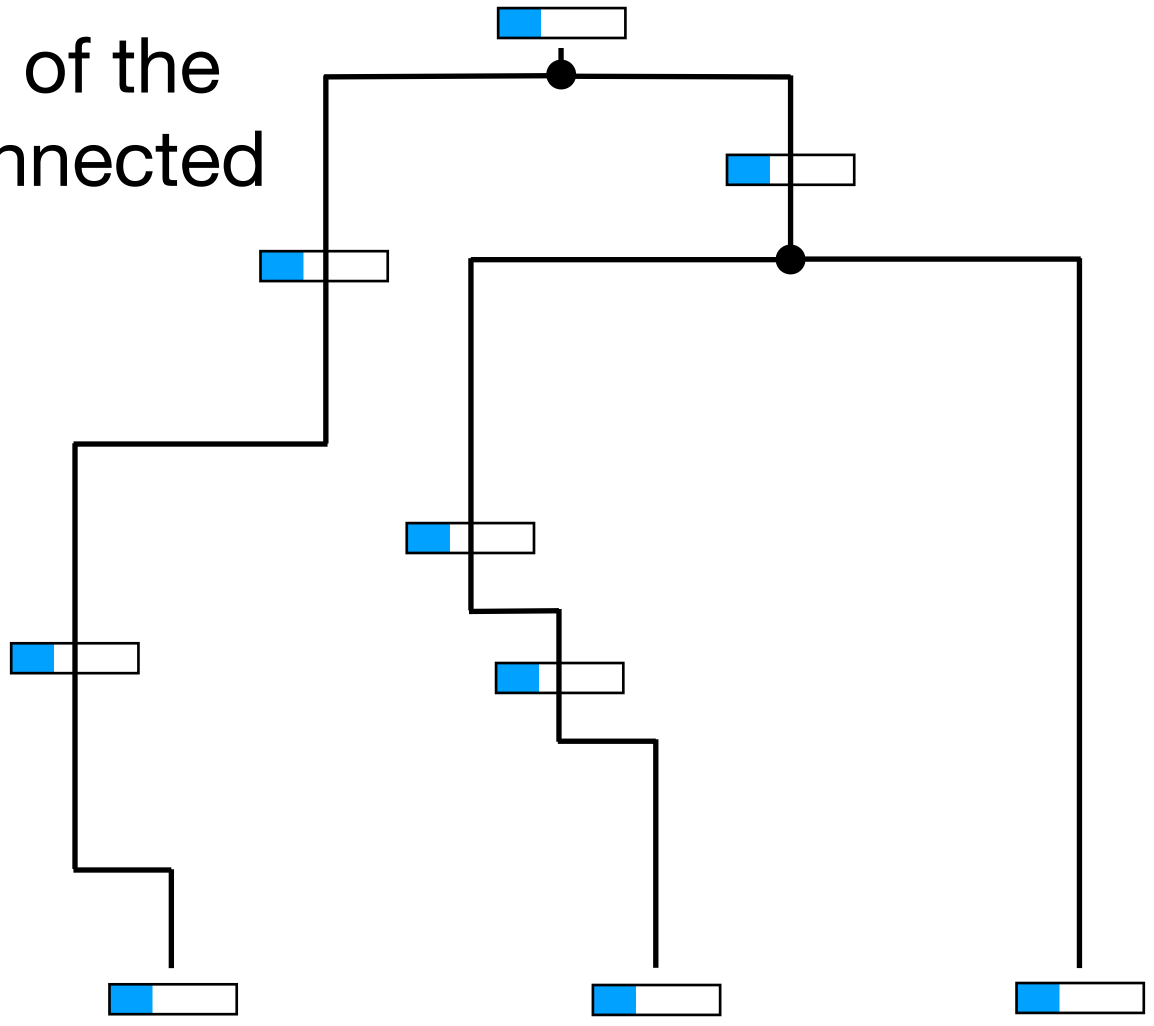
Adding trees, segment by segment



The ARG process

Along the sequence

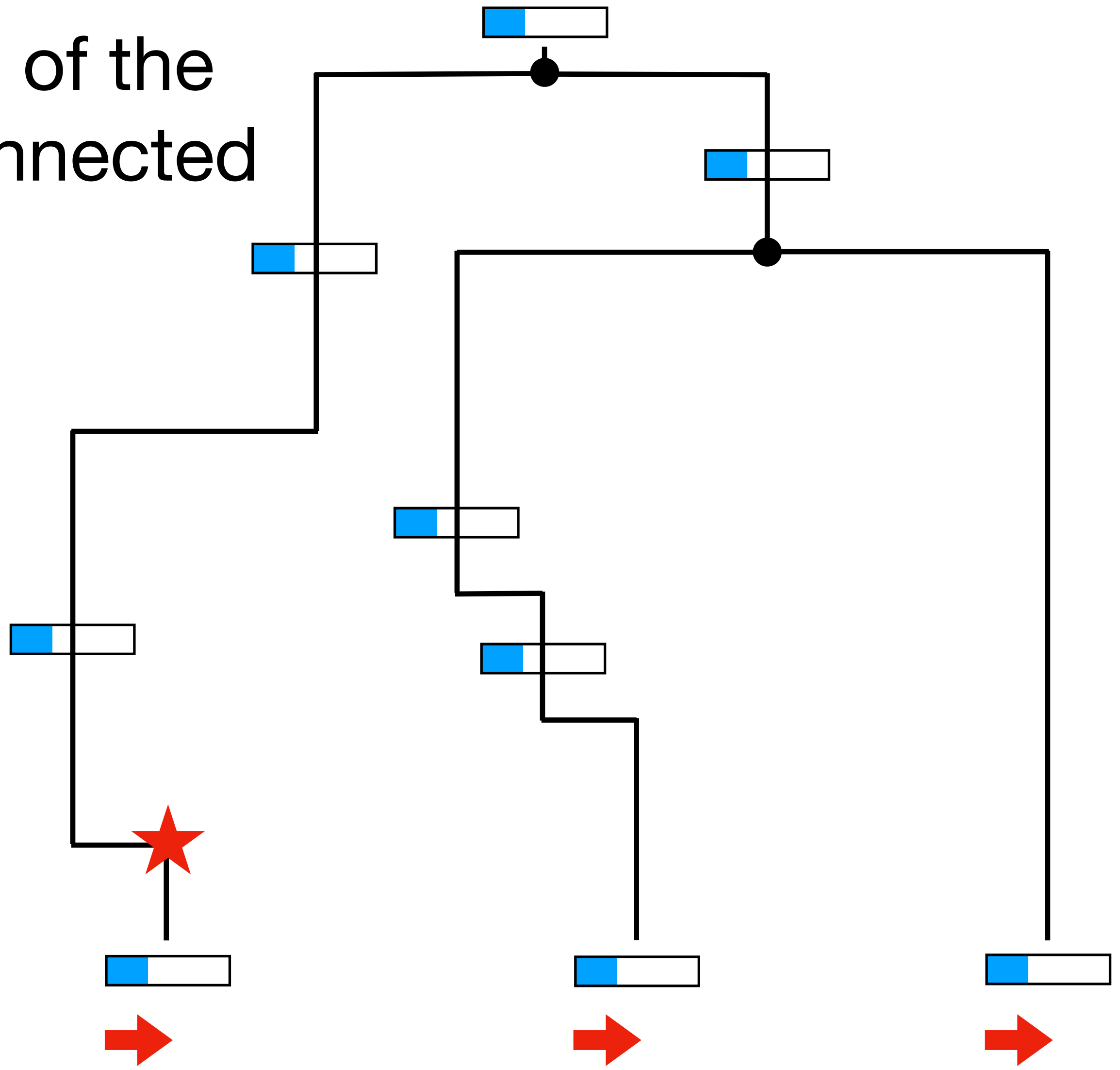
The leftmost parts of the sequences are connected by the this tree:



The ARG process

Along the sequence

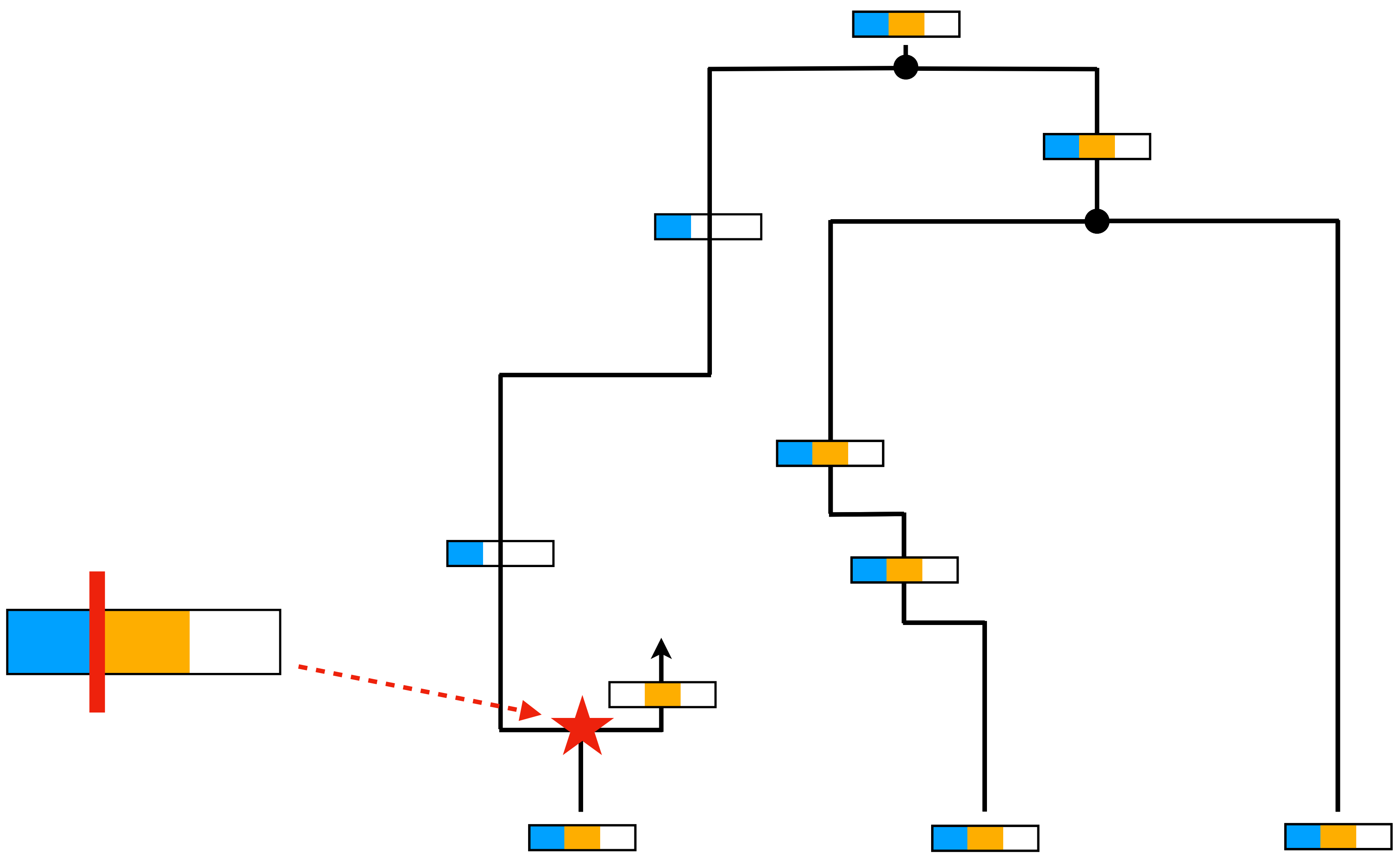
The leftmost parts of the sequences are connected by the this tree:



Move **right** until you encounter a recombination anywhere on this tree

The ARG process

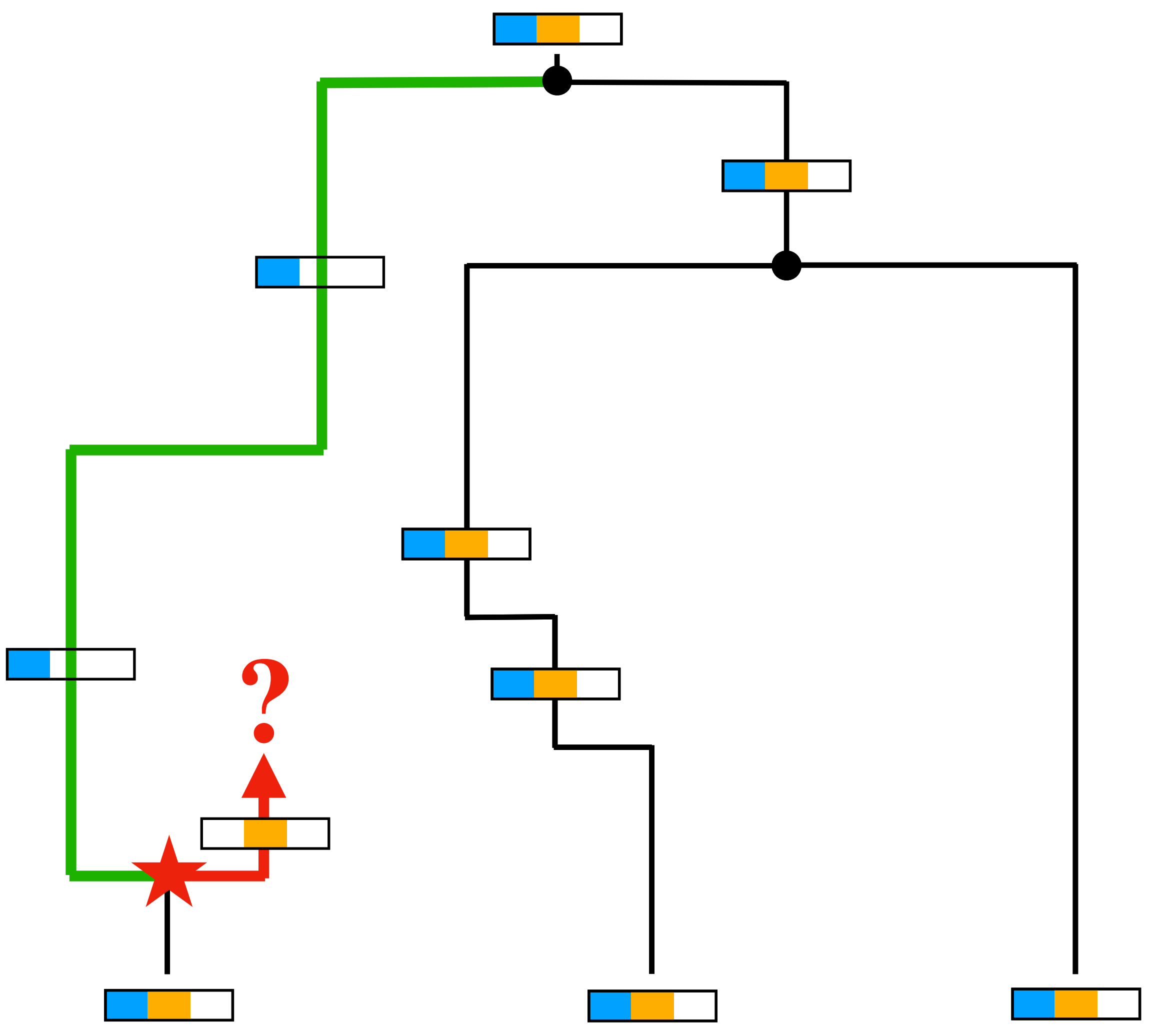
Along the sequence



A recombination event splits a sequence into a the **blue** part and a subsequent **yellow** part carried by a separate lineage.

The ARG process

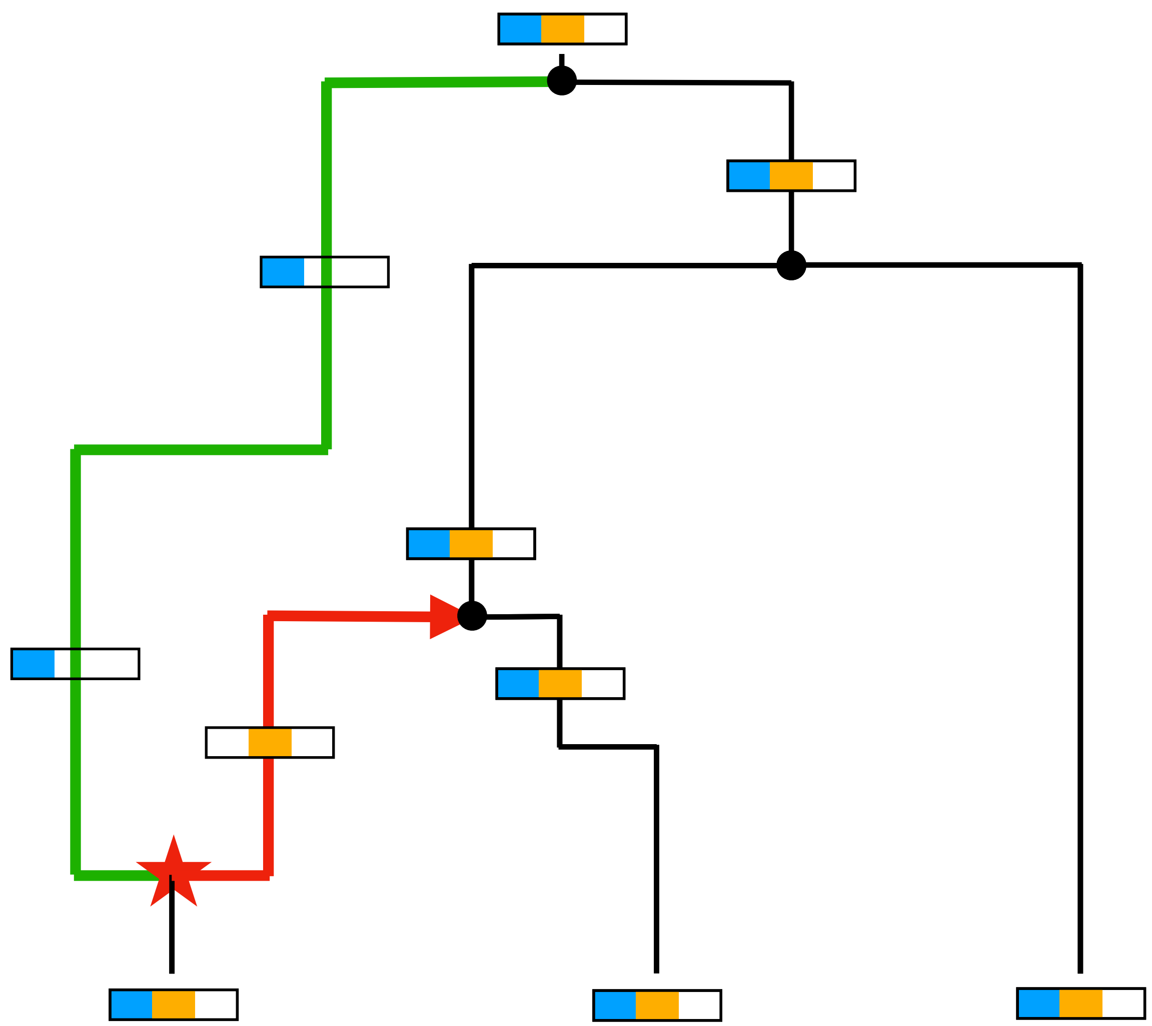
Along the sequence



A new **recombinant** lineage, separate from the **original** lineage, free to coalesce on its own.

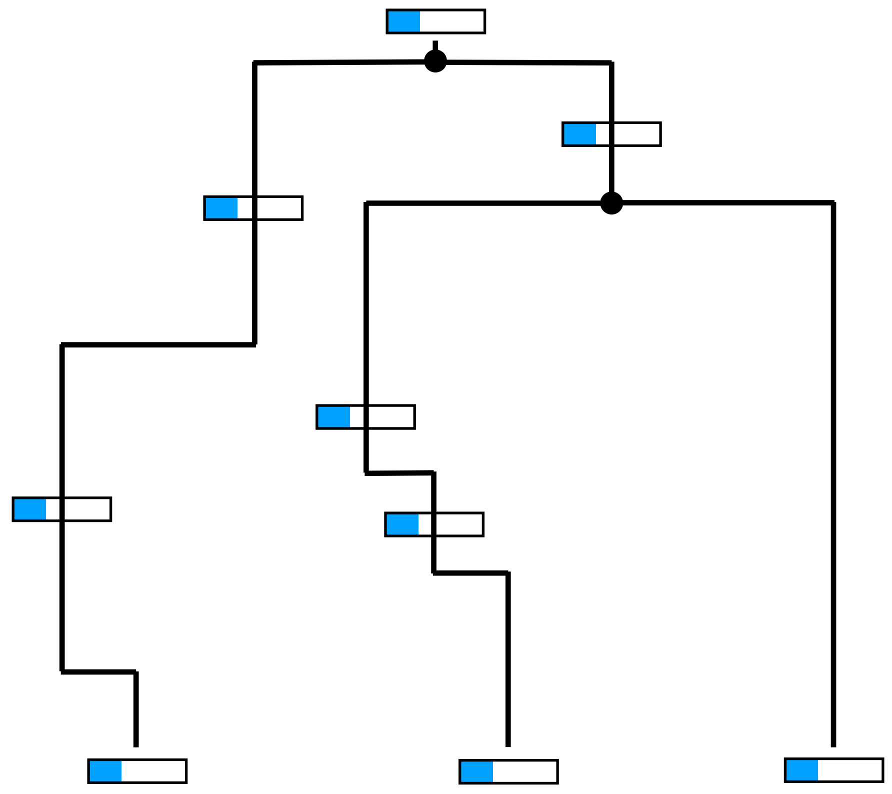
The ARG process

Along the sequence



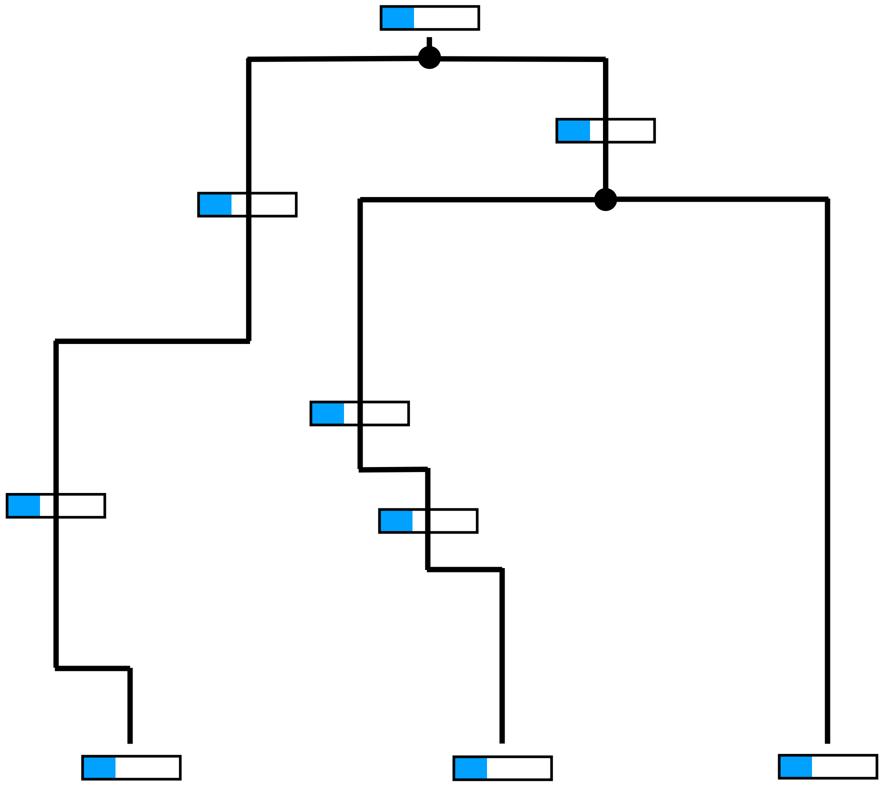
A new **recombinant** lineage, separate from the **original** lineage, free to coalesce on its own.

ARGs:

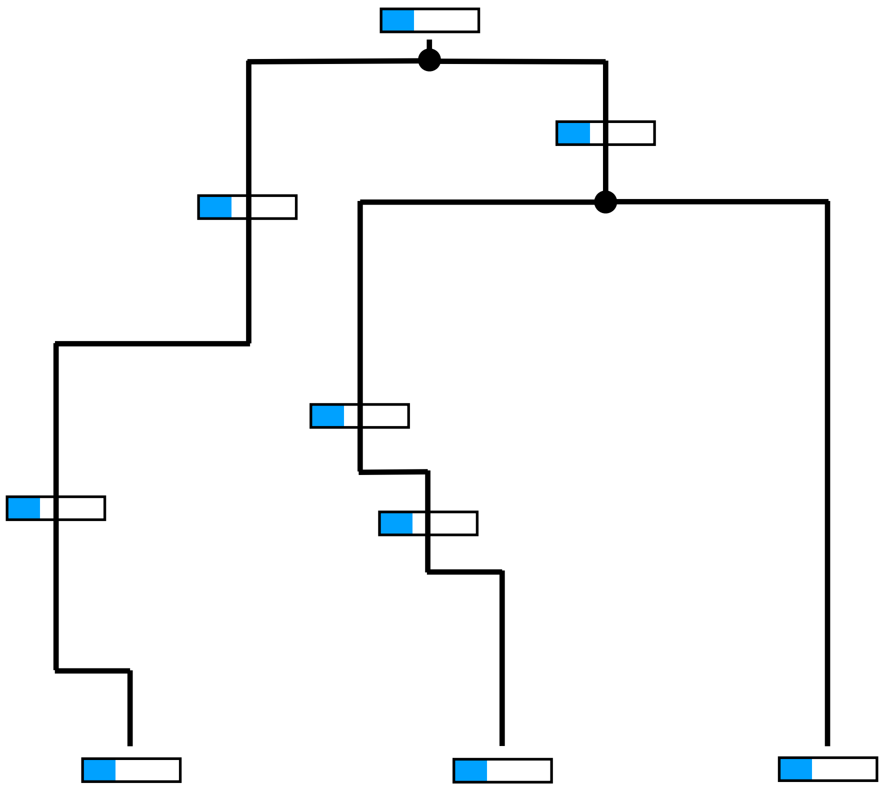


Trees:

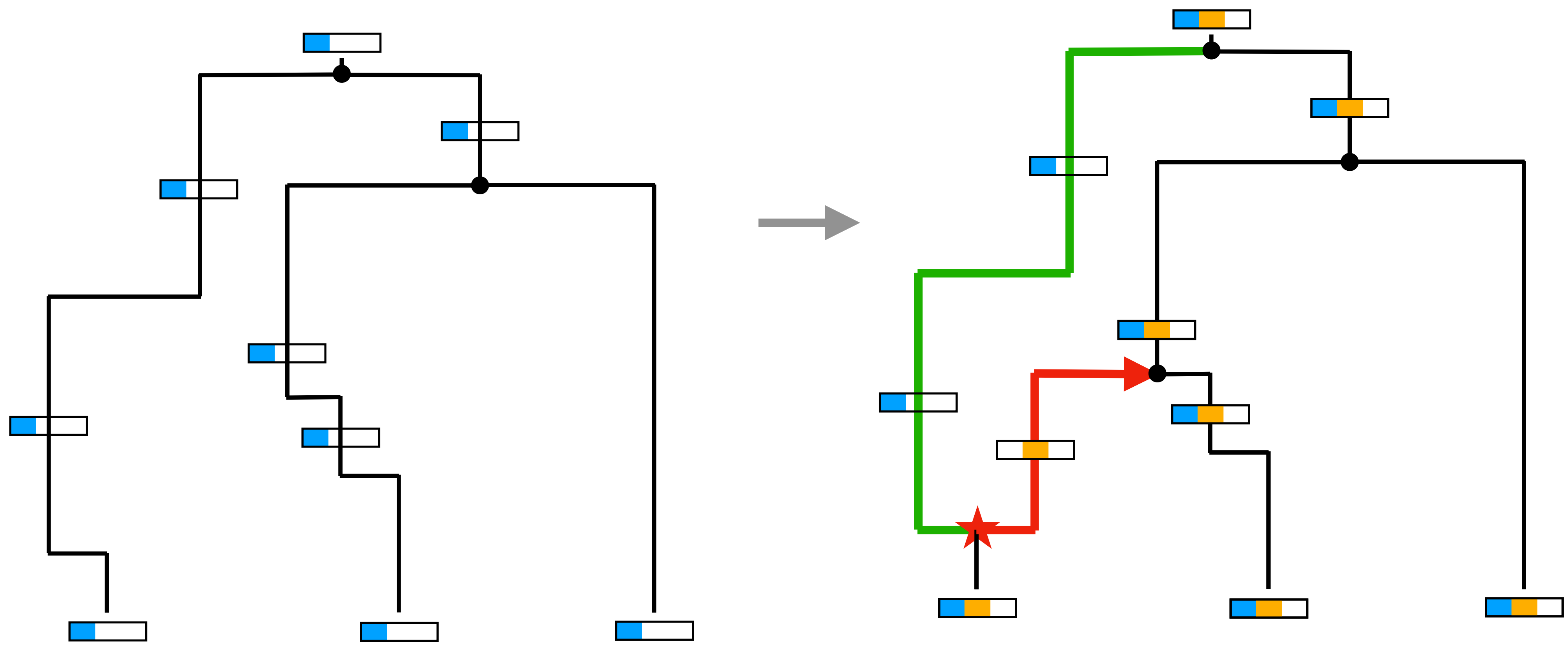
ARGs:



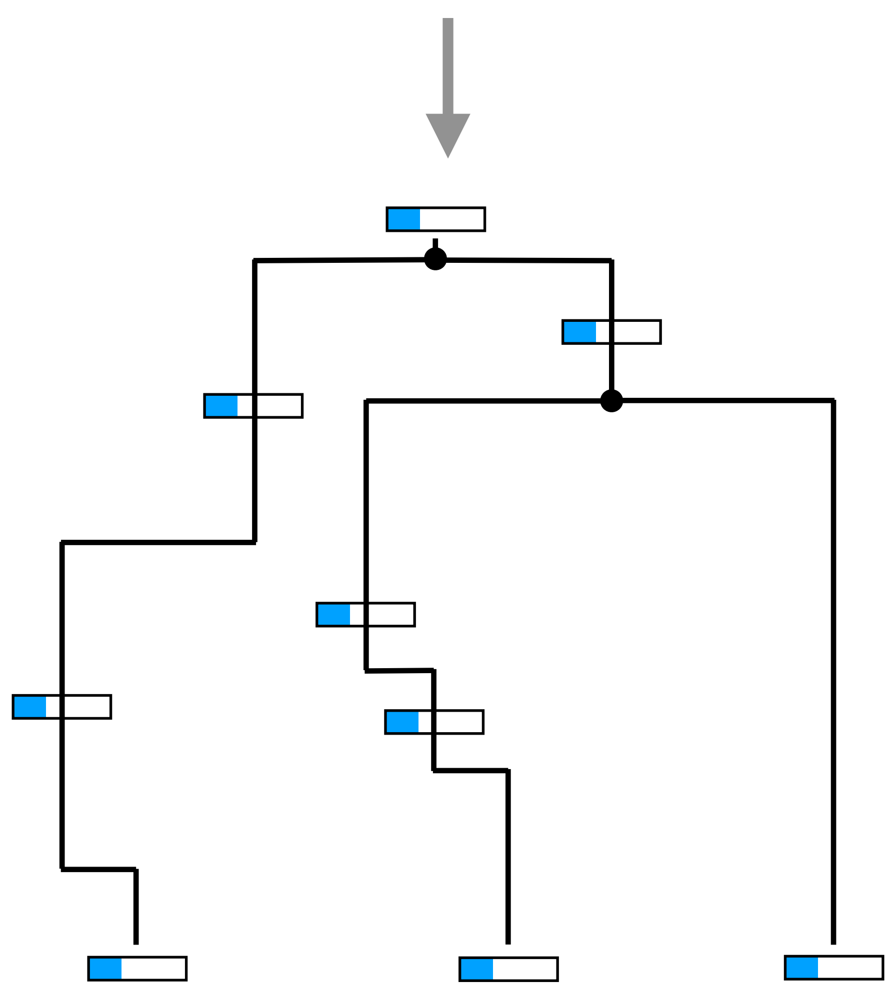
Trees:



ARGs:

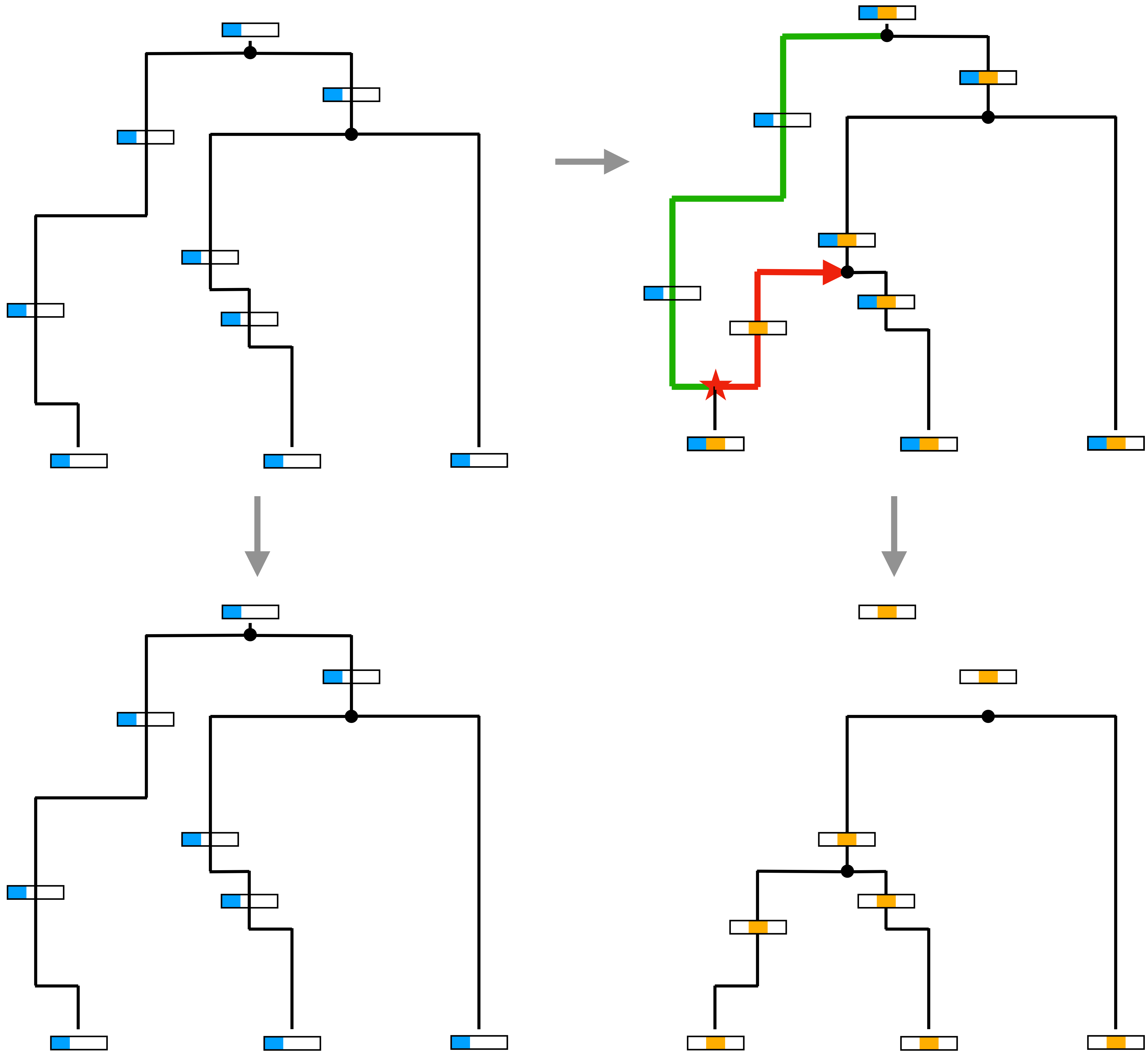


Trees:



ARGs:

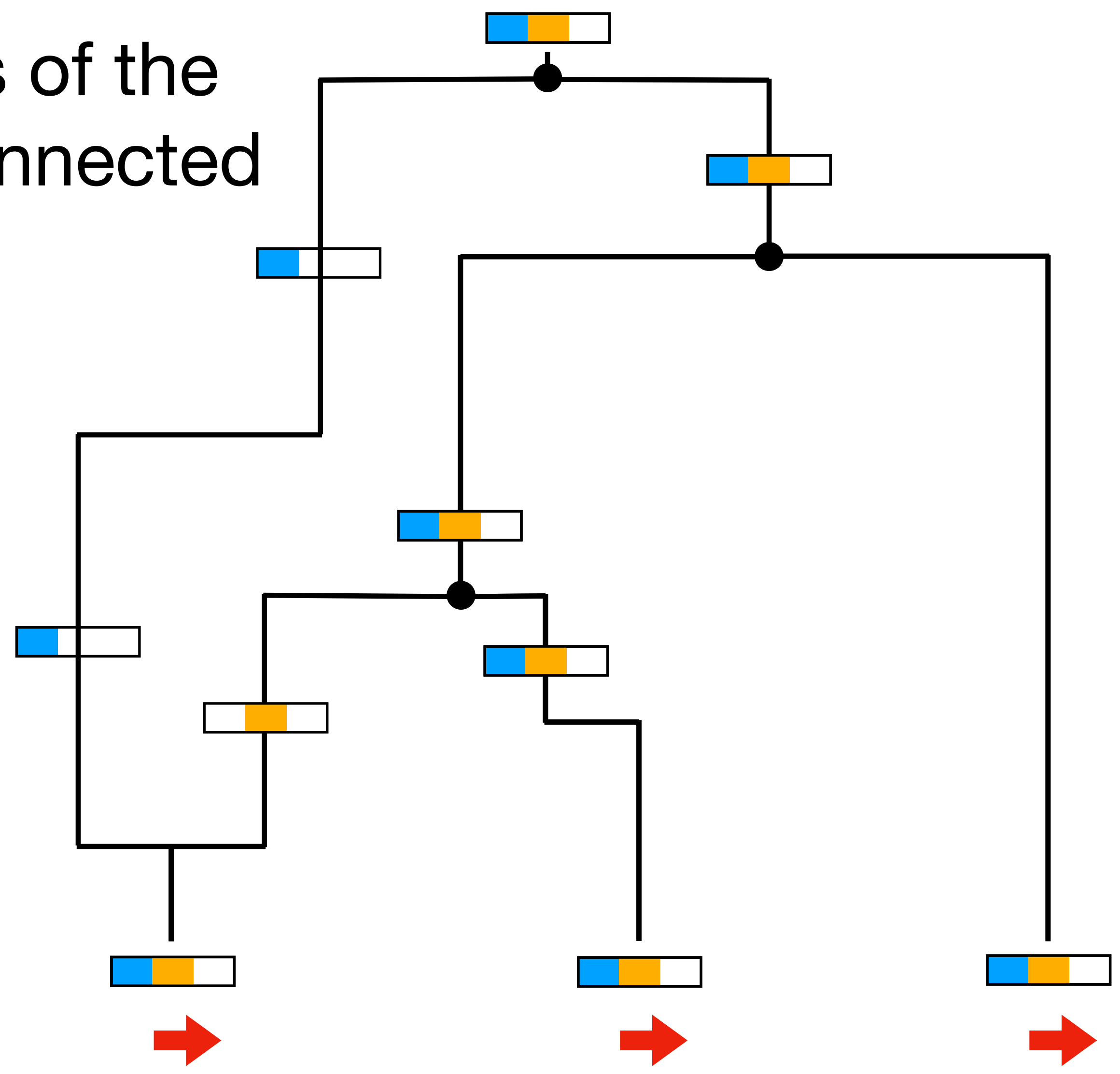
Trees:



The ARG process

Along the sequence

The leftmost parts of the sequences are connected by the this ARG:

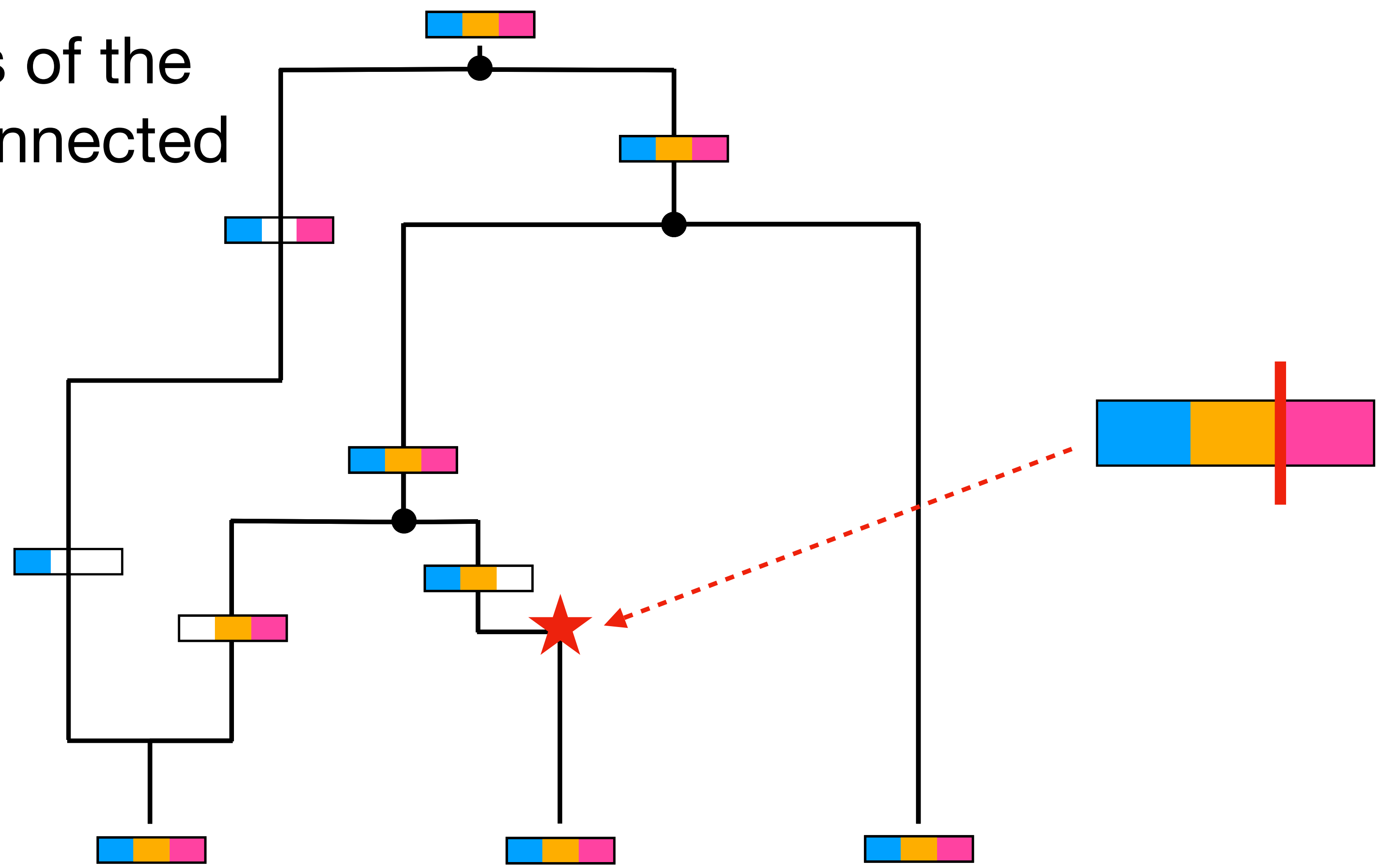


Move **right** until you encounter a recombination anywhere on this tree

The ARG process

Along the sequence

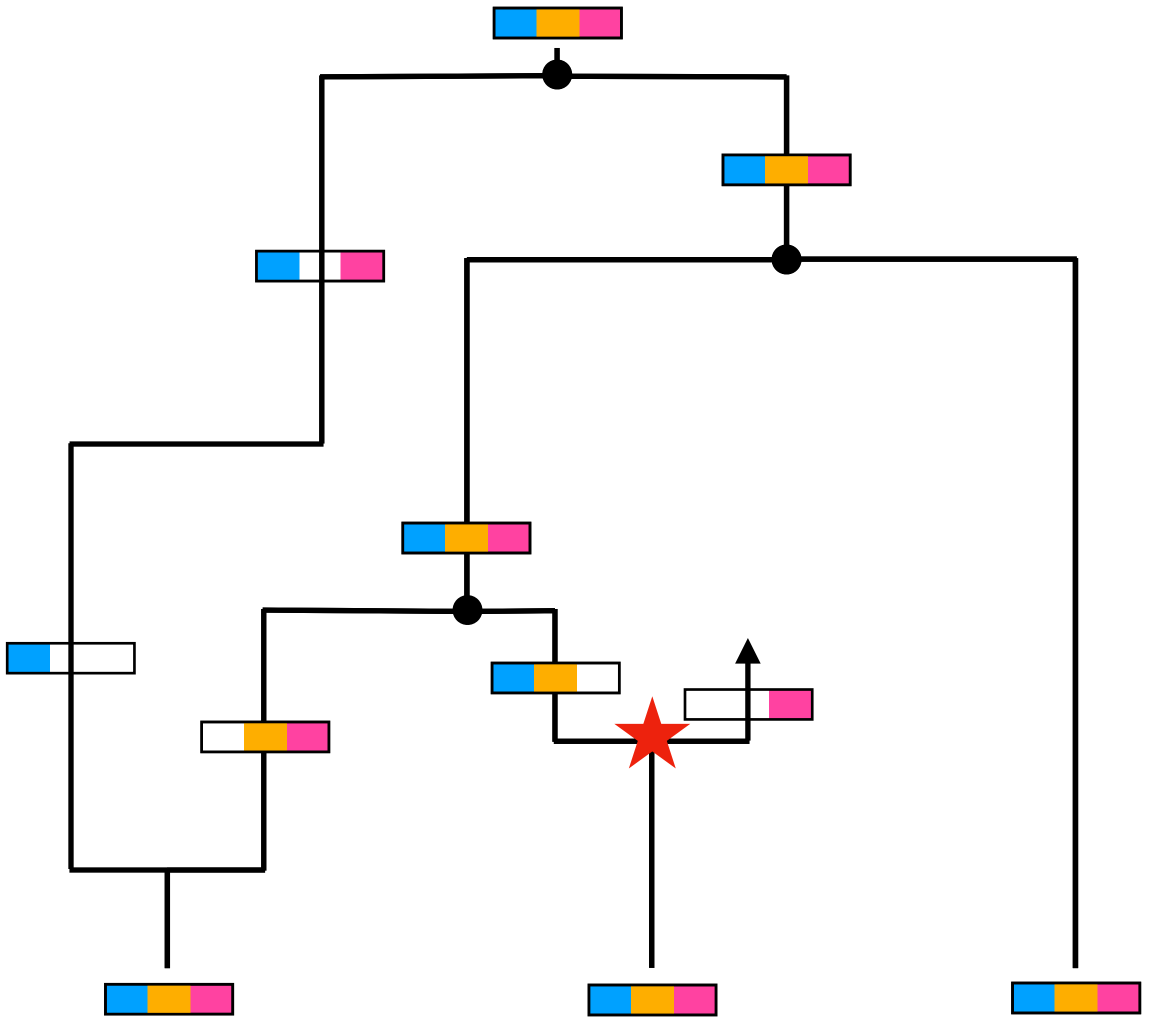
The leftmost parts of the sequences are connected by the this ARG:



Move **right** until you encounter a recombination anywhere on this tree

The ARG process

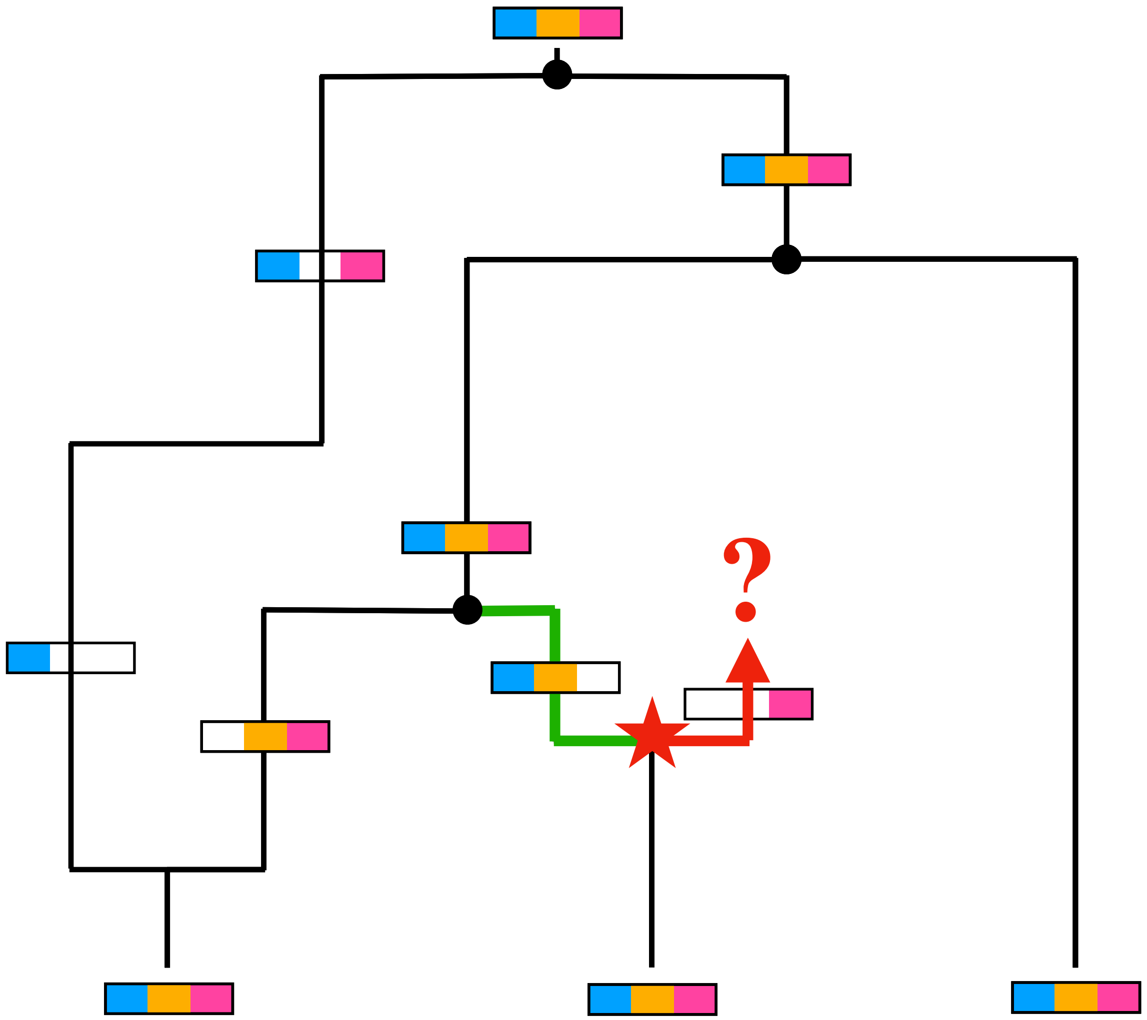
Along the sequence



A recombination event splits a sequence into a the **yellow** part and a subsequent **pink** part carried by a separate lineage.

The ARG process

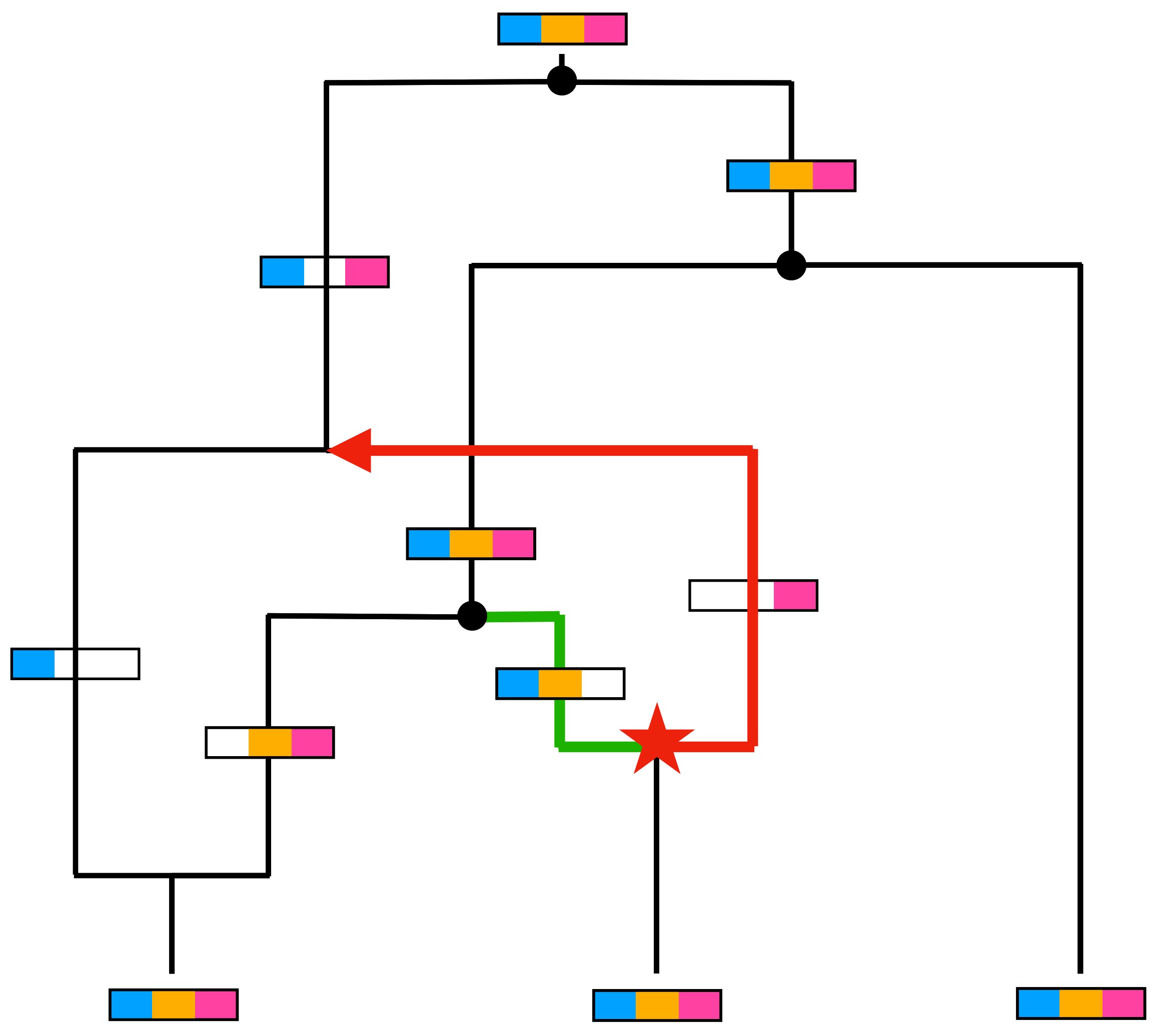
Along the sequence



A recombination event splits a sequence into a the **yellow** part and a subsequent **pink** part carried by a separate lineage.

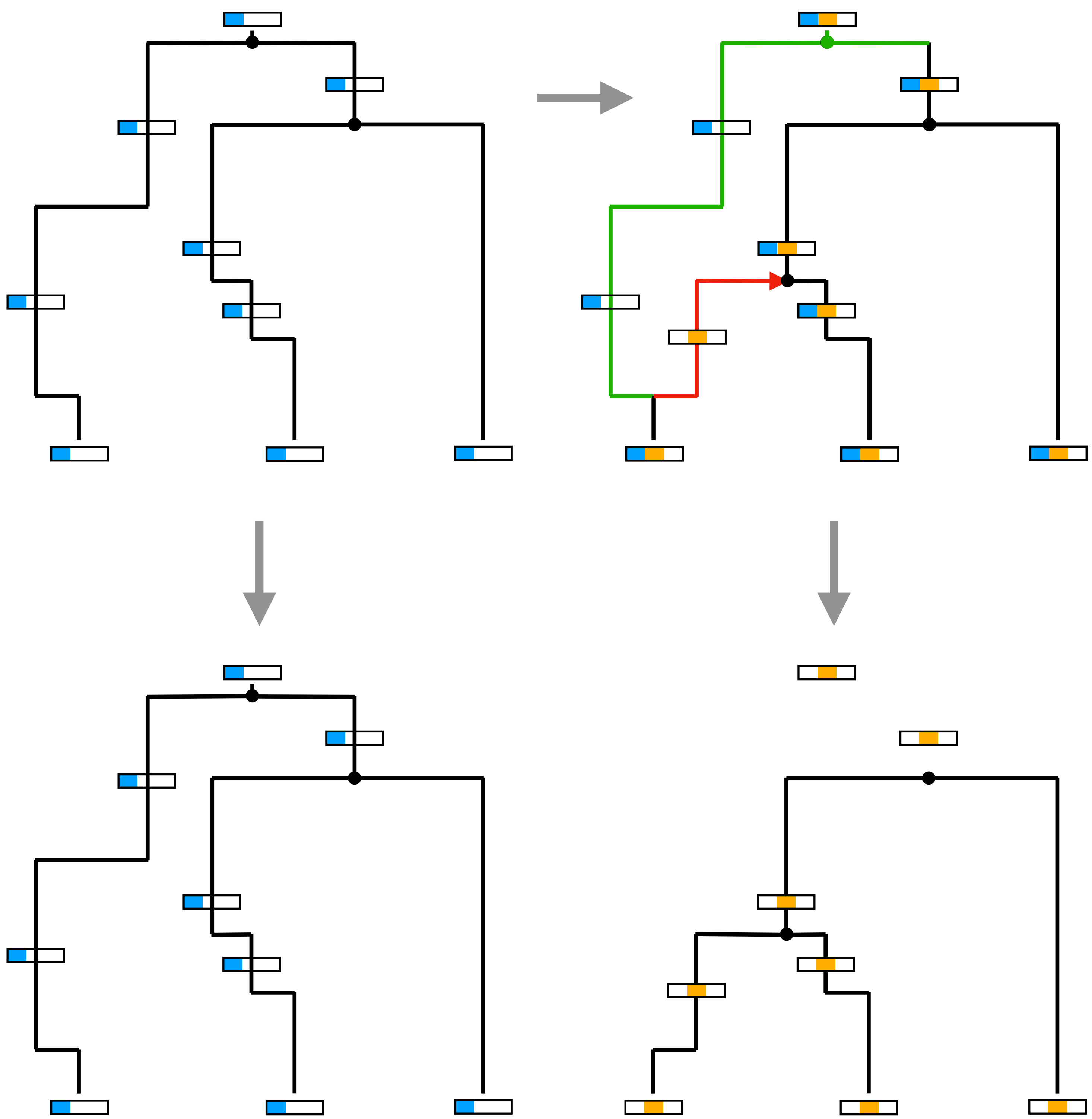
The ARG process

Along the sequence



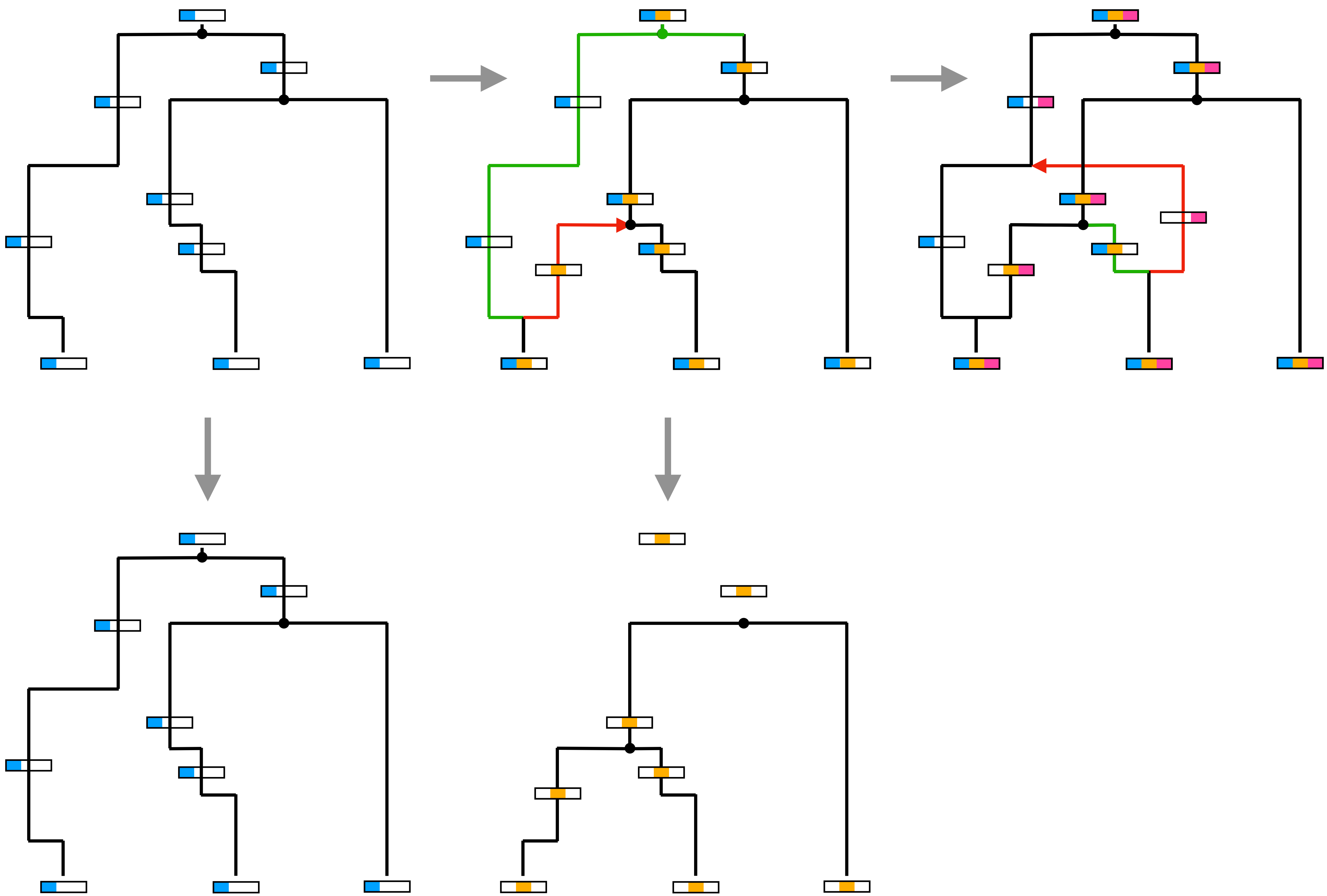
A recombination event splits a sequence into a the **yellow** part and a subsequent **pink** part carried by a separate lineage.

ARG growing one tree at a time



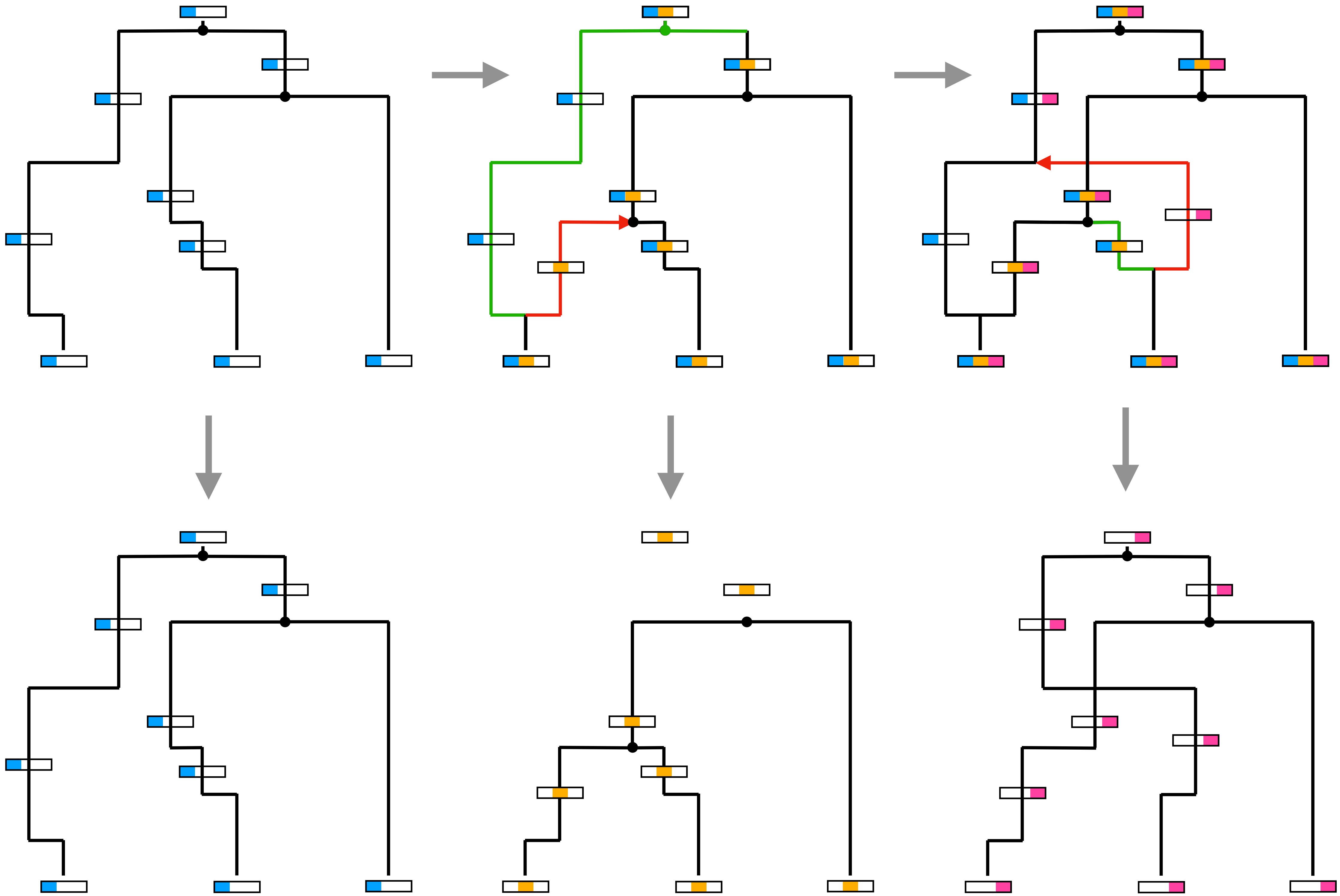
Trees for each segment

ARG growing one tree at a time



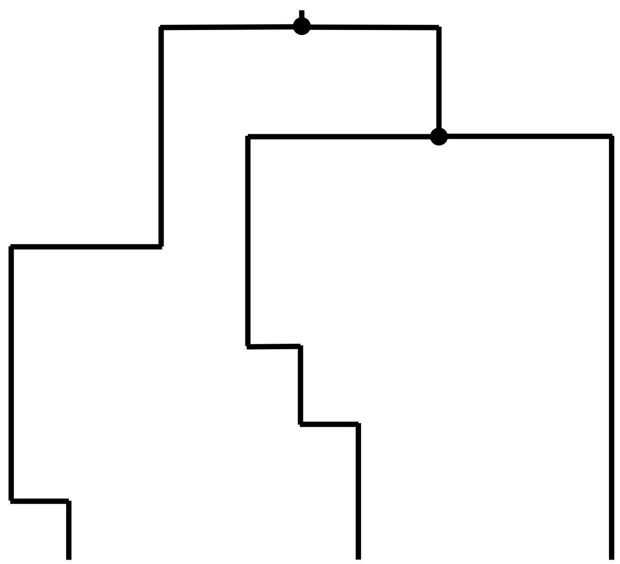
Trees for each segment

ARG growing one tree at a time



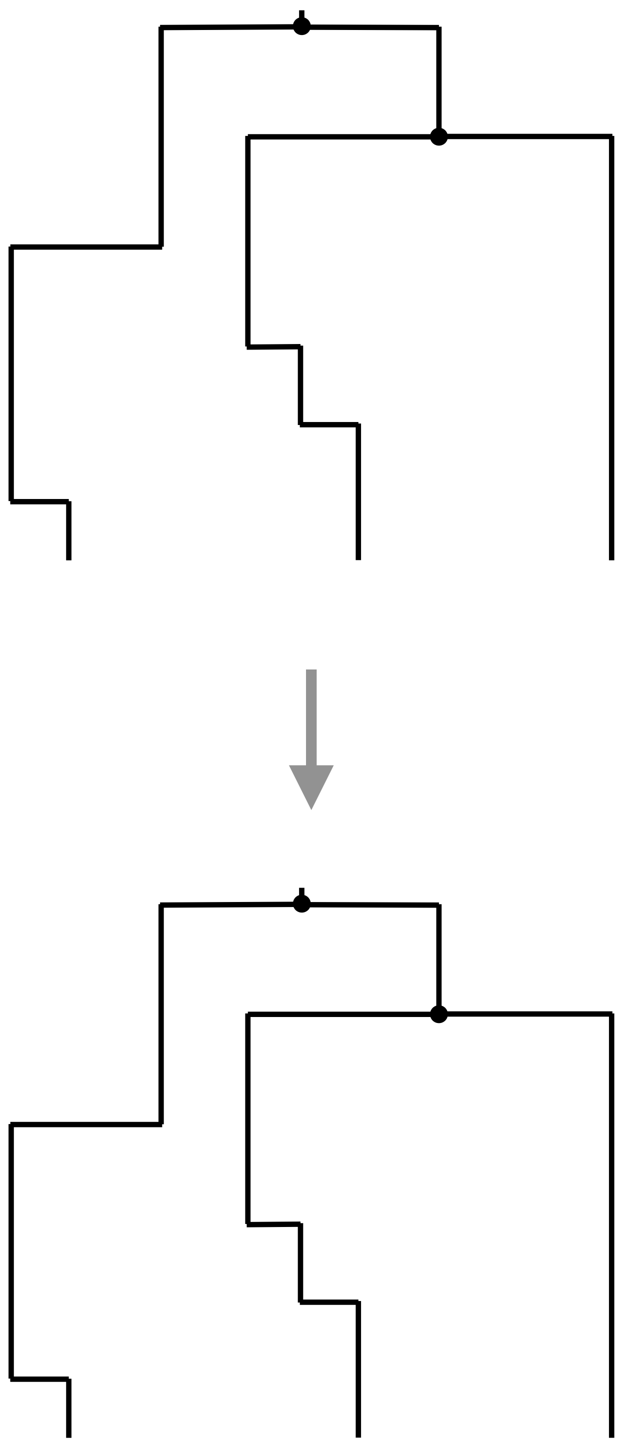
Trees for each segment

ARG growing one tree at a time



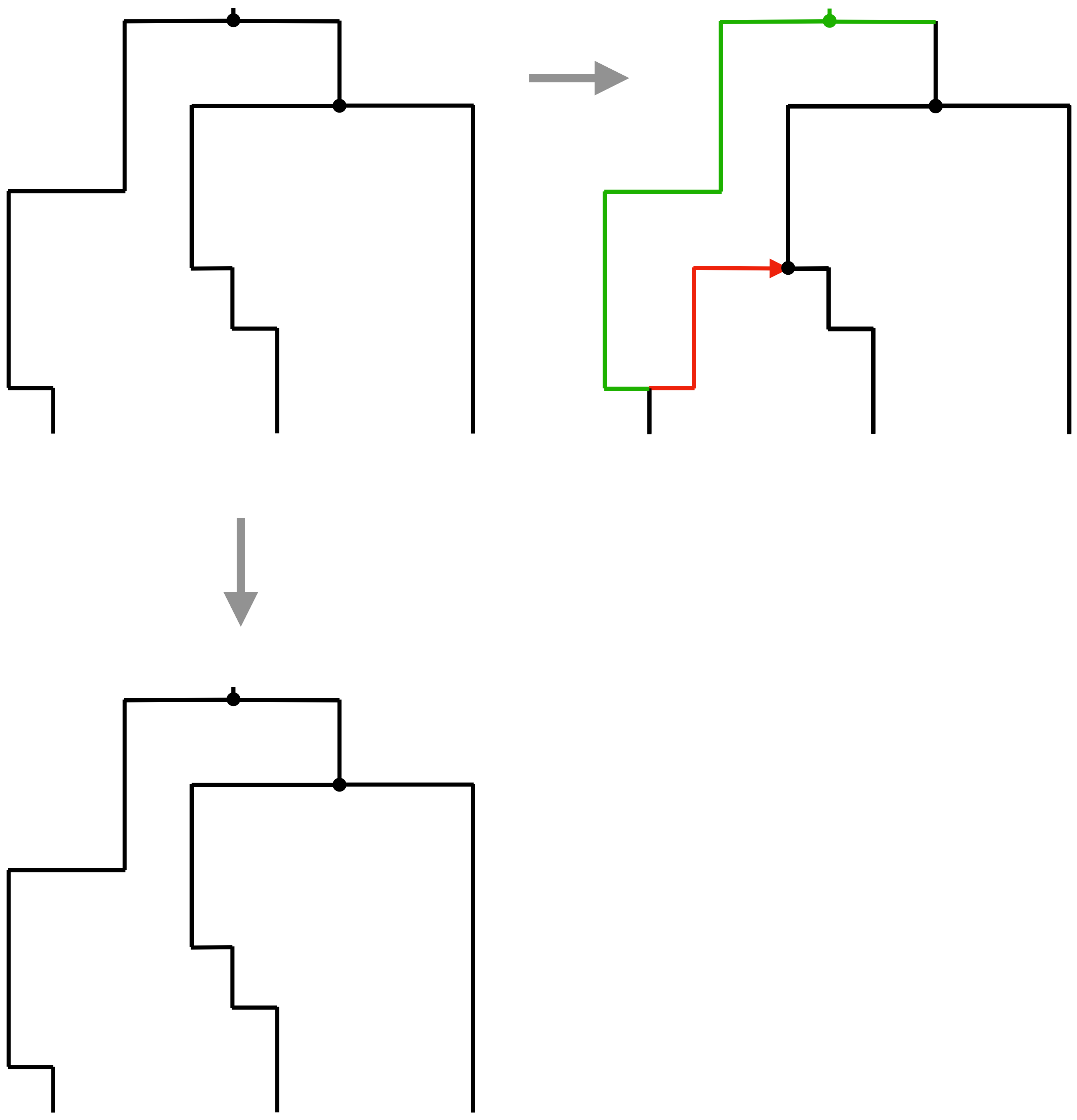
Trees for each segment

ARG growing one tree at a time



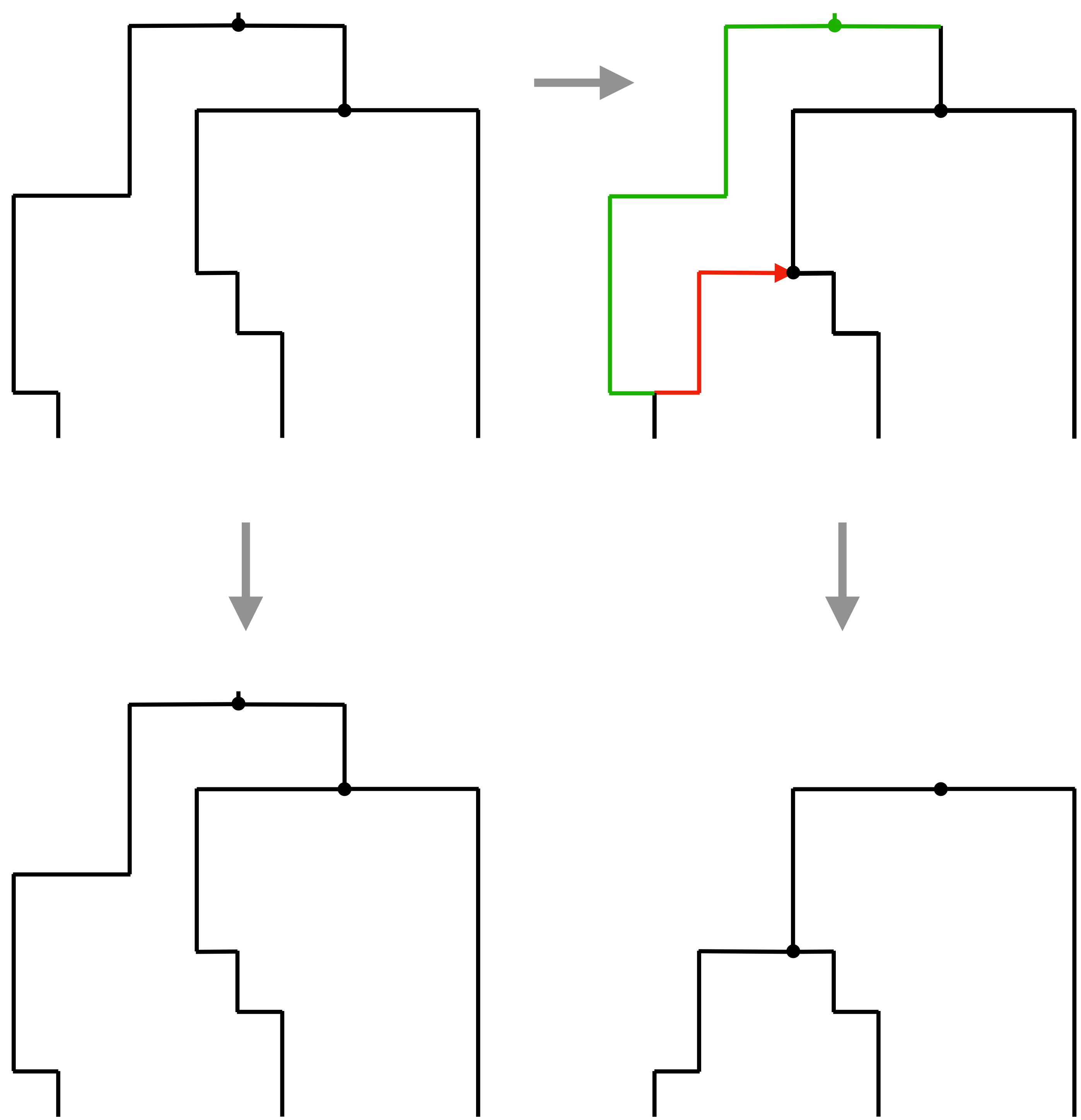
Trees for each segment

ARG growing one tree at a time



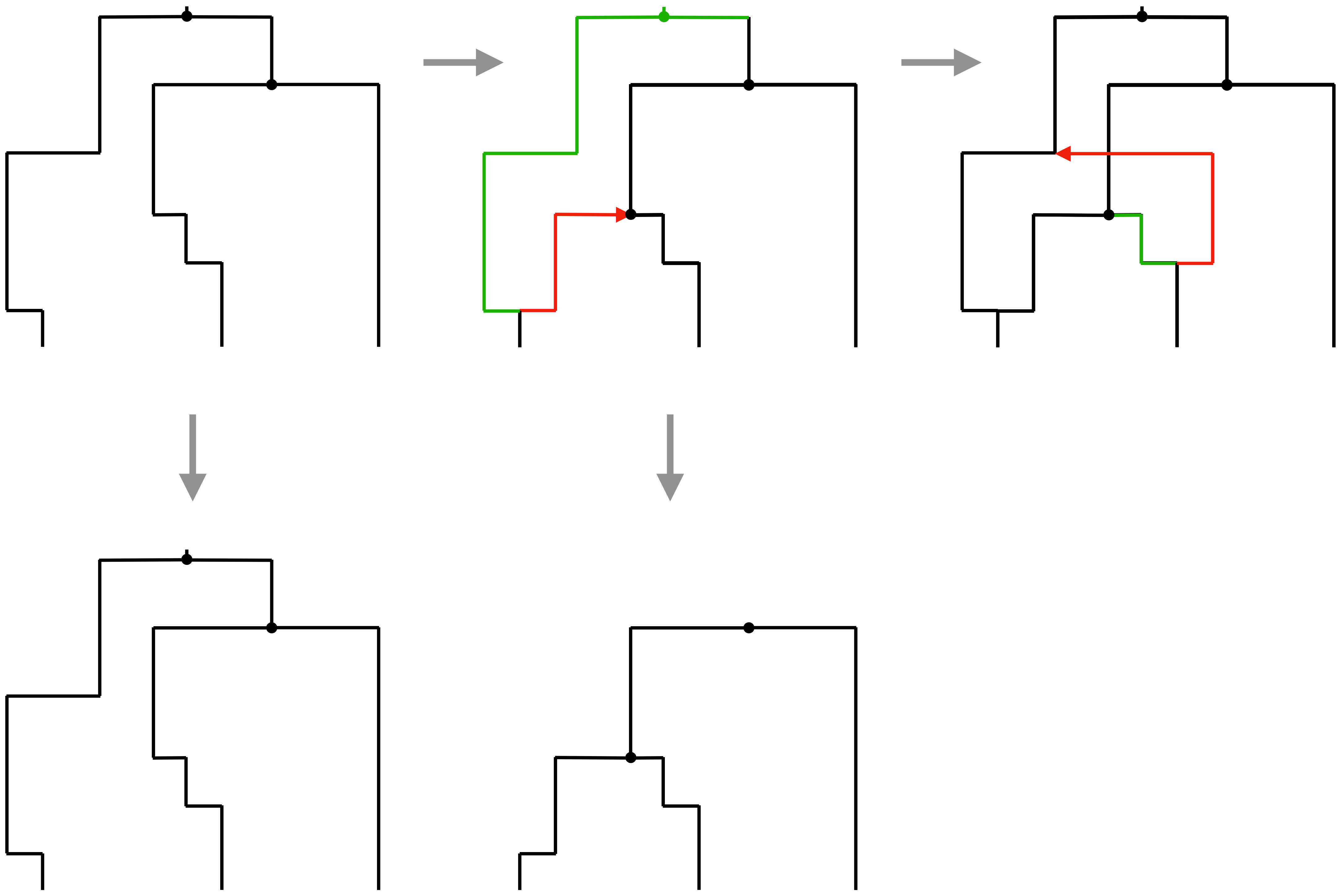
Trees for each segment

ARG growing one tree at a time



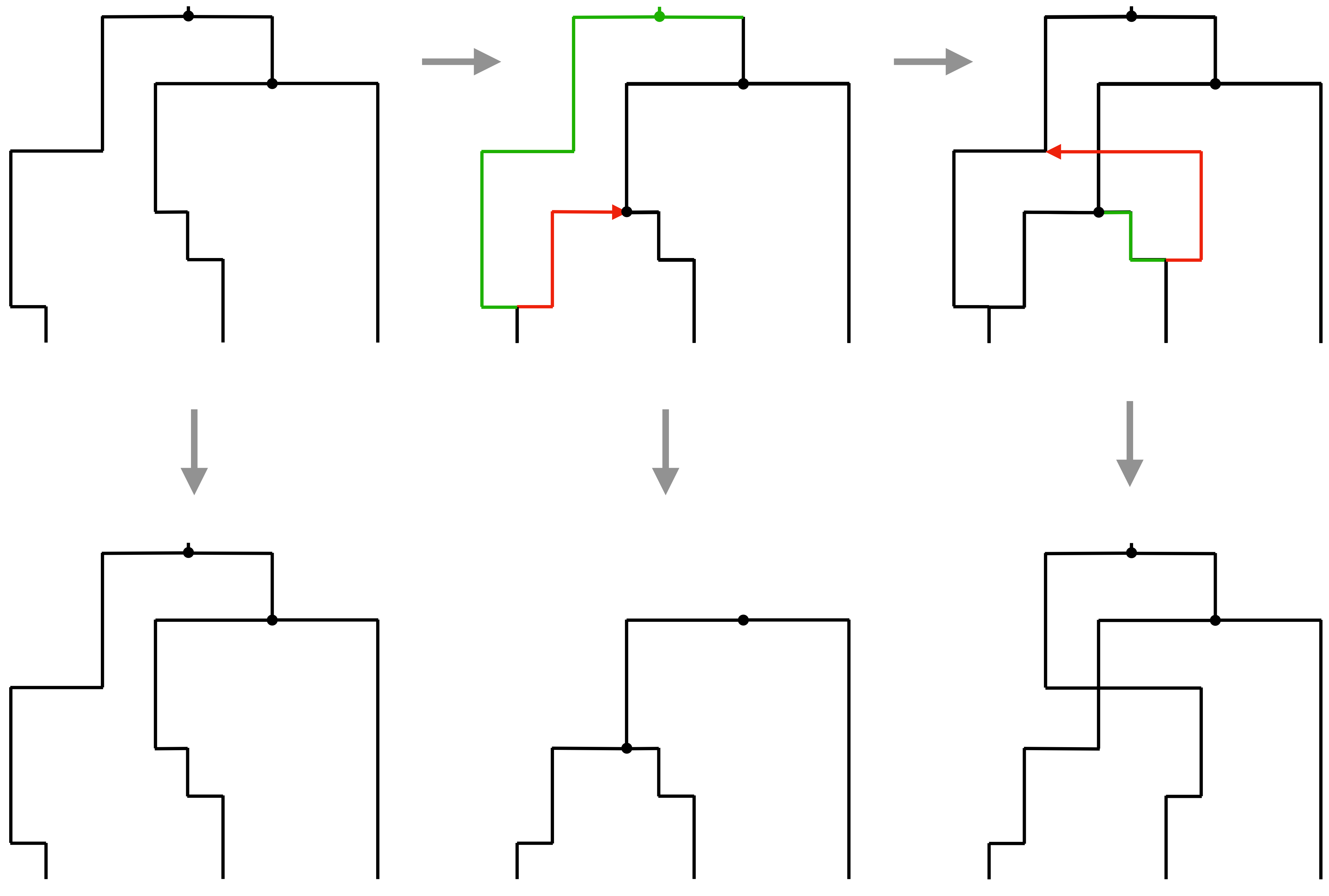
Trees for each segment

ARG growing one tree at a time



Trees for each segment

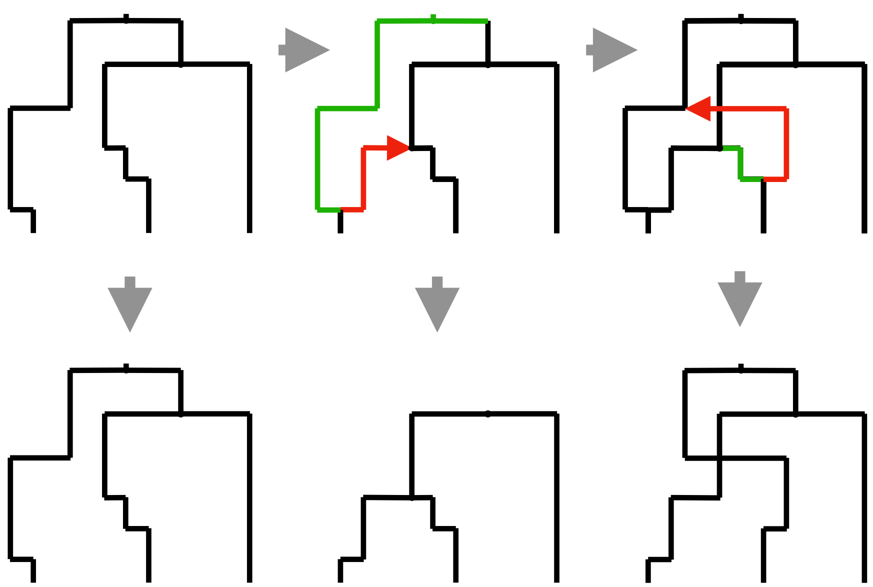
ARG growing one tree at a time



Trees for each segment

The ARG process

Along the sequence



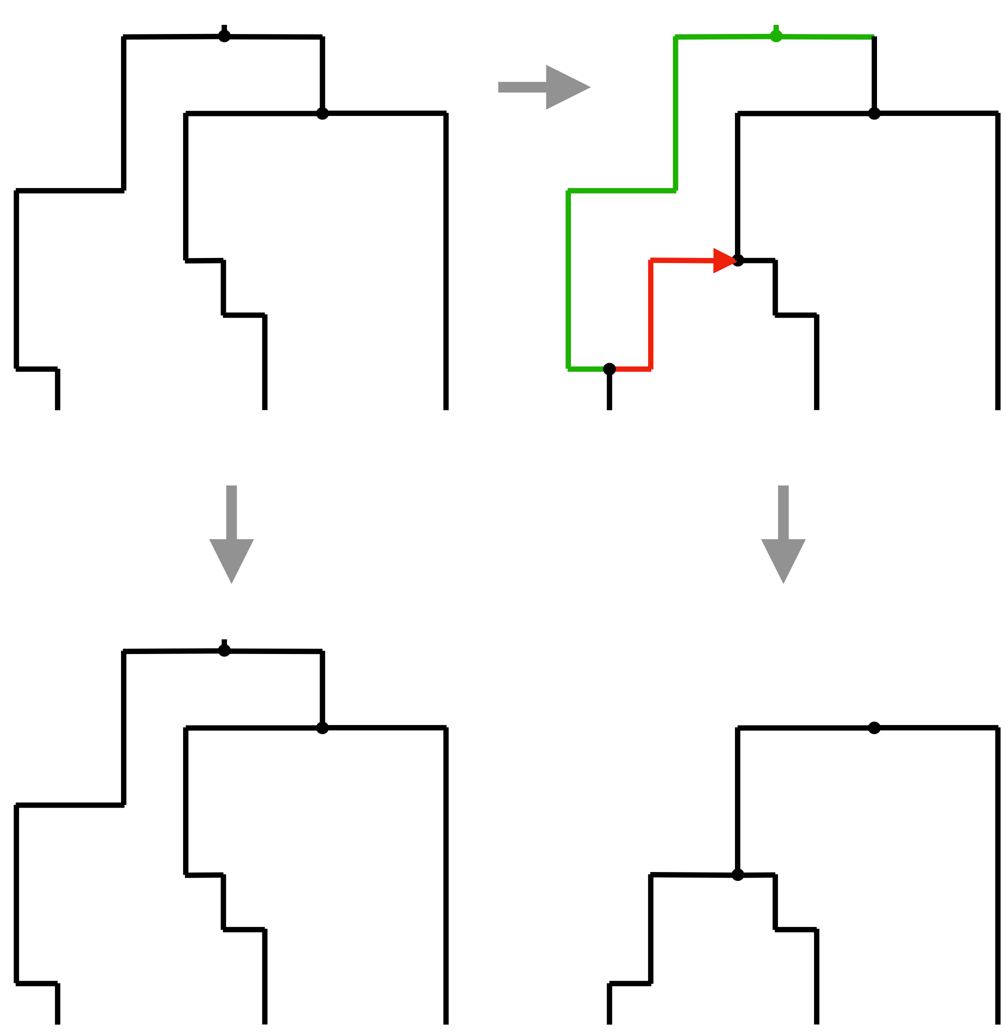
For long sequences, the full ARG becomes to large to handle computationally

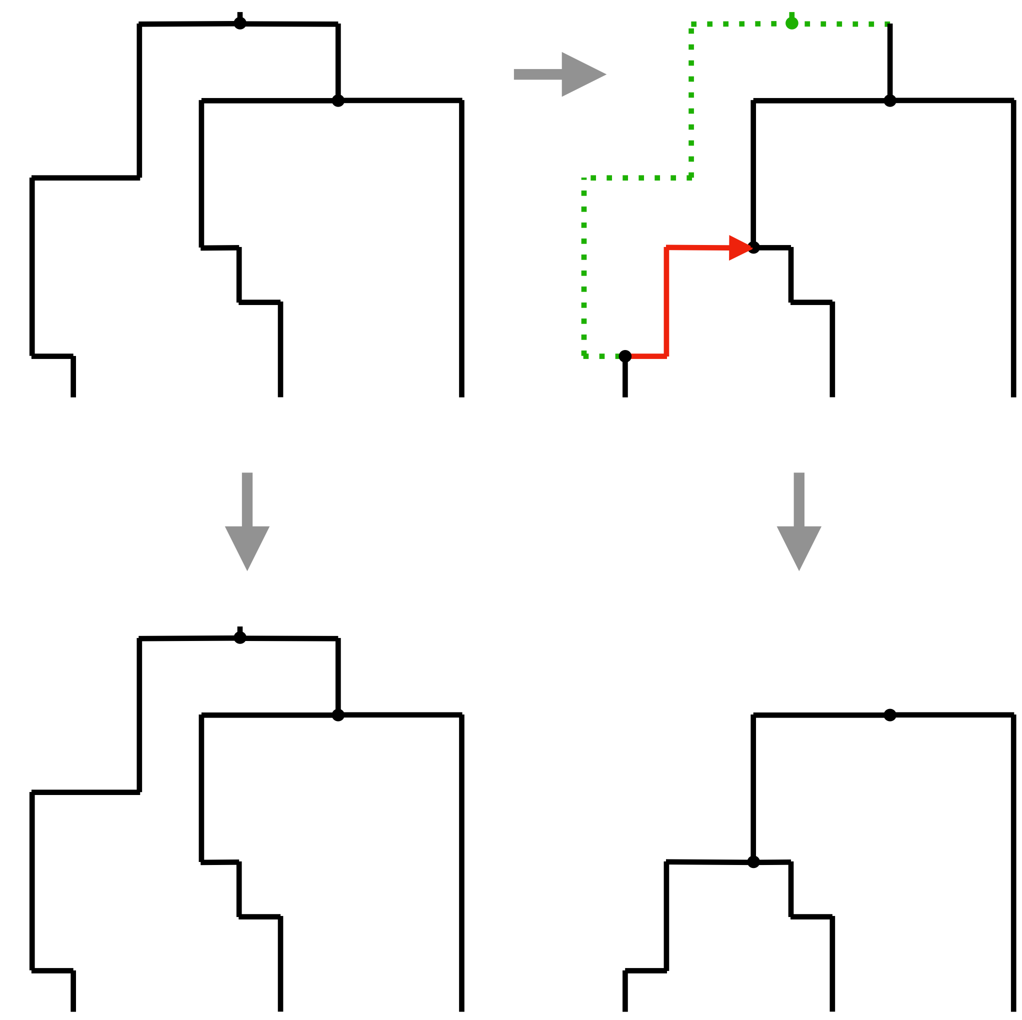
ARG!

Chromosome

SMC

ARG vs. SMC

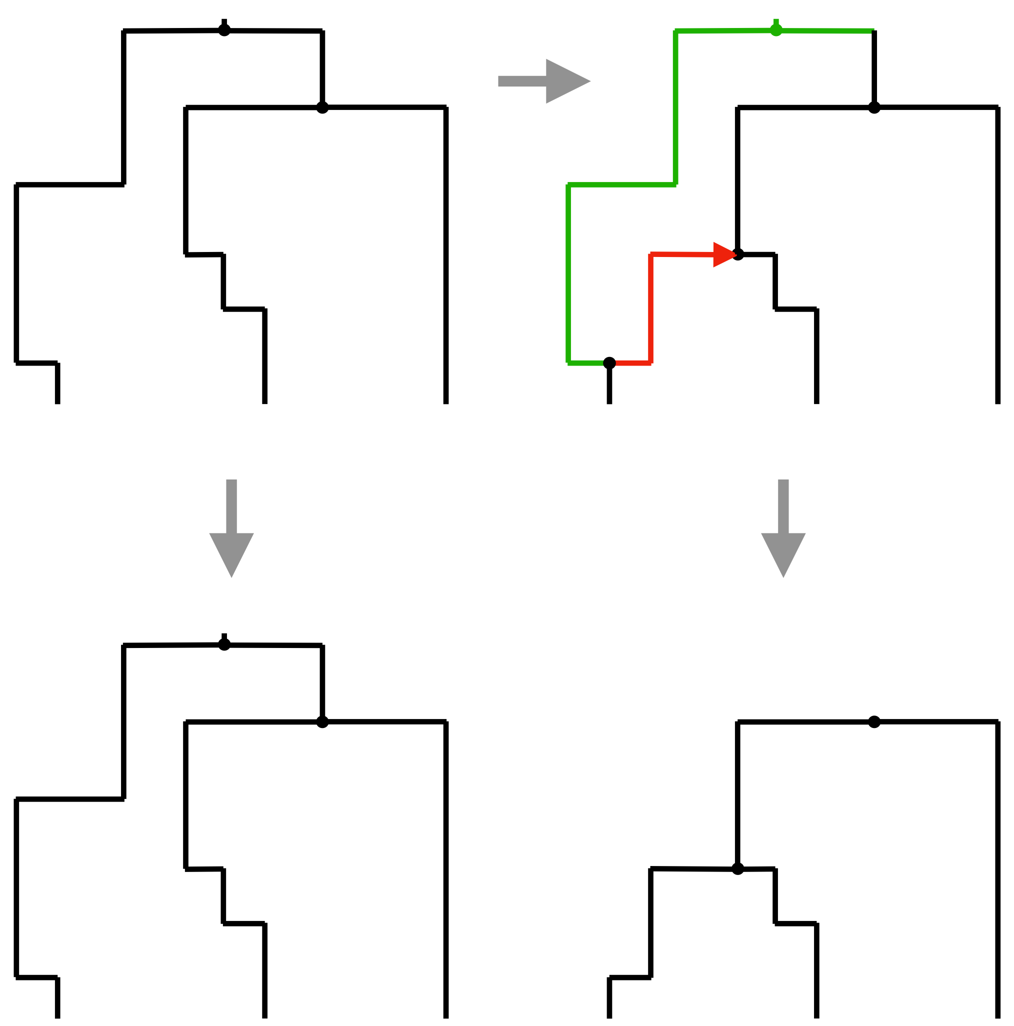




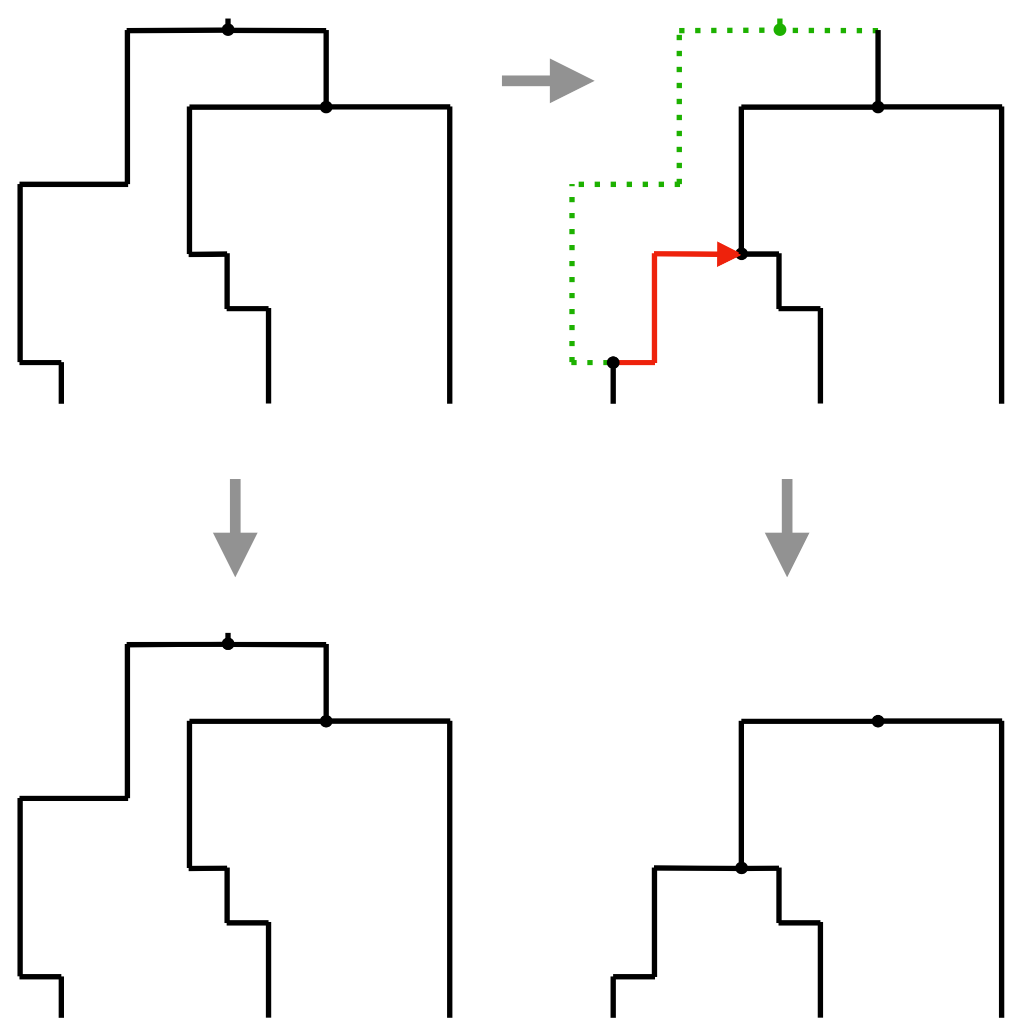
Branch remains in ARG:

Branch is removed:

ARG vs. SMC

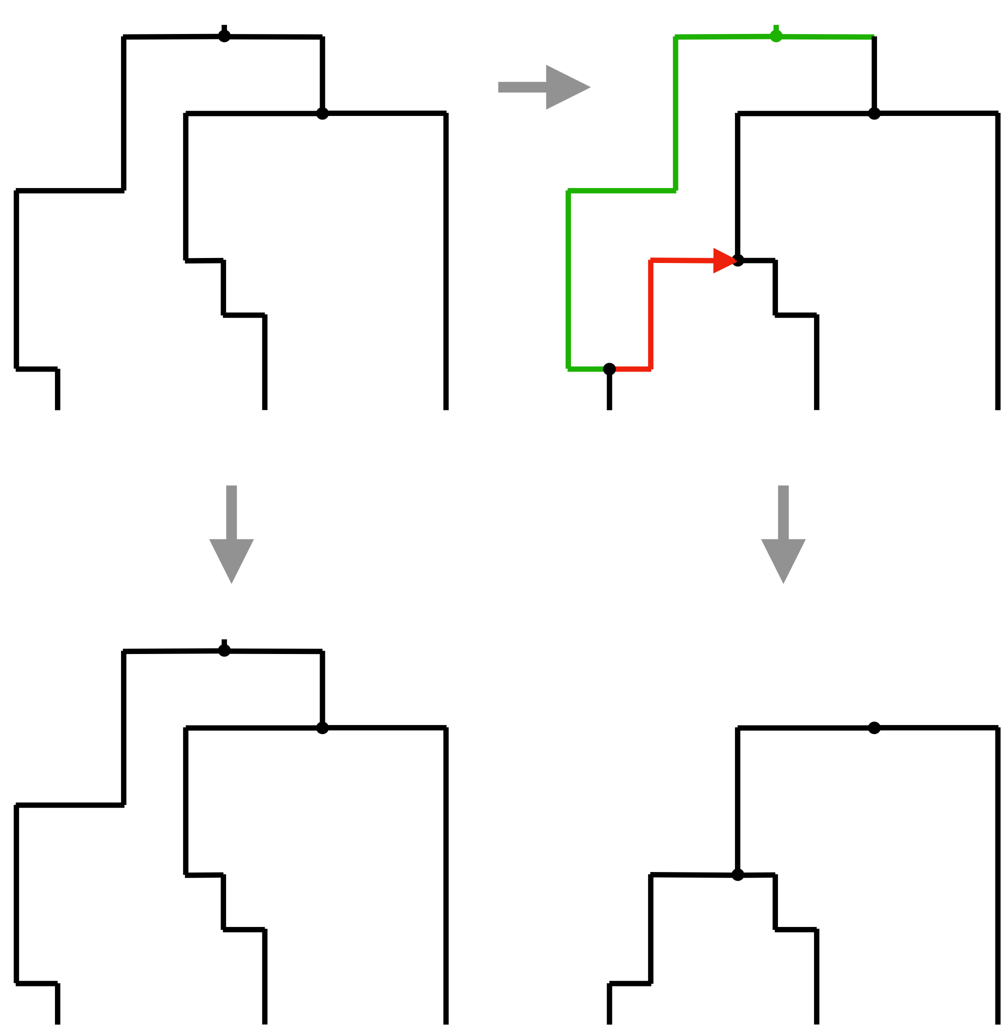


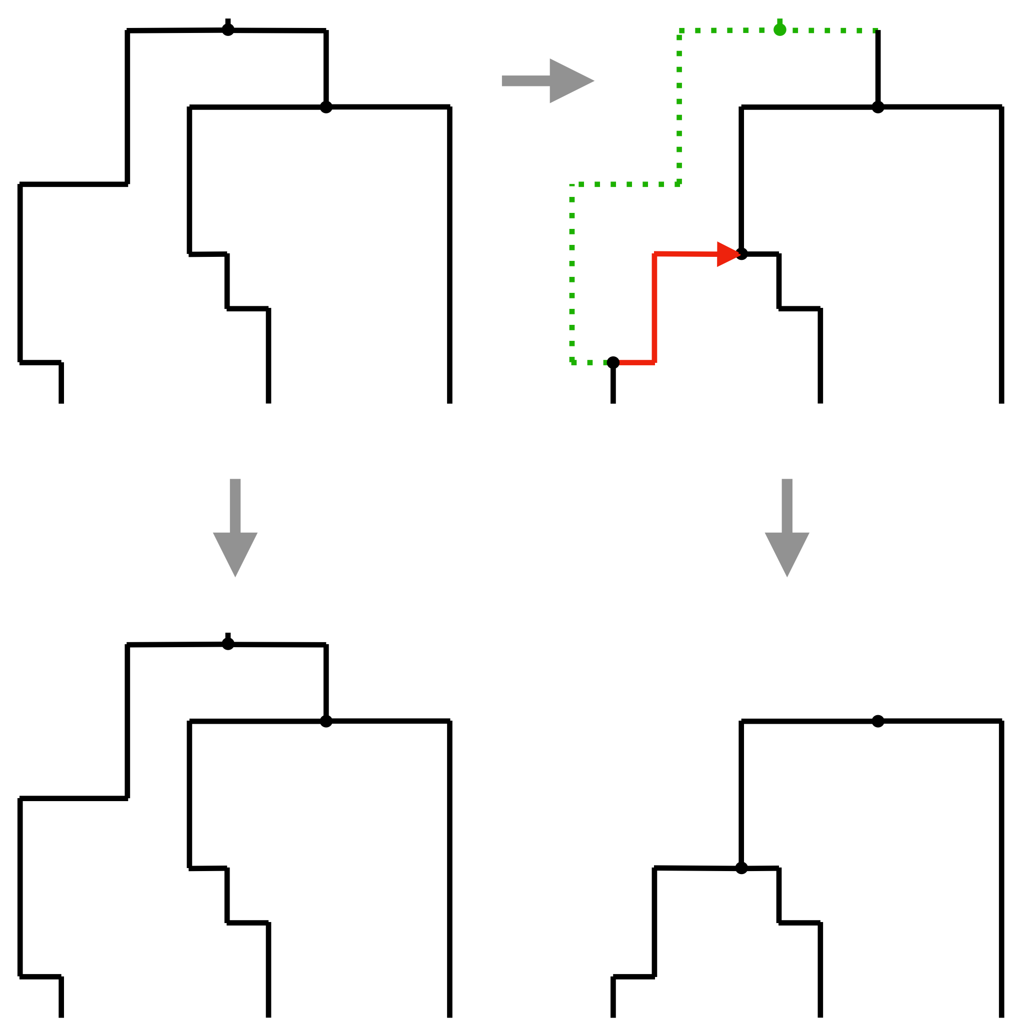
All trees are saved, producing the full ARG at the end of the sequence.



Only the current tree is saved.

ARG vs. SMC



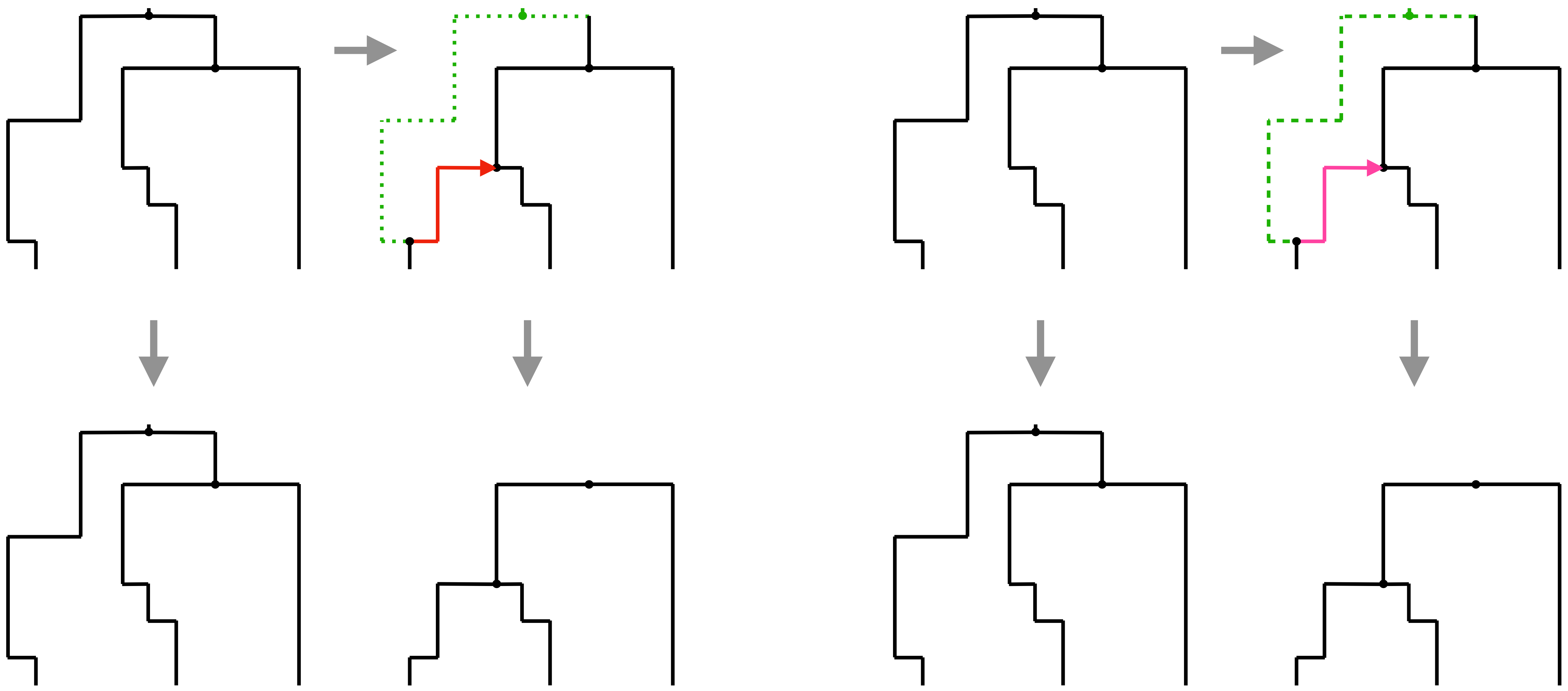


Recombinations can fall on the entire ARG.

Recombinations can only fall on the current tree

SMC’

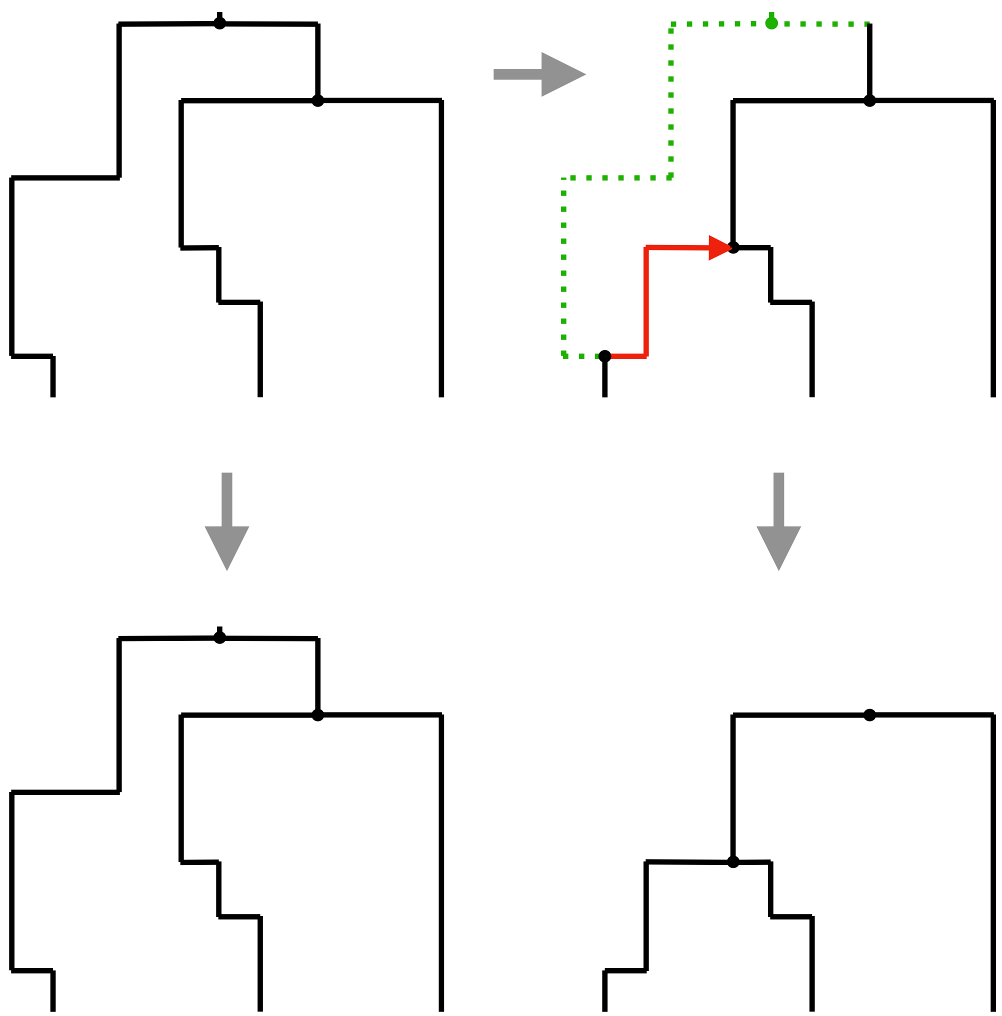
SMC vs. SMC’



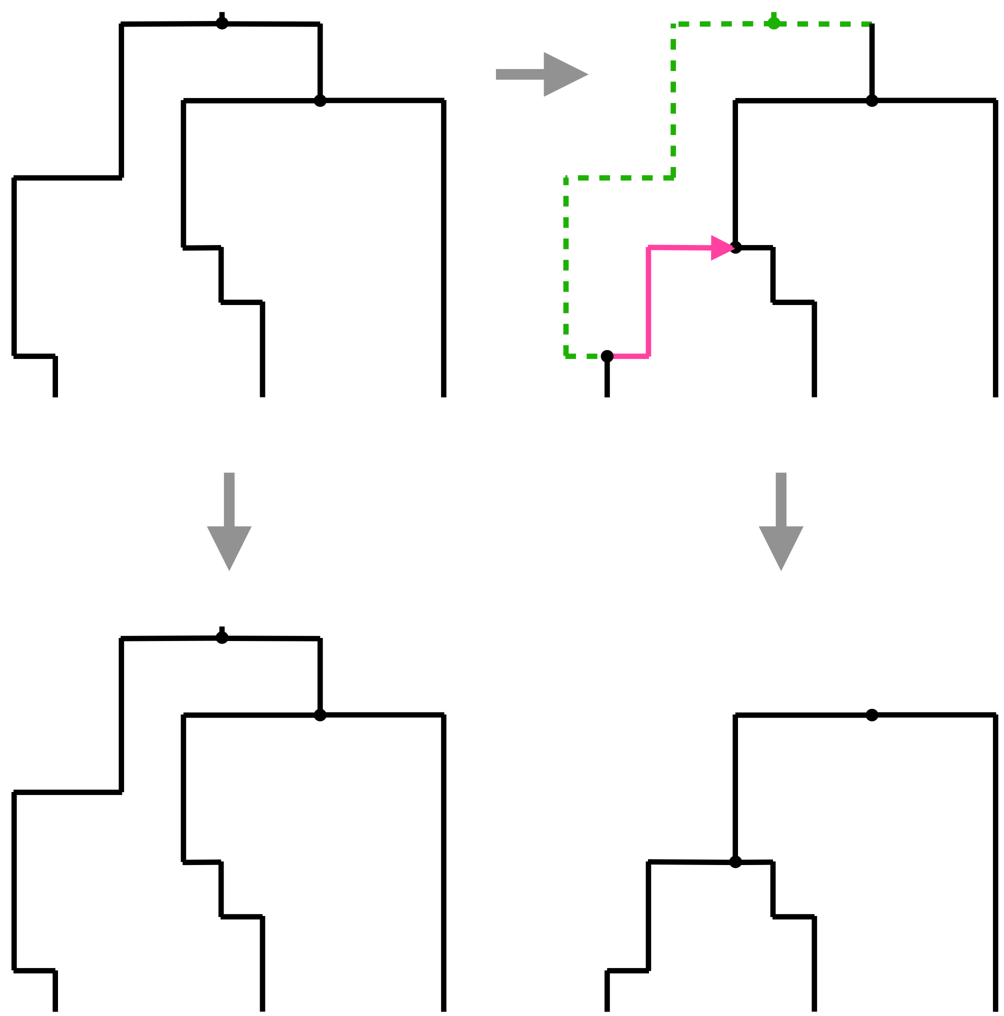
Branch is removed *BEFORE* we find a coalescence for the pink lineage:

Branch is removed *AFTER* we find a coalescence for the pink lineage:

SMC vs. SMC'

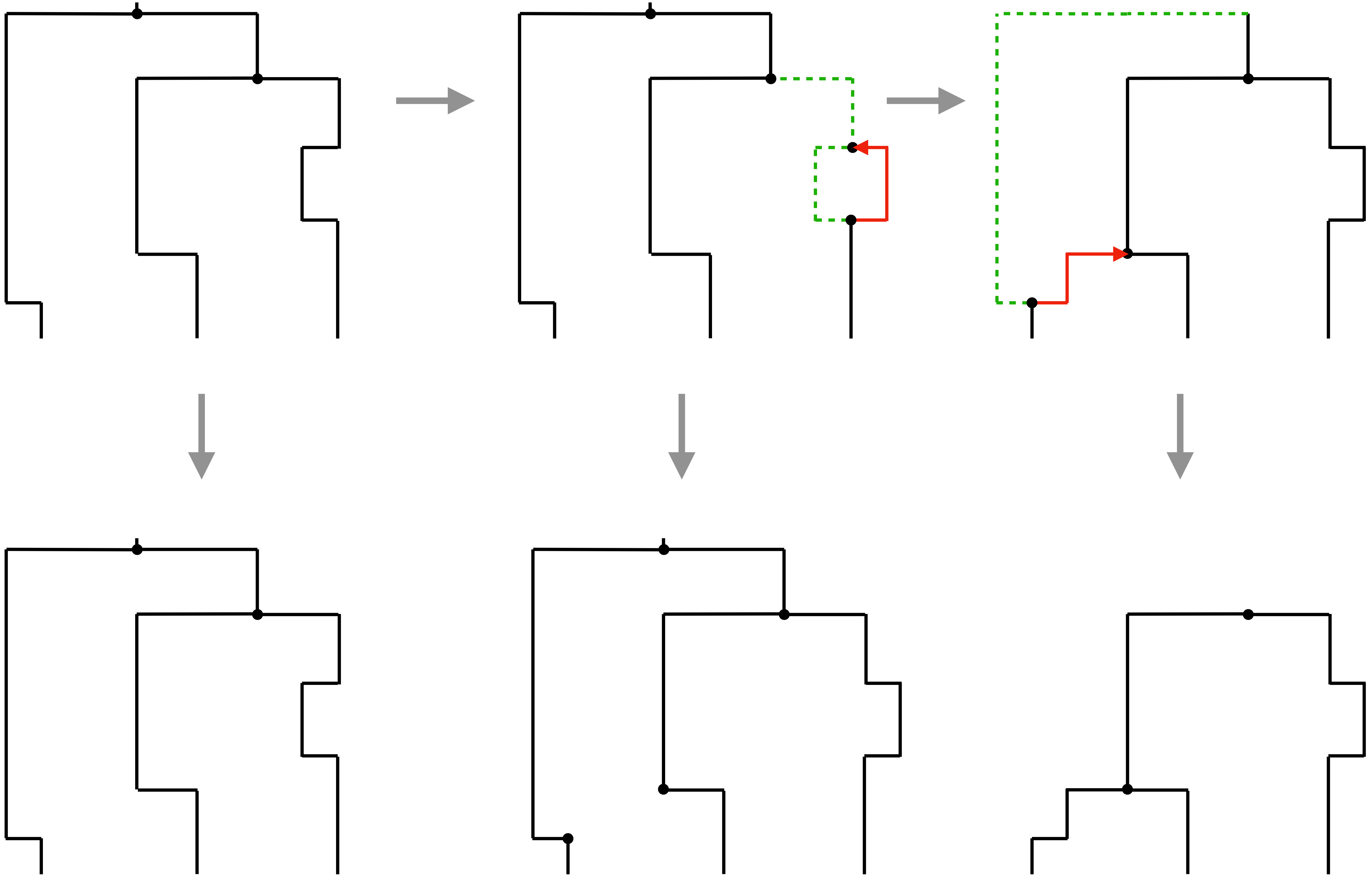


The pink branch CANNOT coalesce back onto the blue branch.

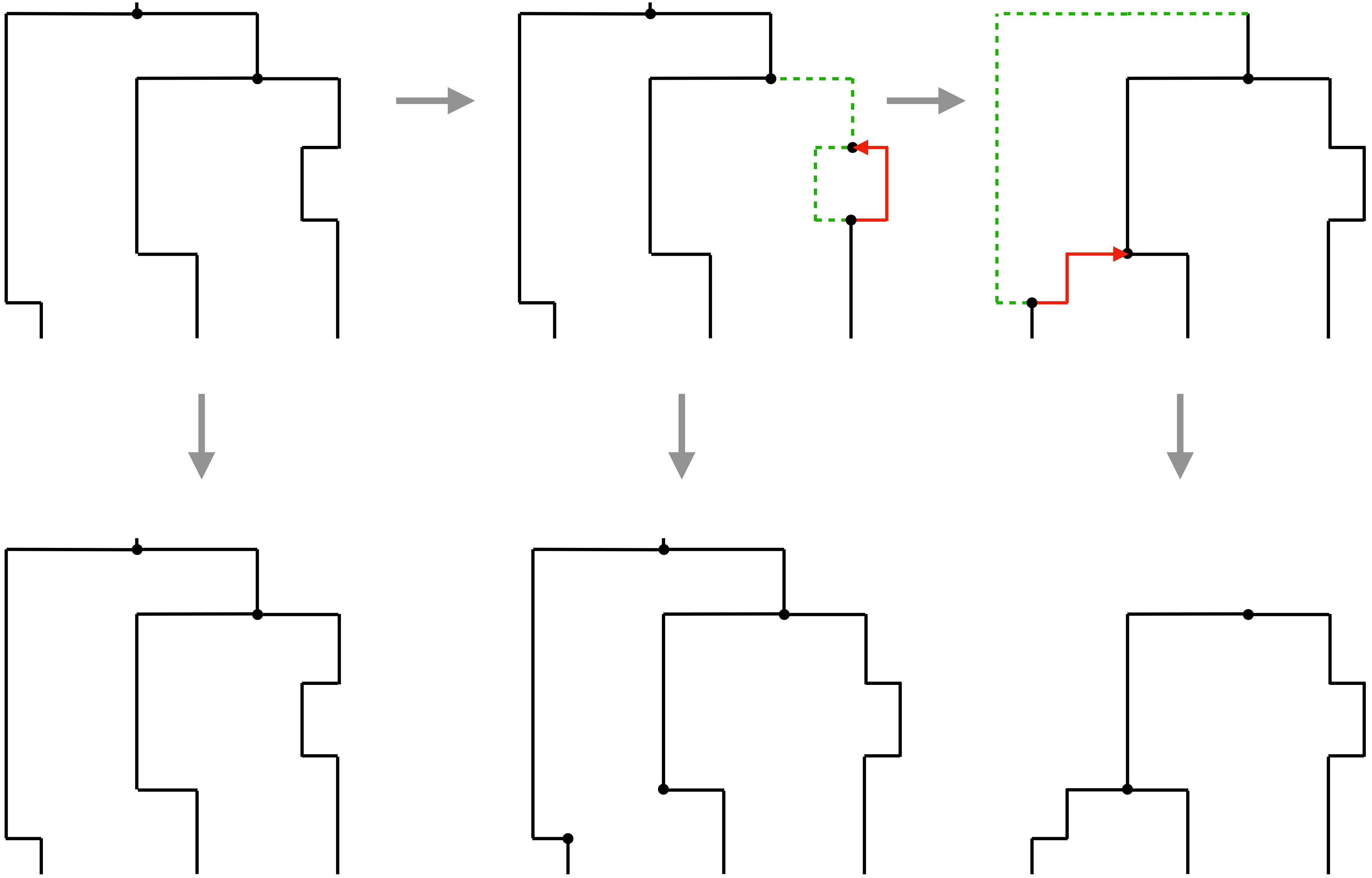


The pink branch CAN coalesce back onto the blue branch.

SMC’



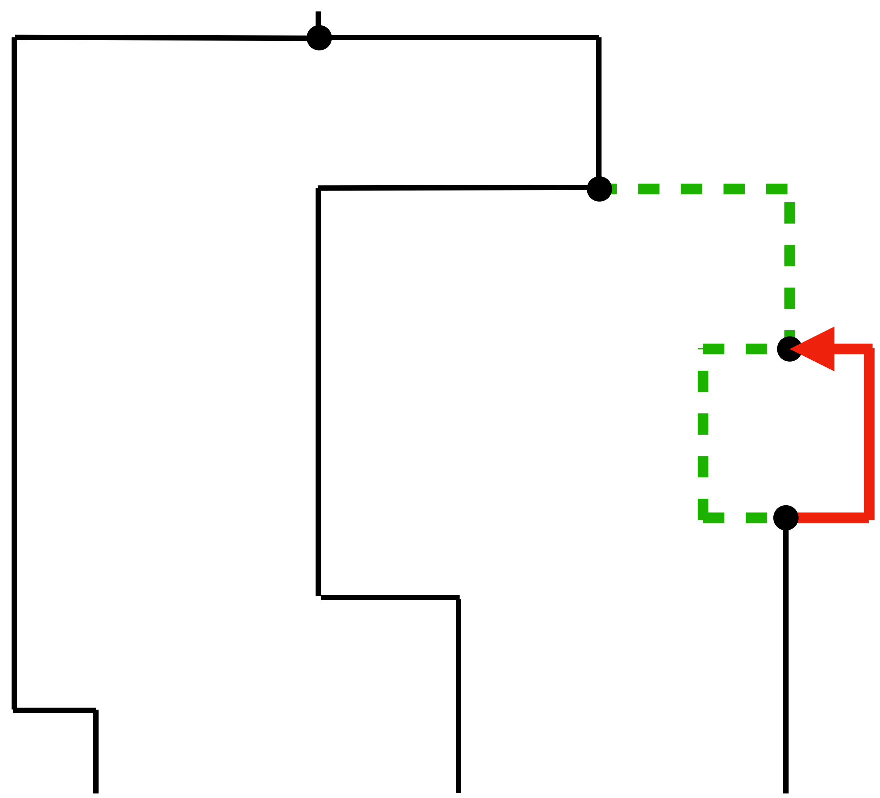
SMC’



Can we distinguish these two trees?

“Diamond” recombinations

Why are they important?



- We need our SMC and SMC’ to explain our data.

- What happens to our interpretation of data if we do not allow “diamond” recombination events?

ARGweaver

MCMC sampling of ARGs

Remove a sample and "weave" it back in:

1. **Sample all new coalescence points:**
2. **Sample new recombination points that reconcile adjacent trees:**