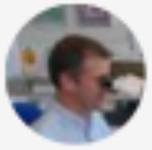
Introduction to phylodynamics

Louise Moncla

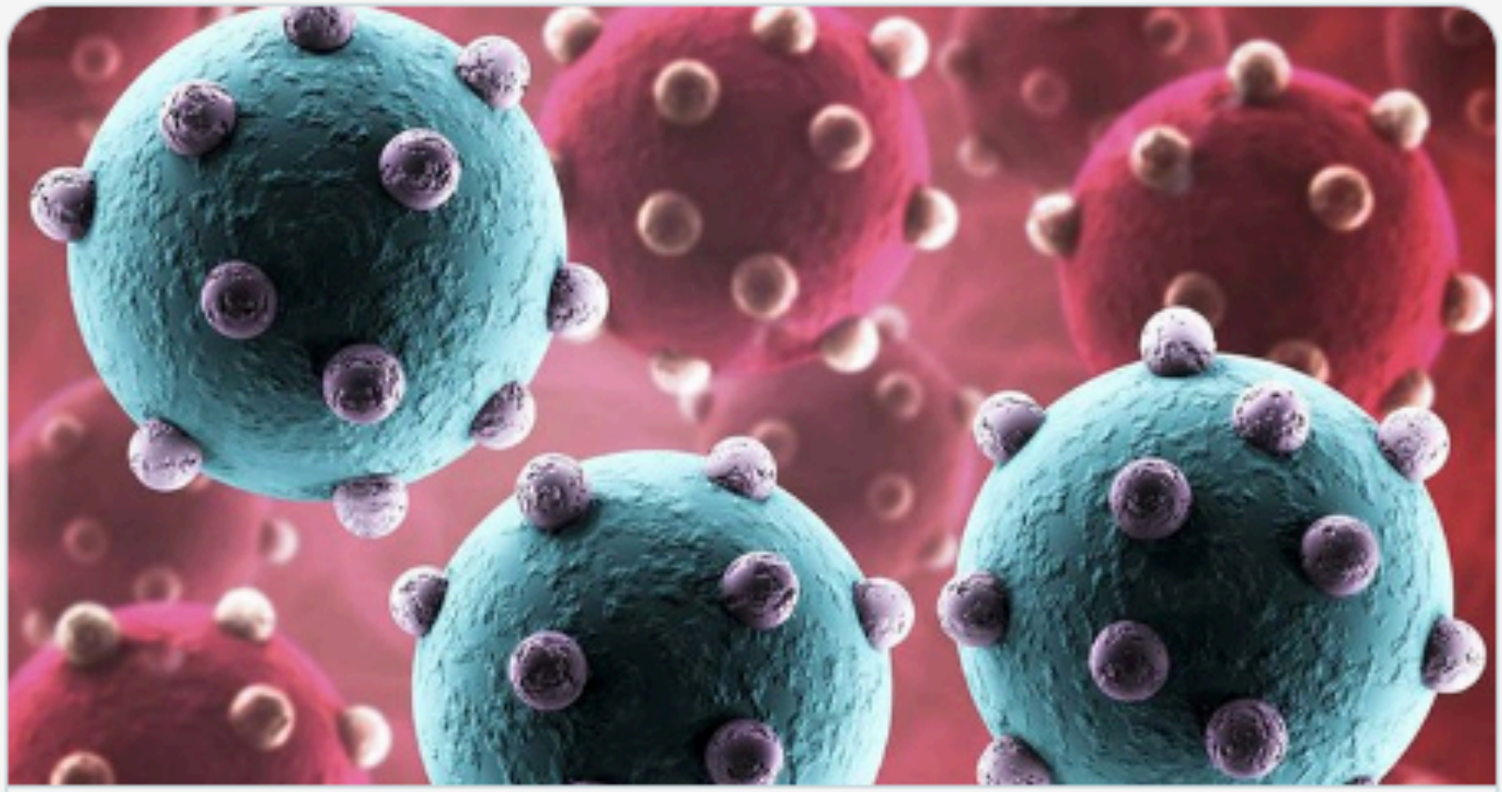
Genomics of Disease in Wildlife Workshop

June 8, 2023



Thijs Kuiken @thijskuiken · Feb 23

Following the death of an 11-year-old girl from highly pathogenic avian #influenza #H5N1 in Prey Veng province, #Cambodia, Ministry of Health reports 12 more infected people, 4 of whom have begun to show symptoms. Results of diagnosis expected tomorrow.

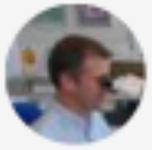


khmertimeskh.com

After death of girl, 12 more possibly detected with H5N1 bird flu in C...

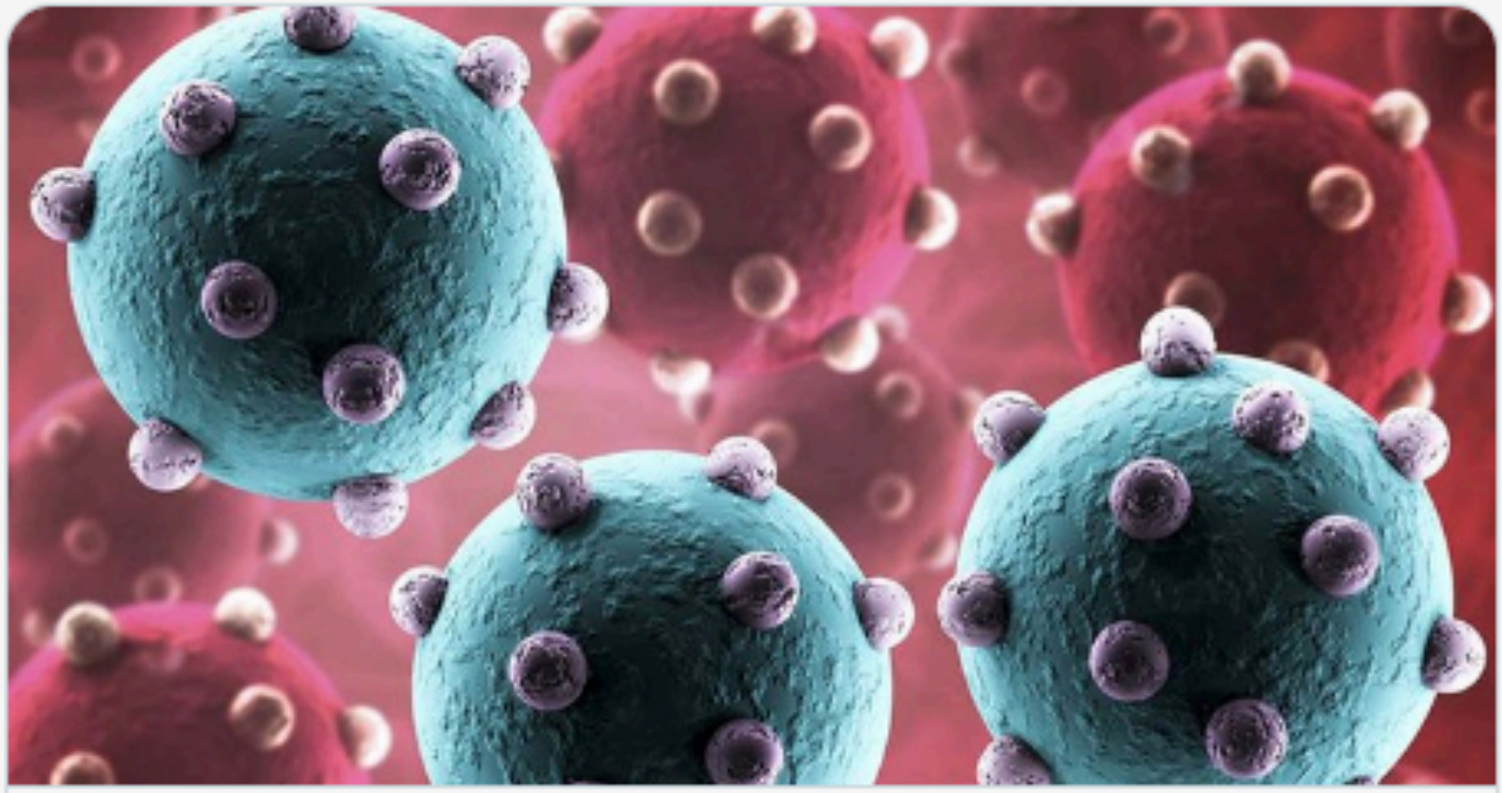
Ms. Youk Sambath, Secretary of State of the Ministry of Health, has confirmed that the Ministry of Health's emergency response team h...

55 1,023 1,507 625K



Thijs Kuiken @thijskuiken · Feb 23

Following the death of an 11-year-old girl from highly pathogenic avian #influenza #H5N1 in Prey Veng province, #Cambodia, Ministry of Health reports 12 more infected people, 4 of whom have begun to show symptoms. Results of diagnosis expected tomorrow.



khmertimeskh.com

After death of girl, 12 more possibly detected with H5N1 bird flu in C...

Ms. Youk Sambath, Secretary of State of the Ministry of Health, has confirmed that the Ministry of Health's emergency response team h...

55 1,023 1,507 625K

Jurre Y Siegers, PhD @jurreysi · Feb 23

There seems to be a translation error from Khmer to English. The report below (in khmer) states there are 12 contacts to the index case and 4 amongst those 12 report flu like symptoms.

The questions: Is this human case related to the ongoing outbreak in Europe/the Americas? Does this reflect a new epidemiological pattern?

Erik Karlsson
@E_A_Karlsson

Happy to announce that the full genome sequence of the Cambodian #H5N1 #avianflu case is now available on @GISAID. This has been an incredible effort by a number of people: (1/)

12:15 AM · Feb 26, 2023 · **532.9K** Views

387 Retweets **74** Quotes **1,615** Likes **85** Bookmarks

Phylogeny

ZOOM TO SELECTED | RESET LAYOUT

Country

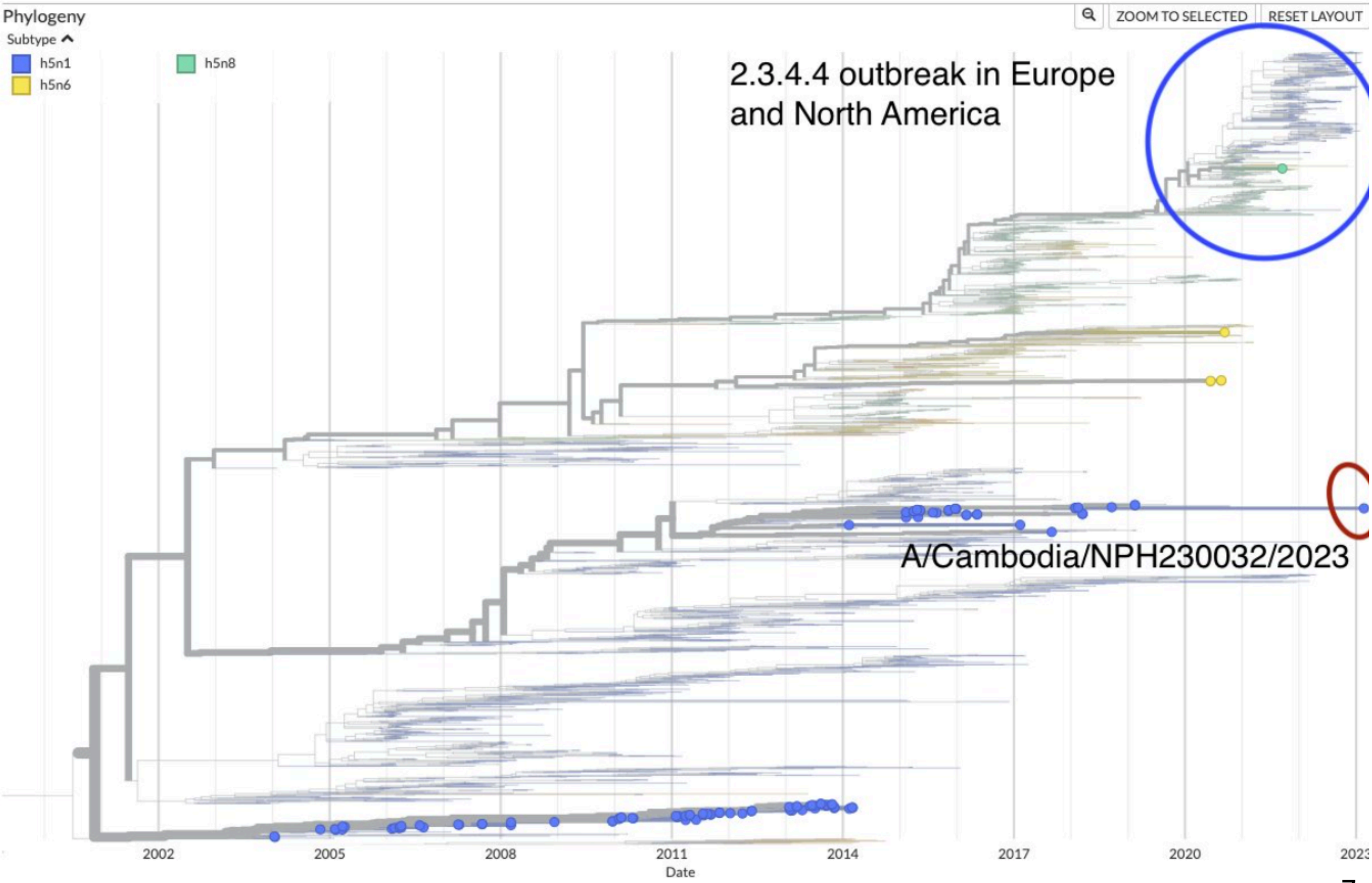
Vietnam
Cambodia
Laos

A/chicken/Vietnam/HU12666/2019
A/chicken/Vietnam/HU12661/2019
A/chicken/Vietnam/HU12668/2019
A/chicken/Cambodia/13/2019
A/chicken/Cambodia/14/2019
A/chicken/LongAn/1428VTC/2019
A/chicken/LongAn/1427VTC/2019
A/duck/Vietnam/HU10-1756/2018
A/chicken/Vietnam/HU10-1051/2018
A/chicken/Vietnam/HU10-1044/2018
A/Muscovy_duck/Vietnam/HU10-1139/2018
A/duck/Lao/NL2072063/2020
A/duck/Lao/NL2072060/2020
A/duck/Lao/NL2072064/2020
A/duck/Lao/NL2072066/2020
A/duck/Lao/NL2072073/2020
A/duck/Vietnam/HU10-1815/2018
A/chicken/Cambodia/12/2018
A/chicken/Cambodia/01/2018
A/Cambodia/NPH2300322/2023
A/Cambodia/2302009/2023
A/Cambodia/NPH230032/2023
A/chicken/Cambodia/02/2018
A/chicken/Cambodia/03/2018
A/chicken/Cambodia/Z9W7M1/2015
A/chicken/Cambodia/Z850W49M1/2015
A/chicken/Cambodia/a0511503/2016
A/chicken/Cambodia/a0511504/2016
A/chicken/Cambodia/11/2018
A/duck/Cambodia/a23W9M1/2016
A/chicken/Cambodia/a27W9M1/2016
A/Muscovy_duck/Vietnam/HU3_584/2015
A/duck/Vietnam/HU5_1571/2016
A/duck/Vietnam/HU5_1574/2016
A/duck/Vietnam/HU5-1574/2016
A/duck/Vietnam/HU5-1573/2016
A/duck/Vietnam/HU8-1561/2017
A/muscovyduck/Vietnam/HU8-1670/2017
A/environment/Vietnam/HU3_1485/2015
A/Muscovy_duck/Vietnam/HU3_1079/2015
A/duck/Vietnam/HU3-16/2015
A/duck/Vietnam/HU3_1265/2015
A/duck/Cambodia/Z765W45M1/2015
A/environment/Cambodia/Z72EW10M3/2015
A/chicken/Cambodia/a17W9M1/2016
A/environment/Cambodia/Z92CW11M3/2015
A/chicken/SanasombounDistrict/NL1942882/2017
A/chicken/SanasombounDistrict/NL1942887/2017
A/muscovyduck/LongAn/88/2014

0.195 0.200 0.205 0.210 0.215 0.220 0.225 0.230 0.235

Divergence

Showing 108 of 3039 genomes sampled between Jan 2004 and Feb 2023. Filtered to Cambodia (108) .



The answers:

- These cases were caused by a lineage endemic to Cambodia, likely by direct interaction with sick, wild birds

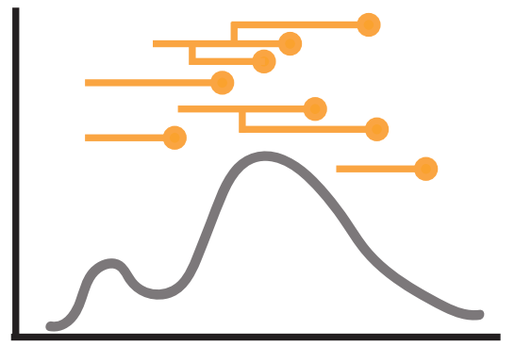
- These cases do not represent evolution of human to human transmission of 2.3.4.4b viruses

Phylodynamics is the study of how epidemiological, immunological, and evolutionary processes act and interact to shape viral phylogenies.

* Volz, Koelle, and Bedford, PLOS Computational Biology, 2013

What types of questions can phylodynamics answer?

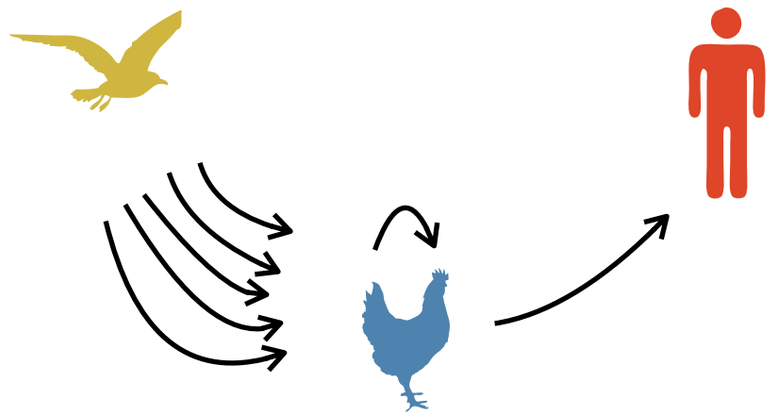
1. When did an outbreak begin, and where? How rapidly is it growing? Who is driving transmission?



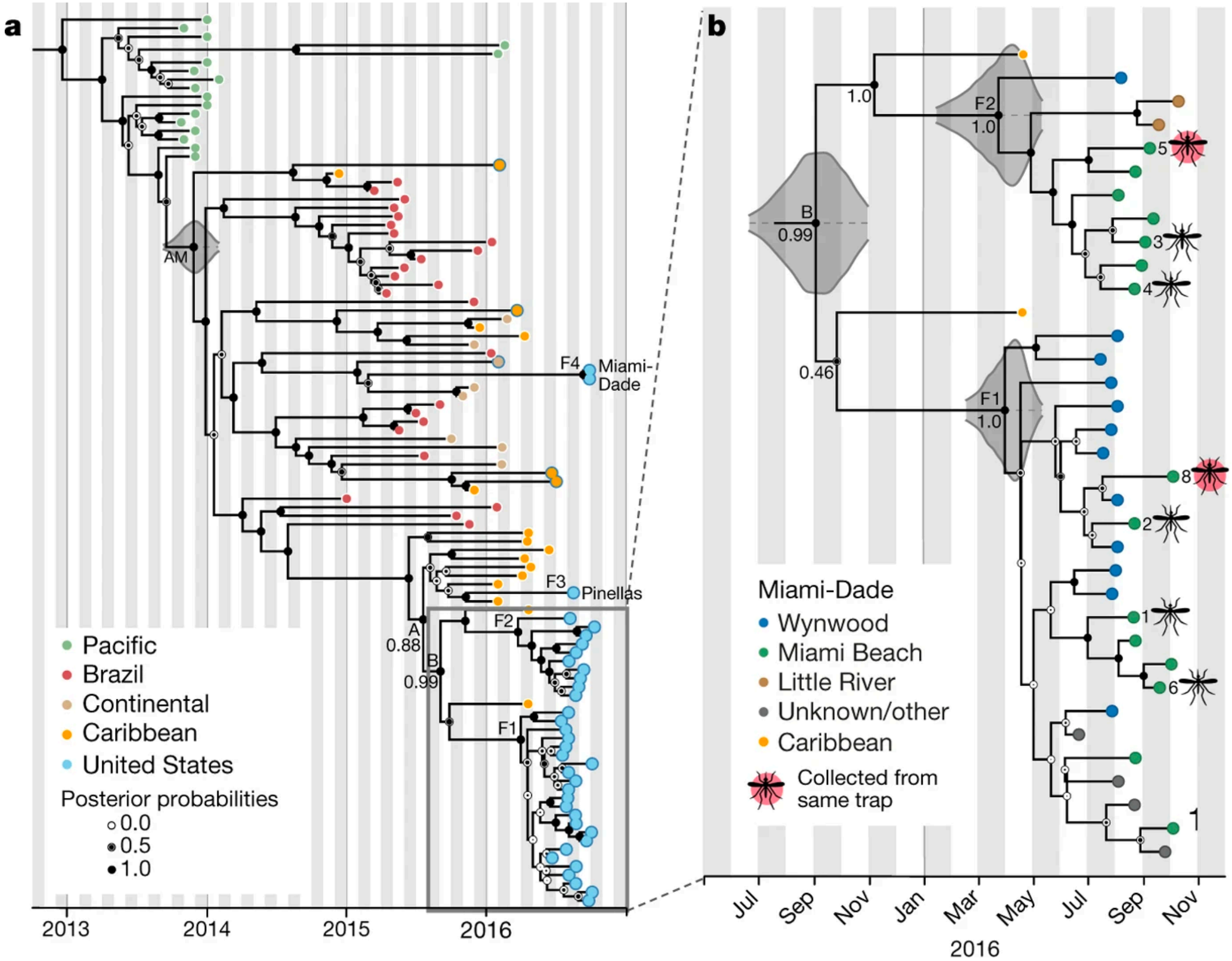
2. Who infected whom?

3. How are pathogen populations shaped by natural selection?

4. How are pathogen populations moving across time, space, and hosts?

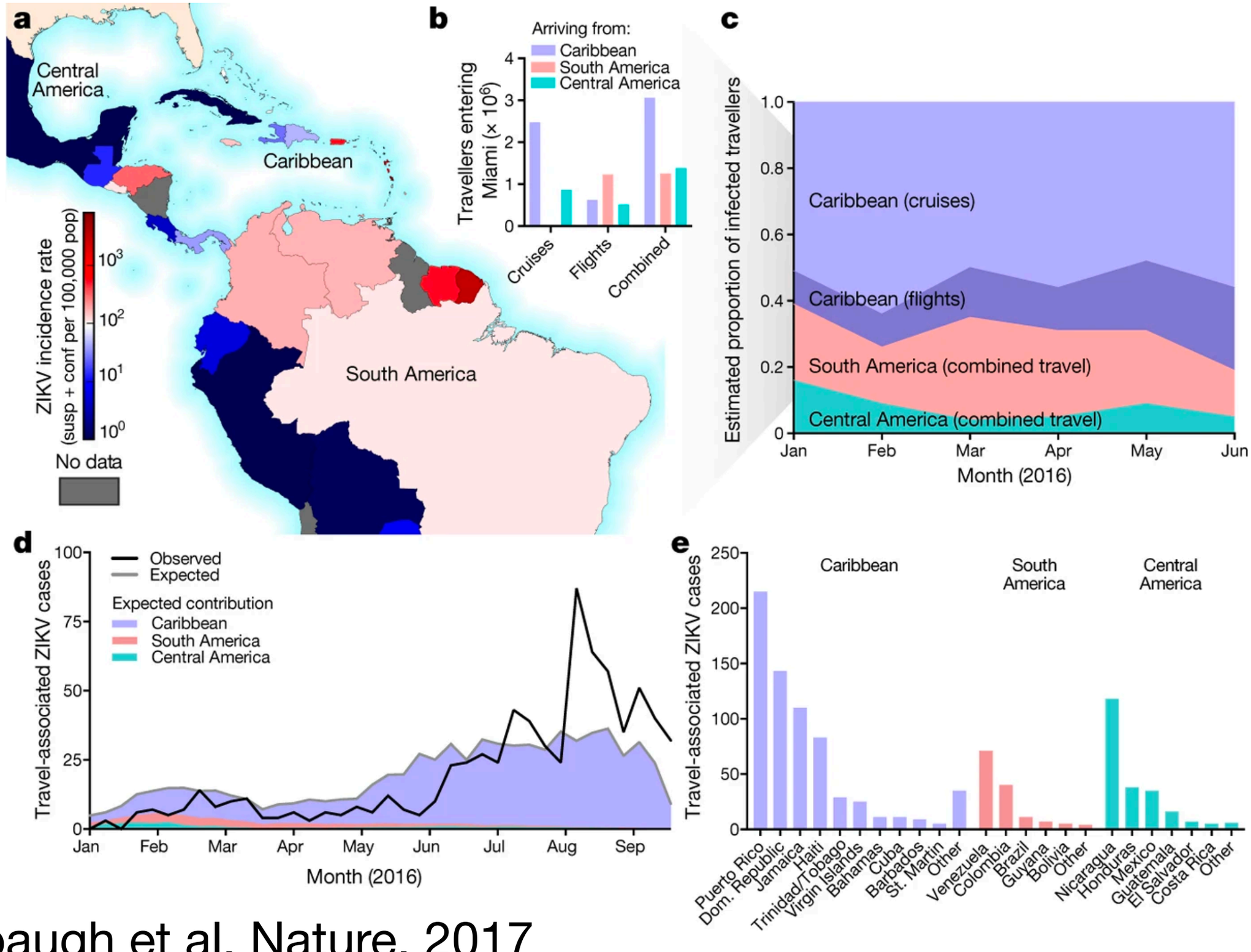


Inferring Zika introductions into Florida



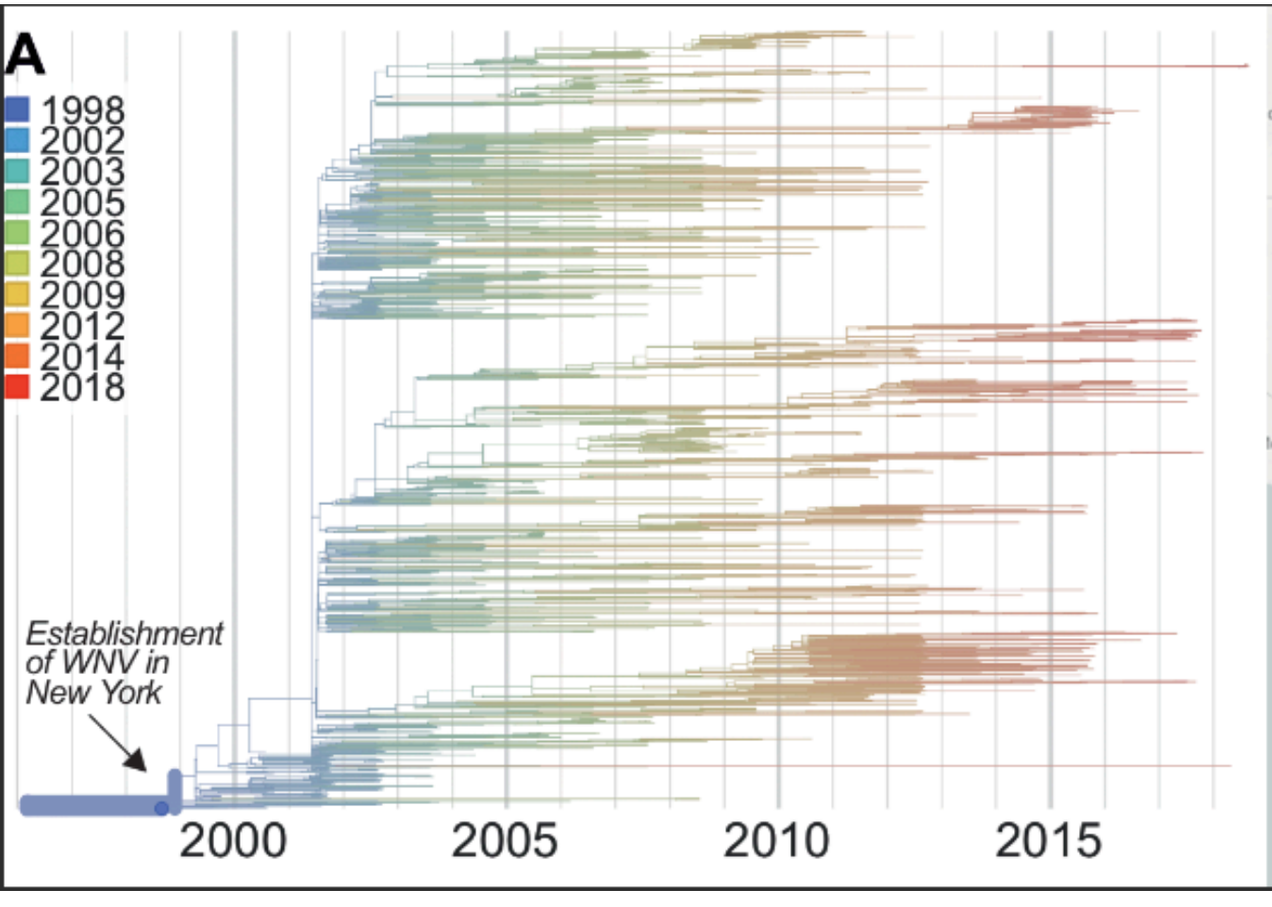
Grubaugh et al, Nature, 2017

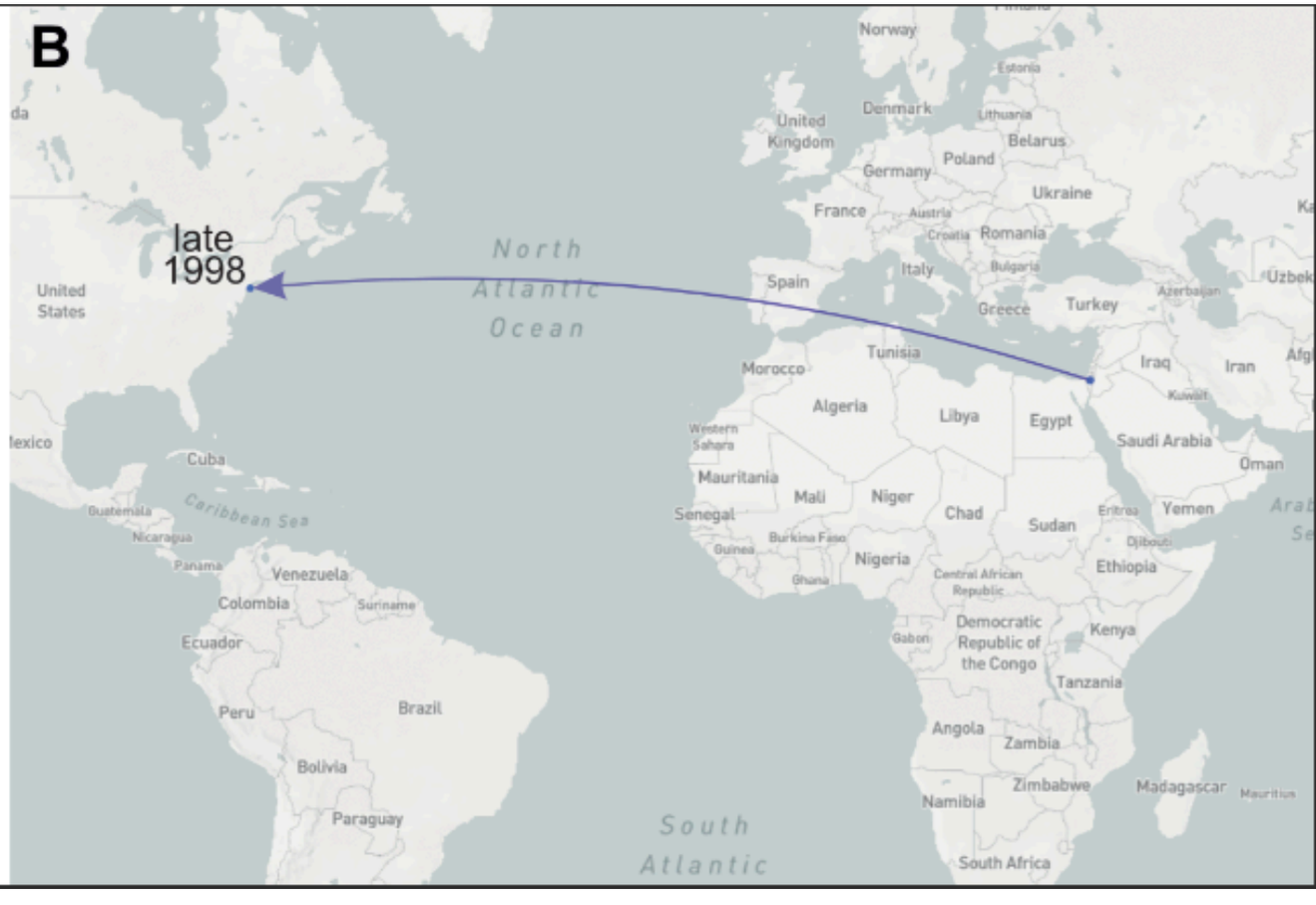
Mosquito abundance and cruise ship traffic made Miami uniquely at risk for Zika



Grubaugh et al, Nature, 2017

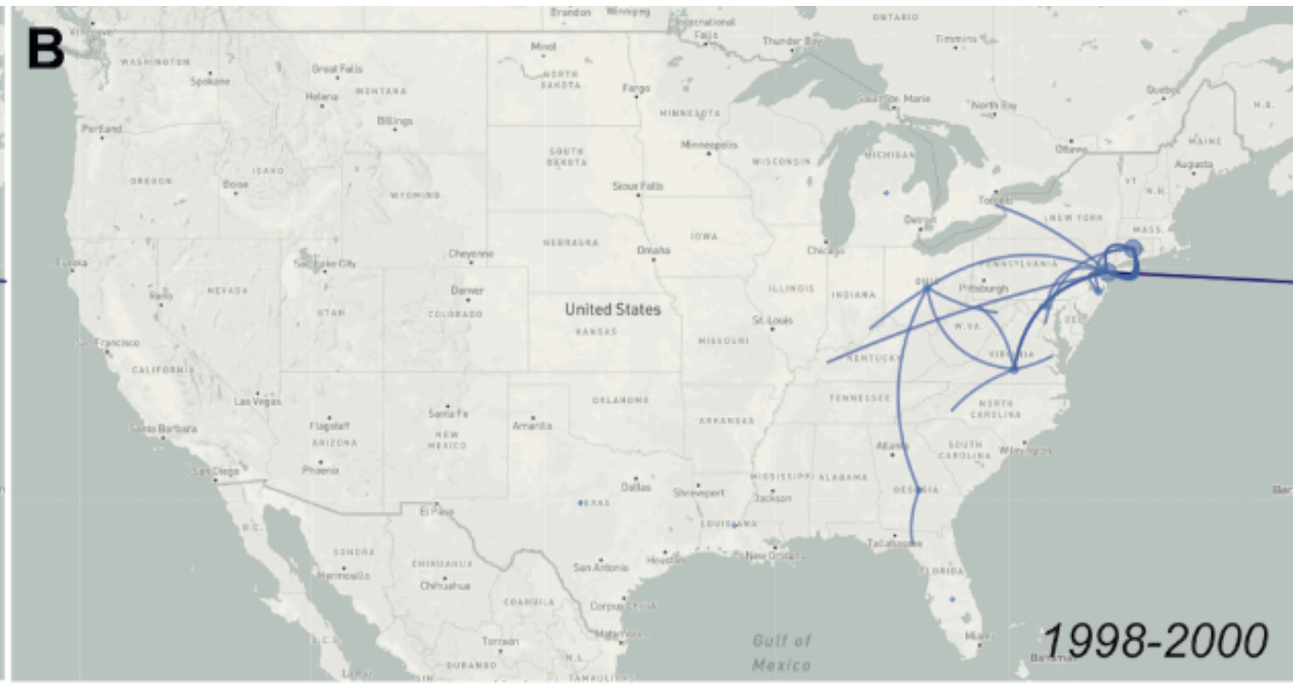
West Nile virus was first introduced into North America in the late 1990s

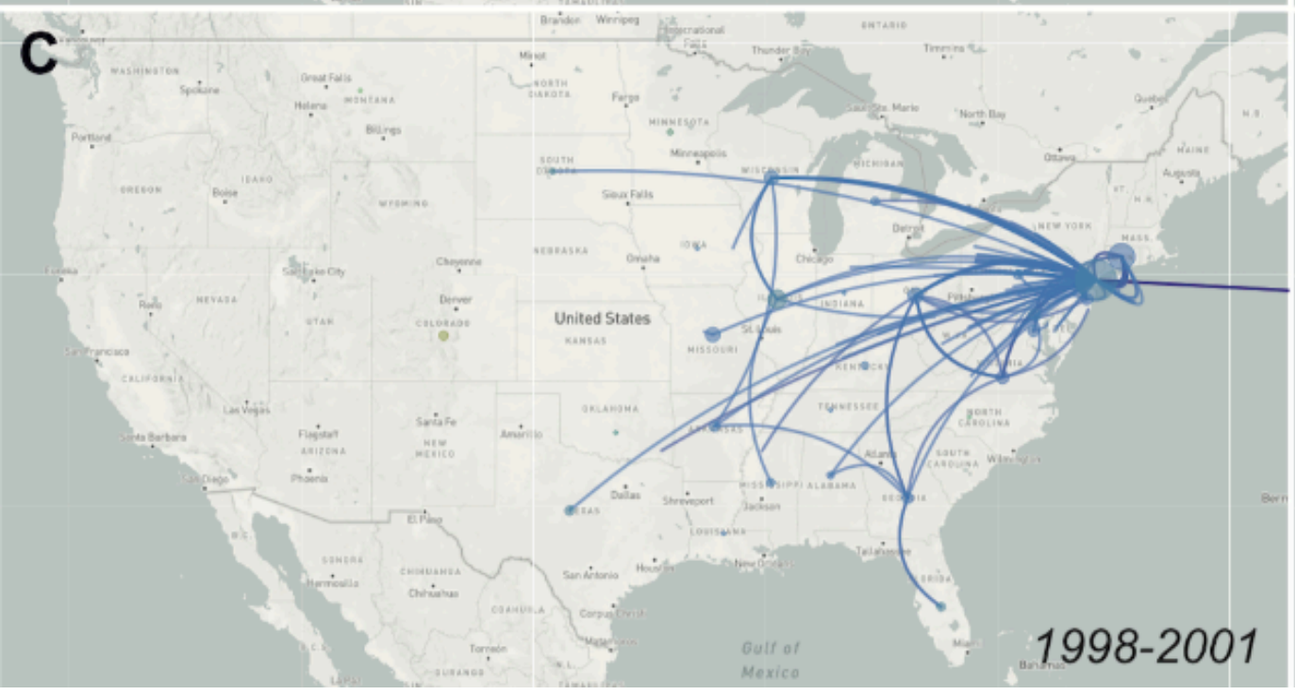


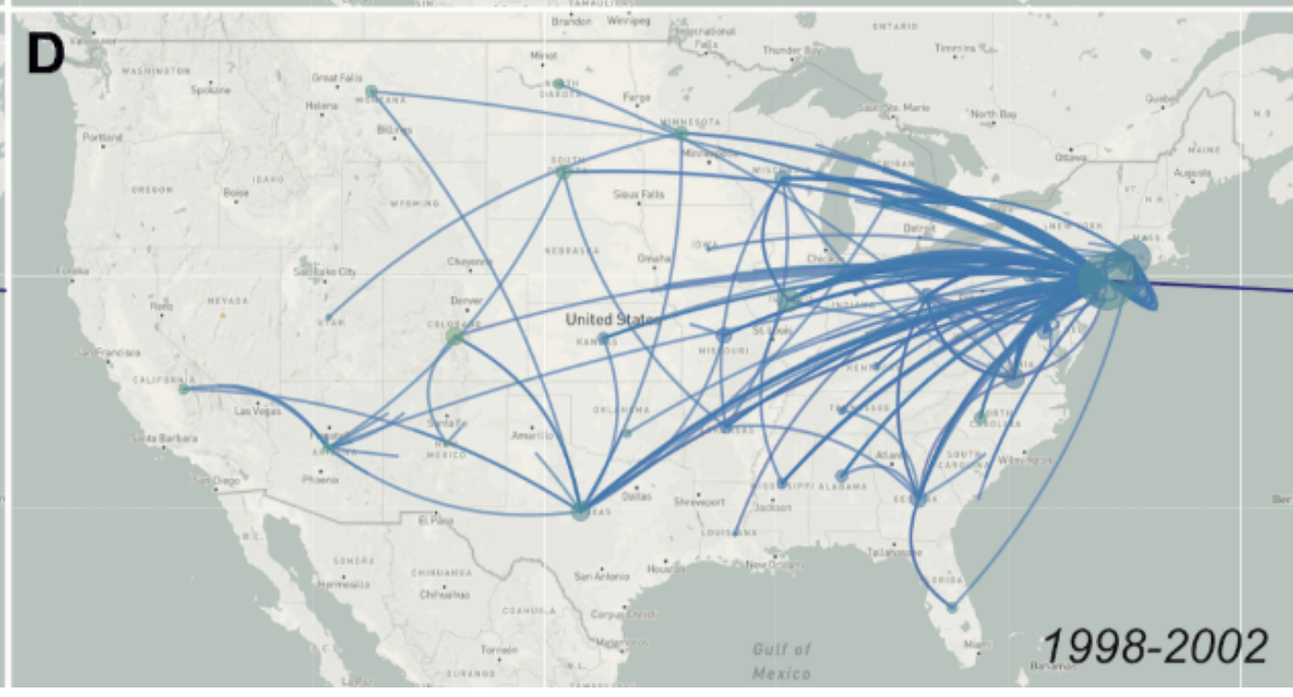


Hadfield et al, PLOS Pathogens, 2019

West Nile virus spread from the east coast







Hadfield et al, PLOS Pathogens, 2019

The key points we'll be making today:

1. Pathogen evolution happens on the same timescale as transmission, meaning that for rapidly evolving pathogens, **evolution, ecology, and epidemiology are linked.**

2. **Coalescent theory** allows us to generate expectations for phylogenies generated by populations that are evolving in particular ways.

3. These features allow us to use trees to infer parameters, and vice versa. We will see **examples** of how others have done this.

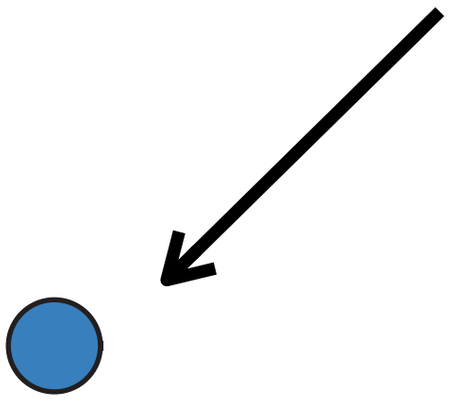
4. How can **we apply** these concepts to our own questions and data?

Phylodynamics is predicated on the fact that for rapidly evolving pathogens, ecological, evolutionary, and epidemiological dynamics all occur on the same timescale.

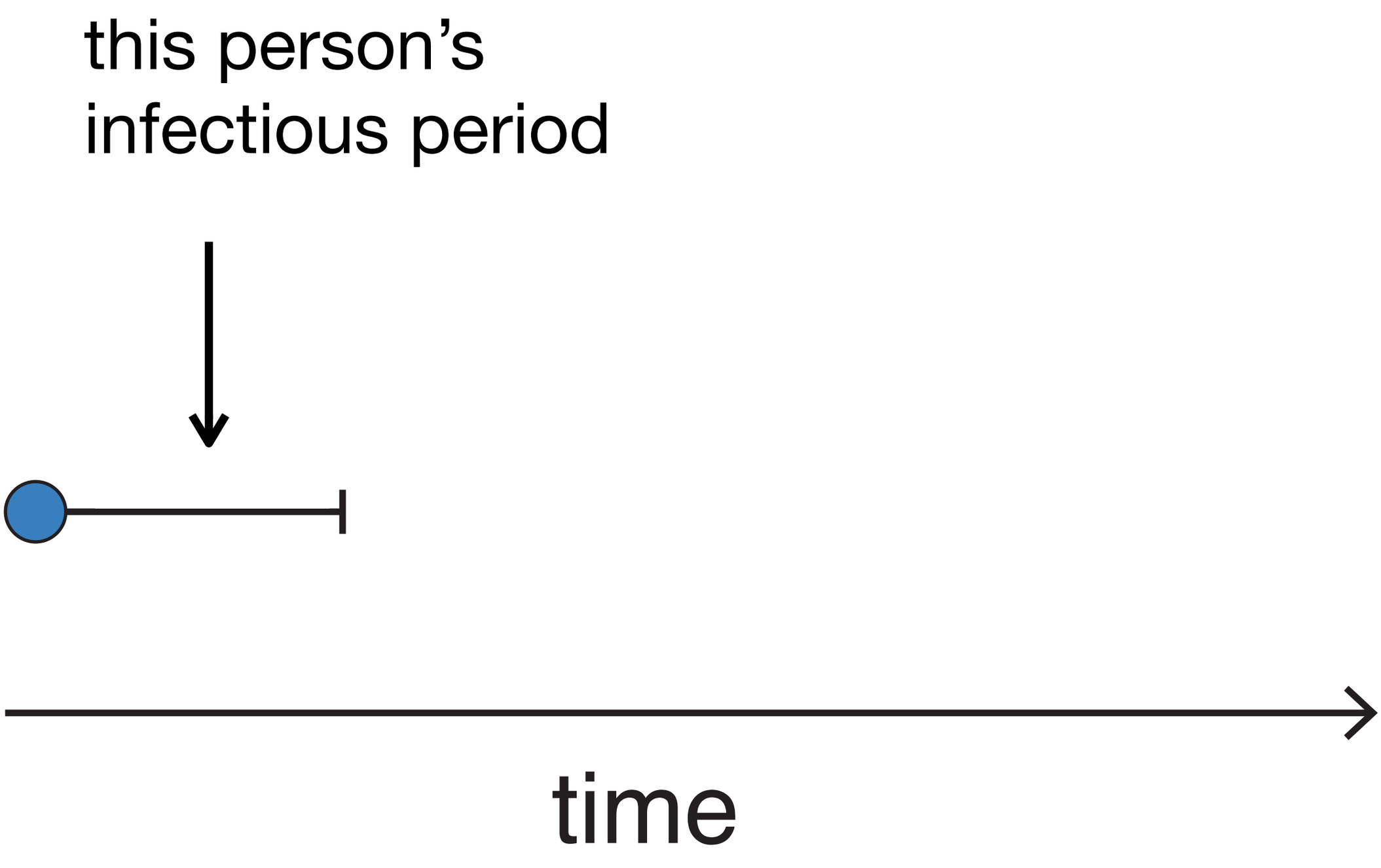
Viral genomes contain records of a virus's transmission history

Viral genomes contain records of a virus's transmission history

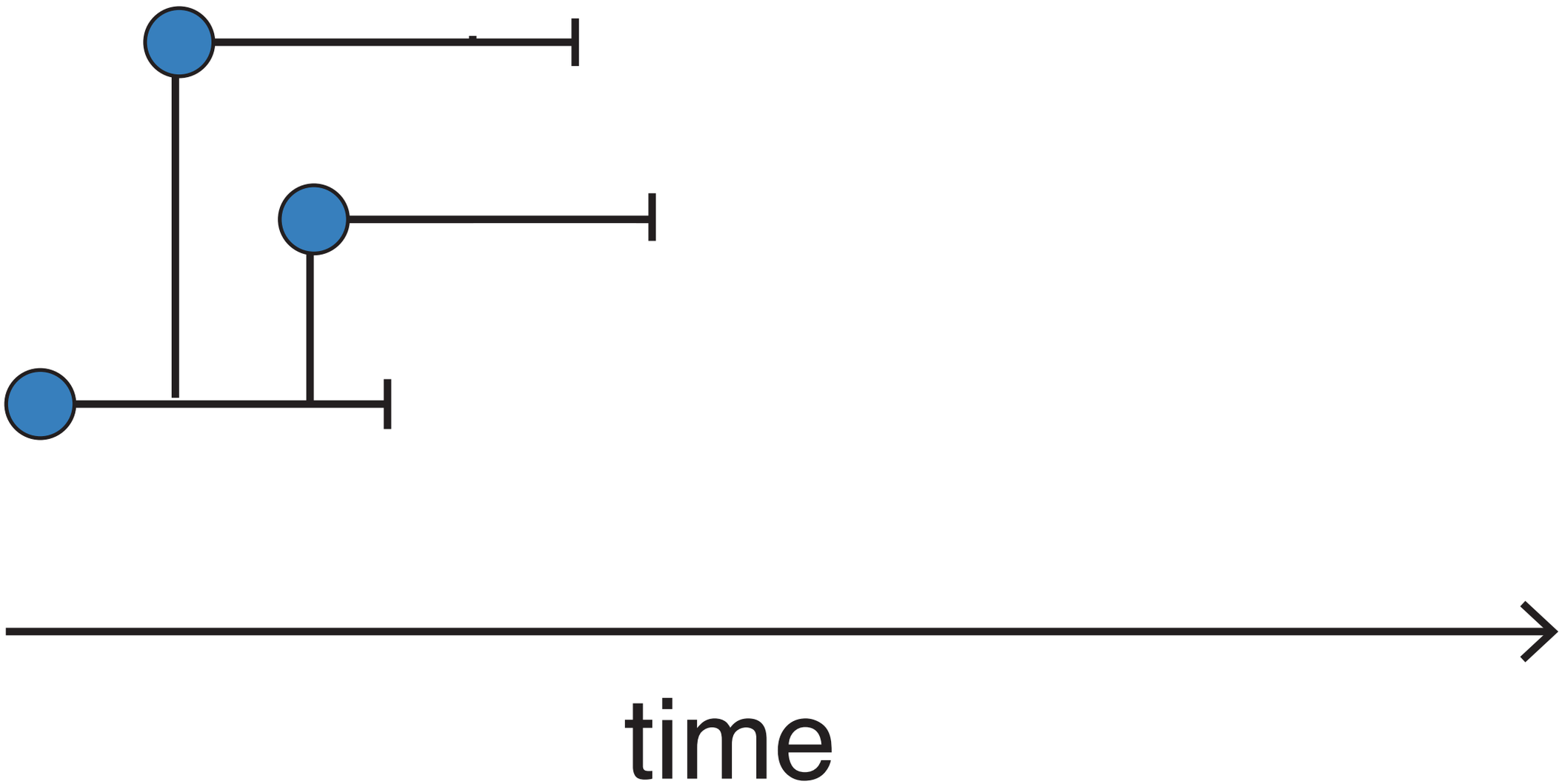
one person
infected with
influenza



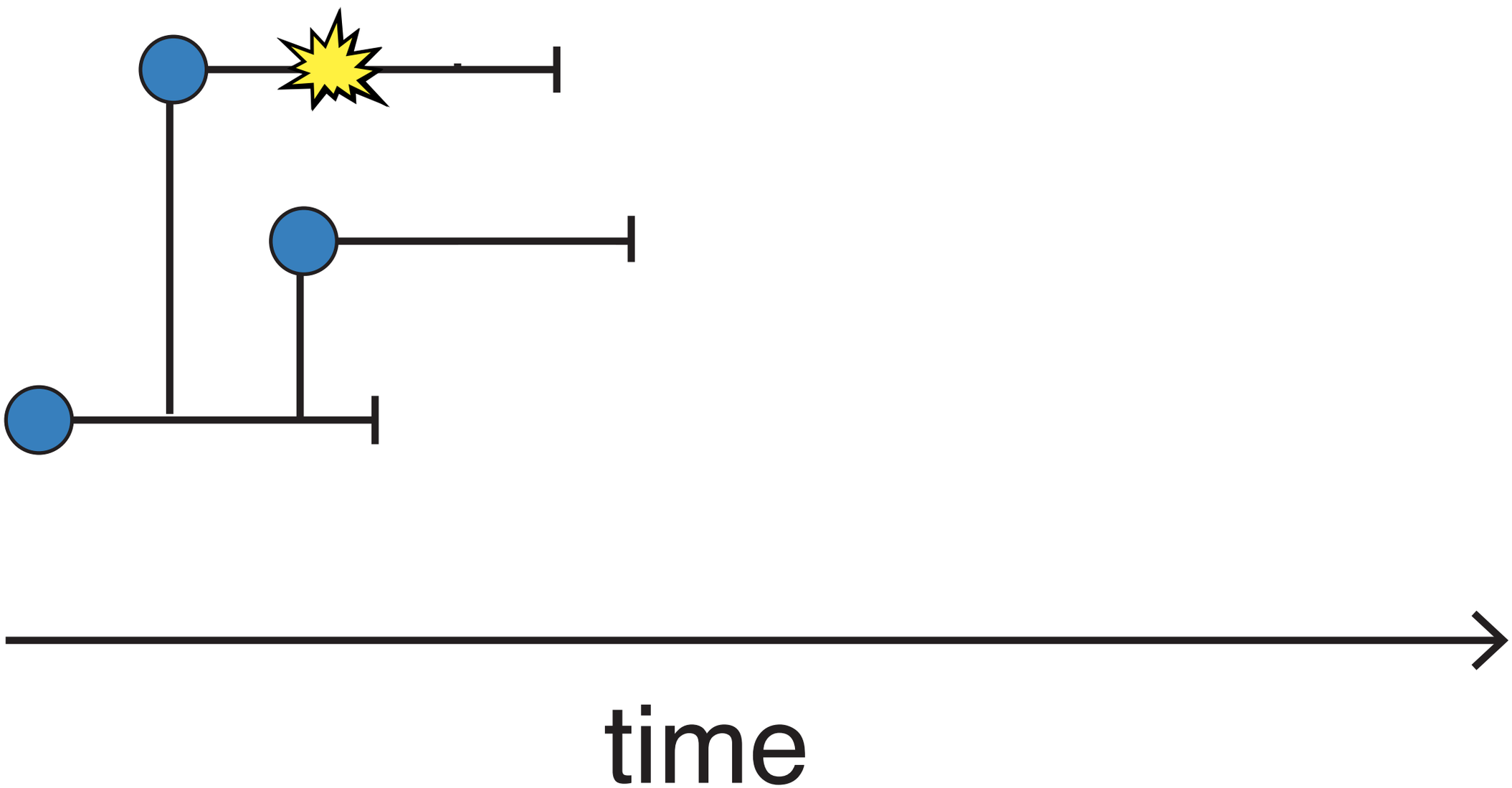
Viral genomes contain records of a virus's transmission history



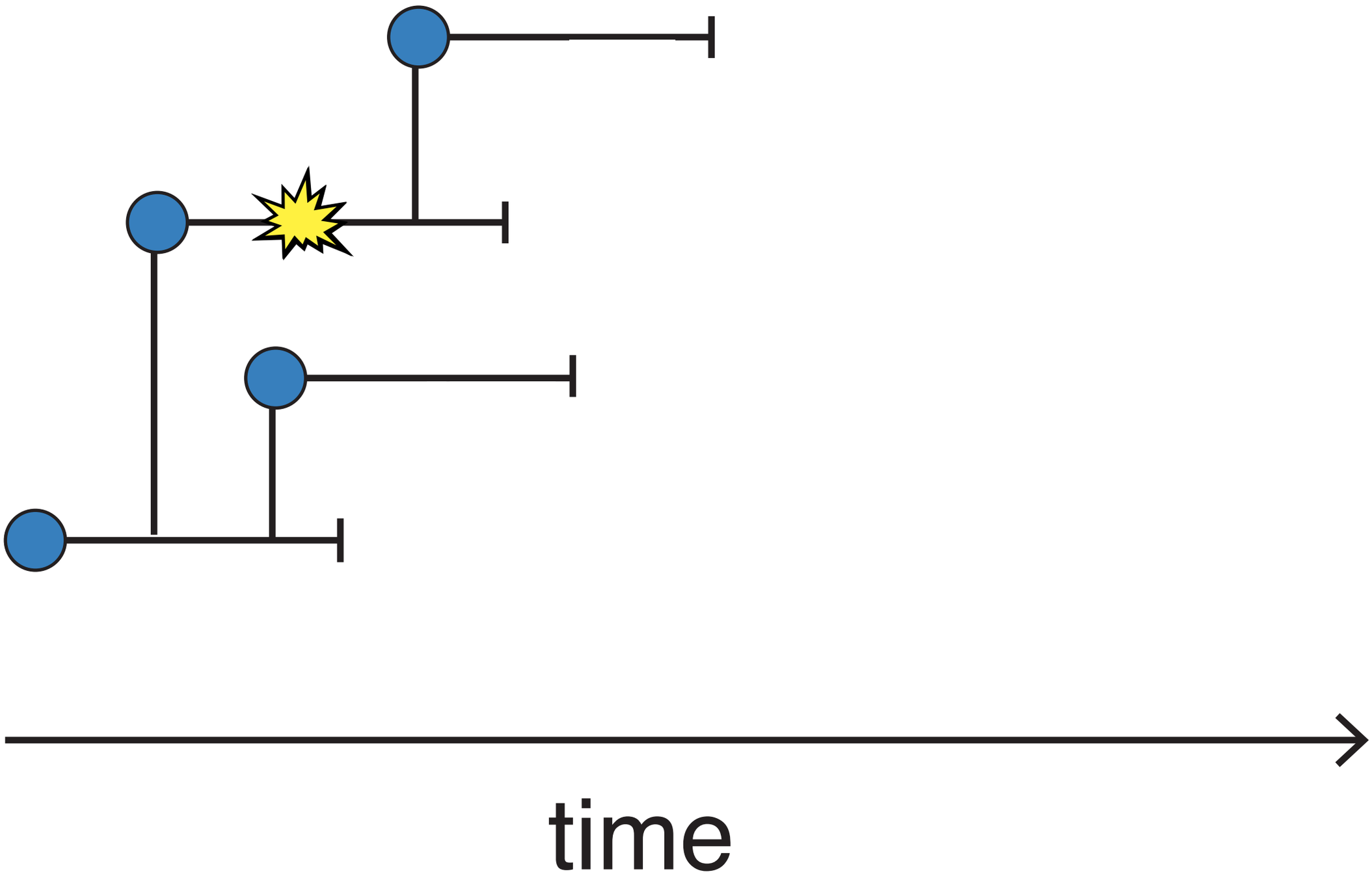
Viral genomes contain records of a virus's transmission history



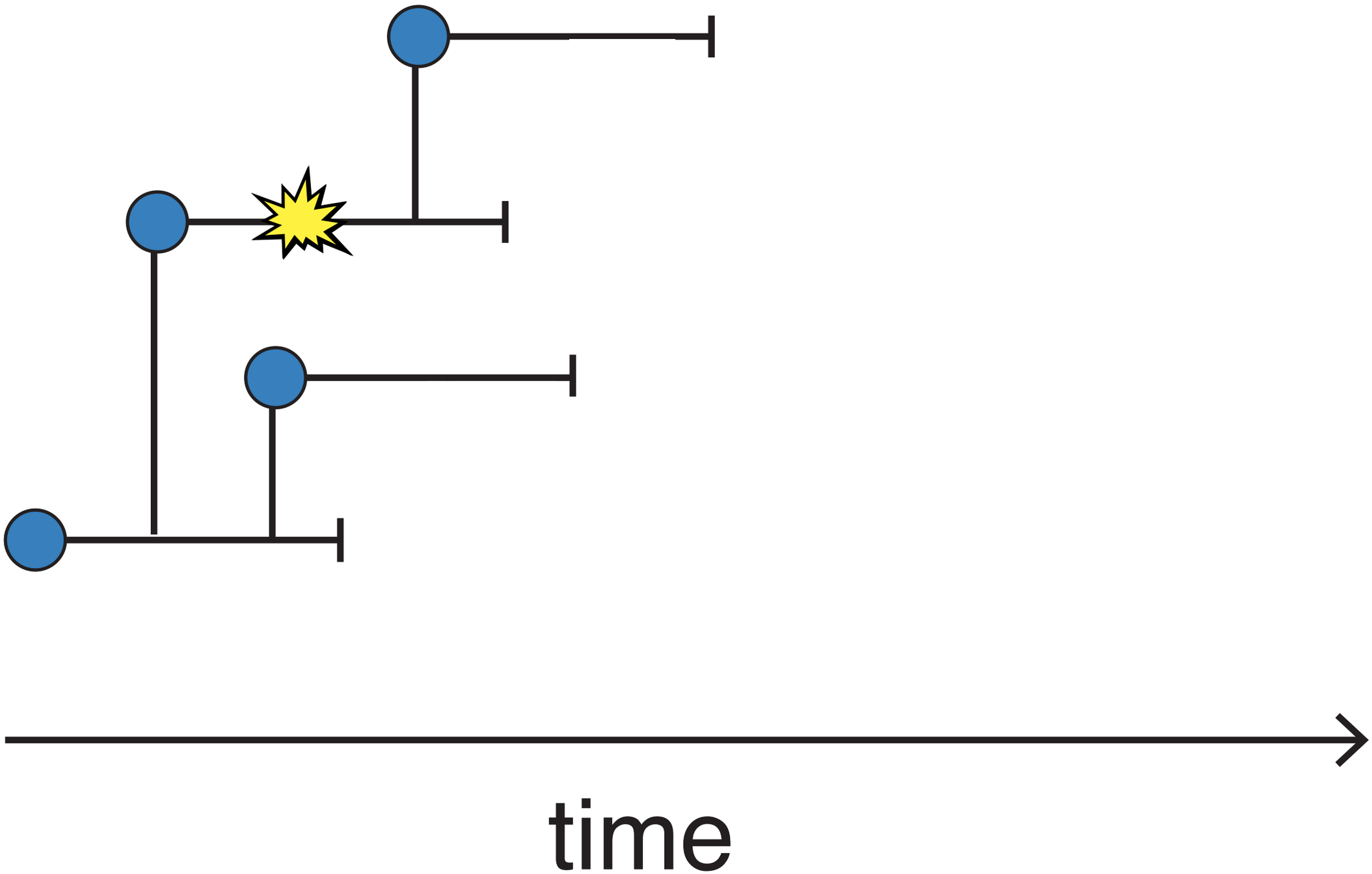
Viral genomes contain records of a virus's transmission history



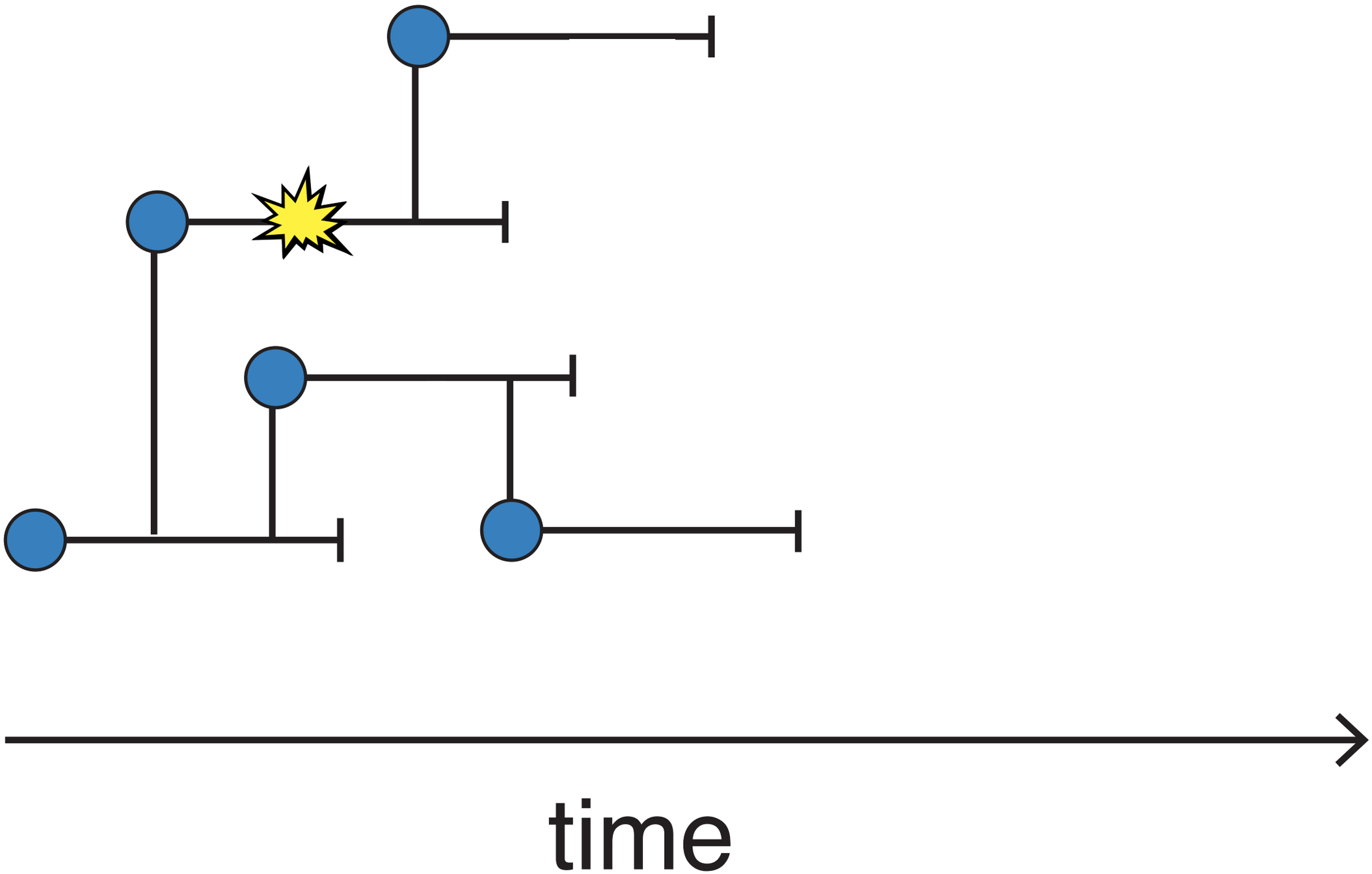
Viral genomes contain records of a virus's transmission history



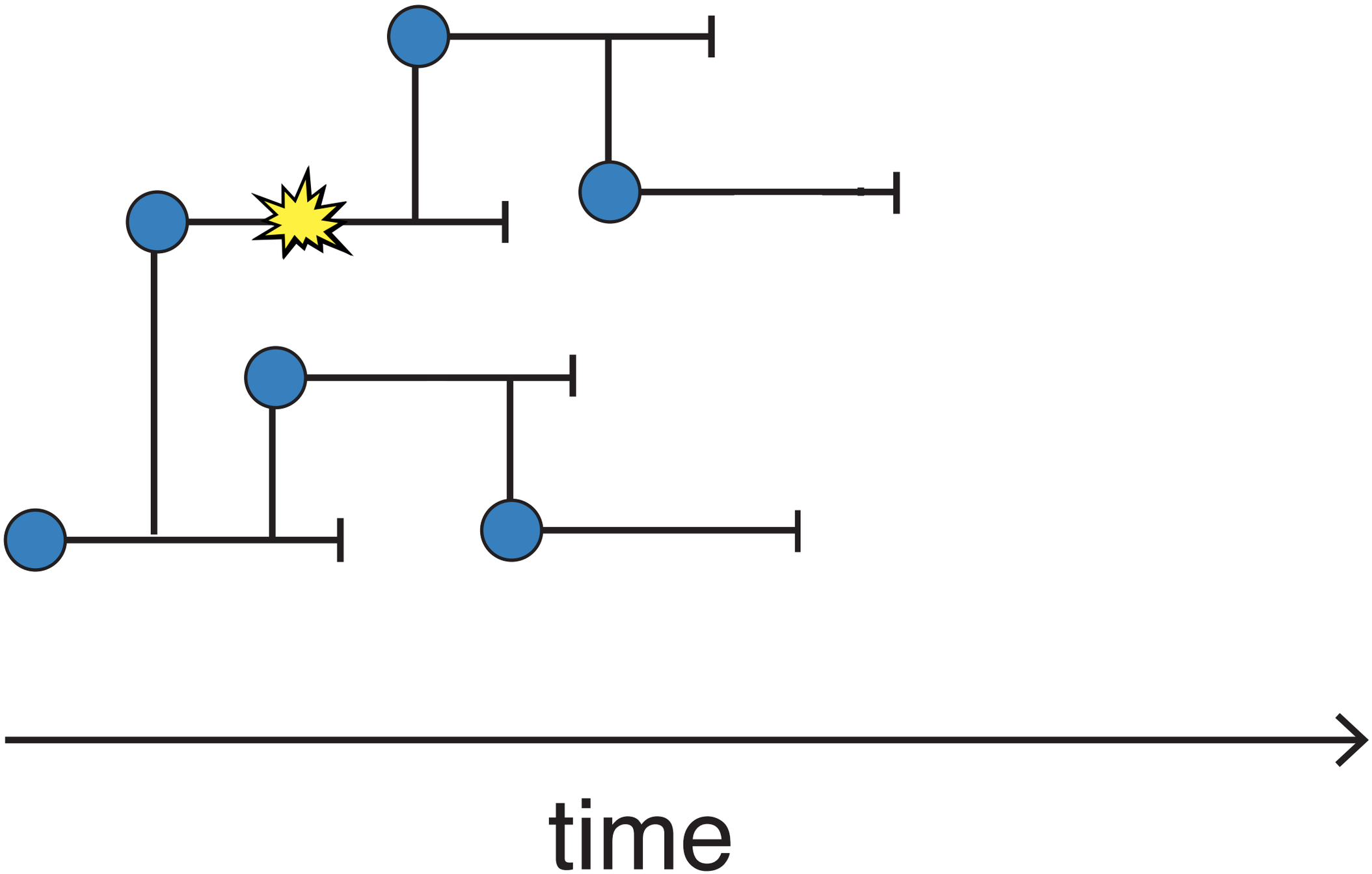
Viral genomes contain records of a virus's transmission history



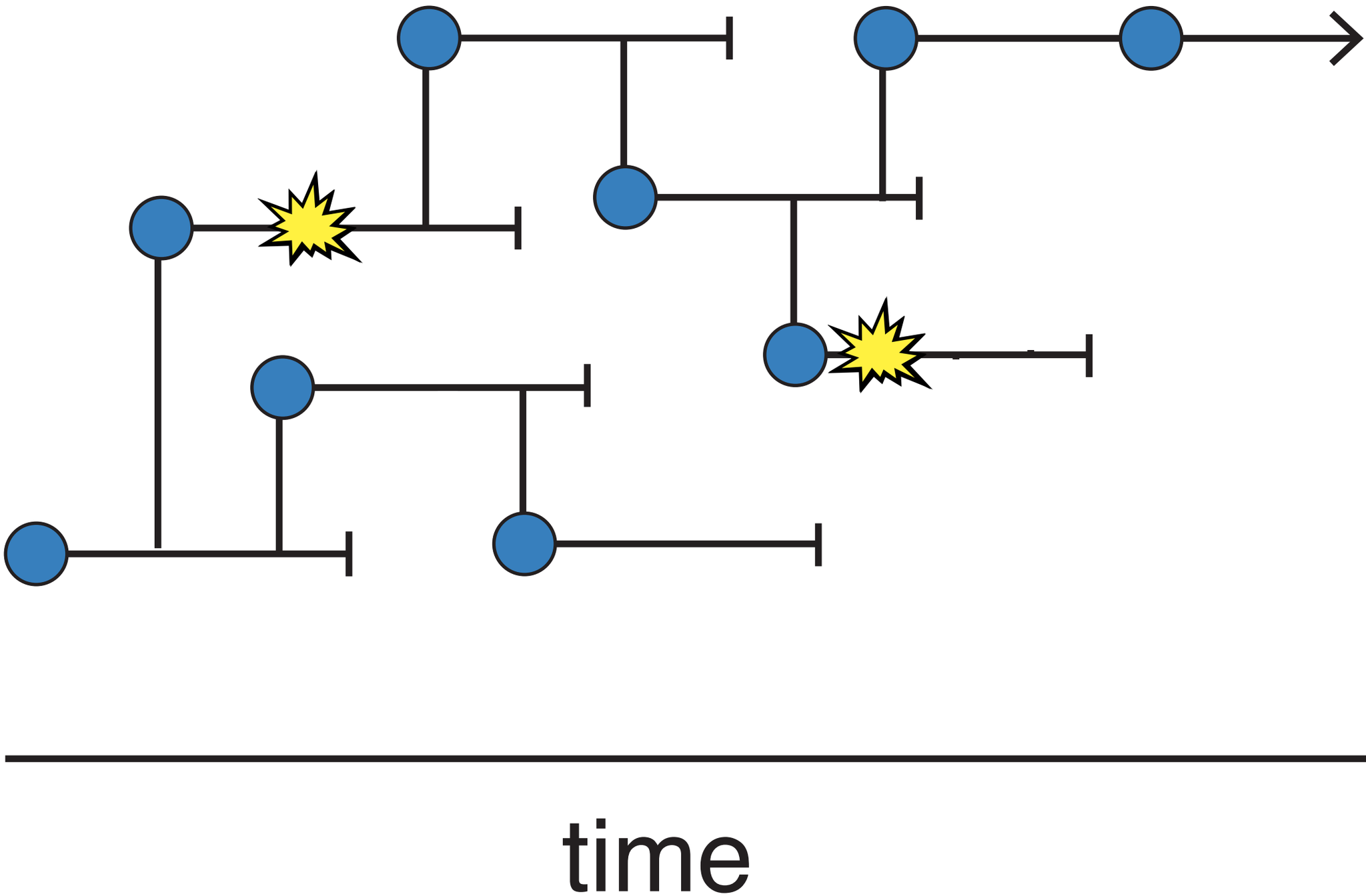
Viral genomes contain records of a virus's transmission history



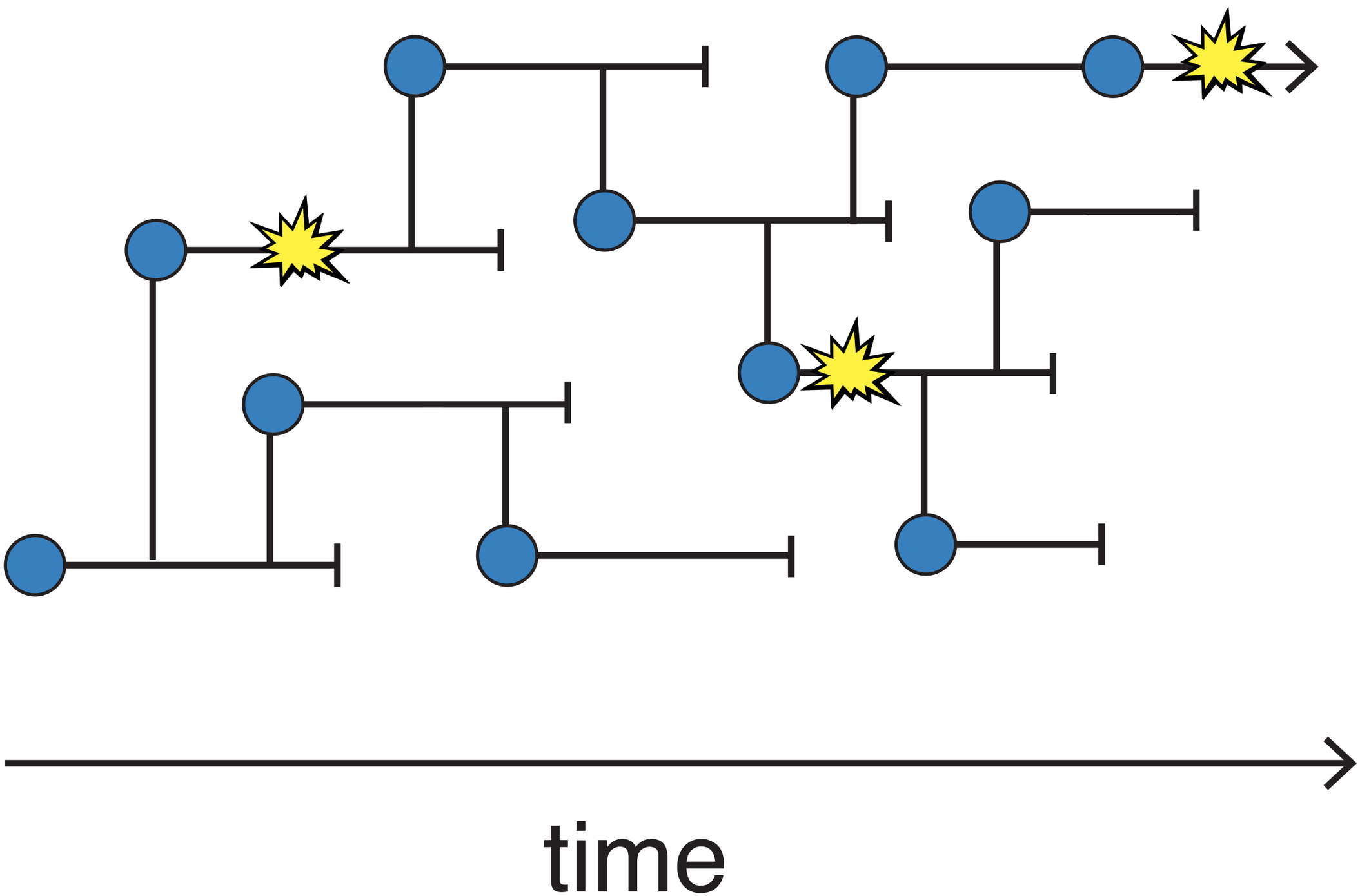
Viral genomes contain records of a virus's transmission history



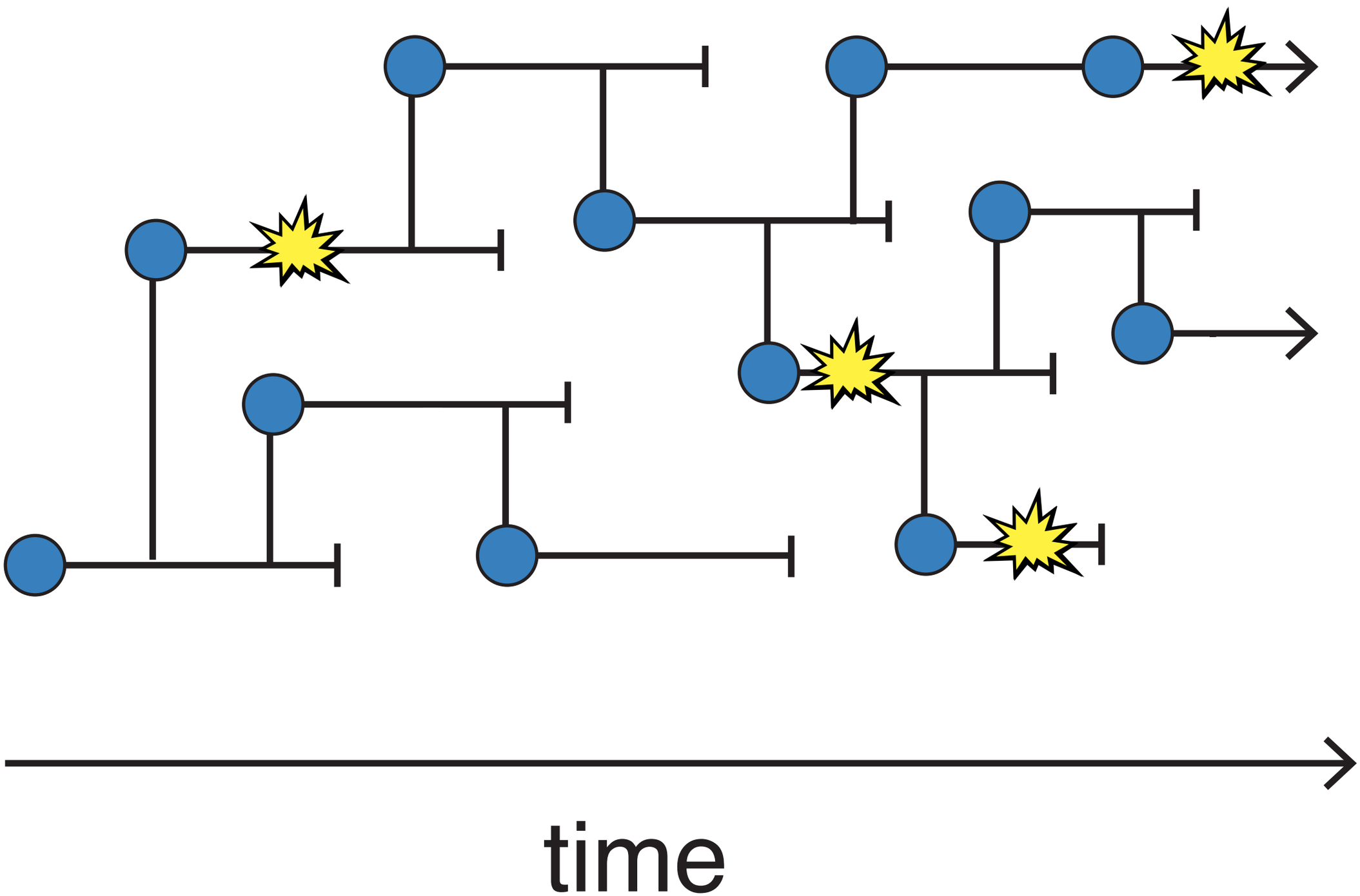
Viral genomes contain records of a virus's transmission history



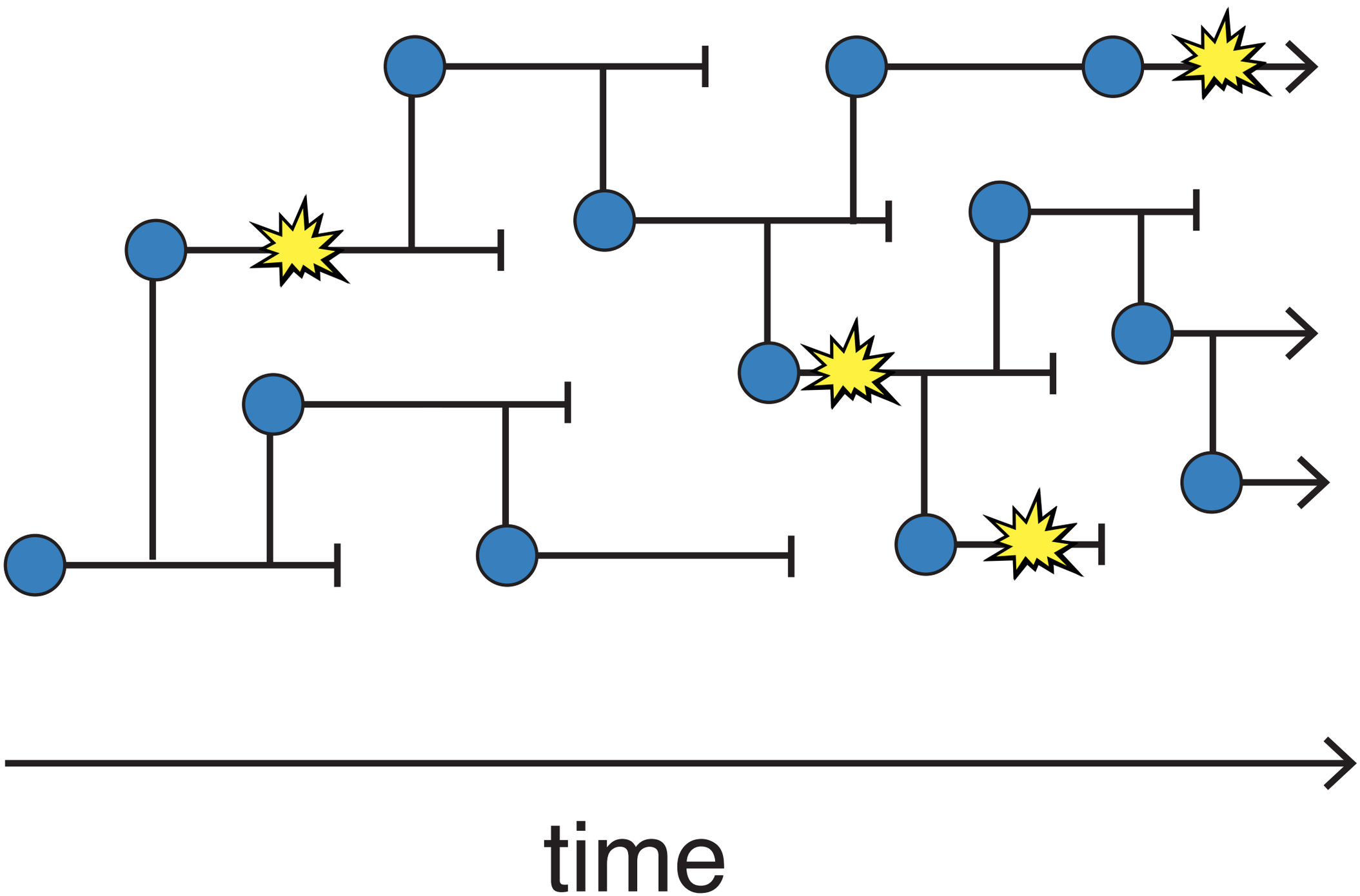
Viral genomes contain records of a virus's transmission history



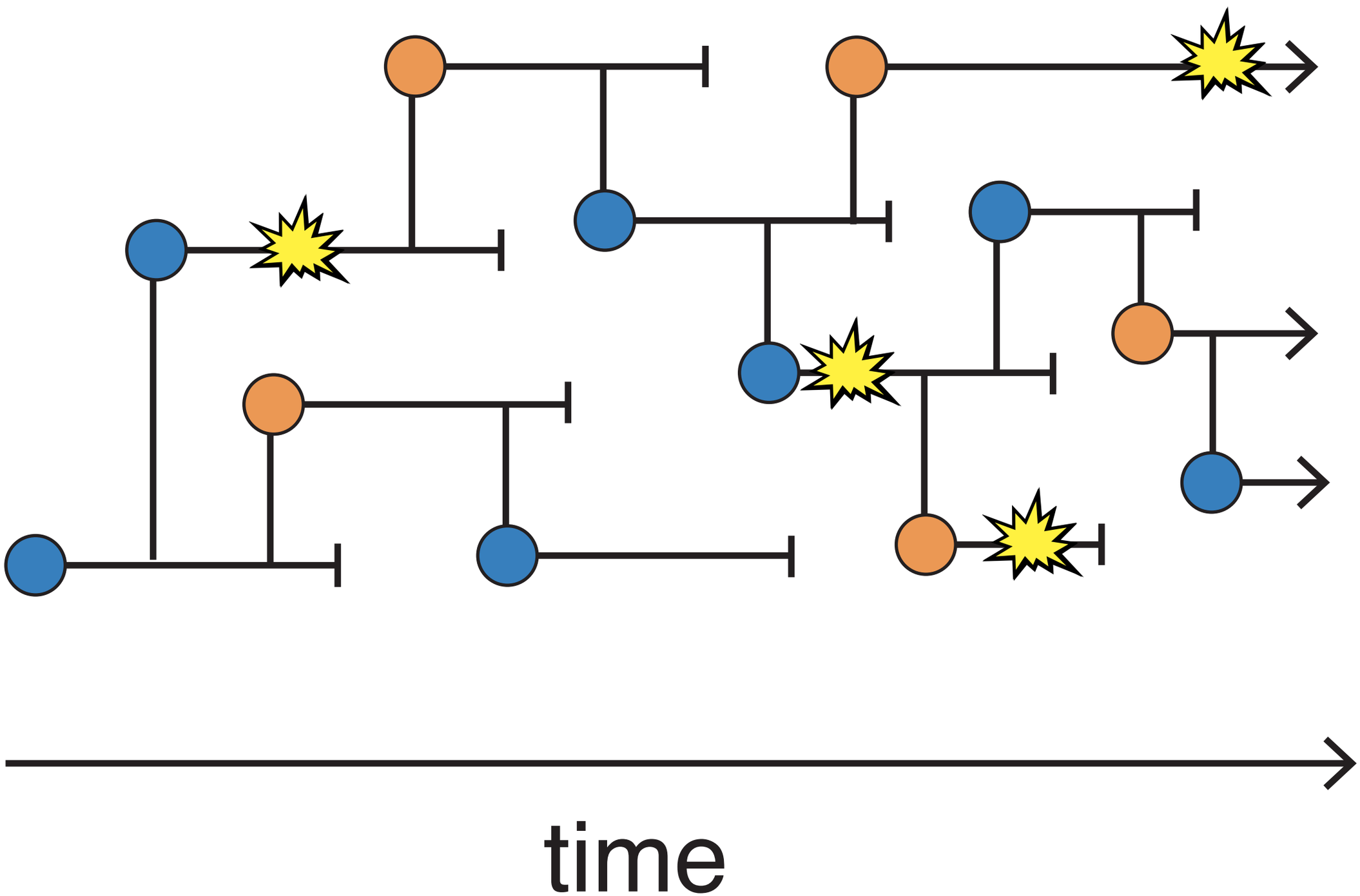
Viral genomes contain records of a virus's transmission history



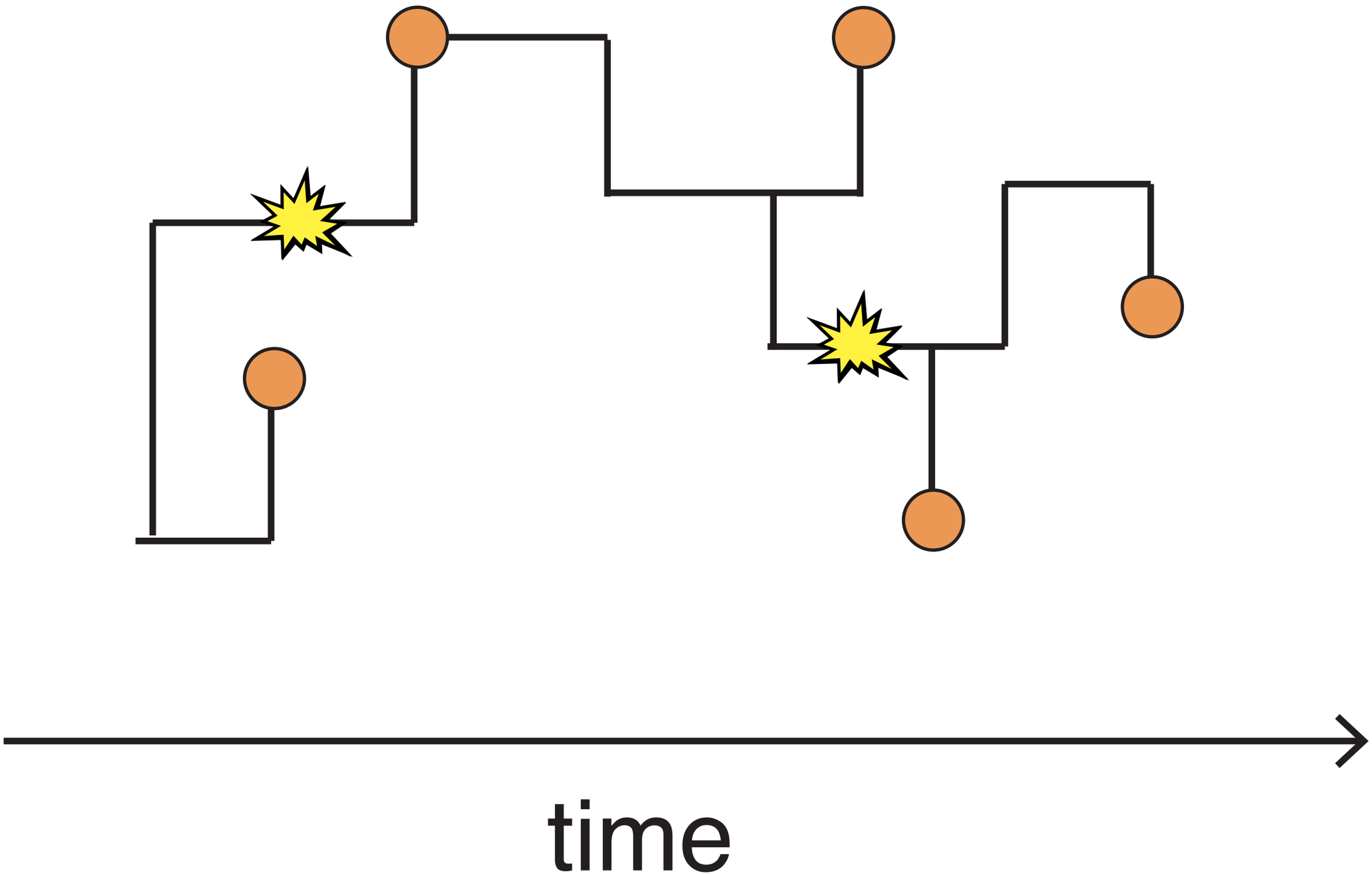
Viral genomes contain records of a virus's transmission history



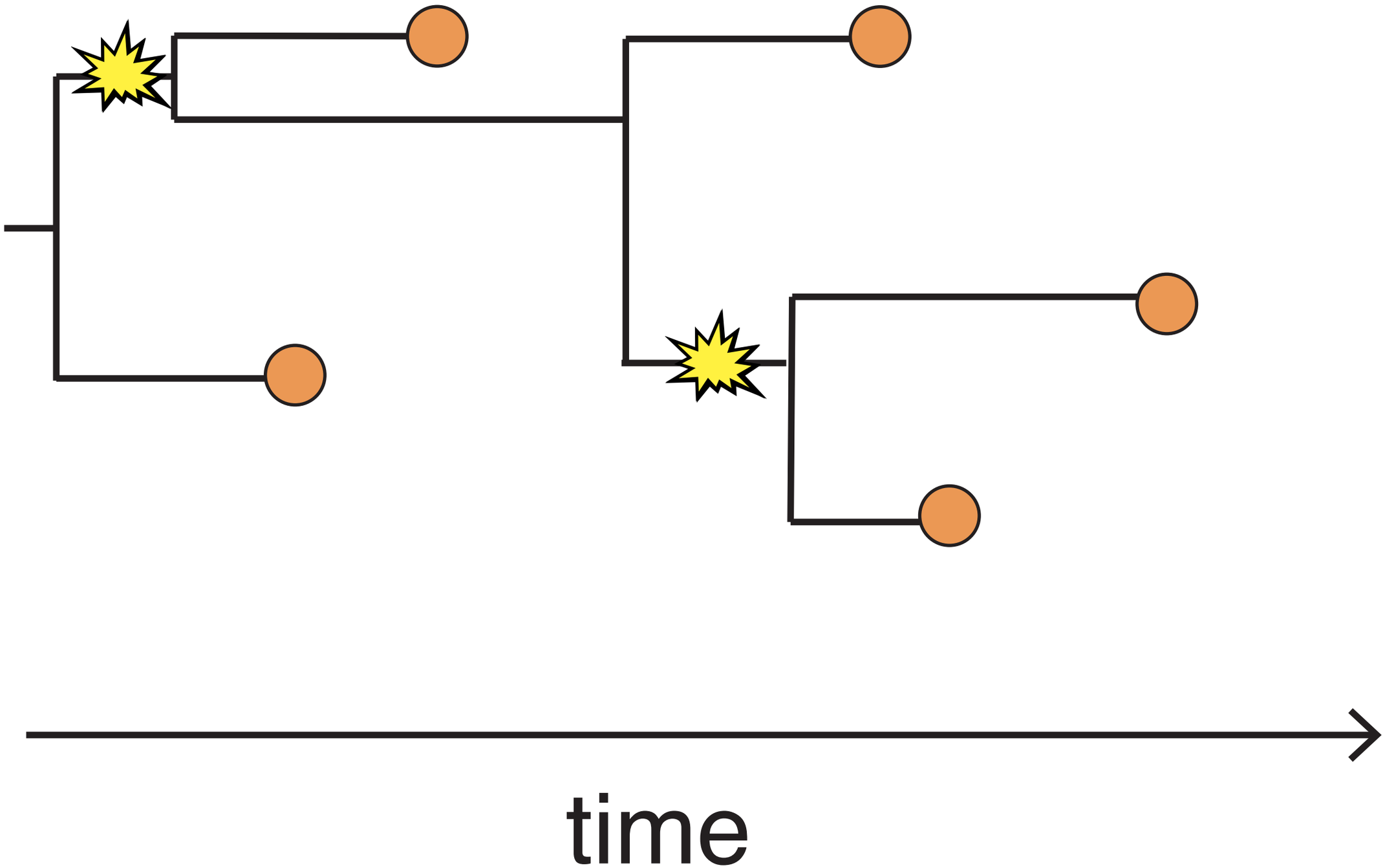
Viral genomes contain records of a virus's transmission history



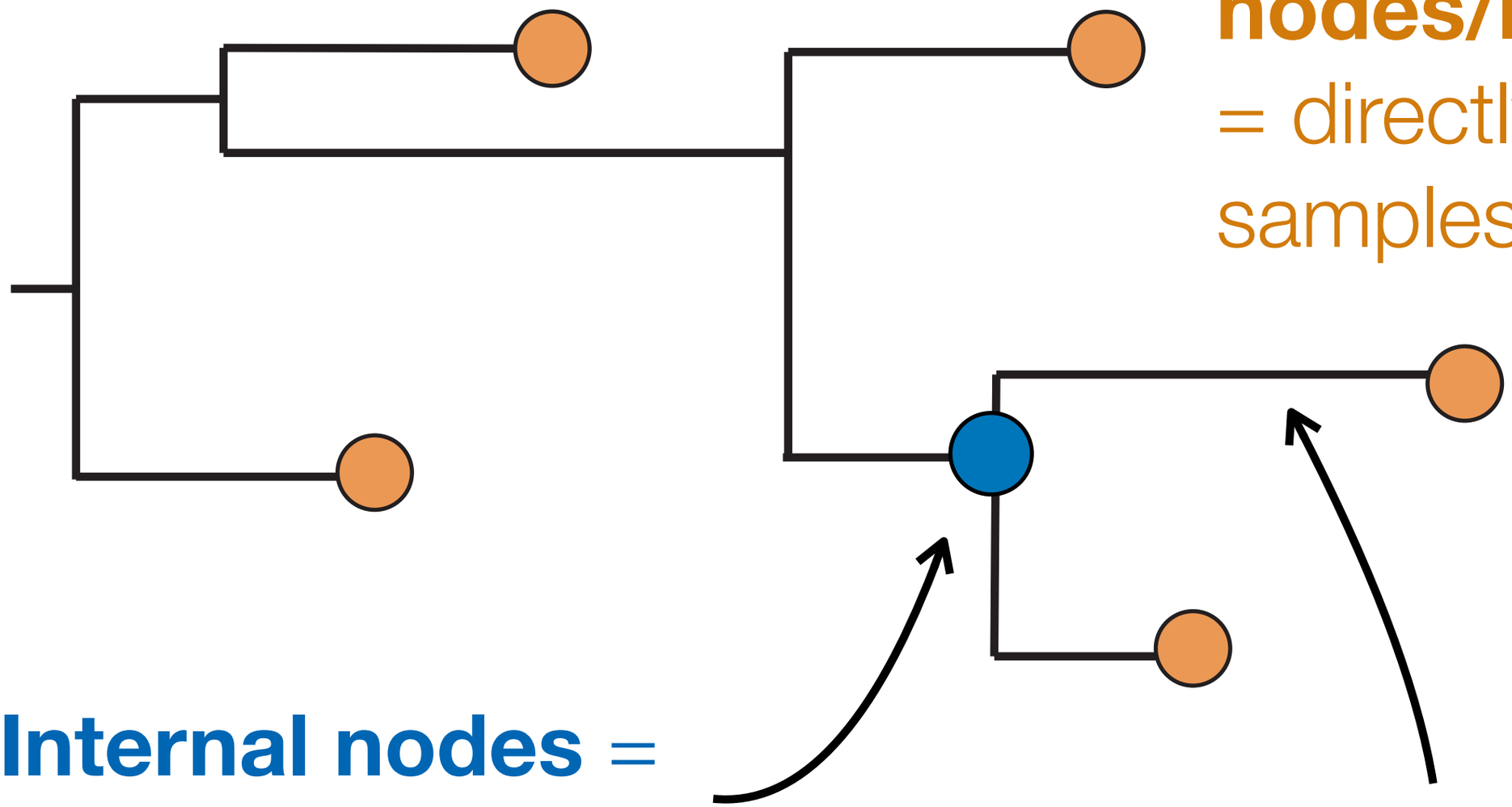
Viral genomes contain records of a virus's transmission history



Viral genomes contain records of a virus's transmission history



A phylogenetic tree is a hypothesis of shared descent

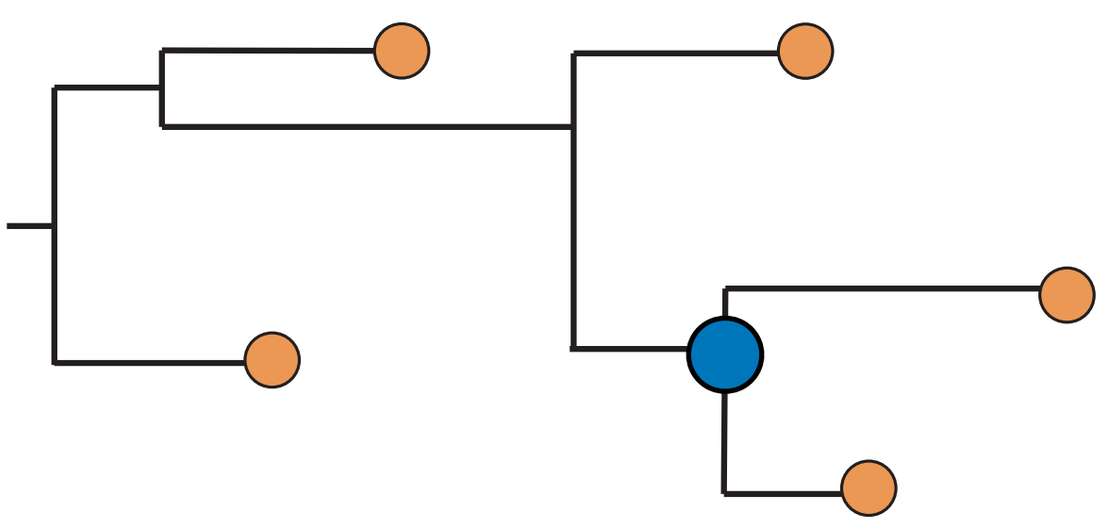


Tips/terminal nodes/leaves/taxa = directly observed samples

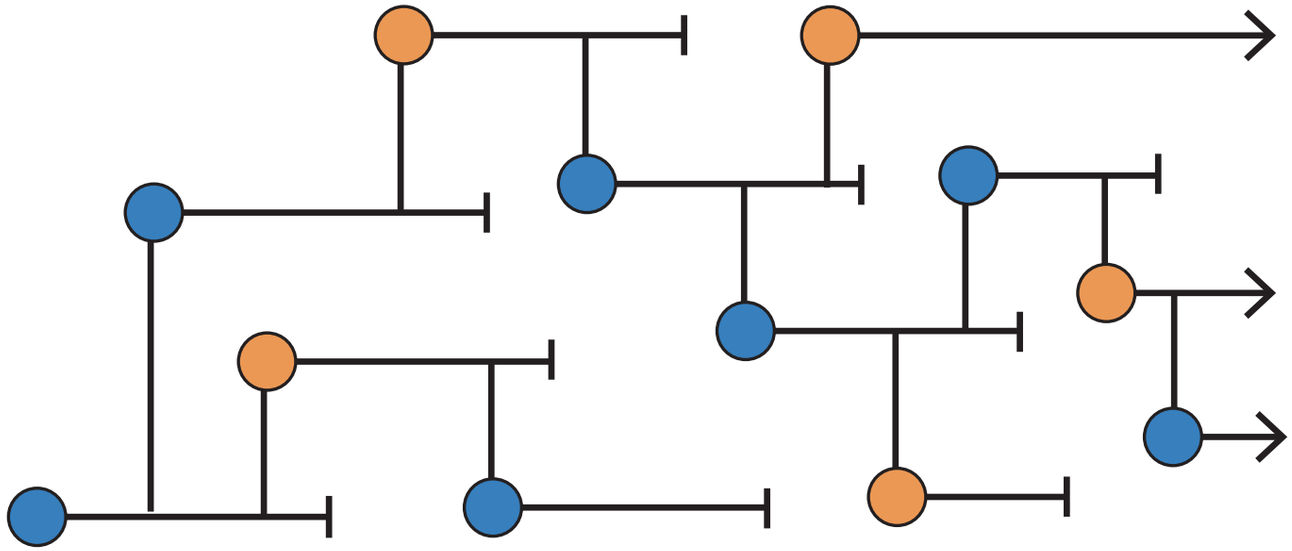
Internal nodes = hypothetical, inferred ancestors

branches = a way to connect tips and nodes; units can be divergence or time

Phylogenetic trees are distinct from transmission trees

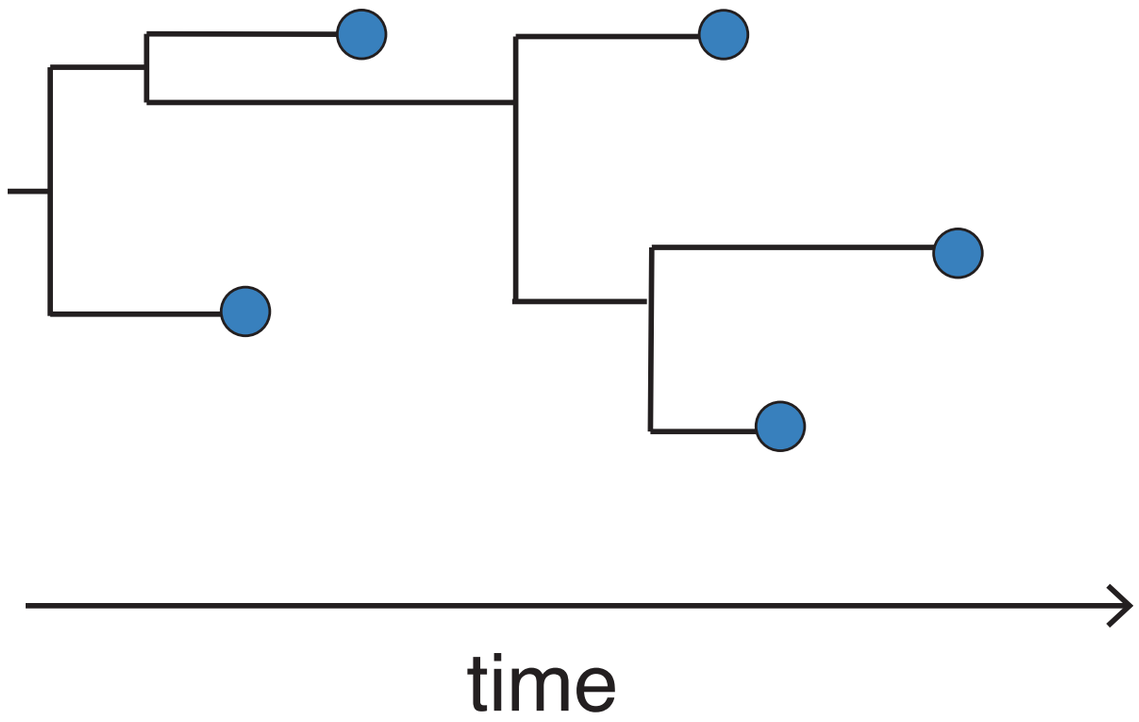


Phylogenetic tree: hypothesis of shared ancestry



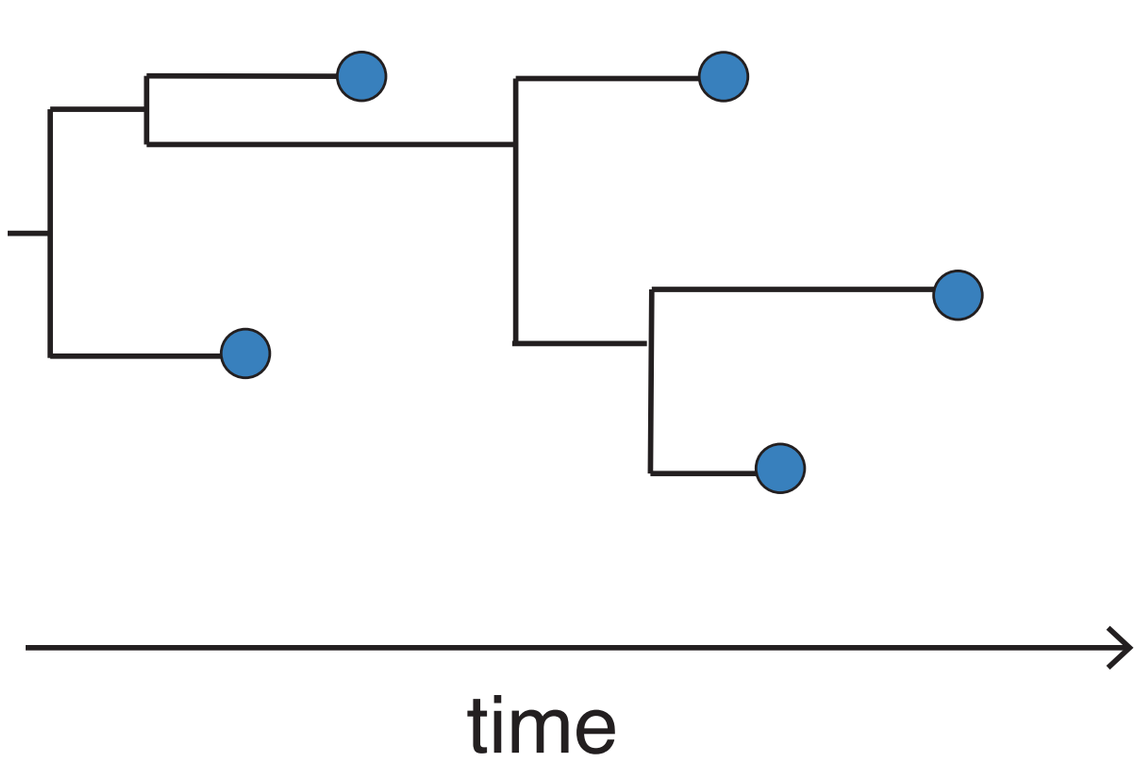
Tranmission tree: true transmission history

The shape of phylogenetic trees changes under different evolutionary, ecological, and epidemiologic scenarios

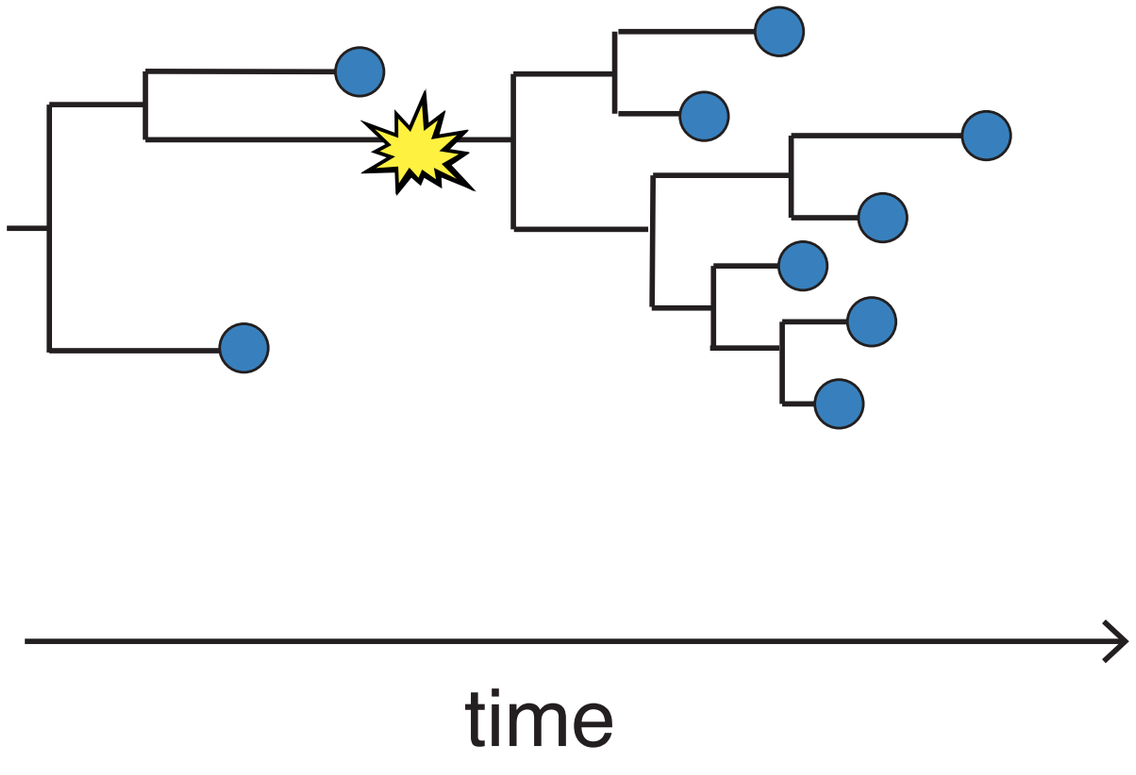


Original tree

The shape of phylogenetic trees changes under different evolutionary, ecological, and epidemiologic scenarios

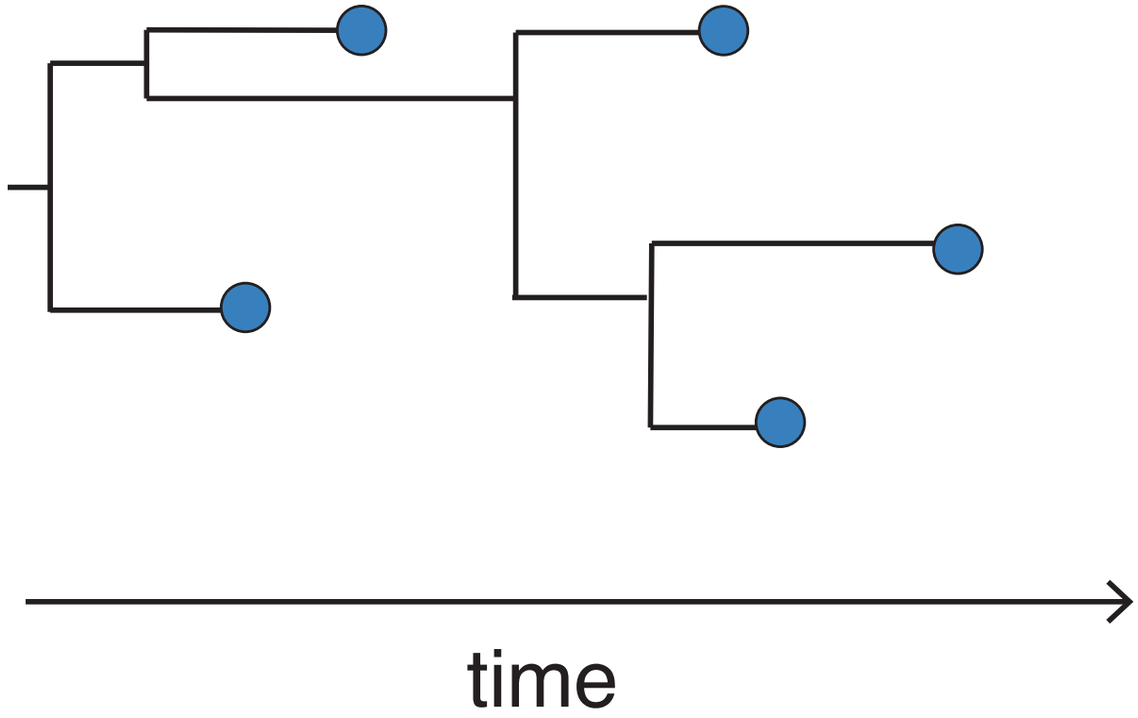


Original tree

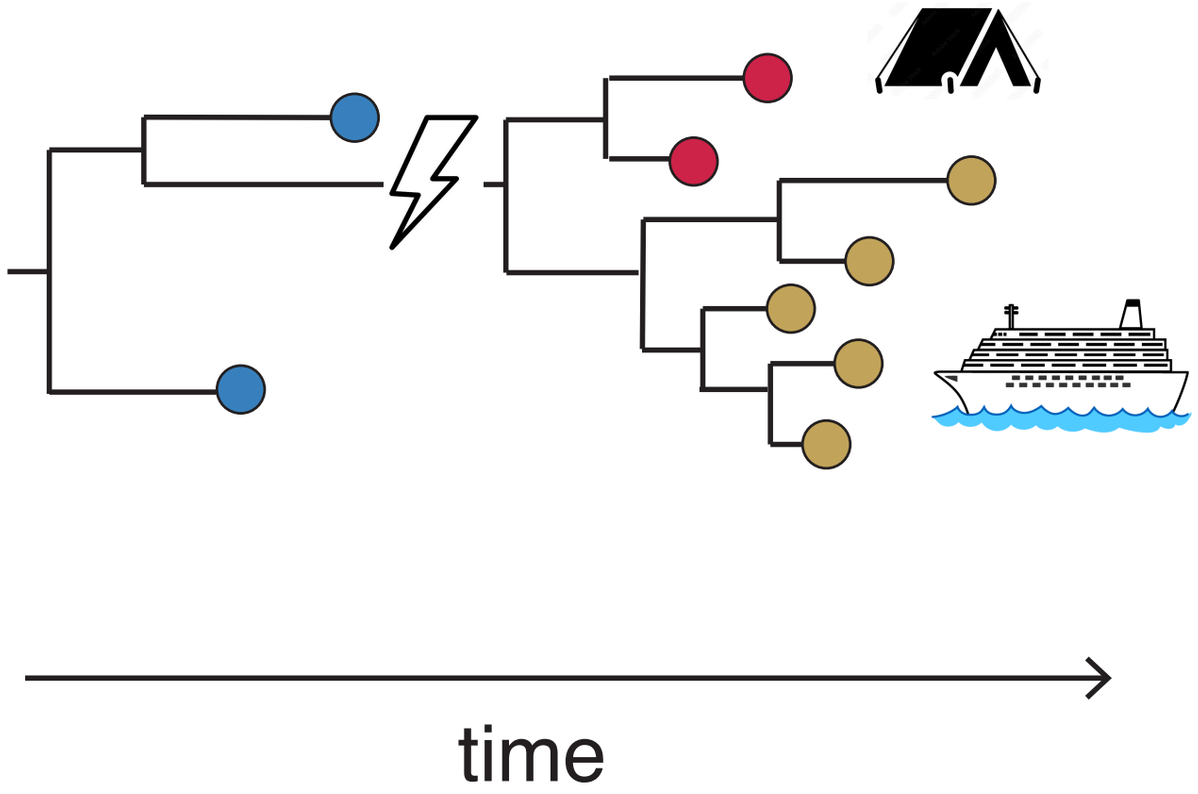


Increased transmissibility

The shape of phylogenetic trees changes under different evolutionary, ecological, and epidemiologic scenarios



Original tree



Population subdivision

… but there's a problem

We don't usually know the evolutionary or transmission history of a pathogen.

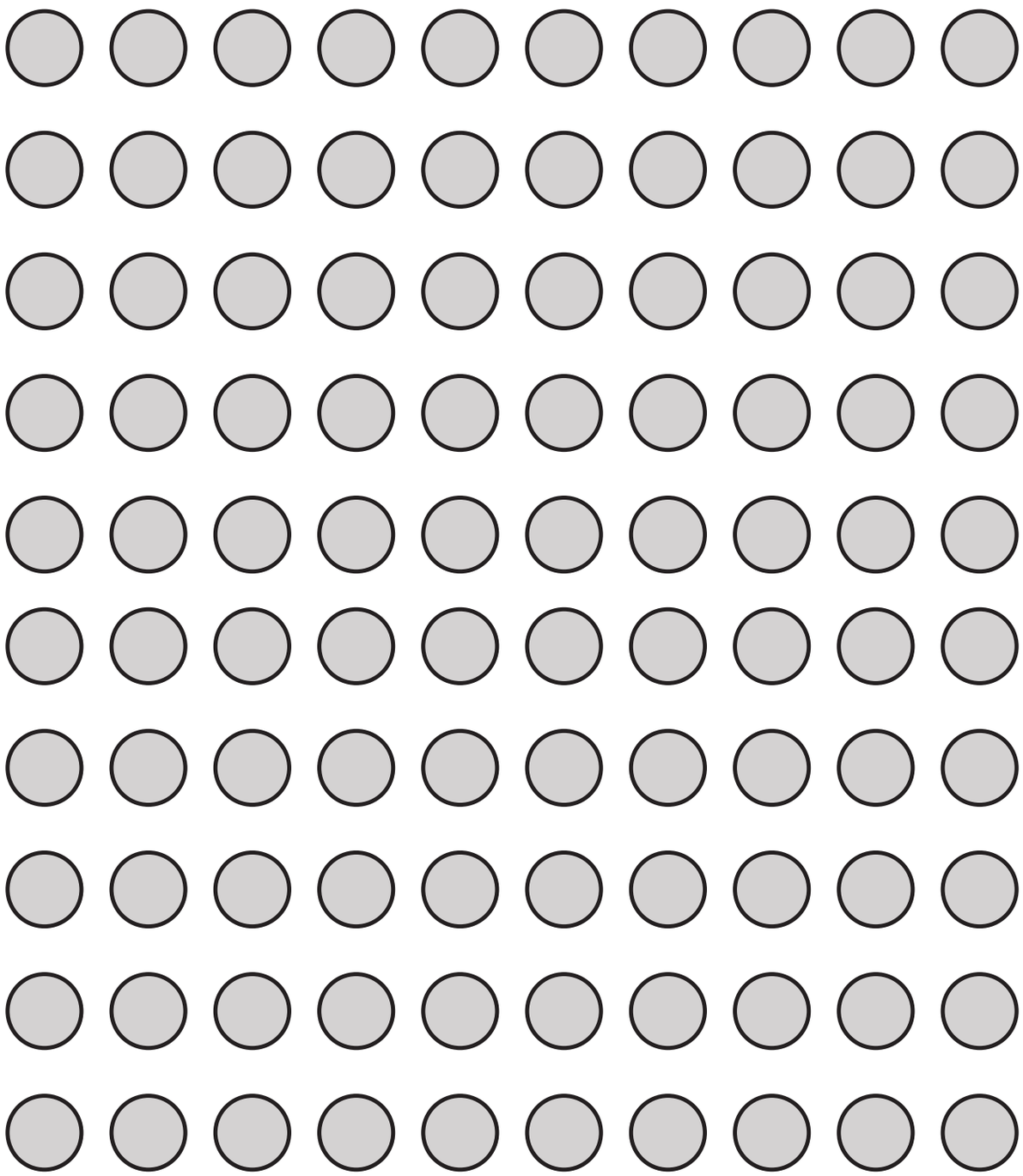
We need a way to quantitatively infer these histories, given a phylogeny

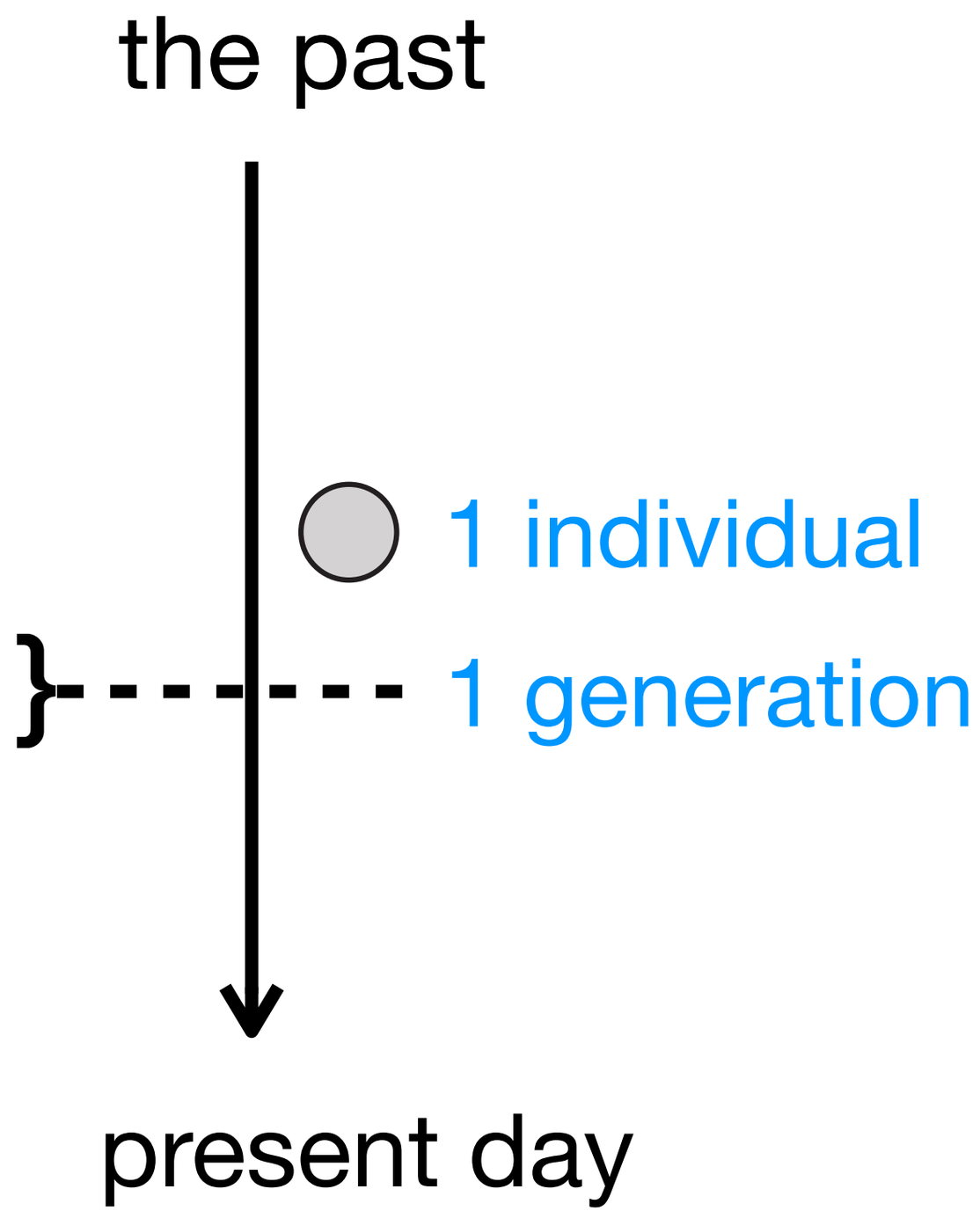

I'm afraid we need to
use...MATH
ALL NEW

The coalescent is a mathematical model that describes the ancestry of a sample of non recombining gene copies.

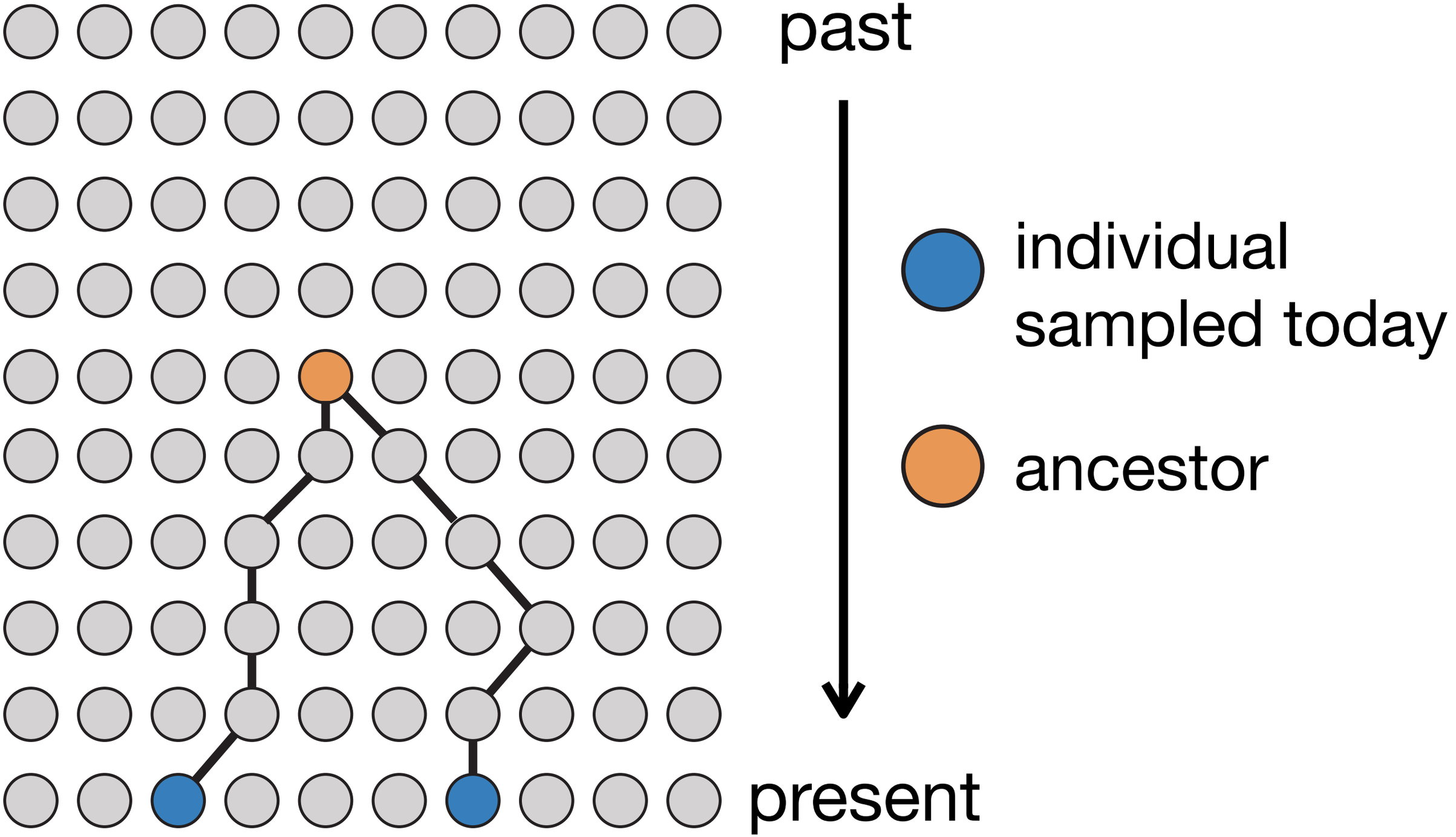
* Volz, Koelle, and Bedford, PLOS Computational Biology, 2013

The coalescent

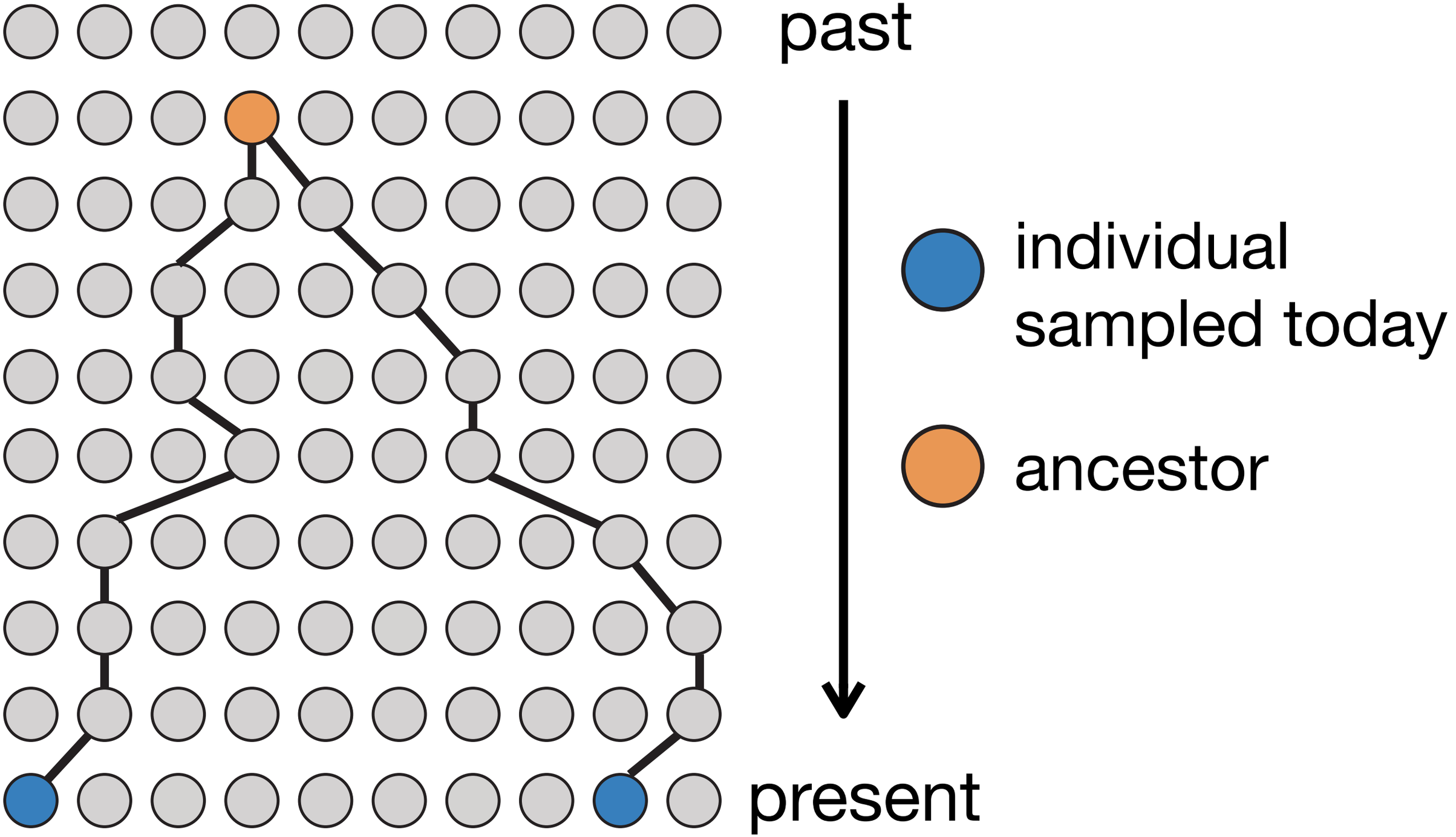




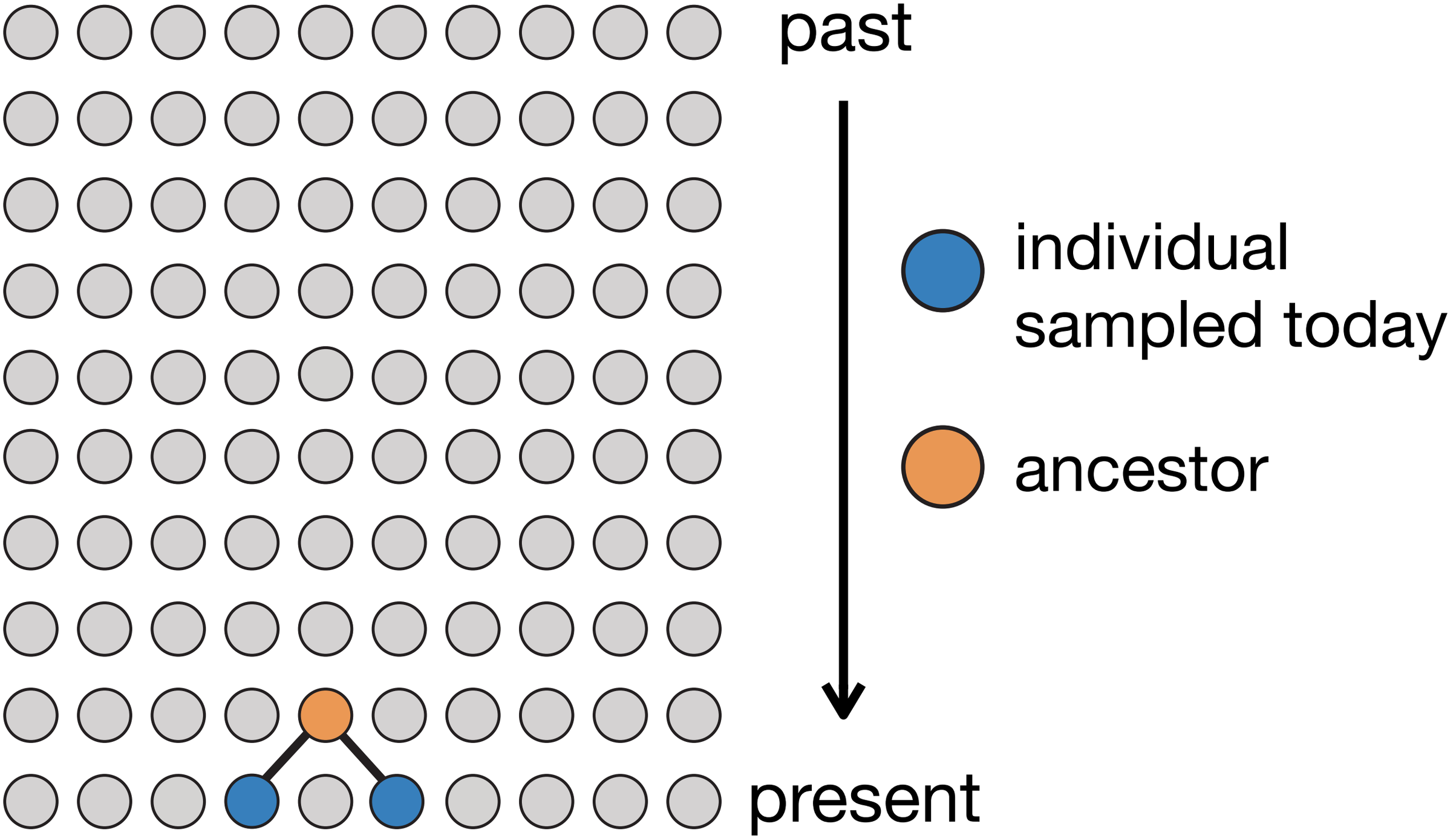
Any 2 individuals sampled today will share a common ancestor in the past



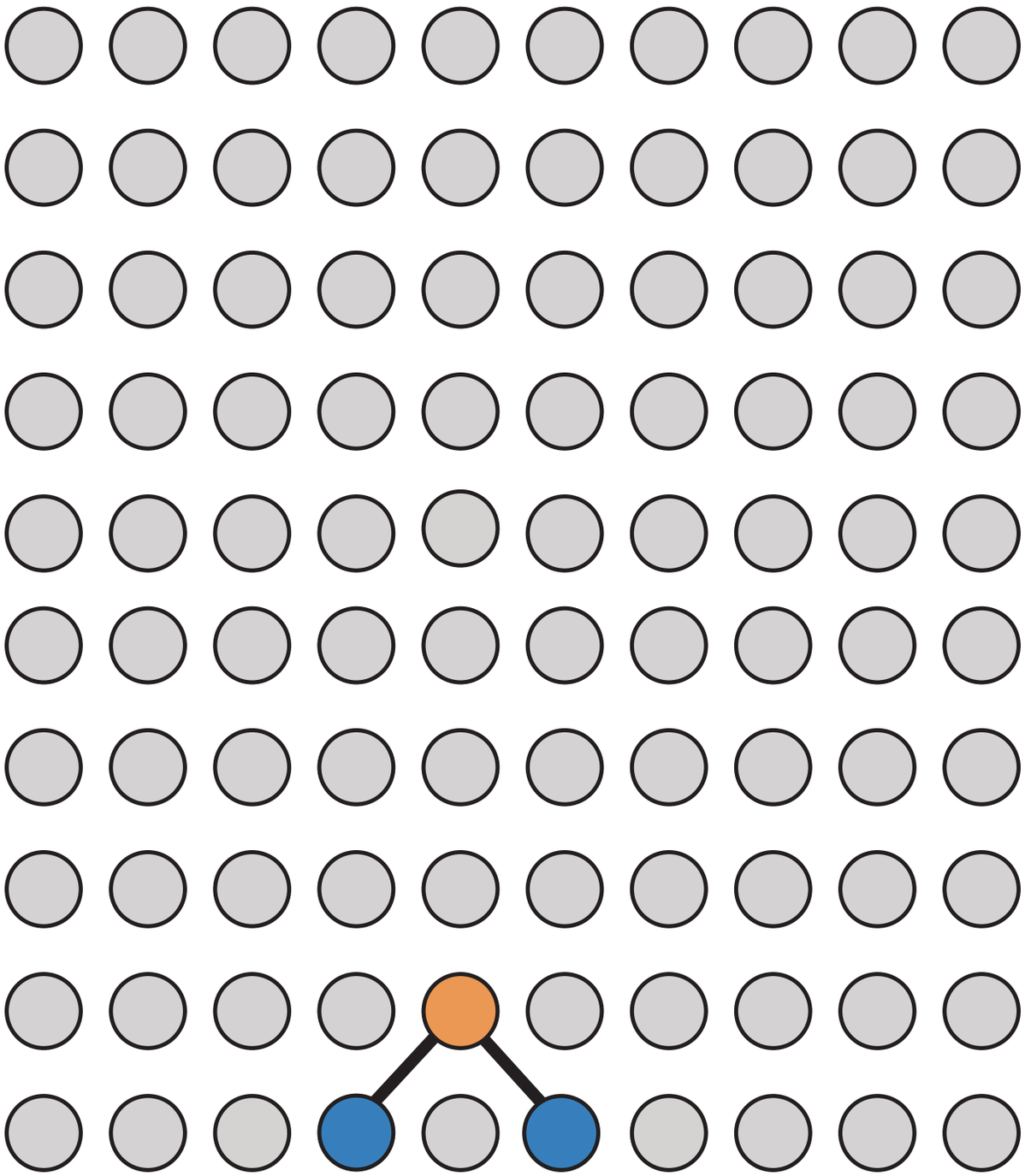
Any 2 individuals sampled today will share a common ancestor in the past



Any 2 individuals sampled today will share a common ancestor in the past



The expected time to coalescence depends on population size



past

present

In any given generation, the probability that 2 individuals coalesce = **1/N** (N = population size)

The expected time to coalescence depends on population size

In any given generation, the **probability** that 2 individuals pick the same parent and coalesce = **1/N**

The expected **waiting time** for 2 lineages to coalesce is geometrically distributed with mean = **N**

* N is in units of generations

How can we scale this to more than 2 lineages?

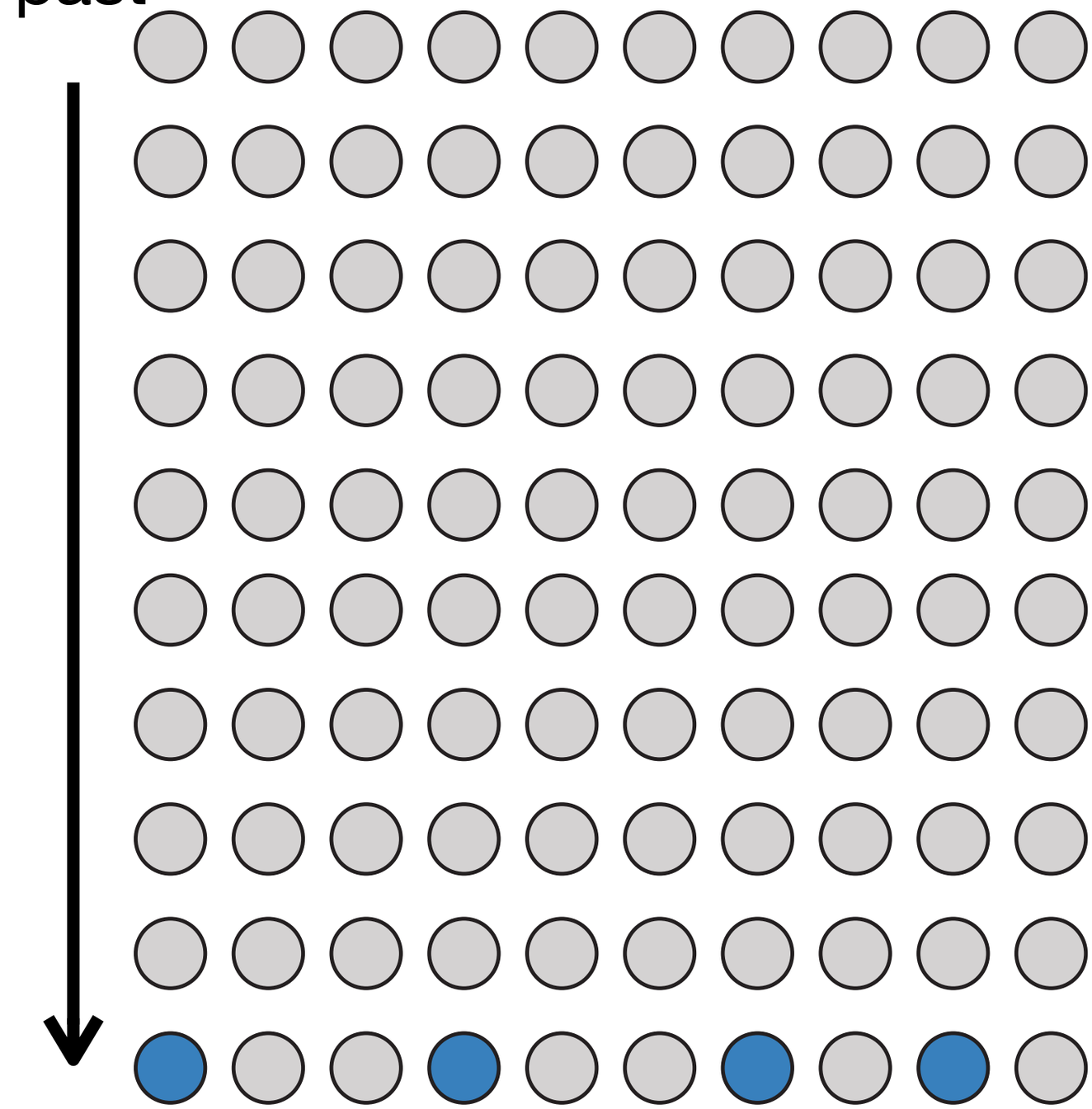
In any given generation, the **probability** that *i* individuals pick the same parent and coalesce:

$$\frac{i(i-1)}{2N}$$

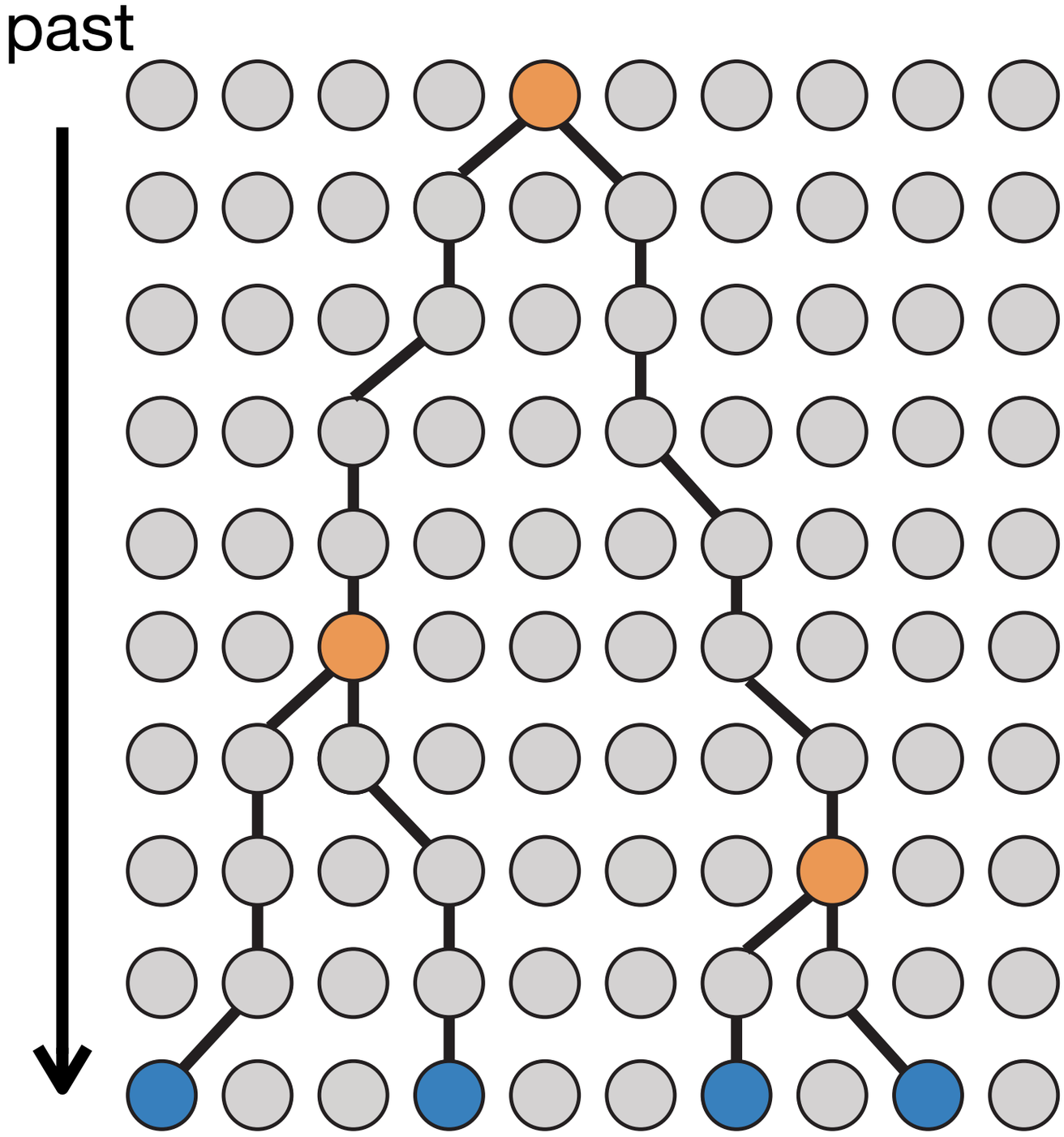
The expected **waiting time** *i* lineages to coalesce is exponentially distributed with a mean of:

$$\frac{2N}{i(i-1)}$$

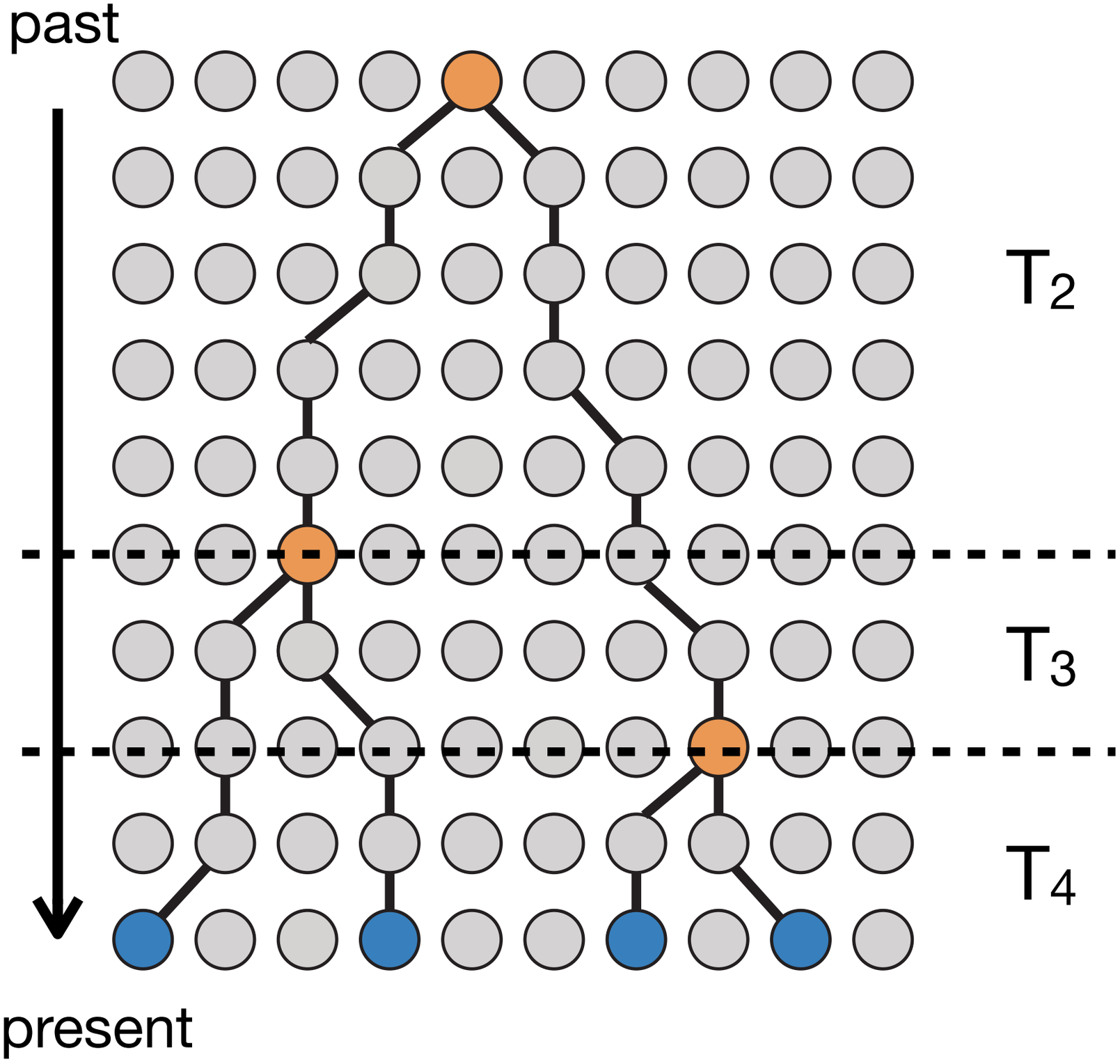
Coalescent times depend on population size and number of “lineages”



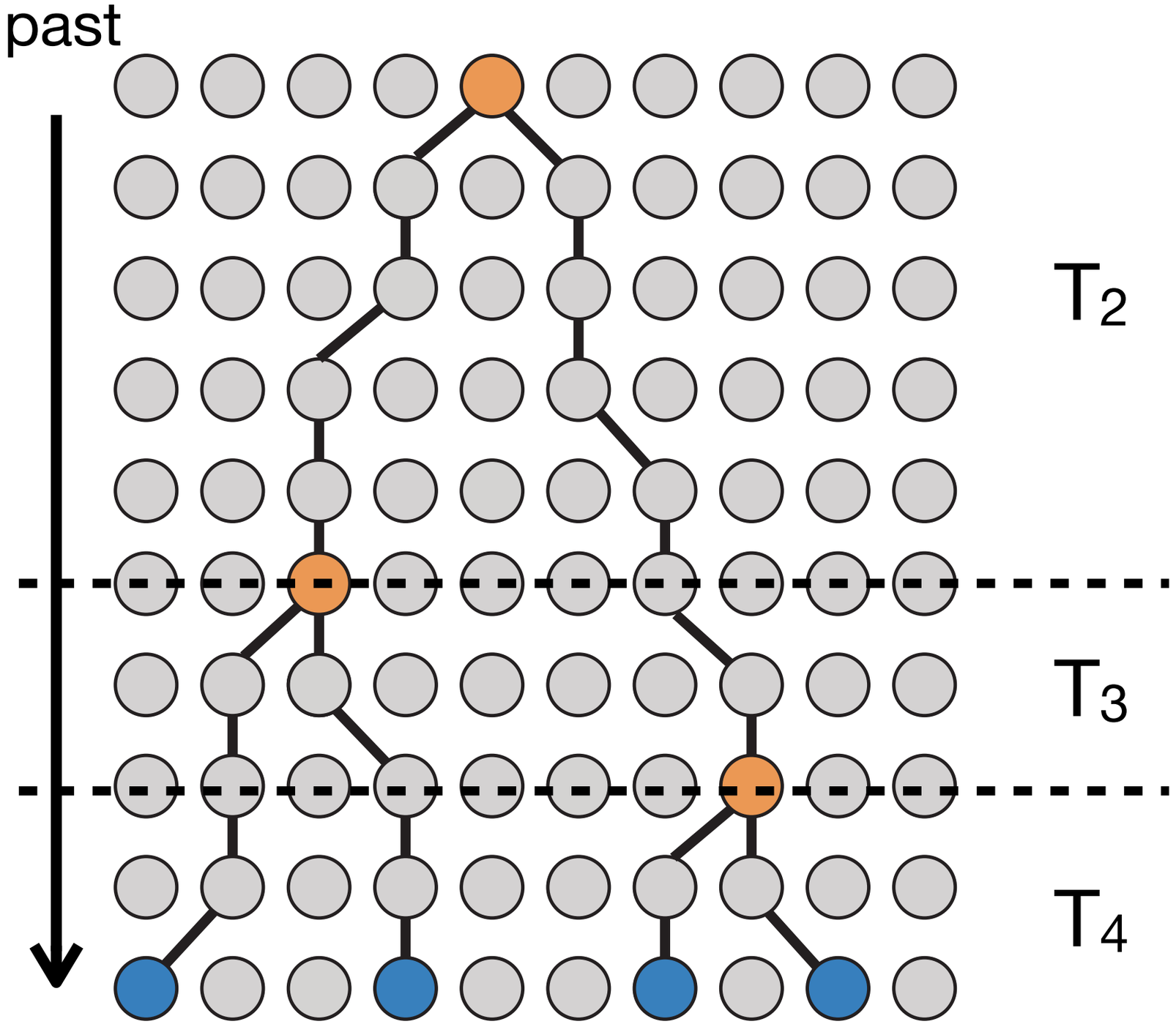
Coalescent times depend on population size and number of “lineages”



Coalescent times depend on population size and number of “lineages”



Coalescent times depend on population size and number of “lineages”



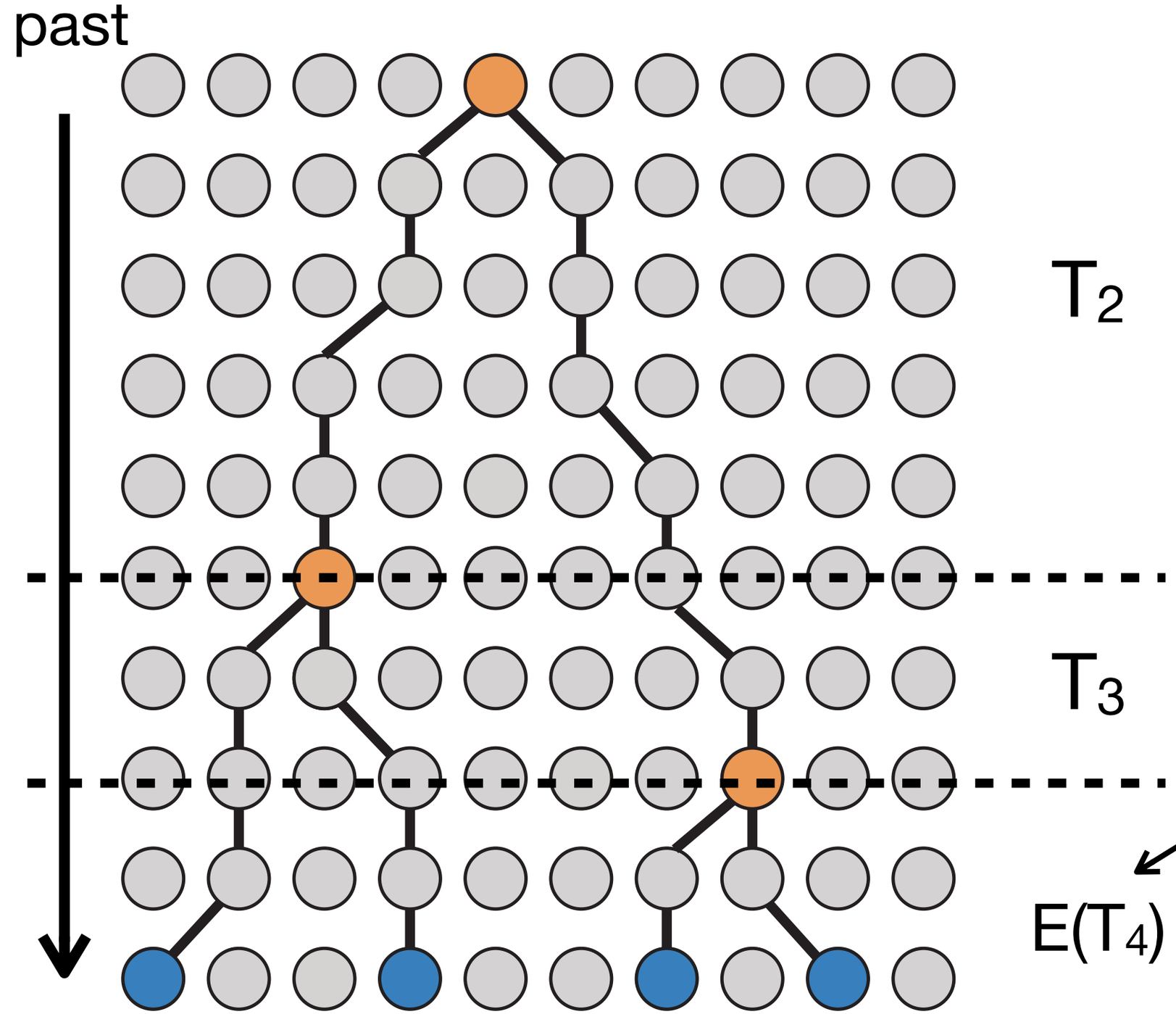
Expected waiting time:

$$\frac{2N}{i(i-1)}$$

i = # of lineages

N = population size, scaled by # of generations

Coalescent times depend on population size and number of "lineages"

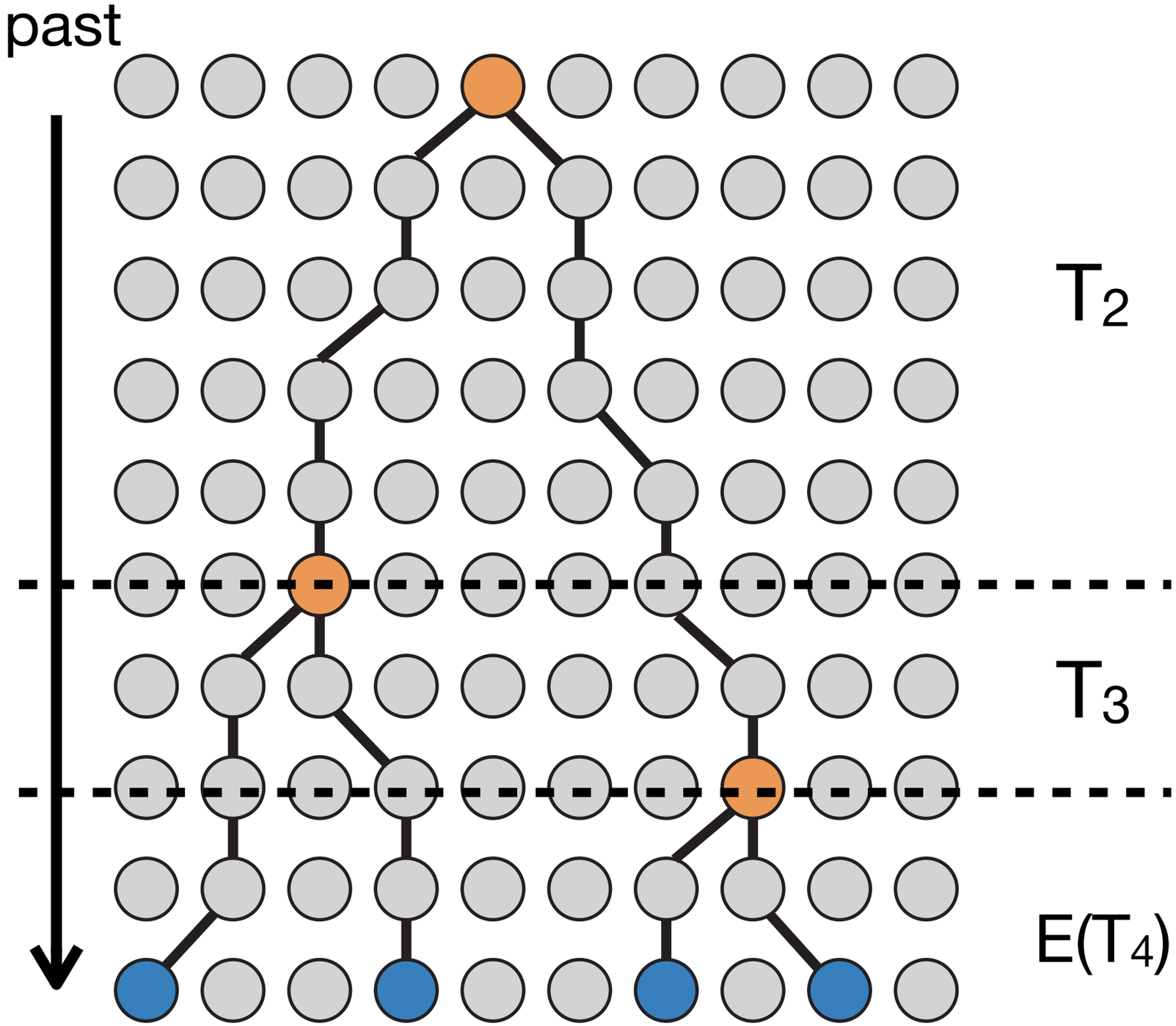


Expected waiting time:

$$\frac{2N}{i(i-1)}$$

"E" here represents an expectation/ expected time

Coalescent times depend on population size and number of “lineages”

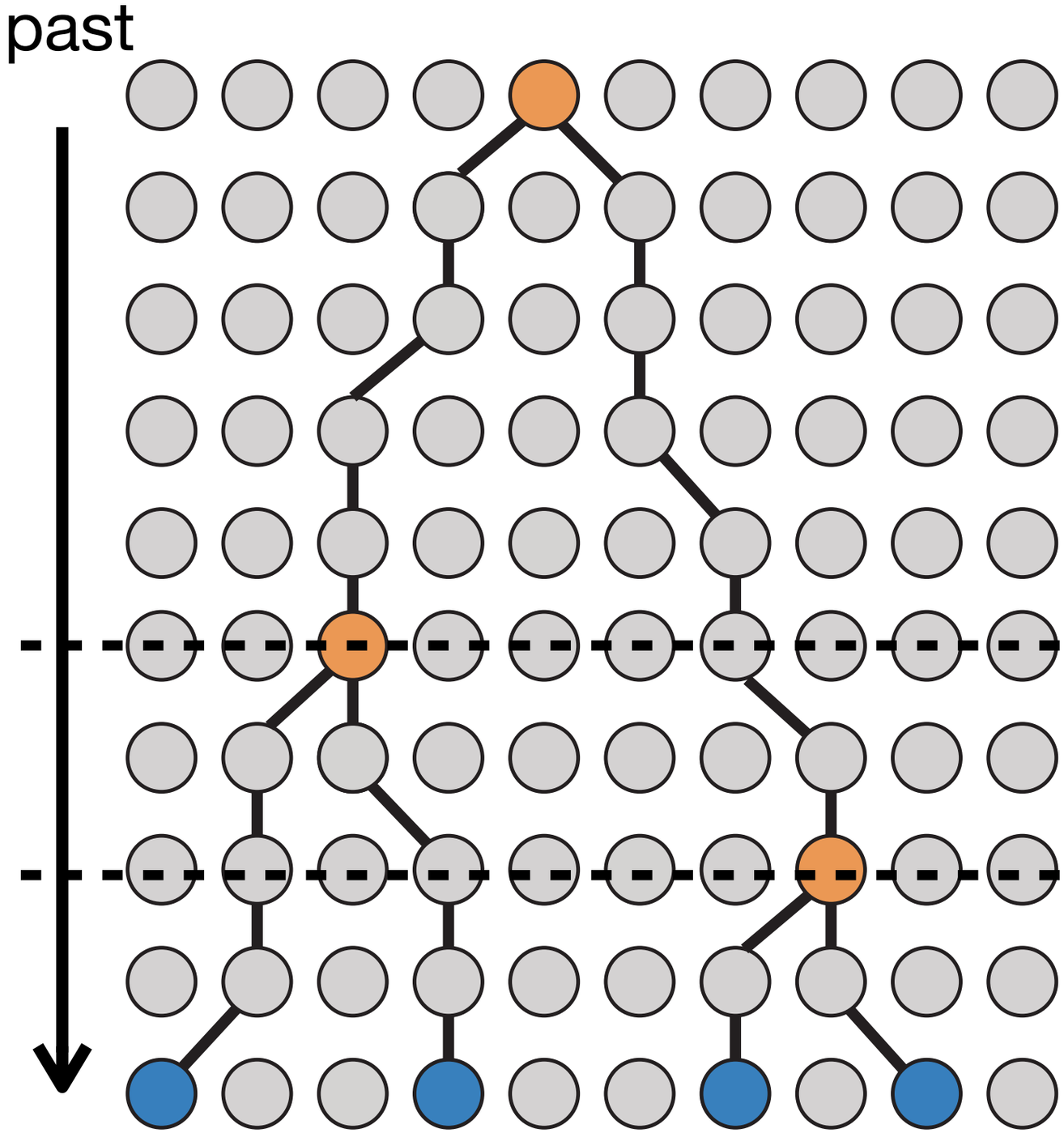


Expected waiting time:

$$\frac{2N}{i(i-1)}$$

$$E(T_4) = \frac{2N}{4(4-1)} = \frac{N}{6} = \frac{10}{6}$$

Coalescent times depend on population size and number of “lineages”



T_2

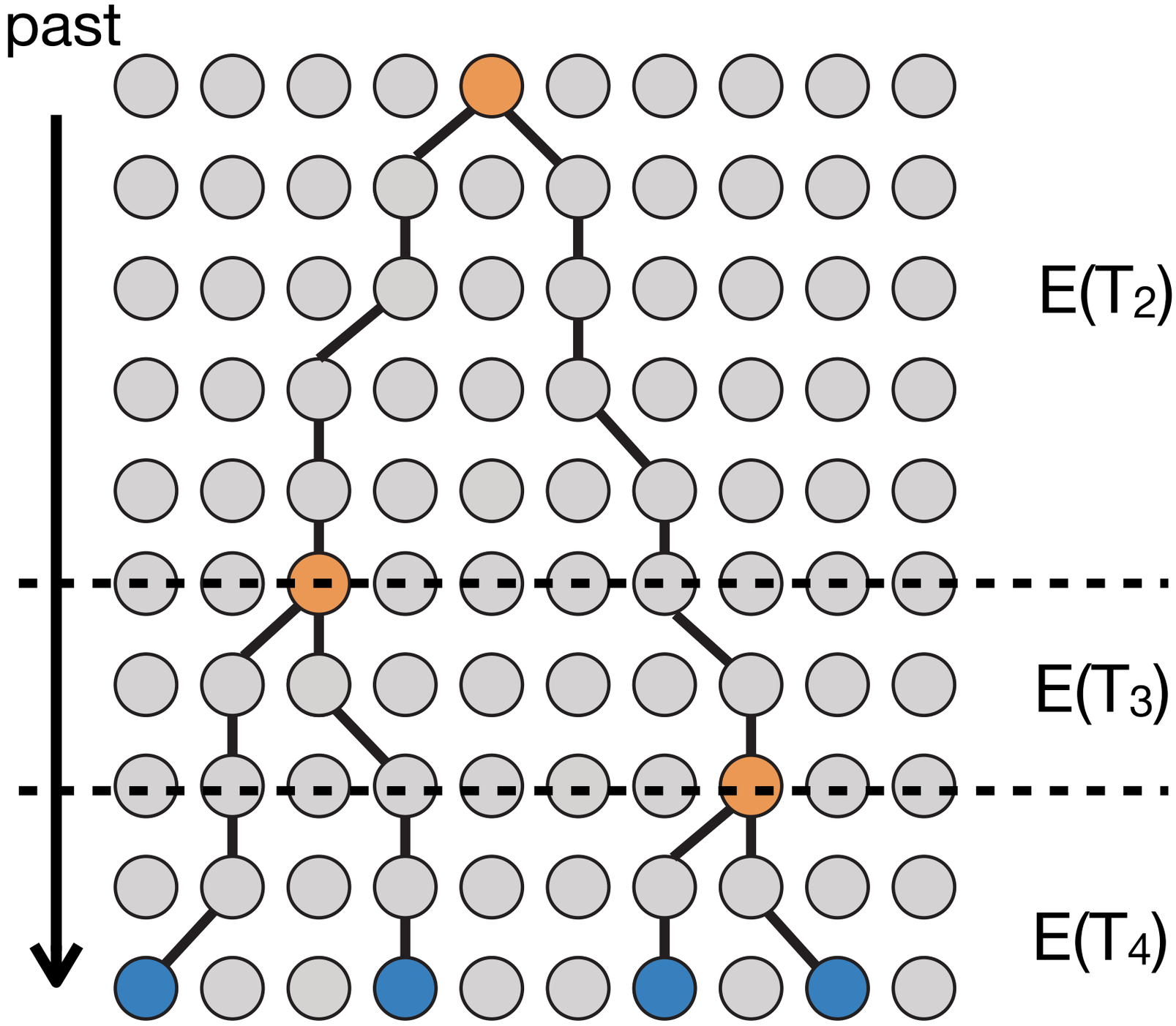
Expected waiting time:

$$\frac{2N}{i(i-1)}$$

$$E(T_3) = \frac{2N}{3(3-1)} = \frac{N}{3} = \frac{10}{3}$$

$$E(T_4) = \frac{2N}{4(4-1)} = \frac{N}{6} = \frac{10}{6}$$

Coalescent times depend on population size and number of “lineages”

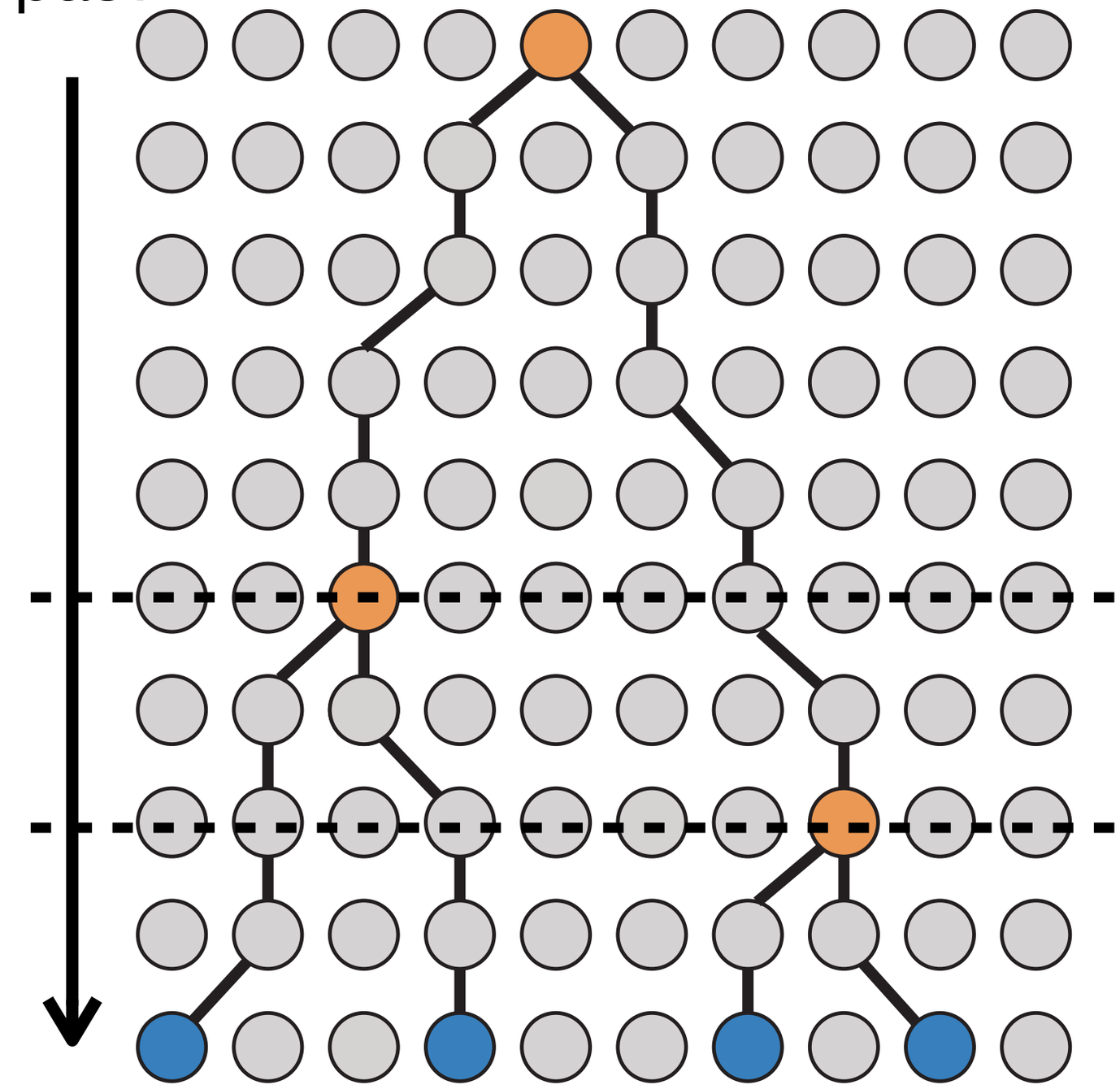


$$E(T_2) = \frac{2N}{2(2-1)} = N = 10$$

$$E(T_3) = \frac{2N}{3(3-1)} = \frac{N}{3} = \frac{10}{3}$$

$$E(T_4) = \frac{2N}{4(4-1)} = \frac{N}{6} = \frac{10}{6}$$

The expected **waiting time** to go from *i* lineages to 1 lineage (TMRCA) is the sum overall all coalescent intervals



$$E(TMRCA) =$$

$$E(T_4) + E(T_3) + E(T_2) =$$

$$\frac{N}{6} + \frac{N}{3} + N =$$

$$\frac{3N}{2} = \frac{30}{2}$$

= 15 generations

The expected **waiting time** to go from *i* lineages to 1 lineage is the sum overall all coalescent intervals:

$$T_{\mathrm{MRCA}} = \sum_{i=2}^{n} T_i \qquad \mathrm{E}[T_i] = \frac{2N}{i(i-1)}$$

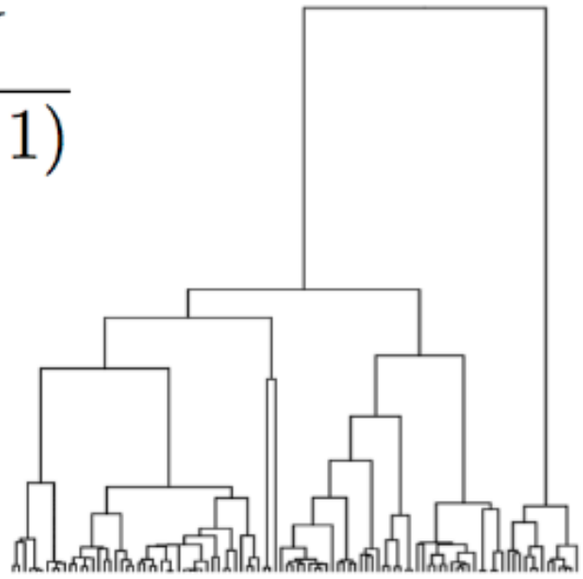
$$\mathrm{E}[T_{\mathrm{MRCA}}] = \sum_{i-2}^{n} \frac{2N}{i(i-1)}$$

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

$$= 2N \left(1 - \frac{1}{2} + \frac{1}{2} - \cdots - \frac{1}{n-1} + \frac{1}{n-1} - \frac{1}{n} \right)$$

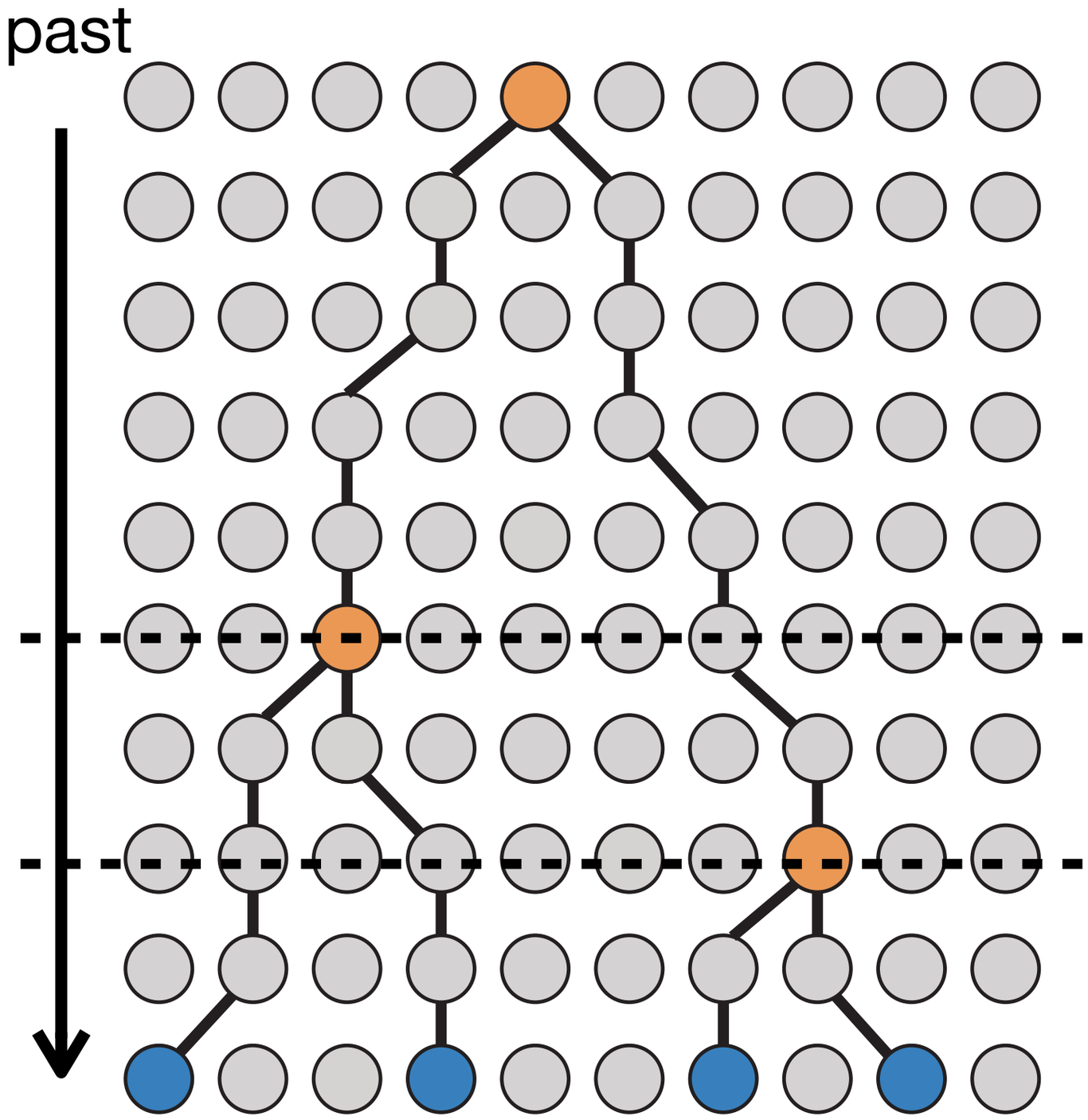
$$= 2N \left(1 - \frac{1}{n} \right)$$

$$\mathrm{E}[T_{\mathrm{MRCA}}] = \lim_{n \to \infty} 2N$$



* Bedford SISMID lecture, Coalescent

The expected TMRCA approaches 2N



N (population size) = 10

n (sampled individuals) = 4

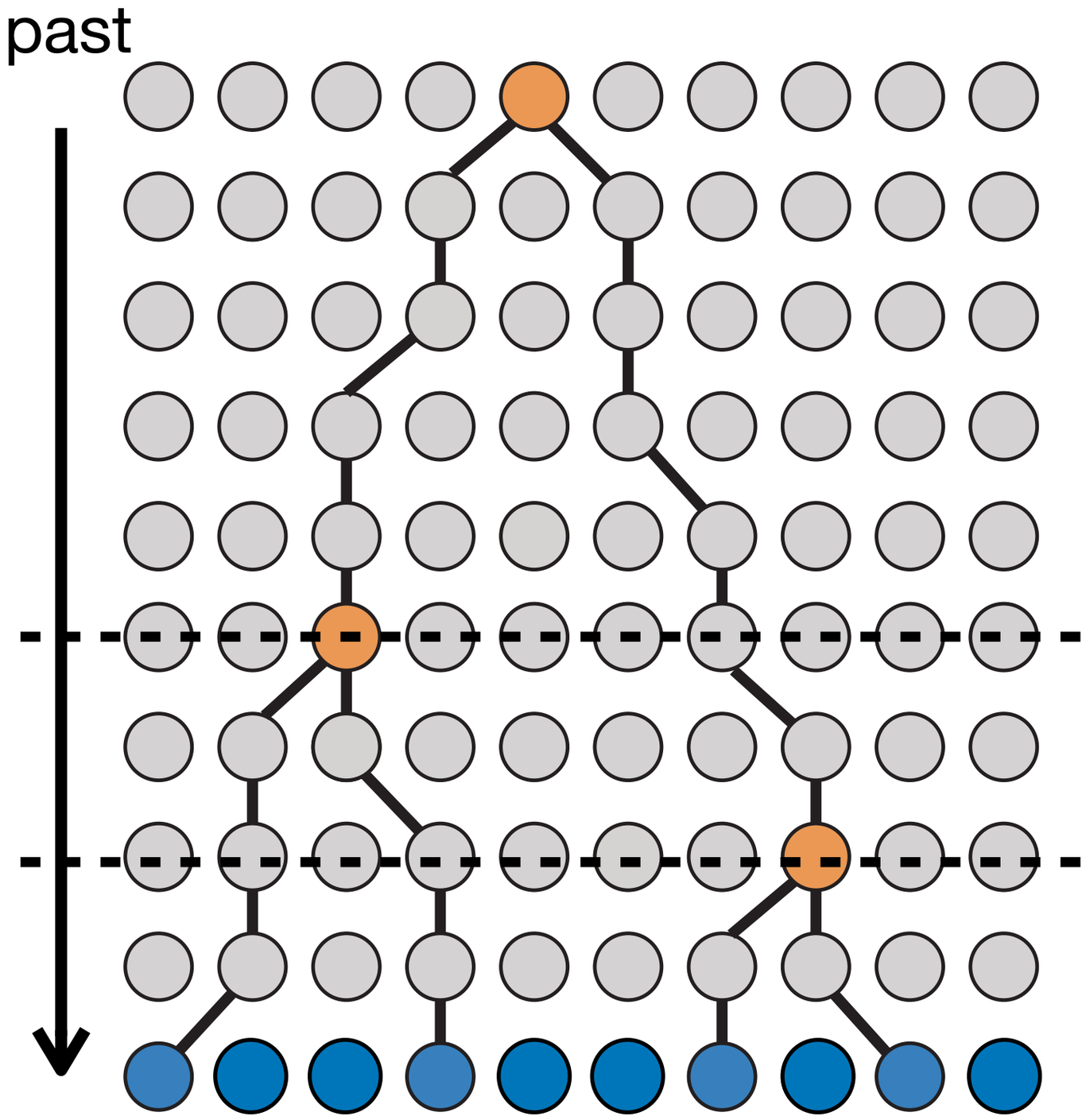
$$E(TMRCA) = 2N\left(1 - \frac{1}{n}\right)$$

$$= 2*10\left(1 - \frac{1}{4}\right)$$

$$= 20(0.75)$$

$$= \mathbf{15\ generations}$$

The expected TMRCA approaches 2N



N (population size) = 10

n (sampled individuals) = 10

$$E(TMRCA) = 2N\left(1 - \frac{1}{n}\right)$$

$$= 2*10\left(1 - \frac{1}{10}\right)$$

$$= 20(0.90)$$

$$= \mathbf{18\ generations}$$

A few key points about coalescent trees:

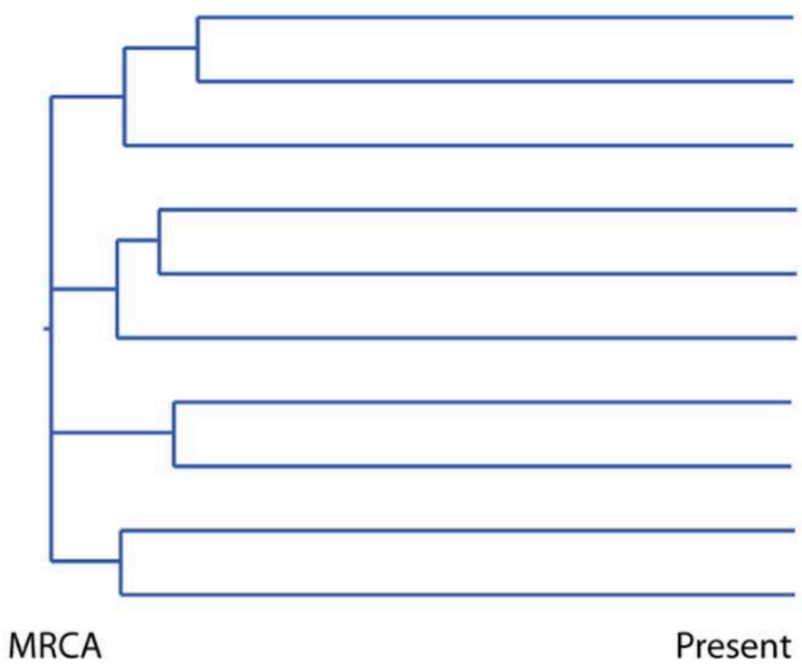
1. Coalescence takes longer in large populations
2. Coalescence takes longer when there are fewer lineages present
3. Most of the tree height is contributed by only 2 lineages
4. The TMRCA for a set of sampled lineages is ~2N
5. The coalescent gives us a framework for setting an expectation for genealogies under different evolutionary scenarios

The coalescent is conceived under a Wright-Fisher model, which makes a lot of assumptions

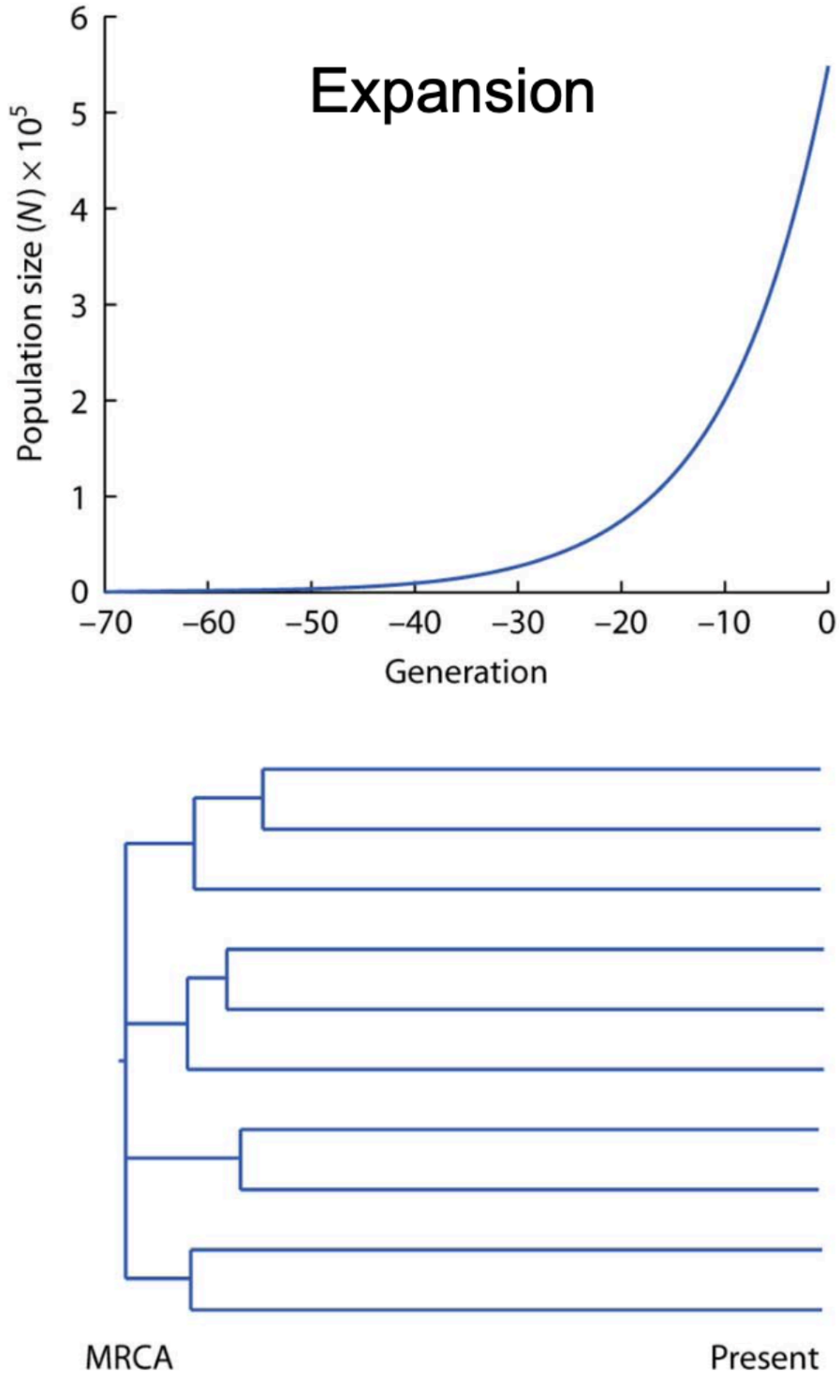
- No selection, migration, or recombination
- Random mating
- Population size is constant over time
- Sampled lineages << N; maximum of 1 coalescent event per generation

We can use our knowledge of expected coalescence times to find evidence of deviation from these assumptions!

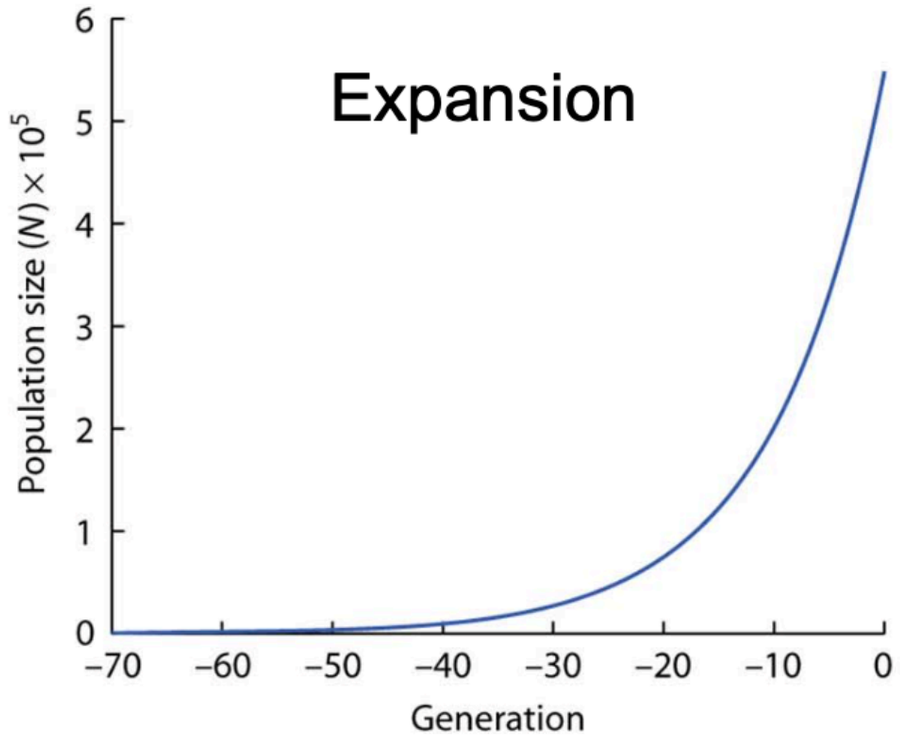
Coalescent trees look very different in large and small populations

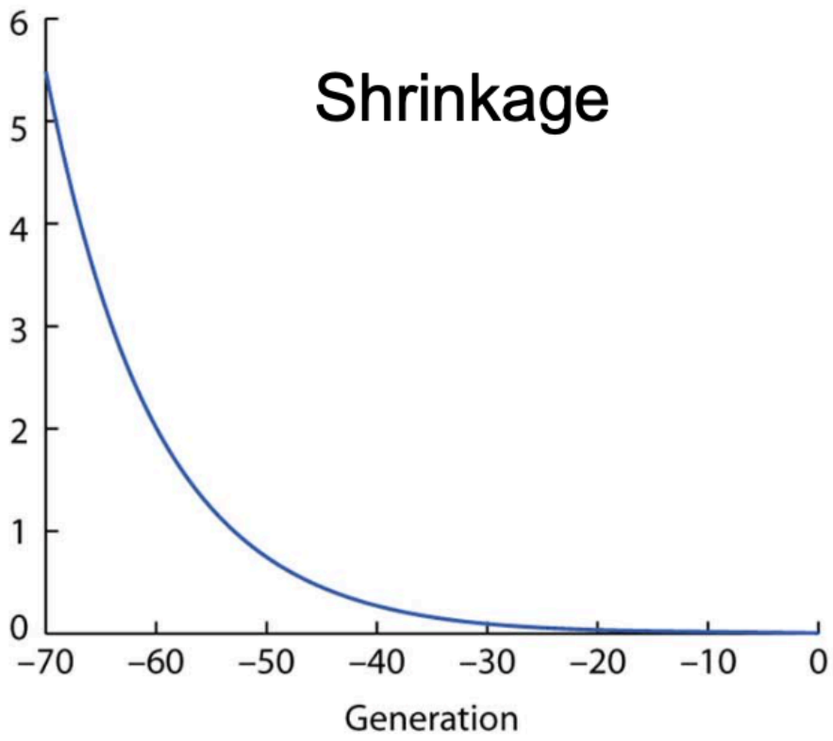


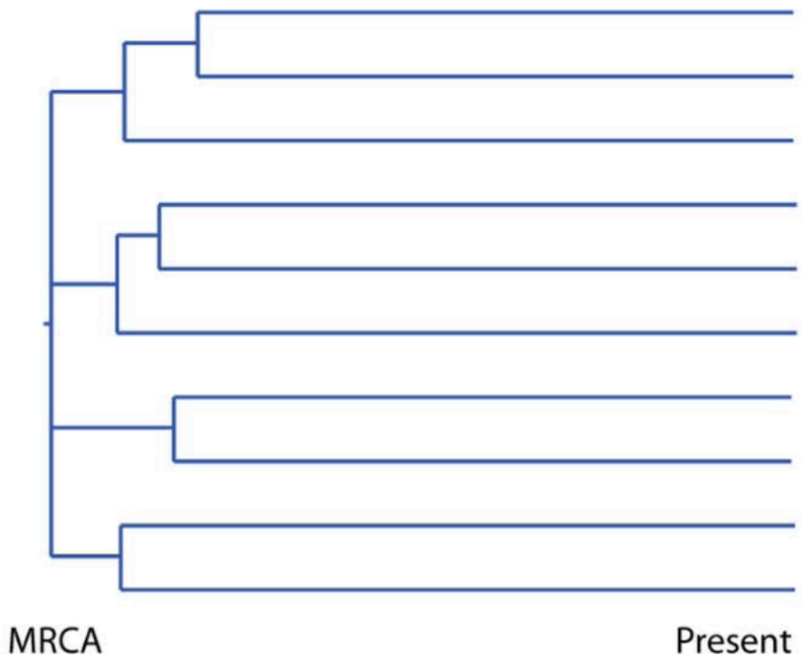
Coalescent trees look very different in large and small populations

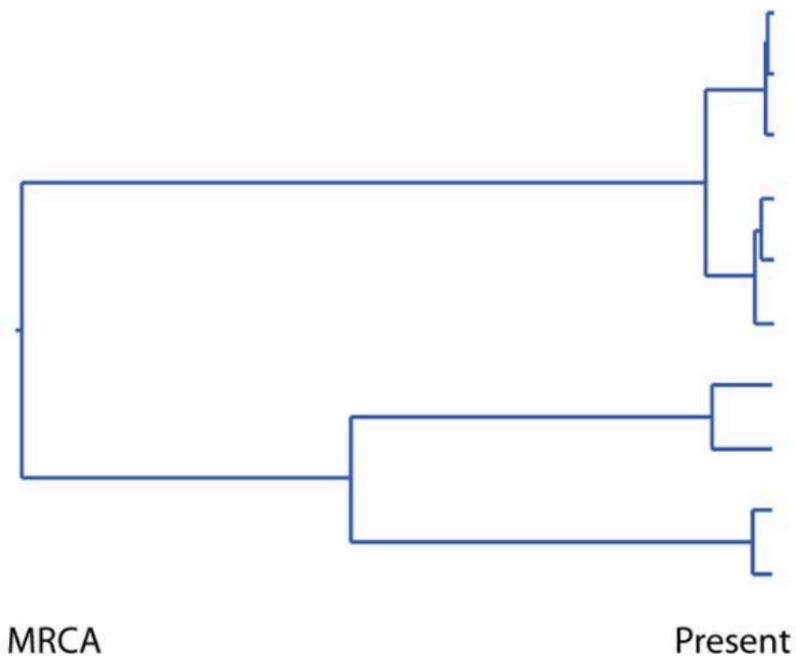


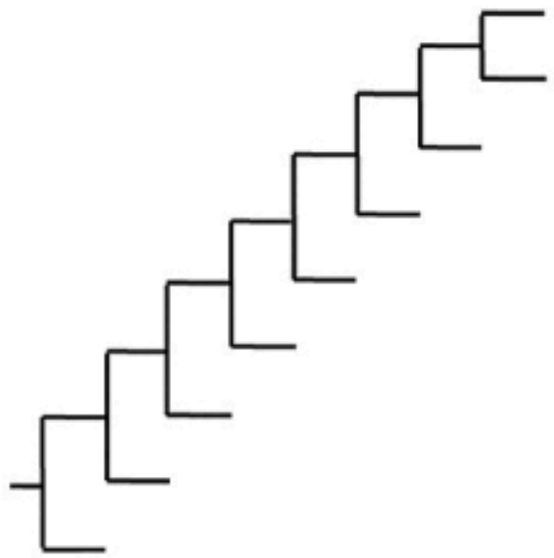
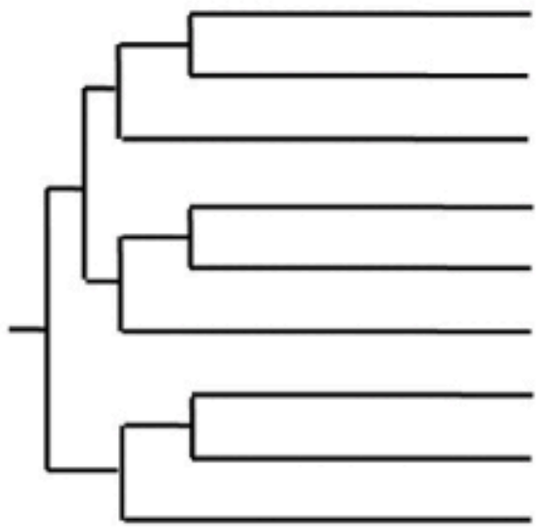
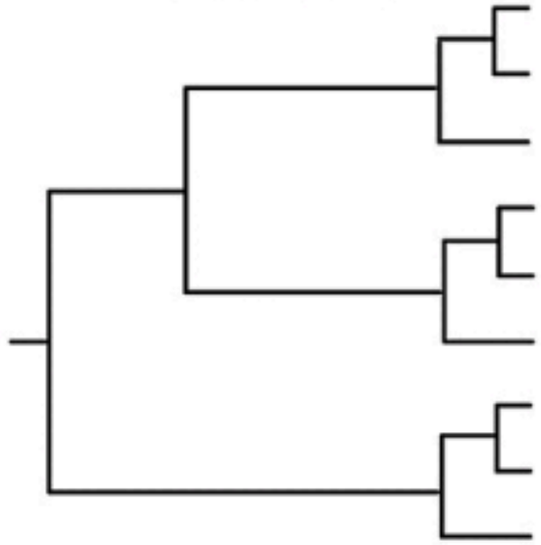
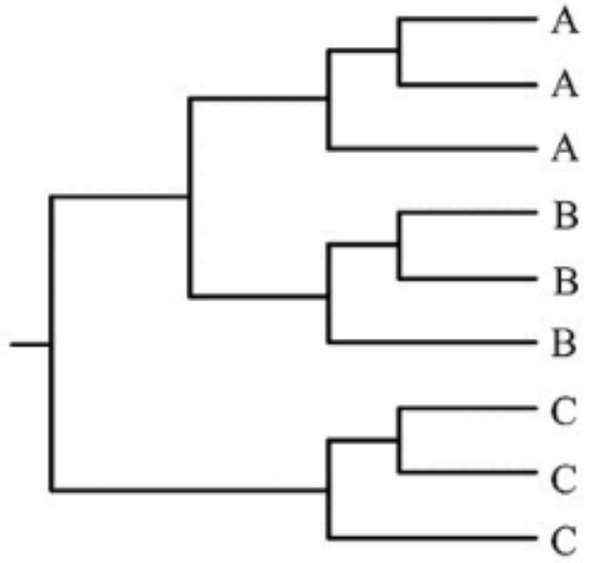
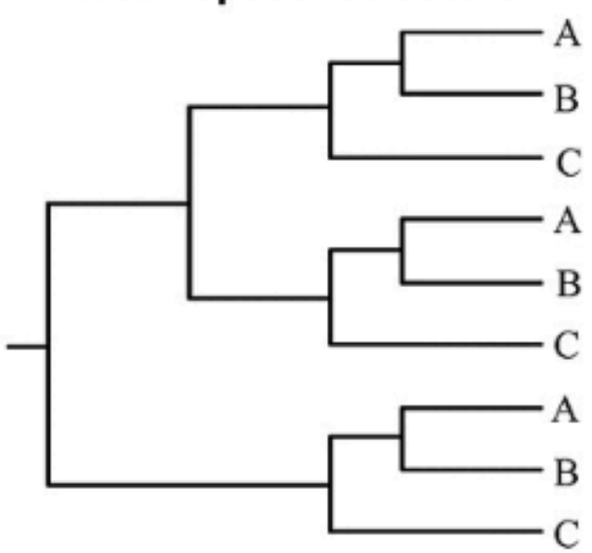
Coalescent trees look very different in large and small populations









	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective population dynamic processes	
		Population size dynamics	Spatial dynamics
Idealized Phylogeny Shapes		**Exponential growth**  **Constant size** 	**Strong spatial structure**  **Weak spatial structure** 
Examples	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV	Measles, rabies inter-host HIV
Tree Inferences	Detection of antigenic escape mutations	Estimation of population growth rates	Estimation of population migration rates

The key points we'll be making today:

1. Pathogen evolution happens on the same timescale as transmission, meaning that for rapidly evolving pathogens, **evolution, ecology, and epidemiology are linked.**

2. **Coalescent theory** allows us to generate expectations for phylogenies generated by populations that are evolving in particular ways.

3. These features allow us to use trees to infer parameters, and vice versa. We will see **examples** of how others have done this.

4. How can **we apply** these concepts to our own questions?

The key points we'll be making today:

1. Pathogen evolution happens on the same timescale as transmission, meaning that for rapidly evolving pathogens, **evolution, ecology, and epidemiology are linked.**

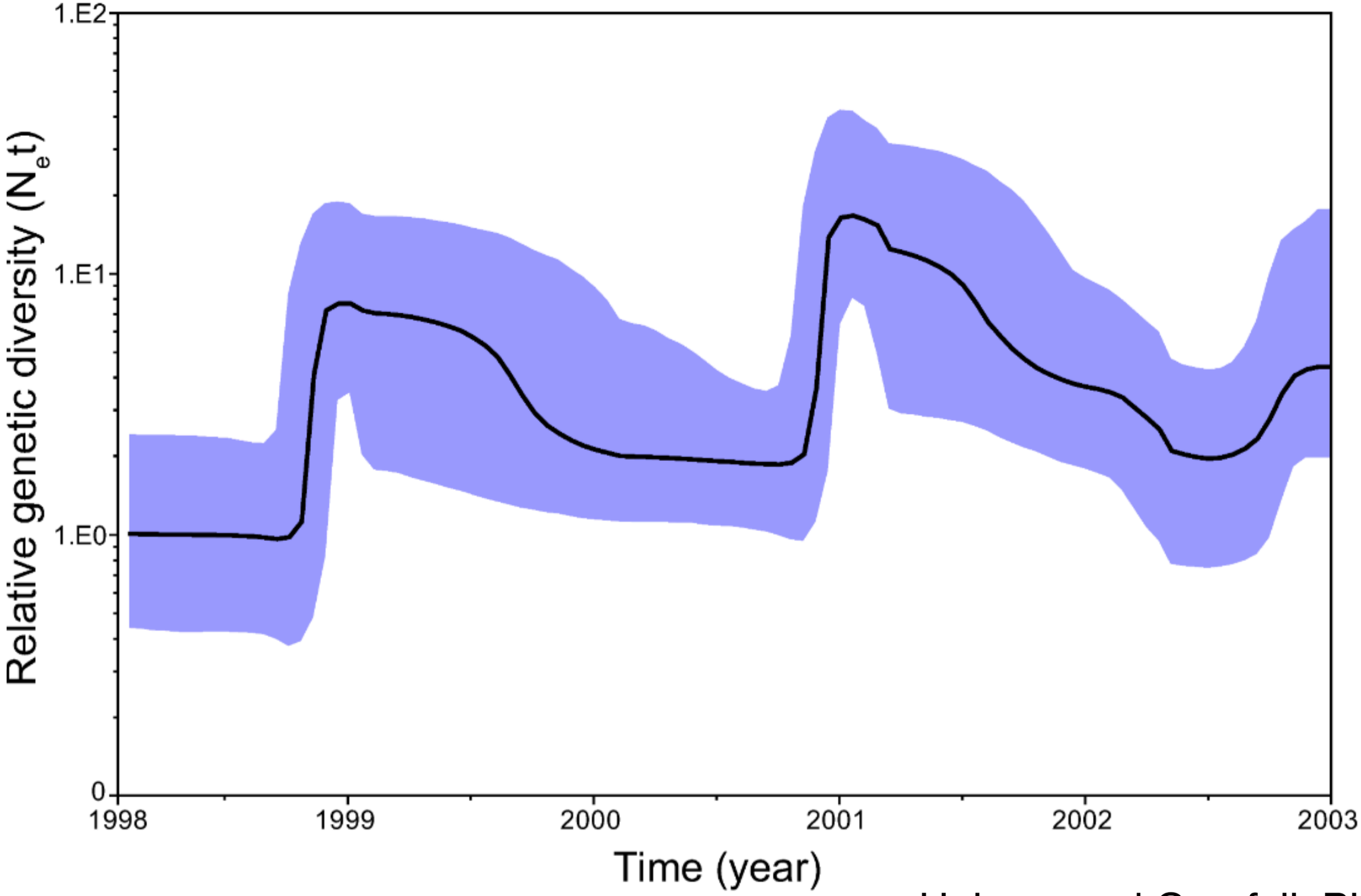
2. **Coalescent theory** allows us to generate expectations for phylogenies generated by populations that are evolving in particular ways.

3. These features allow us to use trees to infer parameters, and vice versa. We will see **examples** of how others have done this.

4. How can **we apply** these concepts to our own questions?

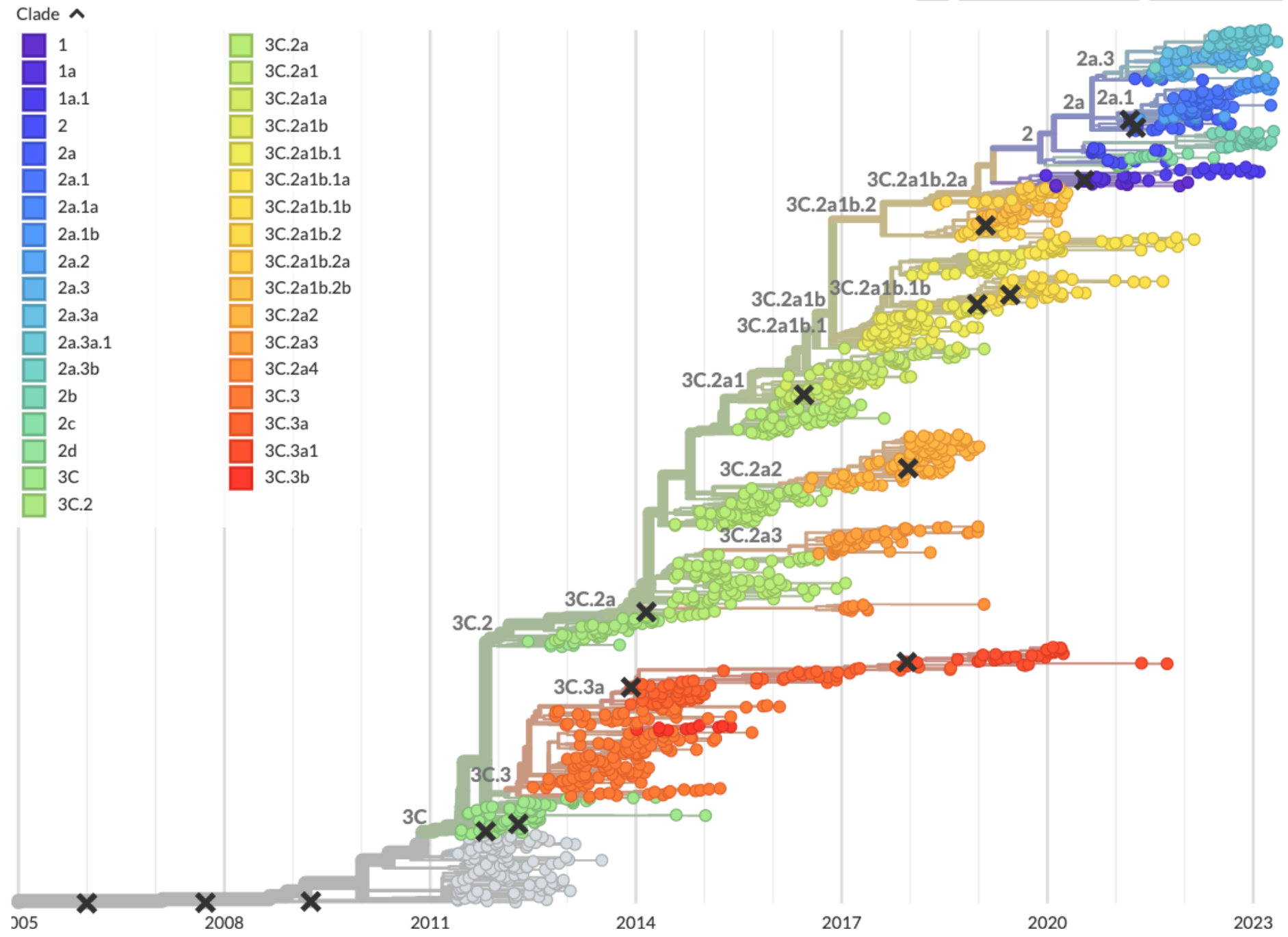
Some real life examples

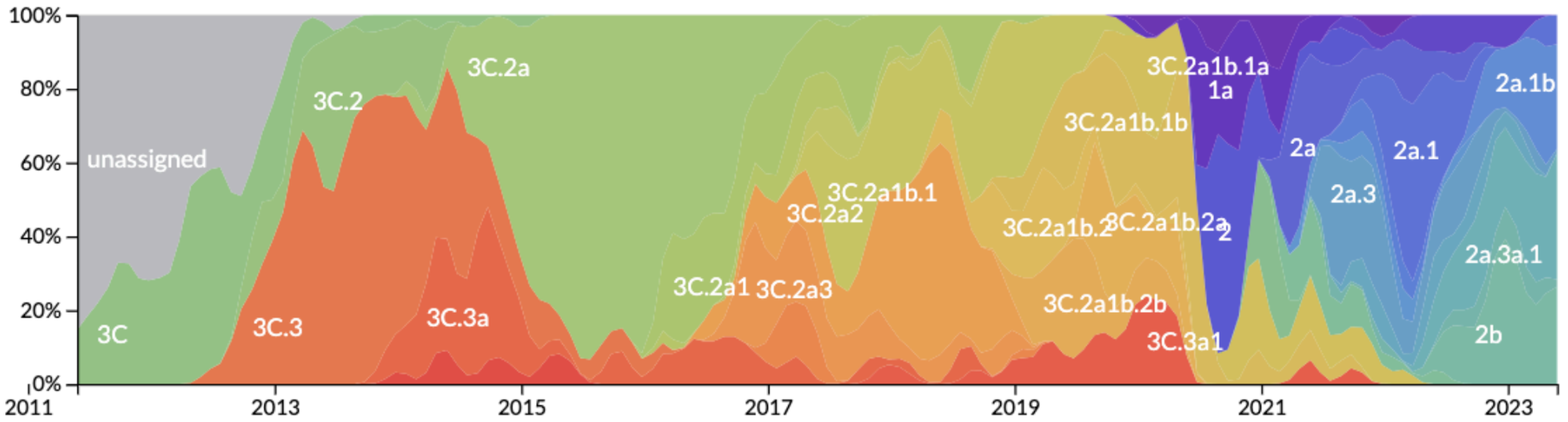
Population size changes in influenza in New York



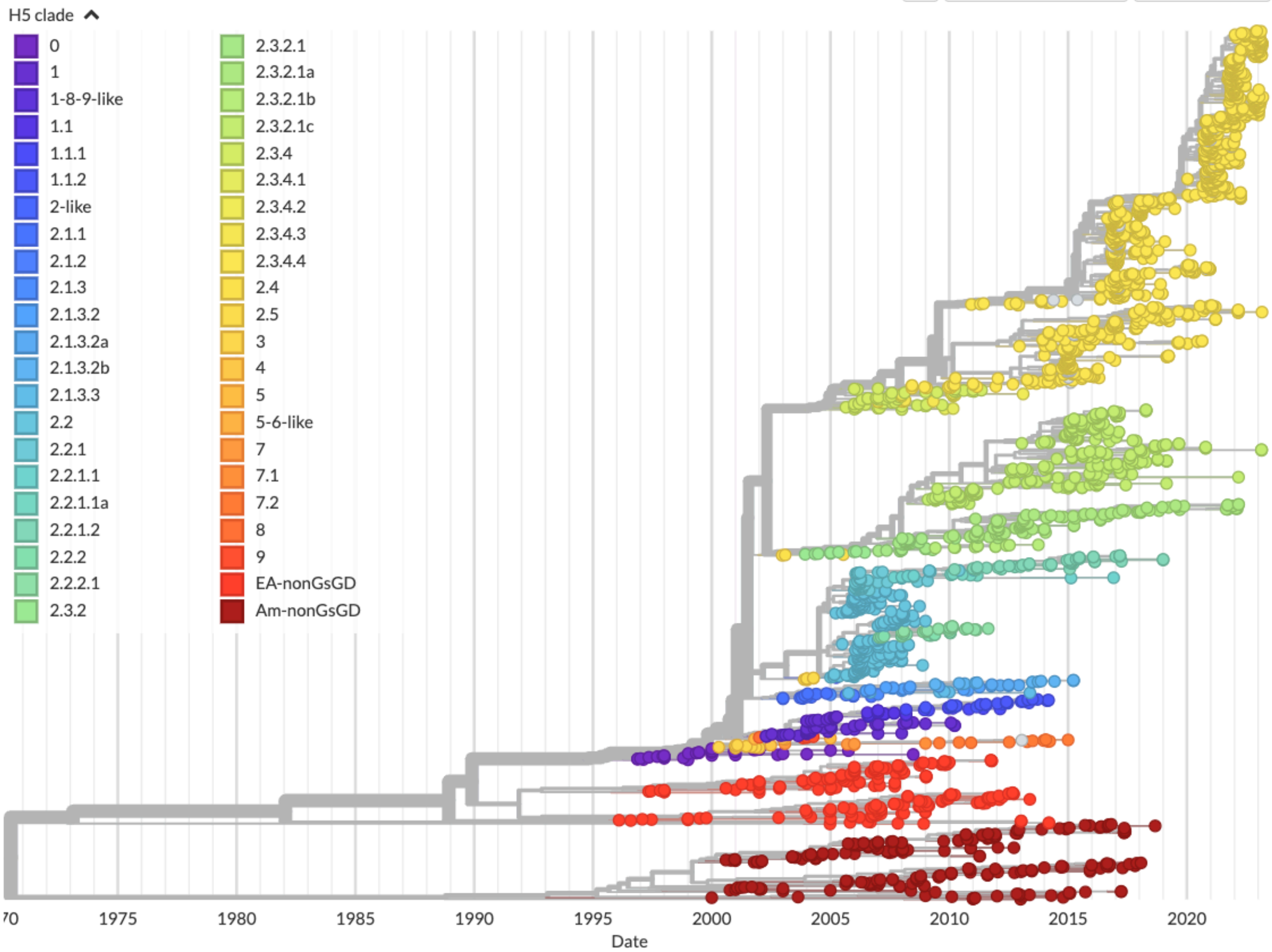
Holmes and Grenfell, PLOS Comp Bio, 2009

Seasonal influenza evolves yearly, necessitating vaccine updates





Avian influenza evolves with less selection, and more geographic structure



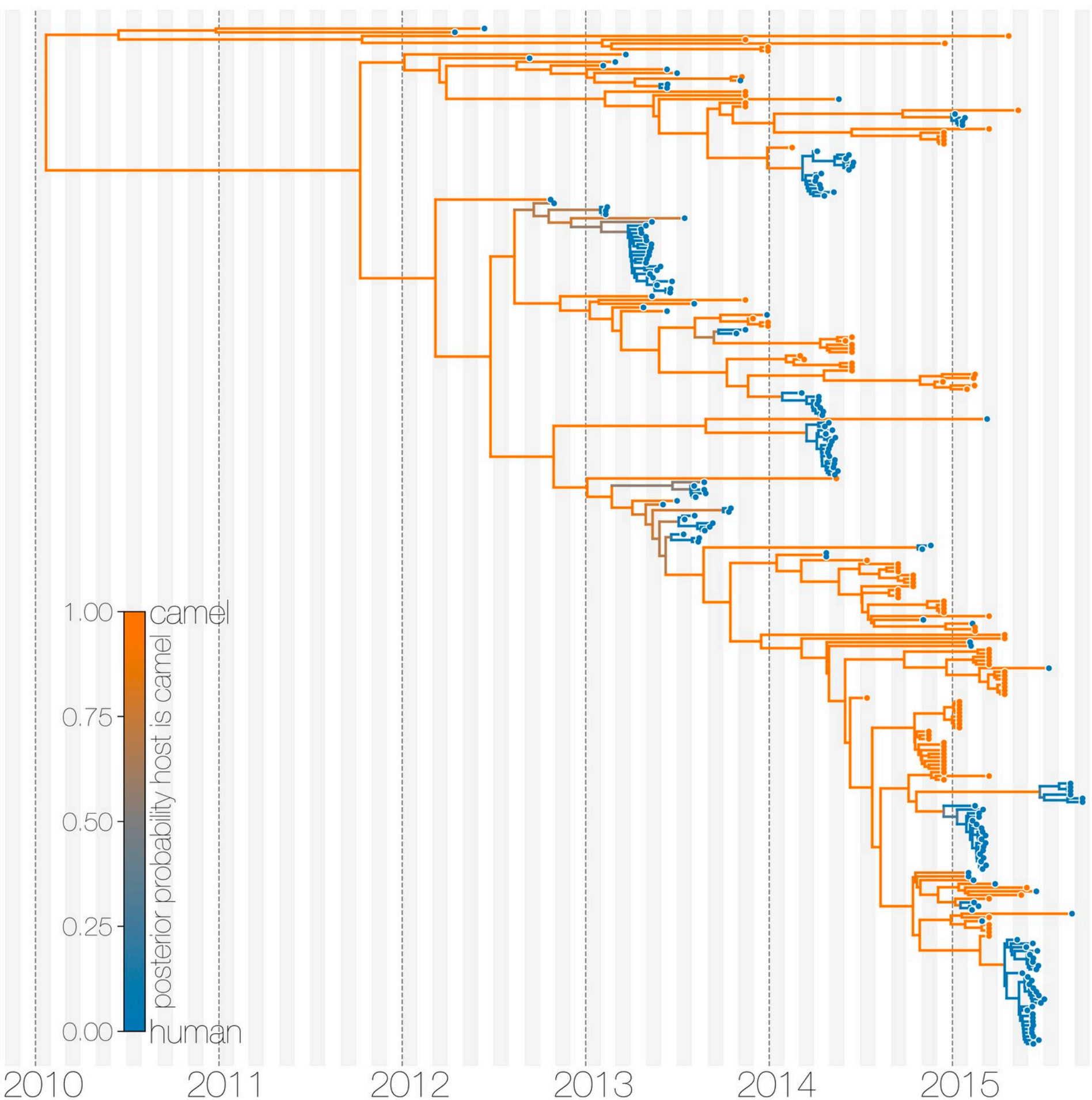
China
Southeast Asia
South Asia
Japan Korea
Oceania
West Asia
Africa
Europe
South America
North America

H3N2 seasonal influenza

2a.3
2a.1
2a
2
3C.2a1b.2a
3C.2a1b.2
3C.2a1b.1b
3C.2a1b
3C.2a1b.1
3C.2a1
3C.2a2
3C.2a
3C.2
3C.3a
3C.3
3C
2010
2015
2020
70

H5Nx avian influenza

1980
1990
2000
2010
2020



Inferring host switching for MERS

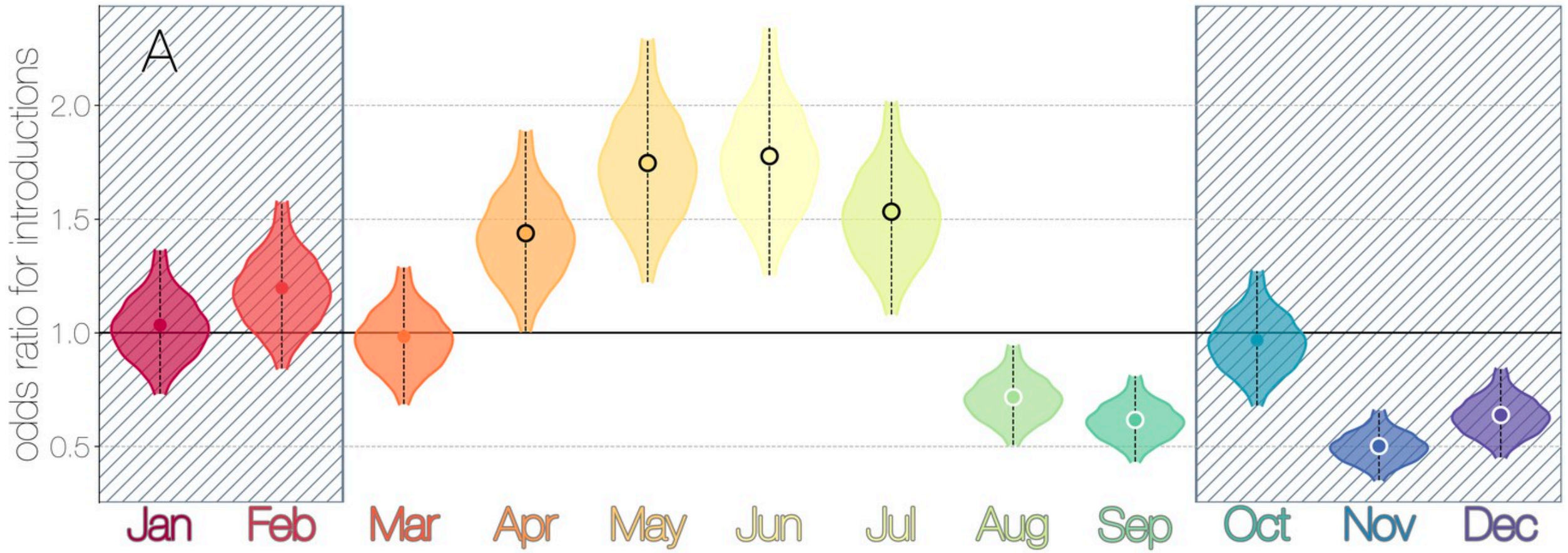
NeTau estimates:

Humans: 0.24 years

Camels: 3.5 years

Dudas et al, eLife 2020

Spring calving season correlates with an increase in the number of cross-species transmission events



Dudas et al, eLife, 2020

Extra resources and cool examples:

- West Nile virus in the Americas: https://nextstrain.org/narratives/twenty-years-of-WNV and https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1008042

- Ebola in West Africa: https://www.nature.com/articles/nature22040

- HIV in humans: https://www.science.org/doi/full/10.1126/science.1256739

- Cryptic transmission of SARS-CoV-2 in Washington state: https://www.science.org/doi/full/10.1126/science.abc0523

-

We will be using 2 pieces of software to explore phylodynamic inference

How can we scale this to more than 2 lineages?

In any given generation, the **probability** that *i* individuals pick the same parent and coalesce:

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

The expected **waiting time** *i* lineages to coalesce is <u>exponentially distributed</u> with a mean of:

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

What the heck is *i* choose 2?

In any given generation, the **probability** that *i* individuals pick the same parent and coalesce:

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

The expected **waiting time** *i* lineages to coalesce is exponentially distributed with a mean of:

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