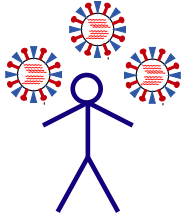


Phylodynamics for recombining pathogens

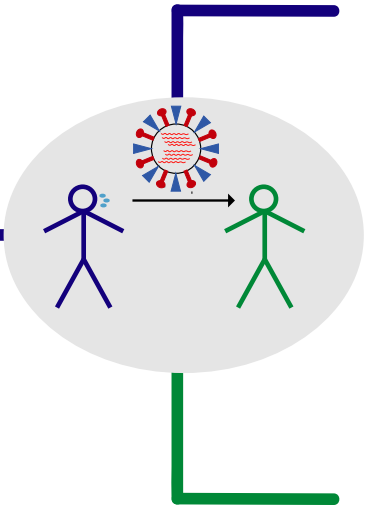
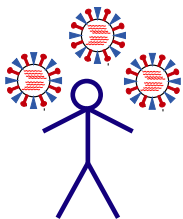
Inferring phylogenetic networks

Nicola Felix Müller, PhD

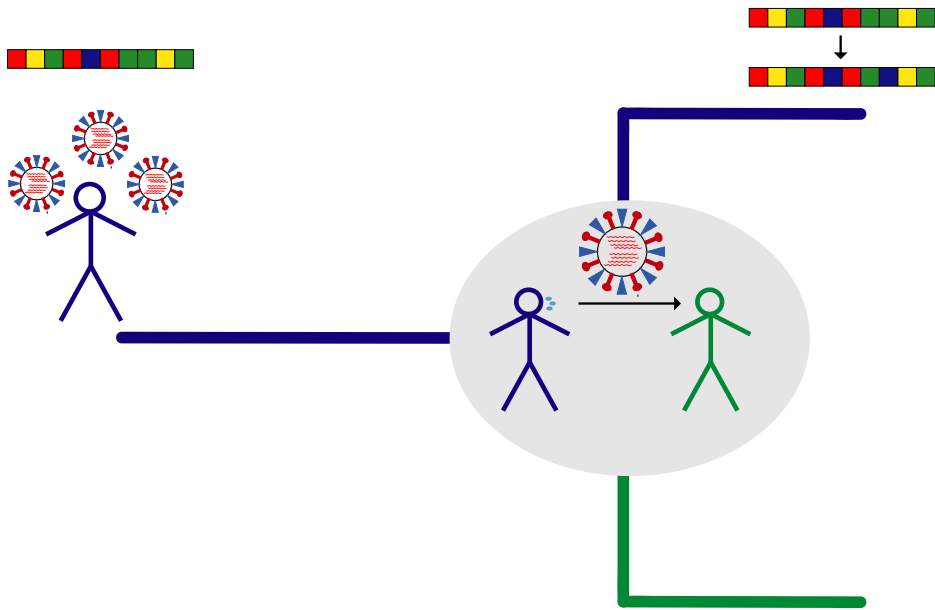
e-mail: nicola.mueller@ucsf.edu



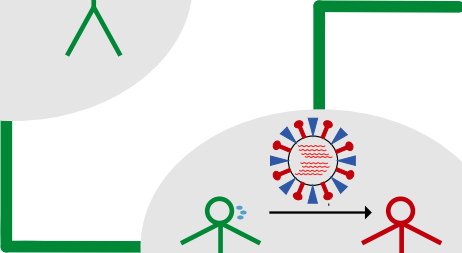
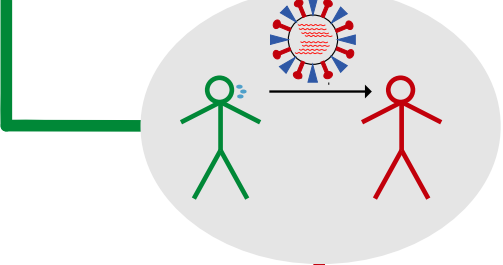
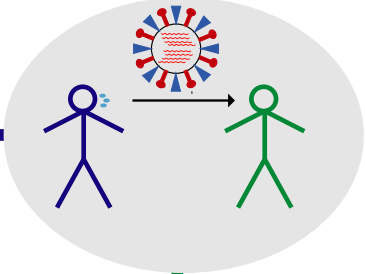
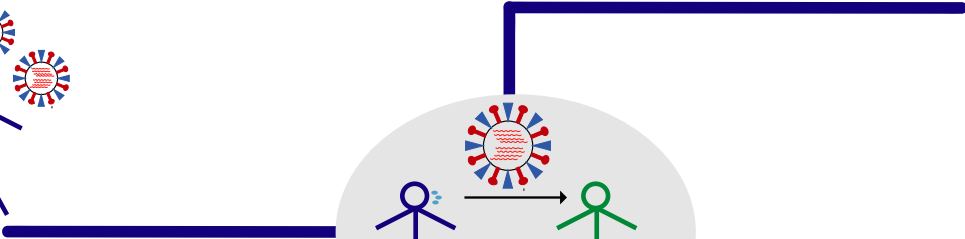
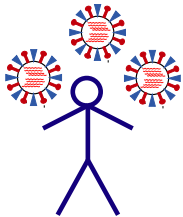
time



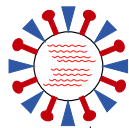
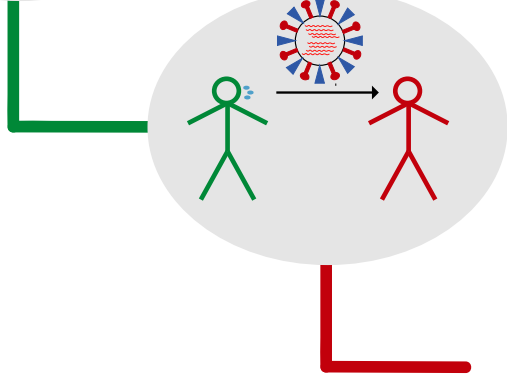
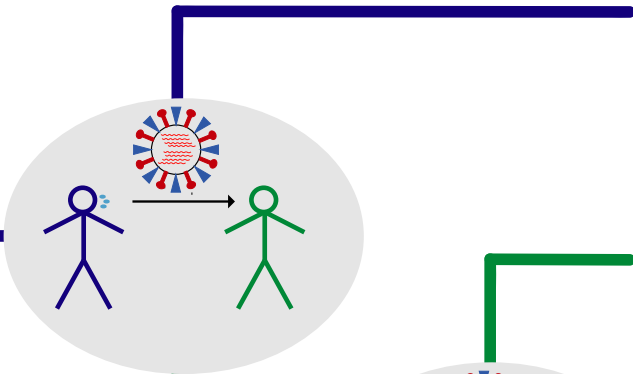
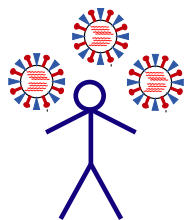
time



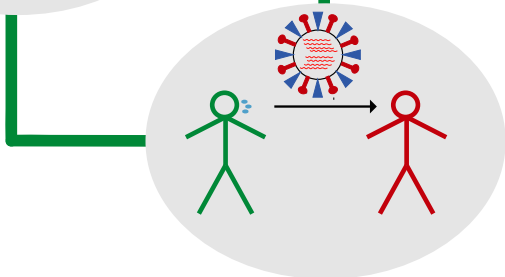
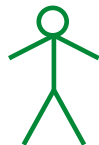
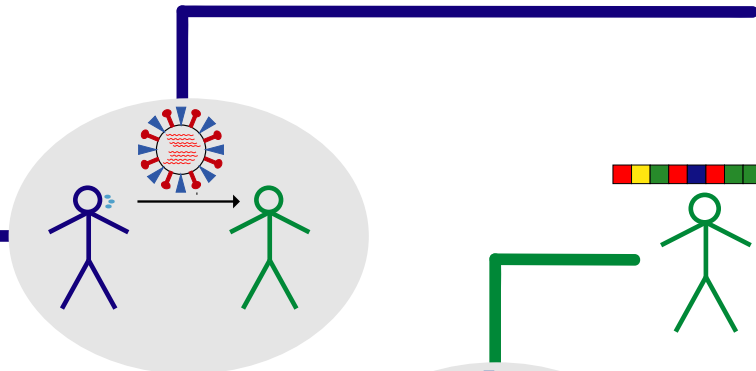
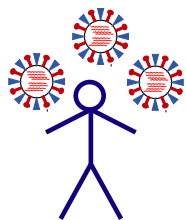
time



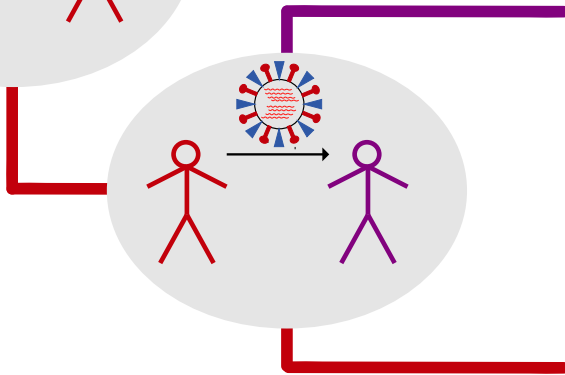
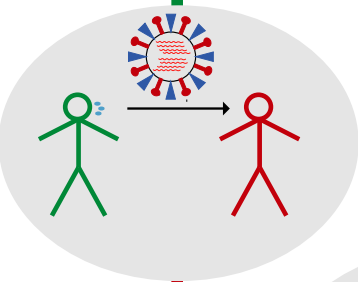
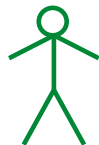
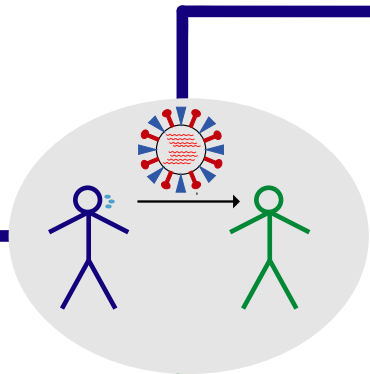
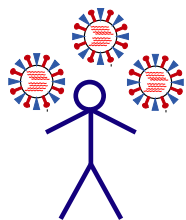
time



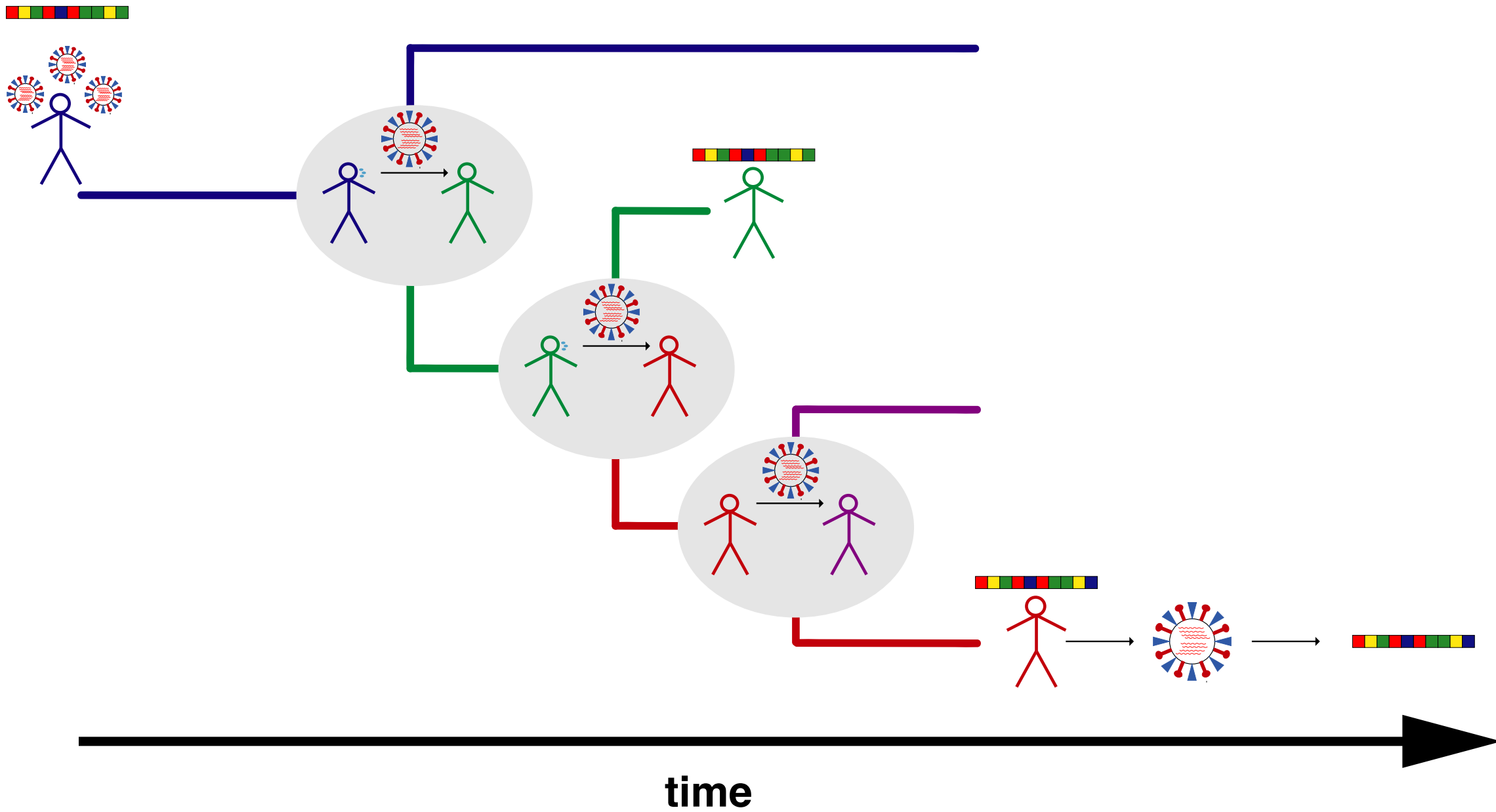
time

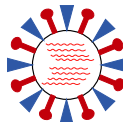
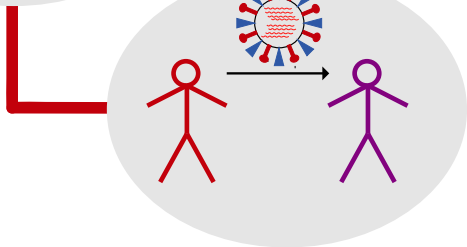
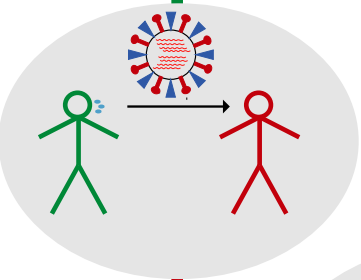
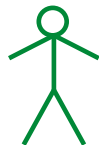
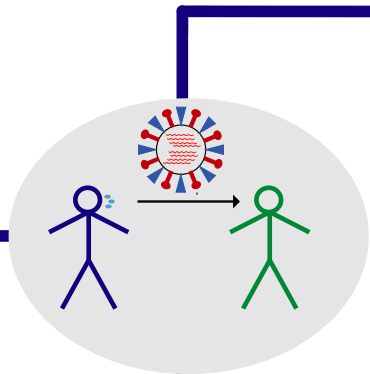
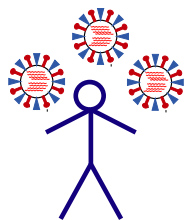


time

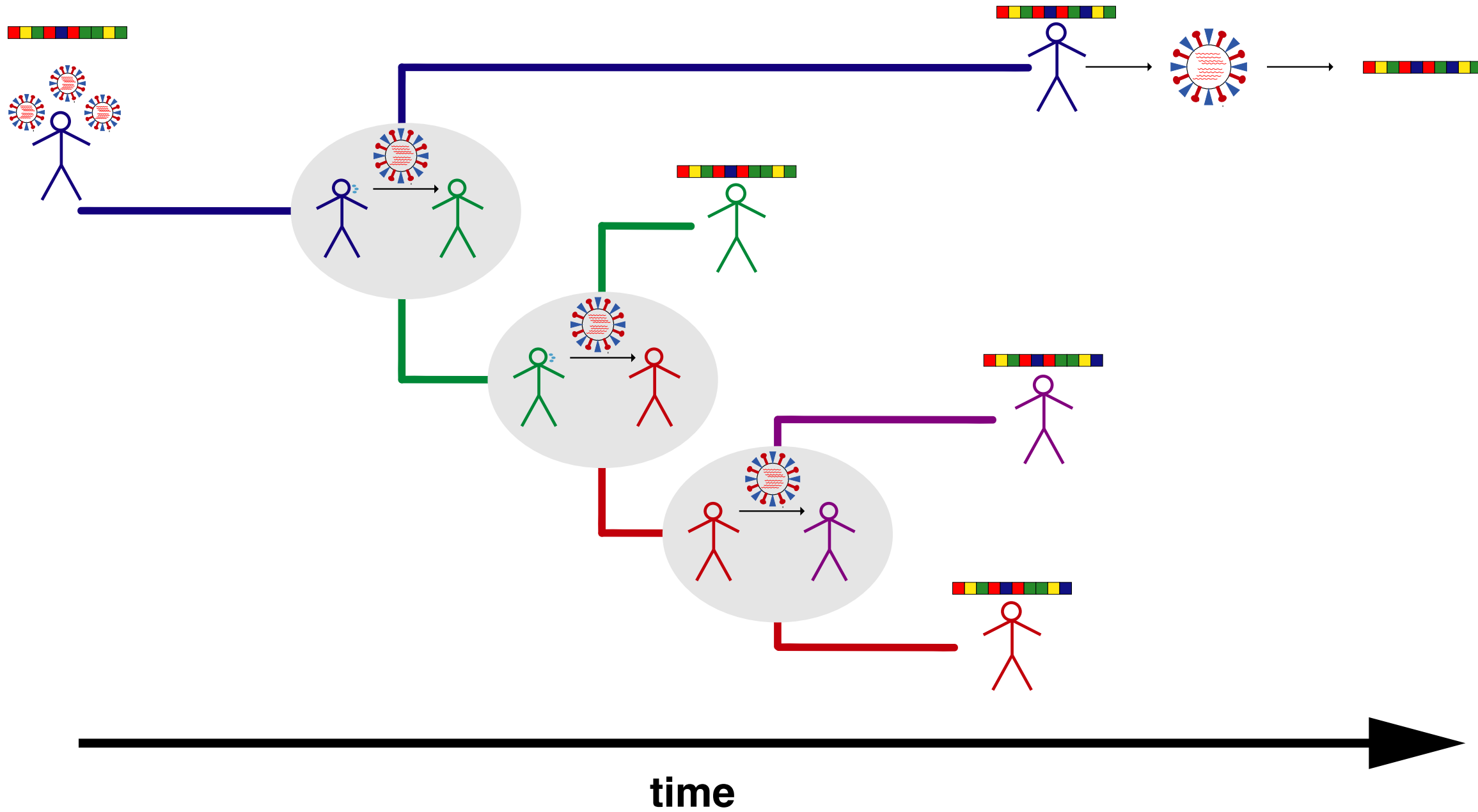


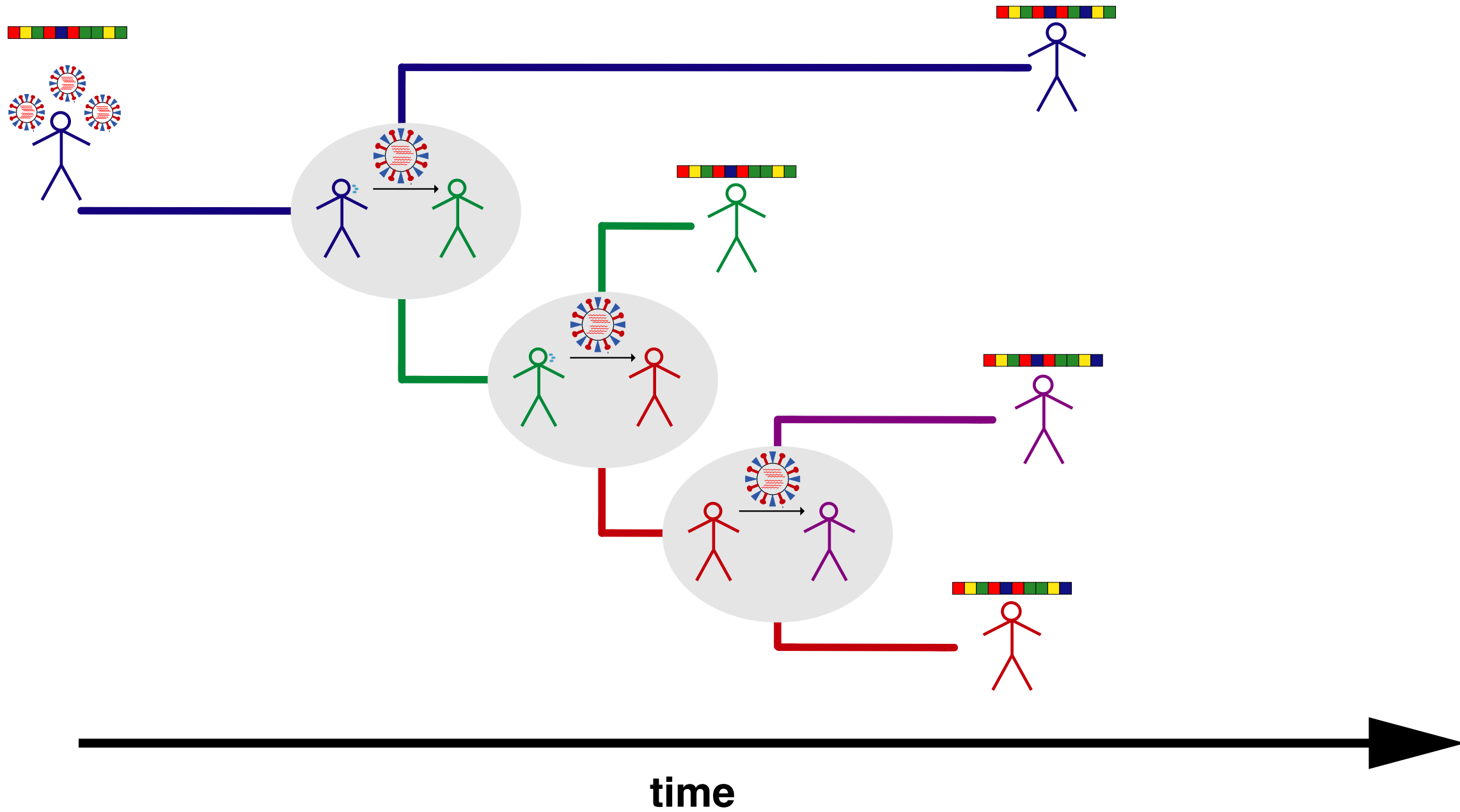
time

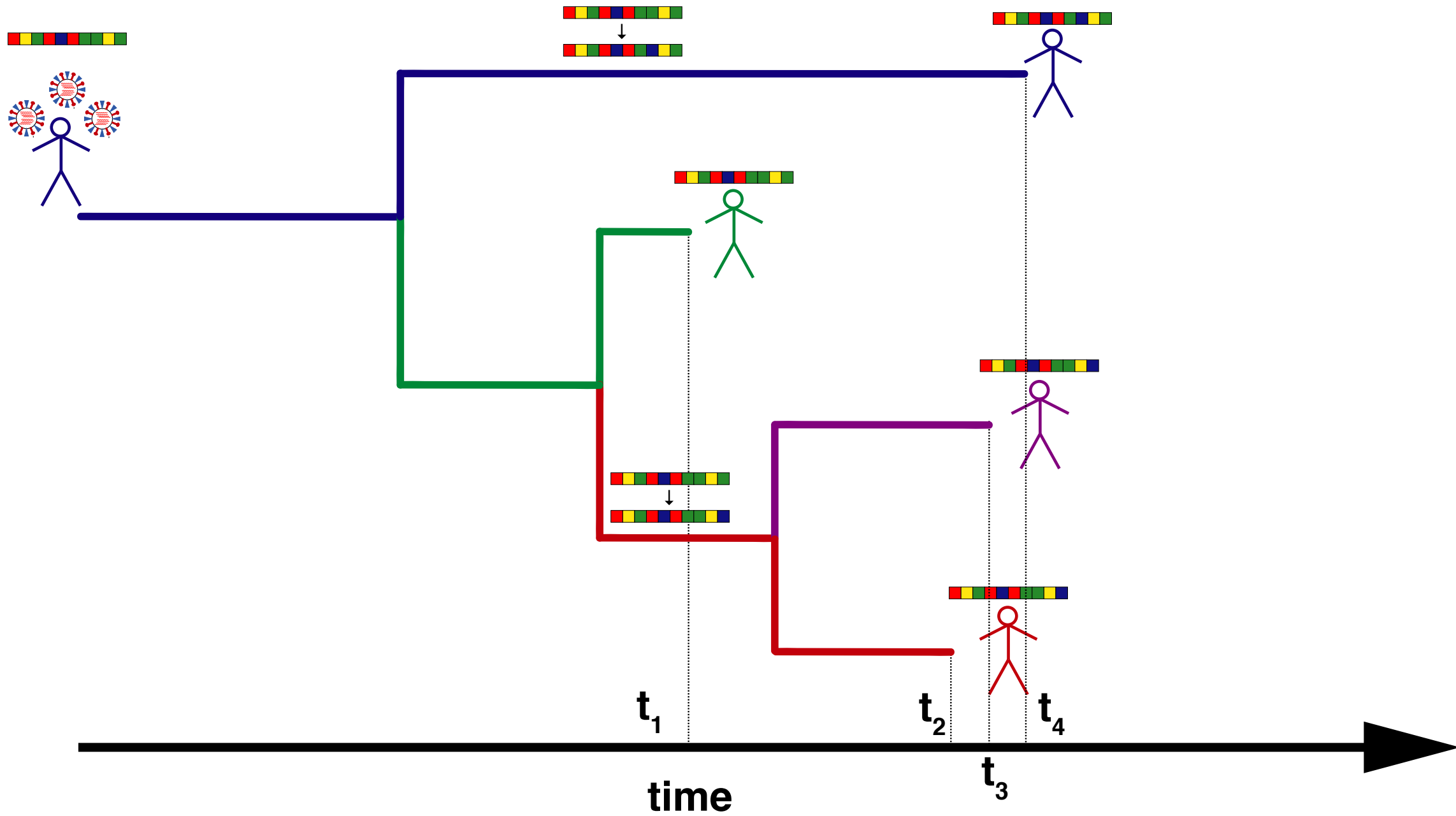


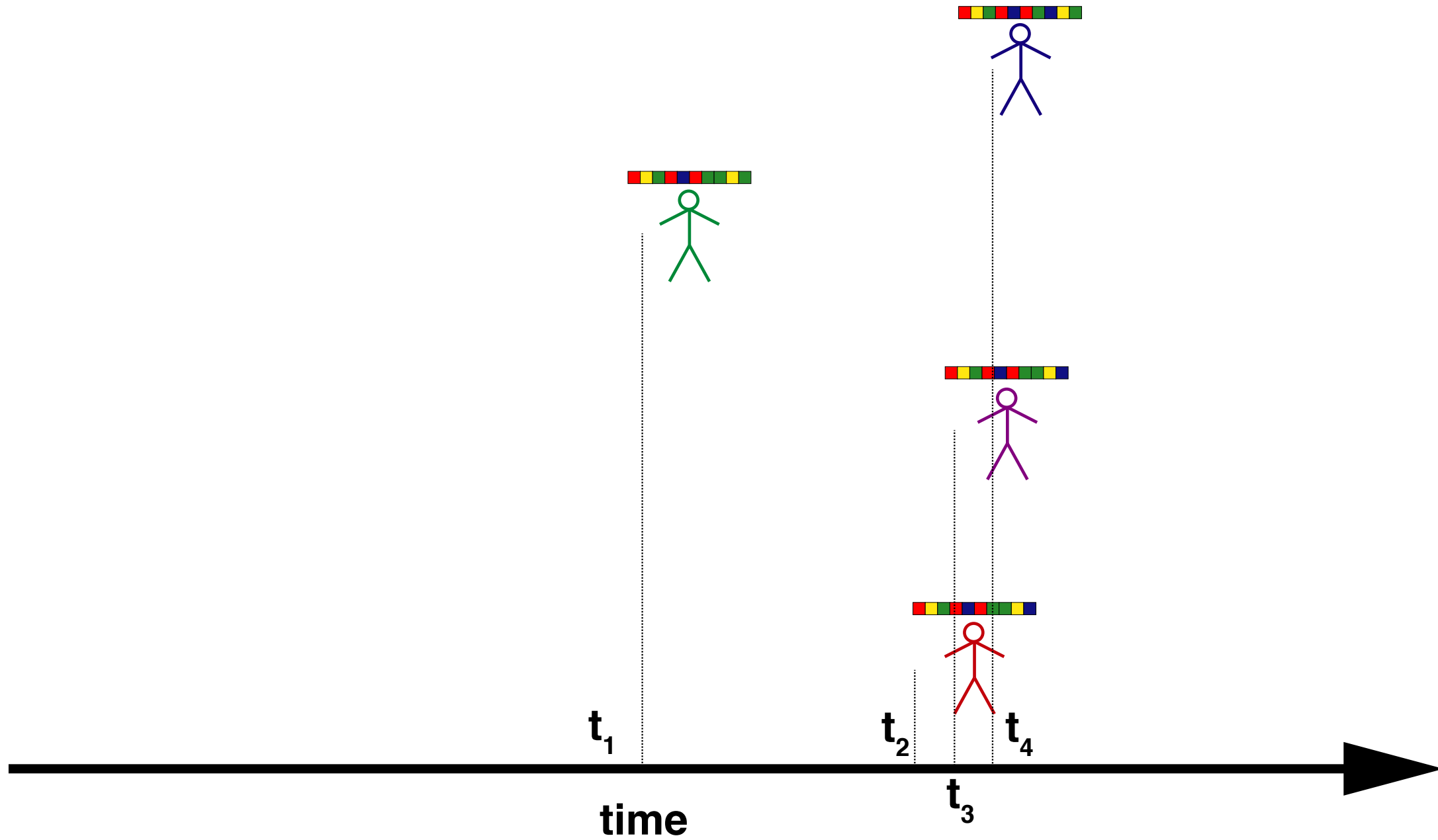


time

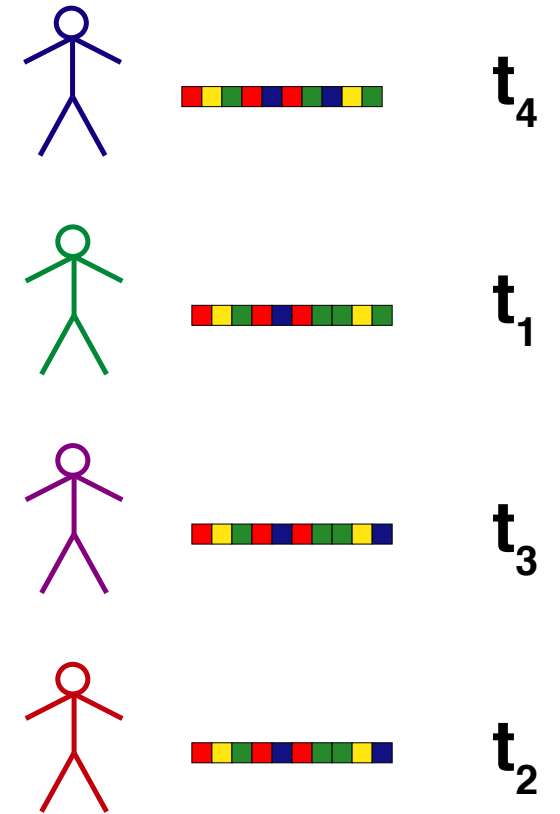




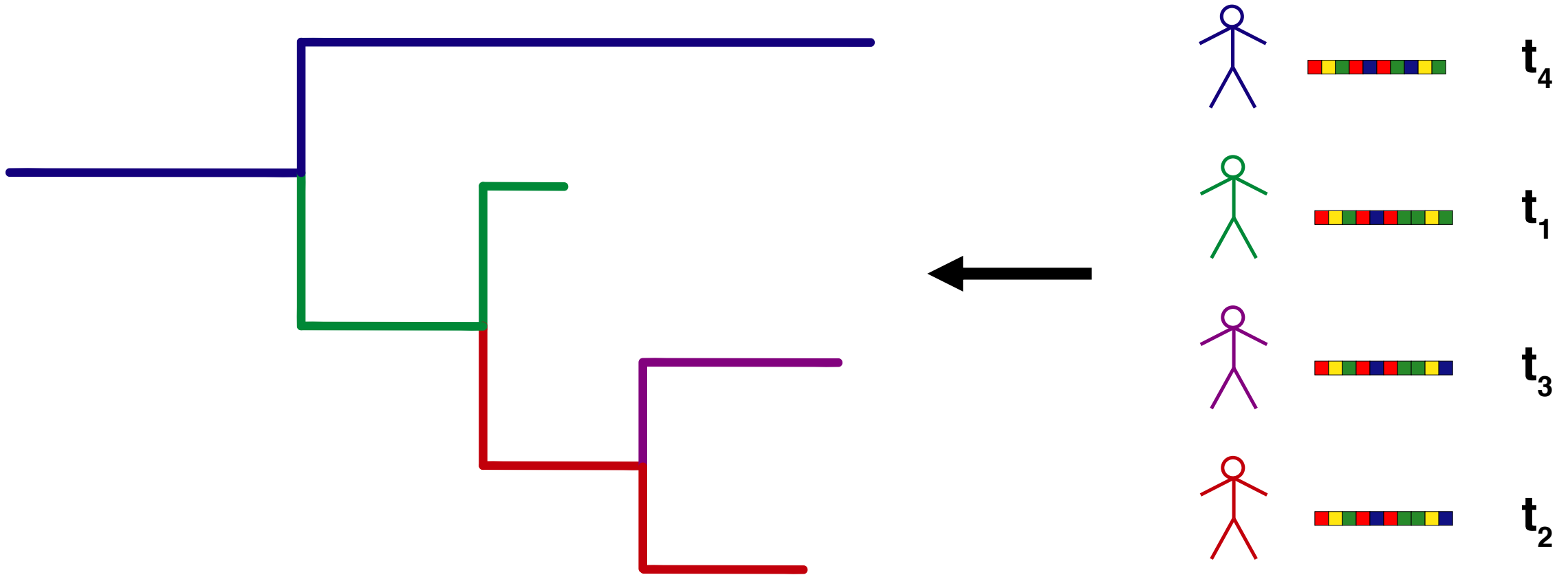




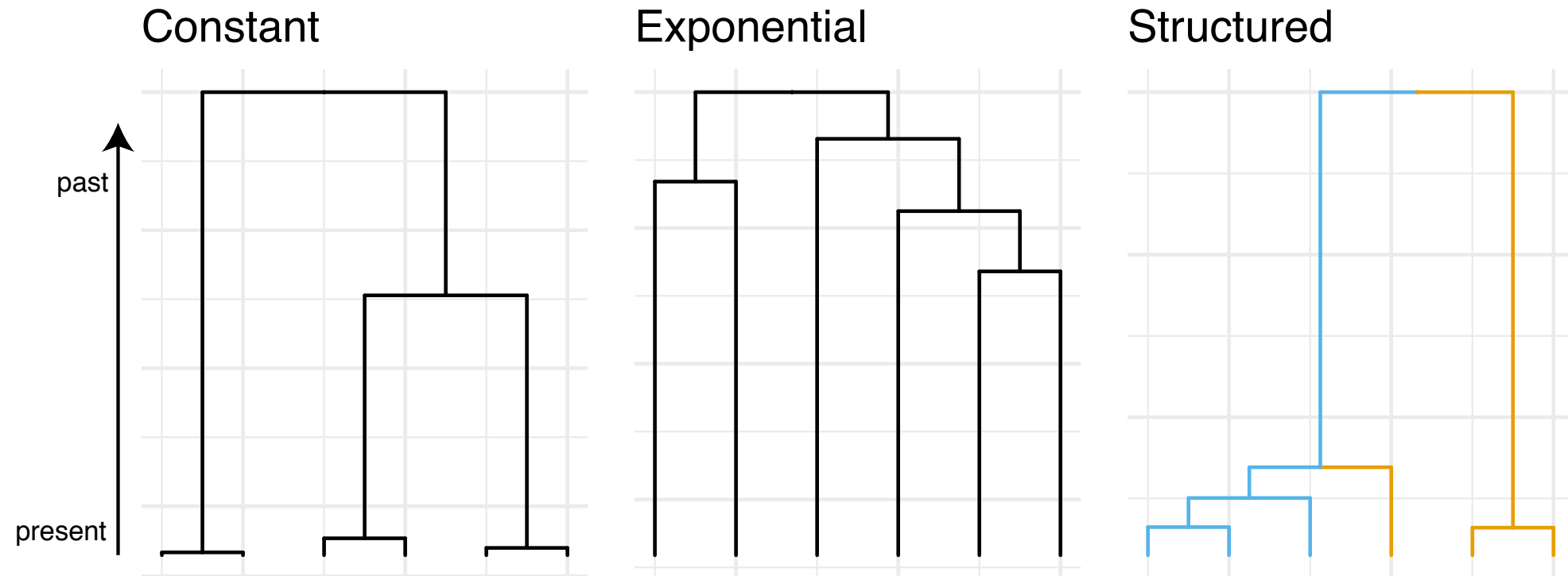
Our data: We know who was infected with which pathogen and sequence and when they were sampled



Phylogenetics allows us to infer the shared ancestral history of the different pathogens



Phylogenetic trees are formed by population processes and contain information about them



Bayesian phylogenetics allows us to jointly infer the phylogenetic trees, evolutionary and Demographics models

$$P(\text{Phylogeny} \mid \text{Sequence Data})$$

Bayesian phylogenetics allows us to jointly infer the phylogenetic trees, evolutionary and Demographics models

Tree generating models



$$P(\text{Tree} \mid \text{Data}) = \frac{P(\text{Tree} \mid \text{Model}) P(\text{Model})}{P(\text{Data})}$$

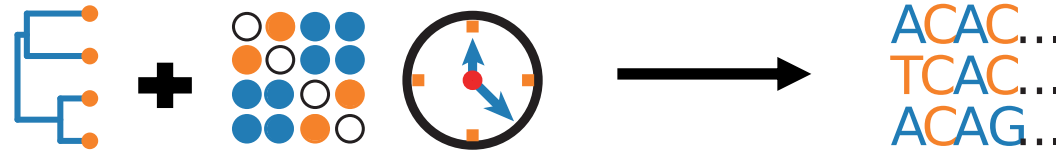
The equation represents the joint inference of a phylogenetic tree, evolutionary model, and demographic model from genomic data. The numerator consists of the likelihood of the tree given the model and the prior probability of the model. The denominator is the marginal likelihood of the data, which is the sum of the numerator over all possible trees and models.

Bayesian phylogenetics allows us to jointly infer the phylogenetic trees, evolutionary and Demographics models

Tree generating models

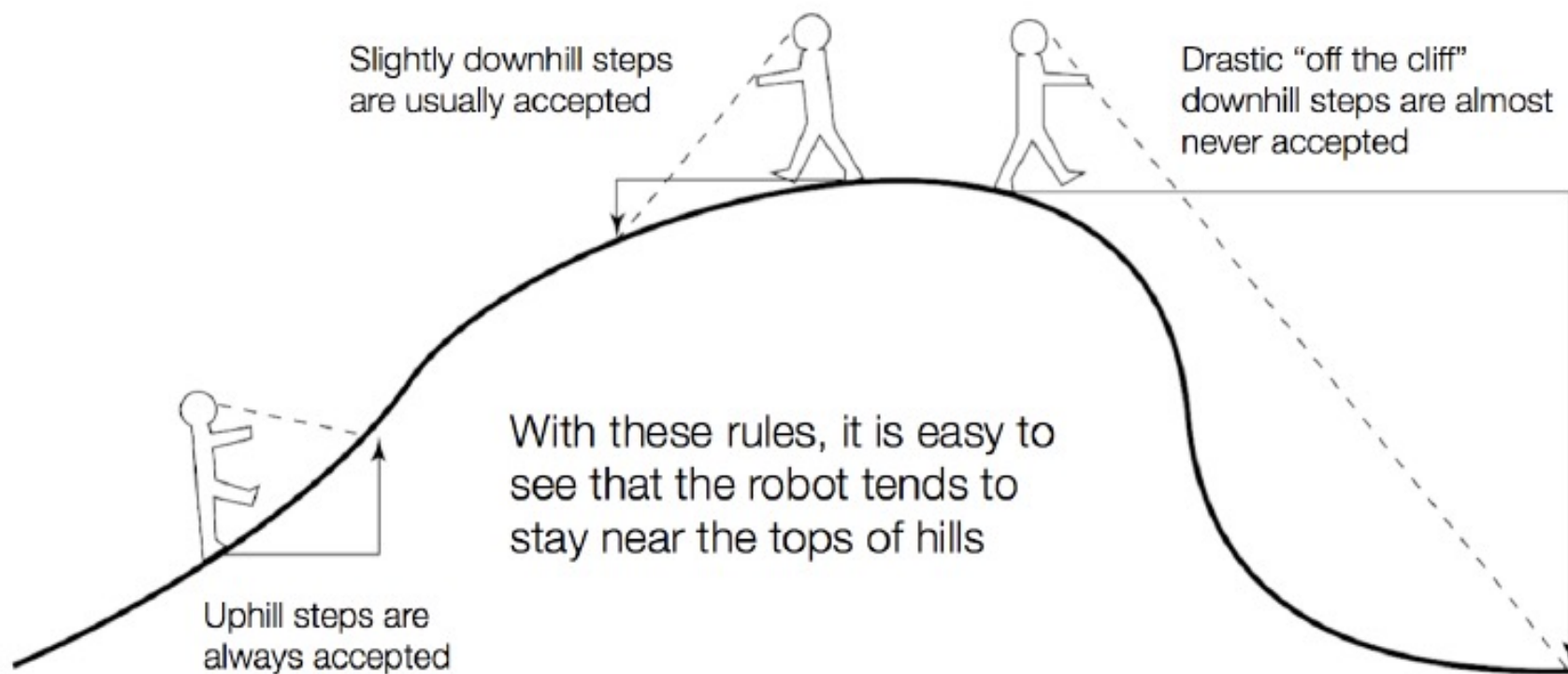


Model of sequence evolution

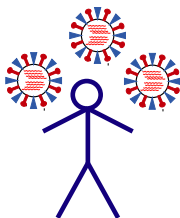


$$P(\text{Tree, Matrix, Clock} \mid \text{Sequences}) = \frac{P(\text{Sequences} \mid \text{Tree, Matrix, Clock}) P(\text{Tree} \mid \text{Coalescent}) P(\text{Coalescent}) P(\text{Matrix}) P(\text{Clock})}{P(\text{Sequences})}$$

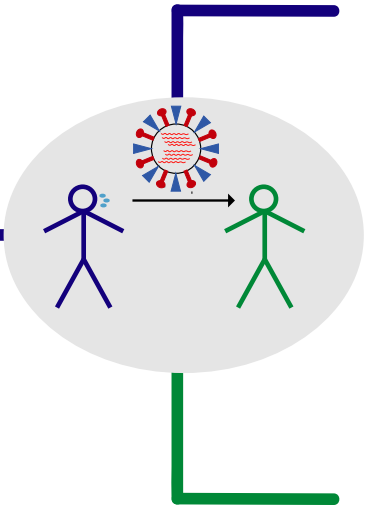
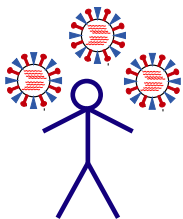
Exploring the posterior probability distribution using Markov chain Monte Carlo (MCMC)



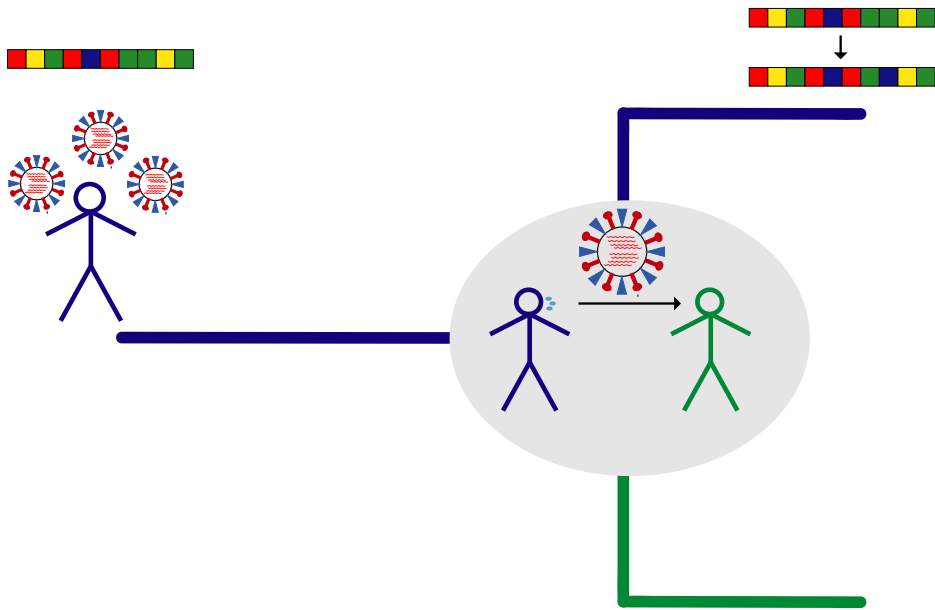
Slide from Paul Lewis



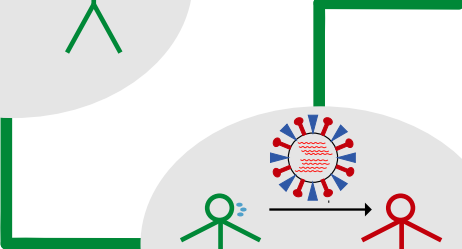
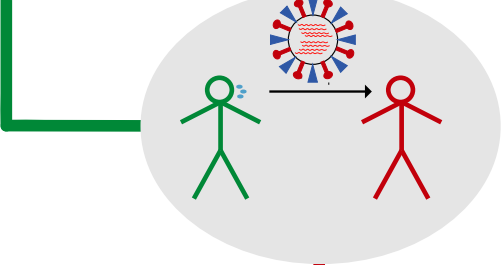
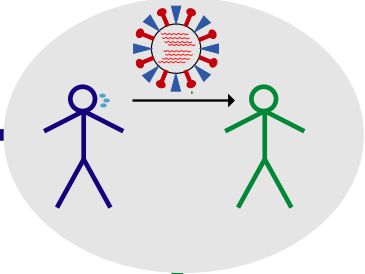
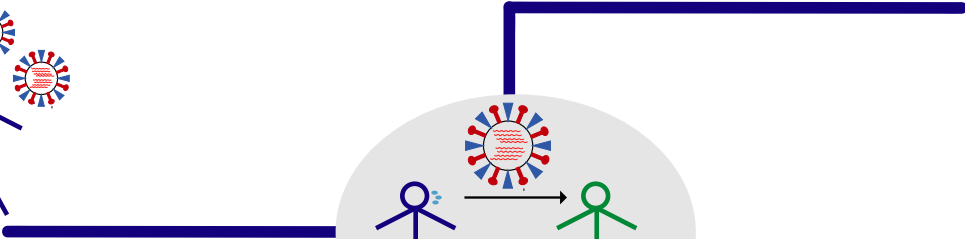
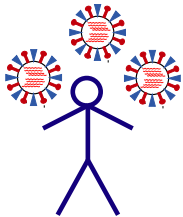
time



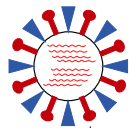
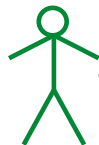
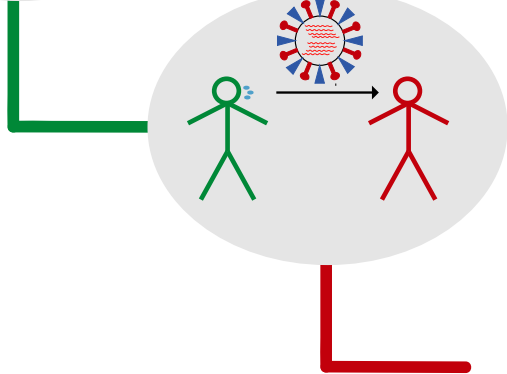
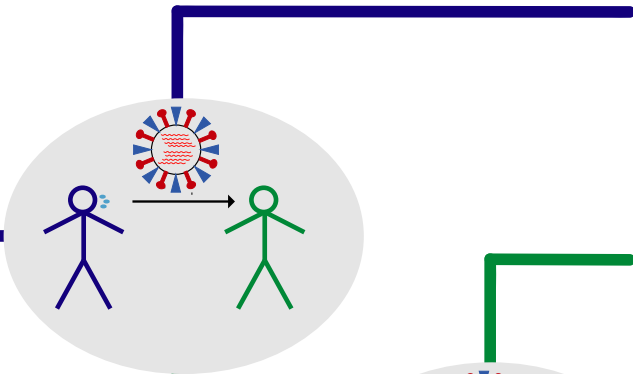
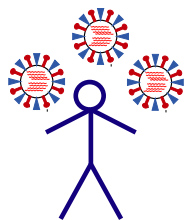
time



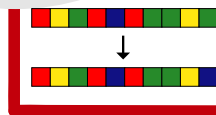
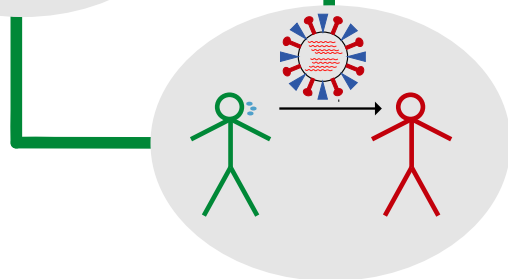
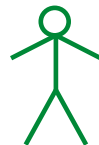
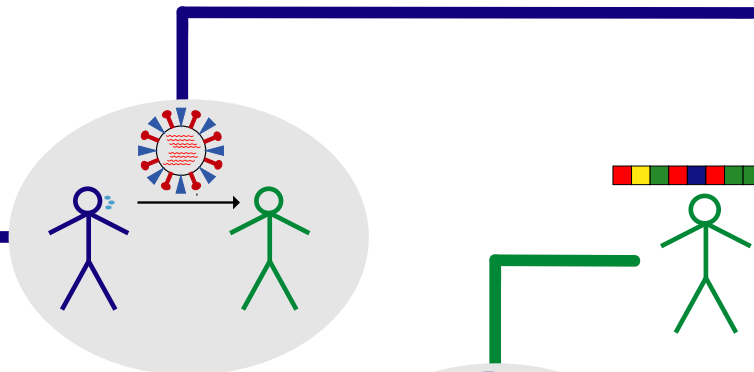
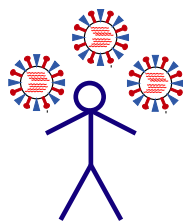
time



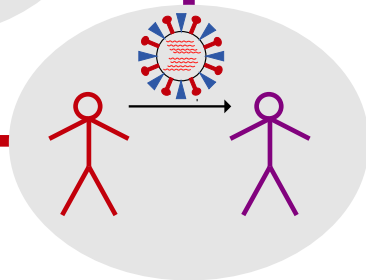
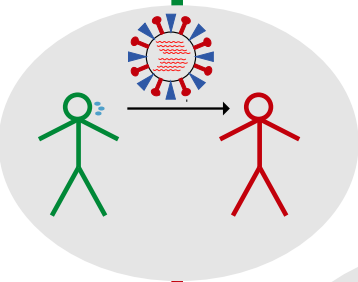
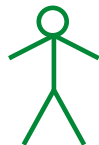
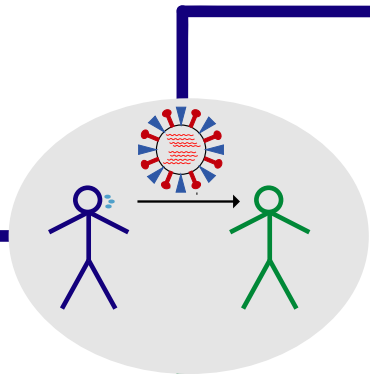
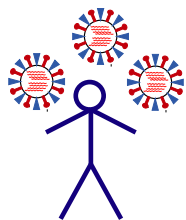
time



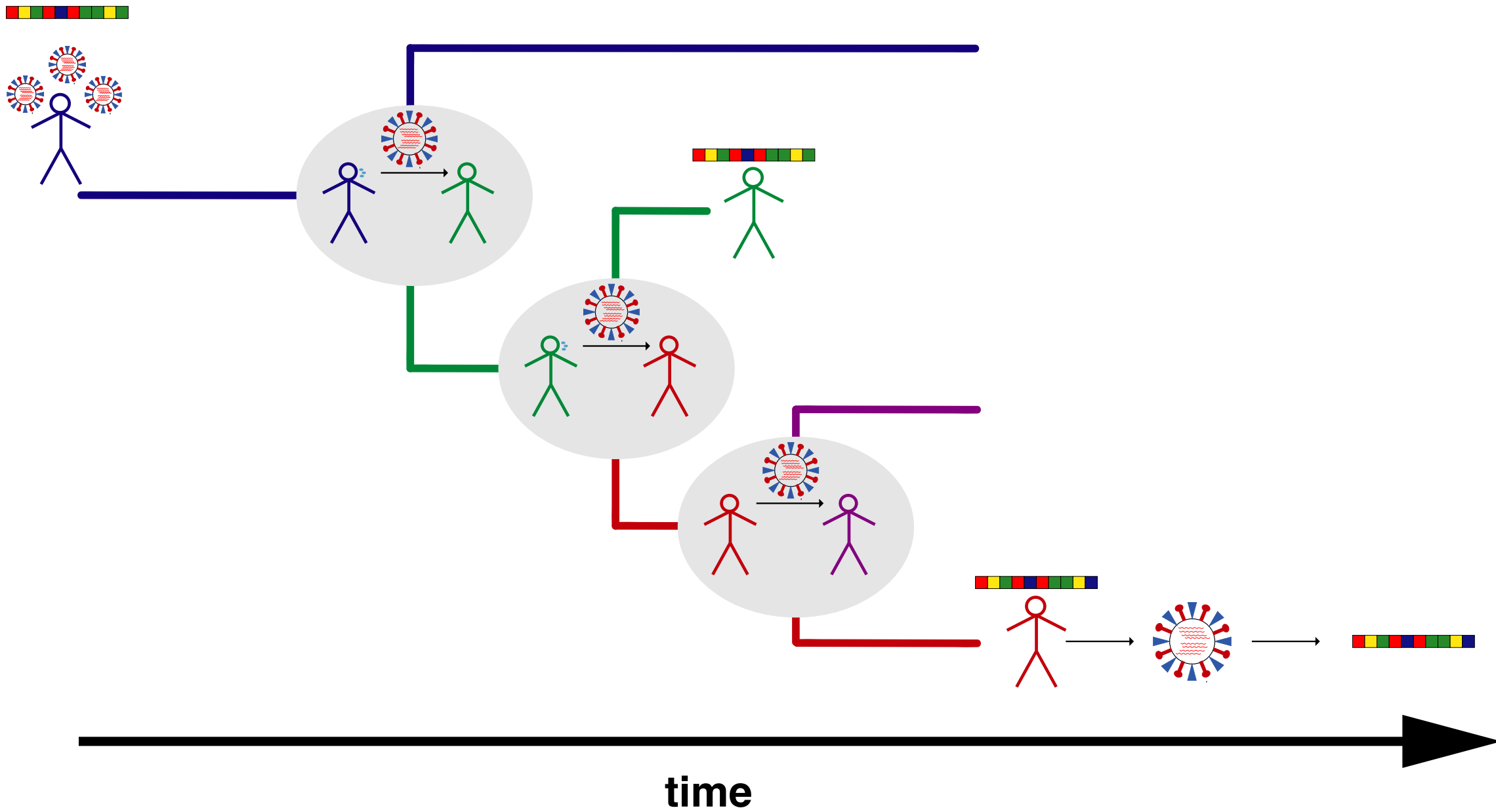
time

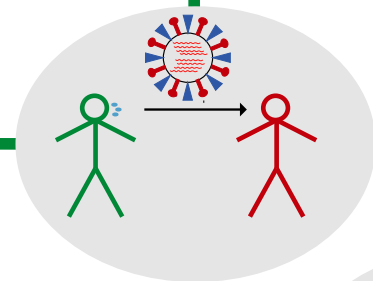
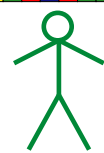
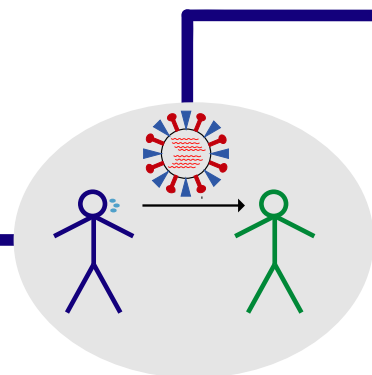
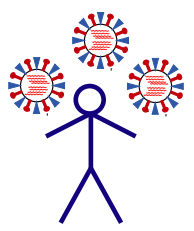


time

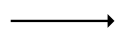
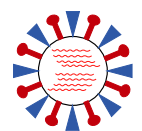
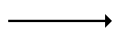
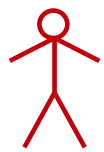
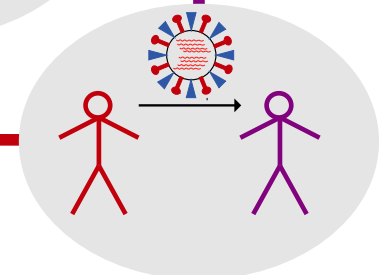


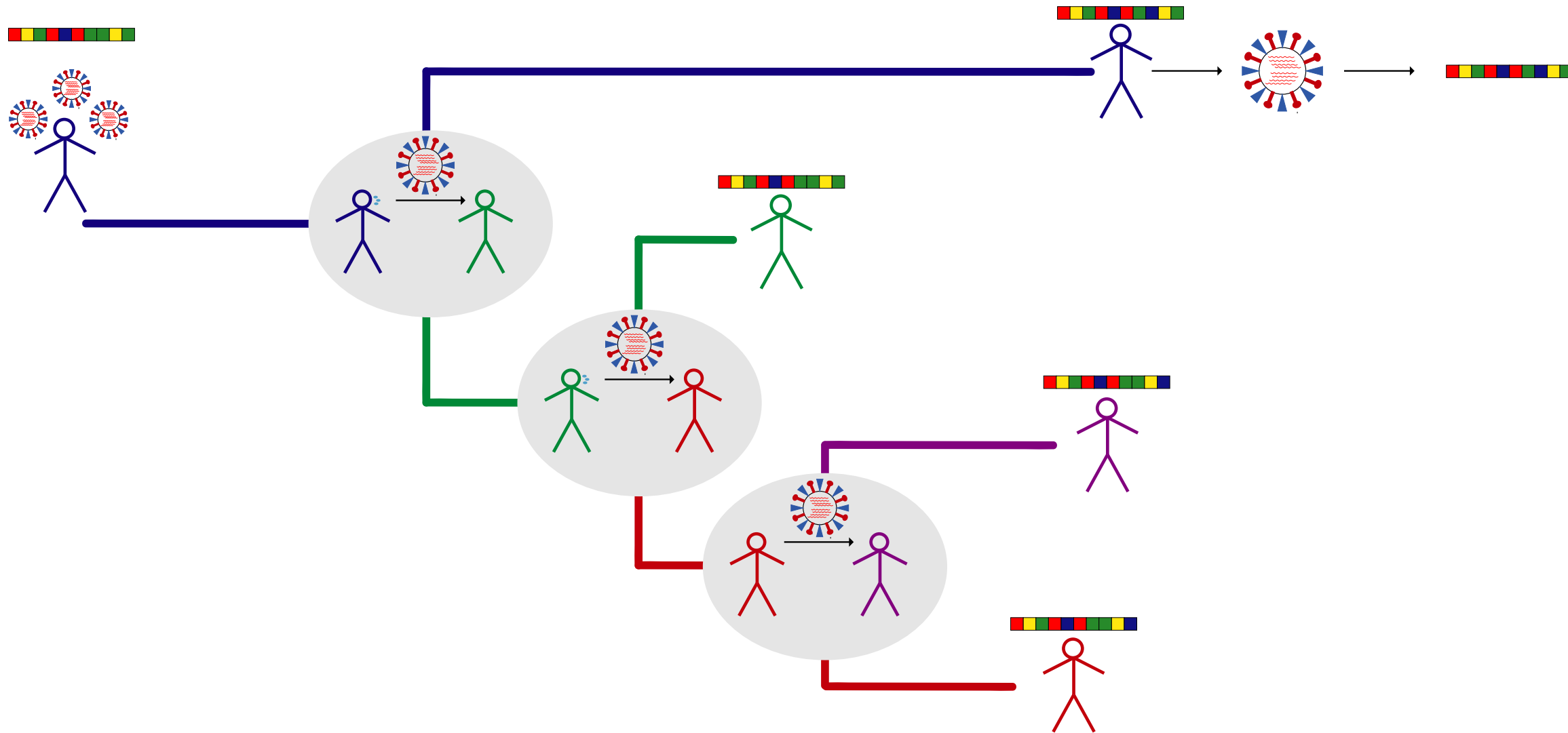
time



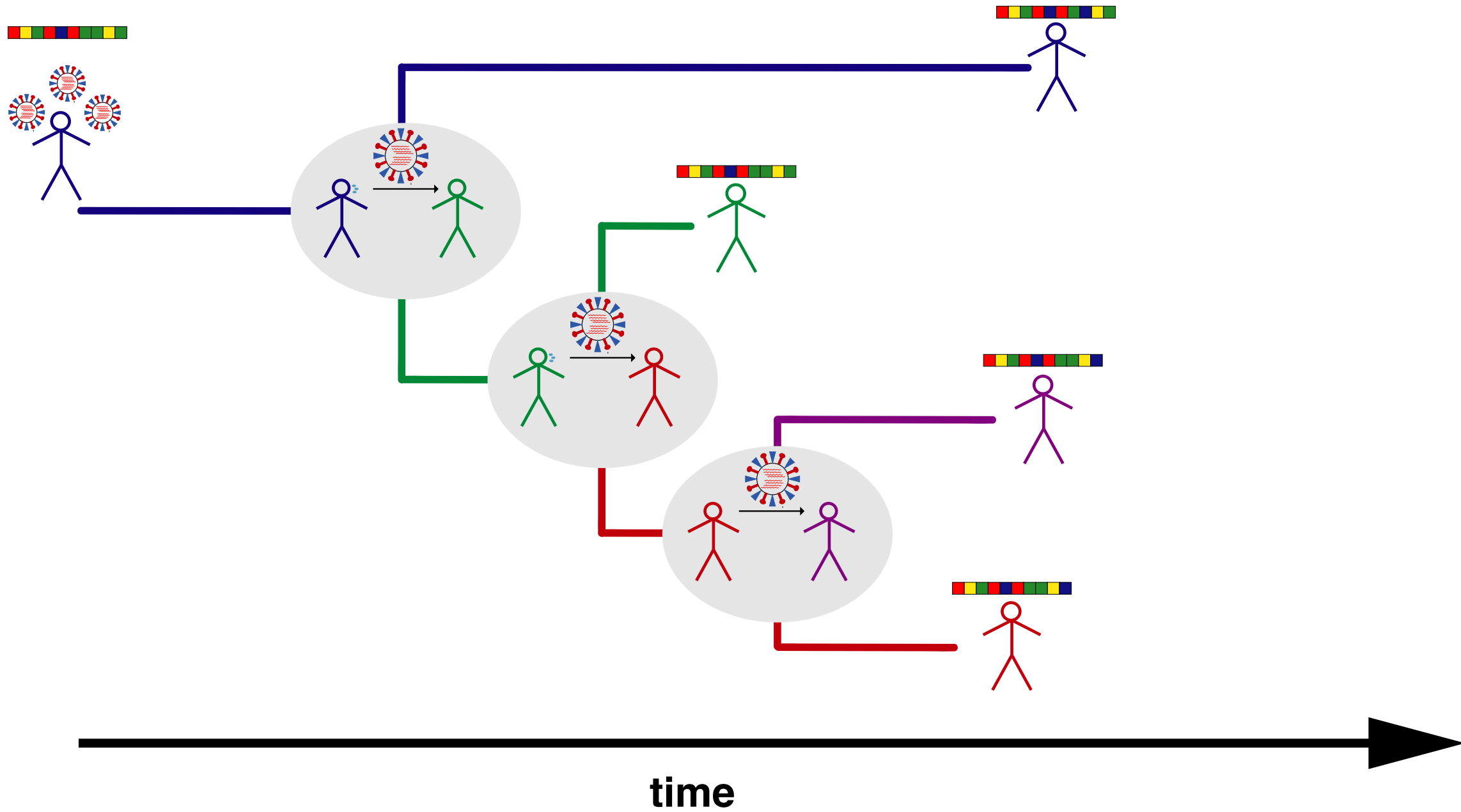


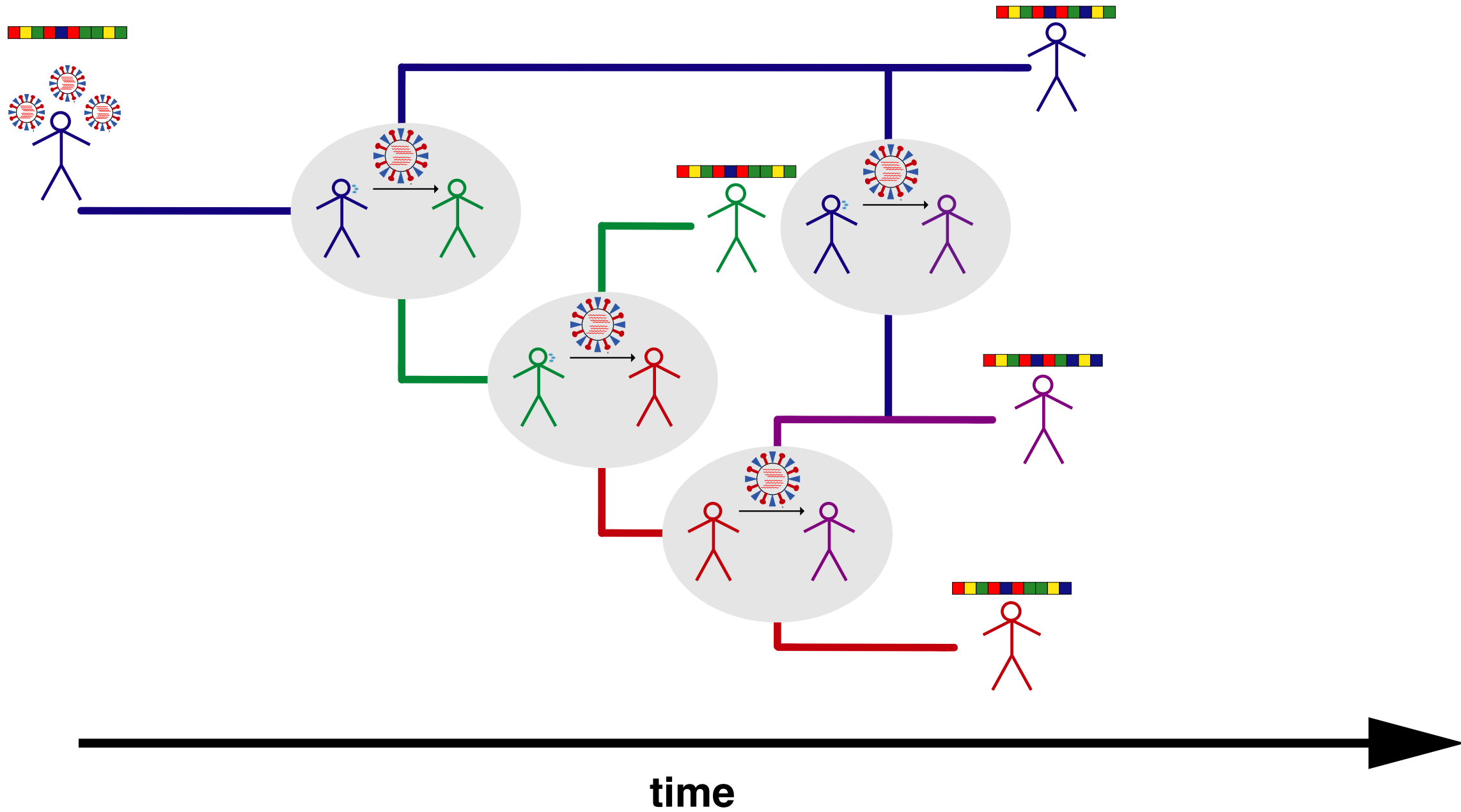
time





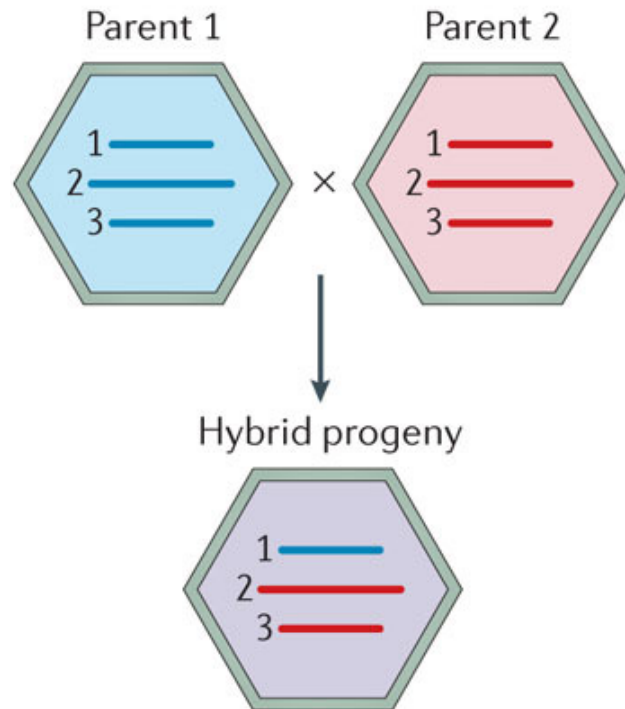
time



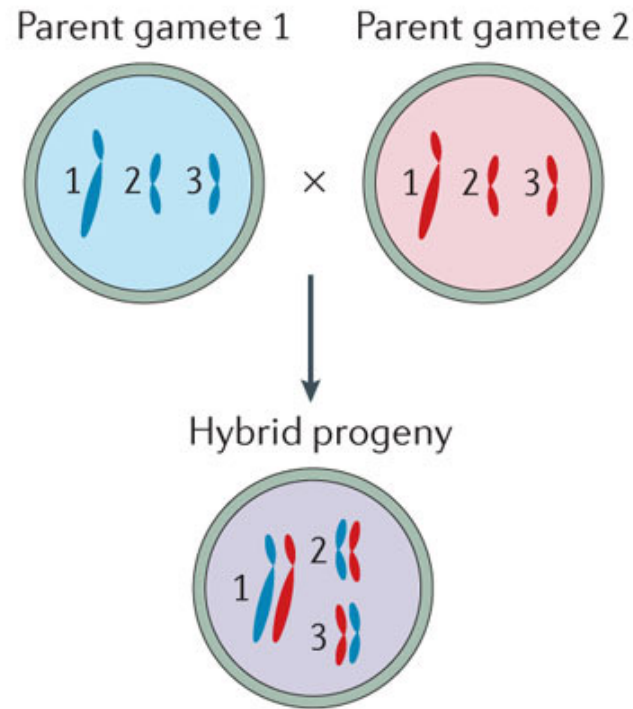


Different processes to “mix” genetic materials from two parent lineages exist

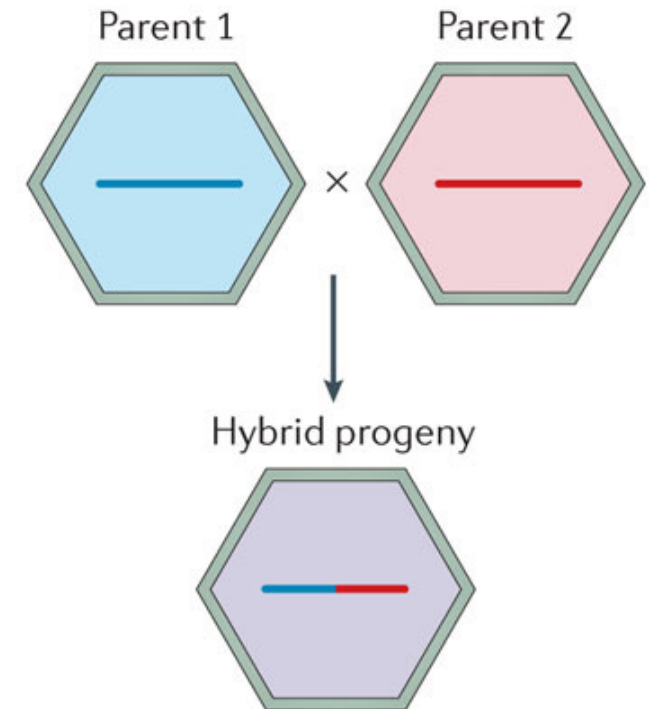
a Reassortment



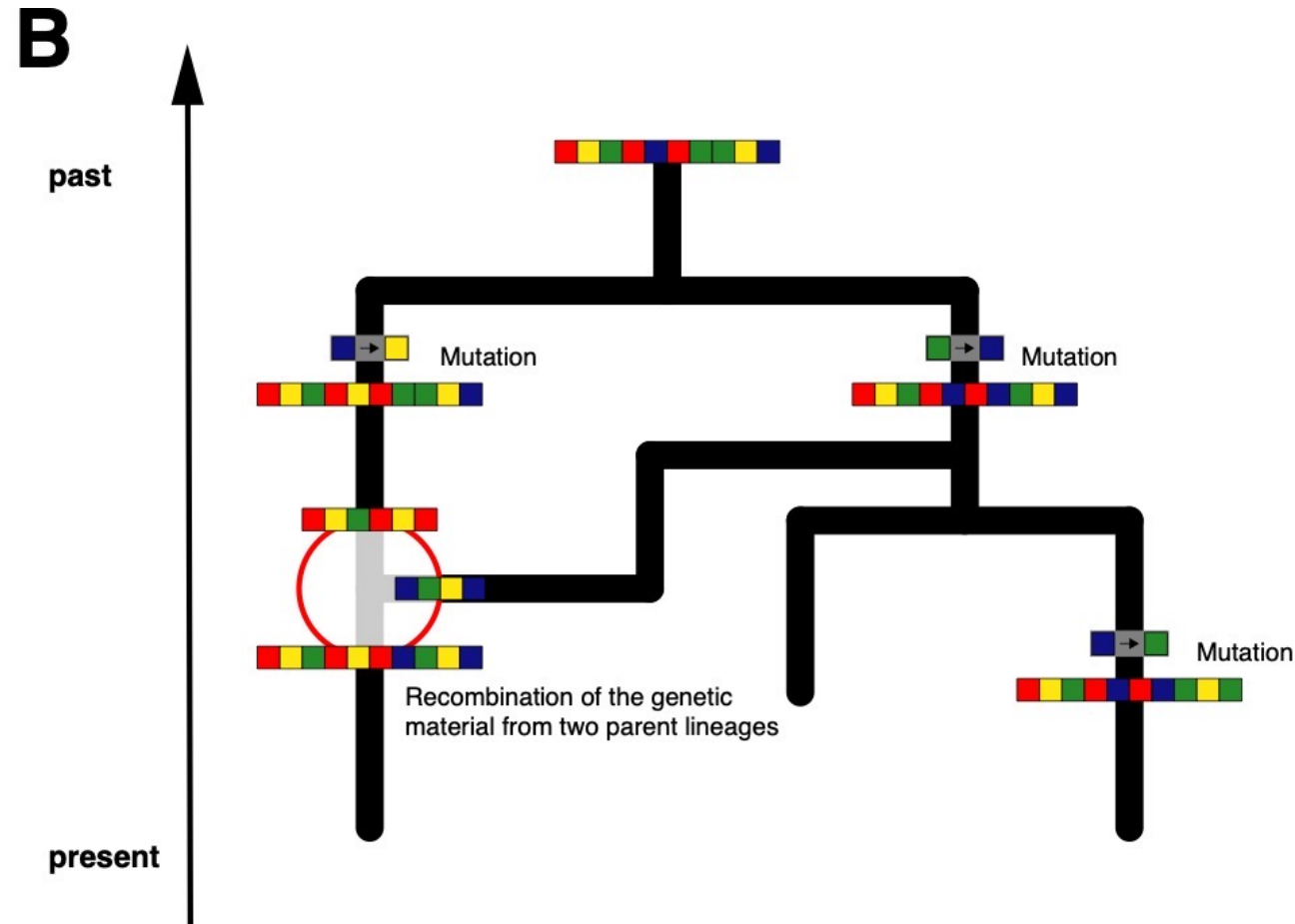
b Sexual reproduction



c Recombination

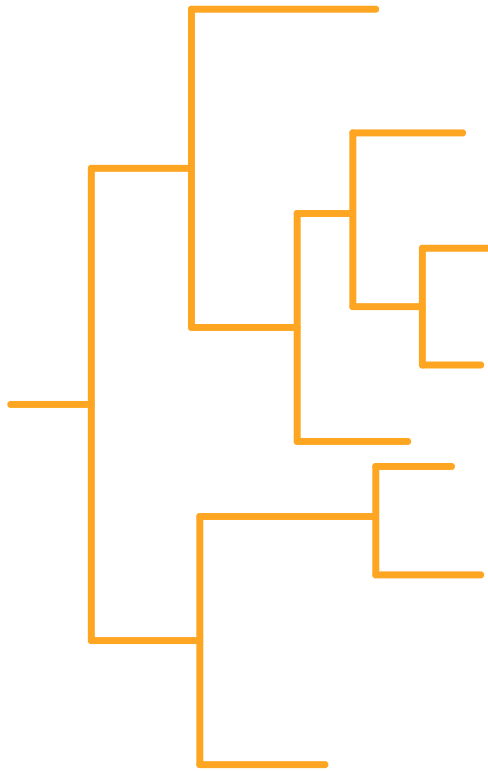


The combined genetic material may represent a network and not a tree

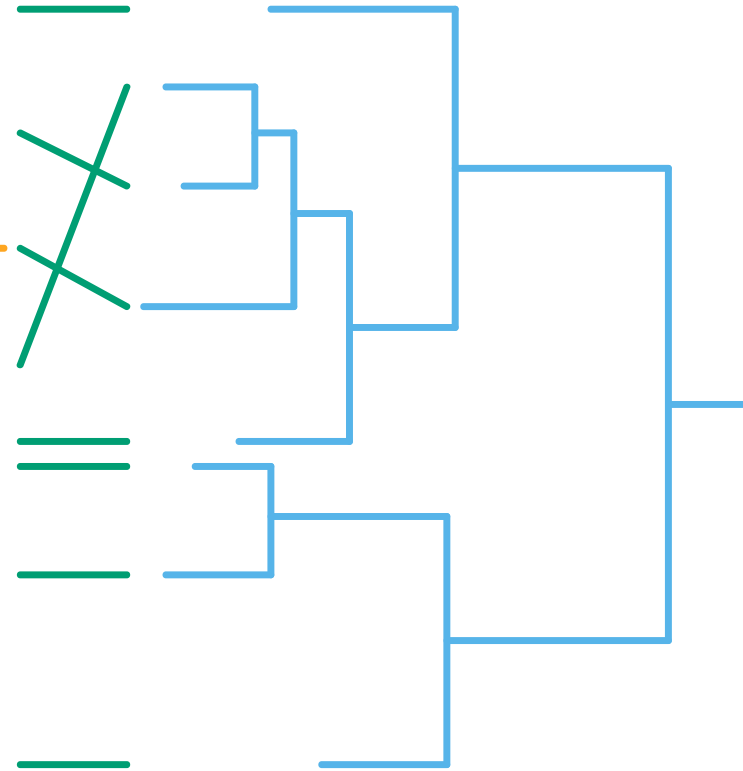


Genetic recombination processes lead to different parts of a genome coding for different histories

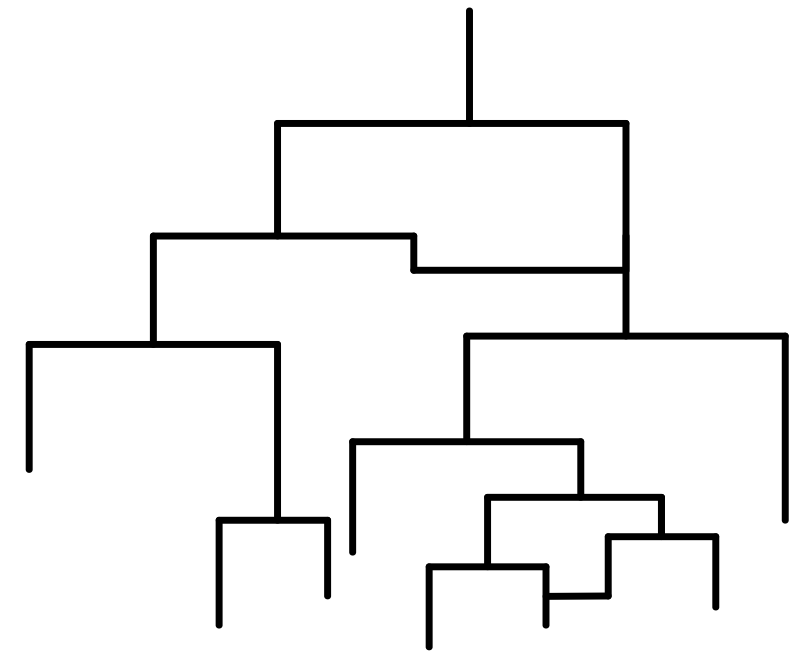
Tree 1



Tree 2



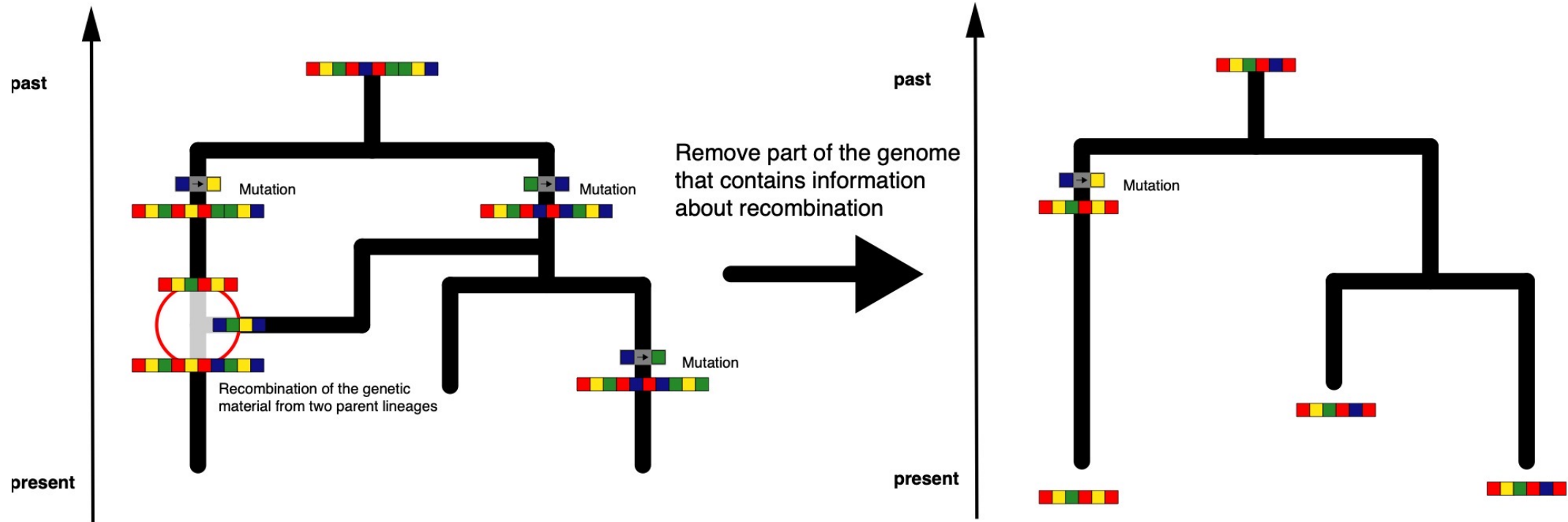
Joint history



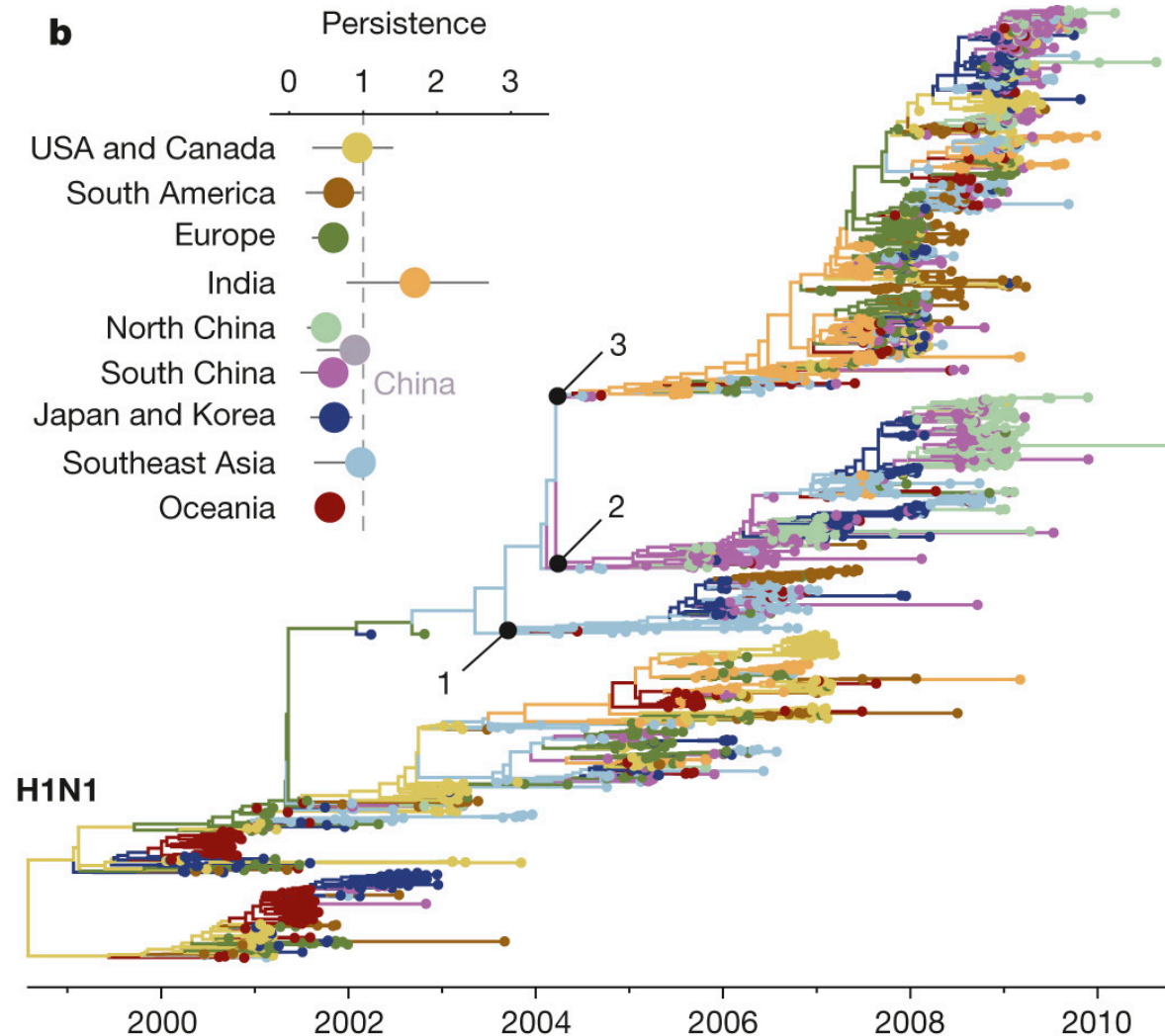
Ways to deal with recombination

- Only use small parts of the genome that code for the same tree (e.g. only use one segment of influenza).
- Consider different parts of a genome to be independent (Bad if they are not).
- Just ignore recombination (bad if there is).
- Infer networks instead of trees. (Best case, but potentially slow)

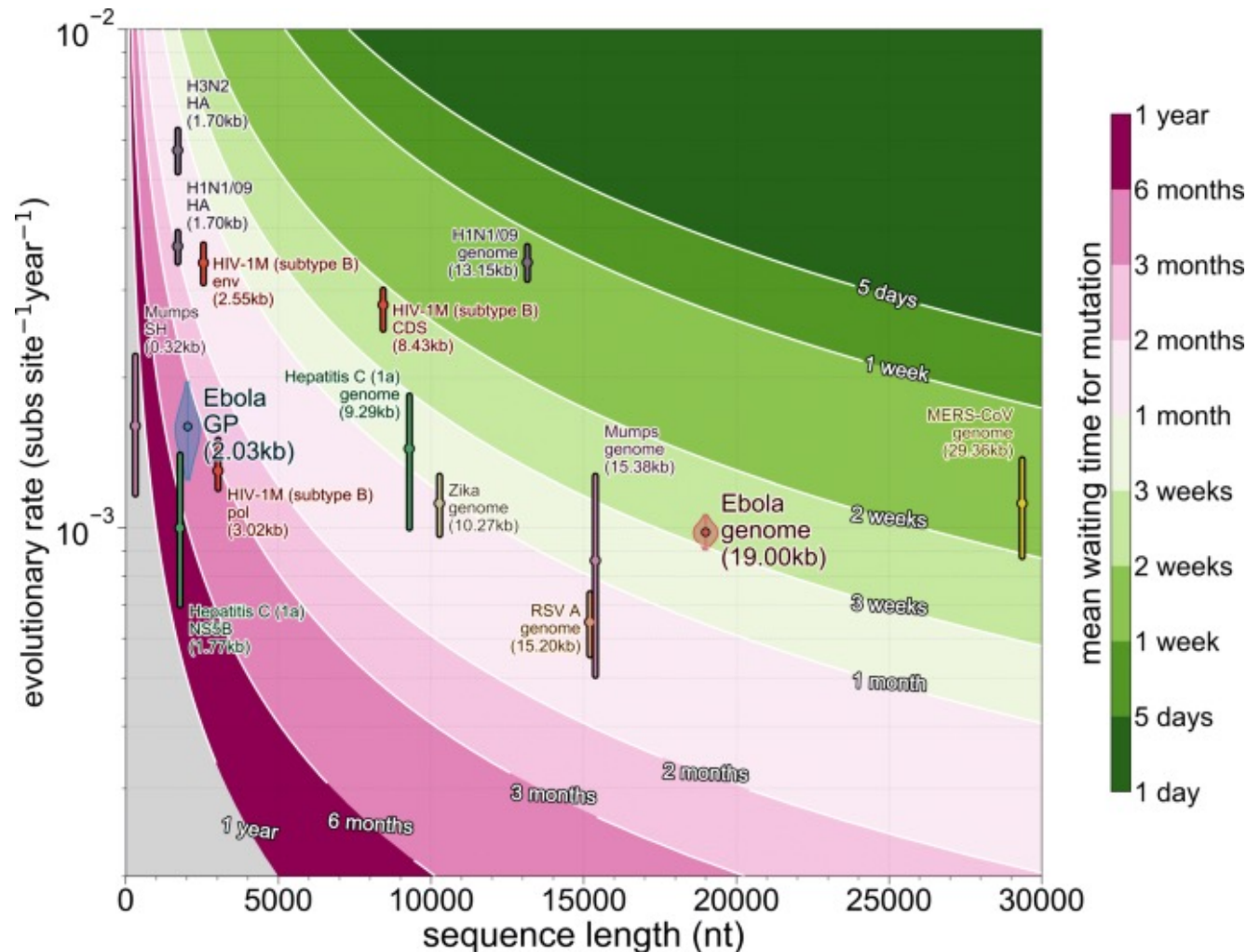
Only use small parts of the genome that code for the same tree.



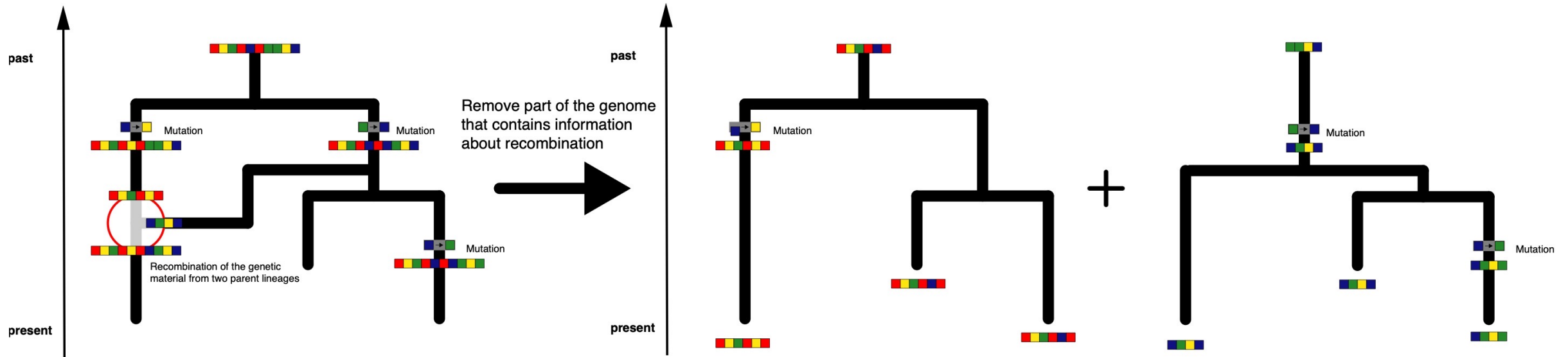
Population dynamics can be inferred from individual influenza segments



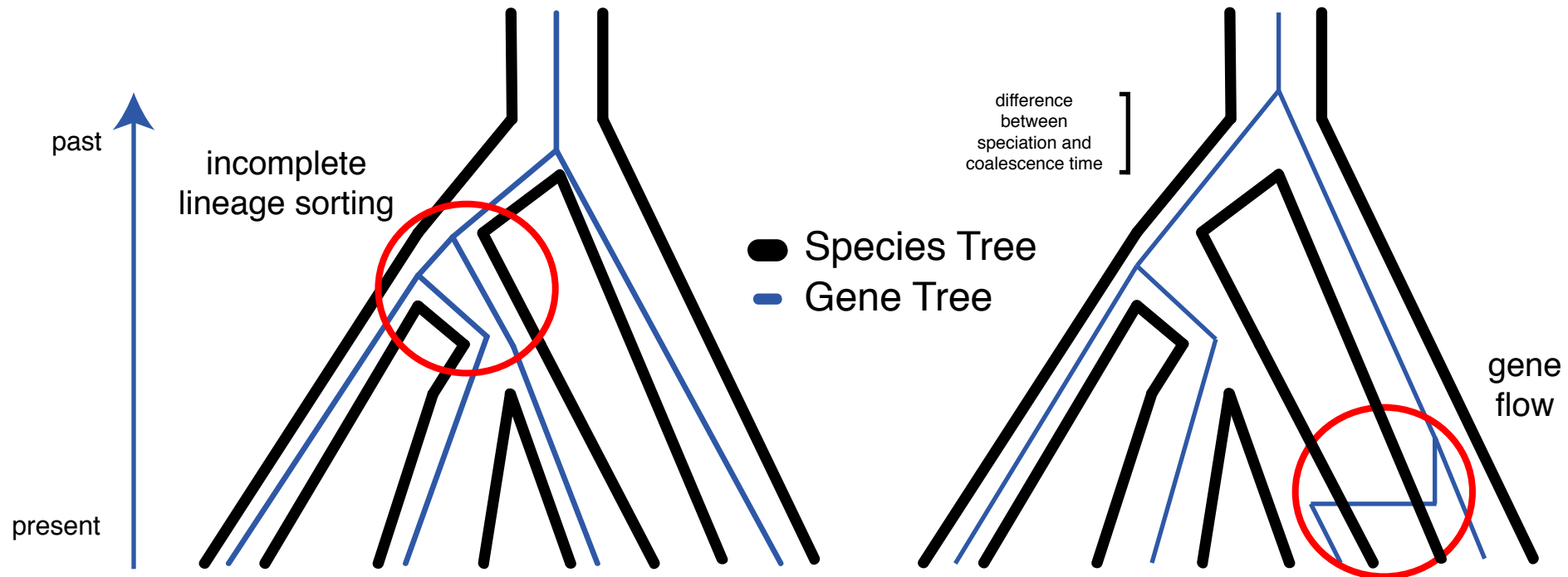
Using only part of pathogen genomes reduces time resolution



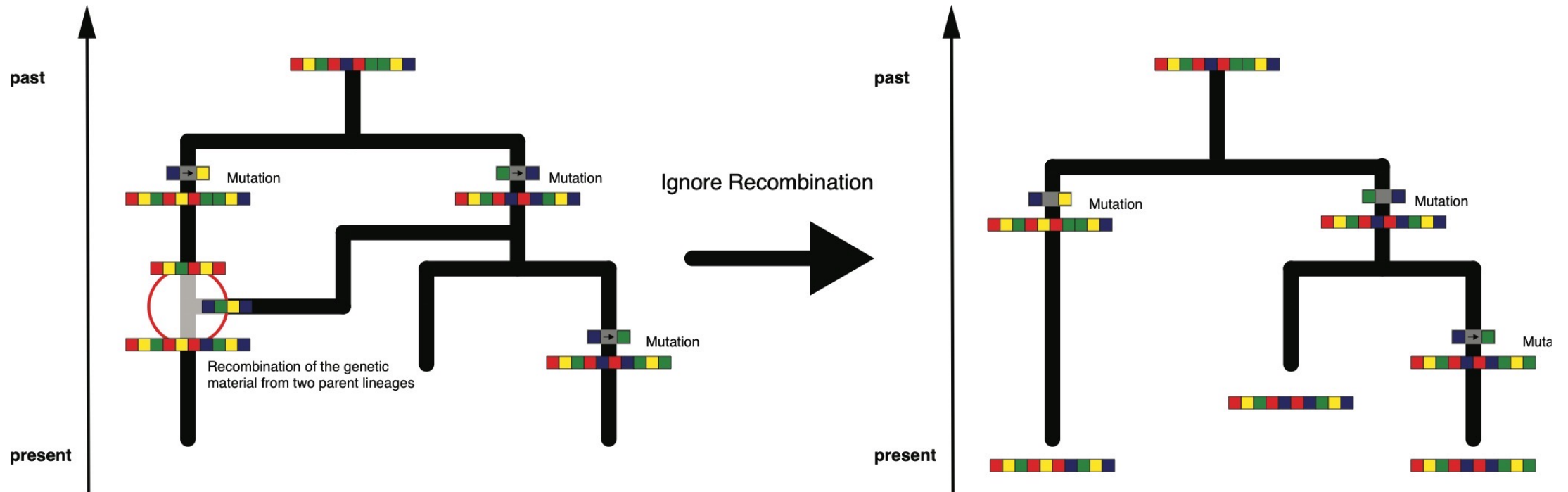
Consider different parts of a genome to be independent. (Sometimes correct)



Species tree inferences typically assume individual parts of the genome to be independent observations of a speciation process



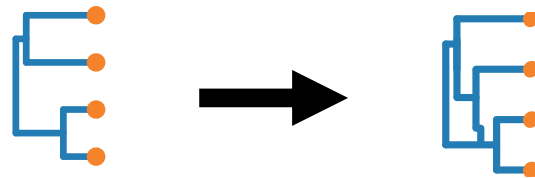
Just ignore recombination. (Always wrong)



QUESTIONS?

To perform inference of phylogenetic networks, we have to introduce a new inference approach

$$P(\text{Tree} \quad \text{Matrix} \quad \text{Clock} \quad \text{Rearrangement} \mid \text{Sequence}) = \frac{P(\text{Sequence} \mid \text{Tree} \quad \text{Matrix} \quad \text{Clock}) \quad P(\text{Tree} \mid \text{Rearrangement}) \quad P(\text{Rearrangement}) \quad P(\text{Matrix}) \quad P(\text{Clock})}{P(\text{Sequence})}$$



$$P(\text{Tree} \quad \text{Matrix} \quad \text{Clock} \quad \text{Rearrangement} \mid \text{Sequence}) = \frac{P(\text{Sequence} \mid \text{Tree} \quad \text{Matrix} \quad \text{Clock}) \quad P(\text{Tree} \mid \text{Rearrangement}) \quad P(\text{Rearrangement}) \quad P(\text{Matrix}) \quad P(\text{Clock})}{P(\text{Sequence})}$$

To infer phylogenetic networks, we need the following

- A way to explore the posterior probability $P(\text{🌳} \text{📊} \text{🕒} \text{🔄} \mid \text{ACAC...} \text{TCAC...} \text{ACAG...})$

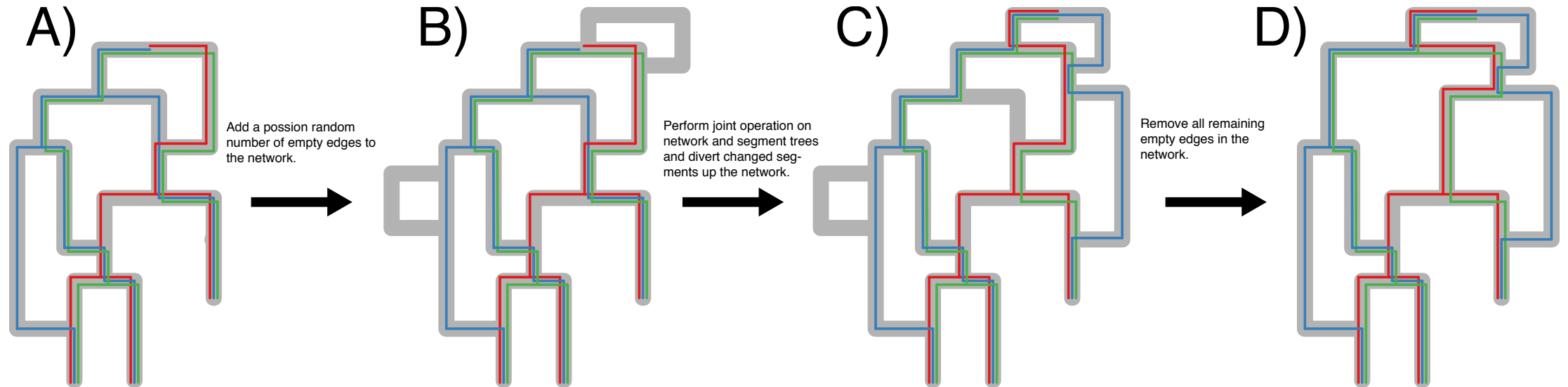
To infer phylogenetic networks, we need the following

- A way to explore the posterior probability $P(\text{🌳} \text{📊} \text{🕒} \text{🔄} \mid \text{ACAC...} \text{TCAC...} \text{ACAG...})$
- The **network likelihood** $P(\text{ACAC...} \text{TCAC...} \text{ACAG...} \mid \text{🌳} \text{📊} \text{🕒})$ can be expressed as a product of tree likelihoods on the different nucleotide positions

To infer phylogenetic networks, we need the following

- A way to explore the posterior probability $P(\text{tree icon} \text{ grid icon} \text{ clock icon} \text{ network icon} \mid \text{sequences icon})$
- The **network likelihood** $P(\text{sequences icon} \mid \text{tree icon} \text{ grid icon} \text{ clock icon})$ can be expressed as a product of tree likelihoods on the different nucleotide positions
- The **network prior** $P(\text{tree icon} \mid \text{network icon})$ requires a network generating model, such as a coalescent process

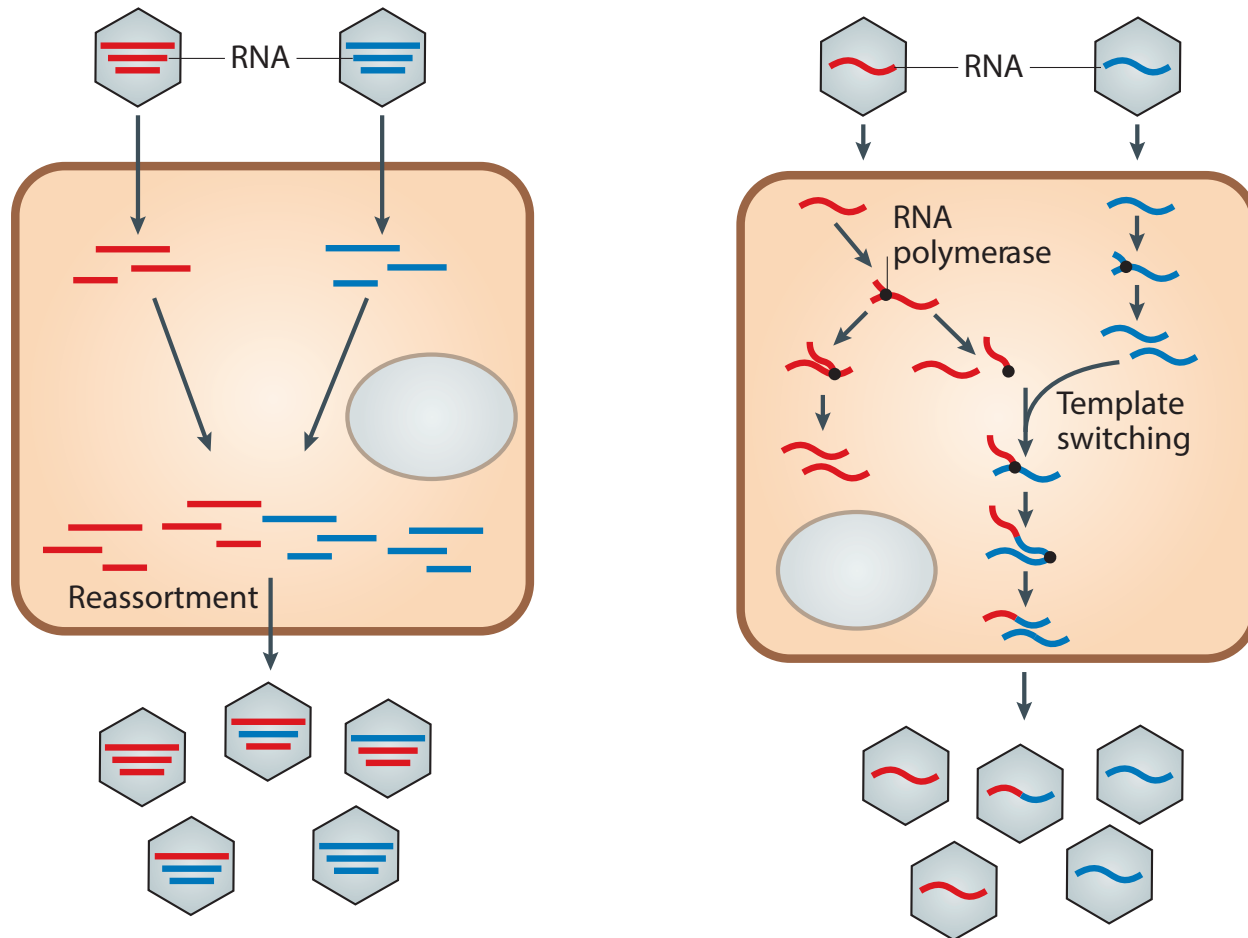
To explore $P(\text{[tree icon] [grid icon] [clock icon] [loop icon] | \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix}})$ of phylogenetic networks,
we need to “operate” on the networks



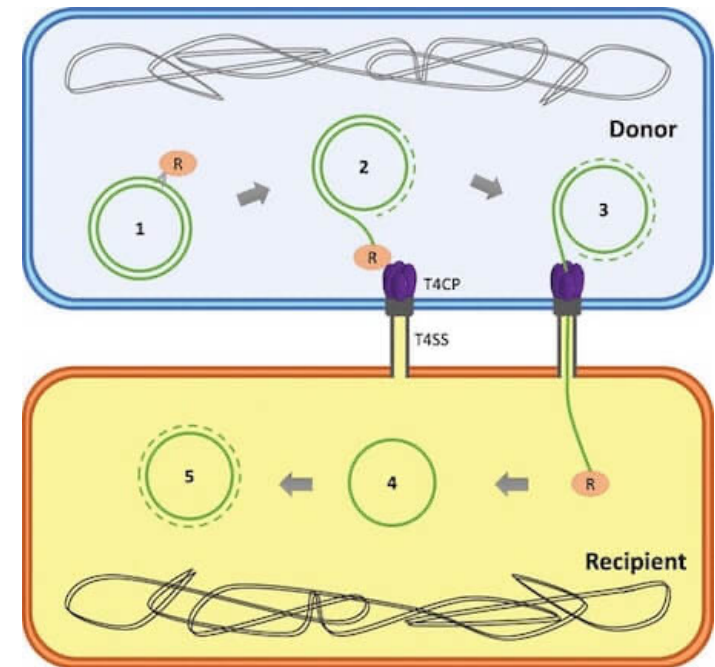
$P(\text{ACAC...} \mid \text{TCAC...} \text{ACAG...})$ can be expressed as the product of tree likelihoods

- The probability of observing each position in this alignment only requires knowing the tree at this position.
- Therefore $P(\text{ACAC...} \mid \text{TCAC...} \text{ACAG...}) = \prod^{sites} P(\text{ACAC...} \mid \text{TCAC...} \text{ACAG...})$

$P(\text{ } | \text{ })$ models as a coalescent and reassortment/recombination/plasmid transfer process

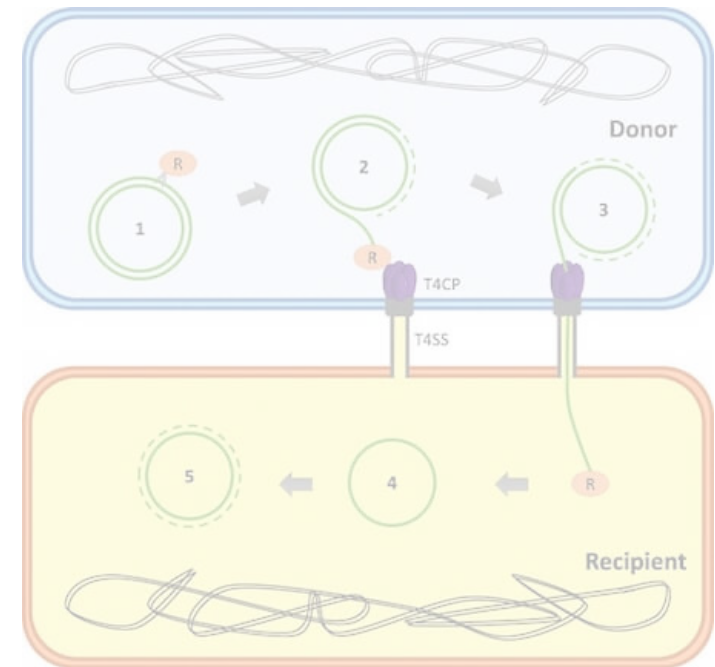
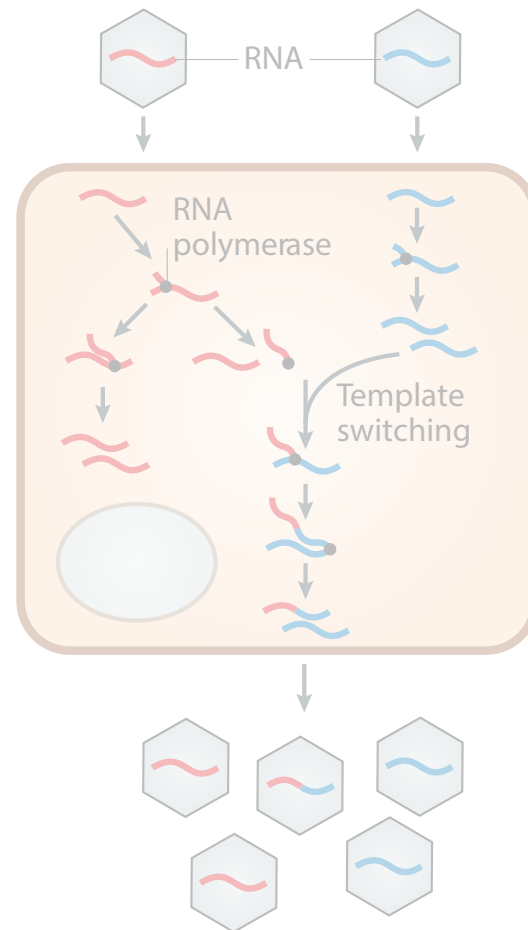
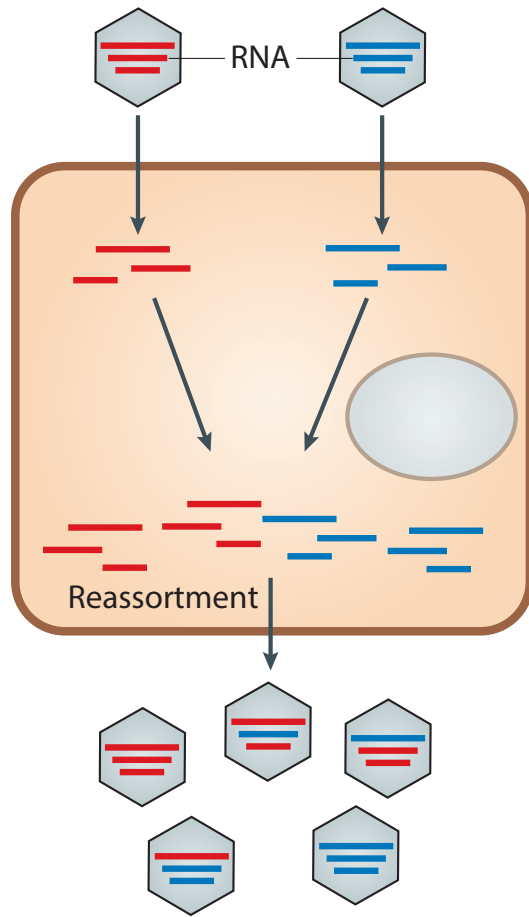


E. Simon-Loriere, E. C. Holmes “*Why do RNA viruses recombine?*” *Nature Reviews Microbiology*, 2011



Getting et al./Microbiology Spectrum, Jan. 2018

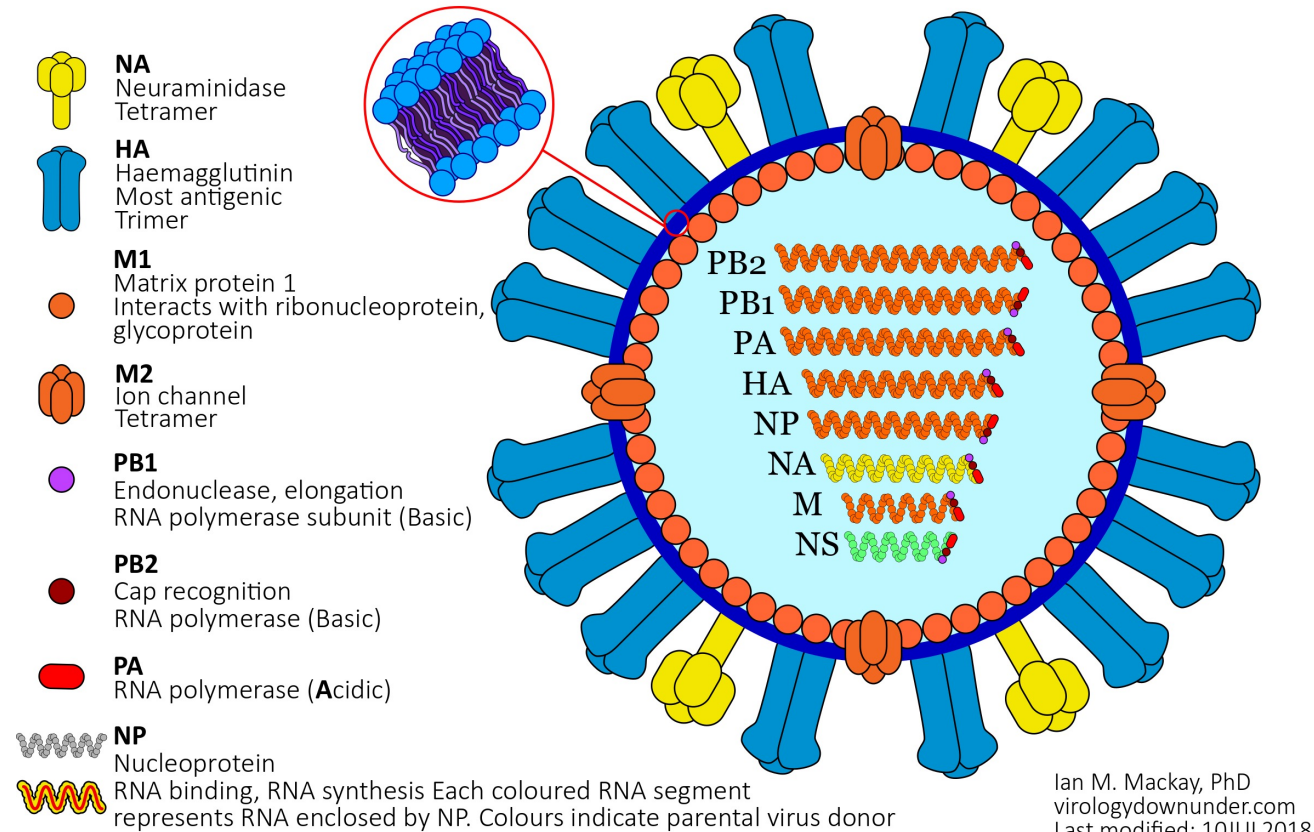
$P(\text{ } | \text{ })$ models as a coalescent and reassortment/recombination/plasmid transfer process



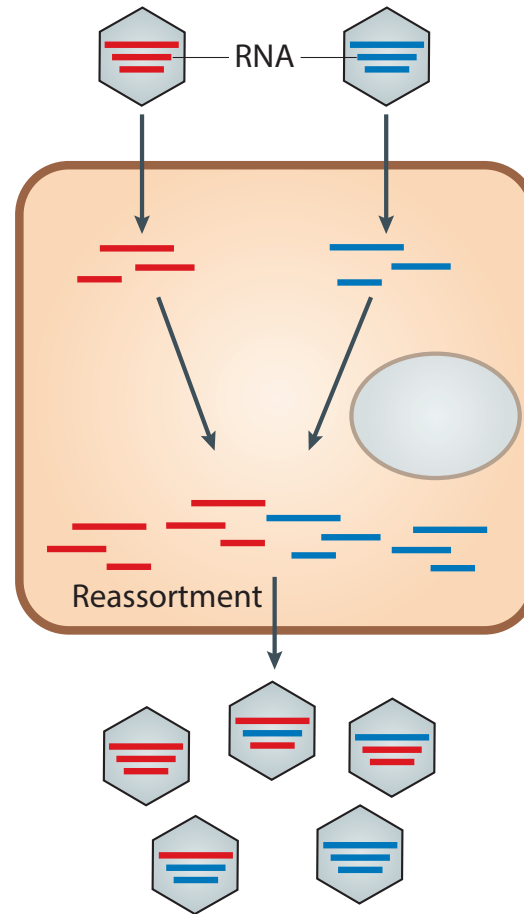
E. Simon-Loriere, E. C. Holmes “*Why do RNA viruses recombine?*” *Nature Reviews Microbiology*, 2011

Getting et al./*Microbiology Spectrum*, Jan. 2018

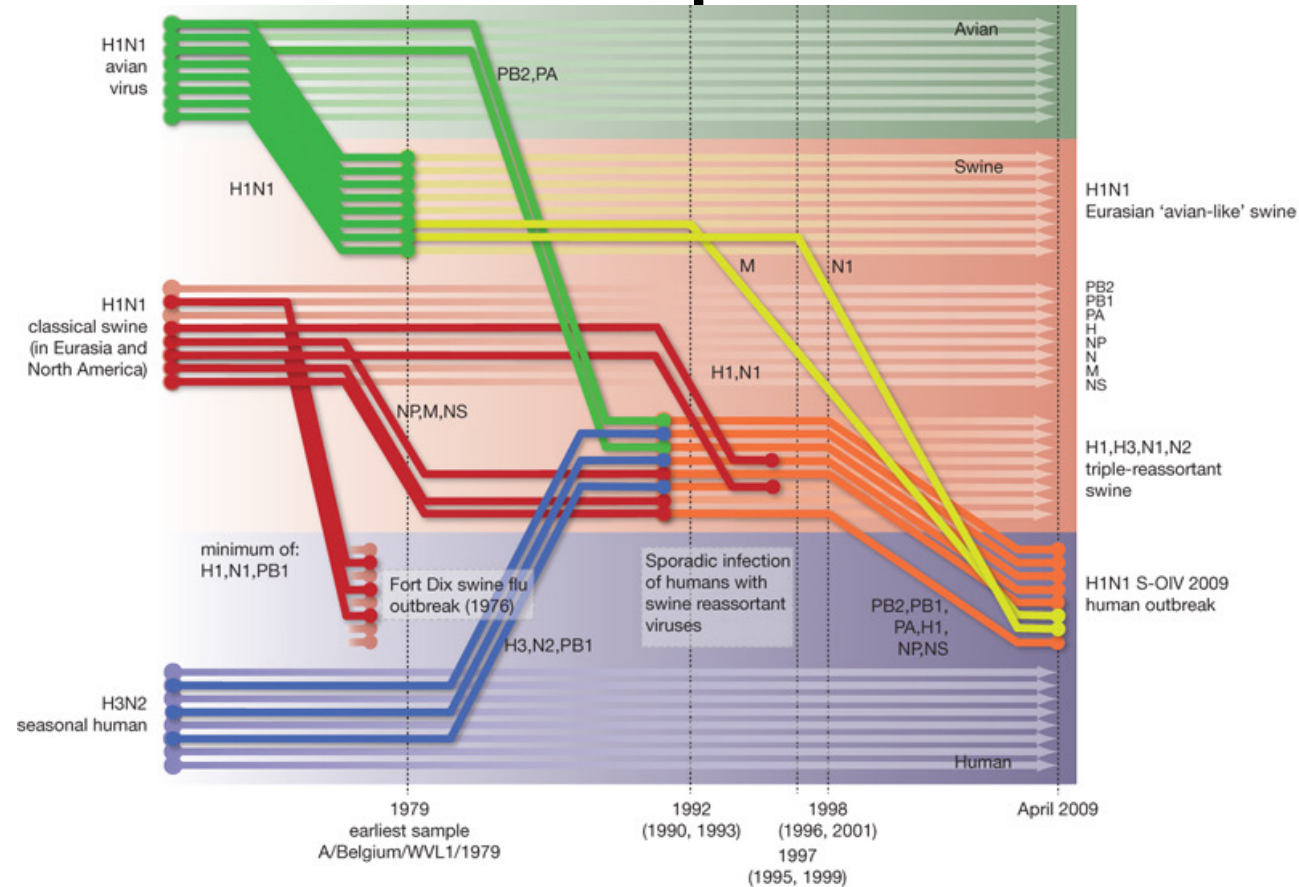
The genome of Influenza is organized in several separated segments



Reassortment leads to a reshuffling of segments from different ancestral lineages



Reassortment between subtypes can create progenitors with segments that originate from different parents



Smith, G., Vijaykrishna, D., Bahl, J. et al "Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic" Nature, 2009

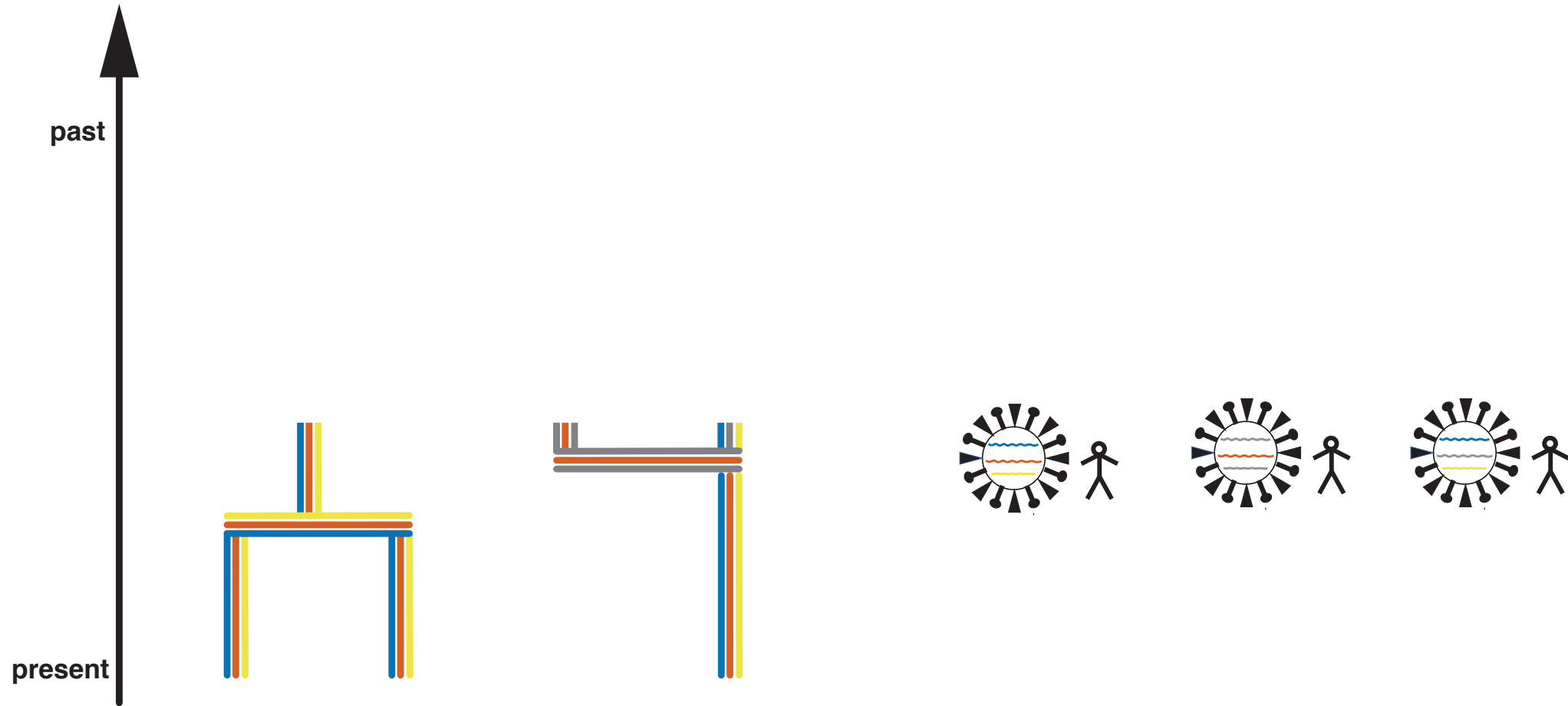
Three samples of a hypothetical influenza virus with three segments



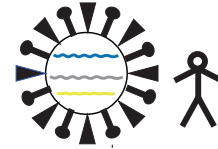
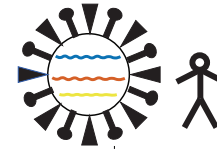
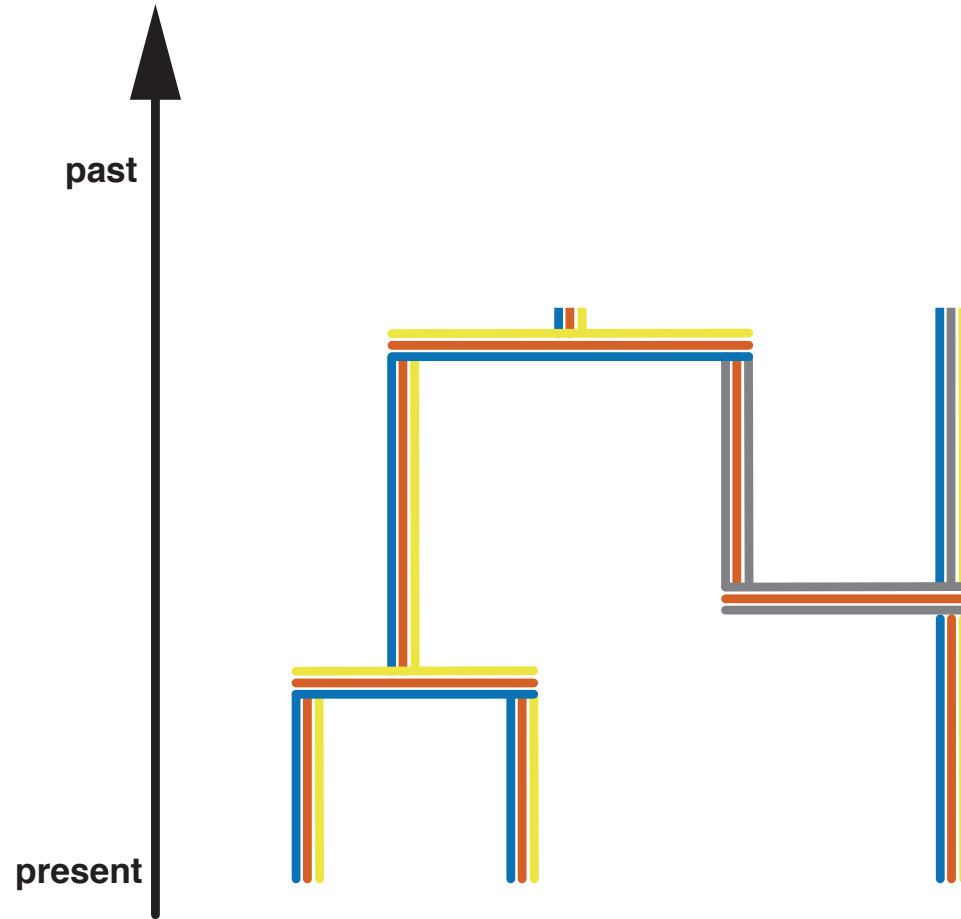
Going back in time, two lineages can share a common ancestor at a rate inverse proportional to N_e



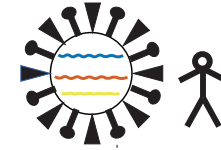
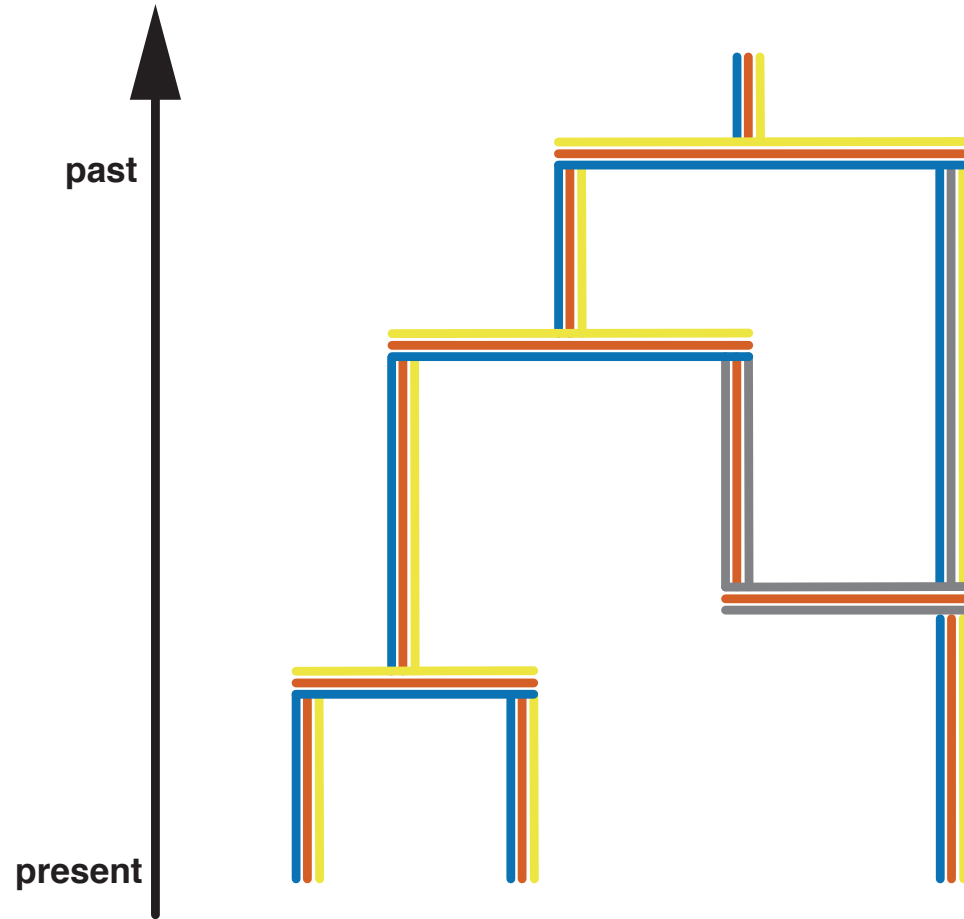
A lineage can reassort at a rate given by the reassortment rate with each segment having originated from one parental lineage



More lineage share a common ancestor



Until all lineage coalescence at the root of the
phylogenetic network



The coalescent with reassortment models a joint coalescent and reassortment process

- Coalescent events between any n coexisting network lineages happen at:

$$\frac{n(n-1)}{2Ne}$$

- Reassortment events that leave a genetic footprint on lineage i happen at a rate of:

$$\rho(1 - 2 * 0.5^{c_i})$$

The reassortment rate is a backwards in time rate of observing reassortment events

- The reassortment rate is a function of:
 - The probability that co-infection occurs
 - The probability of reassortment in a co-infected individual
 - The success/selection of reassortants

Bayesian inference allows us to infer the posterior distributions of:



Reassortment networks

Bayesian inference allows us to infer the posterior distributions of:



Reassortment networks



The embedding of segment trees in those networks

Bayesian inference allows us to infer the posterior distributions of:



Reassortment networks



The embedding of segment trees in those networks



The parameters of evolution of the different segments

Bayesian inference allows us to infer the posterior distributions of:



Reassortment networks



The embedding of segment trees in those networks

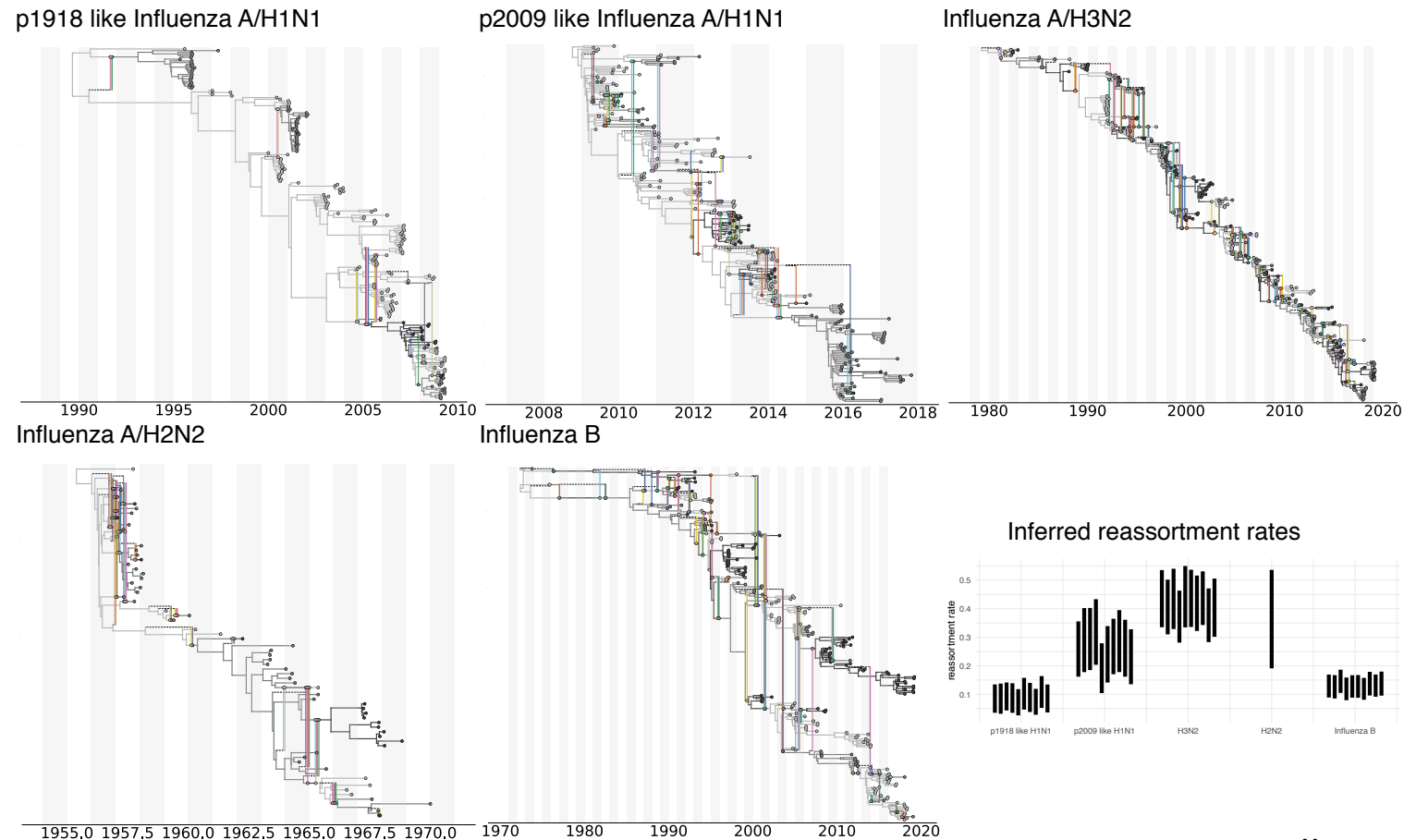


The parameters of evolution of the different segments



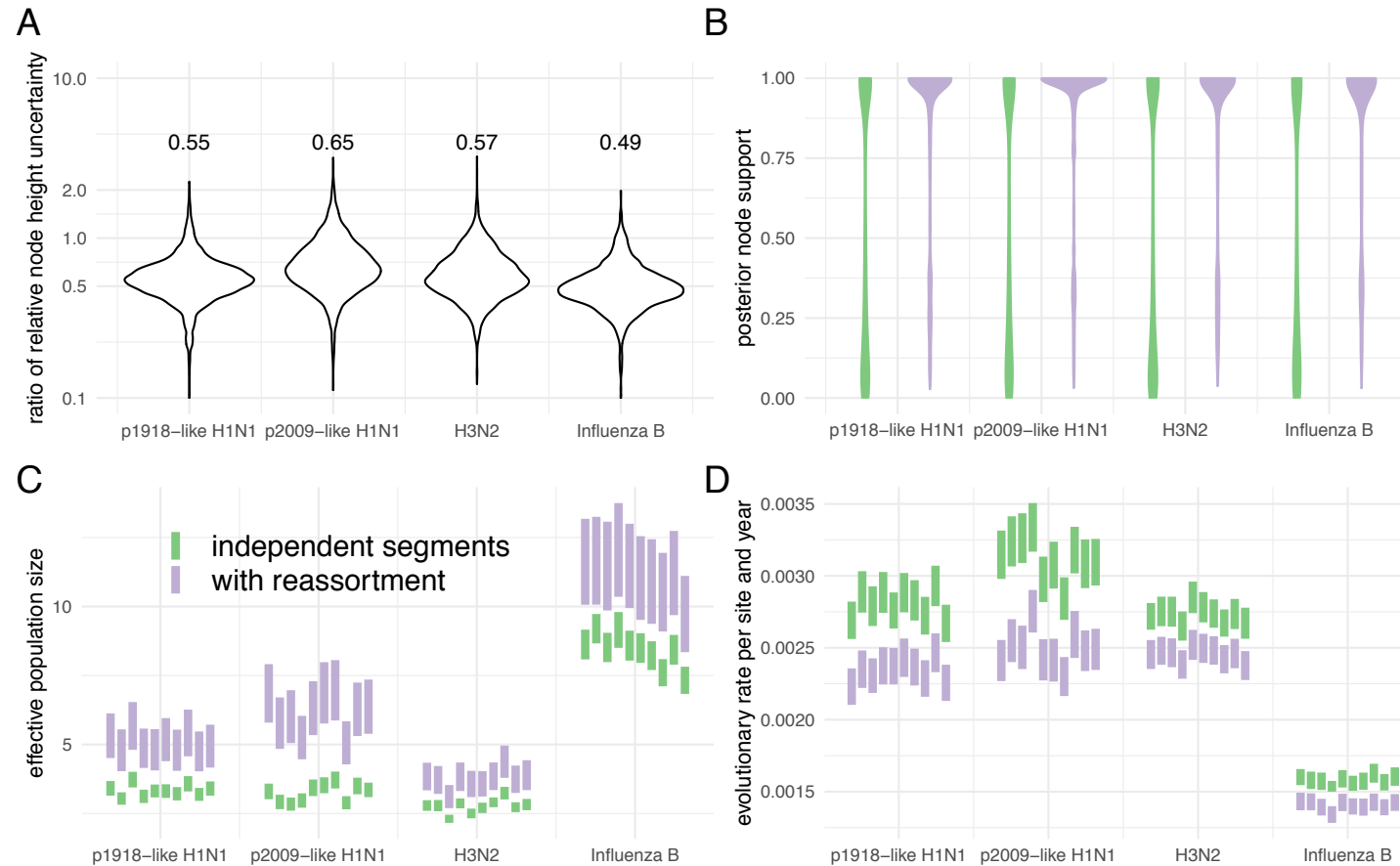
Demographic parameters and rates of reassortment

CoalRe allows inferring reassortment networks, reassortment rates etc. jointly

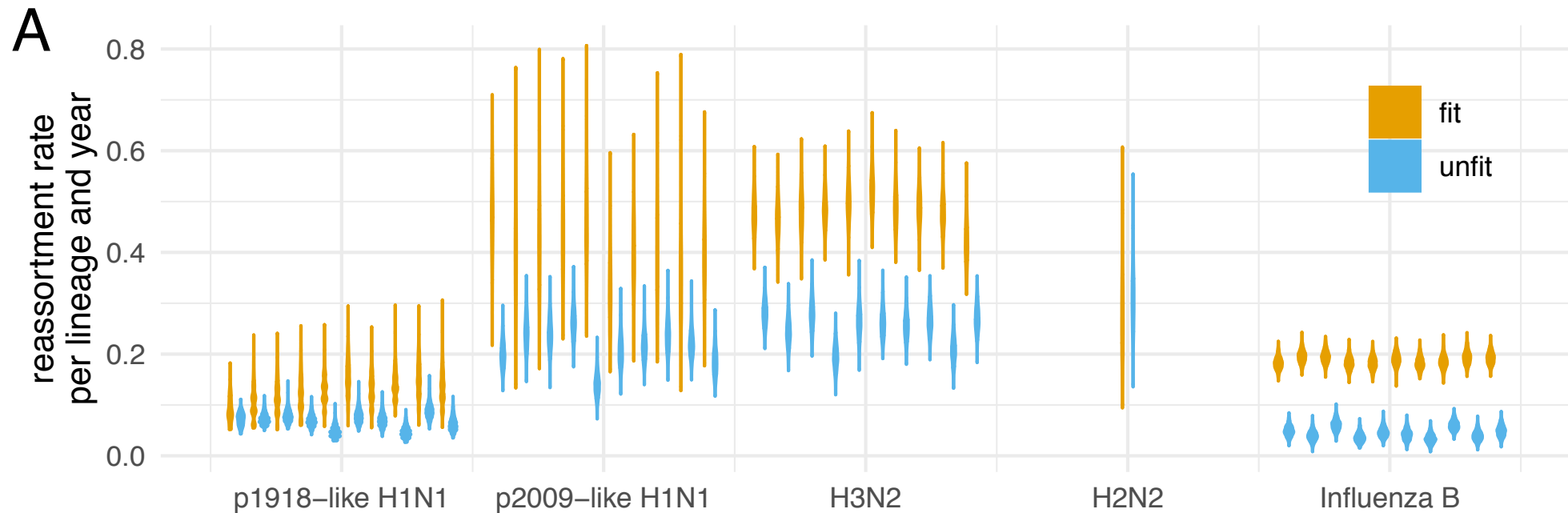


Müller et al. (2020), *PNAS*

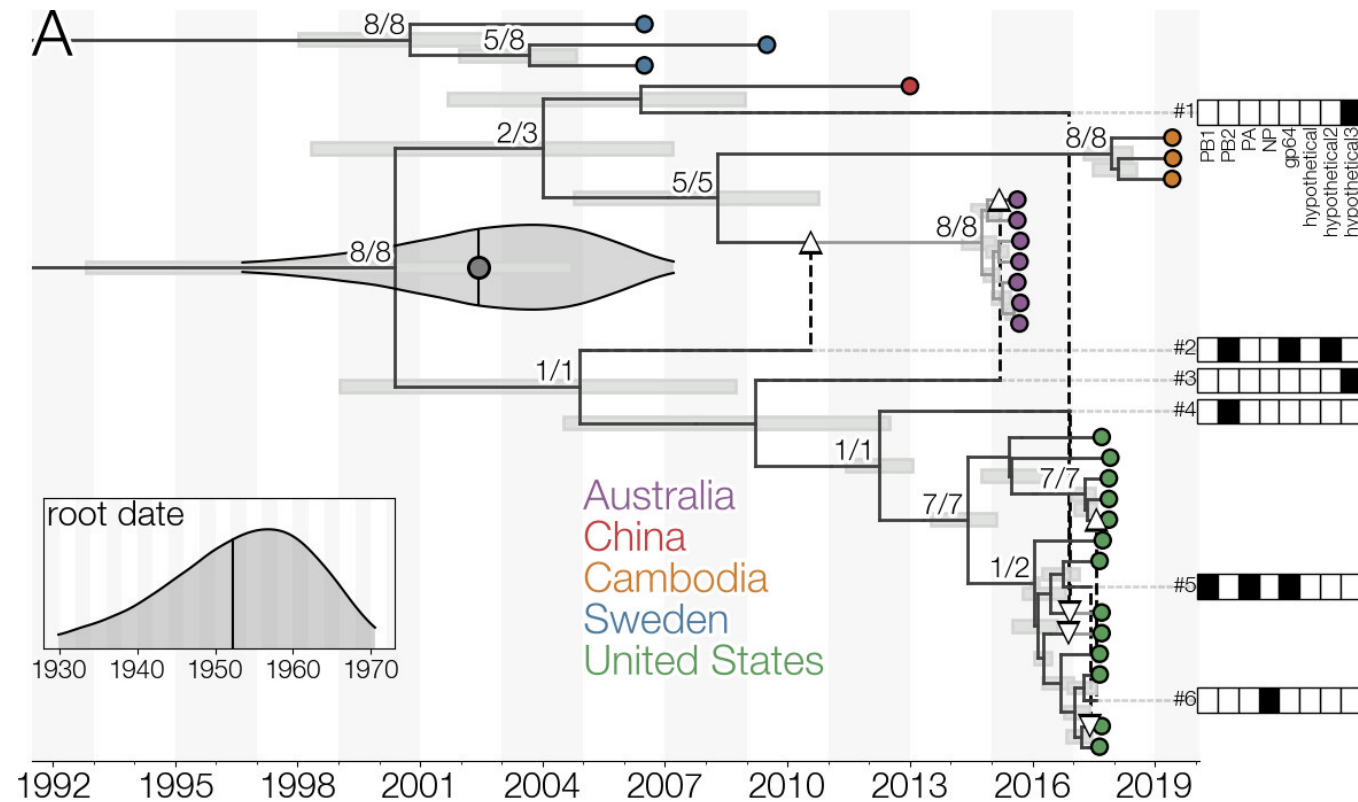
Correctly modelling the reassortment process impact precision and reduces bias



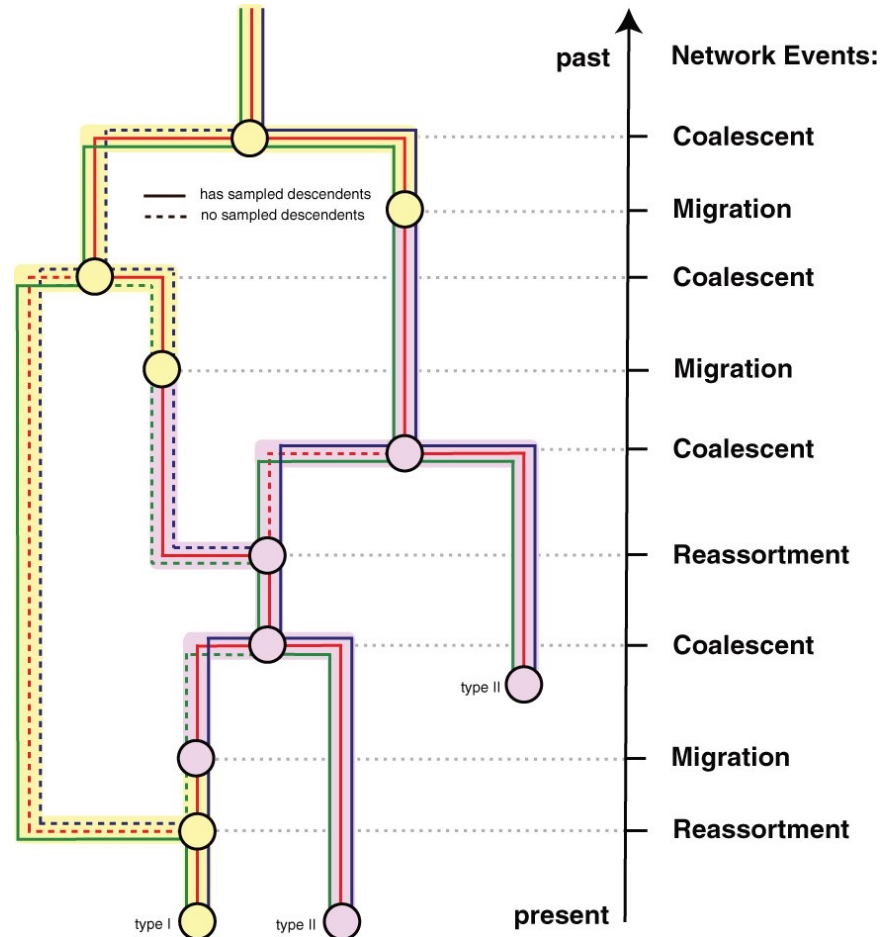
Explicitly modeling when and where reassortment events occurred allows us to investigate whether there are patterns



CoalRe to track the movement and evolution of orthomyxoviruses

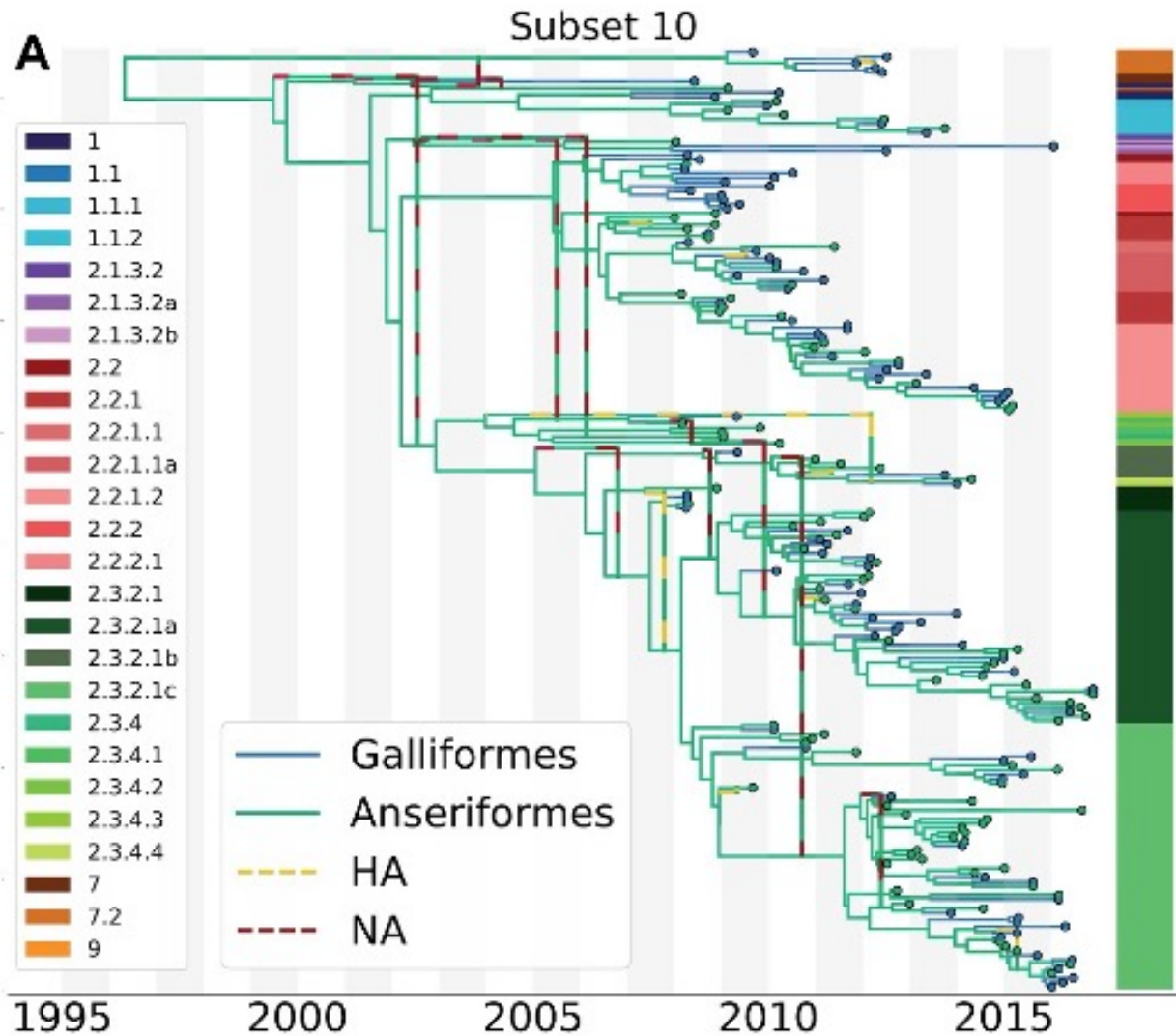


Reassortment and population structure can be reconstructed jointly



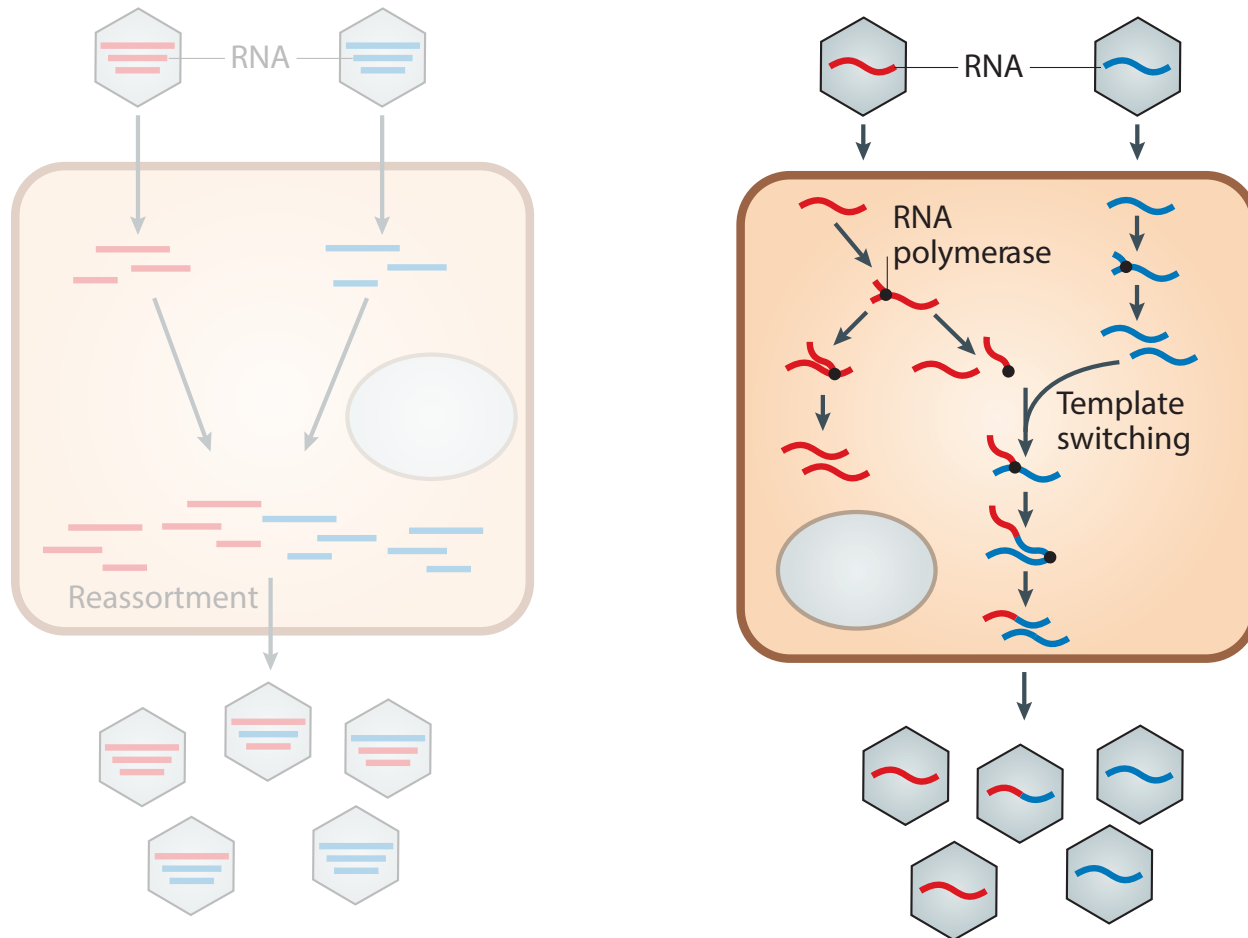
Stolz et al. (2021), Mol. Biol. Evol.

The structured coalescent with reassortment can be used to reconstruct where and when reassortment occurred.

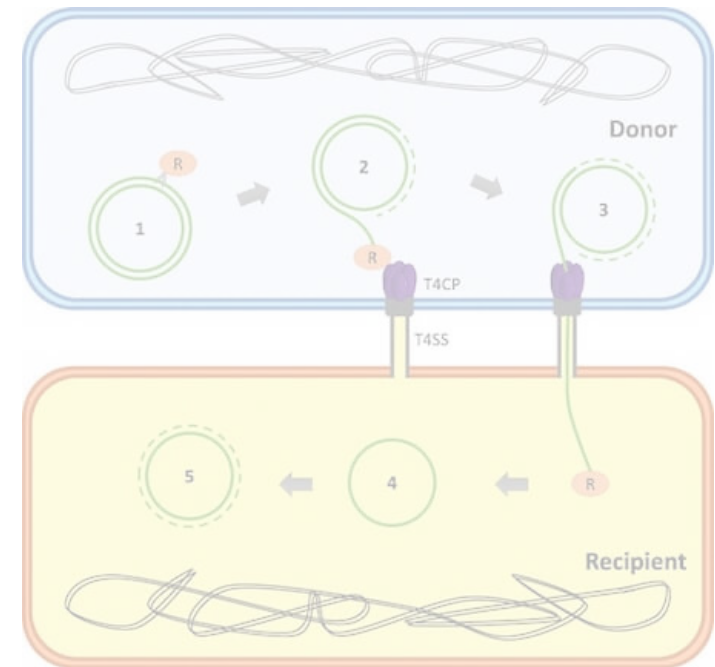


Stolz et al. (2021), Mol. Biol. Evol.

$P(\text{ } | \text{ })$ models as a coalescent and reassortment/recombination/plasmid transfer process

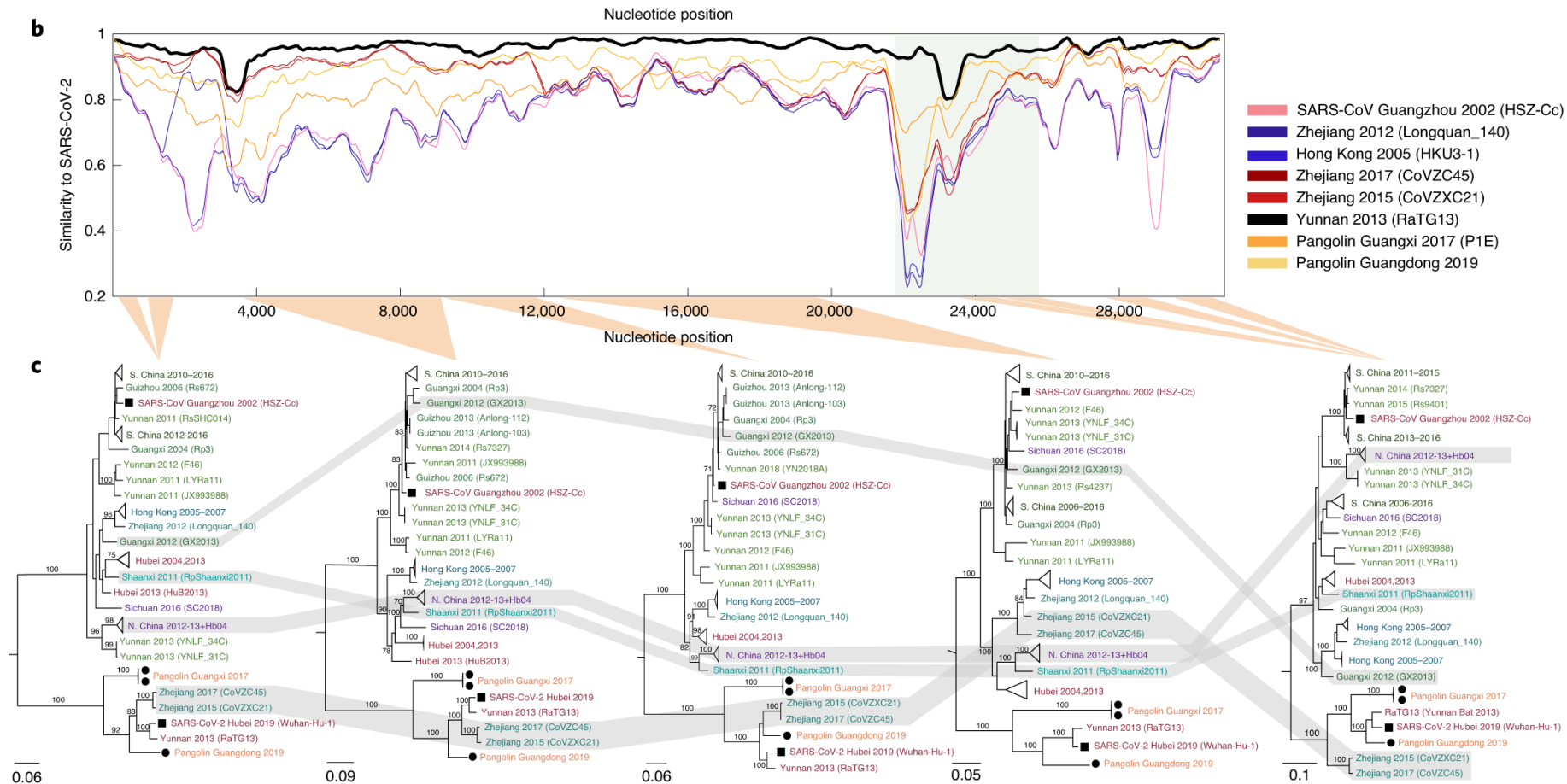


E. Simon-Loriere, E. C. Holmes “*Why do RNA viruses recombine?*” *Nature Reviews Microbiology*, 2011



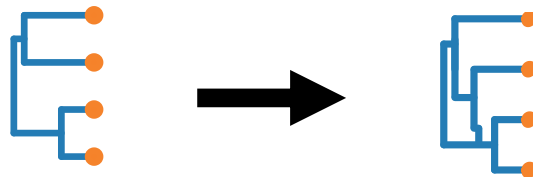
Getting et al./Microbiology Spectrum, Jan. 2018

Different parts of the genome of SARS-like viruses code for different "trees"



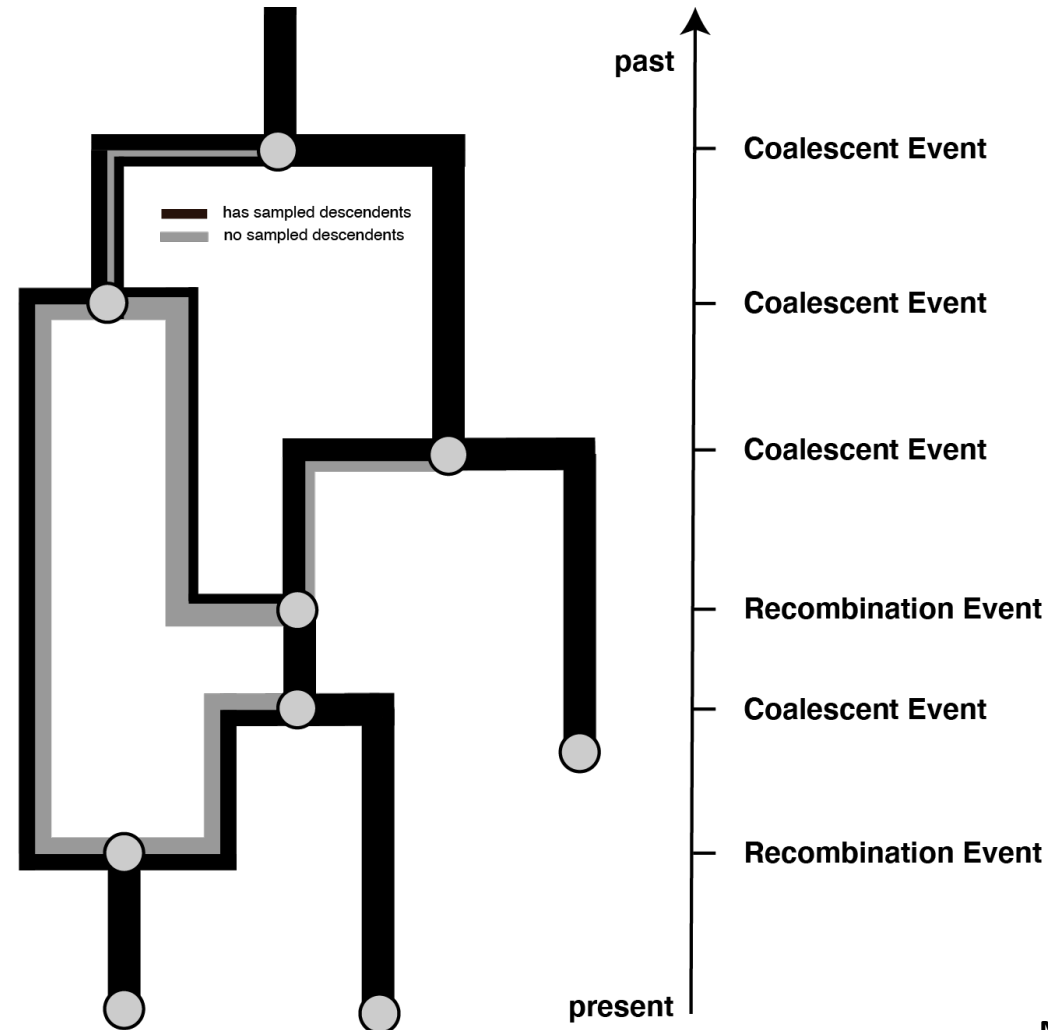
To switch to networks, we conceptually just replace the tree terms with networks

$$P(\text{[tree]} \text{ [grid]} \text{ [clock]} \text{ [loop]} \mid \text{ACAC... TCAC... ACAG...}) = \frac{P(\text{ACAC... TCAC... ACAG...} \mid \text{[tree]} \text{ [grid]} \text{ [clock]}) P(\text{[tree]} \mid \text{[loop]}) P(\text{[loop]}) P(\text{[grid]}) P(\text{[clock]})}{P(\text{ACAC... TCAC... ACAG...})}$$



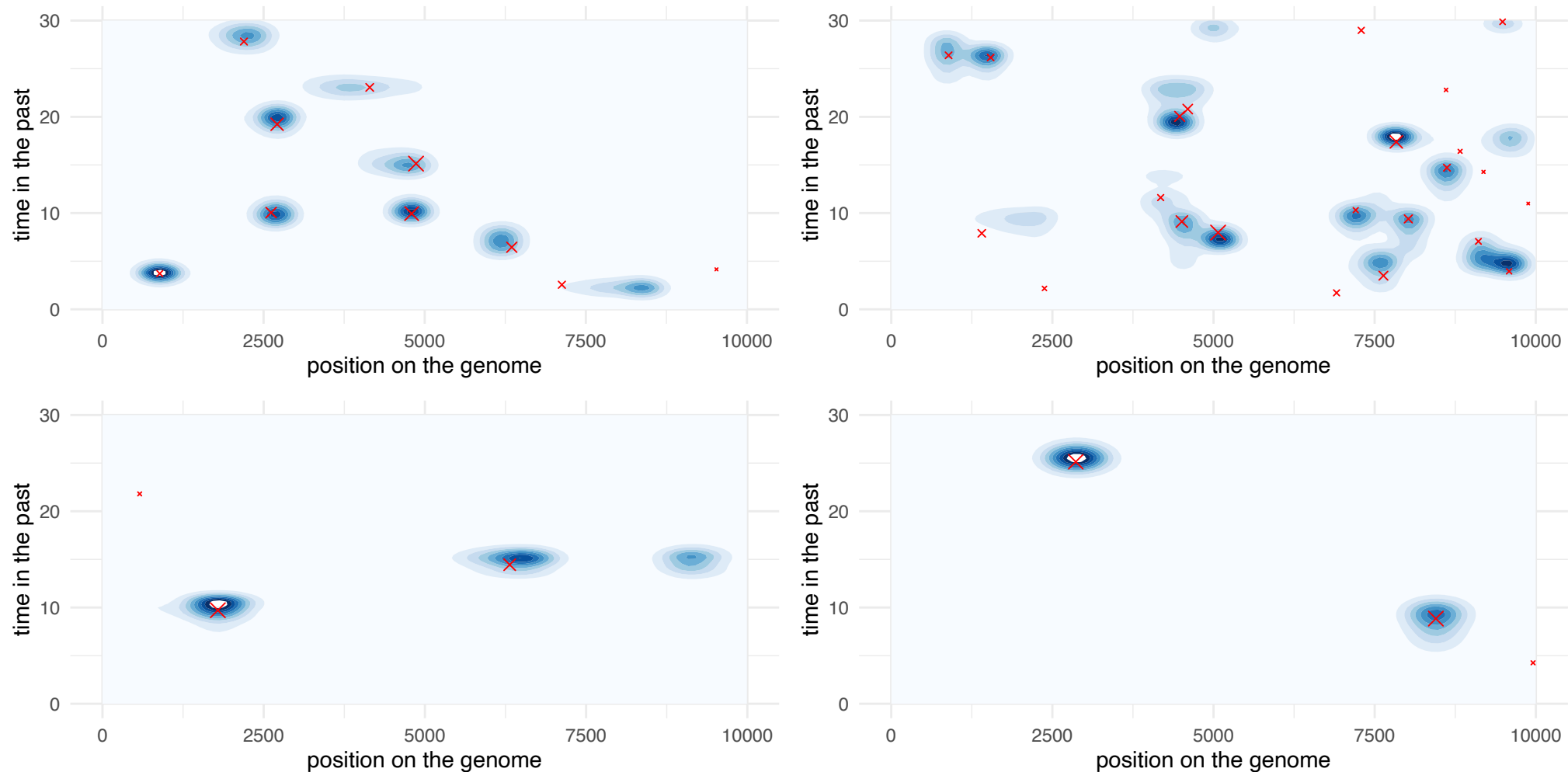
$$P(\text{[tree]} \text{ [grid]} \text{ [clock]} \text{ [loop]} \mid \text{ACAC... TCAC... ACAG...}) = \frac{P(\text{ACAC... TCAC... ACAG...} \mid \text{[tree]} \text{ [grid]} \text{ [clock]}) P(\text{[tree]} \mid \text{[loop]}) P(\text{[loop]}) P(\text{[grid]}) P(\text{[clock]})}{P(\text{ACAC... TCAC... ACAG...})}$$

The complex recombination process can be simplified and modeled as a joint coalescence and recombination process.



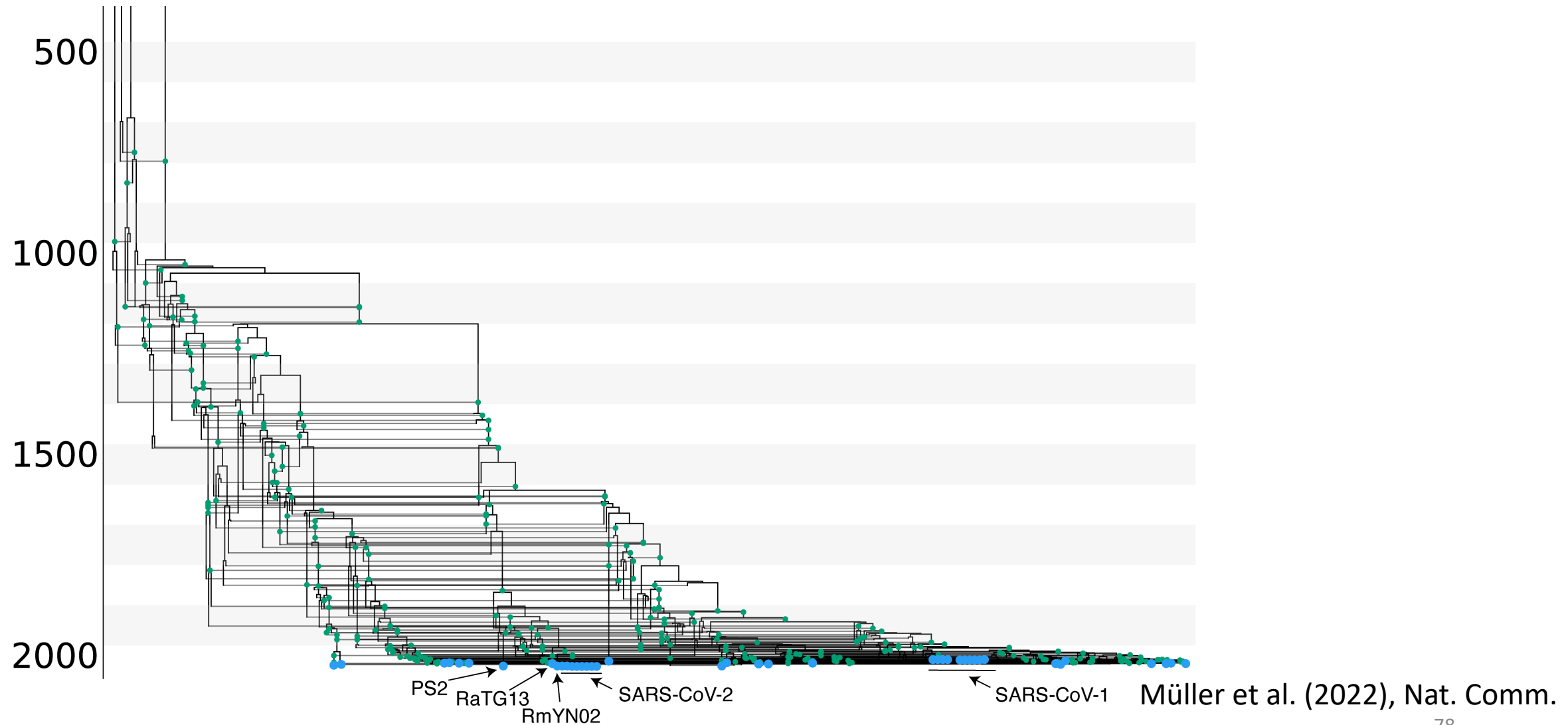
Müller et al. (2022), Nat. Comm.

Recombination events can be inferred well, but getting actual posterior support values is hard

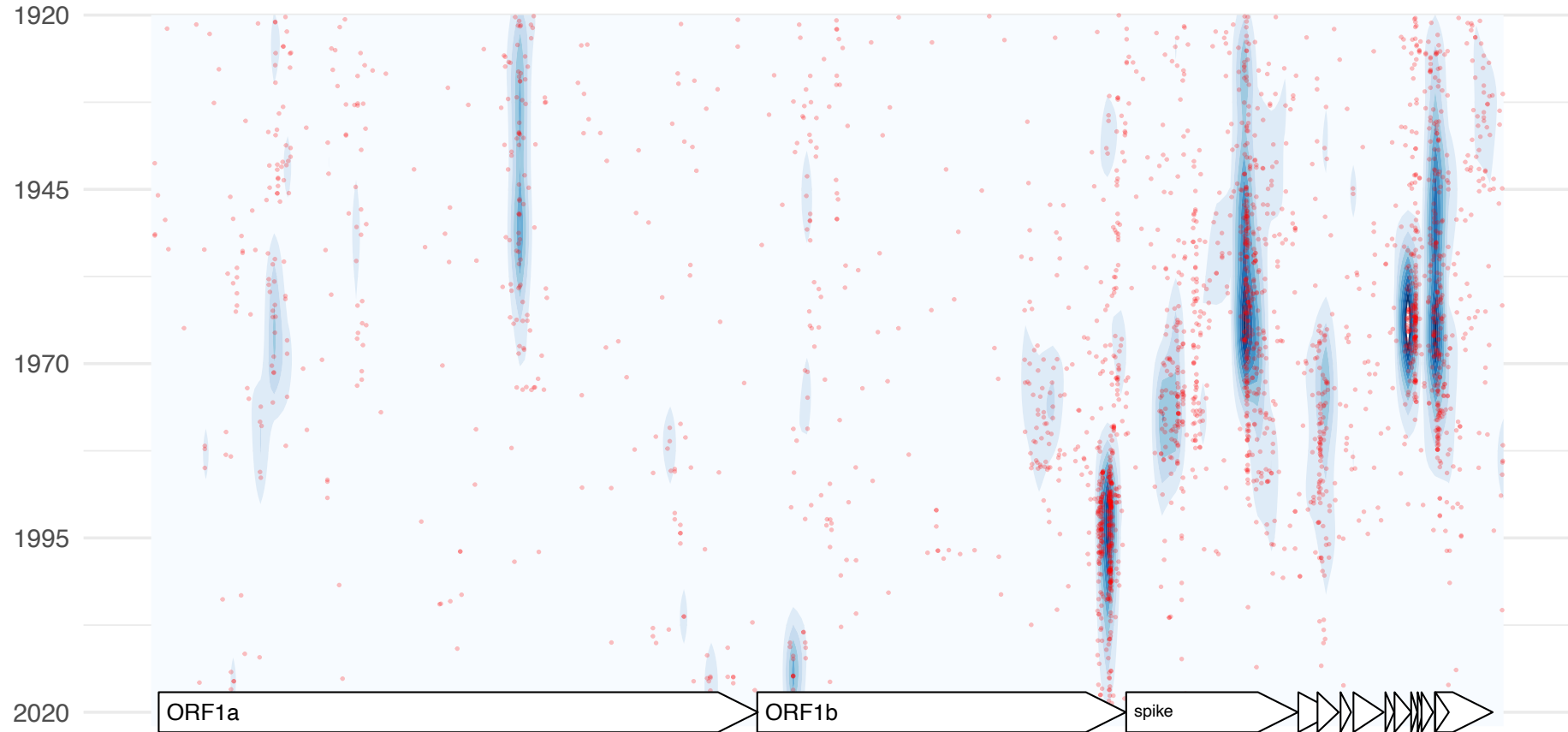


Müller et al. (2022), Nat. Comm.

Recombination events shaped the evolutionary history of SARS-like viruses.

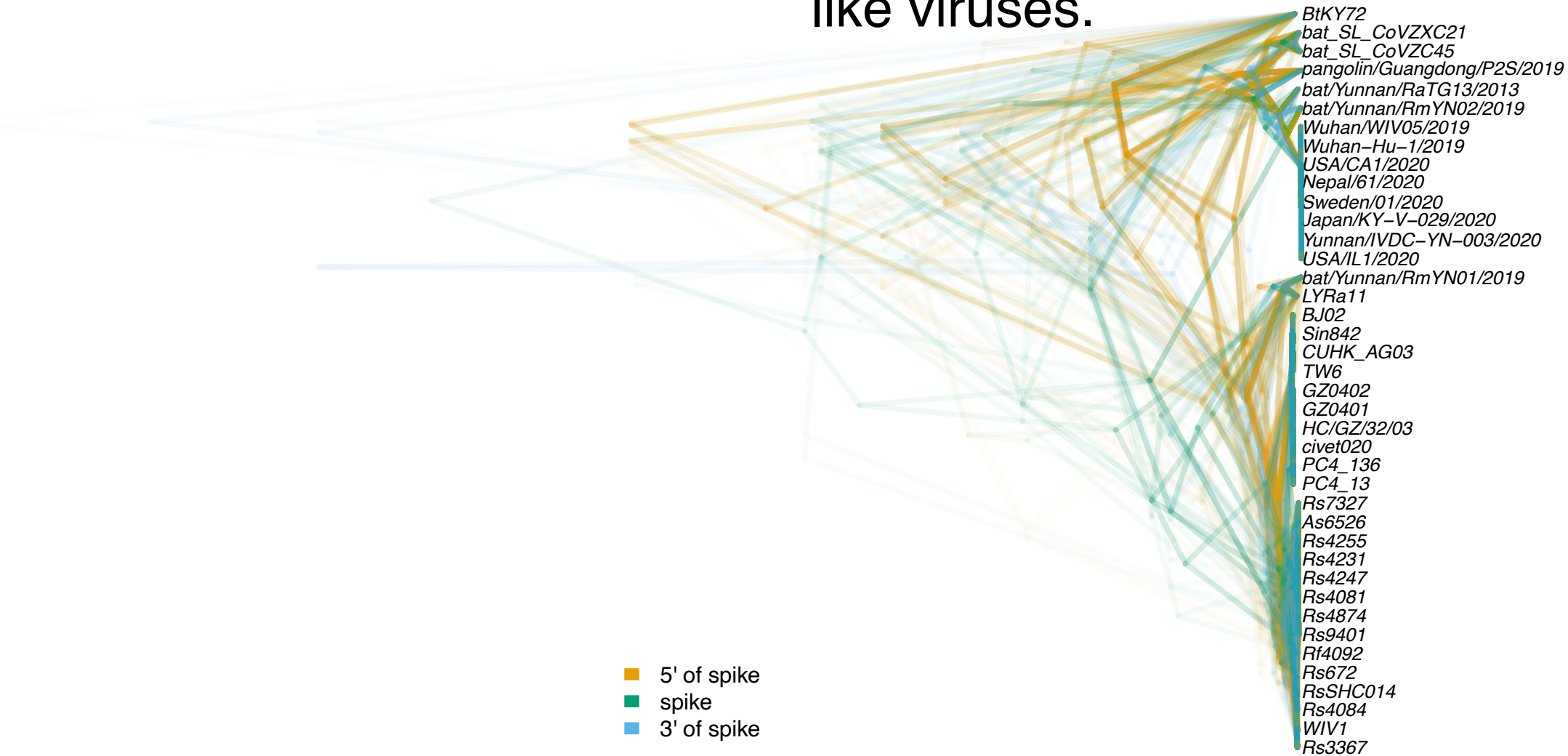


Evidence for a few recombination events in the recent history of SARS-CoV-2 before entering the human population

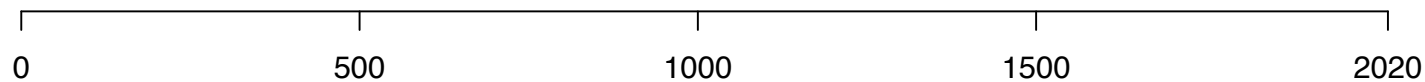


Müller et al. (2022), Nat. Comm.

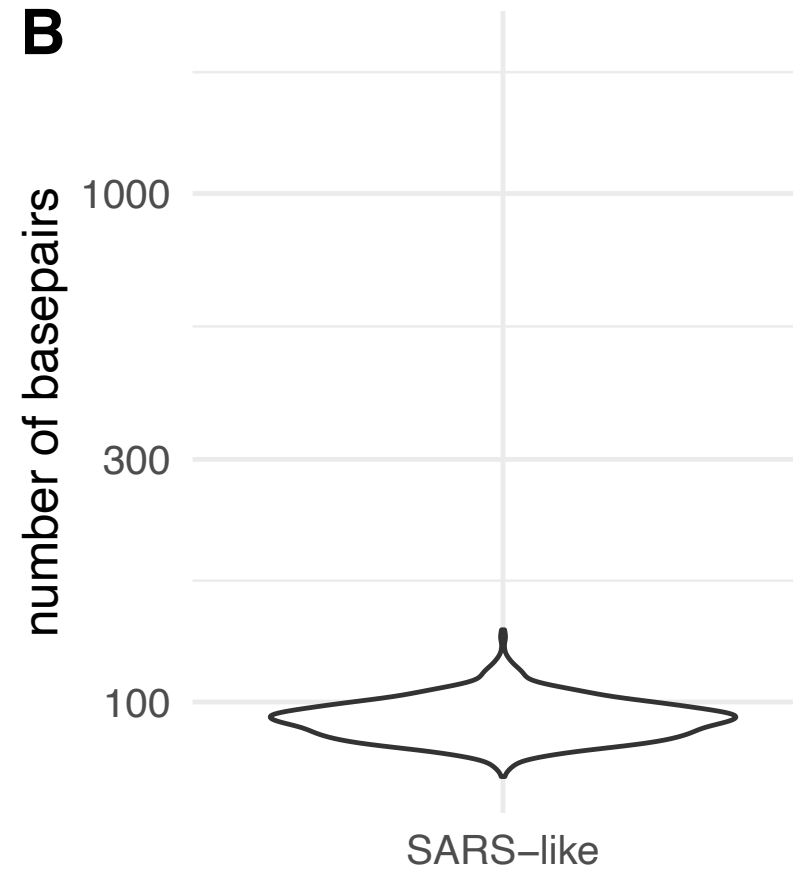
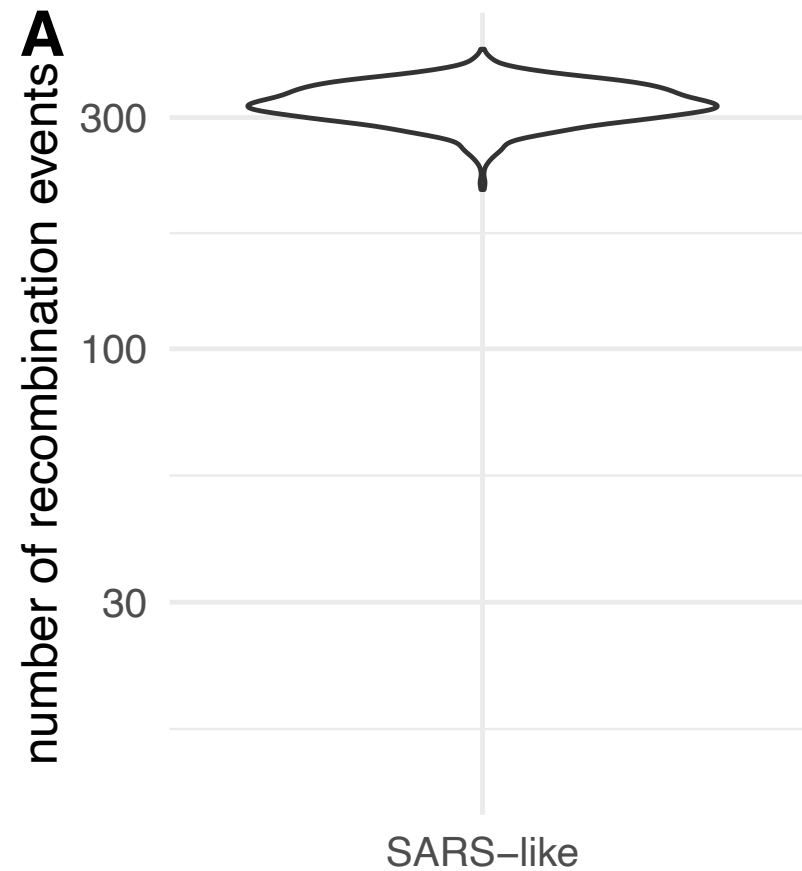
Recombination events shaped the evolutionary history of SARS-like viruses.



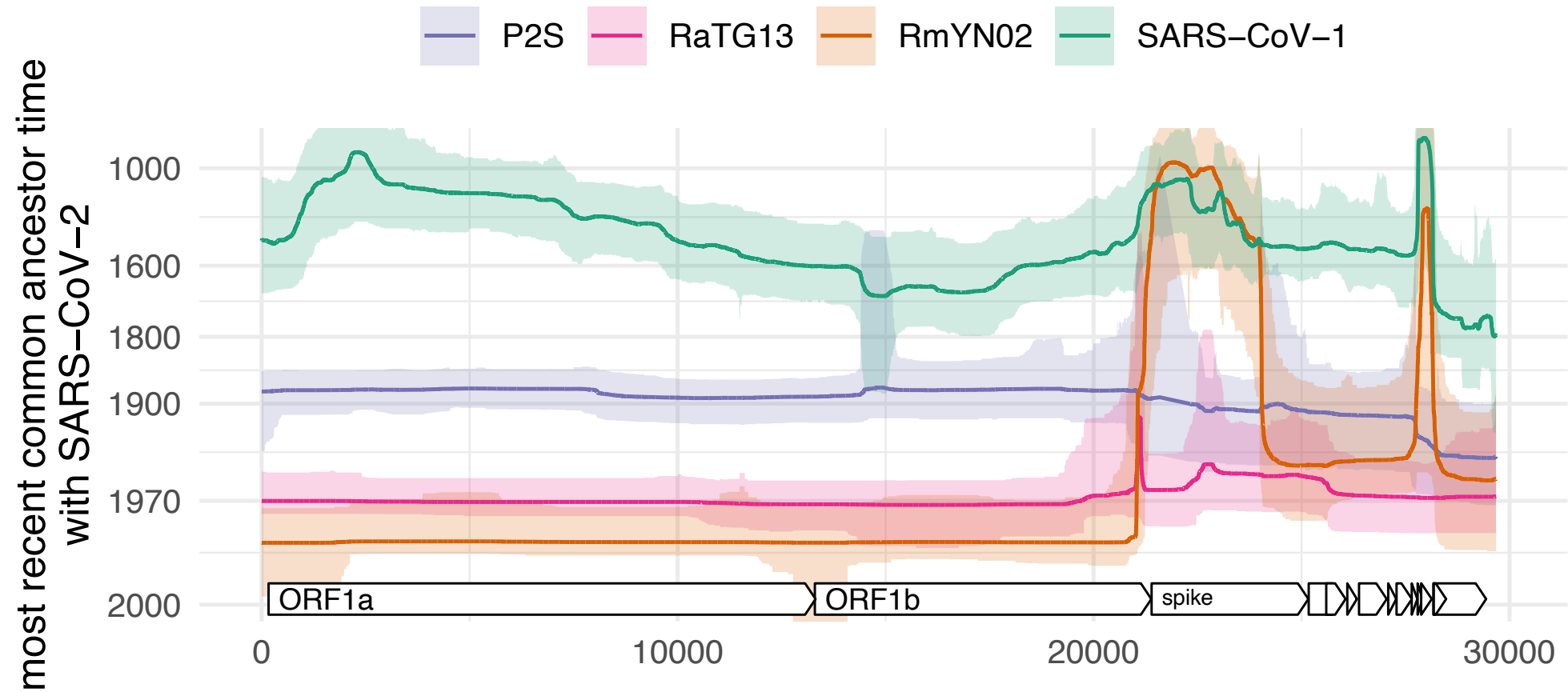
Müller et al. (2022), Nat. Comm.



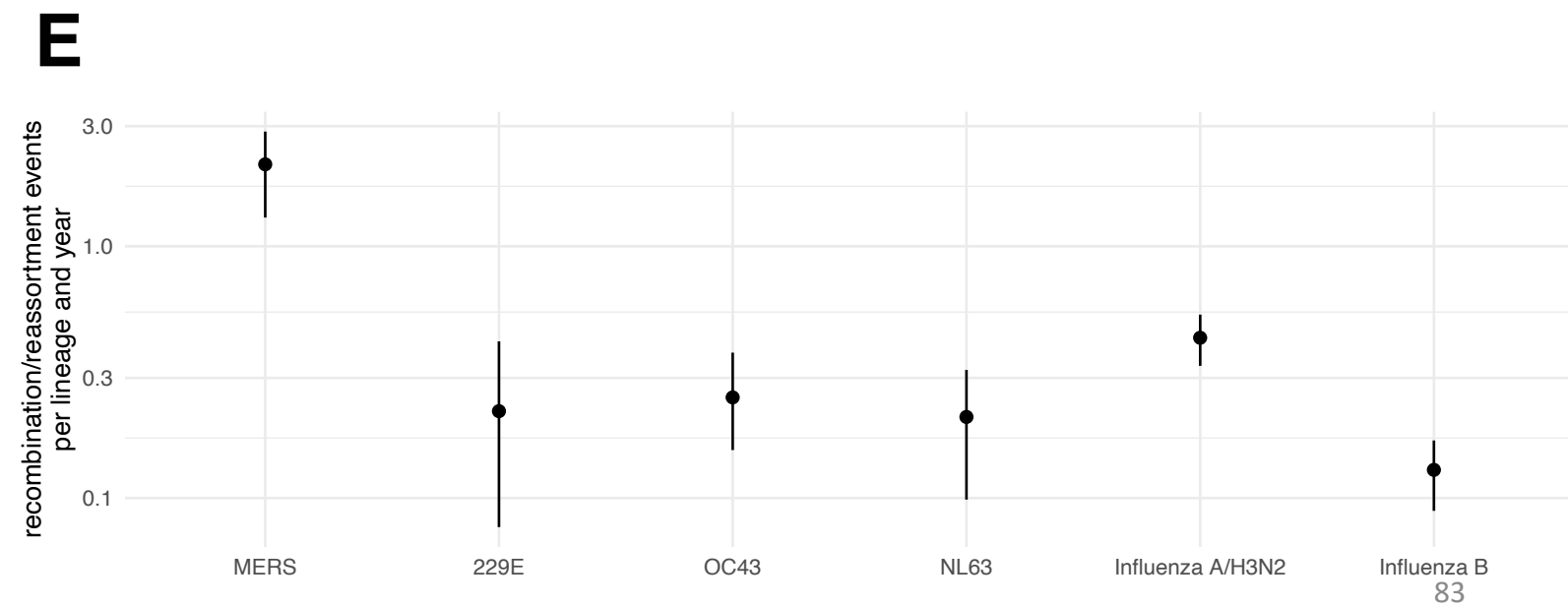
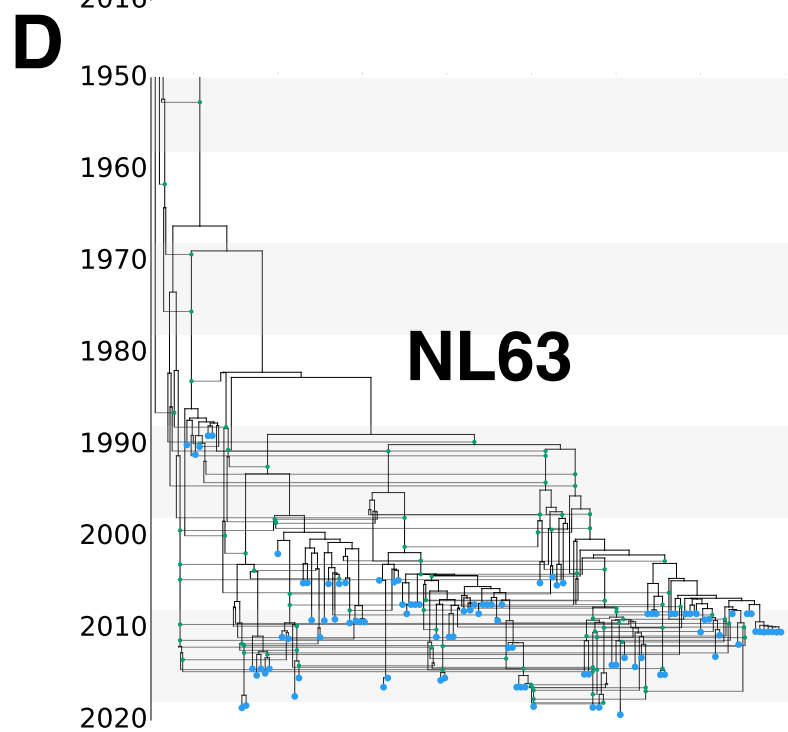
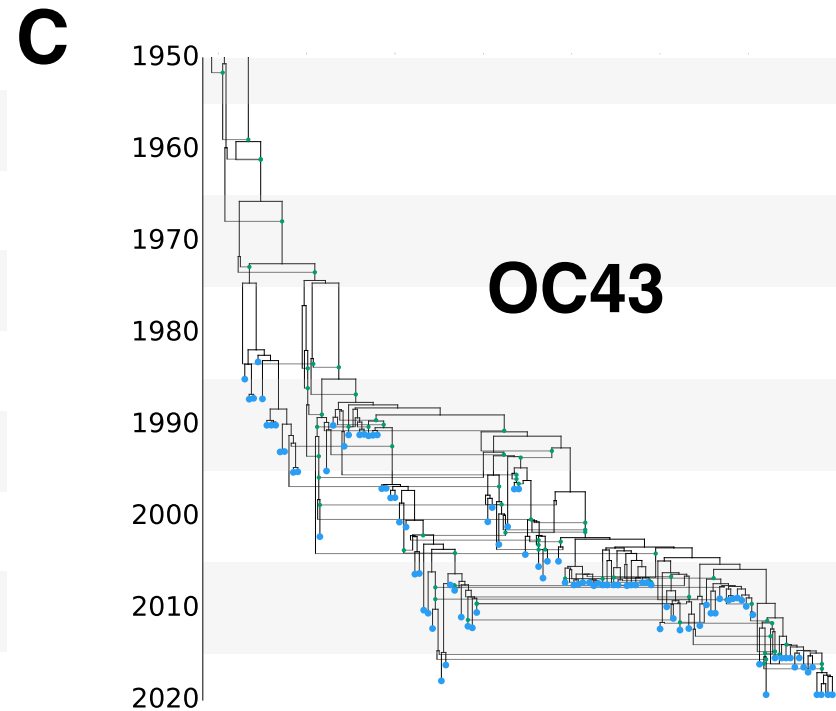
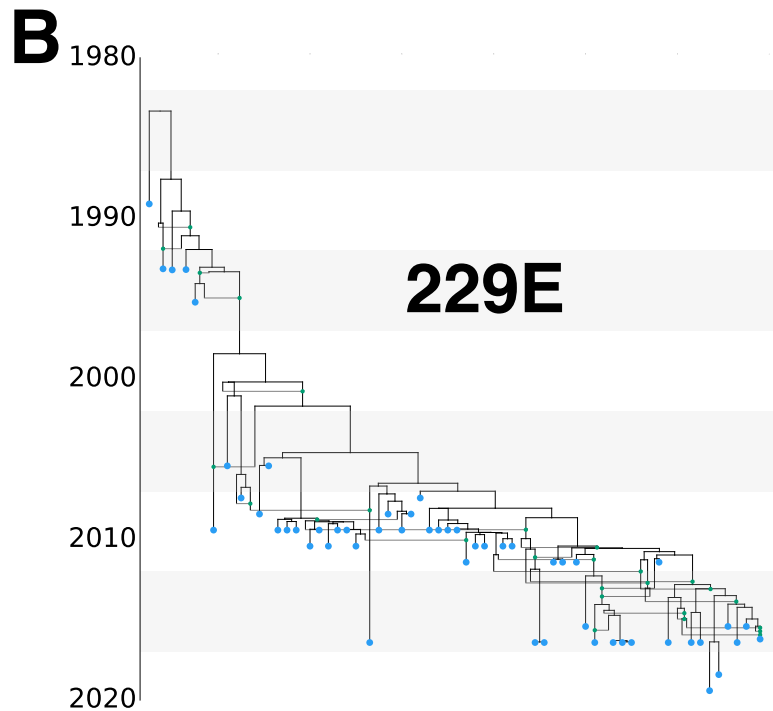
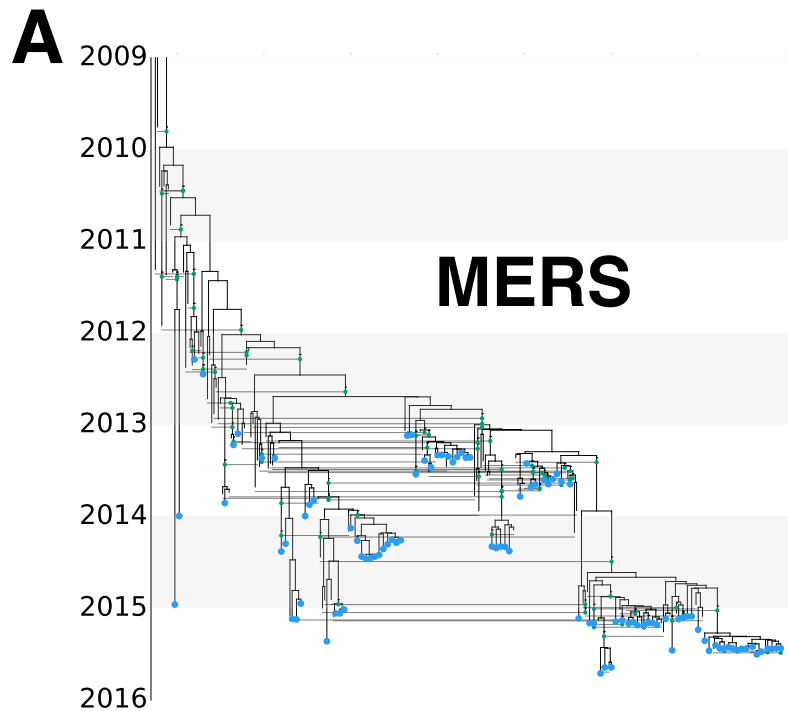
The analyzed SARS-like dataset contains about 300 recombination events, meaning the average number of consecutive basepairs that code for the same tree is 100



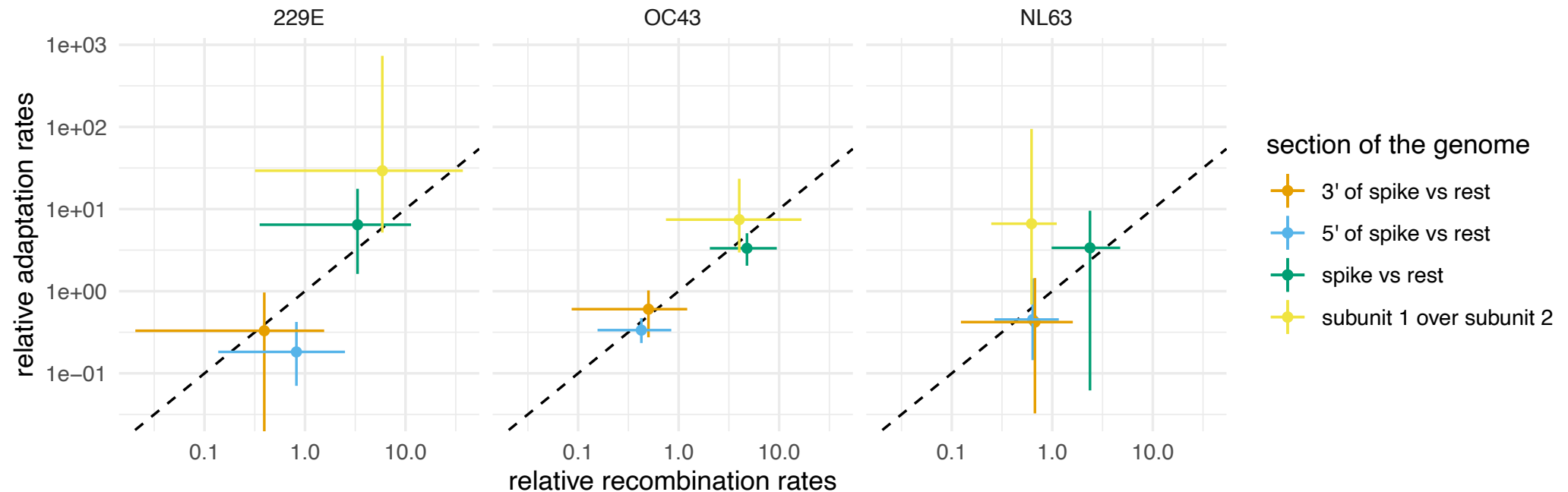
RmYN02 is the closest common ancestor to SARS-CoV-2 on most parts of the genome.



Müller et al. (2022), Nat. Comm.

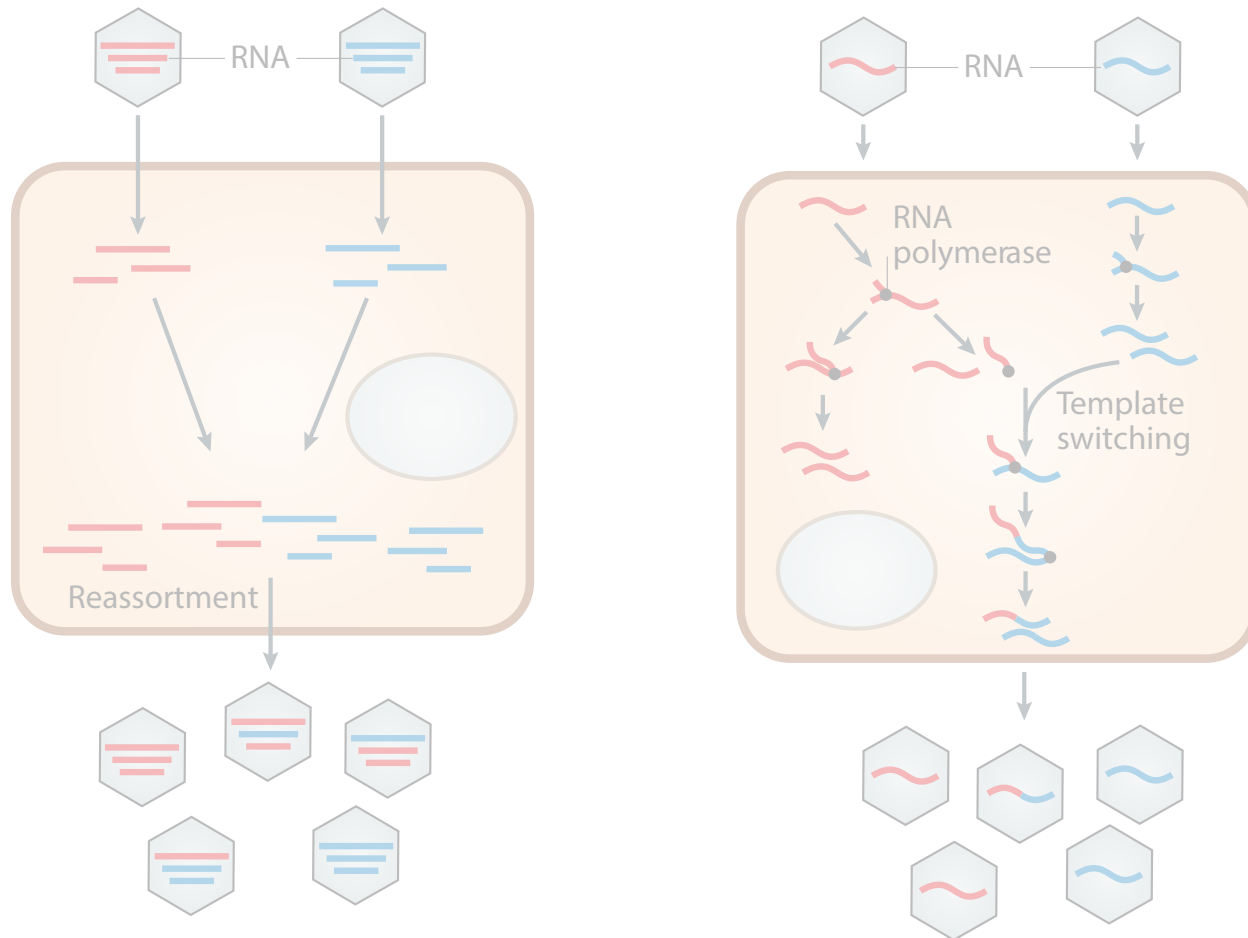


Recombination rates vary with rates of adaptation across the genomes of seasonal coronaviruses

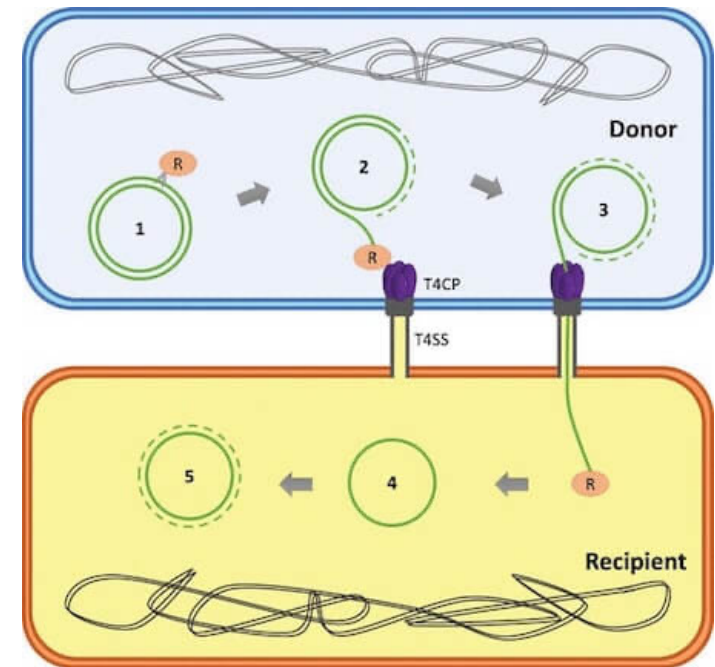


Müller et al. (2022), Nat. Comm.

$P(\text{[icons]} | \text{[icons]})$ models as a coalescent and reassortment/recombination/plasmid transfer process



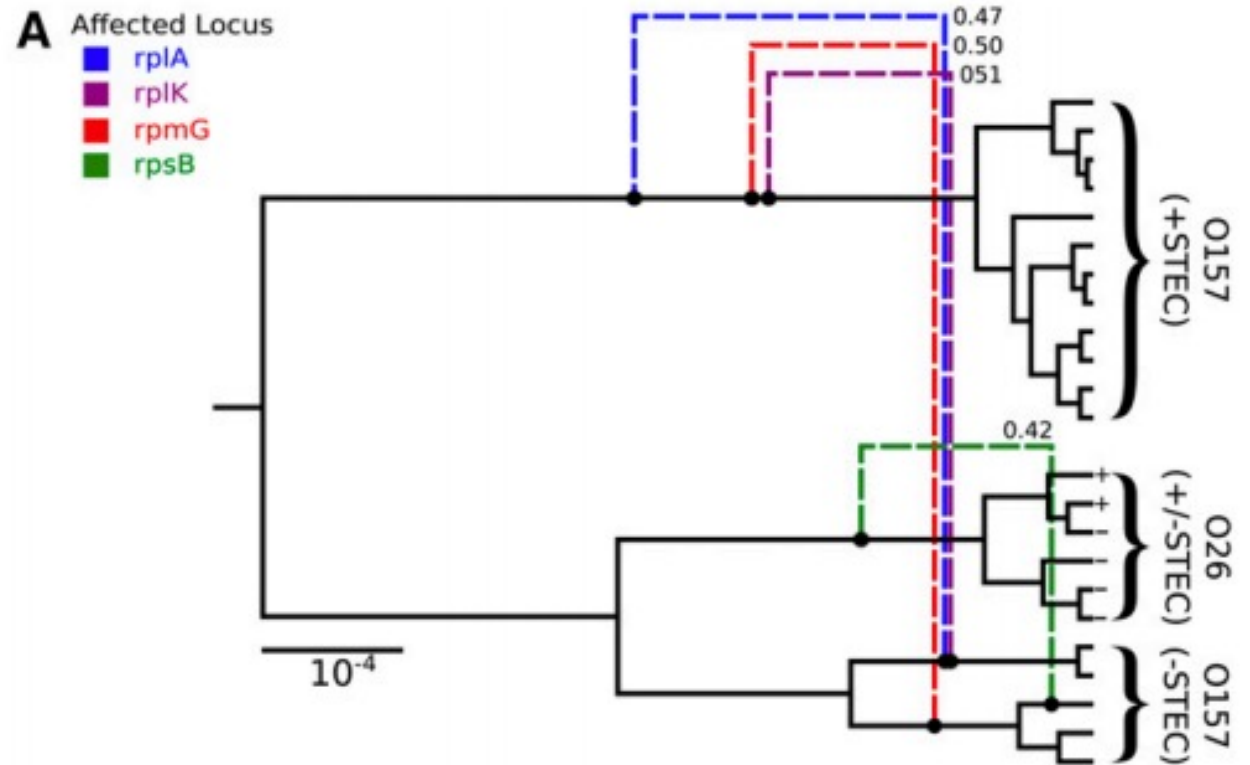
E. Simon-Loriere, E. C. Holmes “Why do RNA viruses recombine?” Nature Reviews Microbiology, 2011



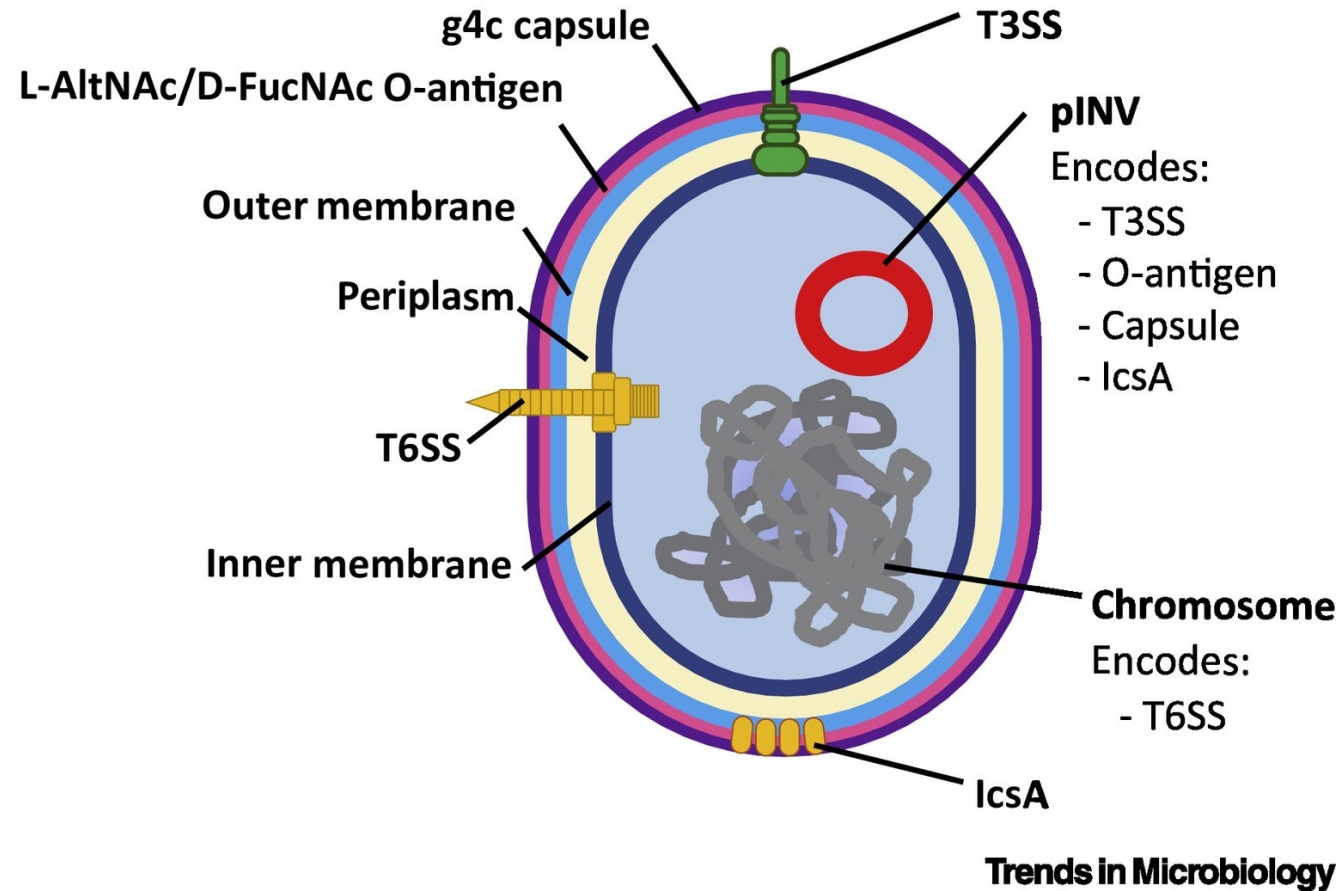
Getting et al./Microbiology Spectrum, Jan. 2018

Ancestral recombination graphs for the coalescent with gene conversion can be inferred in BEAST2 using Bacter

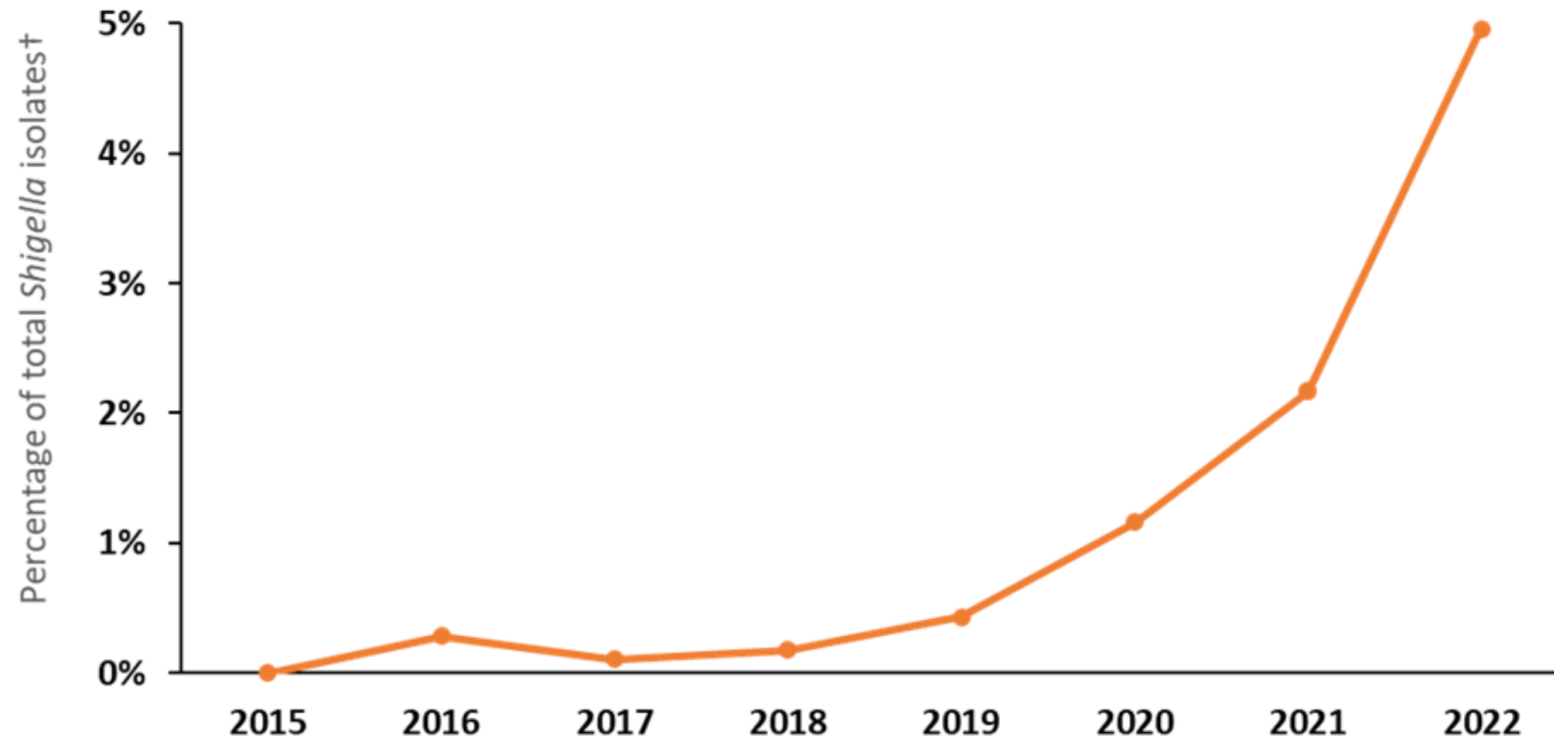
- Allows to estimate tree based networks that have a base tree and edges that detach and re-attach directly to that base tree
- Requires the assumption that only a small part of the genome is subject to recombination



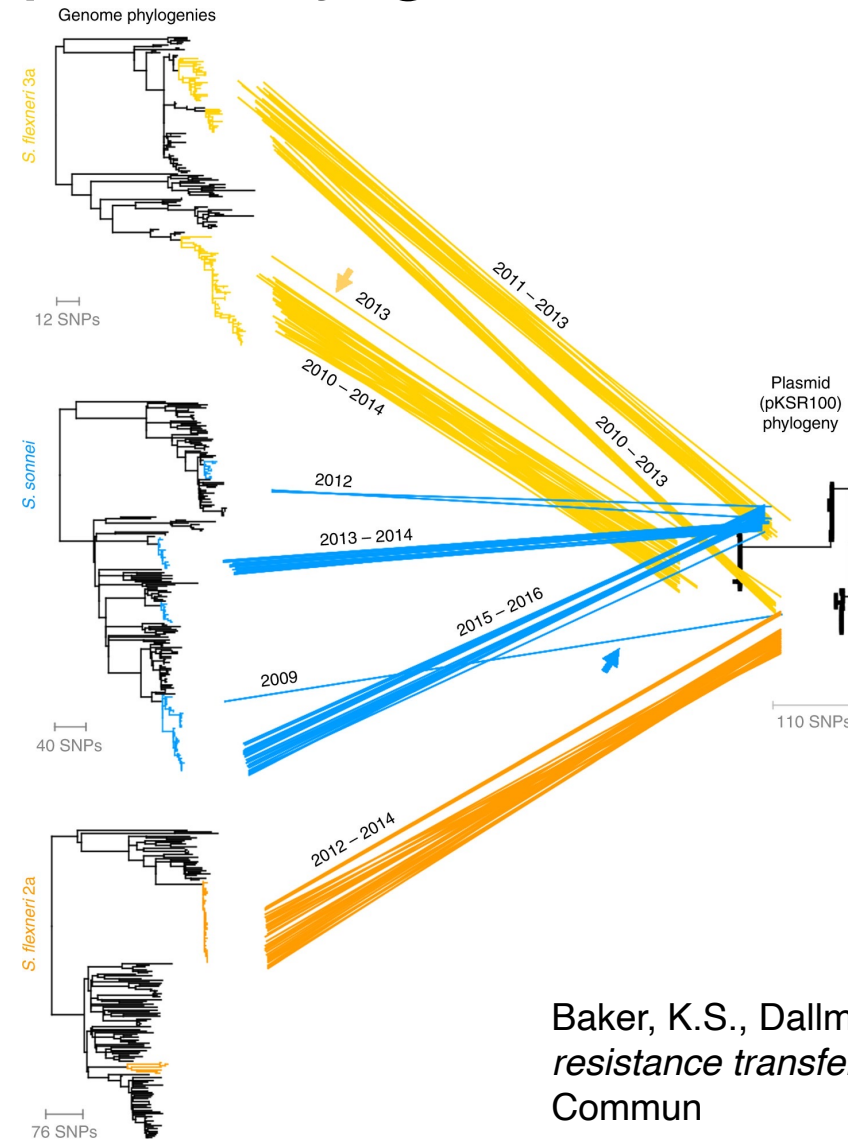
Plasmids encode virulence and antibiotic resistance factors in *Shigella*



XDR *Shigella* is increasing in prevalence in the US

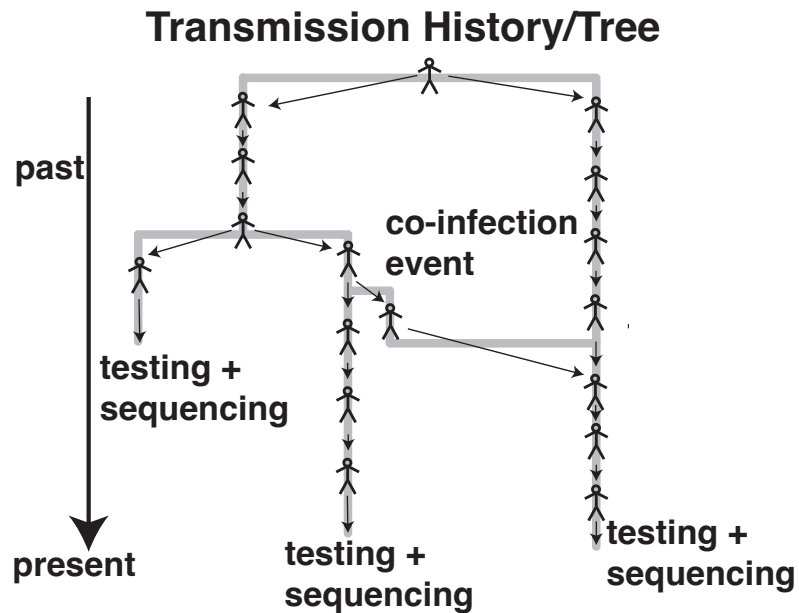


Plasmids can move between different bacterial lineages through, for example, conjugation



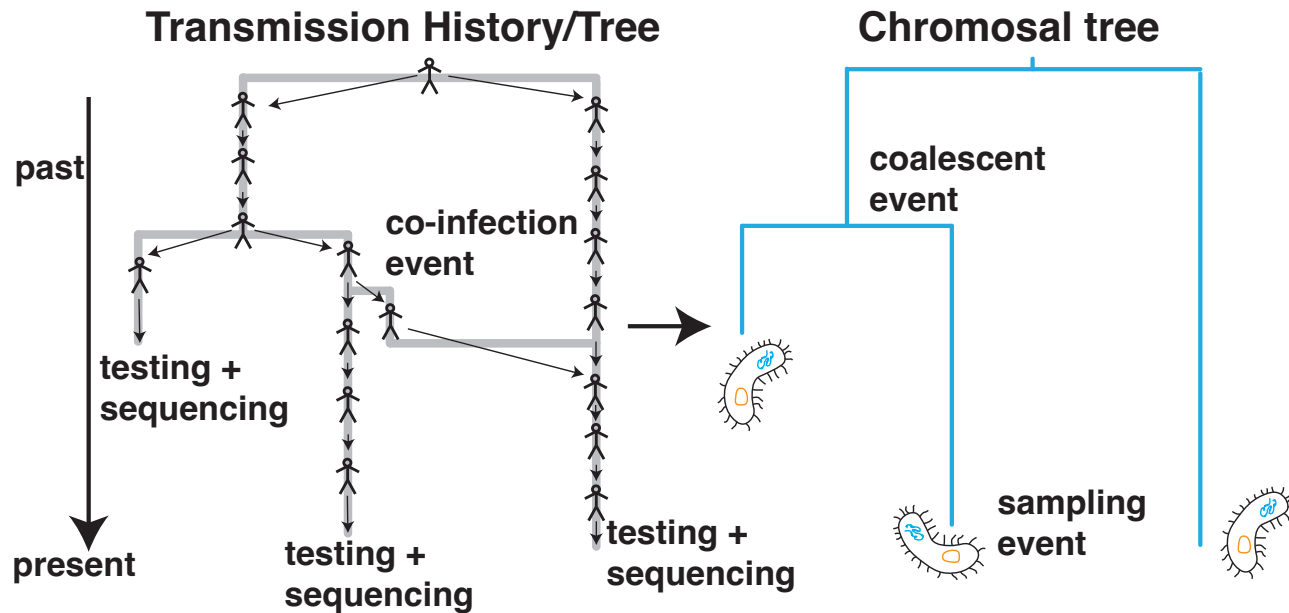
Baker, K.S., Dallman, T.J., Field, N. et al. *Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species*. Nat Commun

The transmission history of bacteria including co-infection can be described by a transmission network



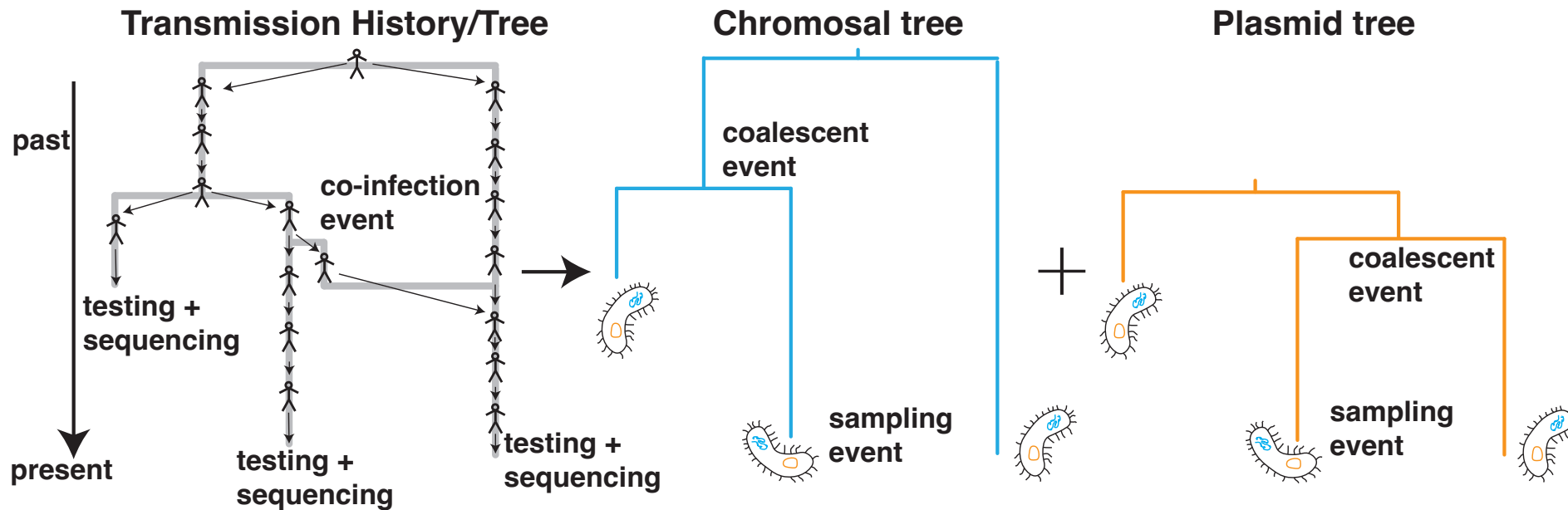
N. F. Müller, Sebastián Duchêne, Deborah A. Williamson, Benjamin Howden, Danielle J Ingle “*Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics*” BioRxiv, 2022

We can reconstruct part of that transmission network from the chromosomal DNA



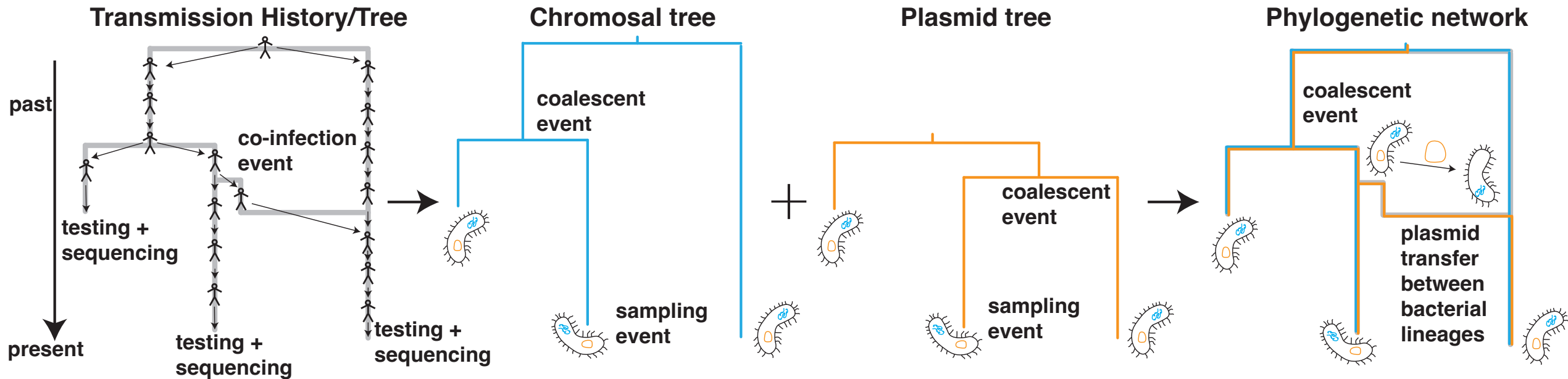
N. F. Müller, Sebastián Duchêne, Deborah A. Williamson, Benjamin Howden, Danielle J Ingle “Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics” BioRxiv, 2022

Discordance between the plasmid and chromosomal tree imply plasmids to have moved between lineages



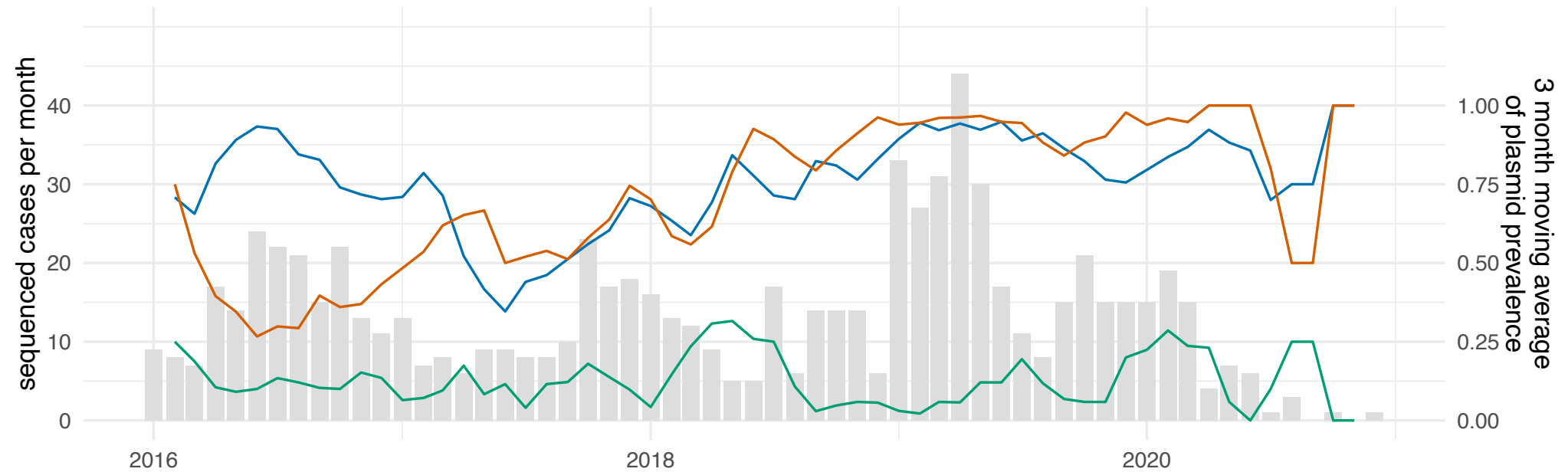
N. F. Müller, Sebastián Duchêne, Deborah A. Williamson, Benjamin Howden, Danielle J Ingle “Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics” BioRxiv, 2022

That shared history of chromosome and plasmid can be denoted by a phylogenetic network

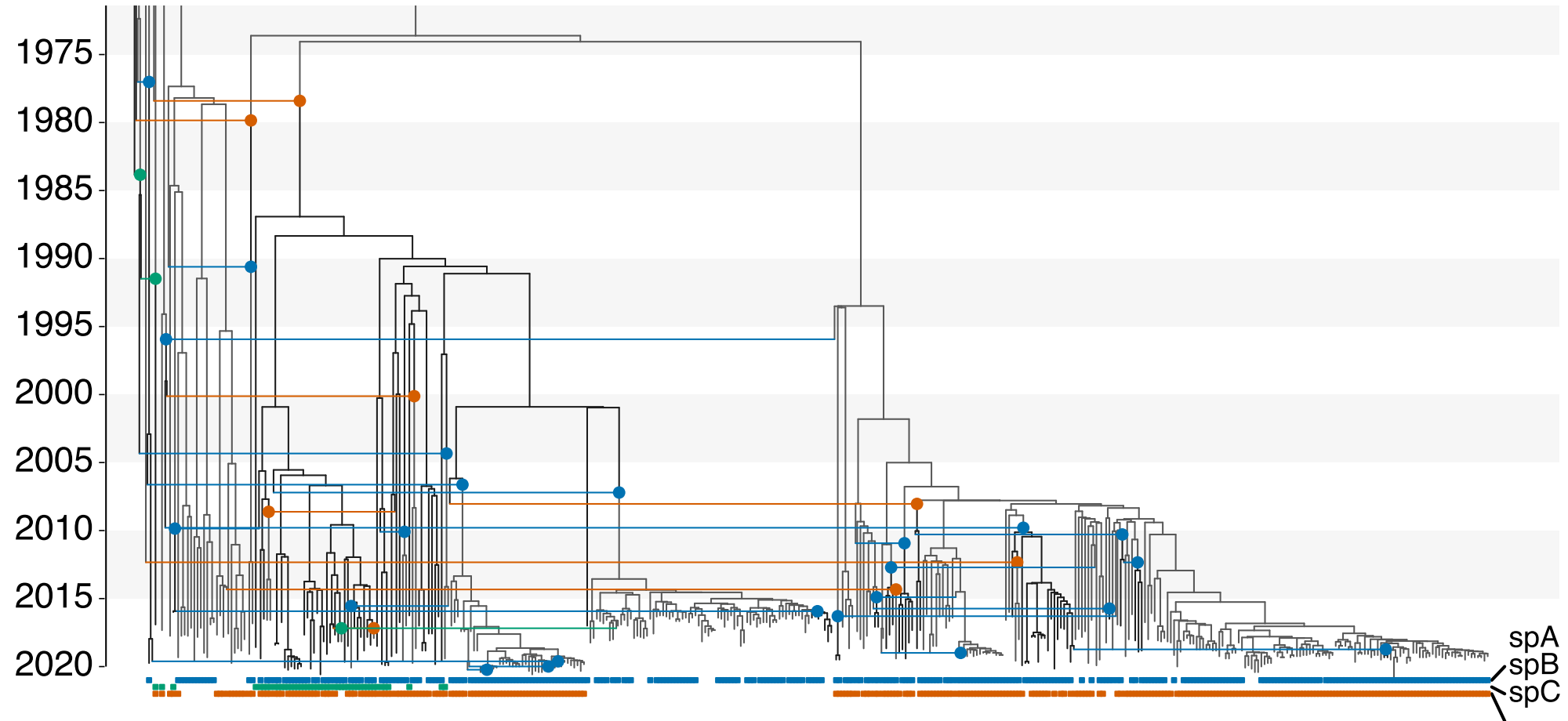


N. F. Müller, Sebastián Duchêne, Deborah A. Williamson, Benjamin Howden, Danielle J Ingle “Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics” BioRxiv, 2022

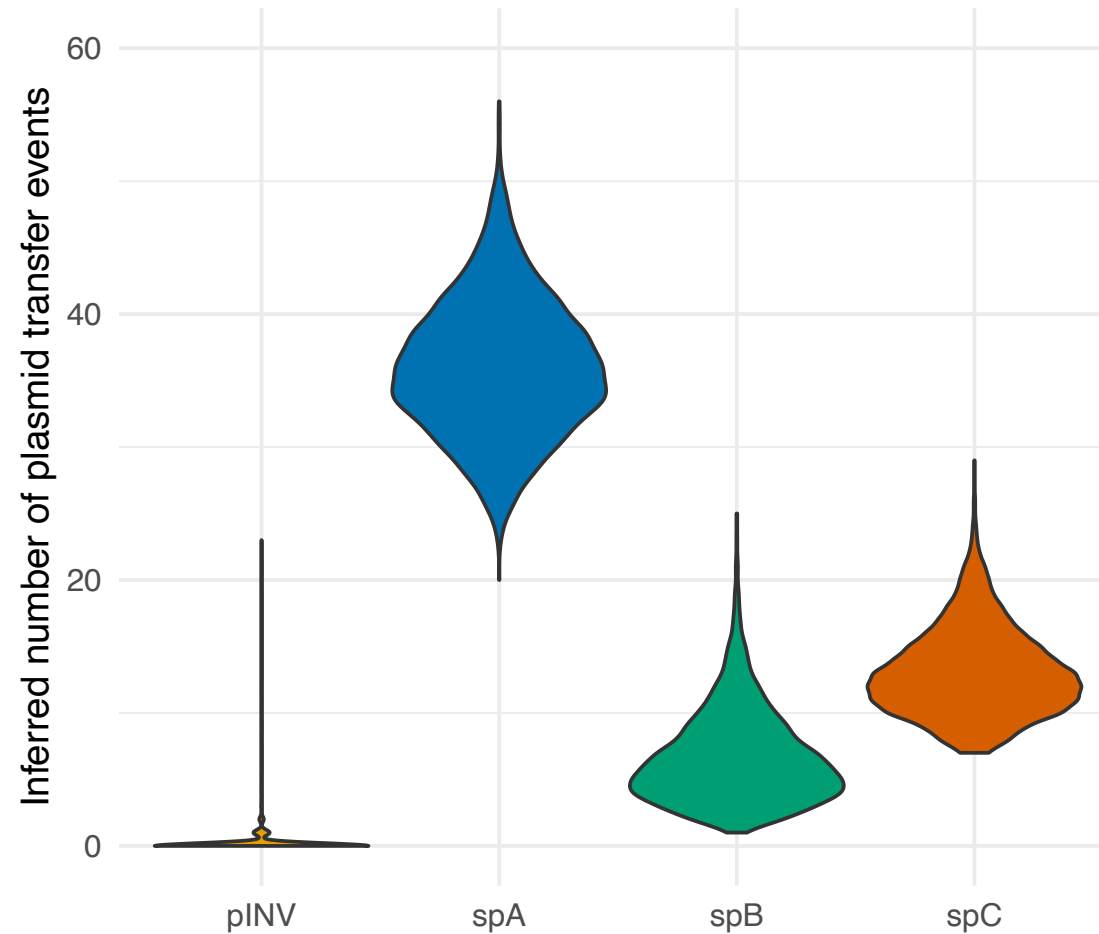
The transfer of plasmids between bacterial lineages can be inferred using phylogenetic networks



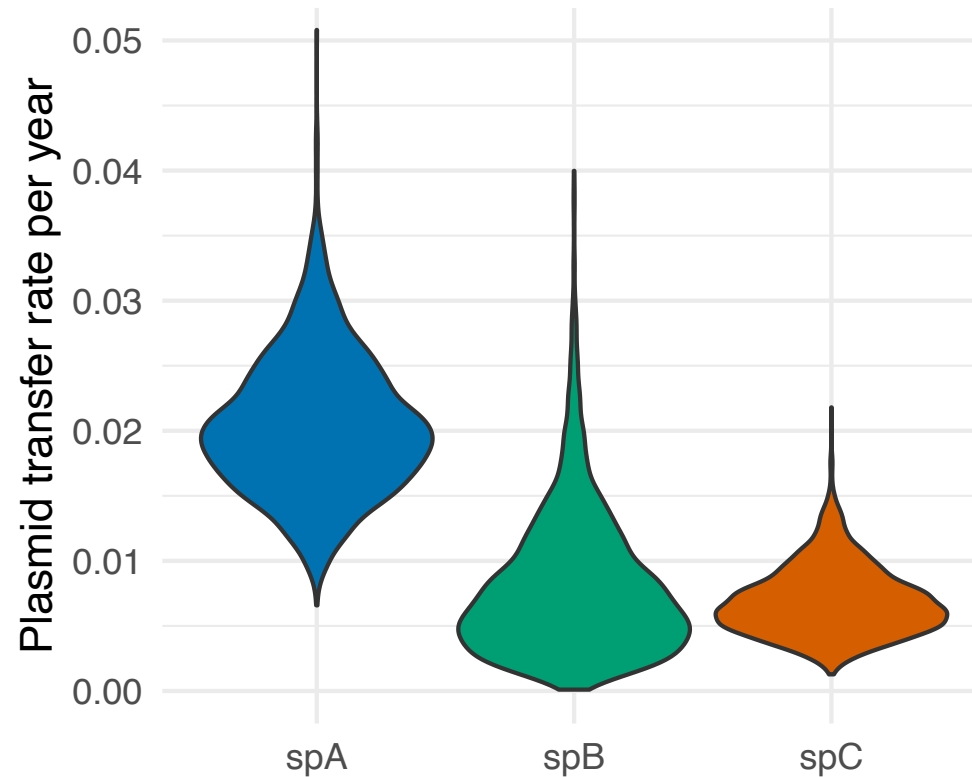
The transfer of plasmids between bacterial lineages can be inferred using phylogenetic networks



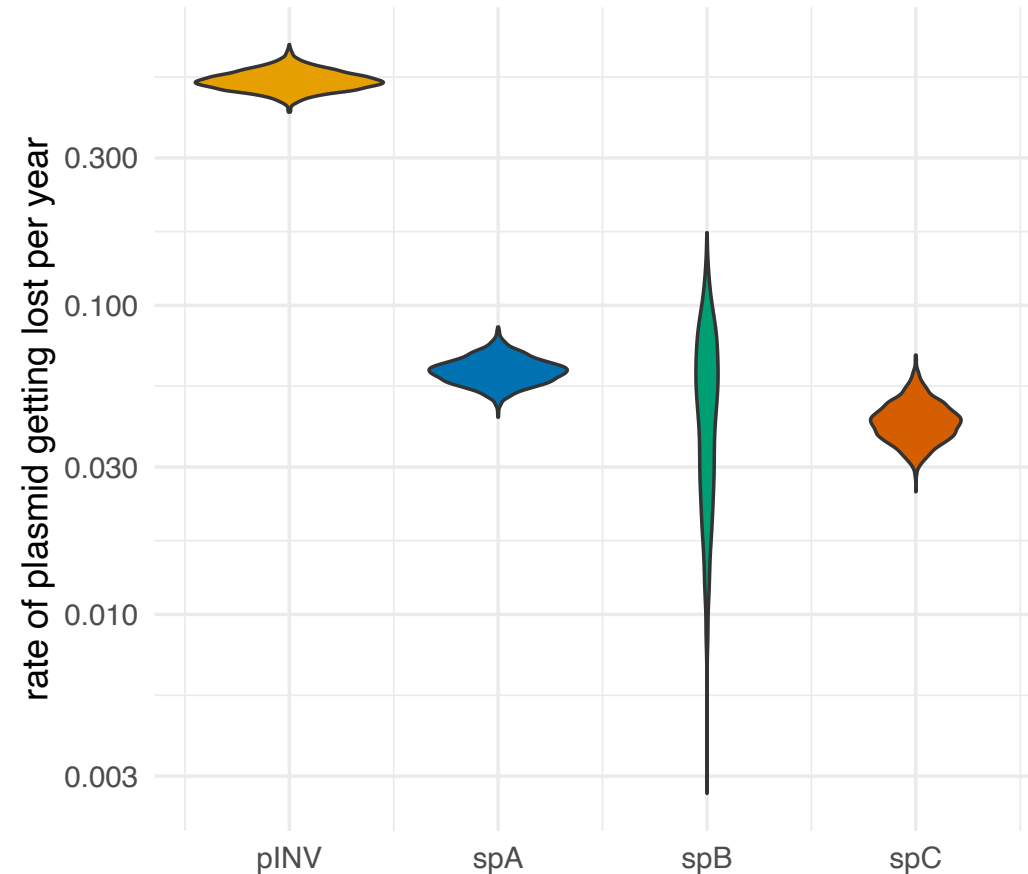
spA is transferred between lineages more often than other plasmids with unknown function (spB, spC)



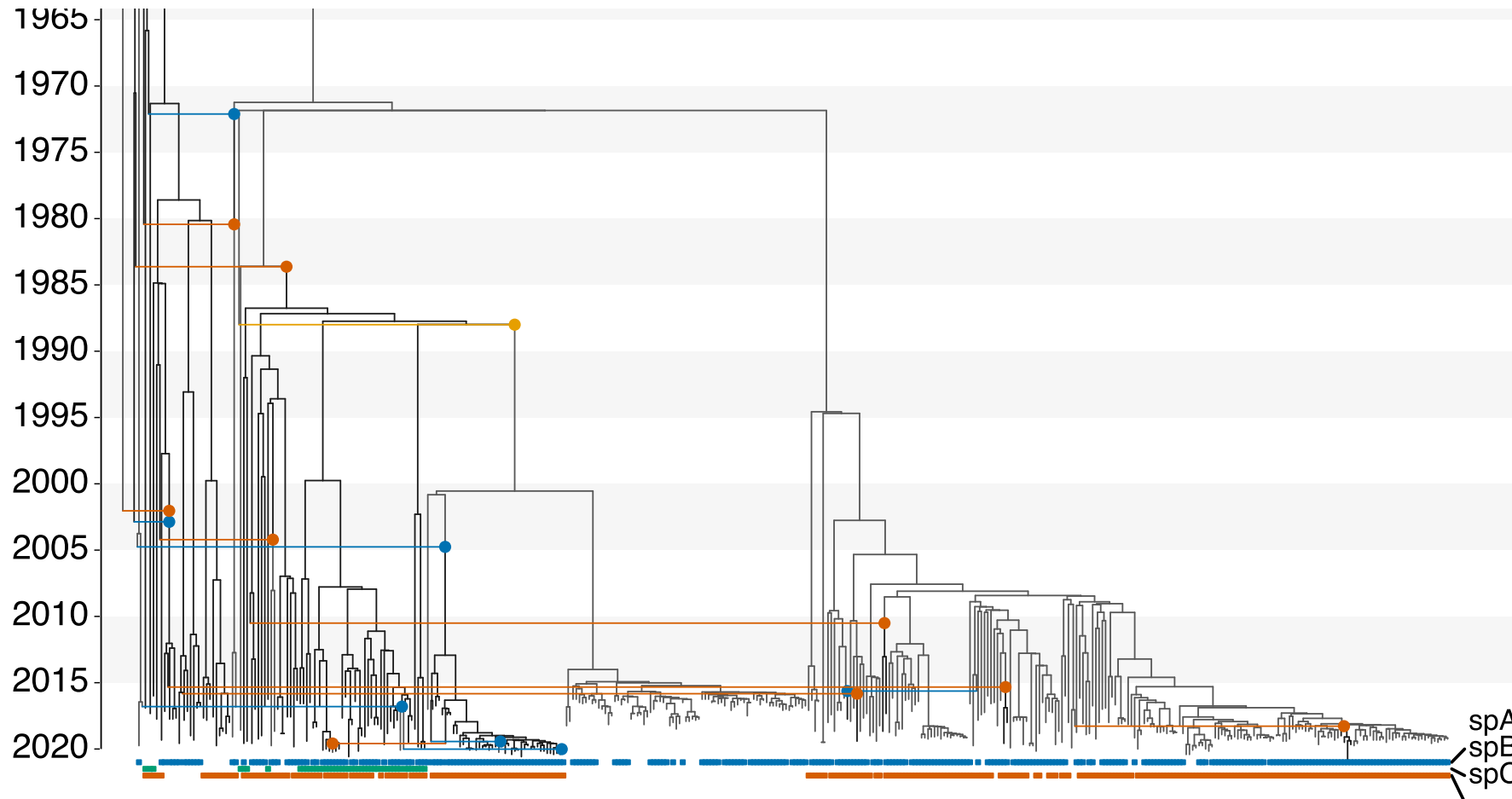
This is reflected in a higher rate of the spA moving between lineages



The transfer of plasmids between bacterial lineages can be inferred using phylogenetic networks

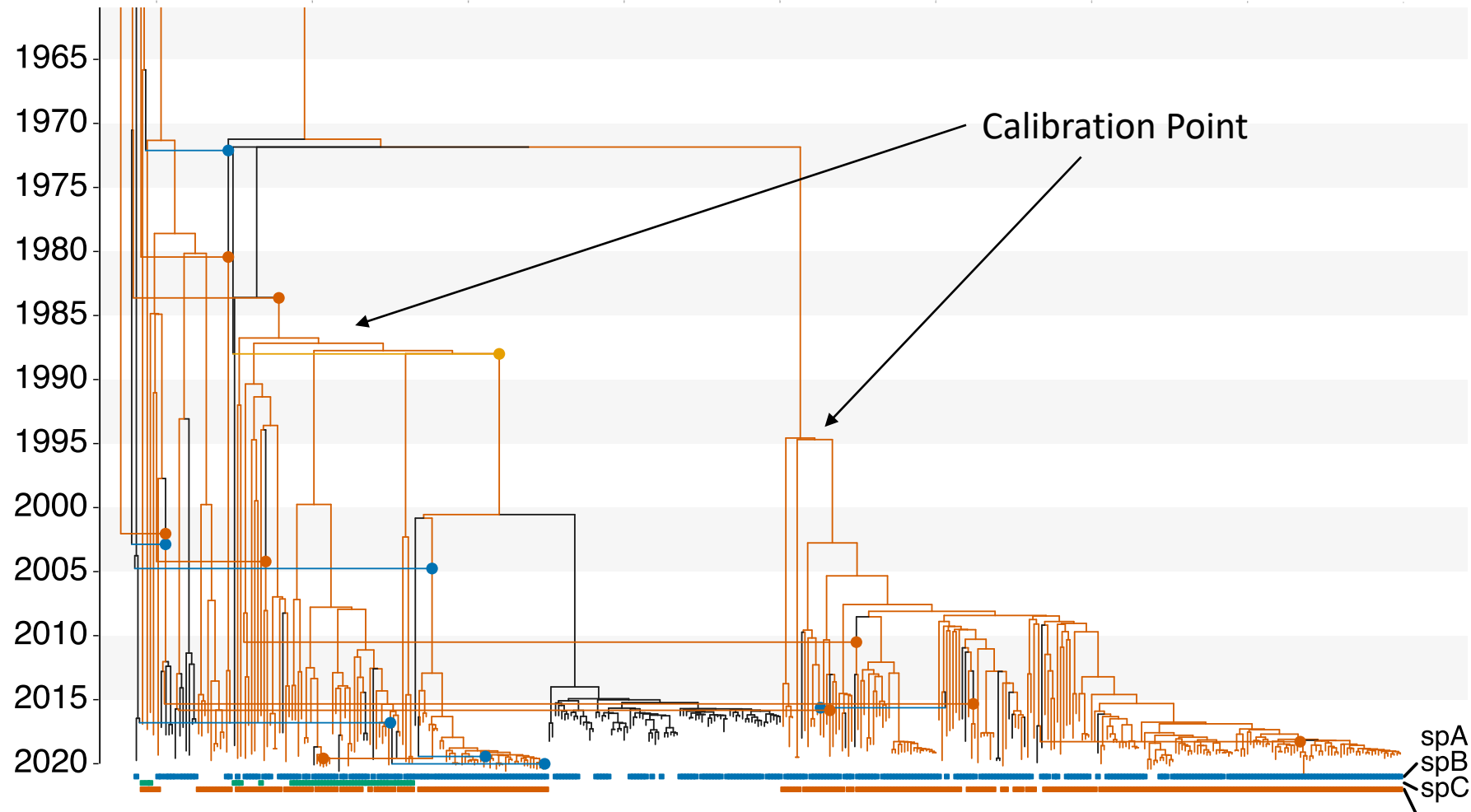


The transfer of plasmids between bacterial lineages can be inferred using phylogenetic networks



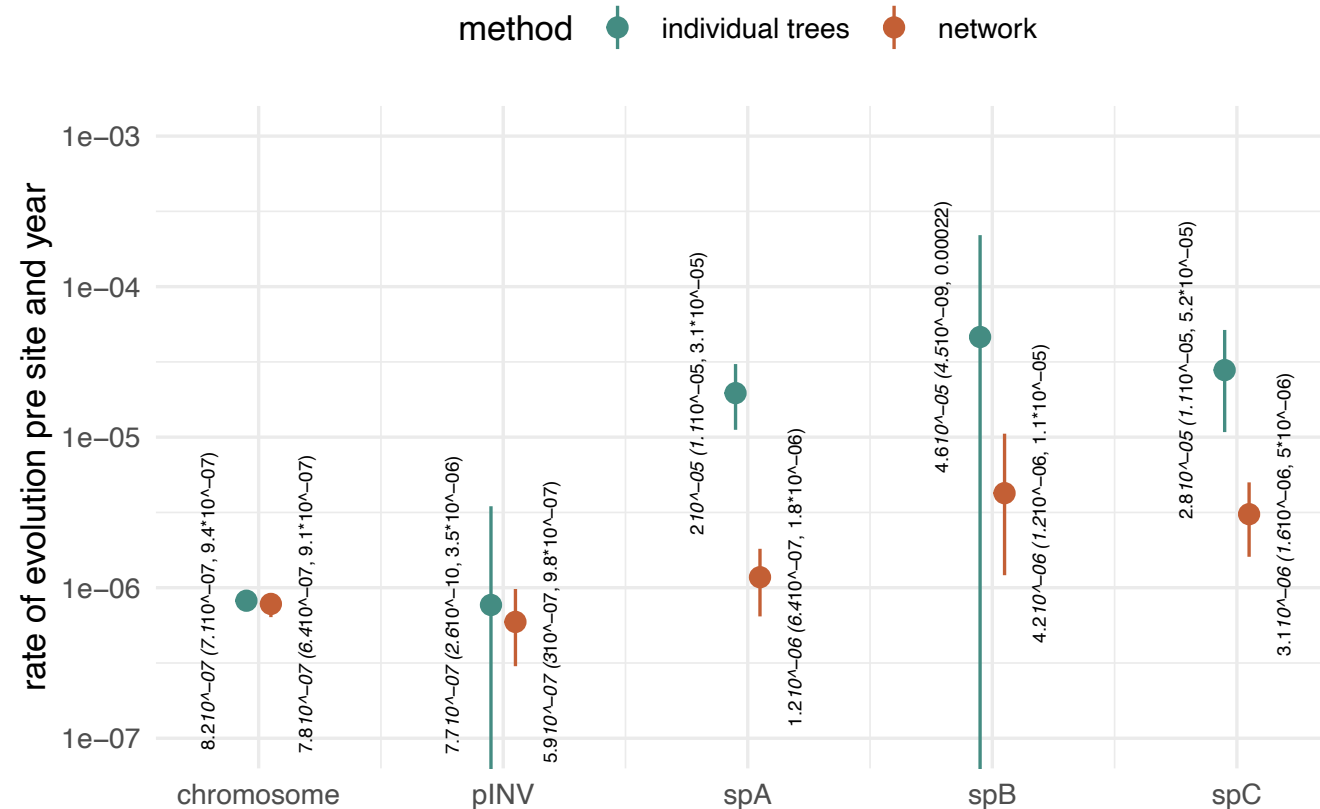
N. F. Müller, Sebastián Duchêne, Deborah A. Williamson, Benjamin Howden, Danielle J Ingle “Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics” BioRxiv, 2022

Chromosome tree provides calibration points for plasmid tree

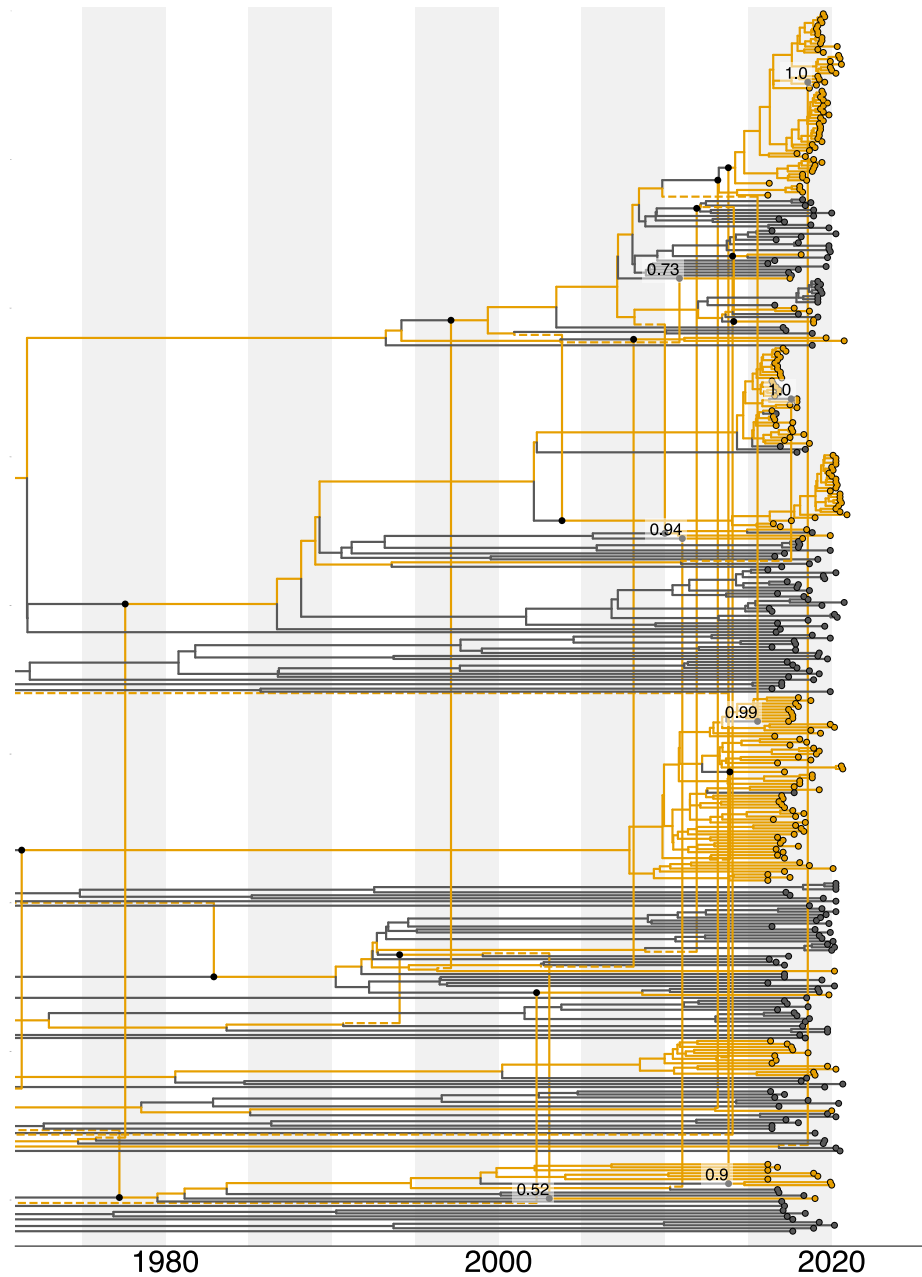


N. F. Müller, Sebastián Duchêne, Deborah A. Williamson, Benjamin Howden, Danielle J Ingle “Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics” BioRxiv, 2022

The evolutionary rates of plasmids can be estimated despite them not “measurably evolving” by themselves



N. F. Müller, Sebastián Duchêne, Deborah A. Williamson, Benjamin Howden, Danielle J Ingle “Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics” BioRxiv, 2022



S. sonnei

S. flexneri

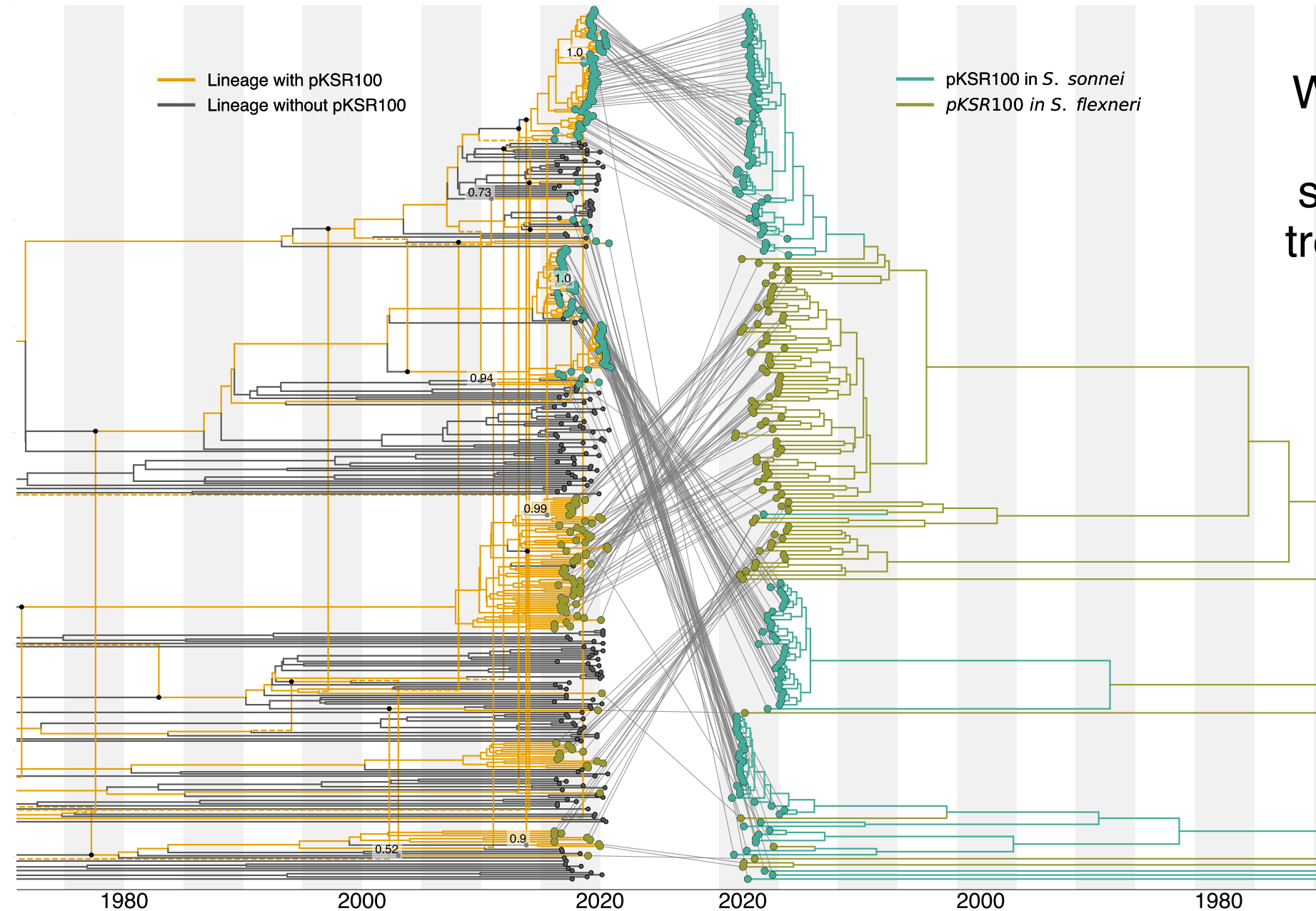
The ancestral history of pKSR100 shows multiple jumps between bacterial lineages of *S. sonnei* and *S. flexneri*.

pKSR100 (in orange) confers resistance to azithromycin and jumped repeatedly between *S. sonnei* and *S. flexneri*

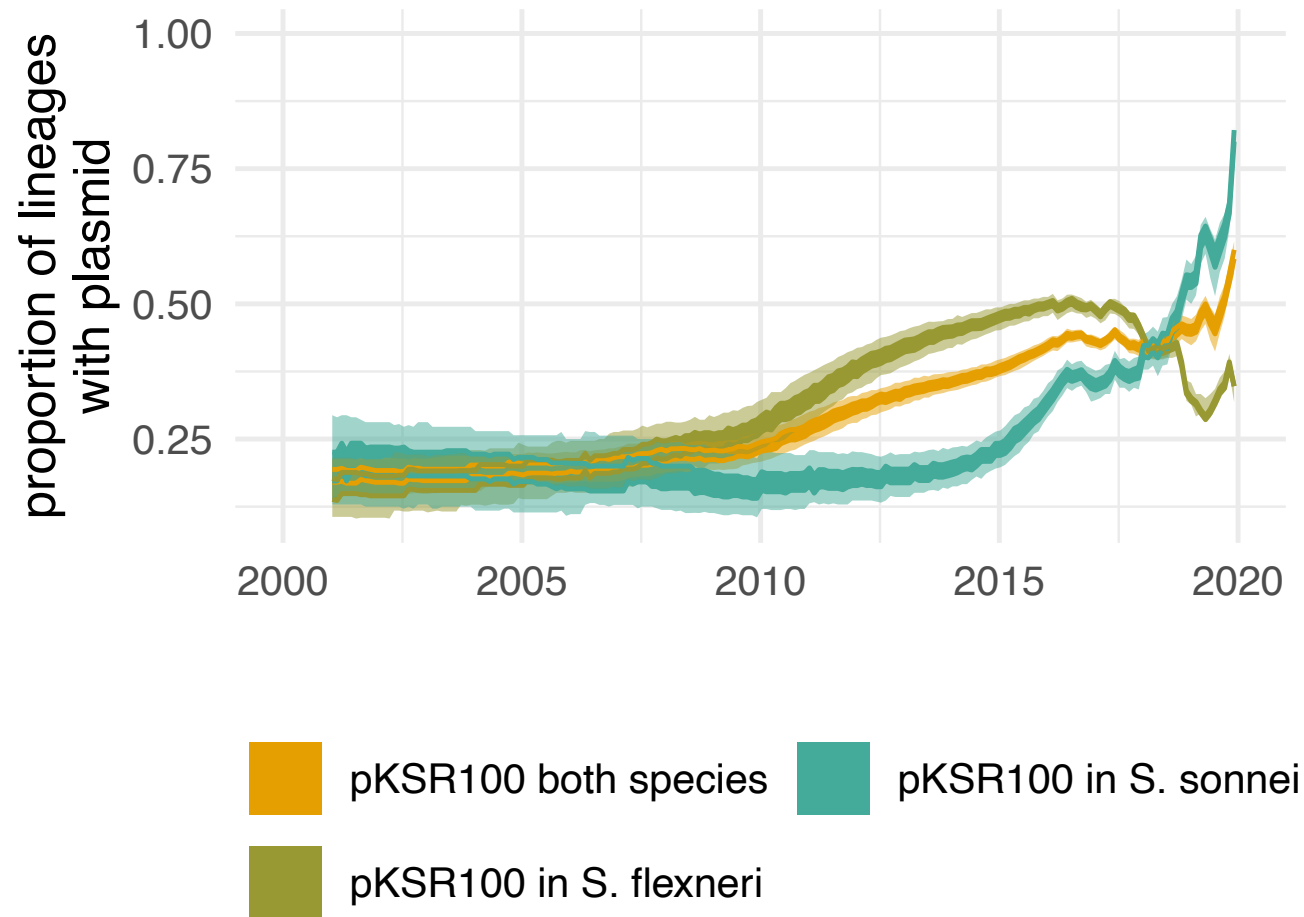
— Lineage with pKSR100
— Lineage without pKSR100

— pKSR100 in *S. sonnei*
— pKSR100 in *S. flexneri*

We can map these jumps of pKSR100 between species onto the plasmid tree and learn when these jumps occurred



Expansion of the number of bacterial lineages carrying pKSR100 over the last decade



QUESTIONS?

Some reading material

- Coalescent with recombination: [https://doi.org/10.1016/0040-5809\(83\)90013-8](https://doi.org/10.1016/0040-5809(83)90013-8)
- Coalescent with gene conversion: <https://www.genetics.org/content/155/1/451.short>
- ARG's for bacteria: <https://www.genetics.org/content/205/2/857>
- Coalescent with reassortment: <https://www.pnas.org/content/early/2020/07/02/1918304117>