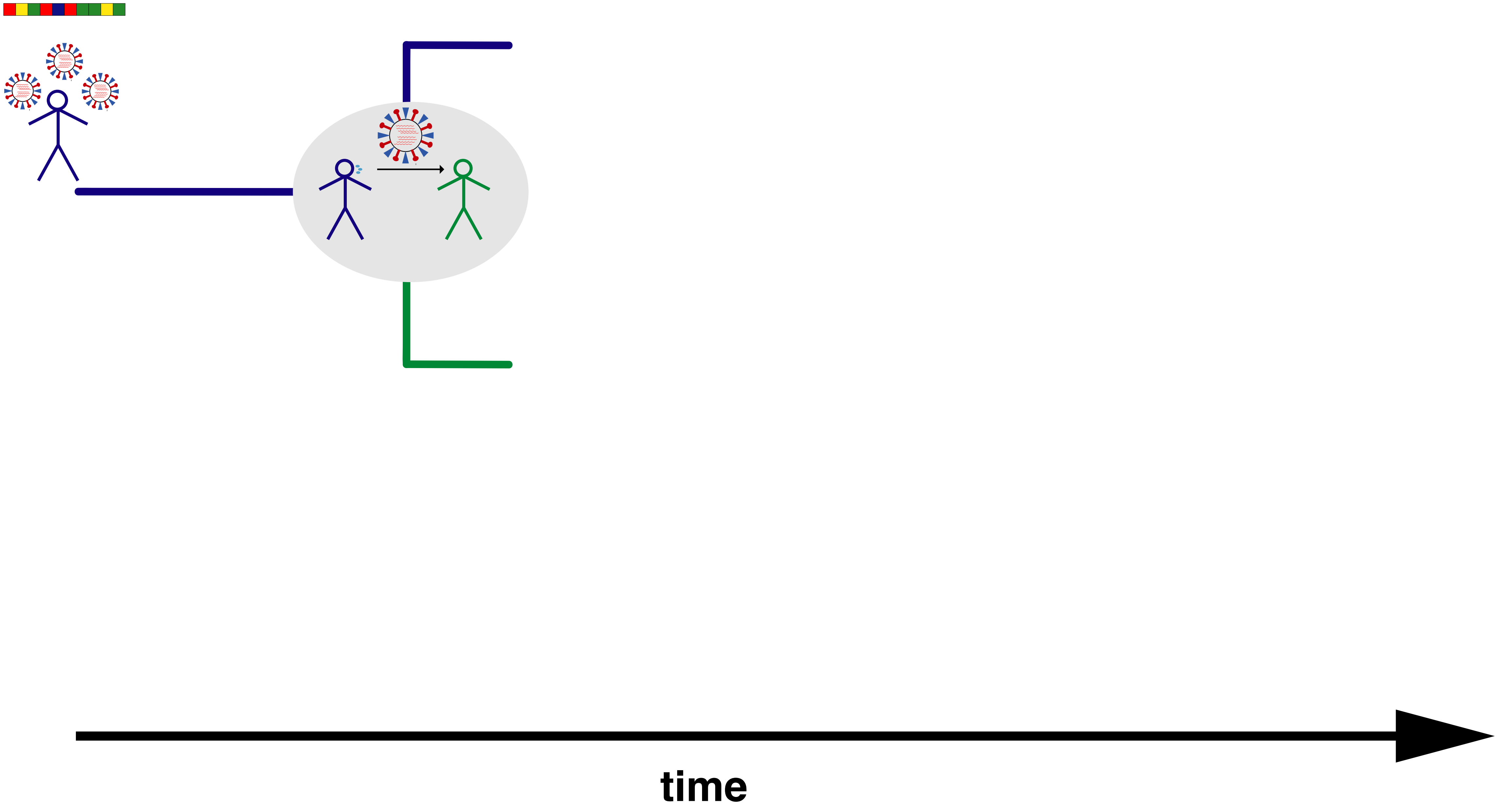
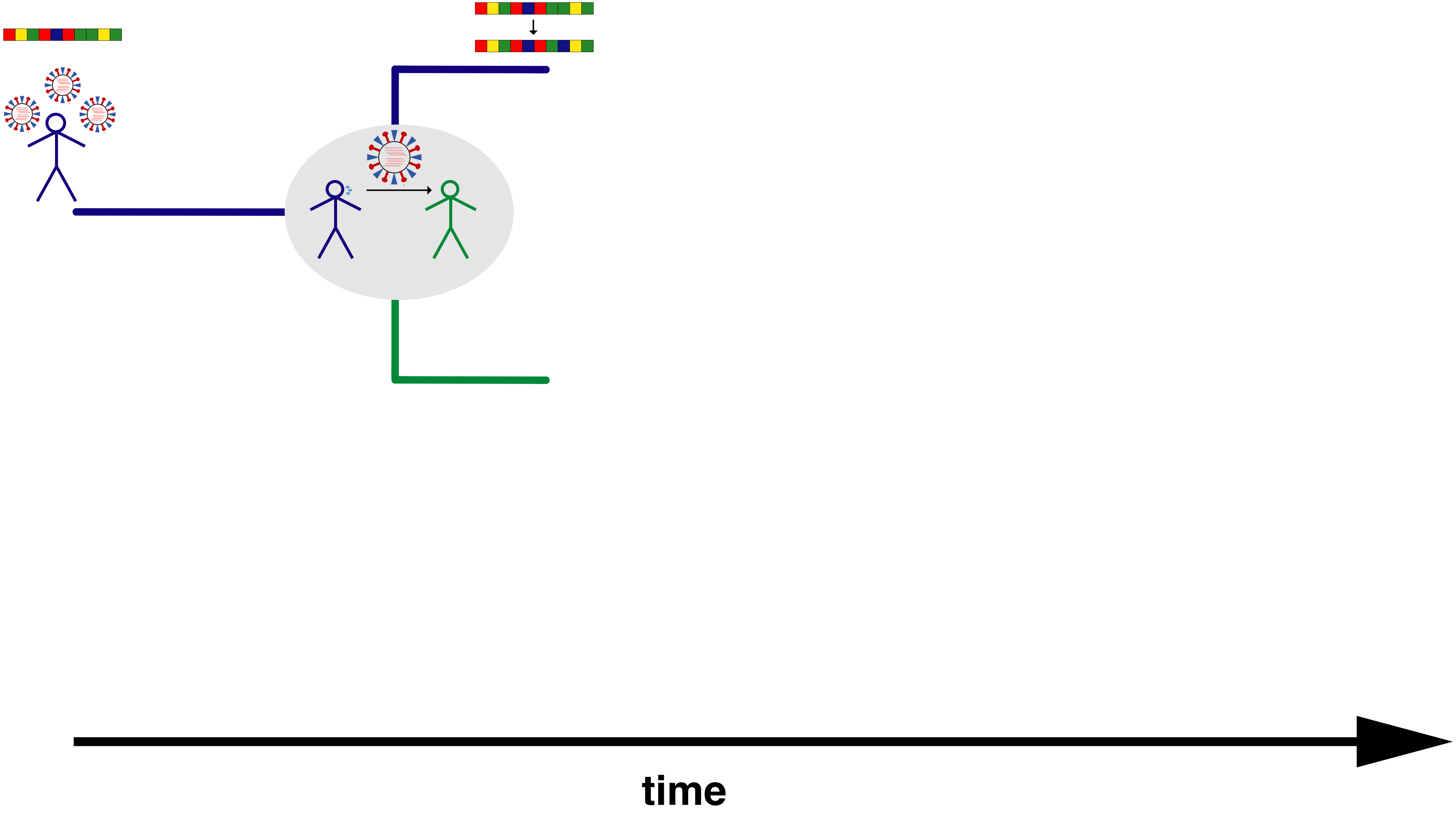
Tracking the evolution of pathogens over time

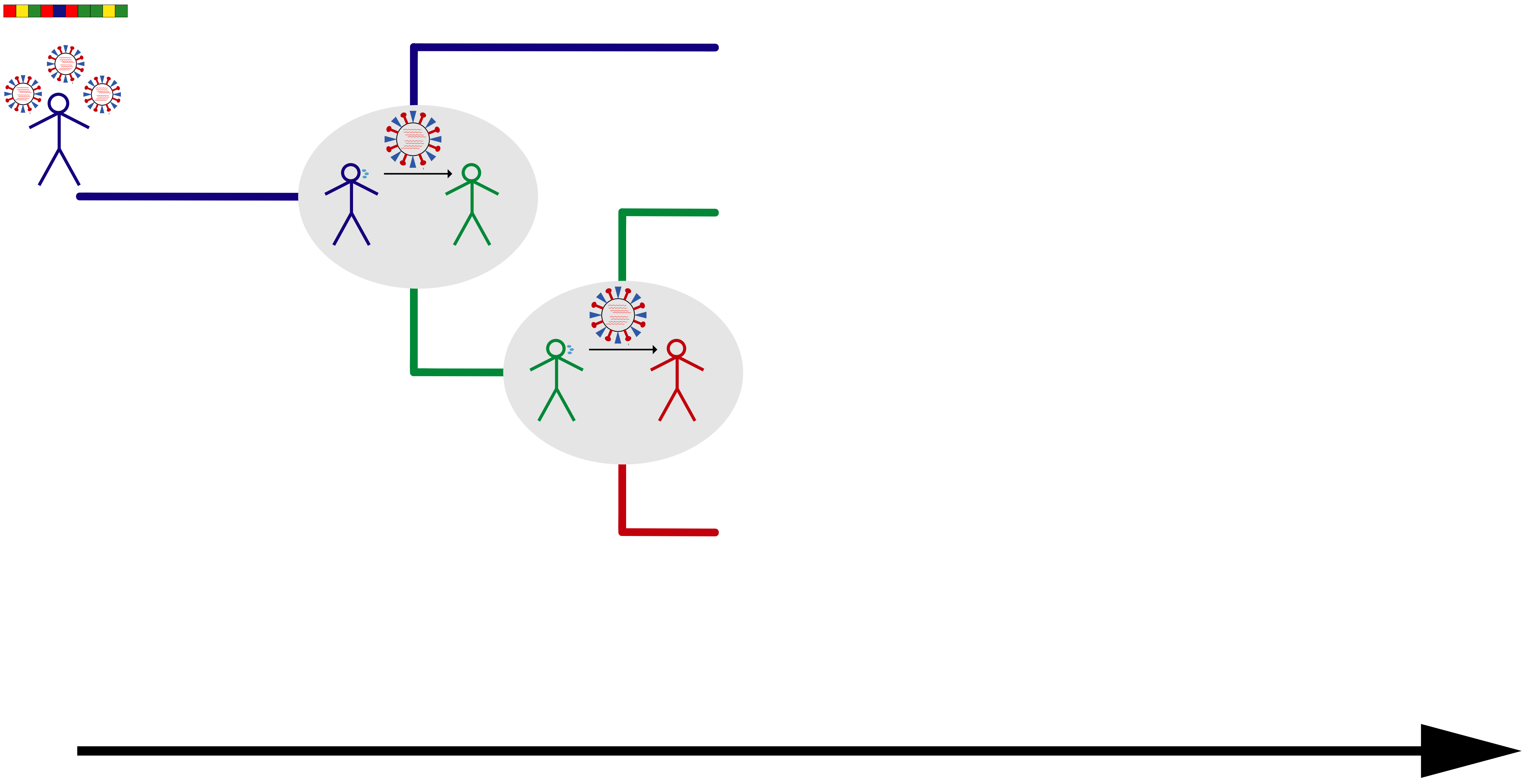
Nicola F. Müller

e-mail: nicola.mueller@ucsf.edu

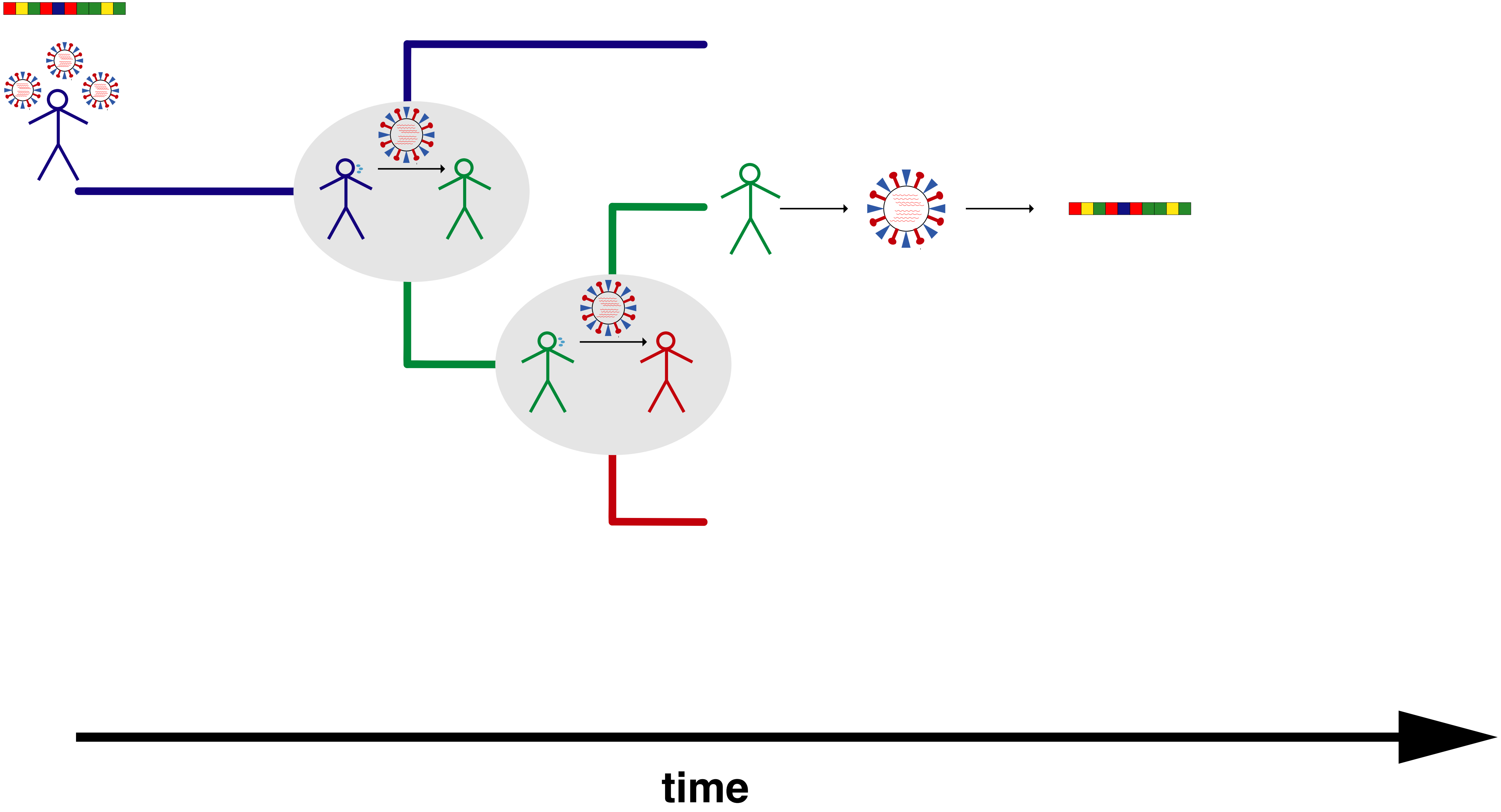
time


time

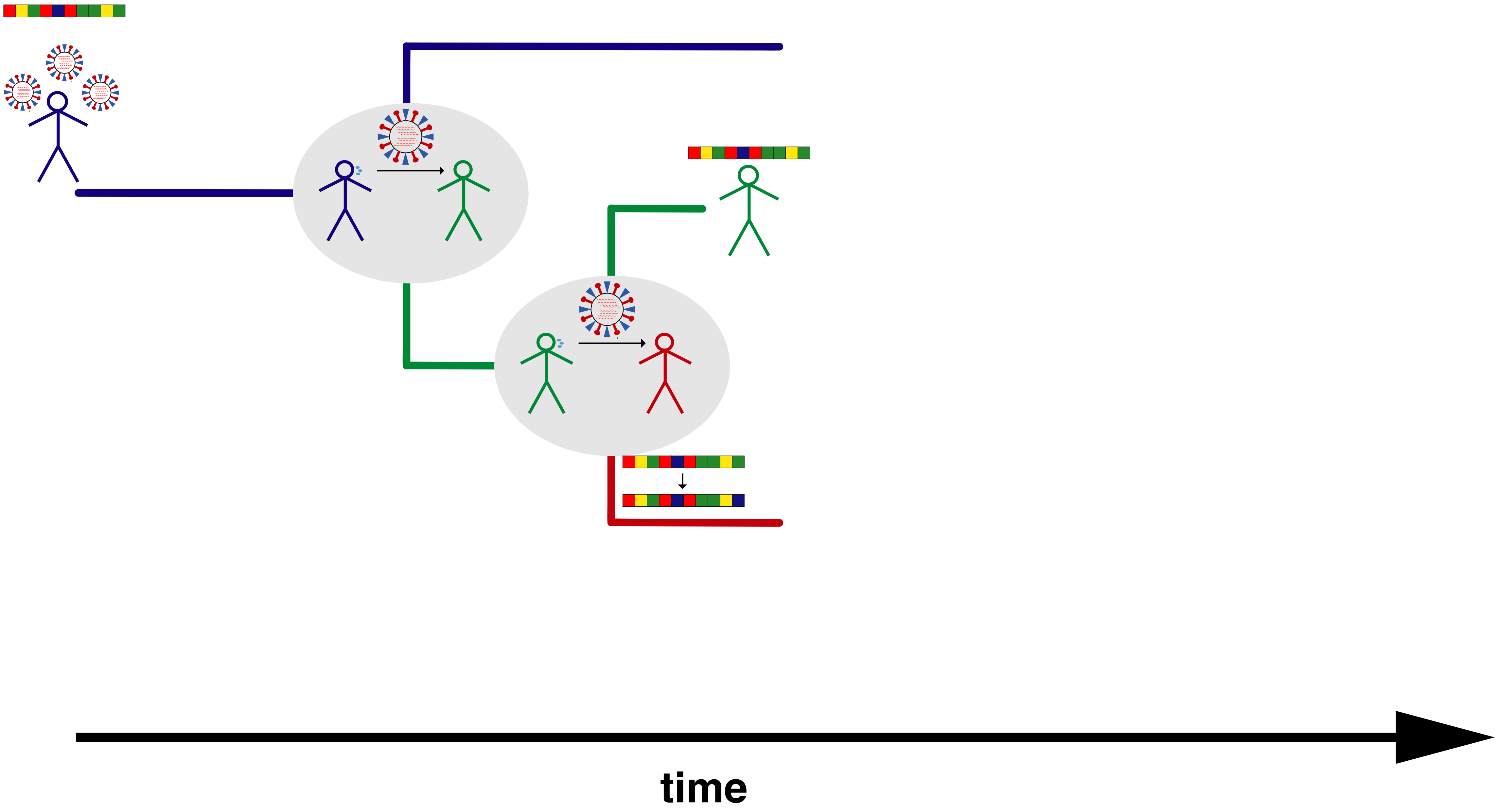

time

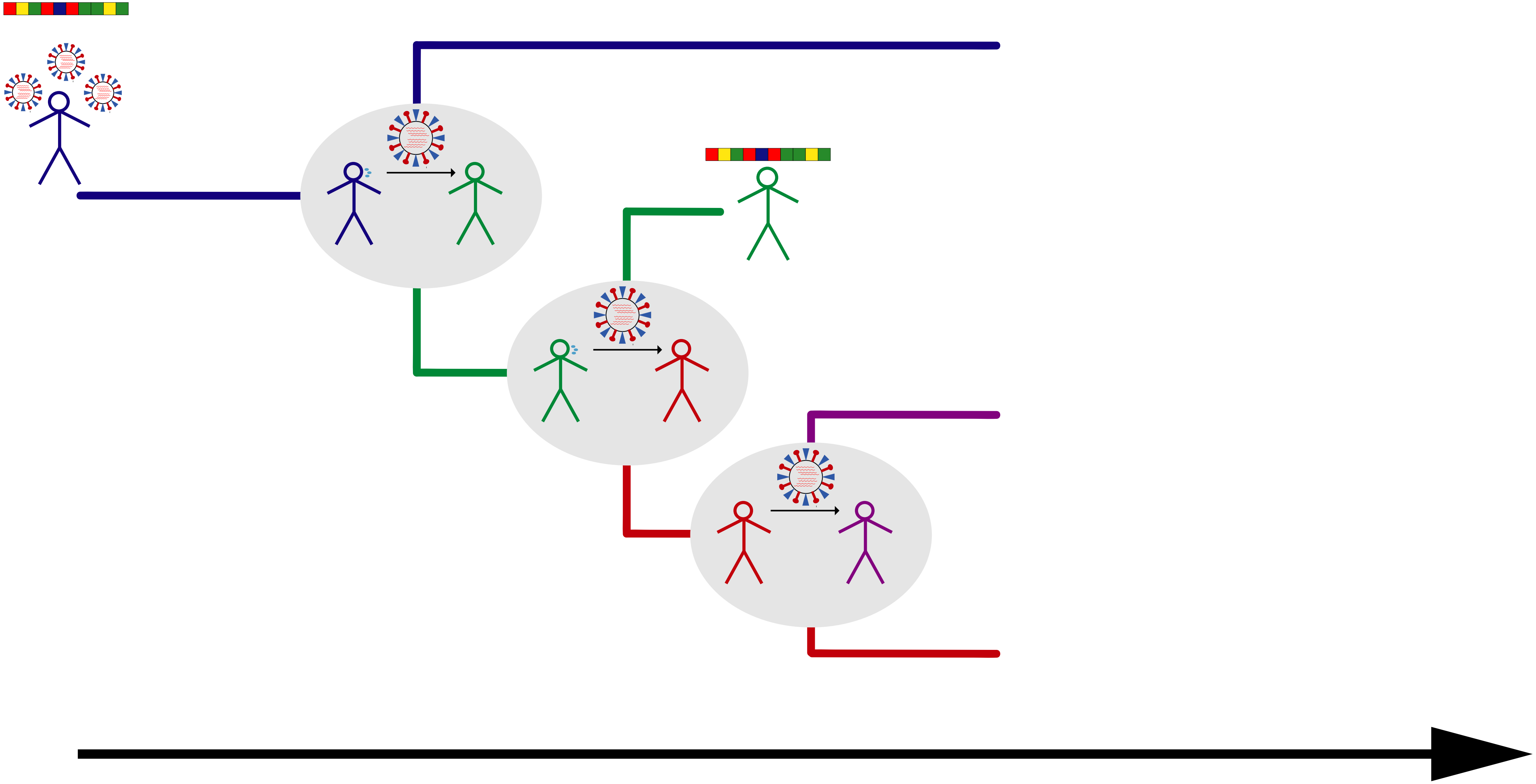


time

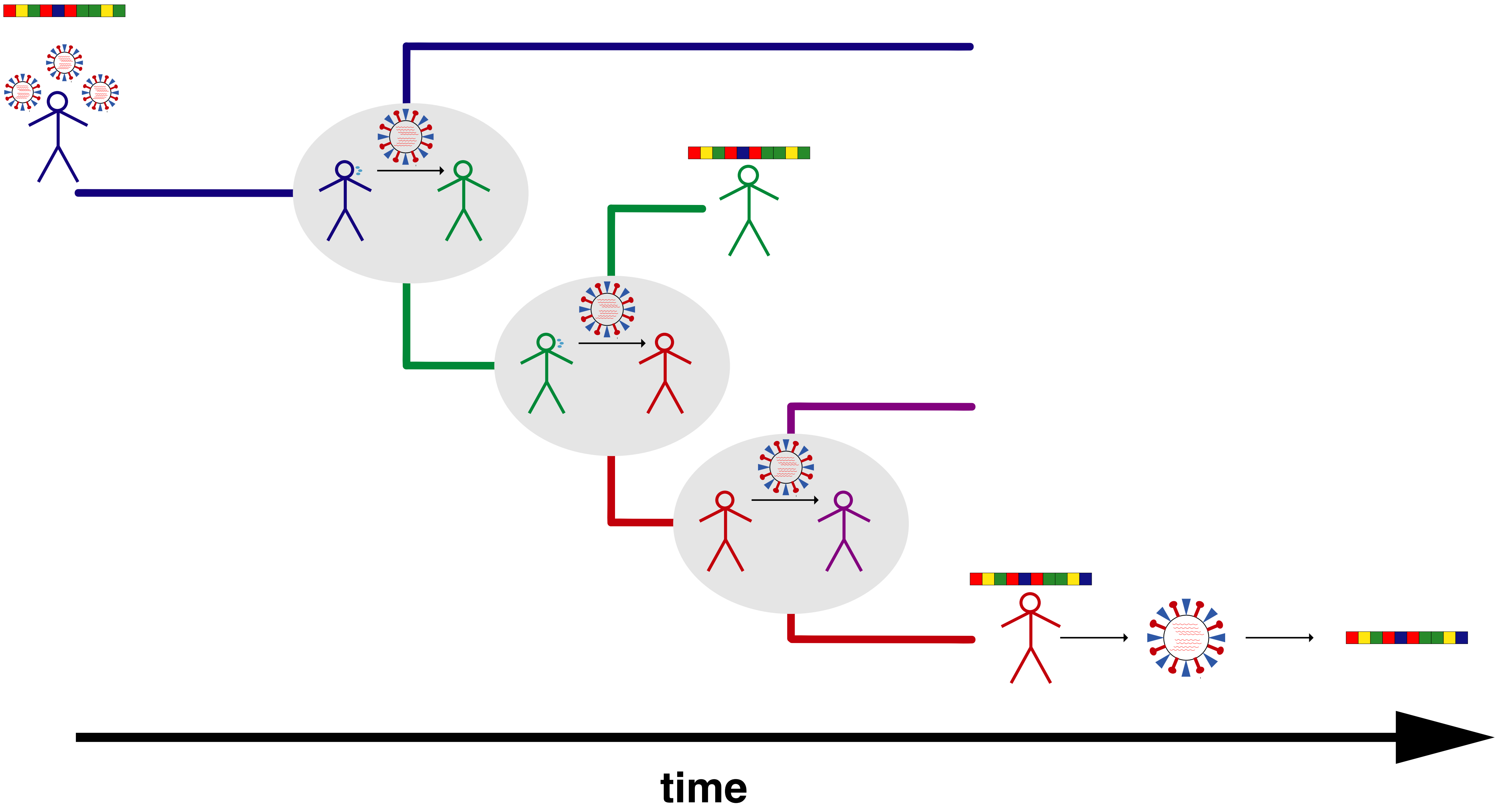


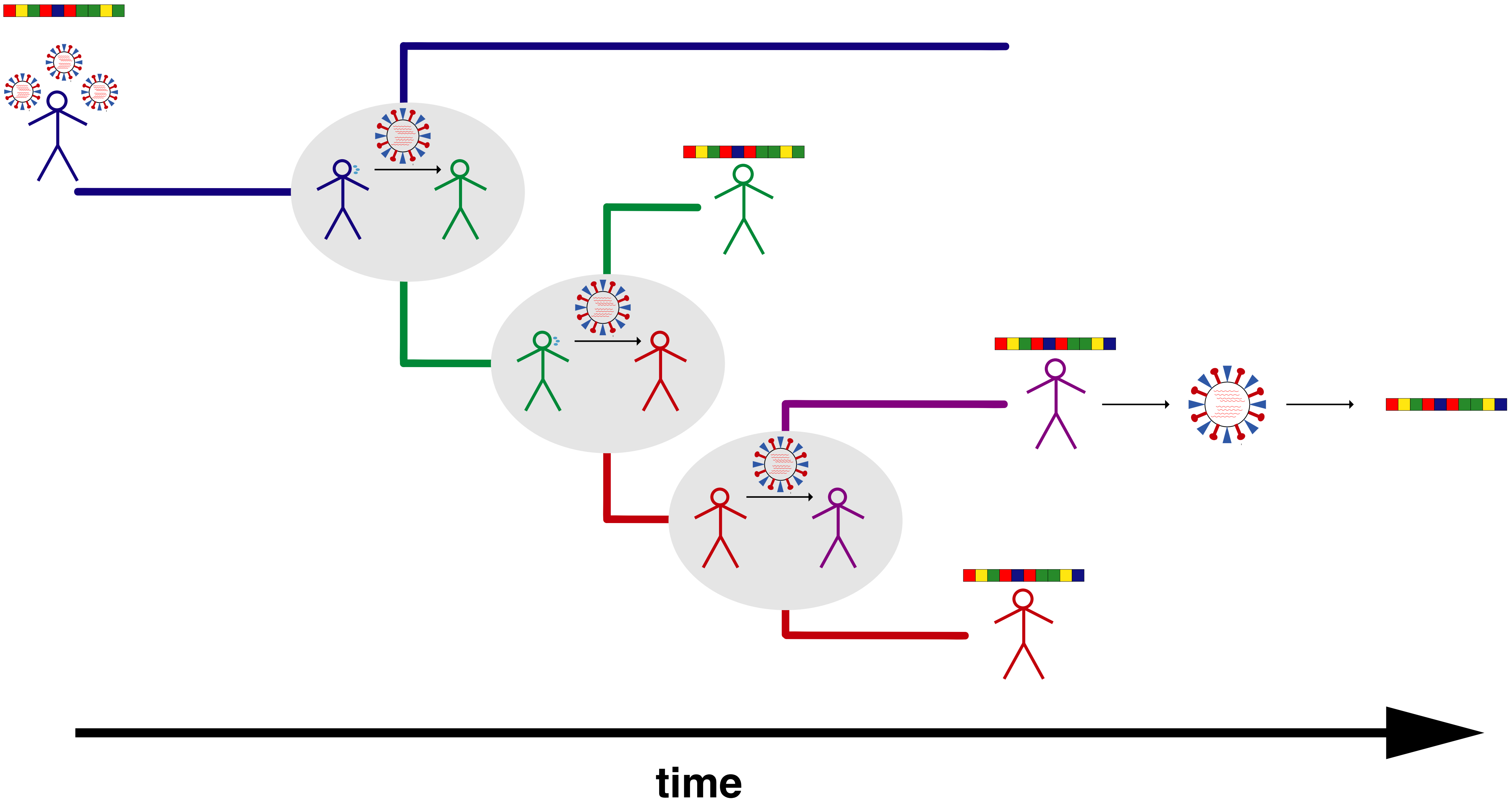
time

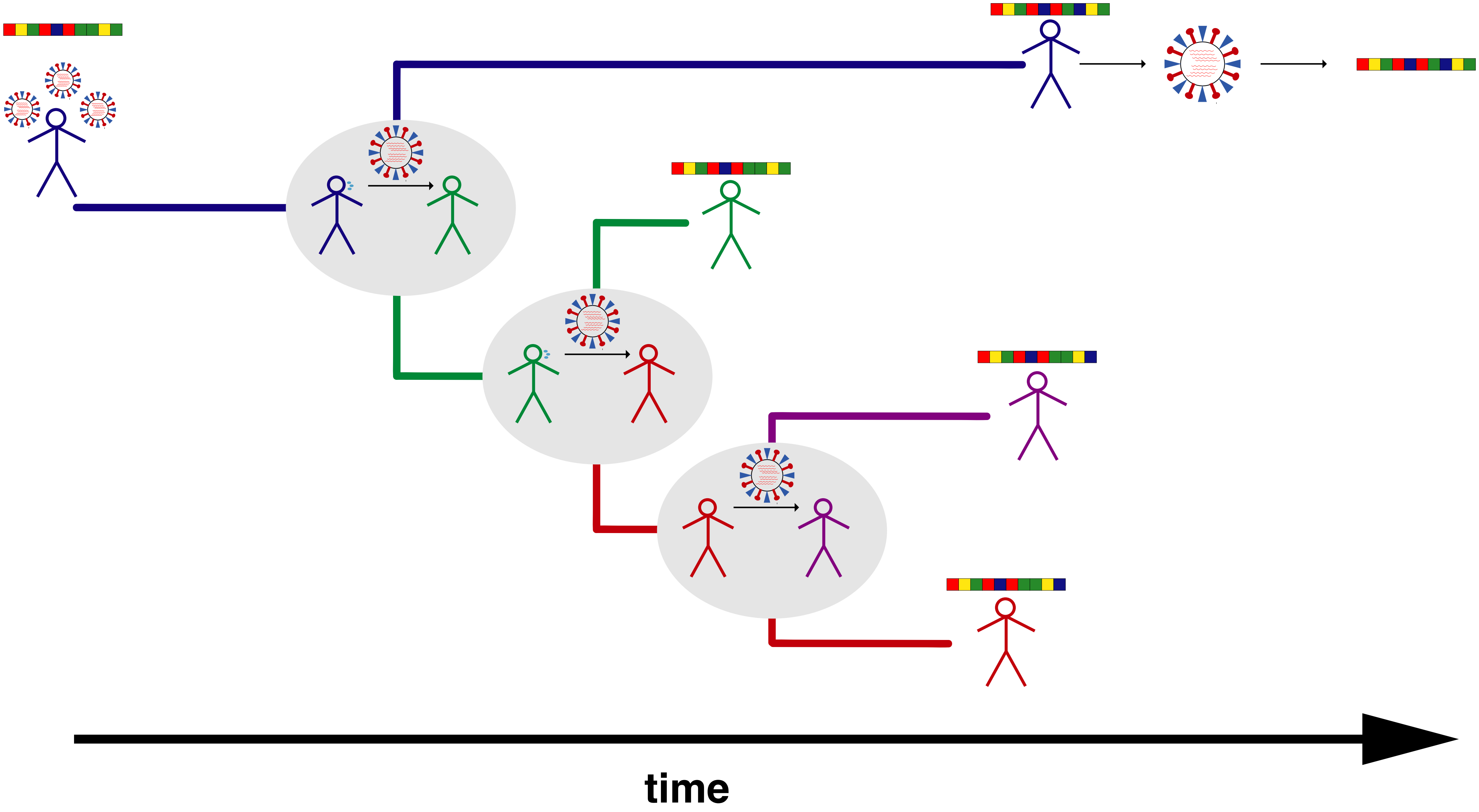

time

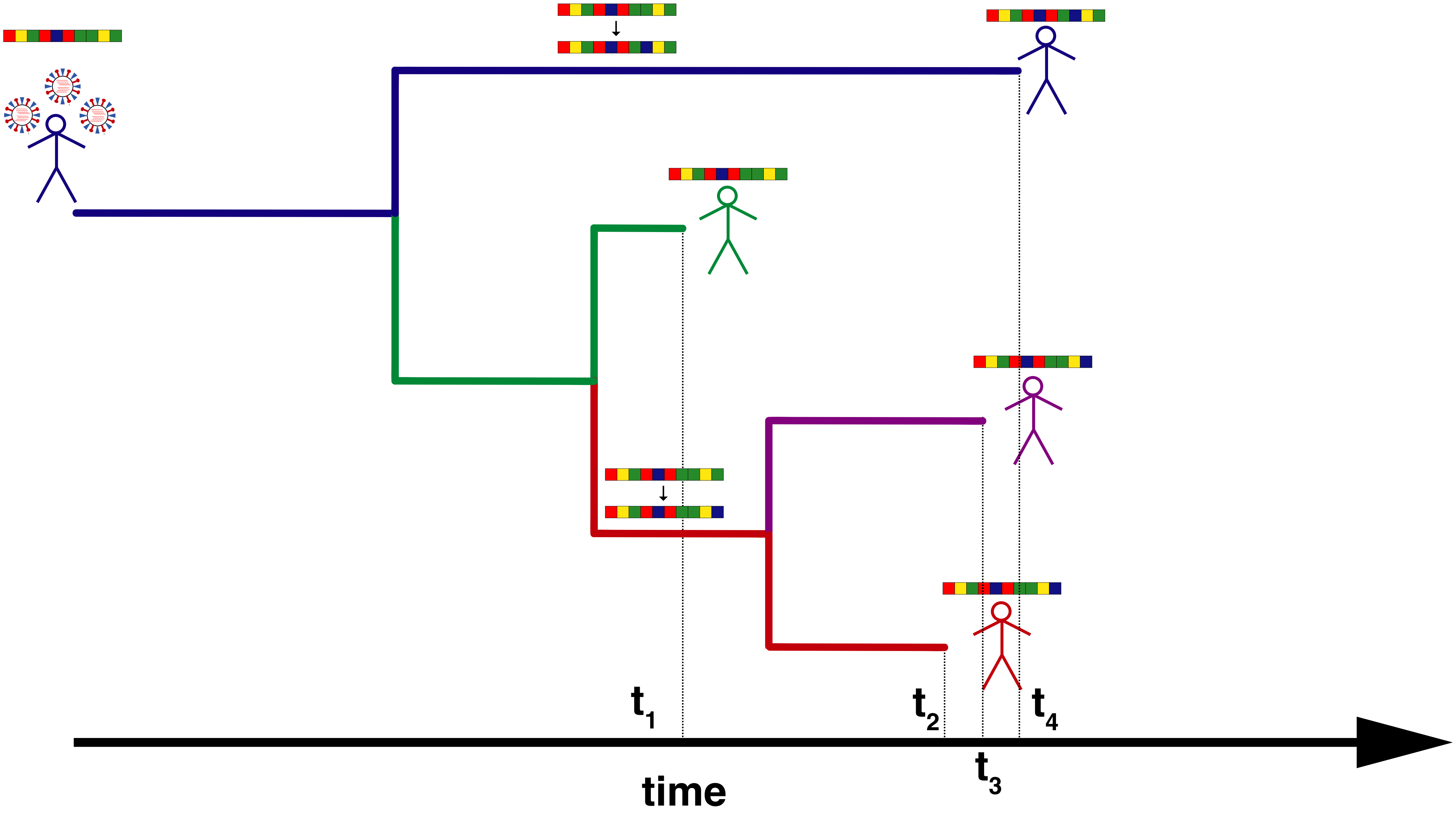


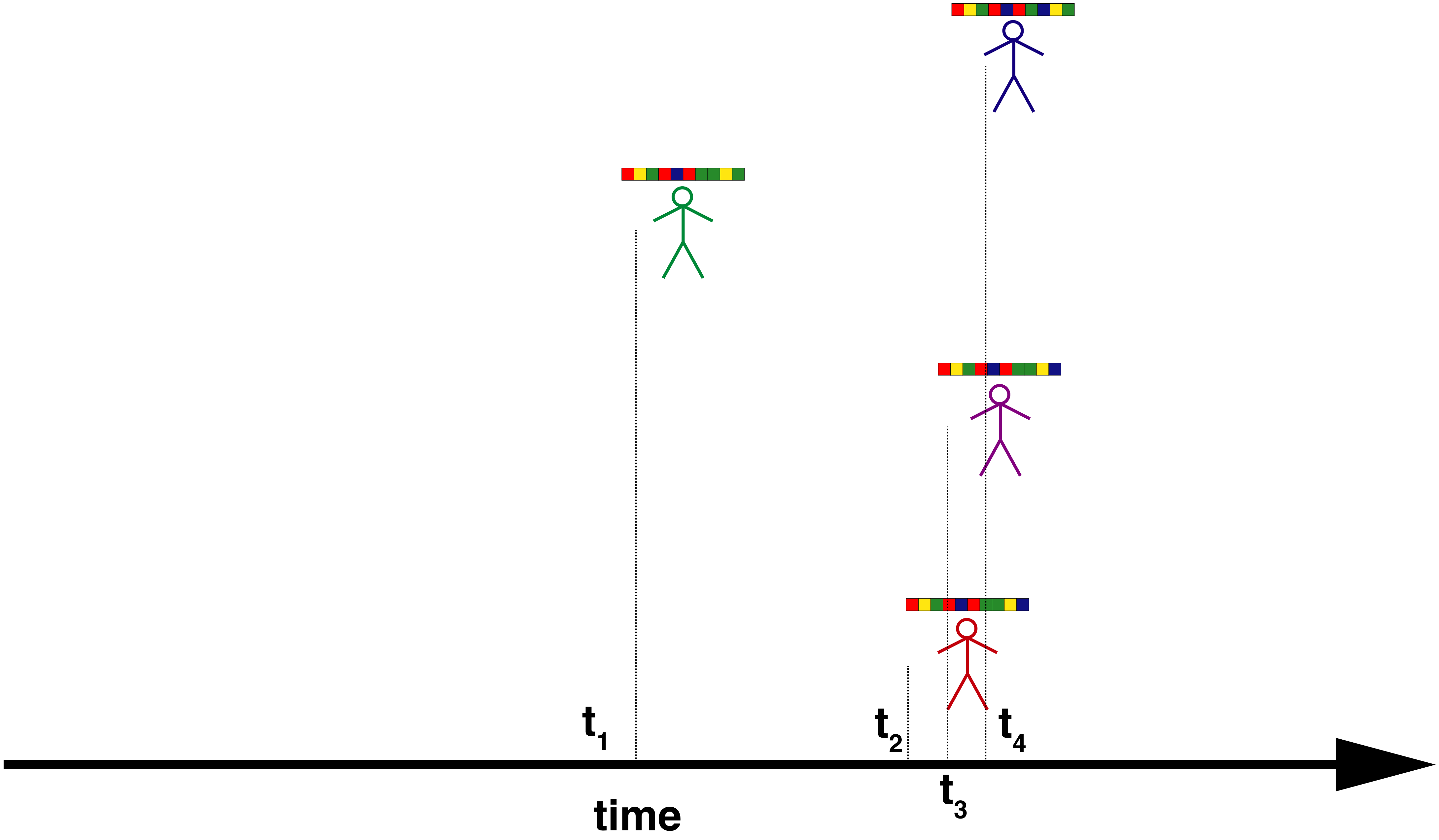
time


time


time

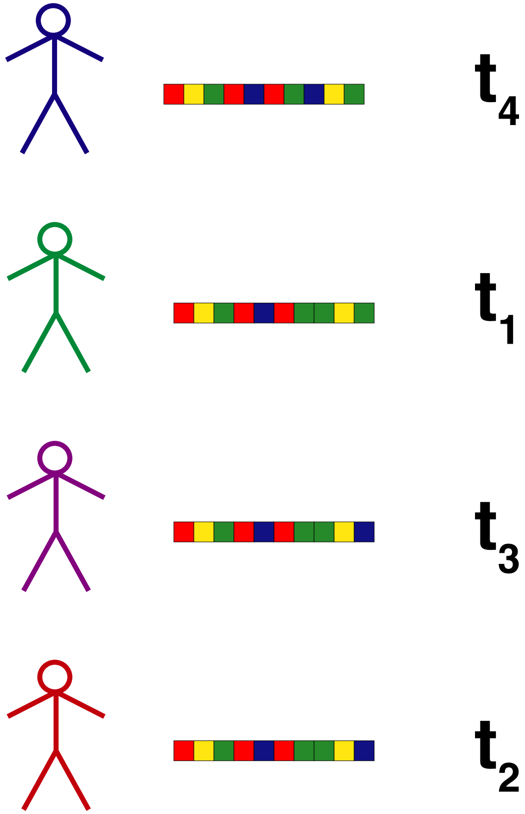

time


t_1
t_2
t_3
t_4
time

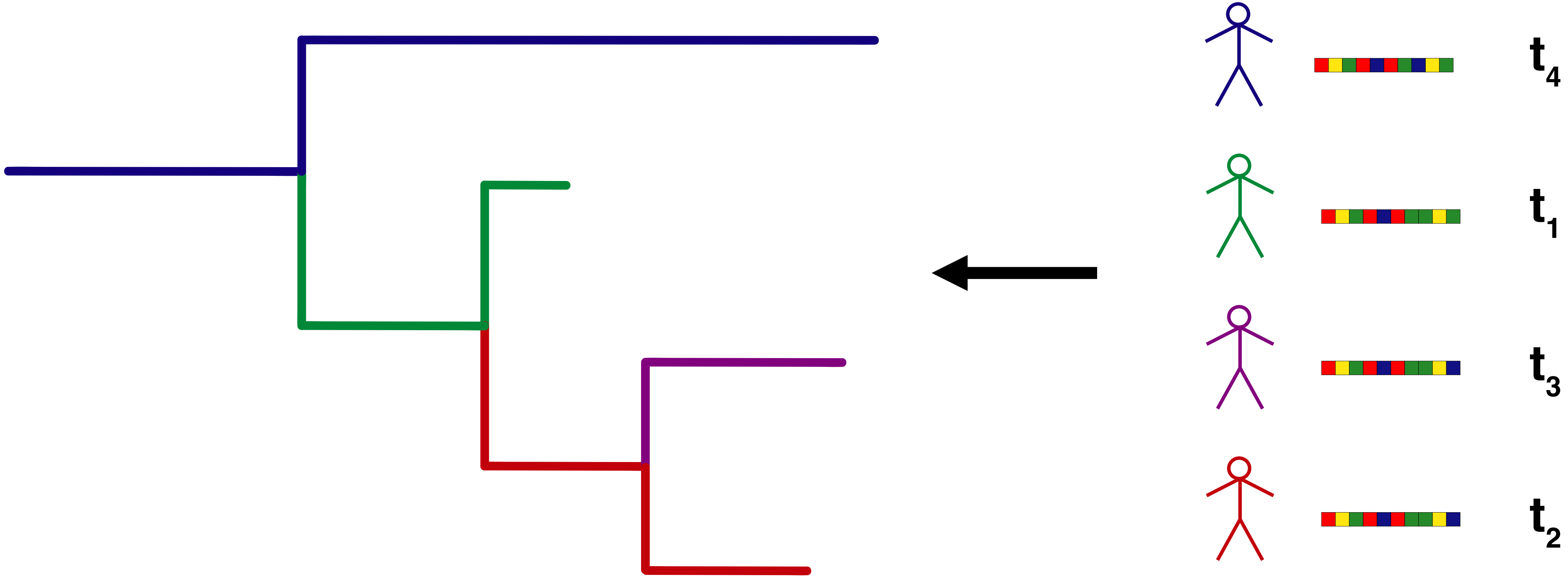


t_1
t_2
t_3
t_4
time

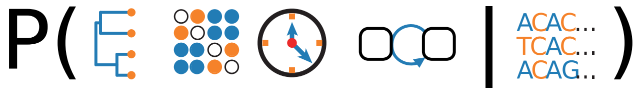
Our data: We know who was infected with which pathogen and sequence and when they were sampled



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



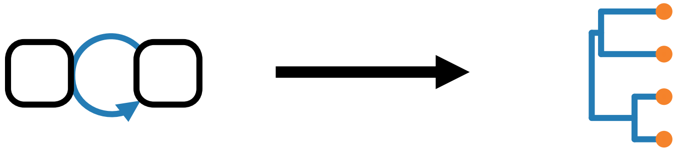
Bayesian phylogenetics allows us to jointly infer the phylogenetic trees, evolutionary and Demographics models

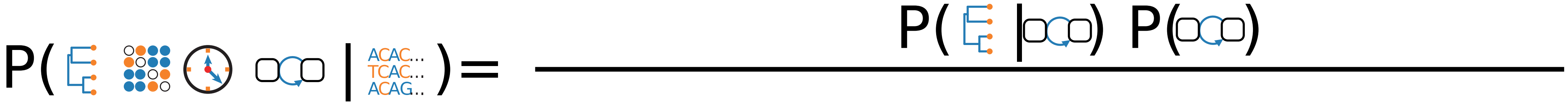


L. Du Plessis, T. Stadler “Getting to the root of epidemic spread with phylodynamic analysis of genomic data” Trends in Microbiology, 2015

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

Tree generating models



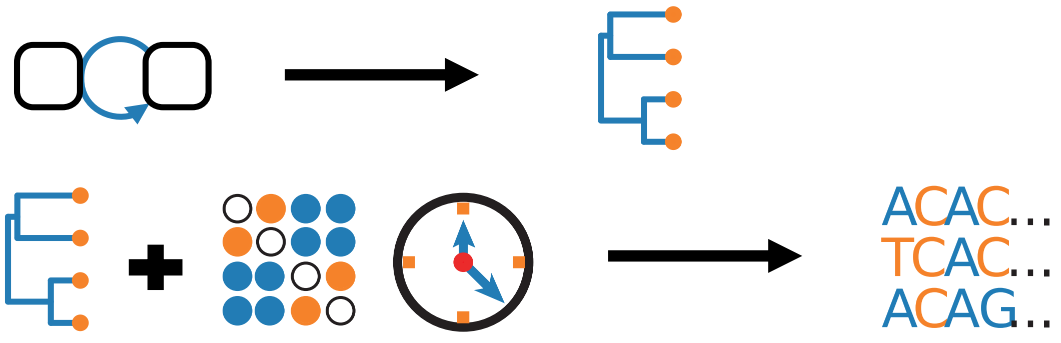


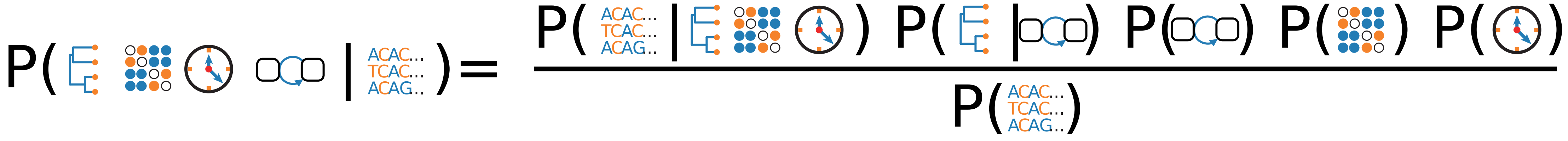
L. Du Plessis, T. Stadler “Getting to the root of epidemic spread with phylodynamic analysis of genomic data” Trends in Microbiology, 2015

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

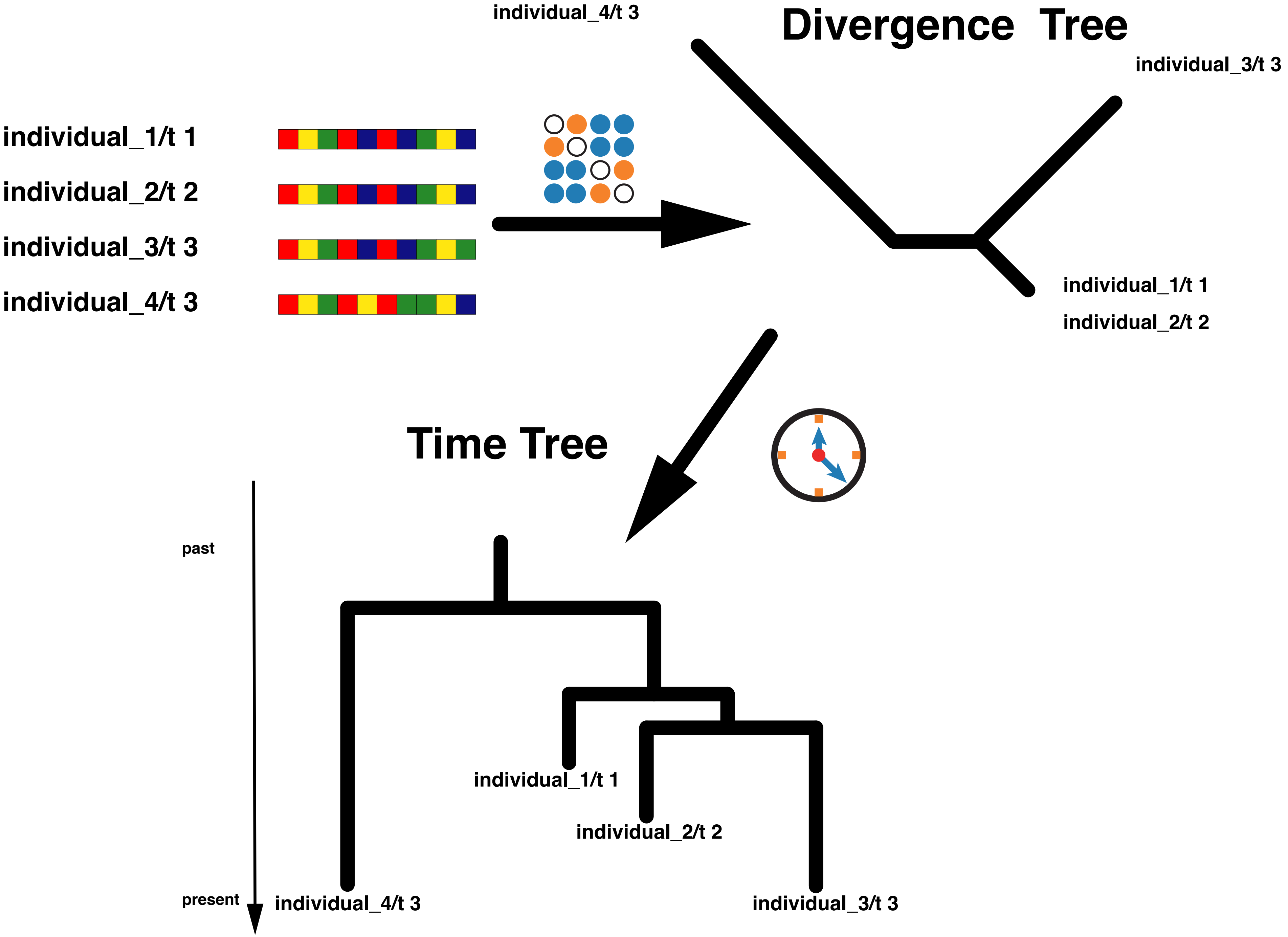
Tree generating models

Model of sequence evolution





L. Du Plessis, T. Stadler "Getting to the root of epidemic spread with phylodynamic analysis of genomic data" Trends in Microbiology, 2015



Divergence Tree

Time Tree

individual_1/t 1

individual_2/t 2

individual_3/t 3

individual_4/t 3

individual_4/t 3

individual_3/t 3

individual_1/t 1

individual_2/t 2

past

present

individual_1/t 1

individual_2/t 2

individual_4/t 3

individual_3/t 3

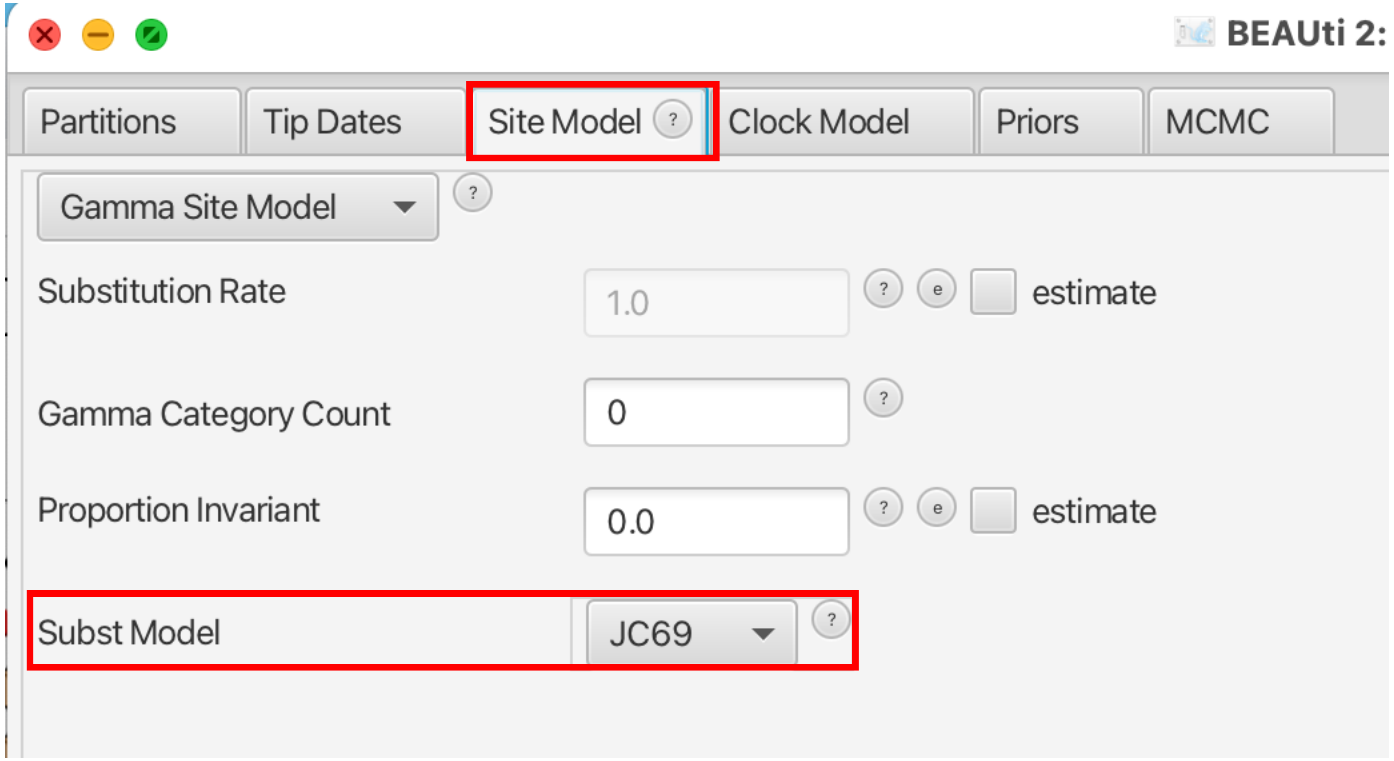
Site models describe the relative change across nucleotides and positions in the alignment and consist of several parts.

- **Substitution models** describe how fast/slow the change from one to another nucleotide happens compared to others

- **Gamma rate heterogeneity + invariant site models** describe how fast/slow some sites in an alignment change bases compared to others

- **Codon positions models** allow for some codon positions to evolve faster/slower than others

Substitution models describe how fast/slow the change from one to another nucleotide happens compared to others



BEAUti 2:

Gamma Site Model

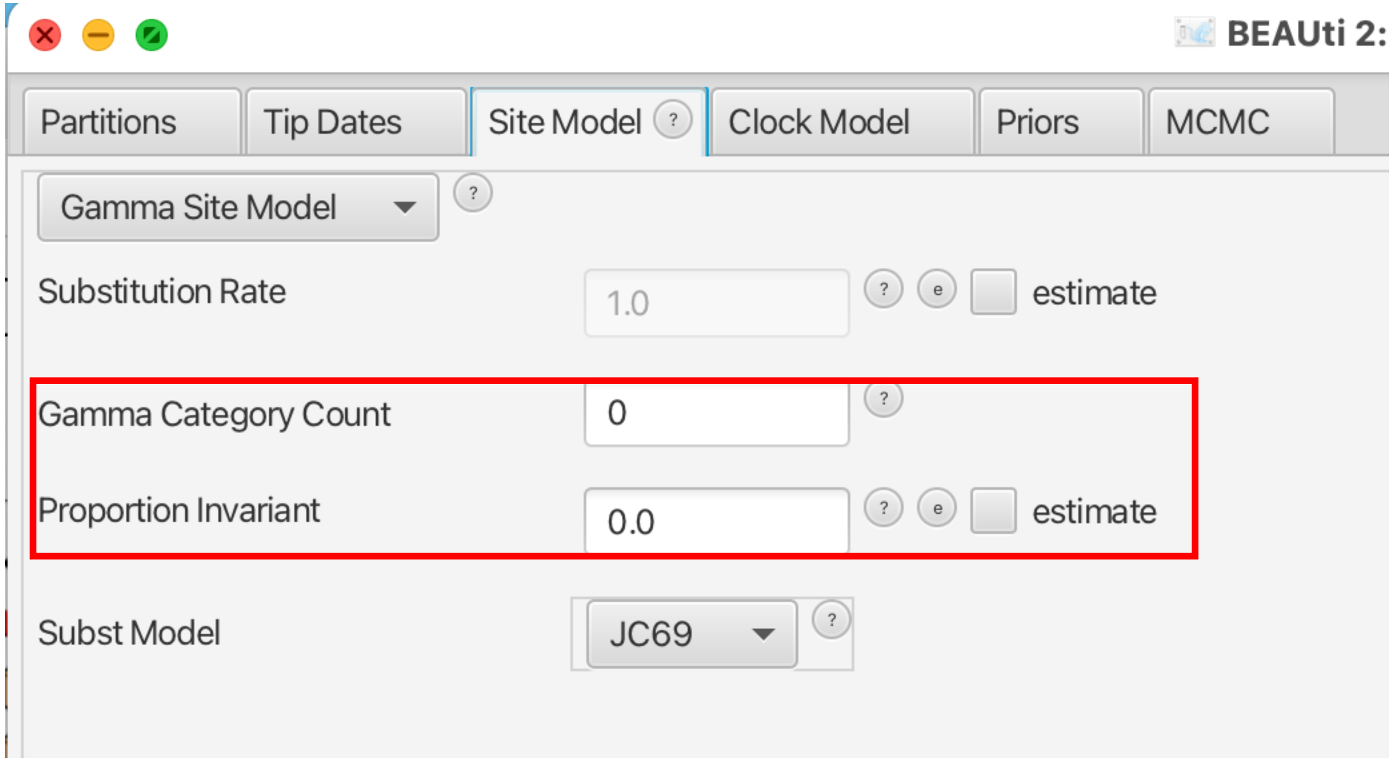
Substitution Rate 1.0 □ estimate

Gamma Category Count 0

Proportion Invariant 0.0 □ estimate

Subst Model JC69

Gamma rate heterogeneity + invariant site models describe how fast/slow some sites in an alignment change bases compared to others



Gamma Site Model

Substitution Rate 1.0 ? e □ estimate

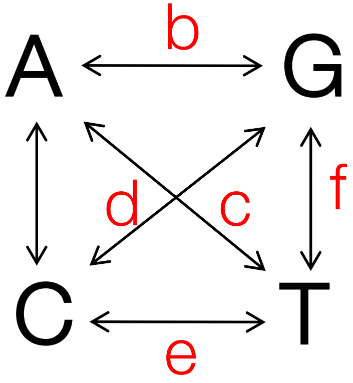
Gamma Category Count 0 ?

Proportion Invariant 0.0 ? e □ estimate

Subst Model JC69 ?

Substitution models allow to account for differences in nucleotide substitution rates.

Rate Matrix



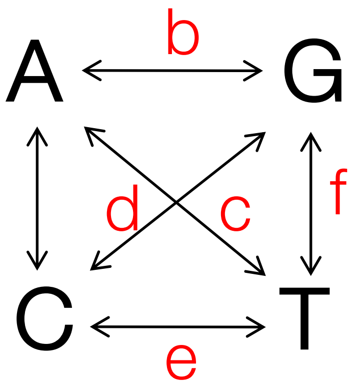
Equilibrium Base Frequencies

$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

(slide from Sebastian Duchene)

Substitution models allow to account for differences in nucleotide substitution rates.

Rate Matrix



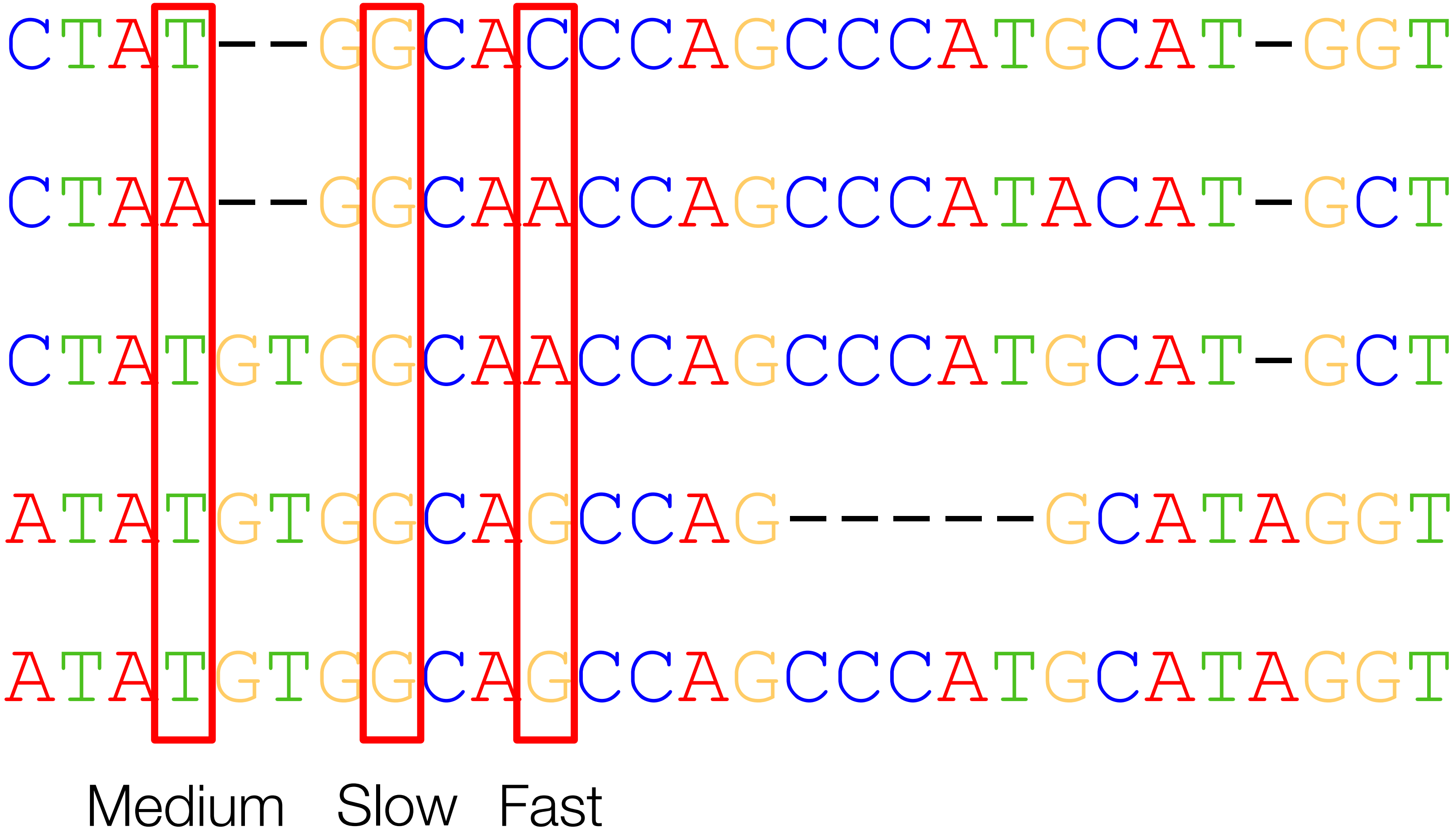
Equilibrium Base Frequencies

$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

JC	**HKY**	**GTR**
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	π_A, π_C, π_G, π_T	π_A, π_C, π_G, π_T
No I or G	No I or G	No I or G
0 free parameters	4 free parameters	8 free parameters

(slide from Sebastian Duchene)

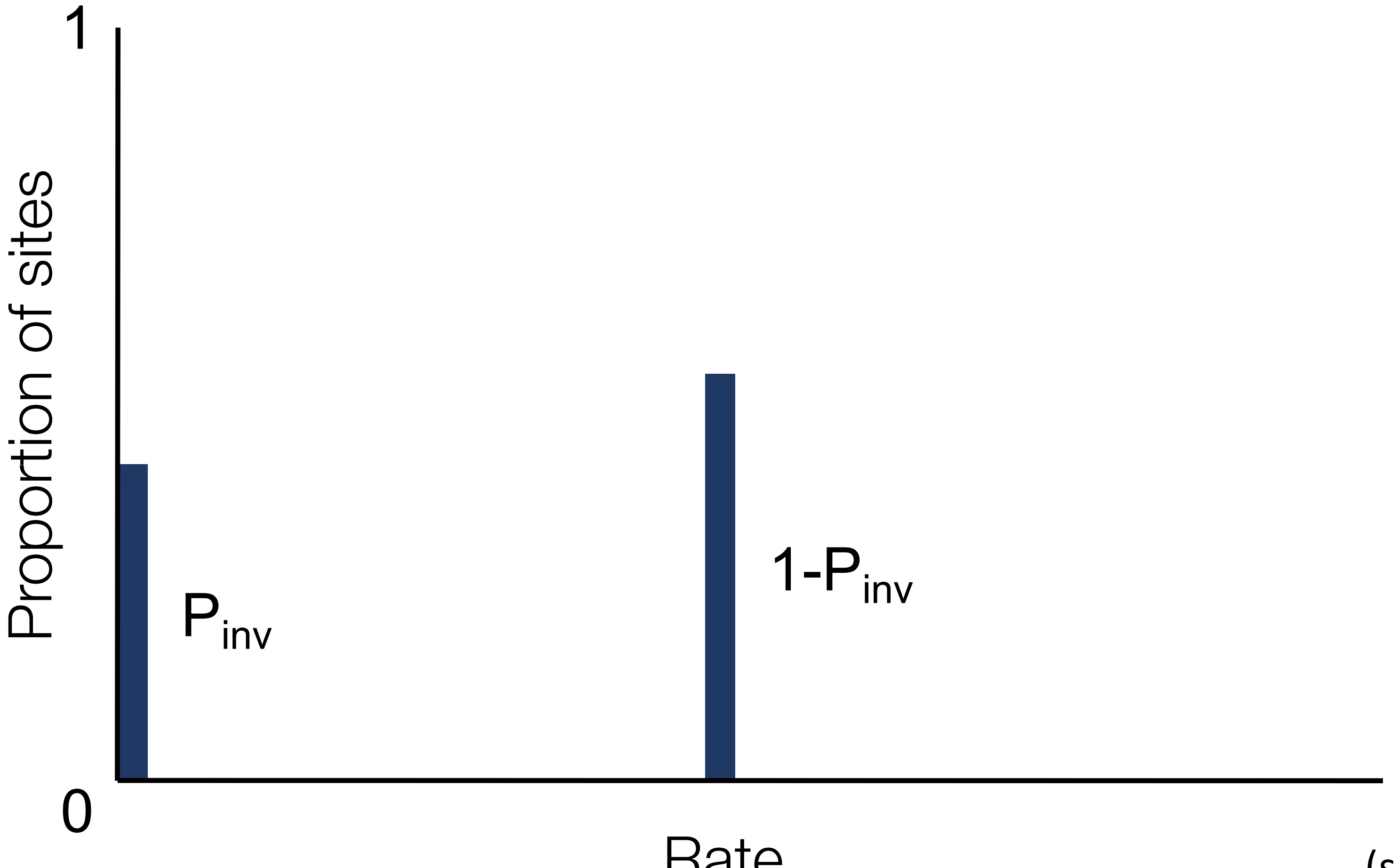
Some site in the alignment might be more flexible and therefore evolve less quickly.



```
CTAT--GGCACCCAGCCCATGCAT-GGT
CTAA--GGCAACCAGCCCATACAT-GCT
CTATGTGGCAACCAGCCCATGCAT-GCT
ATATGTGGCAGCCAG-----GCATAGGT
ATATGTGGCAGCCAGCCCATGCATAGGT
```

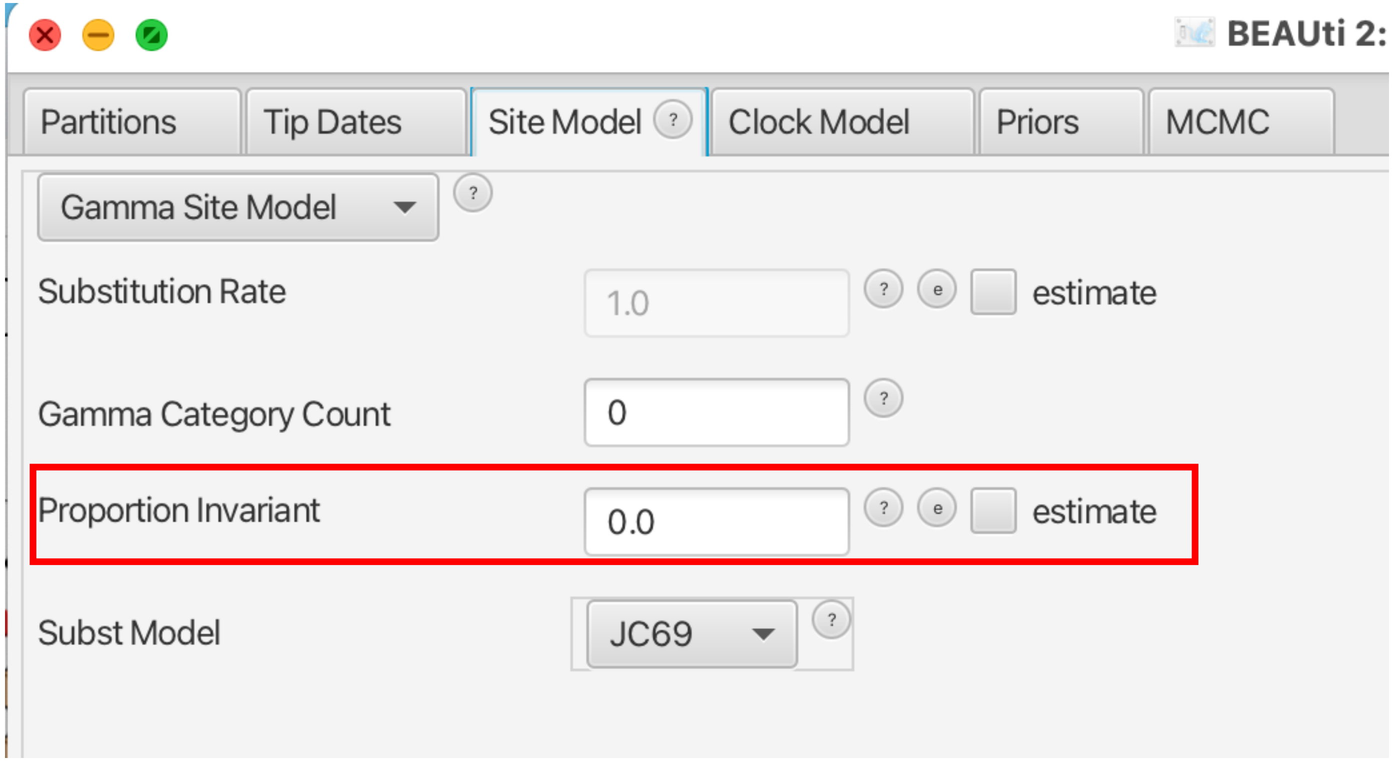
(slide from Sebastian Duchene)

Proportion of invariable sites account for sites that do not change (e.g., HKY+I).



(slide from Sebastian Duchene)

Gamma rate heterogeneity + invariant site models describe how fast/slow some sites in an alignment change bases compared to others



Gamma Site Model

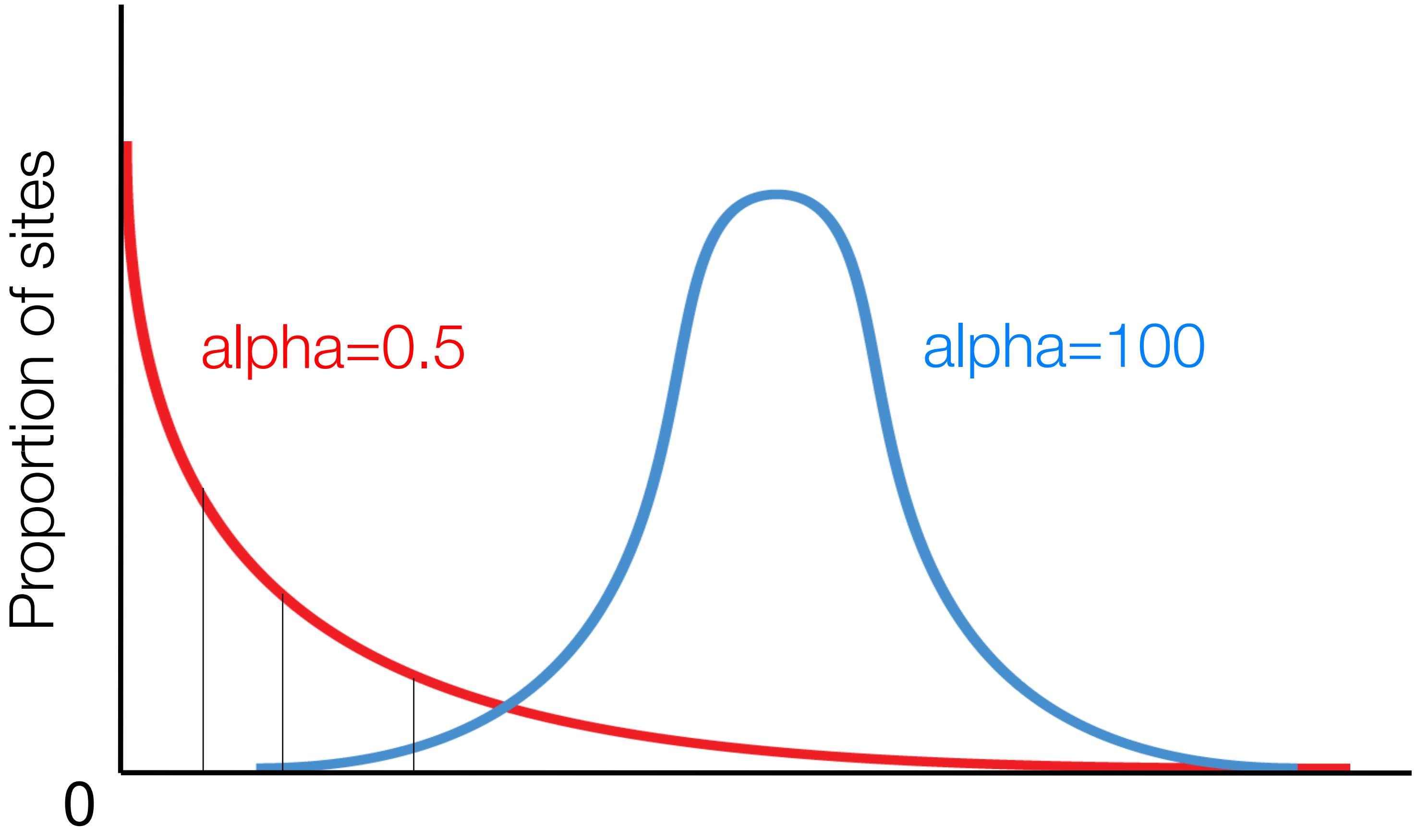
Substitution Rate 1.0 □ estimate

Gamma Category Count 0

Proportion Invariant 0.0 □ estimate

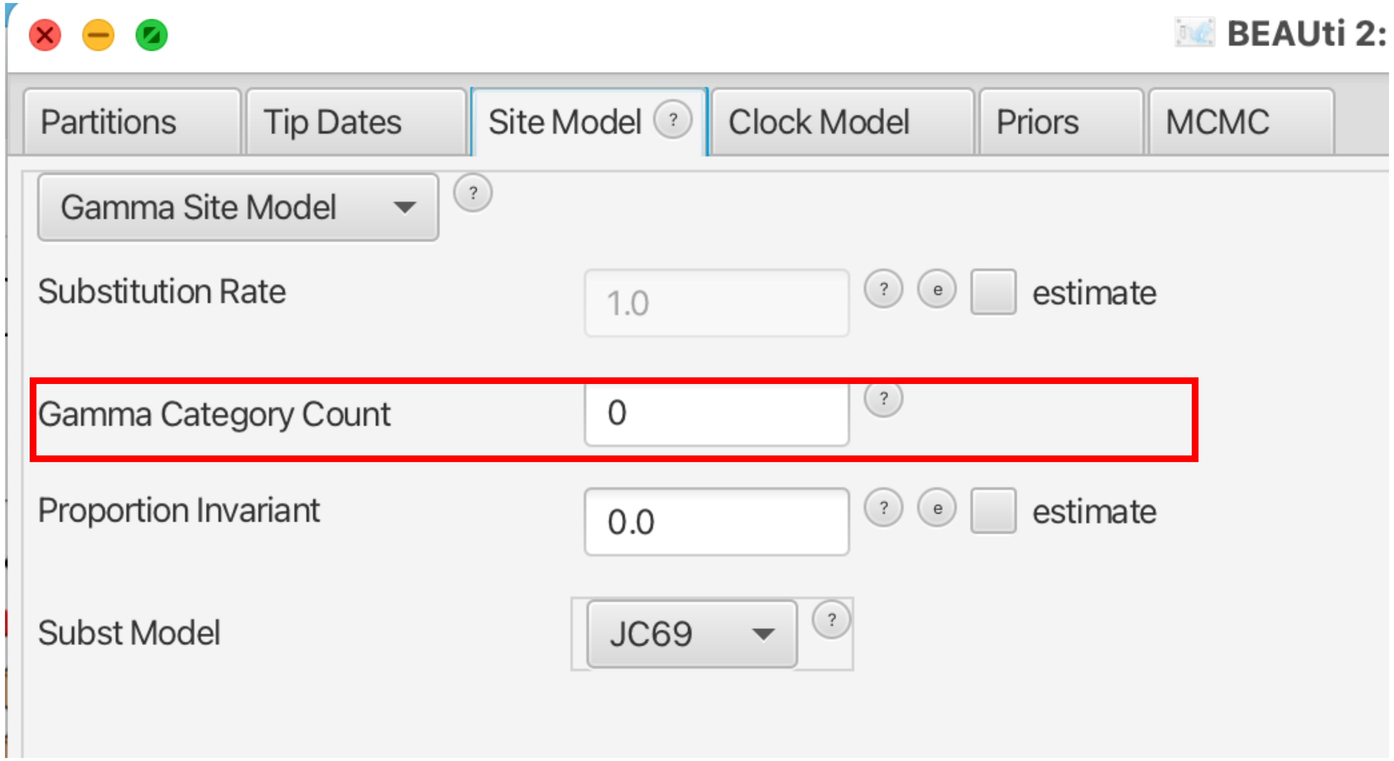
Subst Model JC69

Gamma rate heterogeneity models account for rate differences across sites (e.g. HKY+G_4).



(slide from Sebastian Duchene)

Gamma rate heterogeneity + invariant site models describe how fast/slow some sites in an alignment change bases compared to others



Gamma Site Model

Substitution Rate 1.0 □ estimate

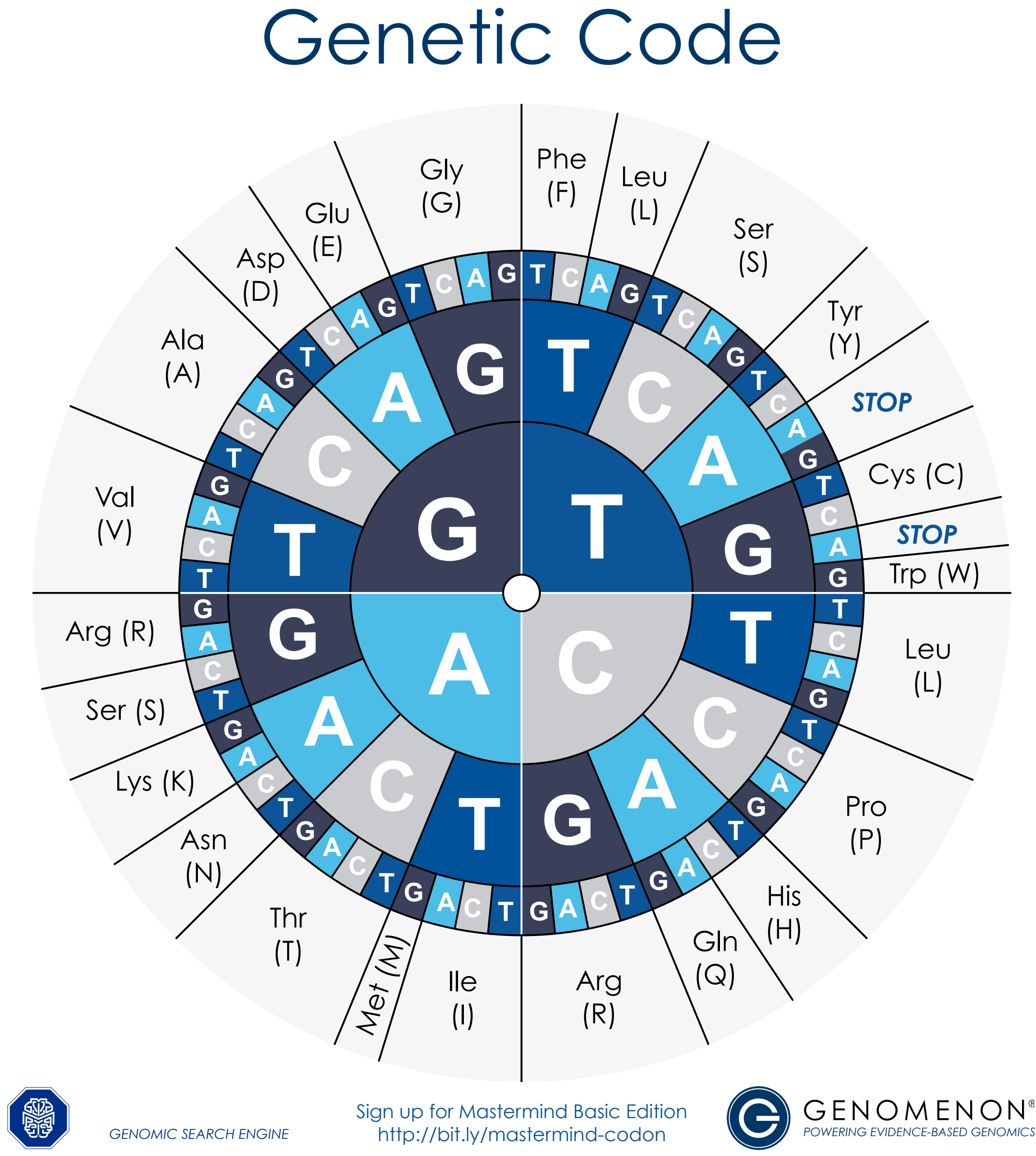
Gamma Category Count 0

Proportion Invariant 0.0 □ estimate

Subst Model JC69

Changes in the third codon position are far more likely to not affect the amino acid

- Splitting up the alignment into different codon positions and allow each having its own site model allows accounting for these differences



https://www.genomenon.com/codon-chart/

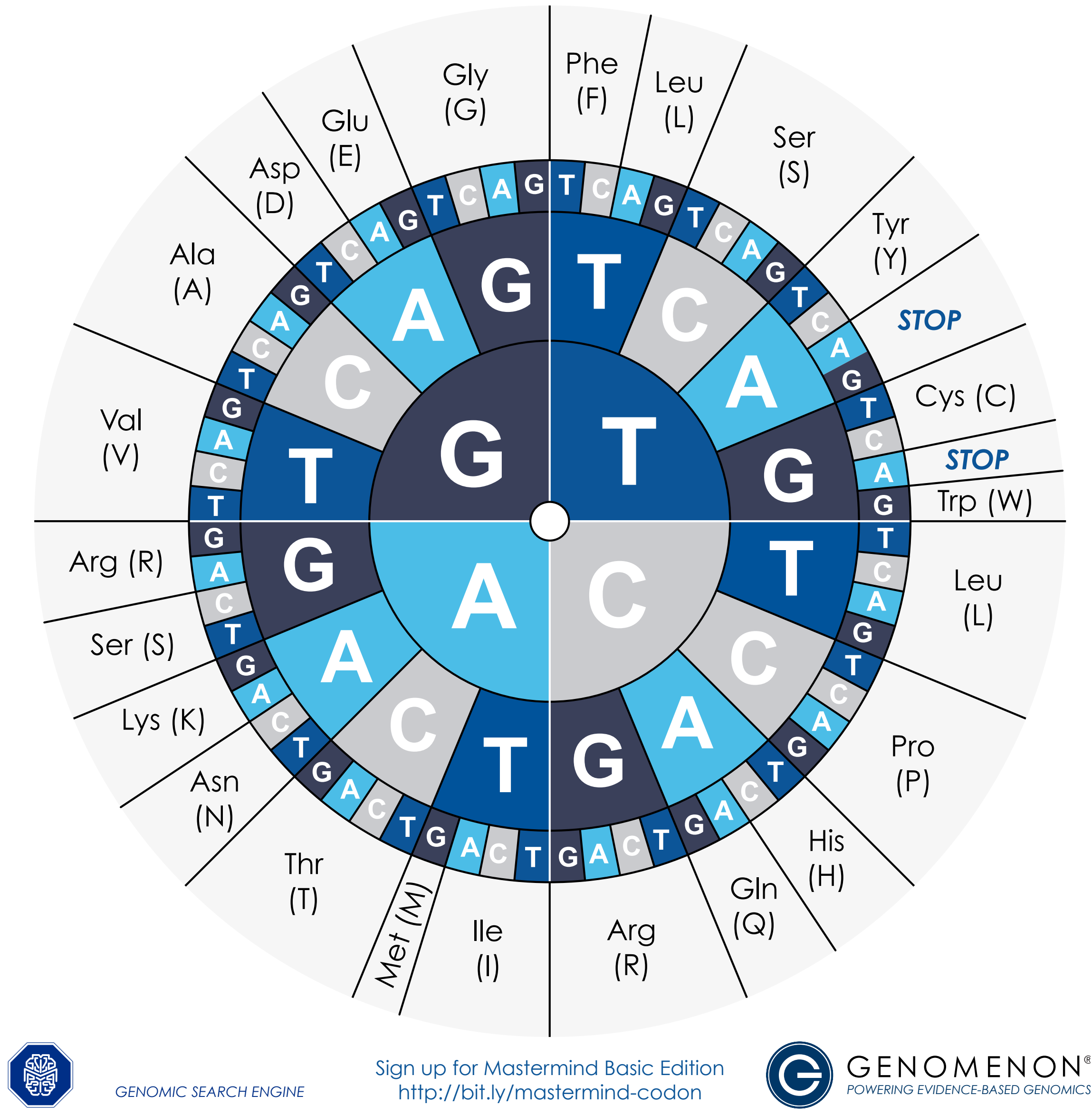
Genetic Code

CTA (Leucine)

First position:
ATA (Isoleucine)
GTA (Valine)
TTA (Leucine)

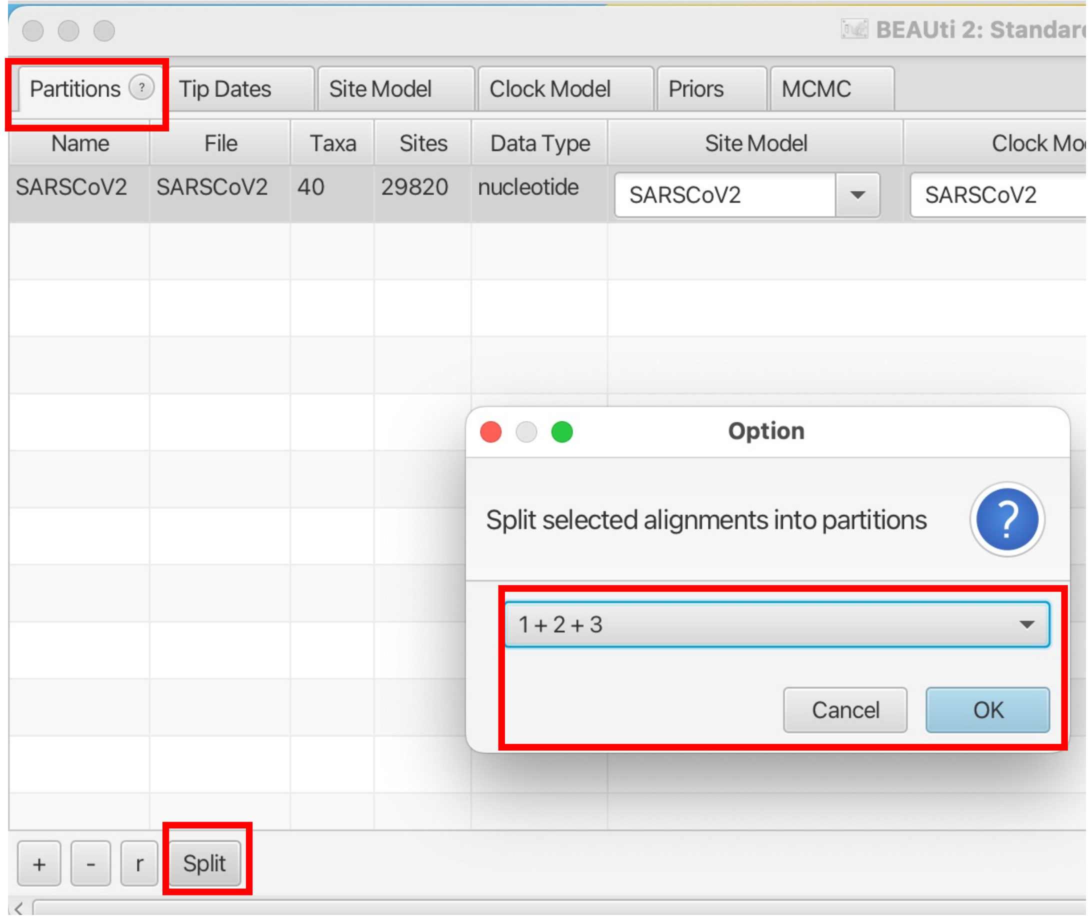
Second position:
CCA (Proline)
CAA (Glutamine)
CGA (Arginine)

Third position:
CTC (Leucine)
CTG (Leucine)
CTT (Leucine)



https://www.genomenon.com/codon-chart/

Codon positions models allow for some codon positions to evolve faster/slower than others



Name	File	Taxa	Sites	Data Type	Site Model	Clock Mo
SARSCoV2	SARSCoV2	40	29820	nucleotide	SARSCoV2	SARSCoV2

Split selected alignments into partitions

1 + 2 + 3

Practical considerations

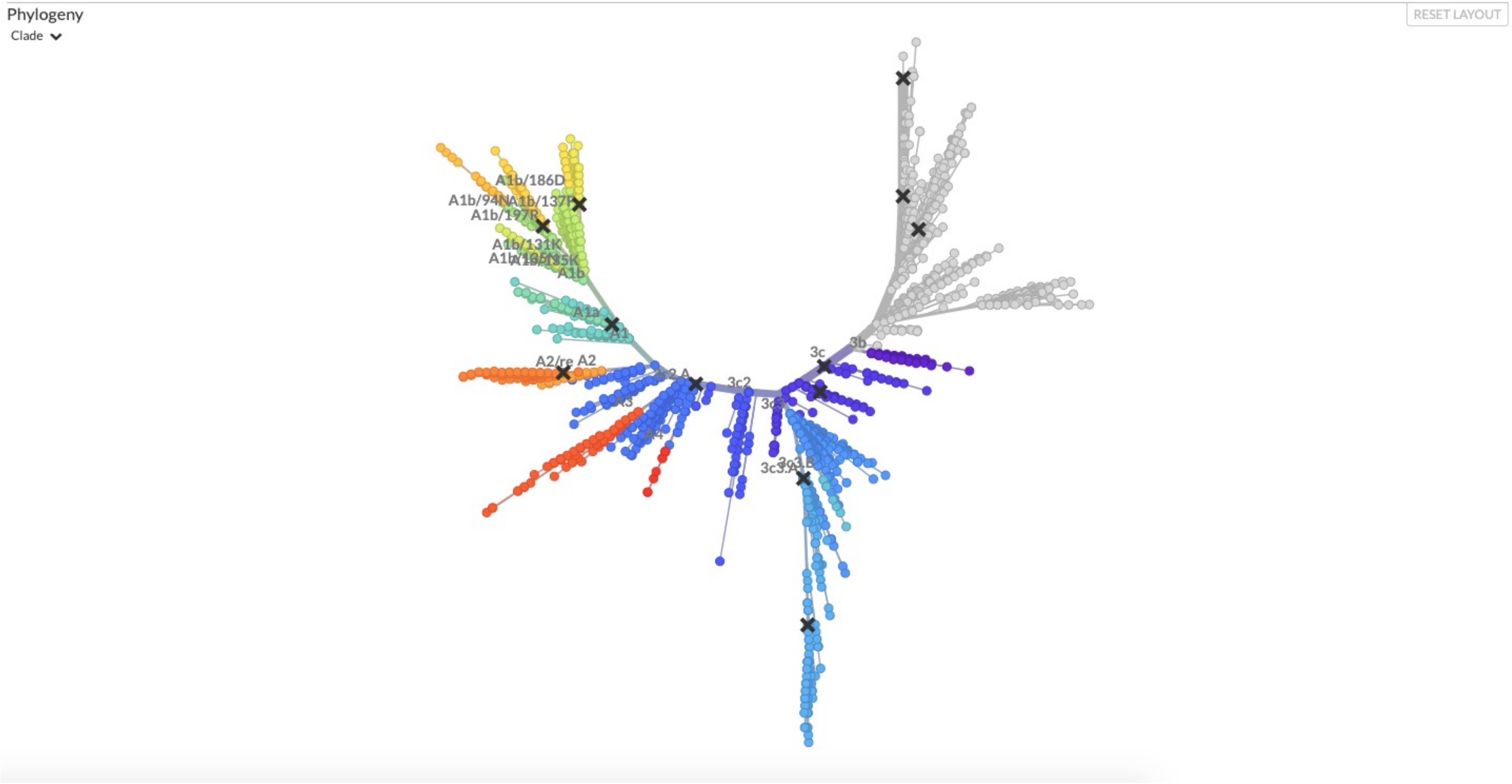
- Everything affects everything in Bayesian phylogenetics. Wrong evolutionary models lead to wrong trees, which leads to wrong population parameters.

- Overparameterization is better than under-parameterization. In doubt, use the more complex site model, such as a GTR+ model (Abadi et al., 2019, Nat. comm.)

- Never forget to account for rate heterogeneity (experience). Also, the JC69 model is hardly ever appropriate.

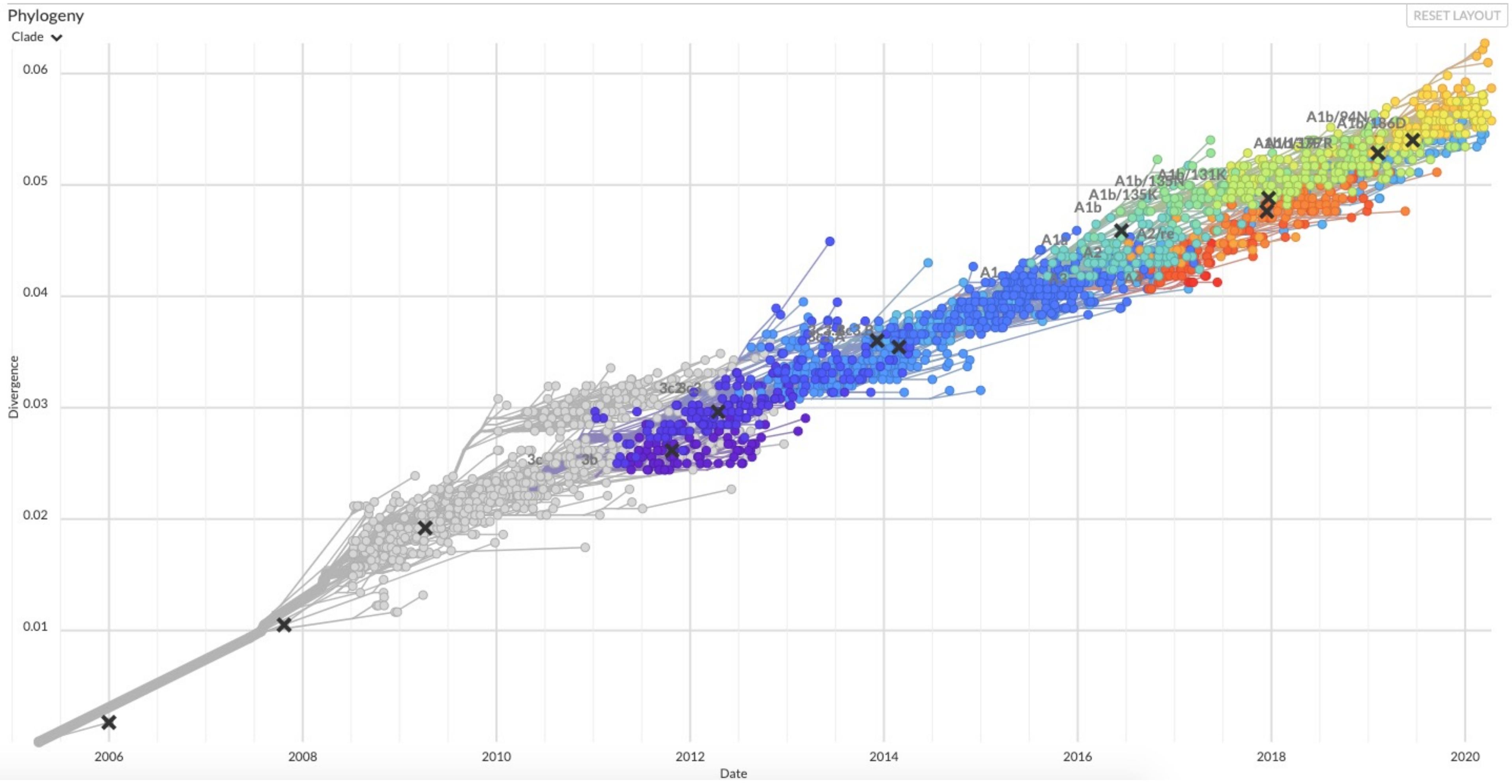
The molecular clock

Using site models, we can get from alignments to divergence trees



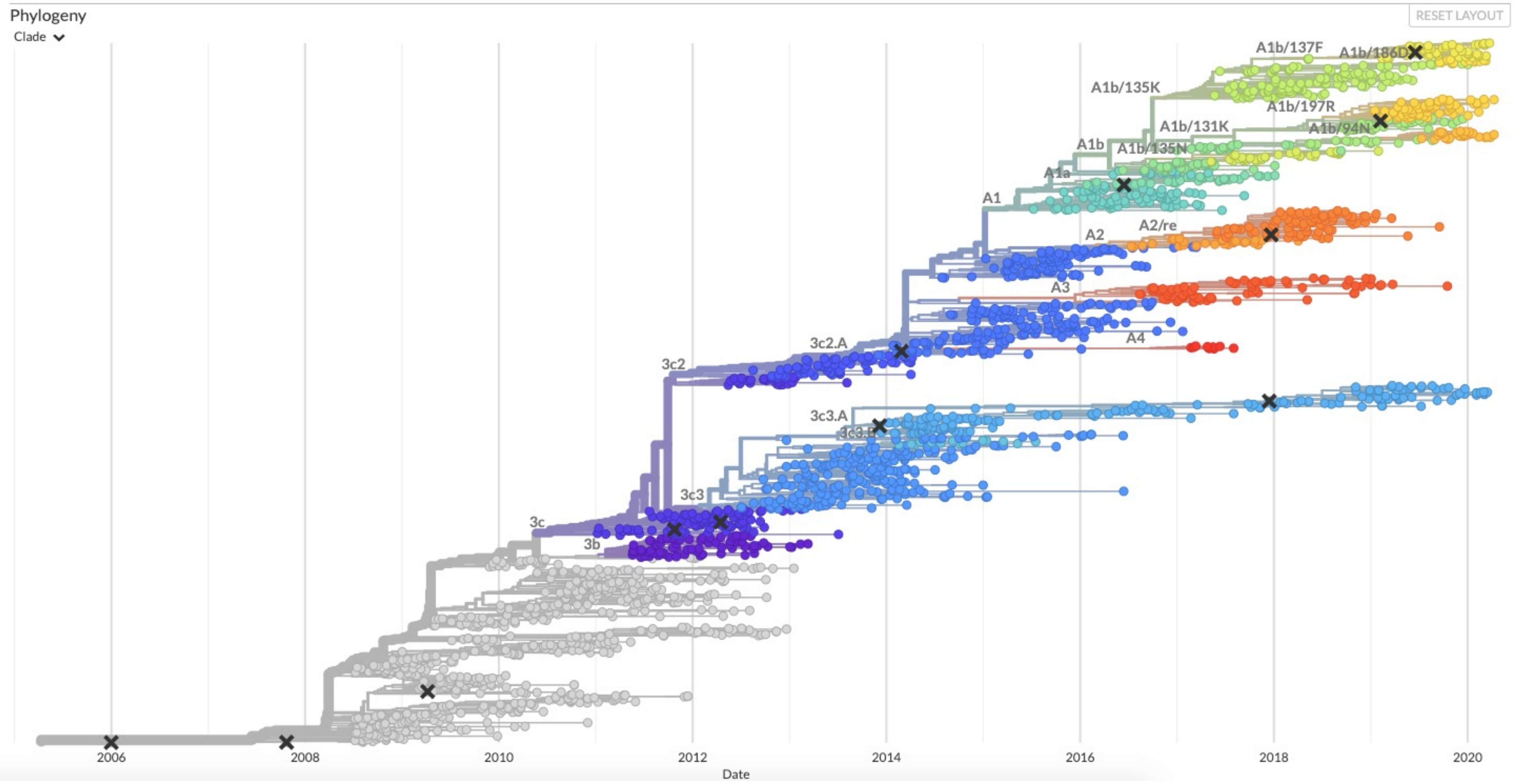
nextstrain.com

Sequences that are further apart in time are more diverged



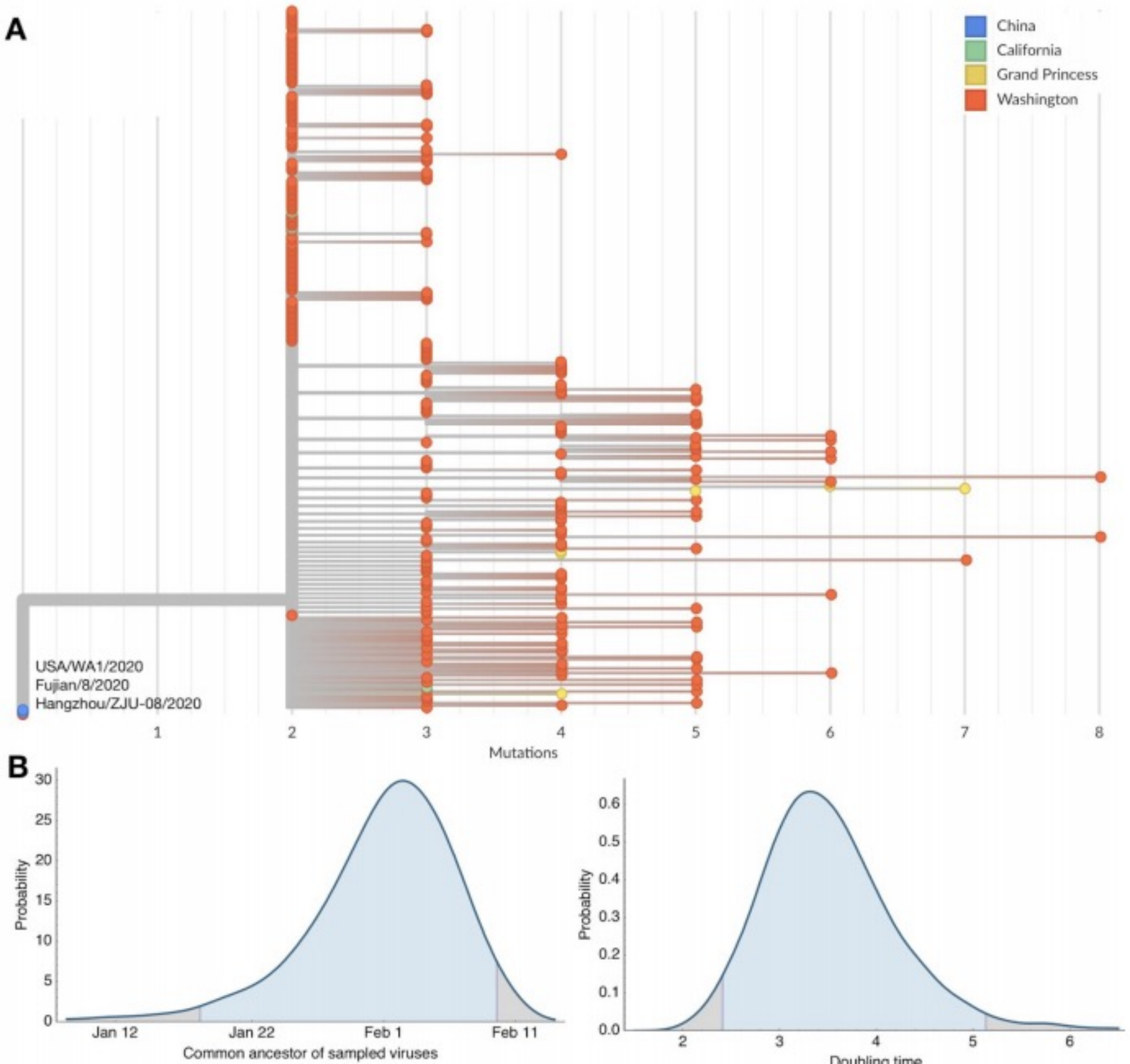
nextstrain.com

Molecular clocks bring us from divergence trees to time trees



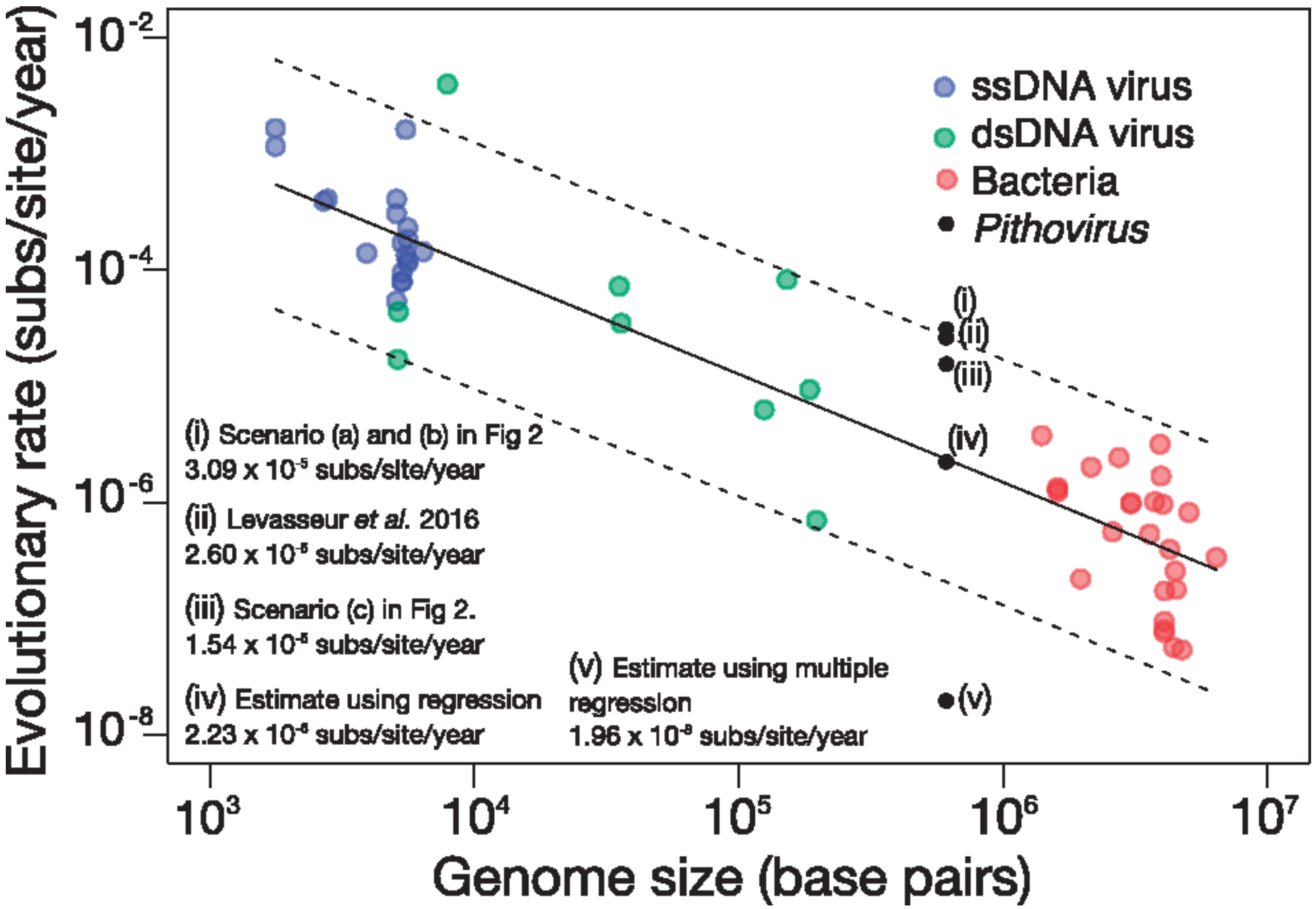
nextstrain.com

Evolutionary models can be used to estimate common ancestor times



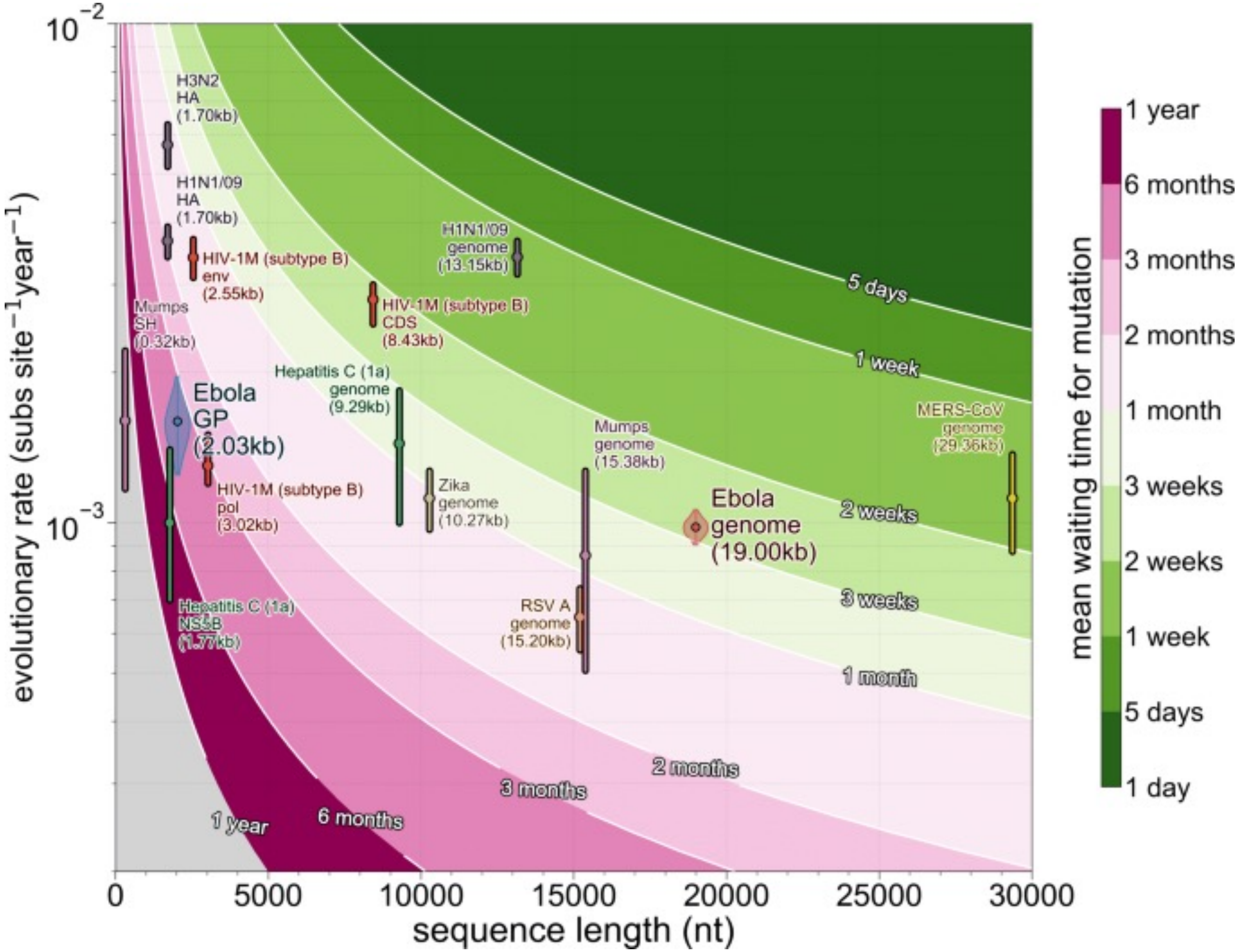
Bedford et al., 2020, Science

Different organisms have vastly different rates of evolution



Duchene et al., 2018, Vir. Evol.

The time resolution is dependent on the size of the genome and the rate of evolution

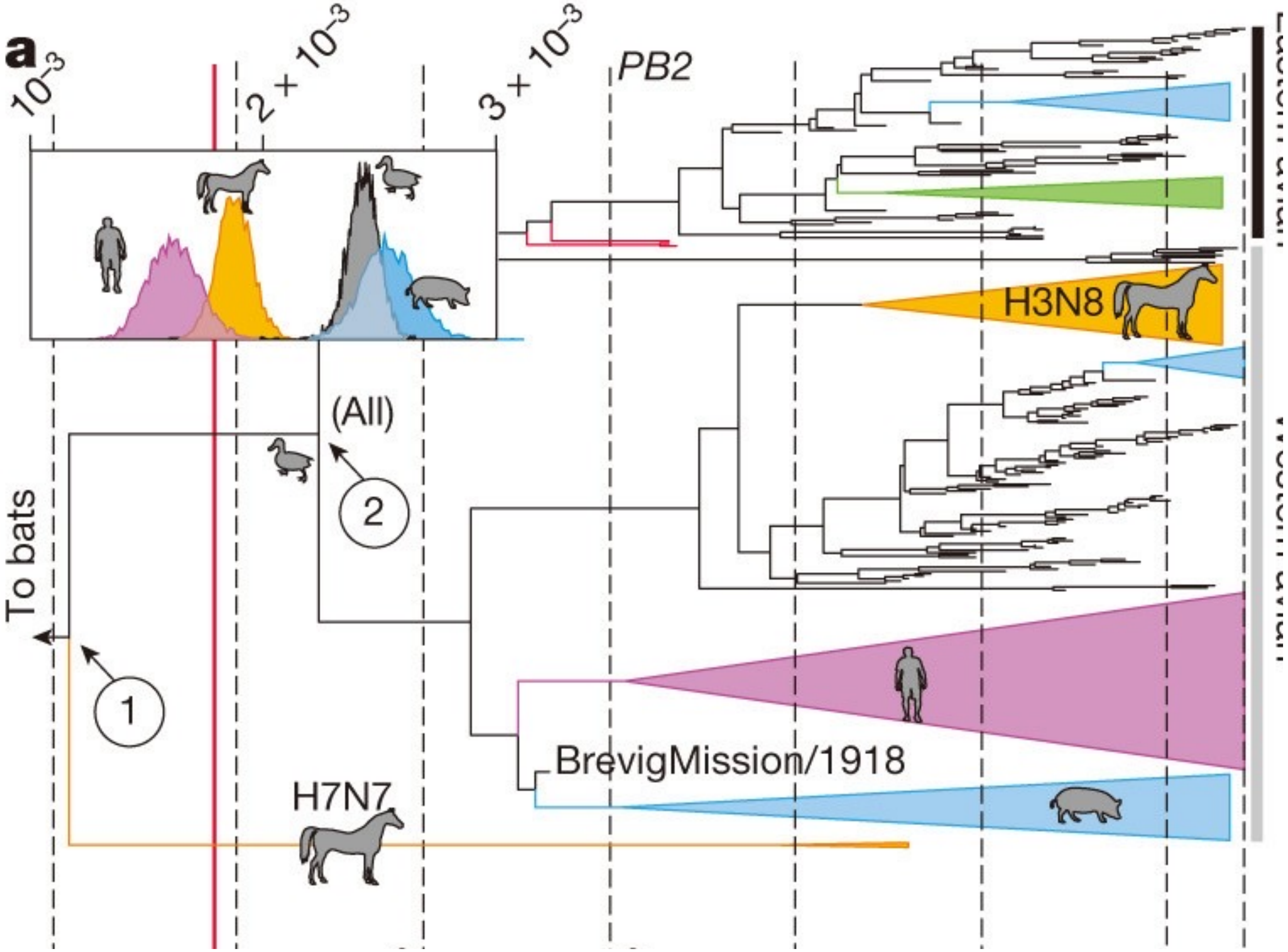


Dudas et al., 2019, *BMC Eco. Evo.*

What is my clock rate? (from Holmes et al. 2016

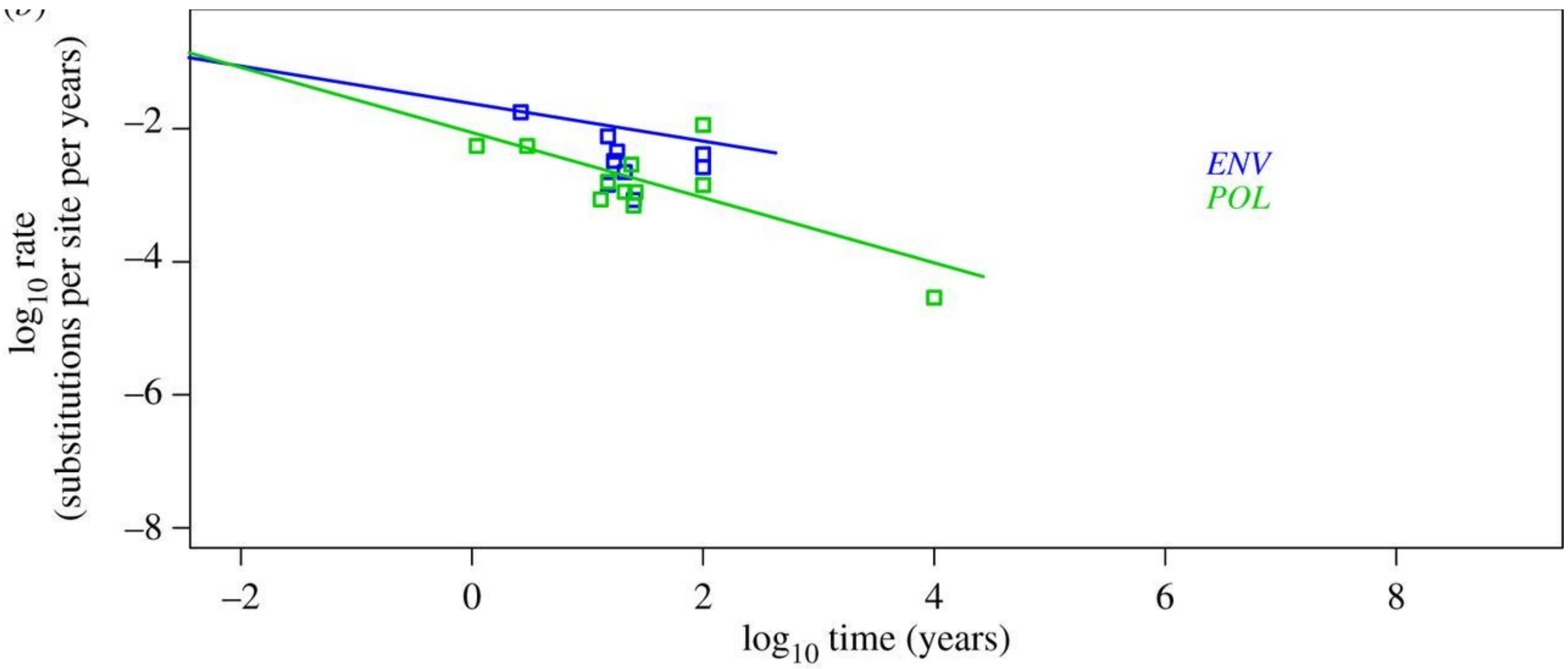
- **Mutation rate (short term, faster):** Error rate in replication.
- **Substitution rate (long term, slower):** Long term rate of evolution.
- **Evolutionary rate (=clock rate):** Measured rate of change. Result of mutation rate and population processes, such as selection. Typically sits between the mutation and substitution rate (The notation in BEAST is confusing with regards to what is what).

Rates of evolution can vary due to different hosts



Worobey et al. (2014), Nature

Rates of evolution can vary in the short term and long term



Duchene et al. (2014), Proceedings B

Mathematically clock models are functions that have divergence as input and time as output

- **Strict clock models** assume that evolution happens equally fast on each branch of a tree. Strict clocks are mostly used to study pathogens over rather short times (a few years)

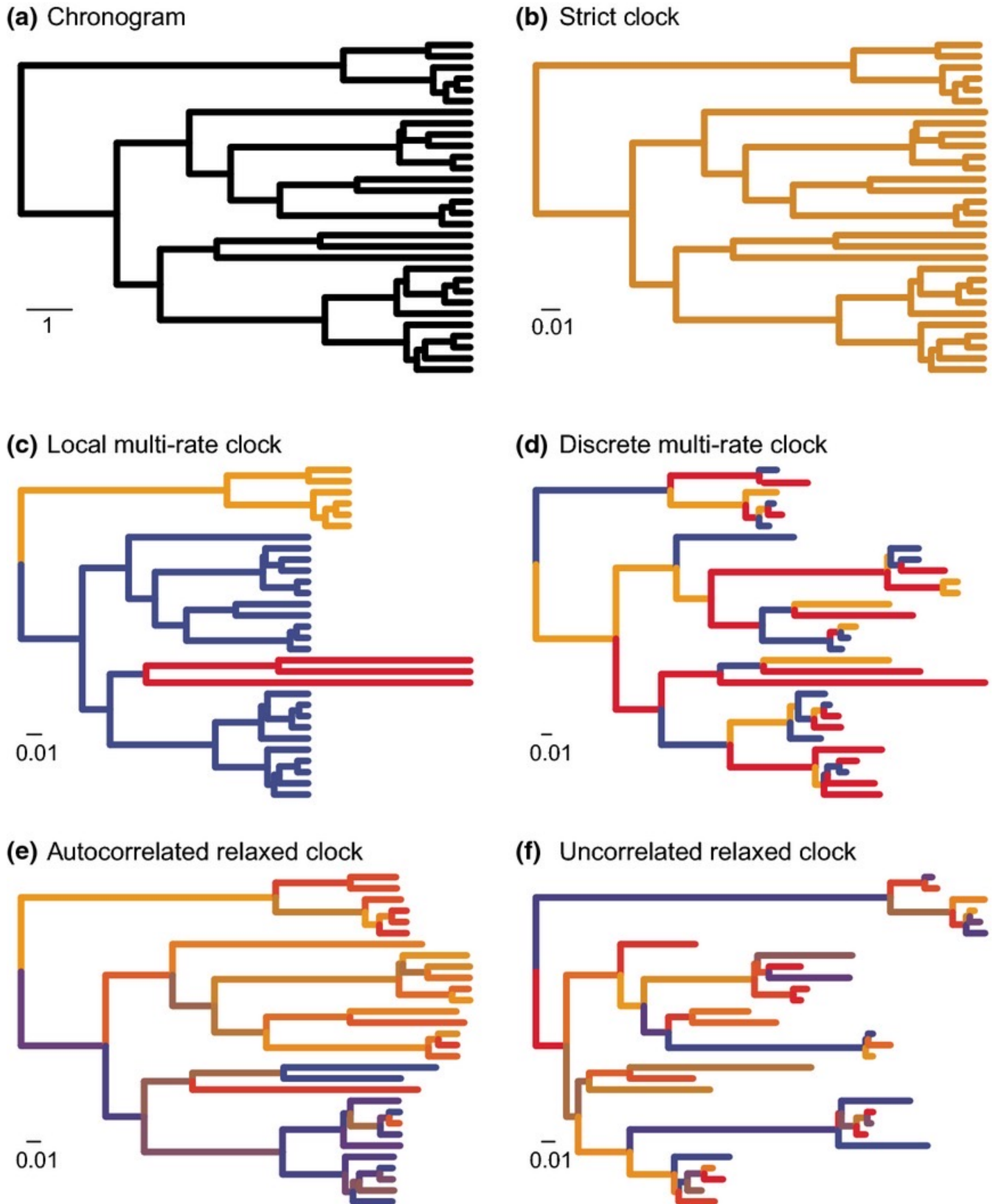
- **Random clock models** allow different branches of a phylogenetic tree to have different rates (speed) of evolution. These are more prevalent when analyzing datasets that were sampled over longer time periods

Random clock models can be separated into two classes

- **Local random clock models.** Clock rates change on a few branches in the tree only. Once they change, all the descendent branches will inherit the same rate

- **Uncorrelated clock models.** Clock rates can vary on each branch (completely uncorrelated). Each branch has a clock rate that is considered a random draw from some distribution (typically Exponential or lognormal).

Molecular clock models map genetic divergence into time



(a) Chronogram

(b) Strict clock

(c) Local multi-rate clock

(d) Discrete multi-rate clock

(e) Autocorrelated relaxed clock

(f) Uncorrelated relaxed clock

- Clock models can account for variation in the speed of evolution across an evolutionary history.

Ho et al. (2014), Molecular Ecology

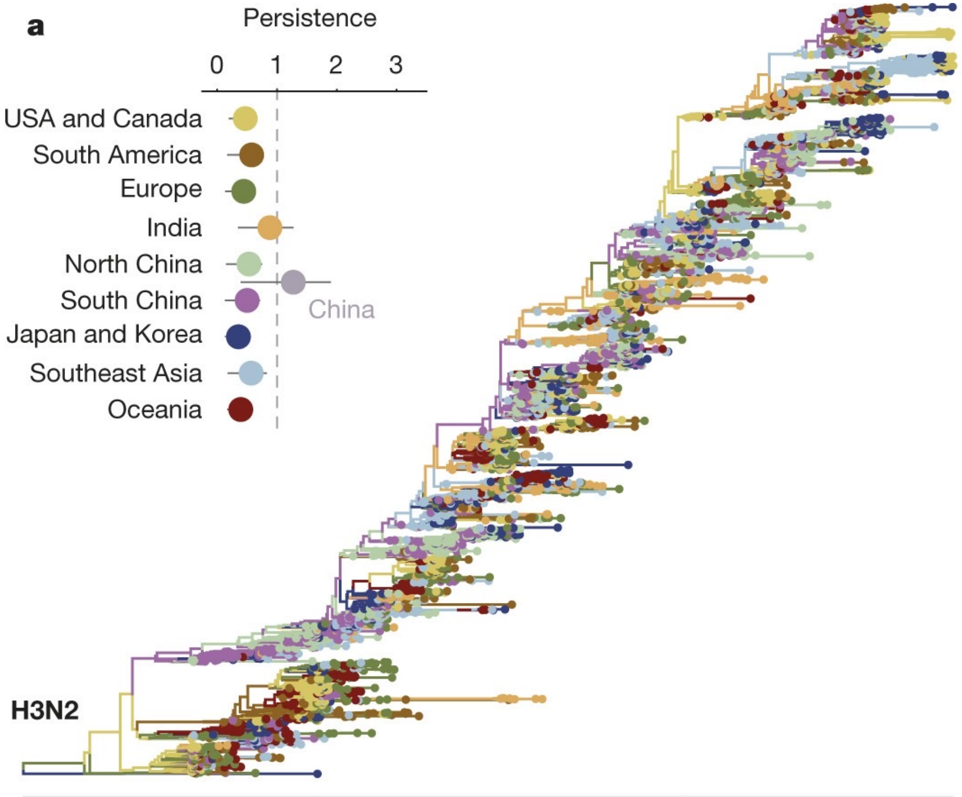
Where does the time information come from to inform the molecular clock?

- Calibrations -> Knowing something about when two lineages shared a common ancestor.

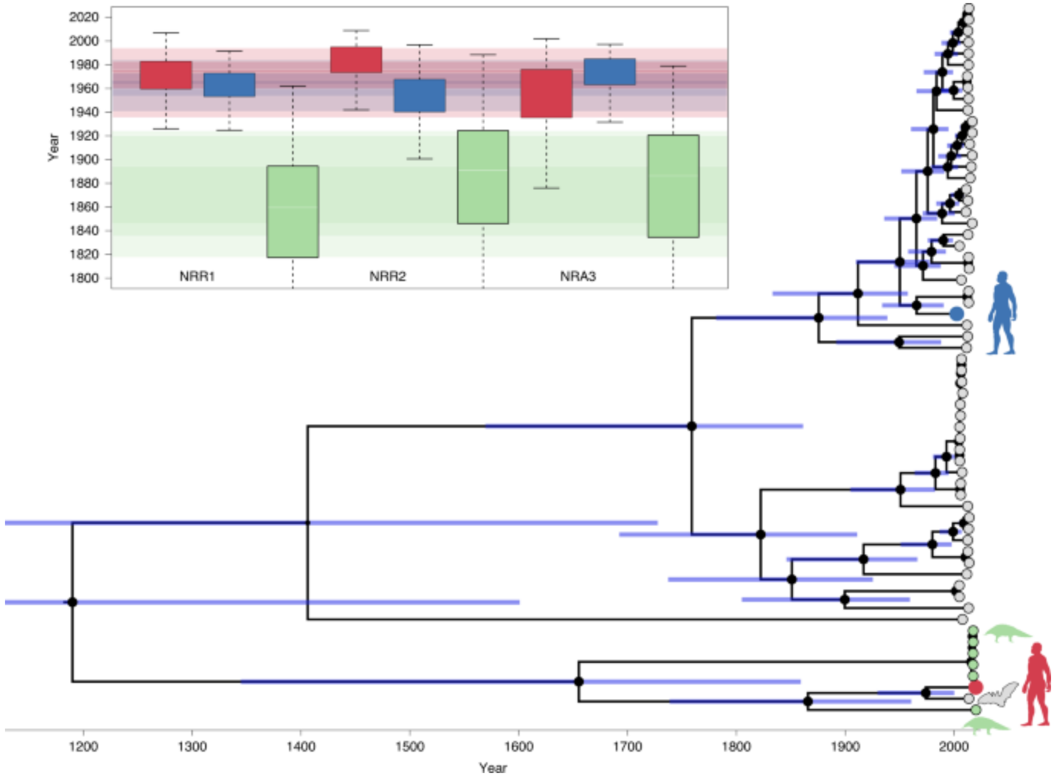
- Sampling through time of “measurably evolving pathogens”.

- Just “knowing” the rate, that is from the prior.

The time information should be in the same order of magnitude as the evolutionary history

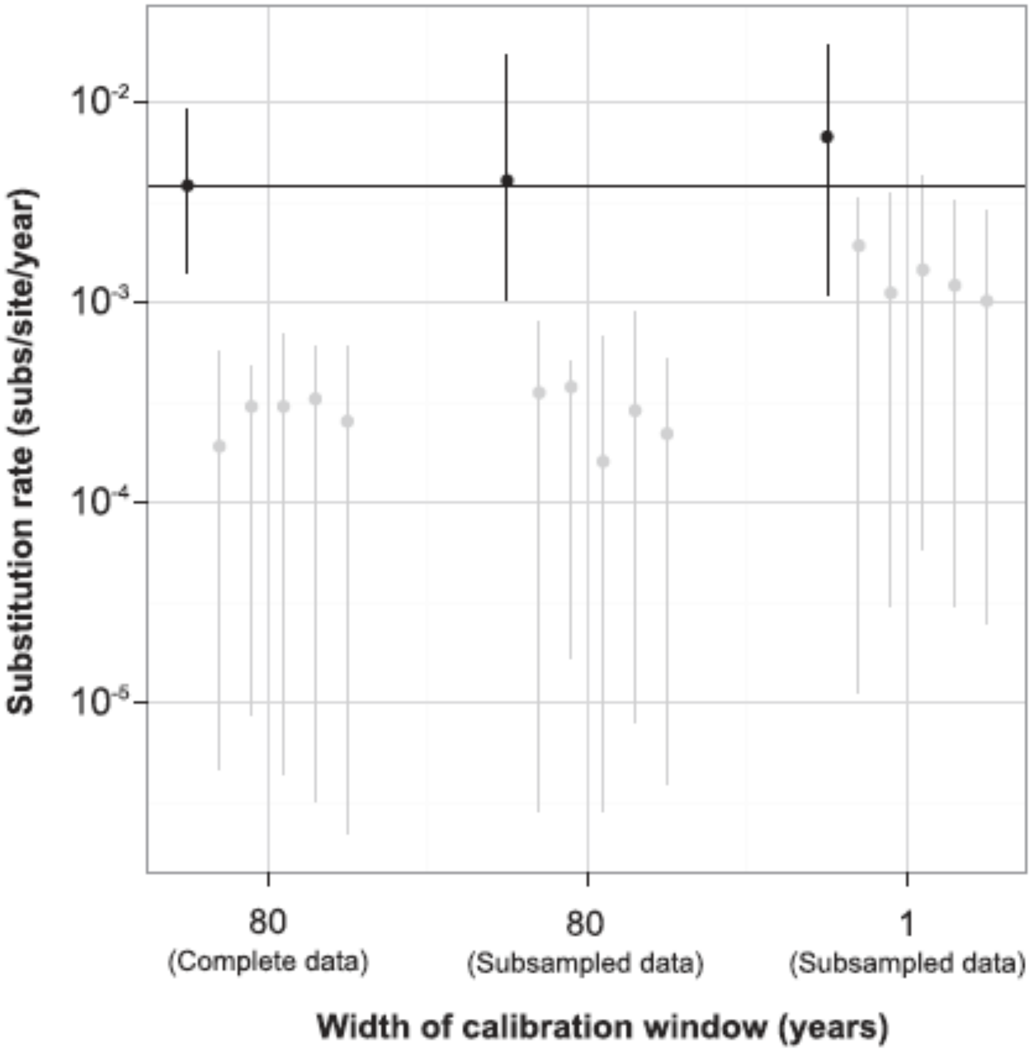


Bedford et al. (2015), Nature



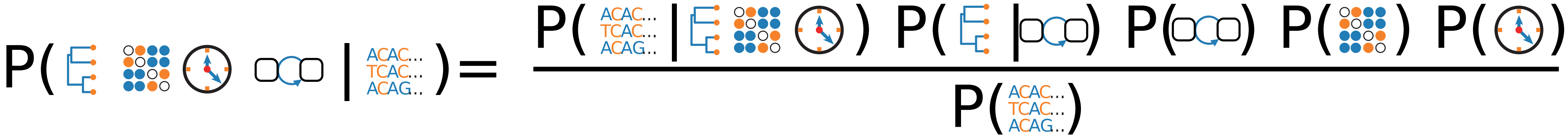
Boni et al. (2020), Nat. Mic.

Is there enough signal to estimate evolutionary rates? Tip Randomization

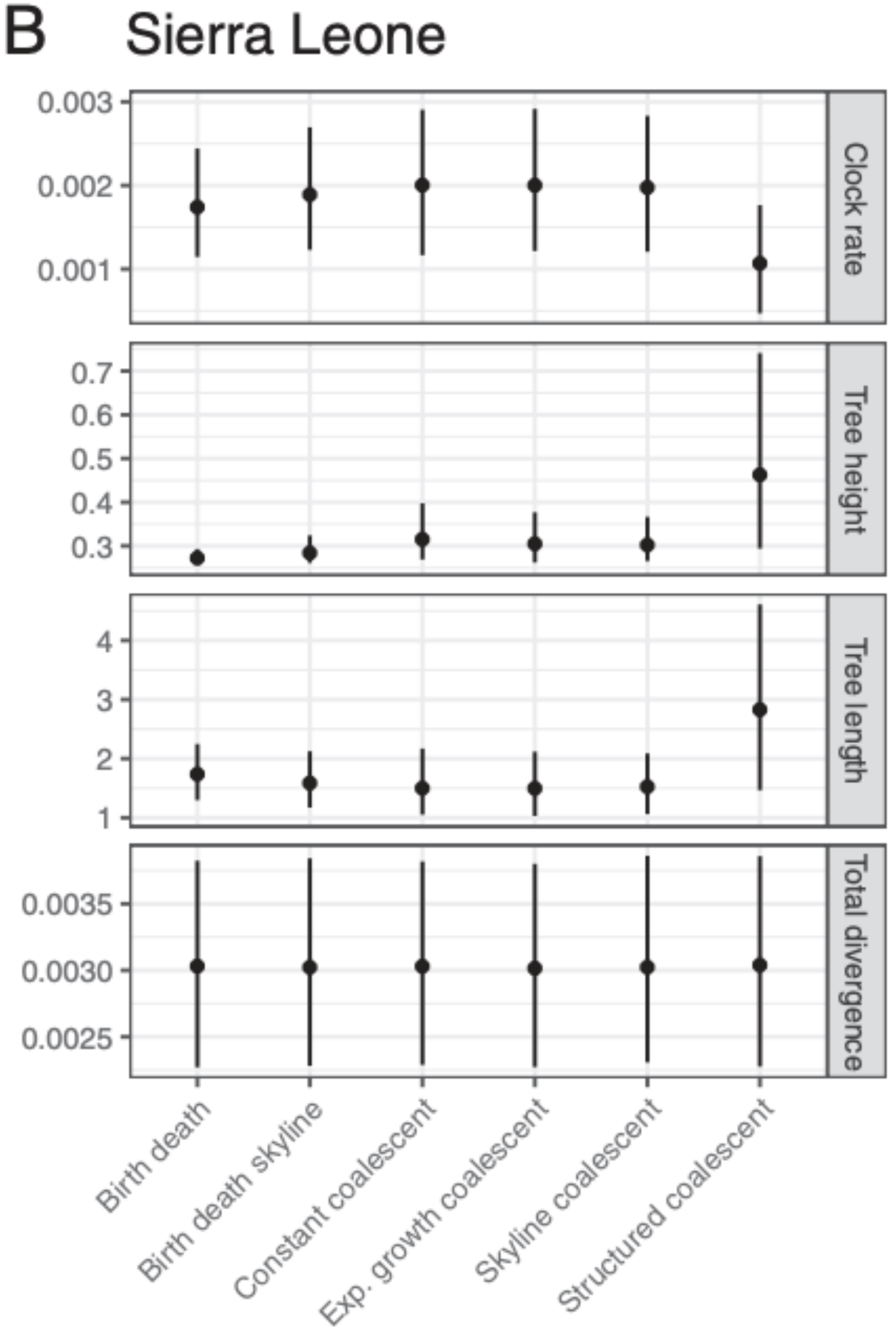


Duchene et al., 2015a, *MBE*

In the posterior probability, modelling the evolution of sequences feeds into the tree likelihood.


P(| ACAC... TCAC... ACAG..)=
P(ACAC... TCAC... ACAG.. |) P(|) P( ) P( ) P( )
P(ACAC... TCAC... ACAG..)

DuPlessis, L. et al. (2015),
Trends in Microbiology



Model choices not directly related to clock models can impact rate estimates

Möller et al., 2018, *PNAS*

Some reading material

- Accounting for codon positions: https://doi.org/10.1093/molbev/msj021
- Overfitting site models is ok: https://www.nature.com/articles/s41467-019-08822-w
- Posterior predictive simulations to evaluate clock signal: https://academic.oup.com/mbe/article/32/11/2986/981260
- Time randomization to evaluate clock signal: https://academic.oup.com/mbe/article/32/7/1895/1016979
- Rates of evolution in EBOV: https://www.nature.com/articles/nature19790