



BILL &
MELINDA
GATES
foundation

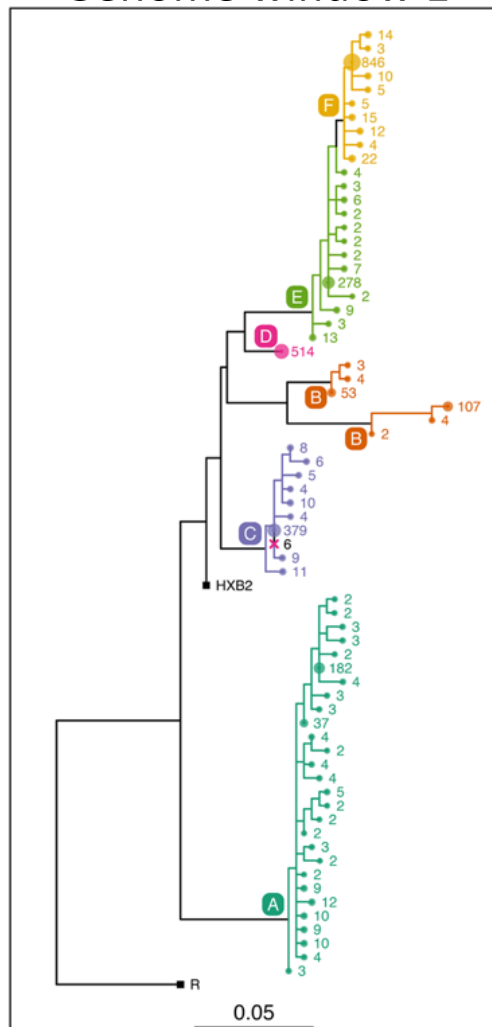


How to interpret a viral phylogeny

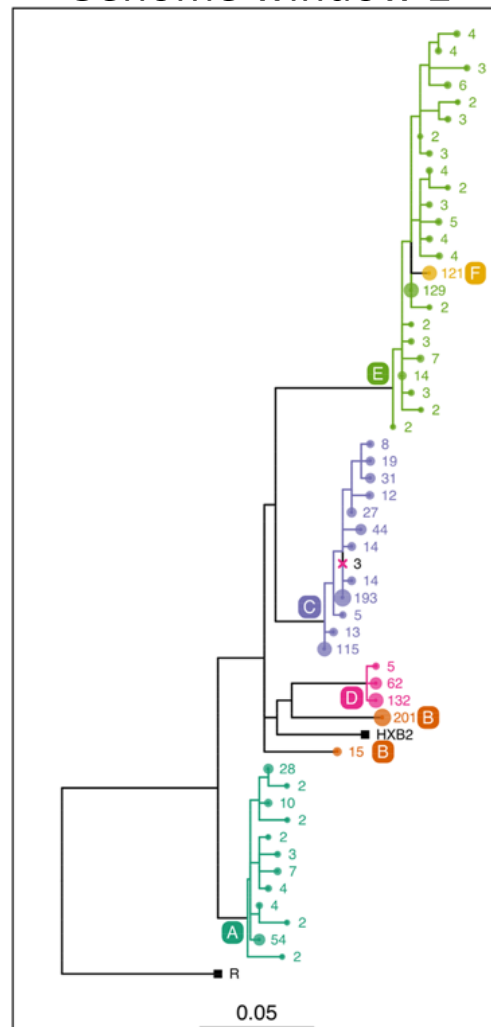
a PANGEA webinar

Christophe Fraser

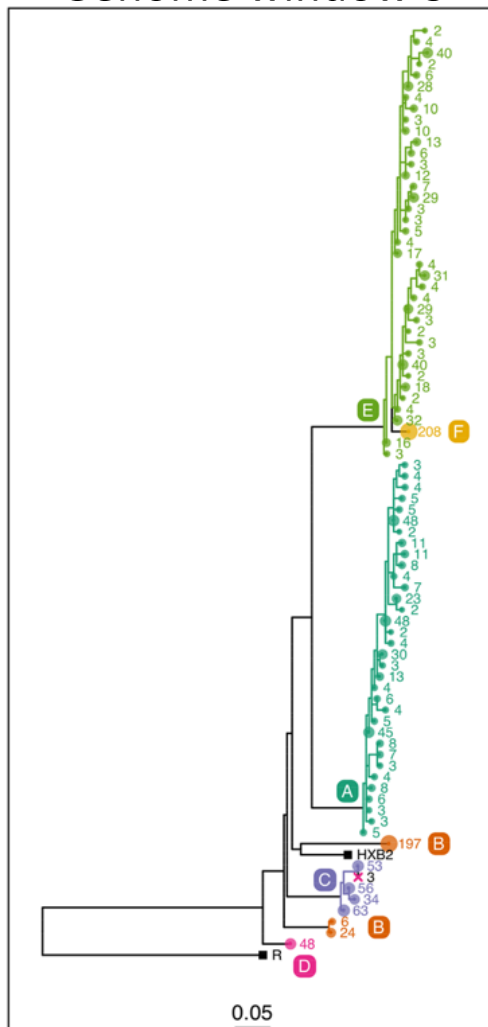
Genome window 1



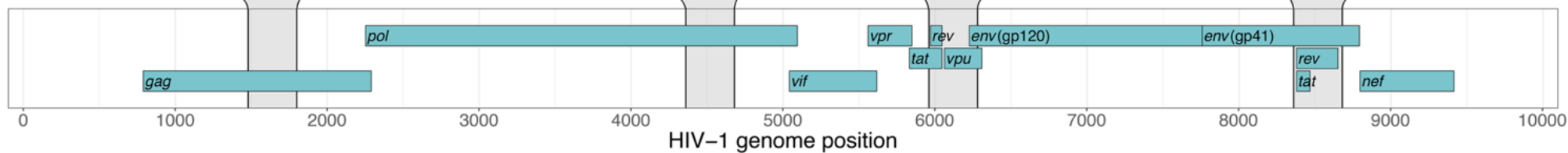
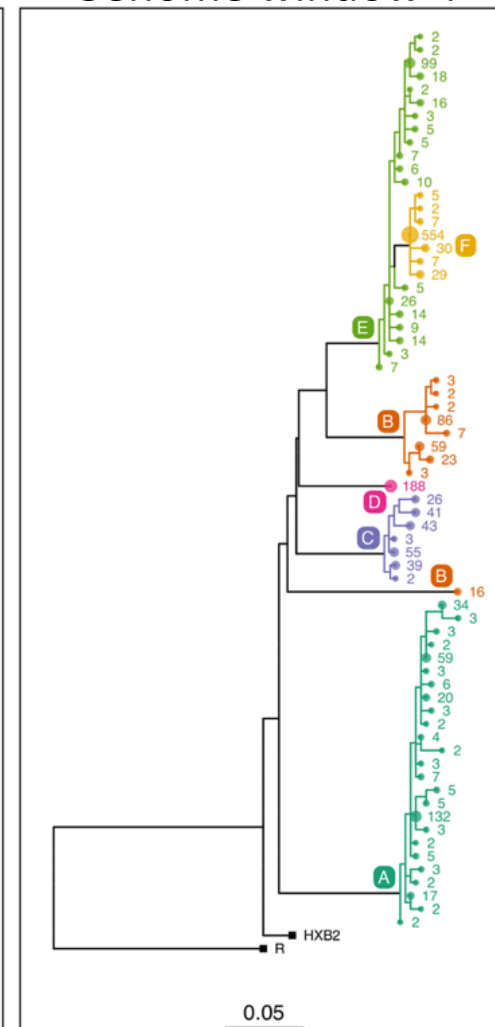
Genome window 2



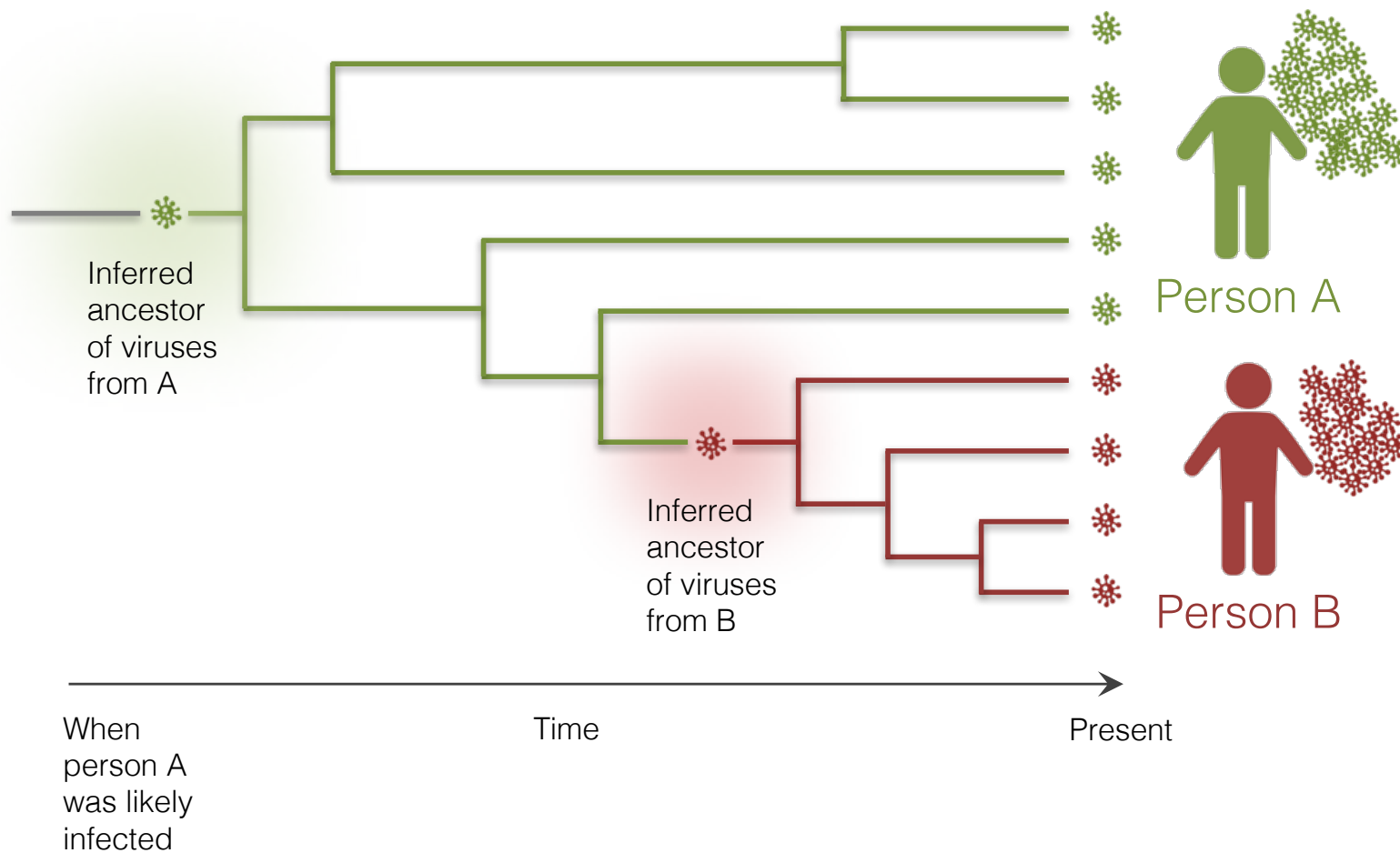
Genome window 3



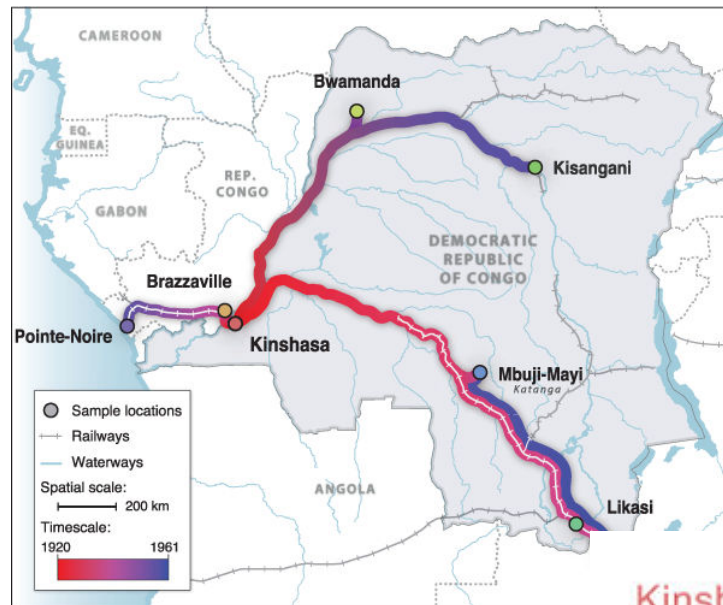
Genome window 4



Using within-host diversity to infer direction of transmission:
Equivalent to inferring ancestral state of virus populations.

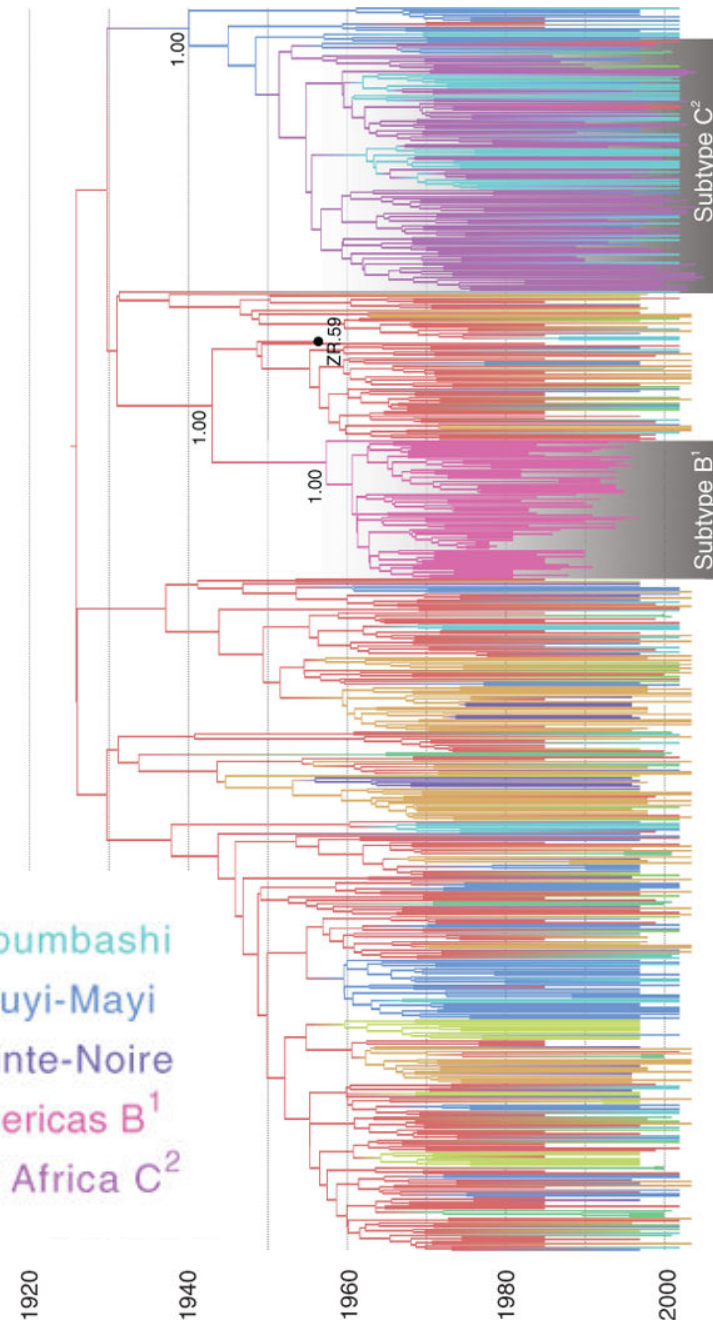


Inferring ancestries in early HIV diversification and spread:



Kinshasa
Brazzaville
Bwamanda
Kisangani
Likasi

Lubumbashi
Mbuyi-Mayi
Pointe-Noire
Americas B¹
SE Africa C²

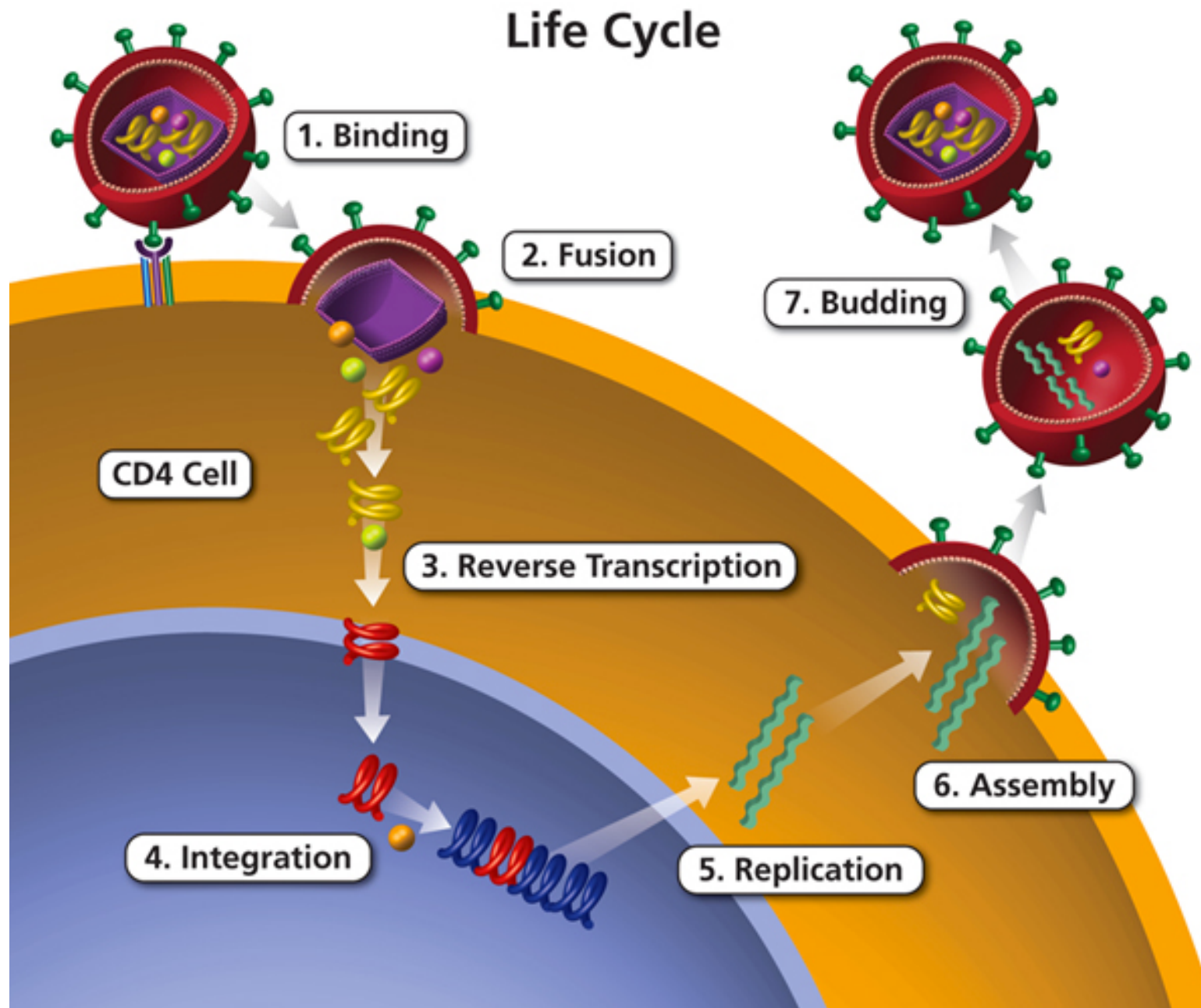


Faria et al, Nature 2014

Why do we represent relatedness of
viruses as trees?

Why not 'networks', 'clusters', 'maps'?

A: because HIV is a biological replicator.

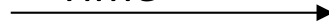


Day 0

A single virus




Time





Time



A black arrow points to the right, indicating the progression of time.

Day 2

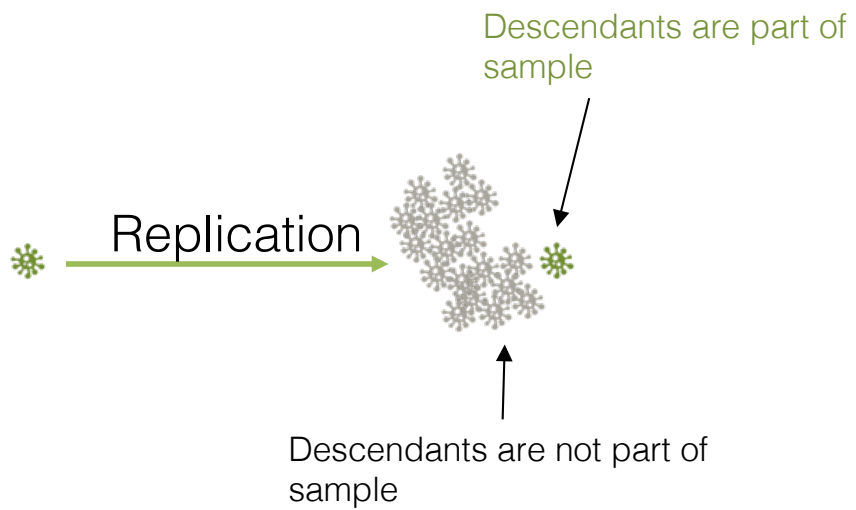
On average, each
replication cycle takes ~2
days, and the virus acquires
>1 mutation



Time →

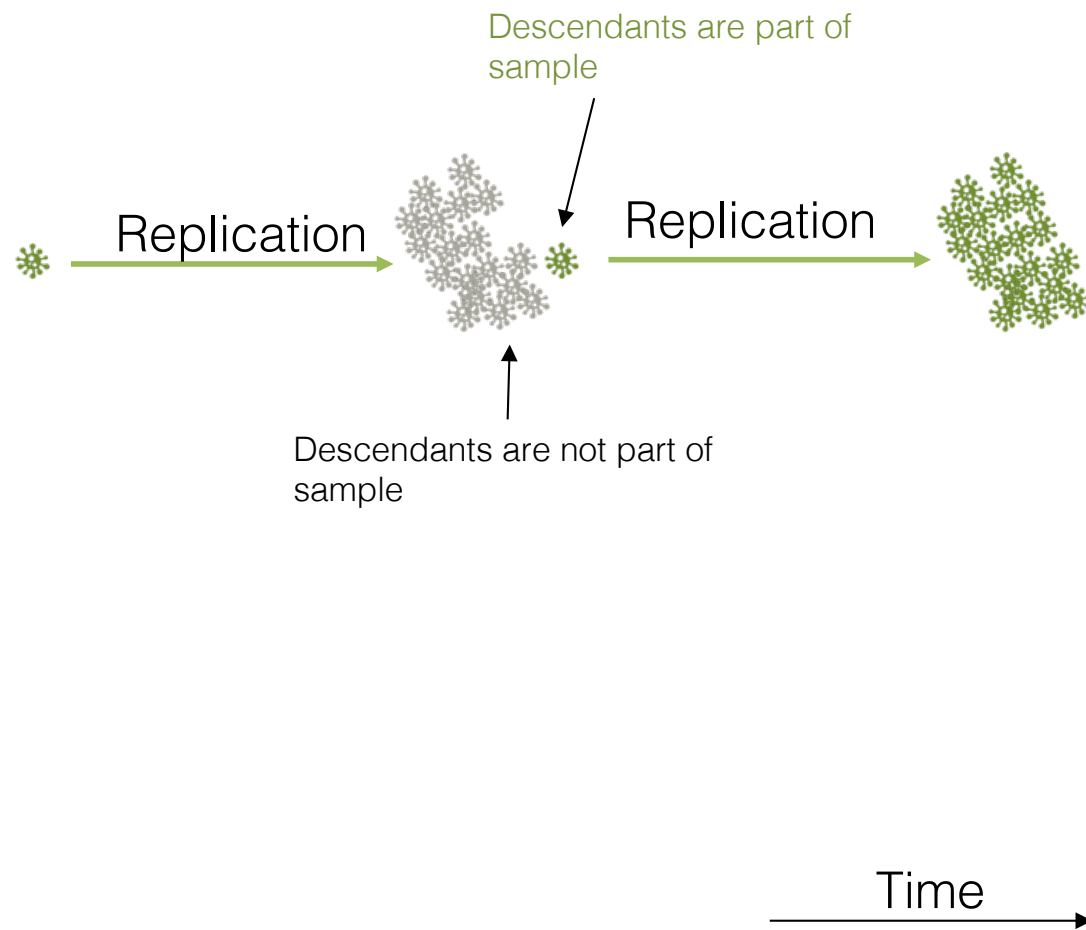
In tracing an ancestral lineage, we only keep track of viruses, whose descendants survive to become part of our sample of interest.



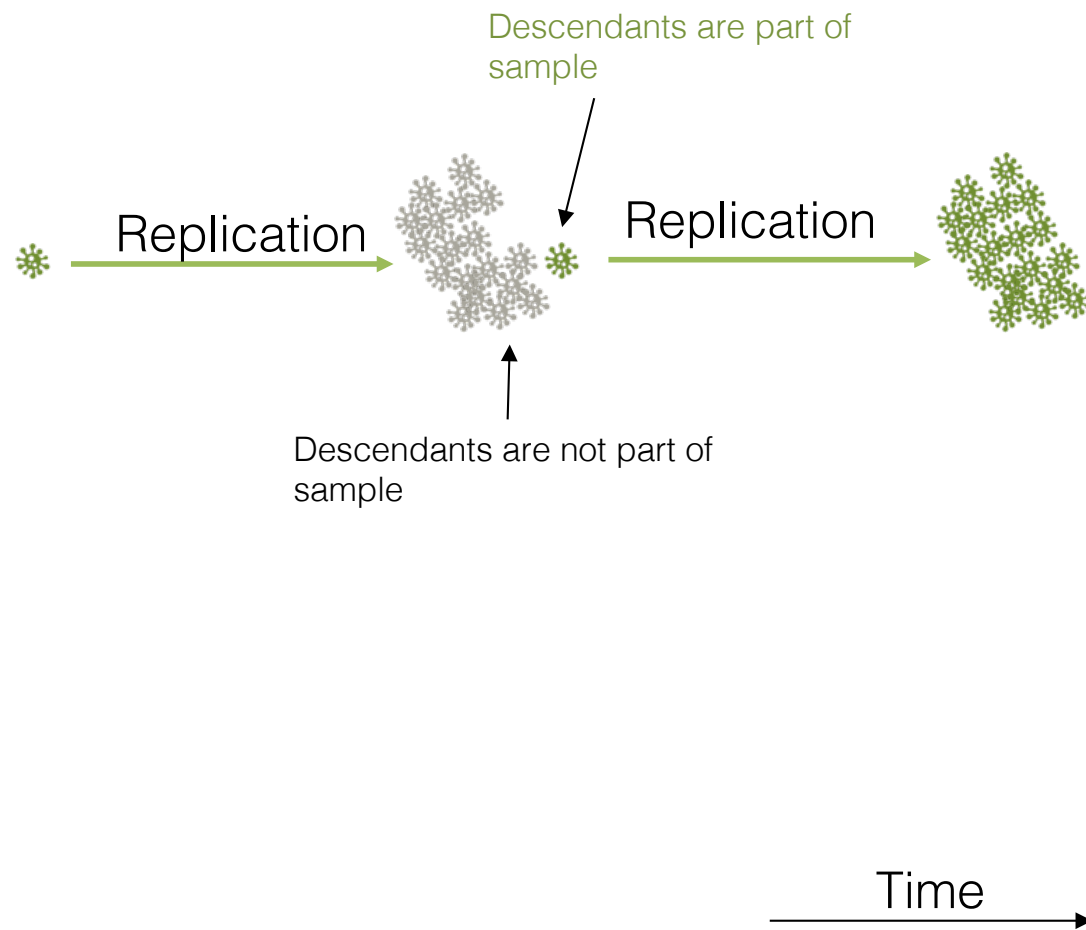


Time →

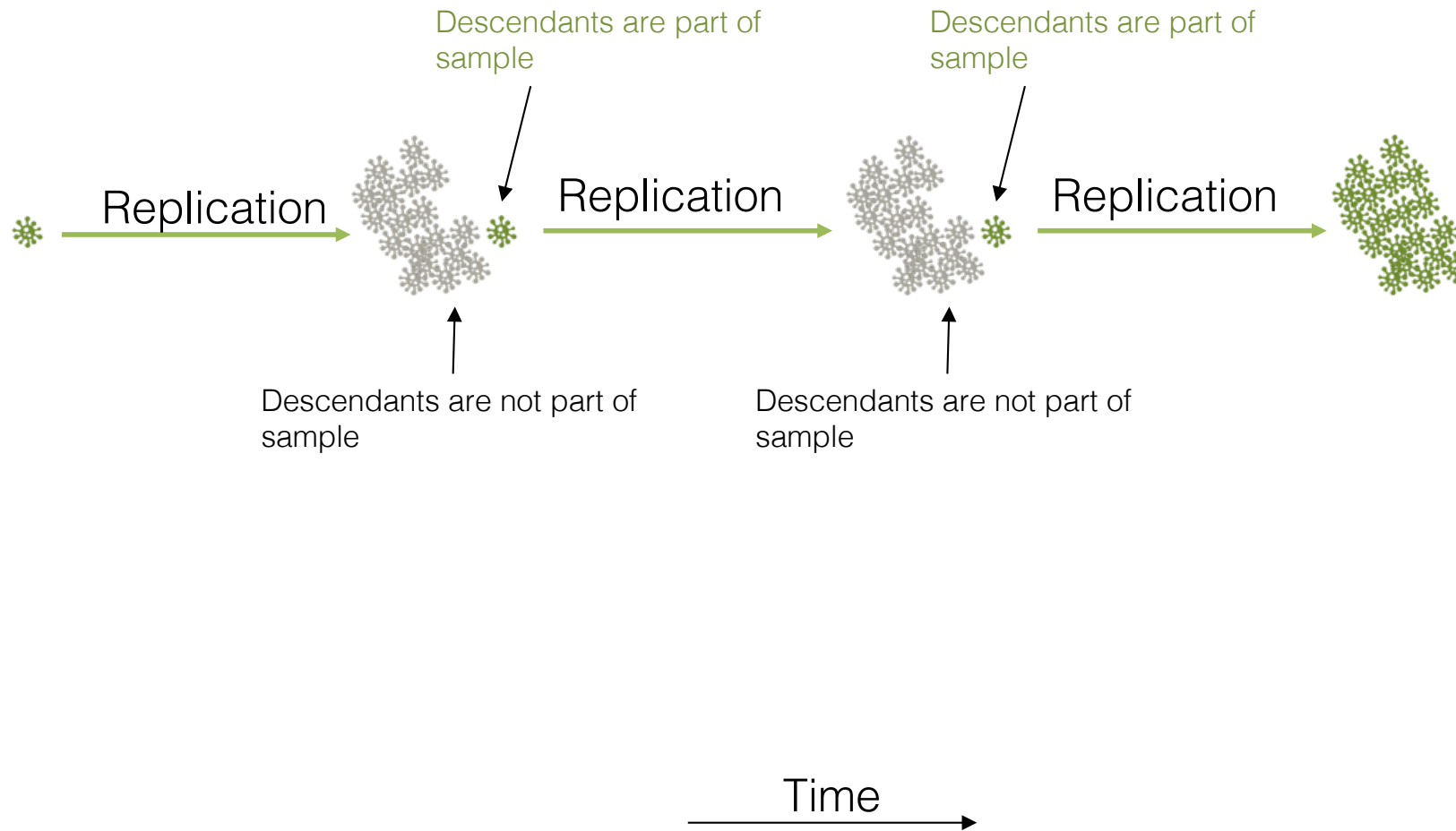
Day 4



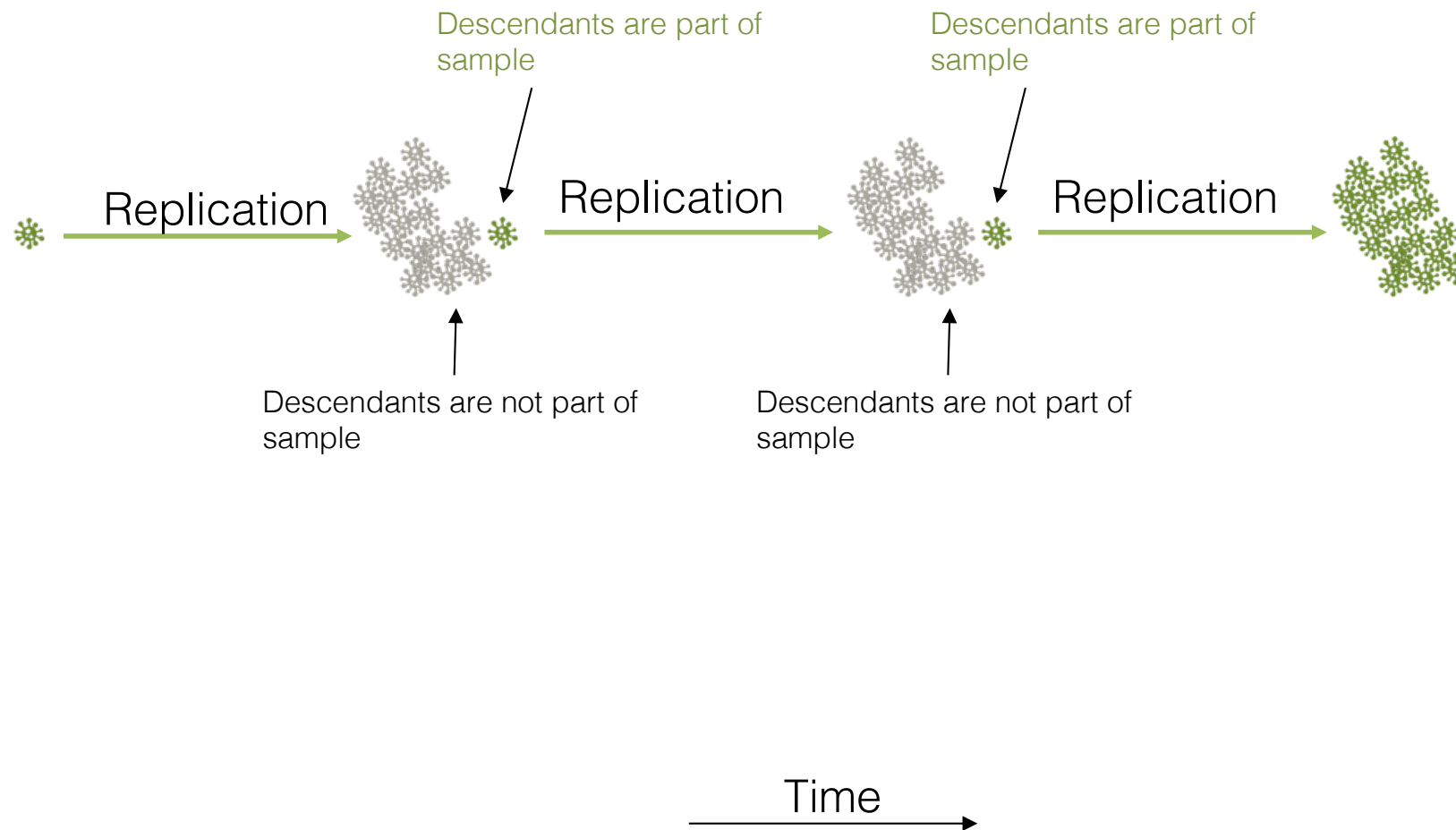
Day 4



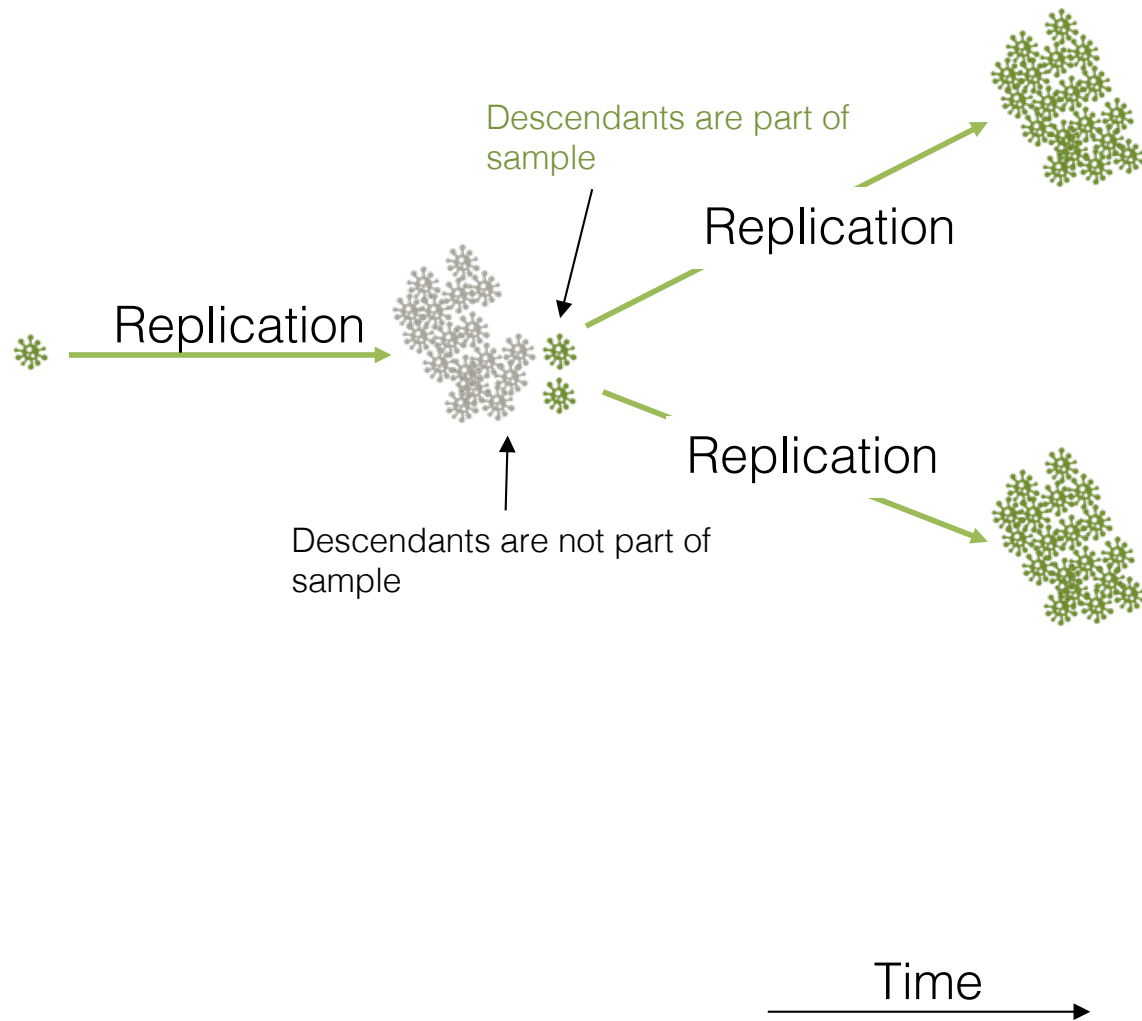
Day 6



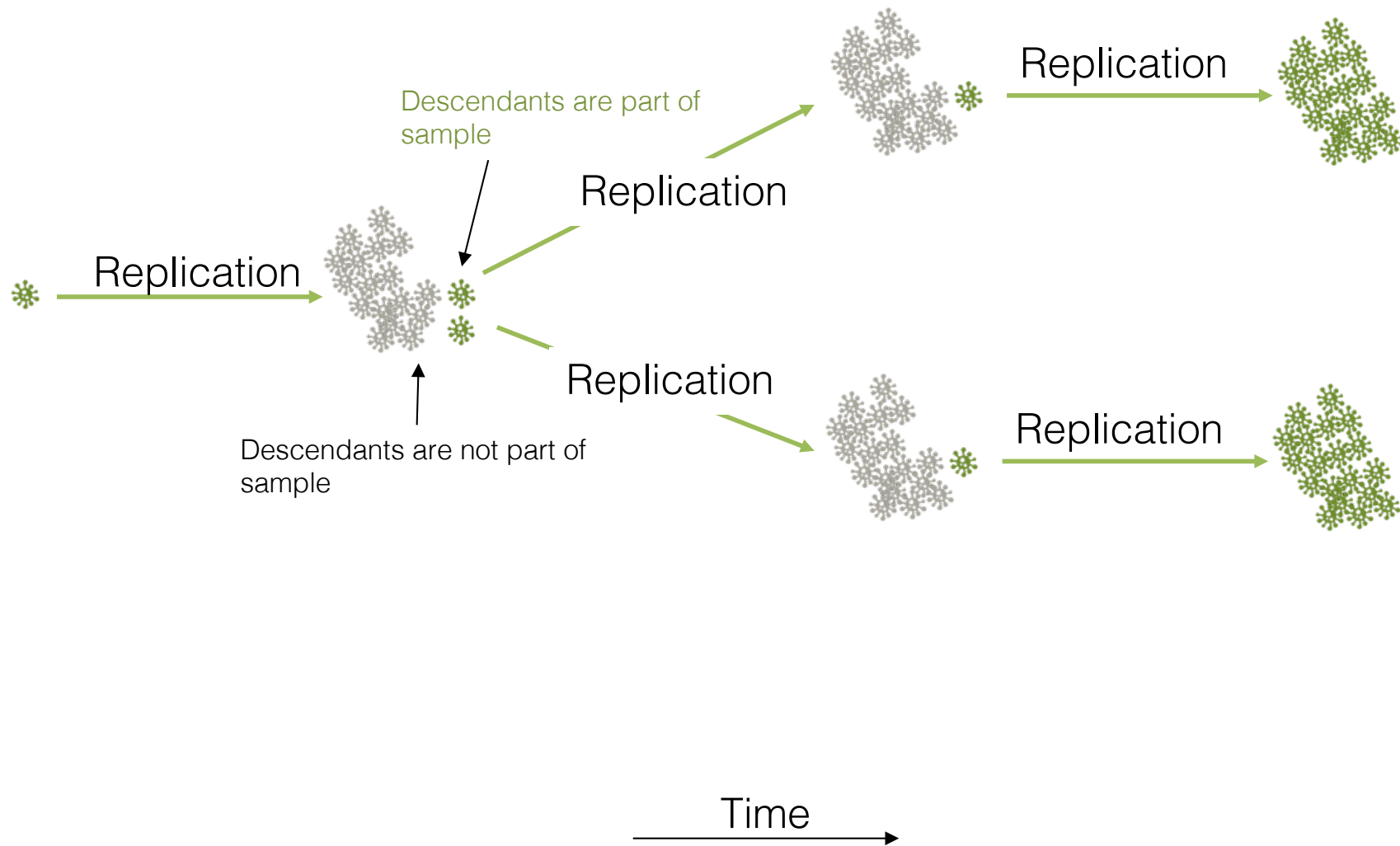
What happens if more than one virus has descendants in a sample?



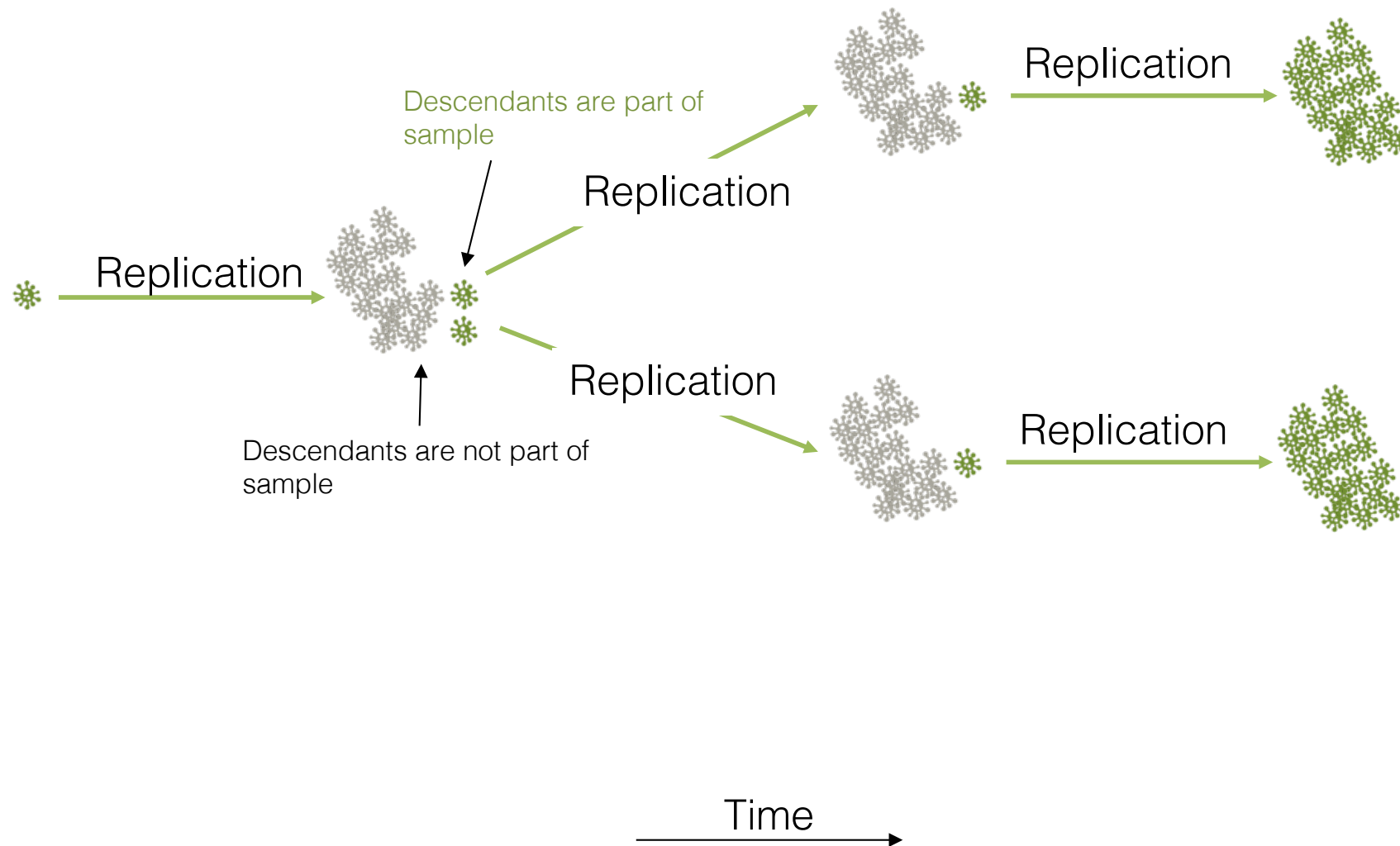
Day 4



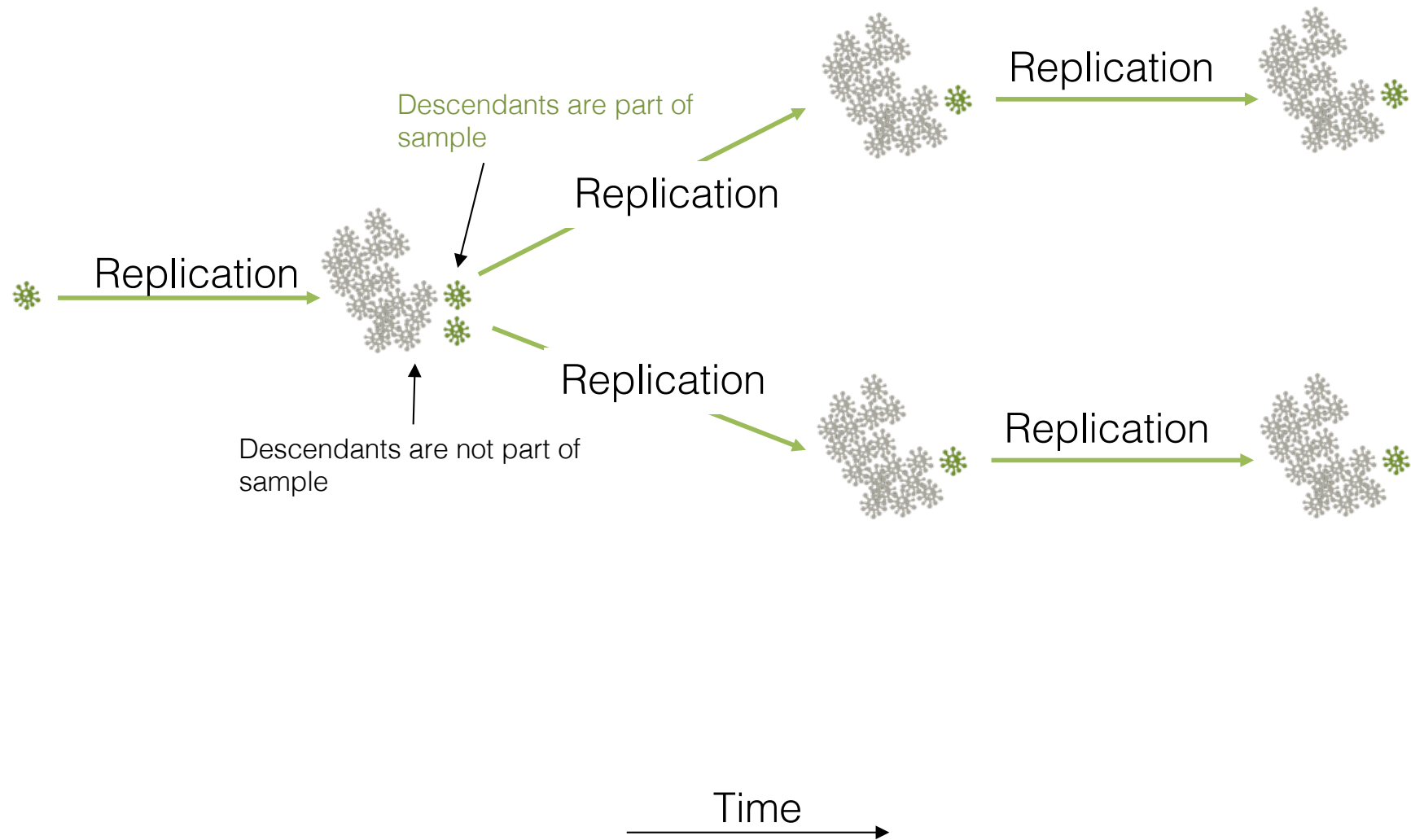
Day 6



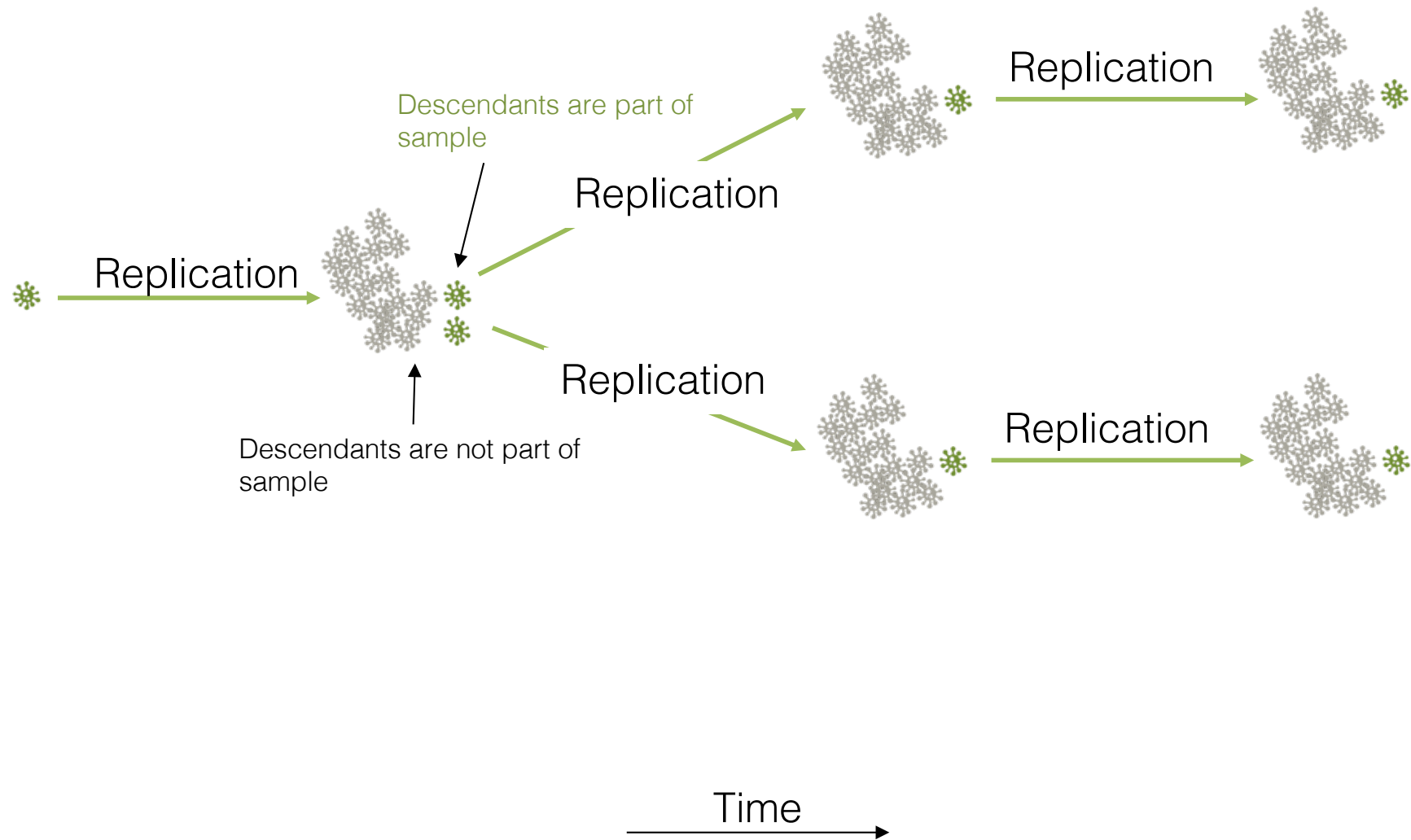
Tracing an ancestral lineage consists of keeping track of which viruses leave behind descendants in the sample



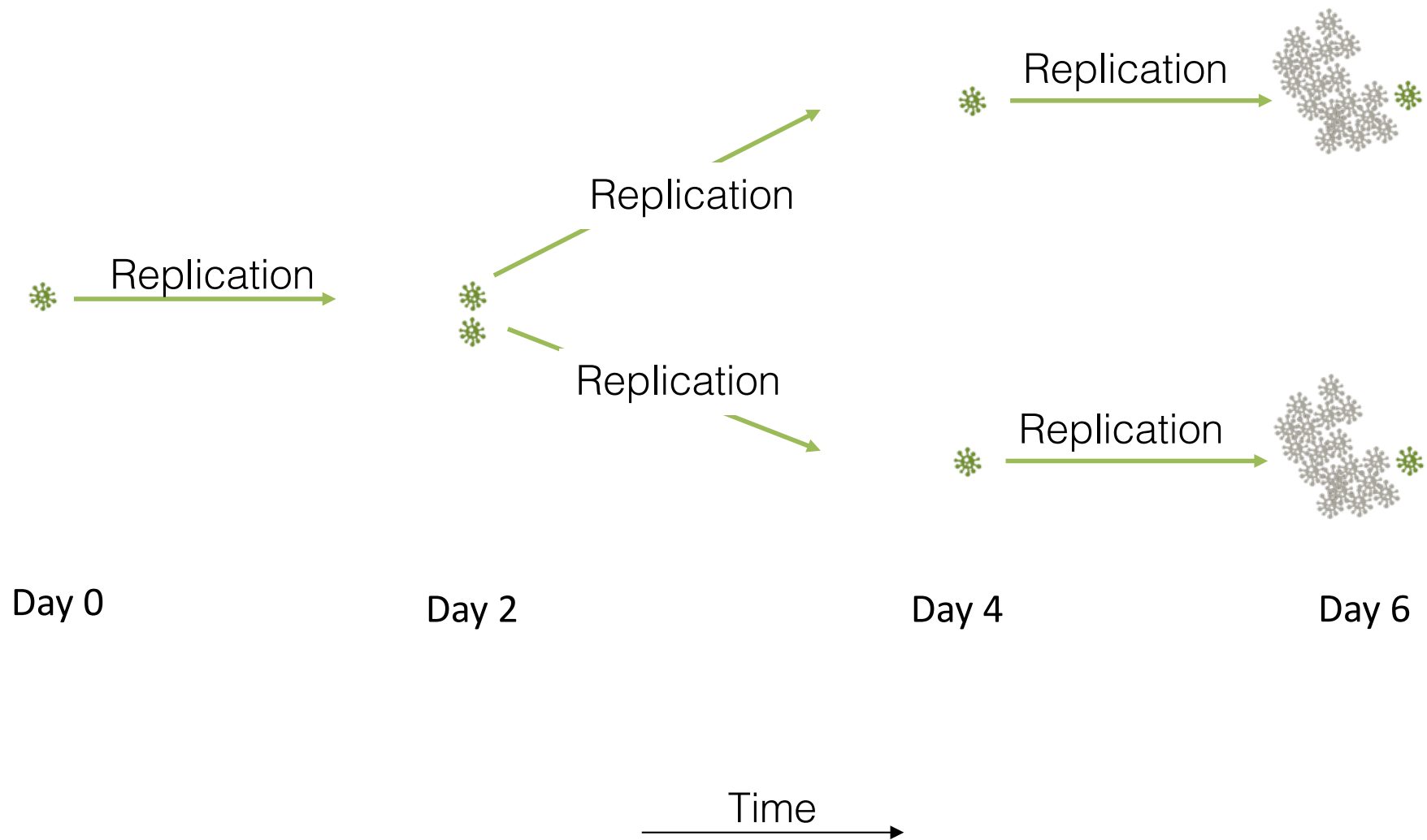
Tracing an ancestral lineage consists of keeping track of which viruses leave behind descendants in the sample.



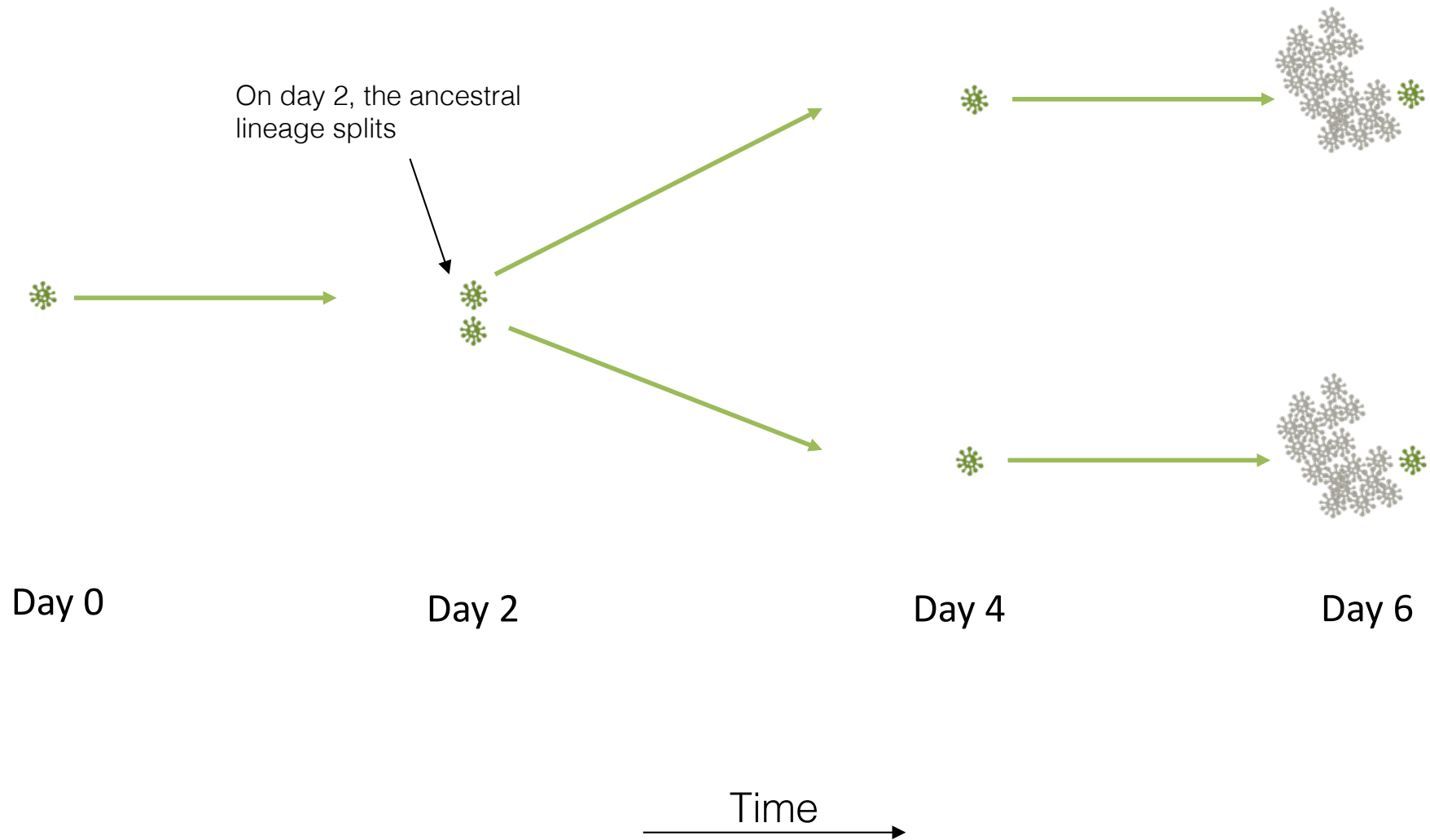
We ignore viruses that don't leave behind descendants:



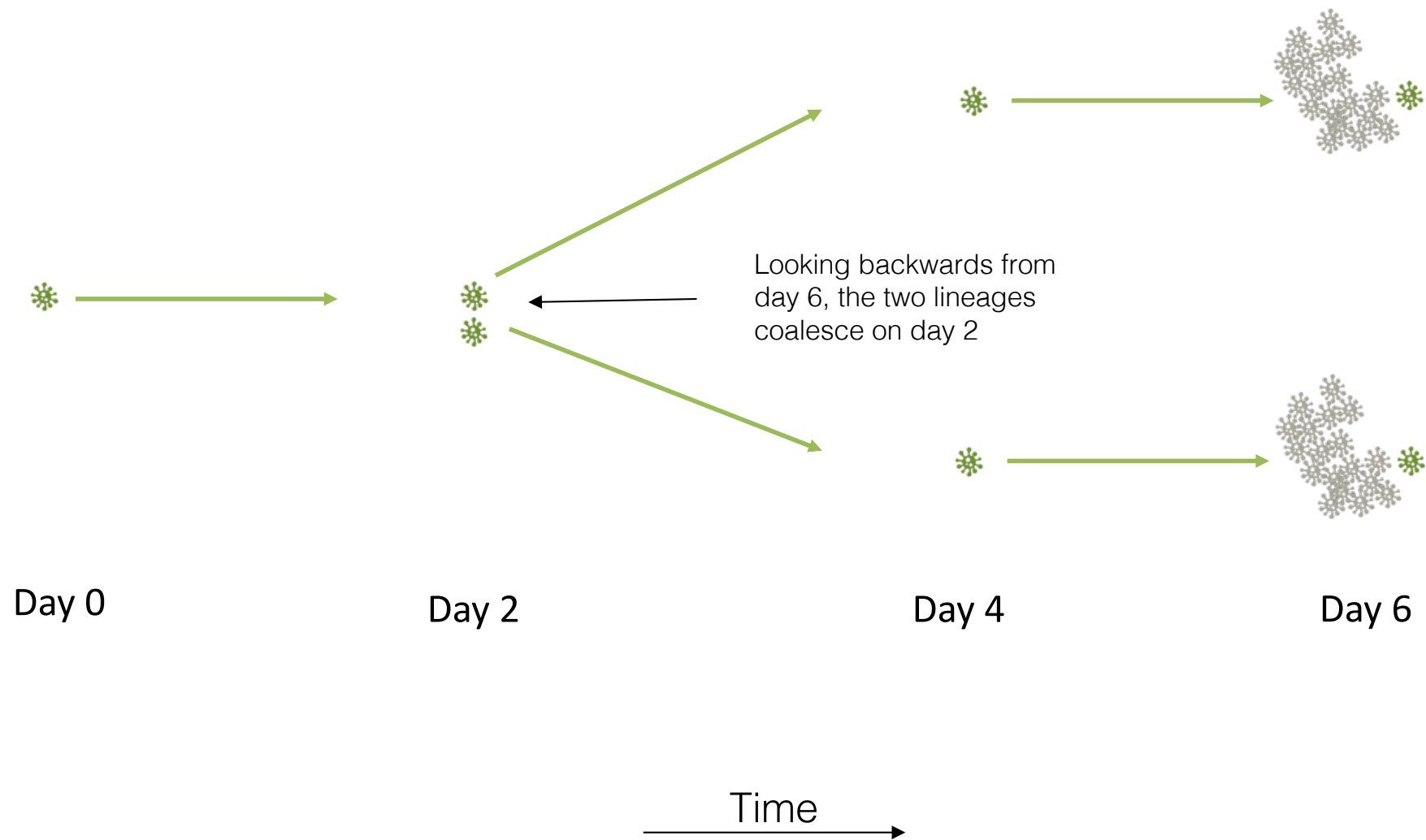
We ignore viruses that don't leave behind descendants:



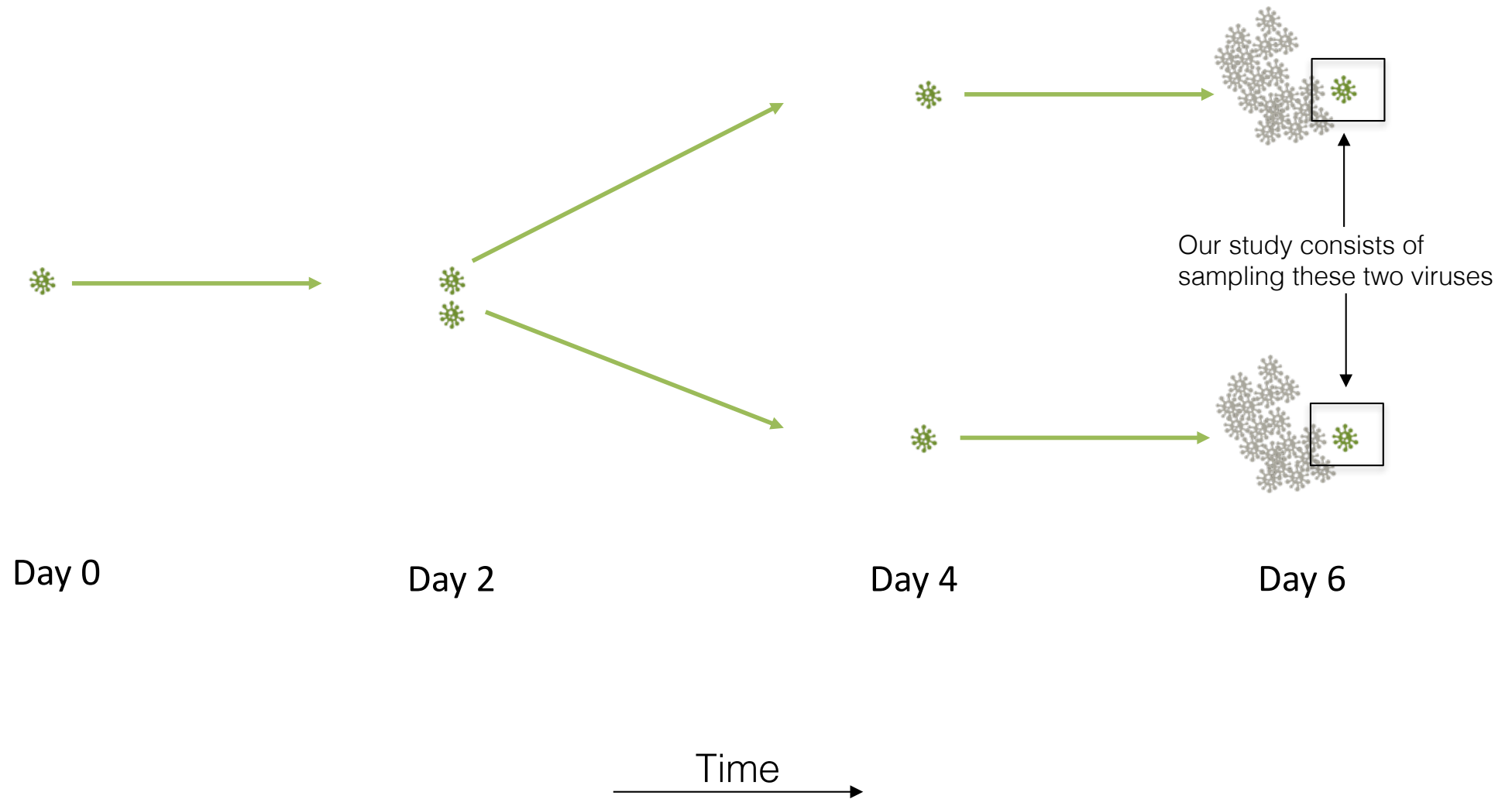
The history of the ancestral lineage has splits in time when more than one virus leaves behind descendants:



Because ancestral lineages are constructed backwards in time from the present, the splits are often called 'coalescences'

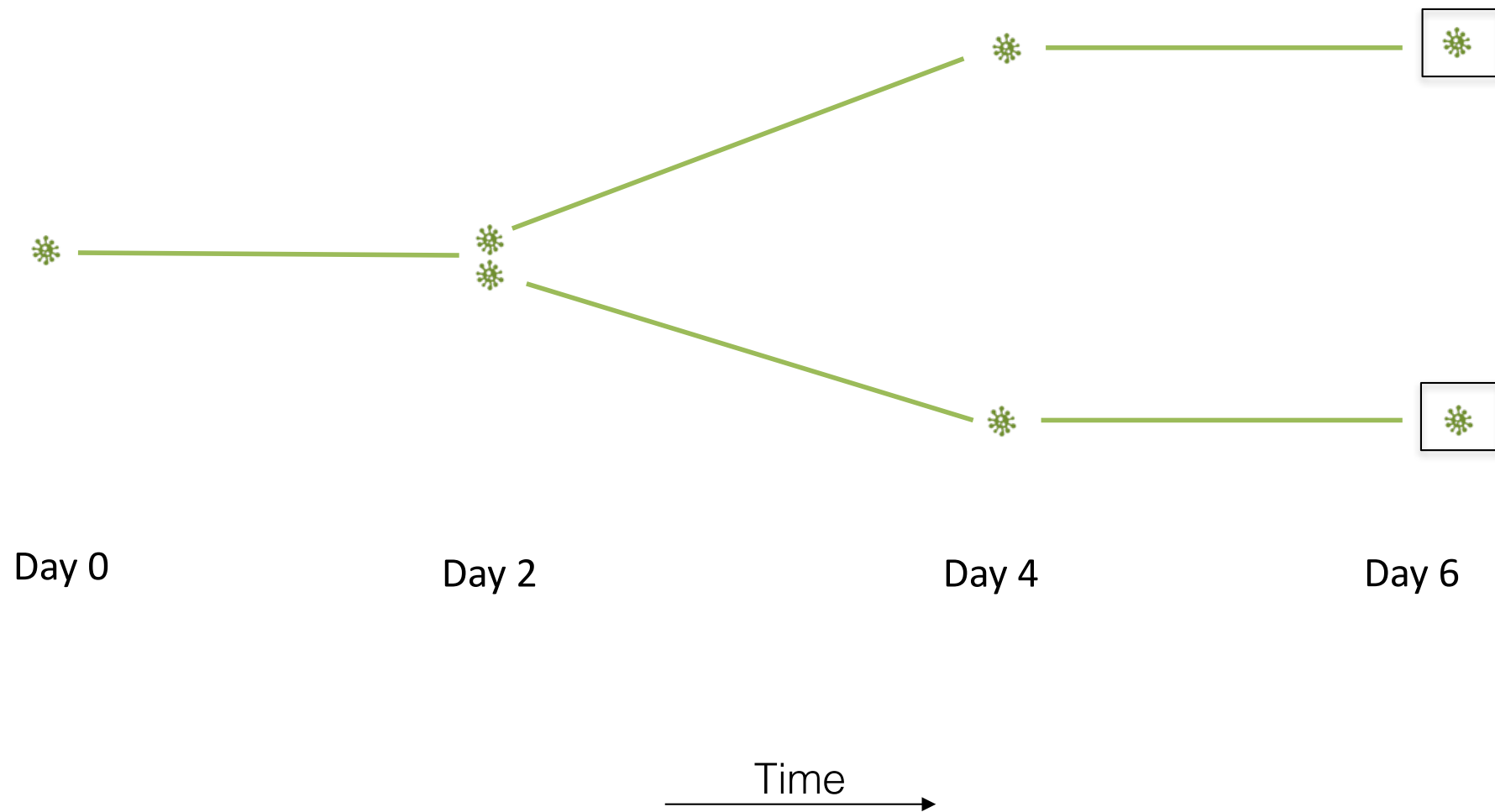


We usually infer this process indirectly from a sample:

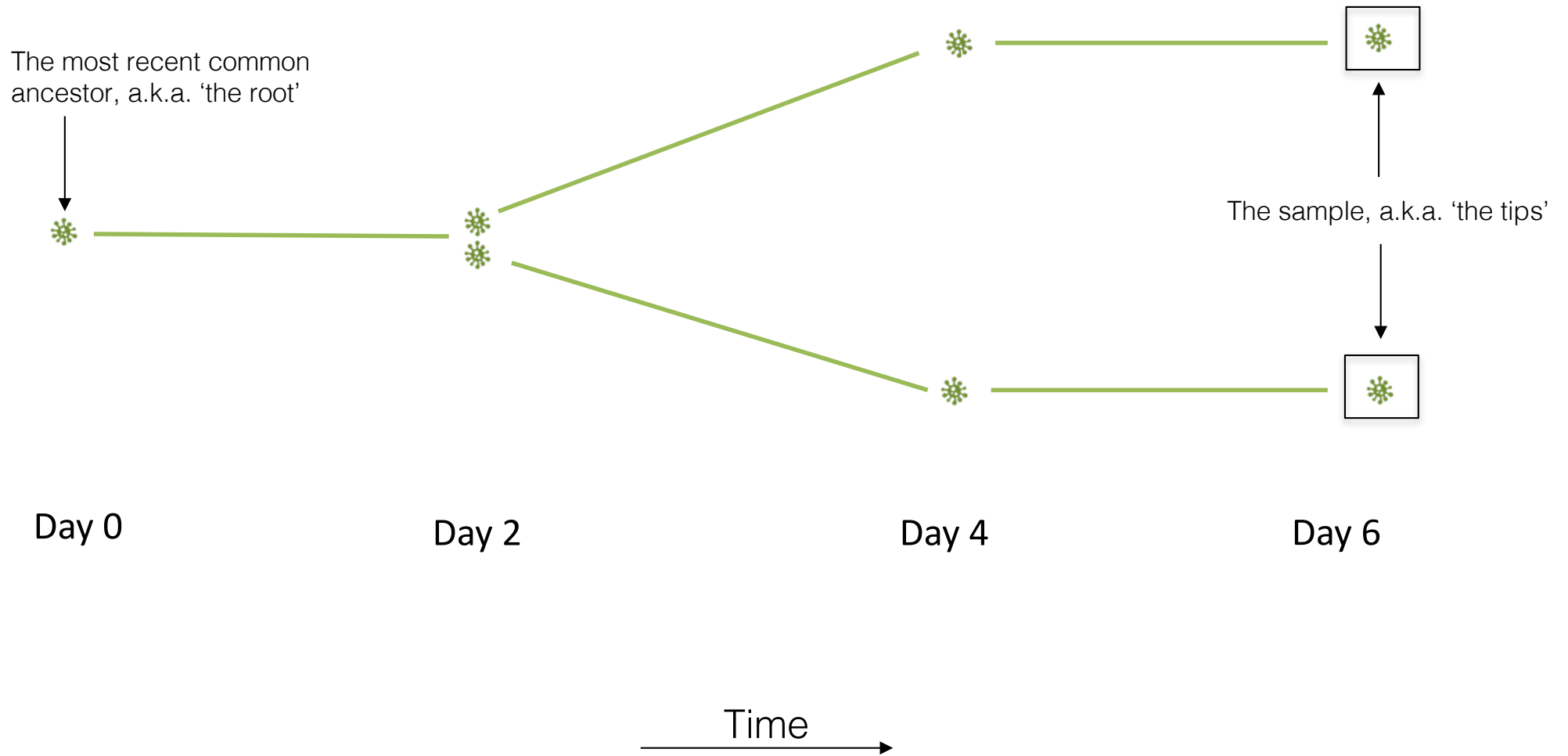


This object is the phylogeny of the two sampled viruses.

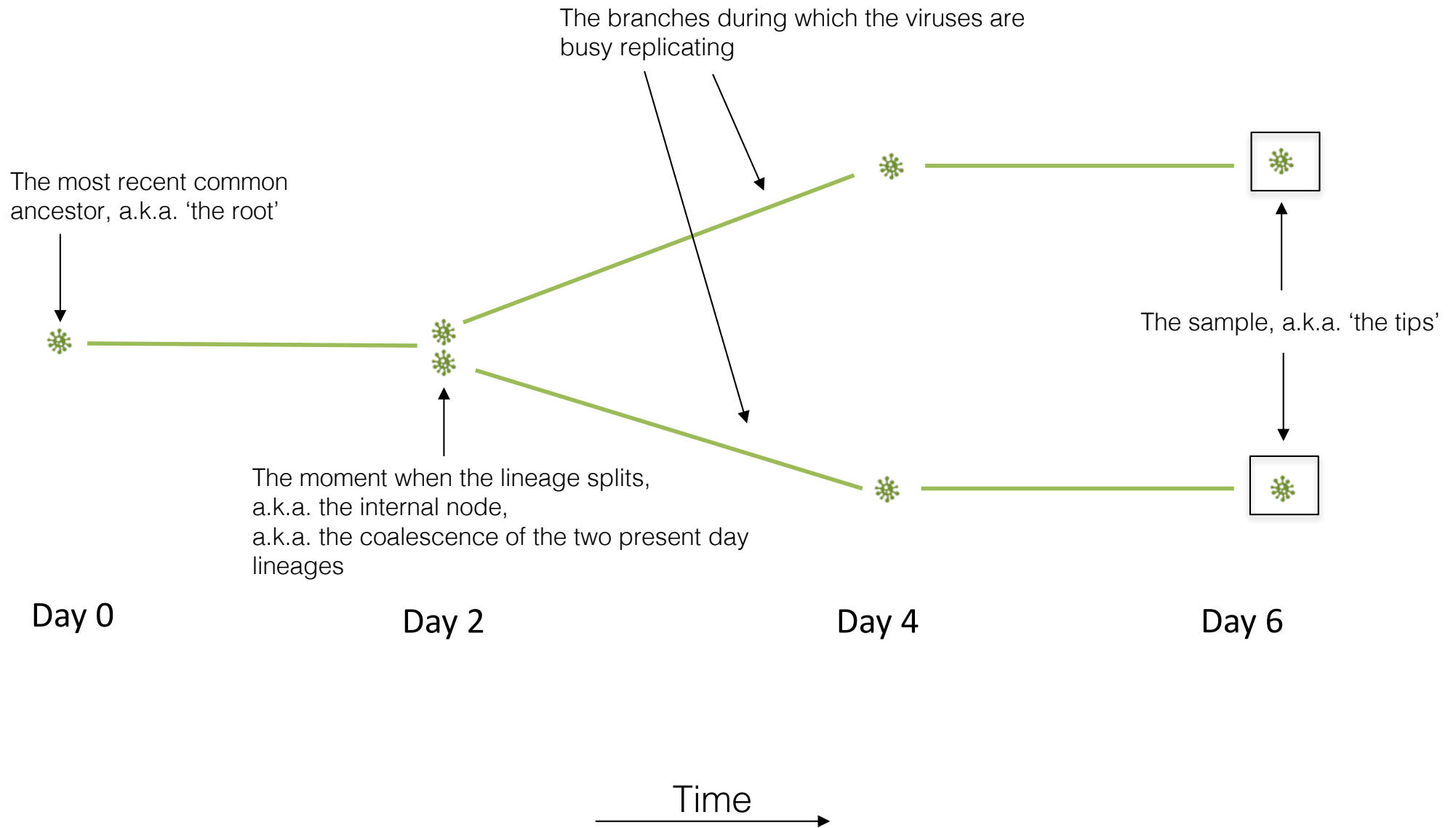
We have to infer this from data obtained from the two sampled viruses in the boxes.

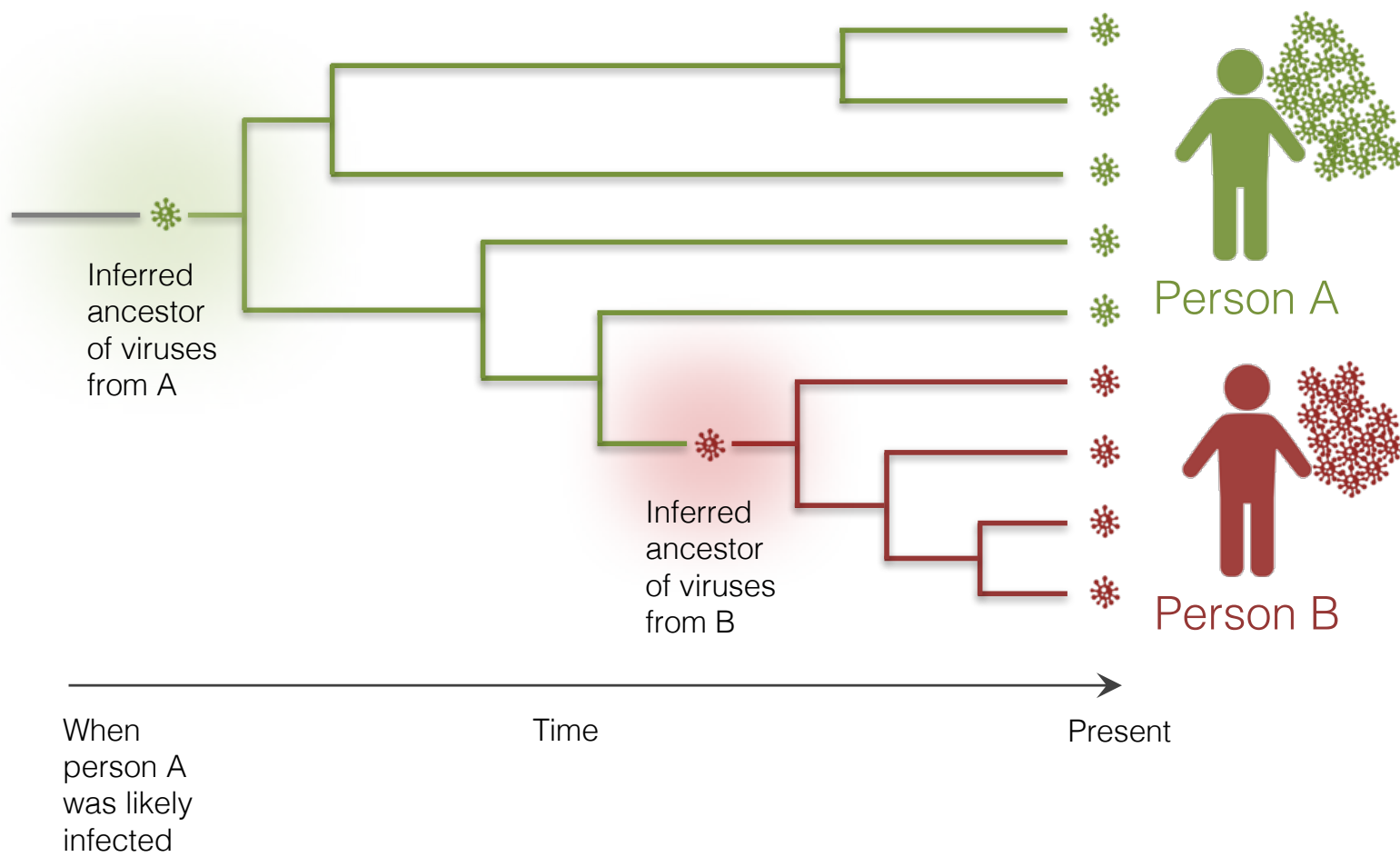


A simple phylogeny

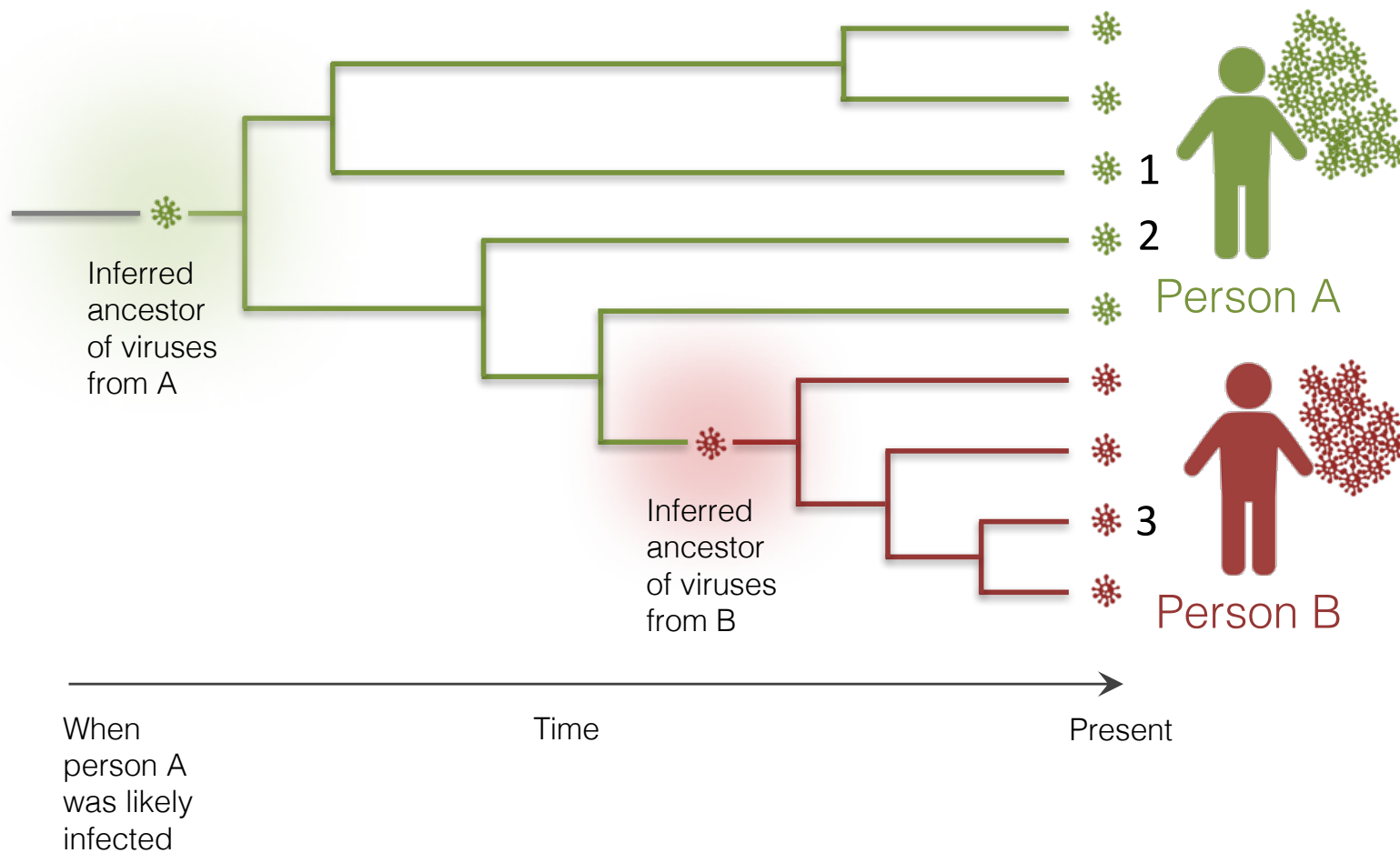


A simple phylogeny



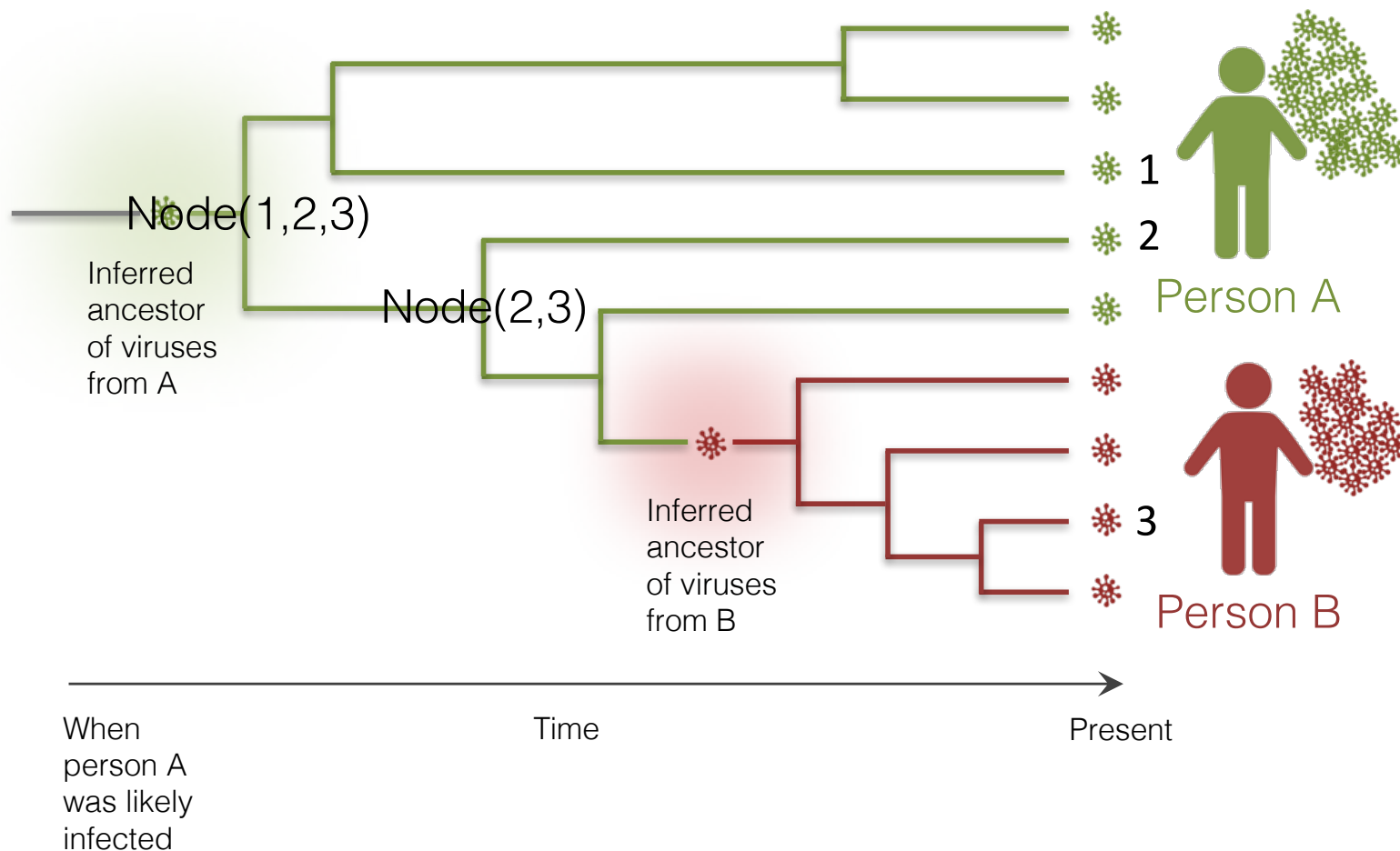


Correctly interpreting trees is not difficult, but requires a bit of practice as they are a bit visually misleading.



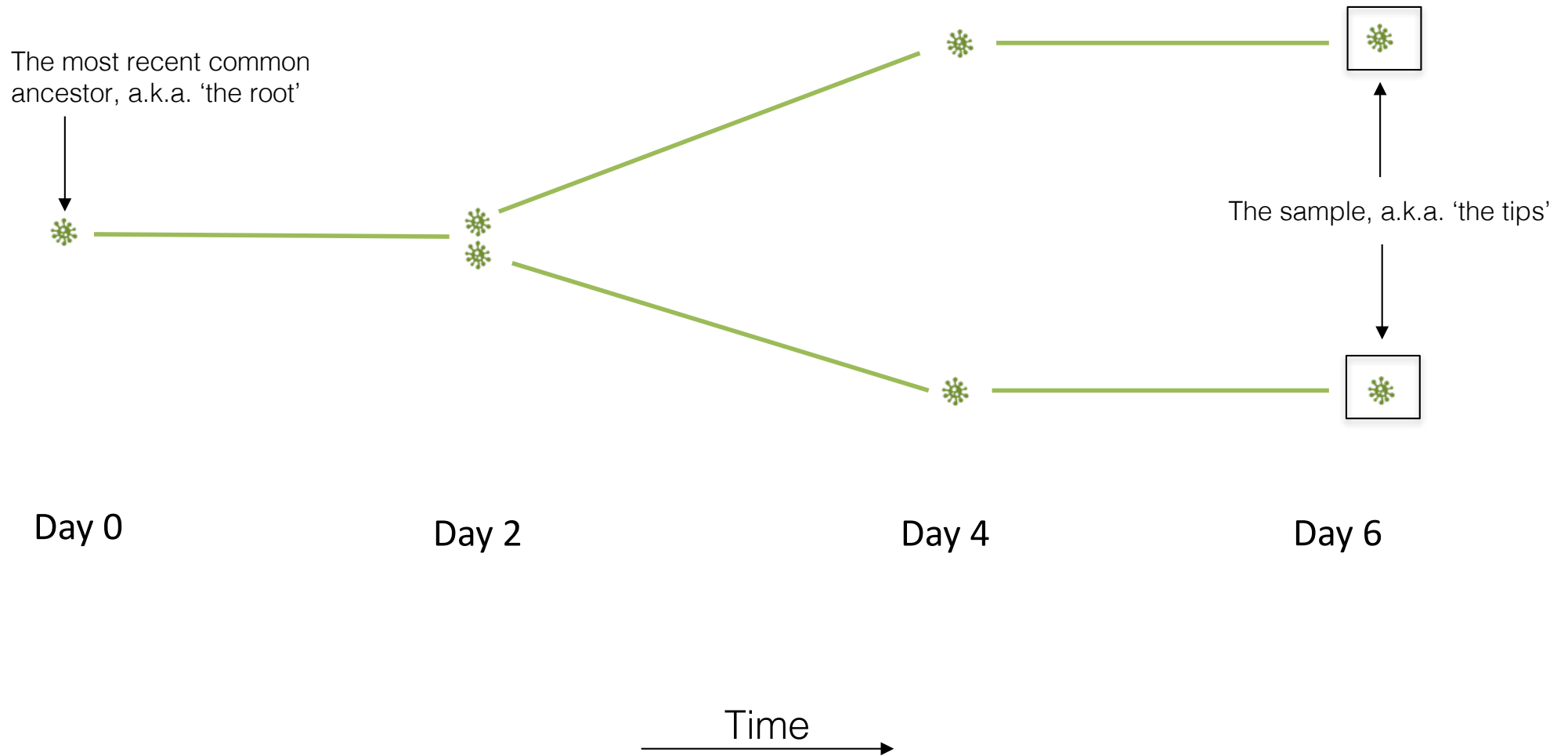
Q: which are more closely related of the viruses 1, 2 & 3?

Correctly interpreting trees is not difficult, but requires a bit of practice as they are a bit visually misleading.

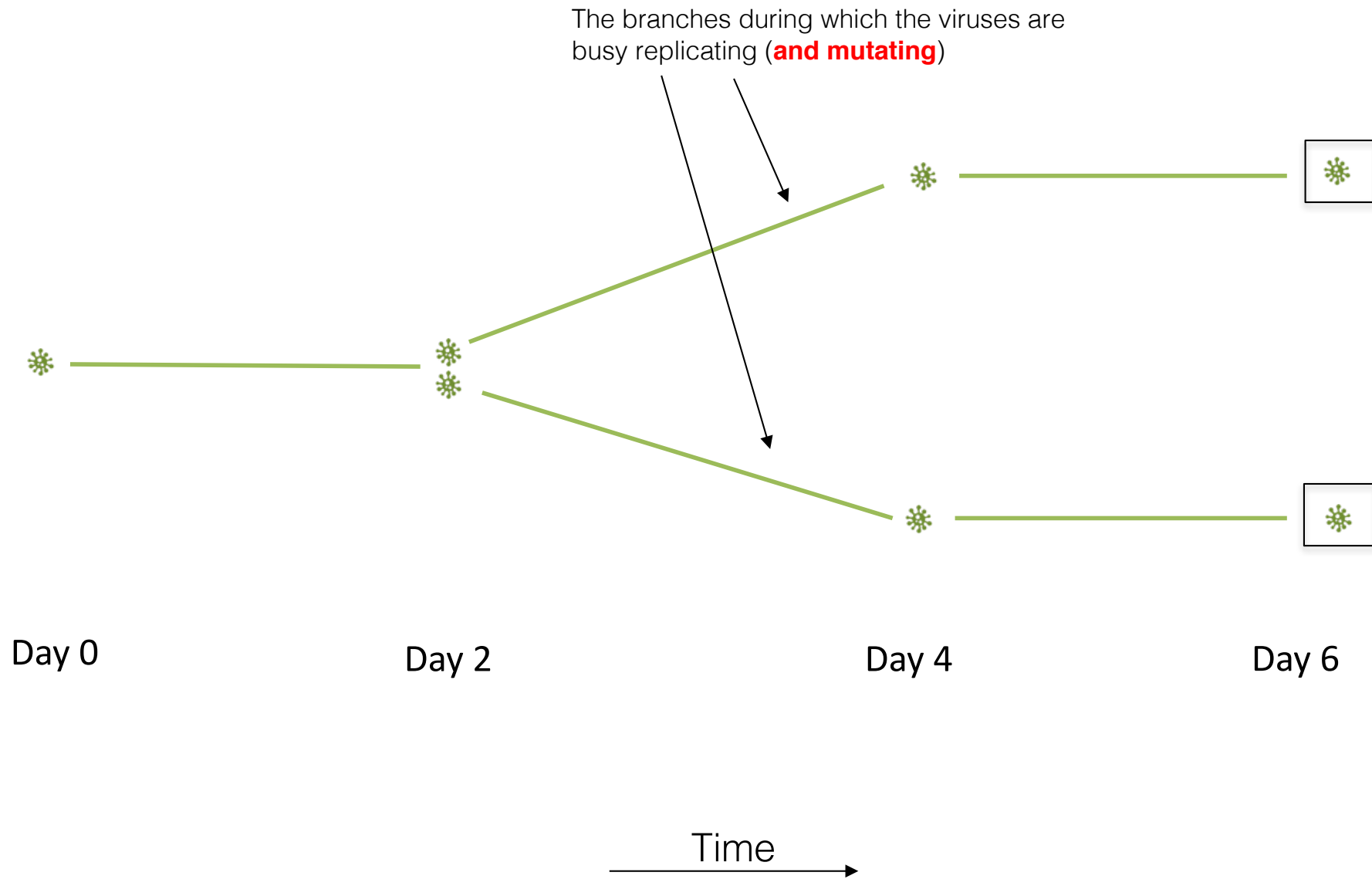


A: Viruses 2 and 3 share a more recent common ancestor than either shares with virus 1, and are expected to be more similar

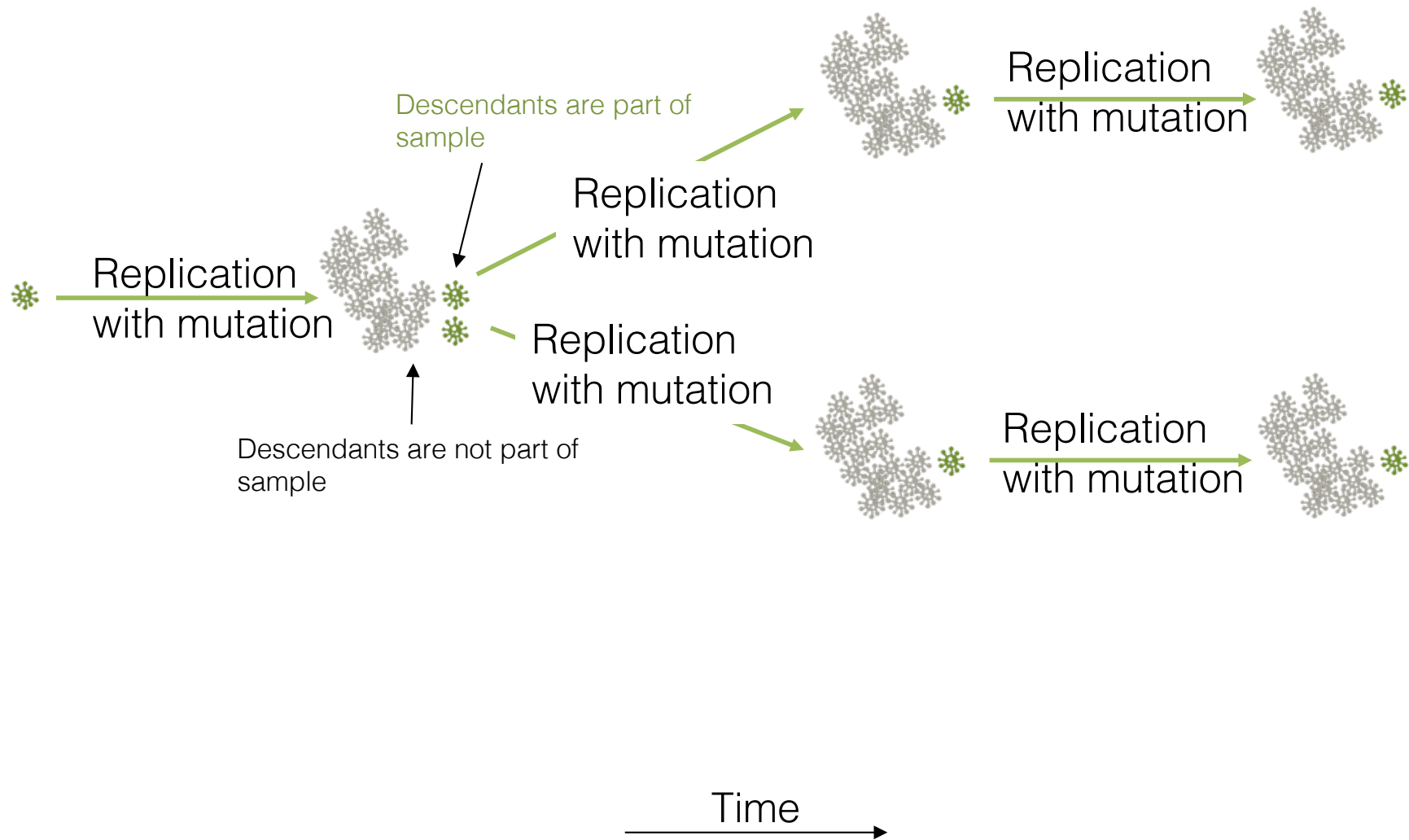
What makes our phylogeny uncertain is that we don't observe it.
Instead, we infer it from data collected only at the tips



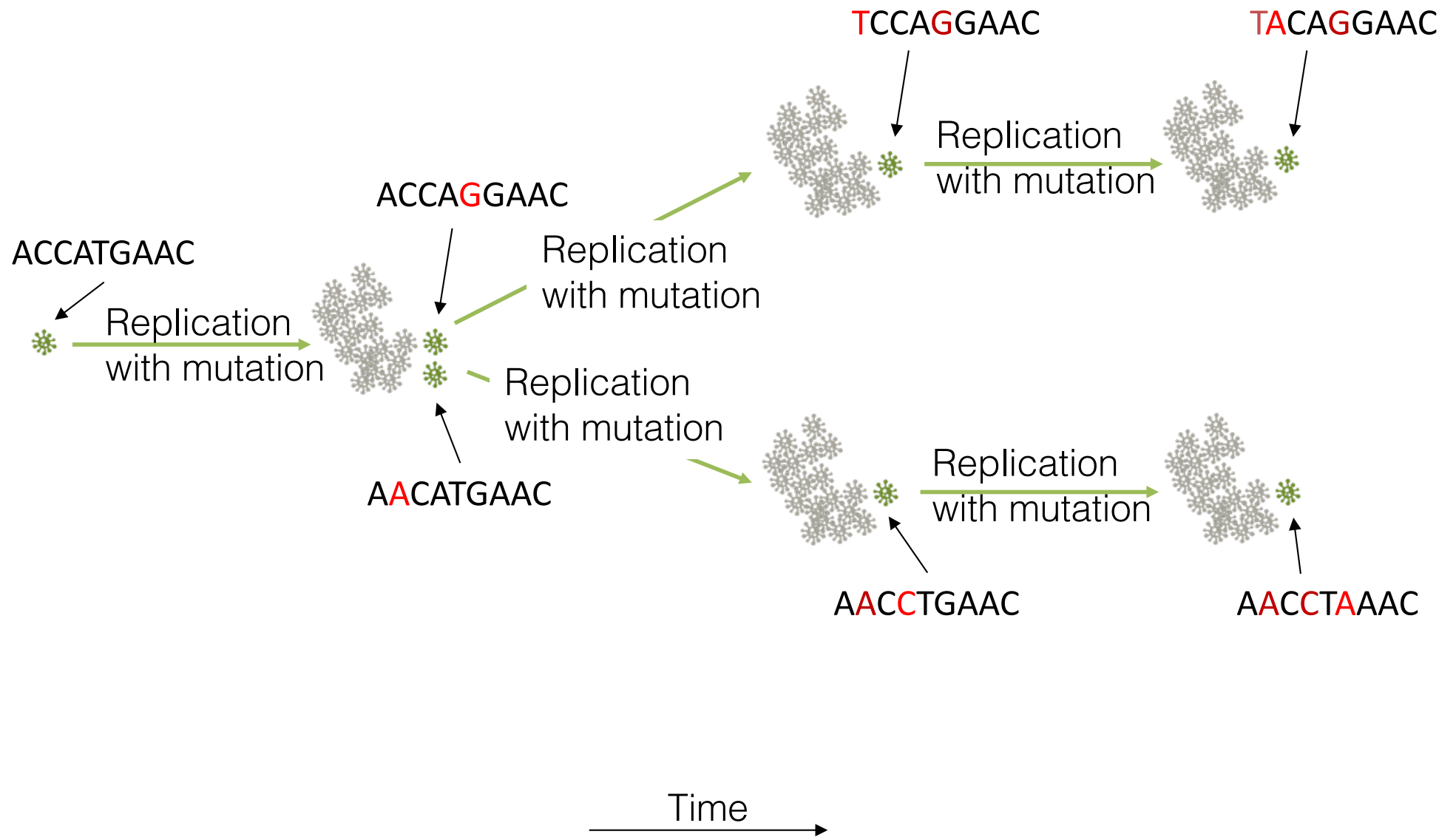
A simple phylogeny



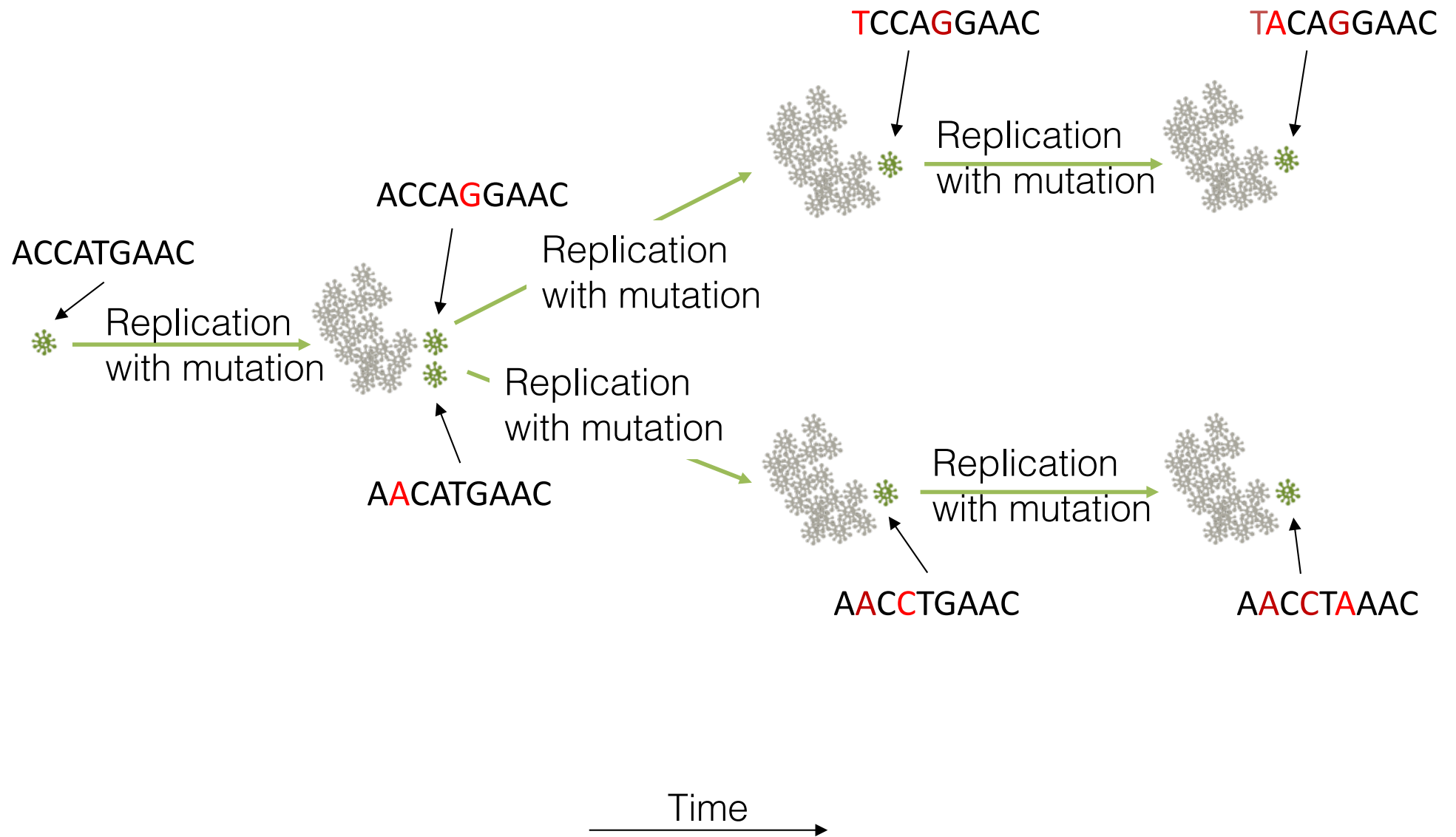
Replication with mutation



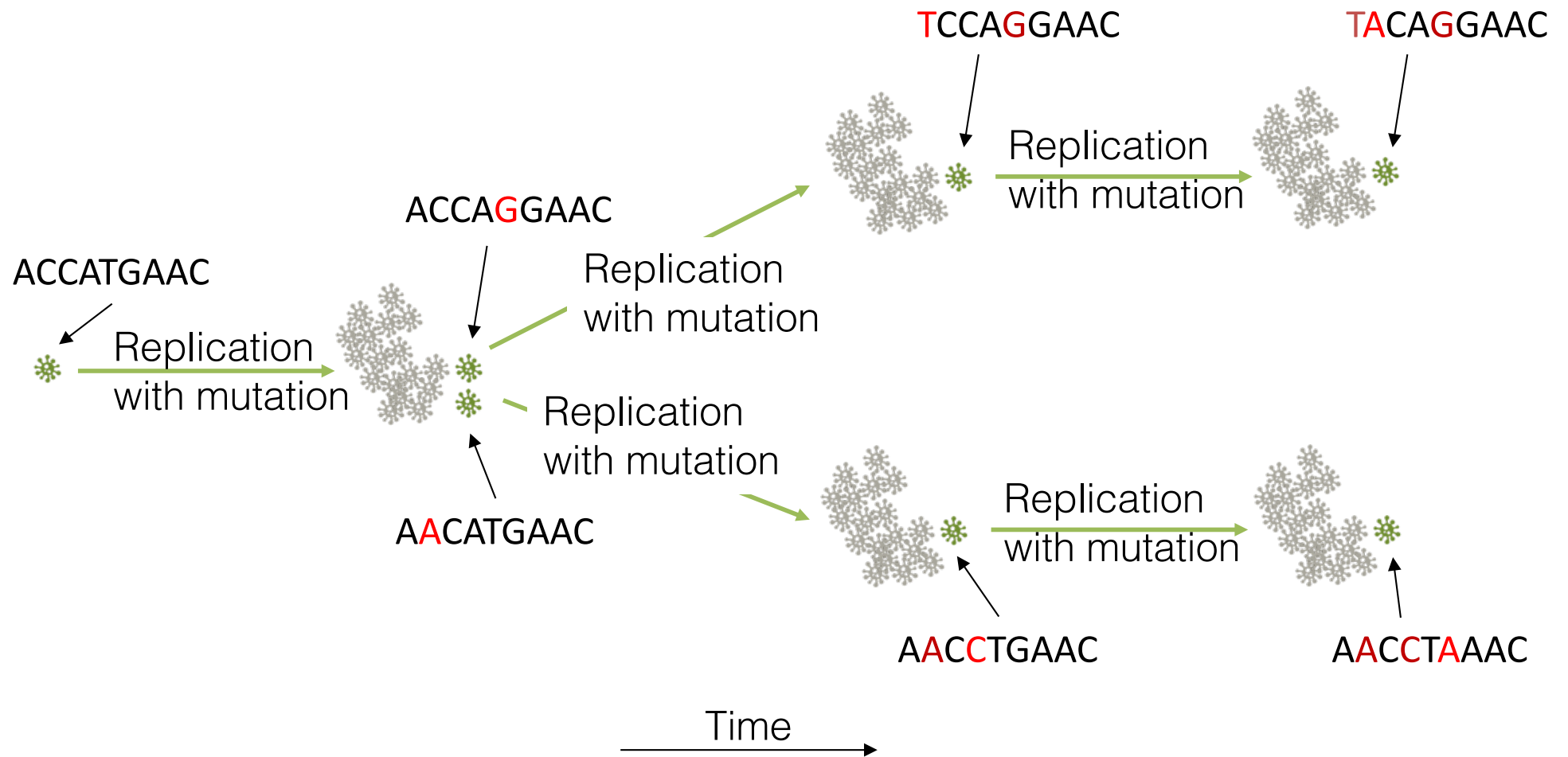
Look at the genome



A substitution is a mutation that survives in the ancestral lineage



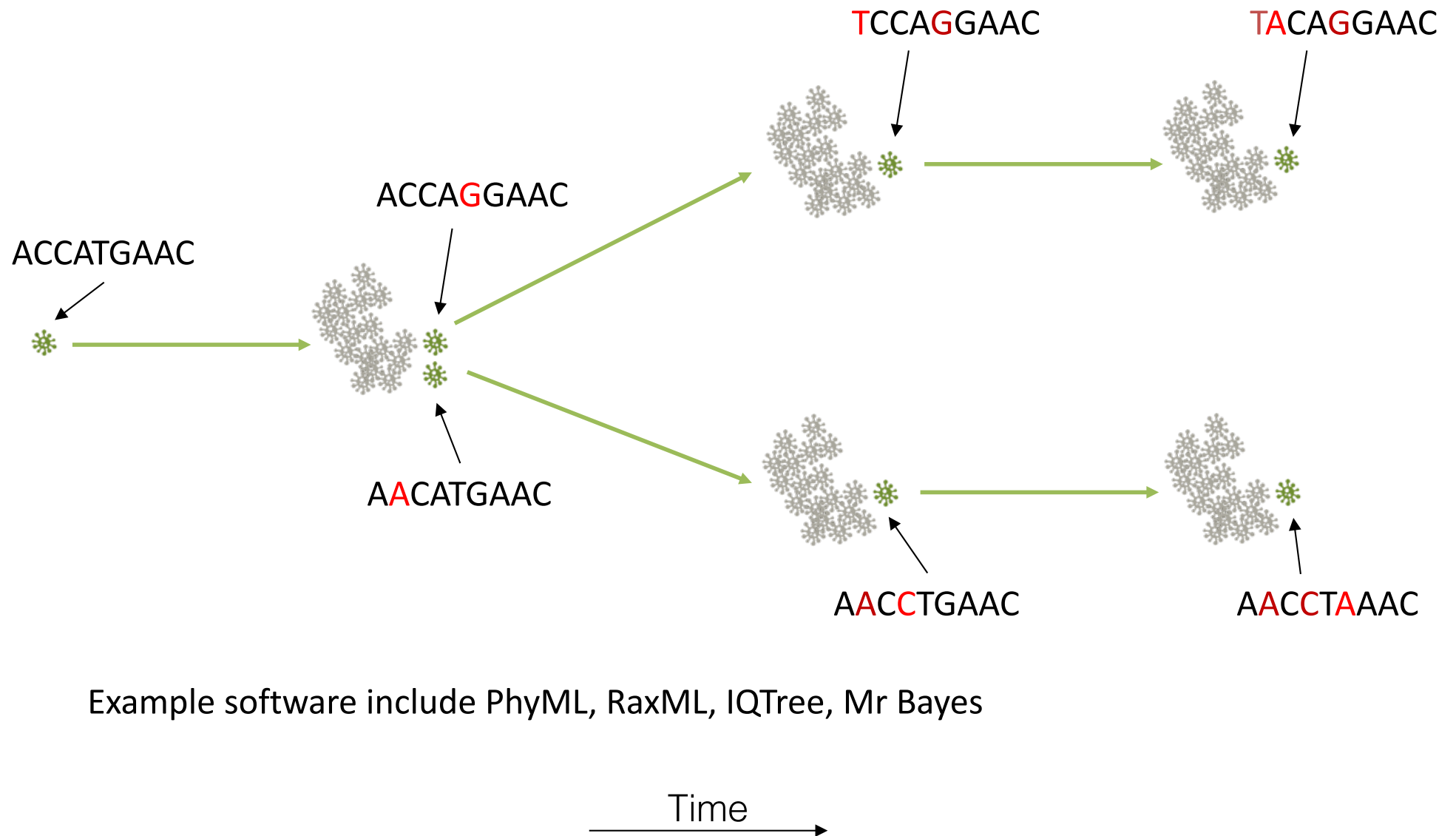
A substitution is a mutation that survives in the ancestral lineage



Substitution rates are typically much lower than mutation rates (~10 substitutions/year versus ~1 mutation per replication cycle & 150 cycles per year).

- Most mutations are harmful to the virus.
- The virus sometimes gets 'stuck' in a non-replicating latent state for years.

Phylogenetic algorithms infer the phylogeny based on molecular models of how the viruses accumulate substitutions (e.g. the relative rate of A>C versus T>G, etc.).

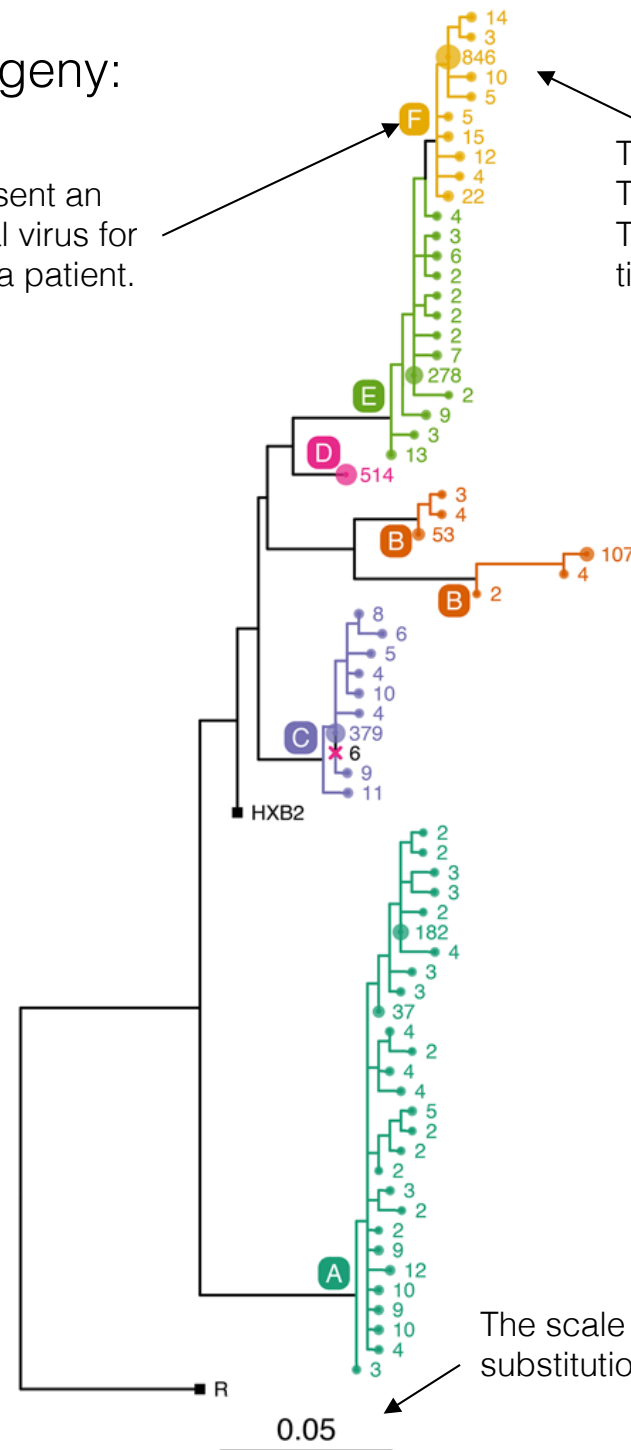


Example software include PhyML, RaxML, IQTree, Mr Bayes

An example real phylogeny:

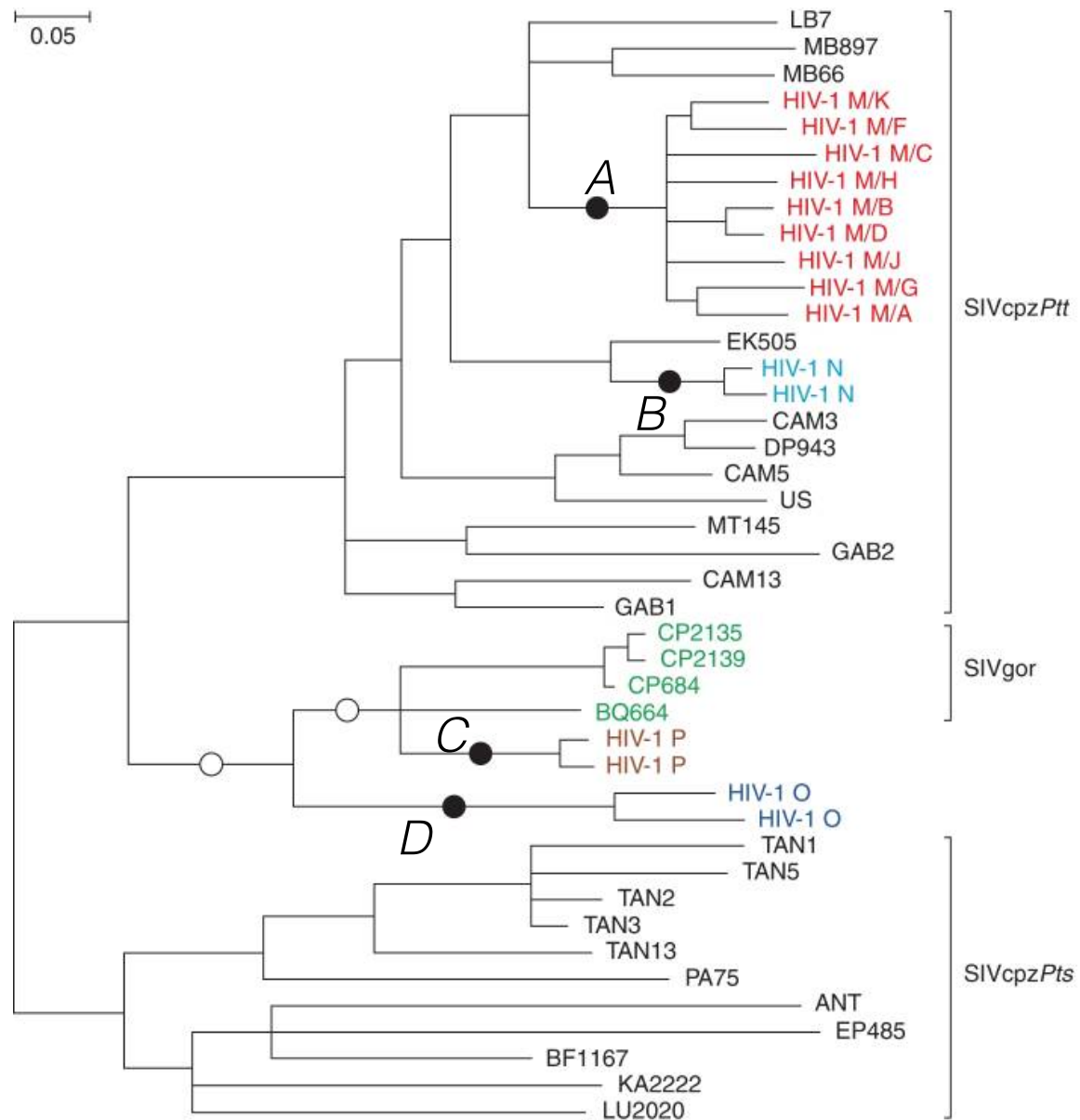
The letters represent an inferred ancestral virus for all the viruses in a patient.

The colors represent patients.
The tips represent a virus.
The number represent the number of times the same genotype was found.

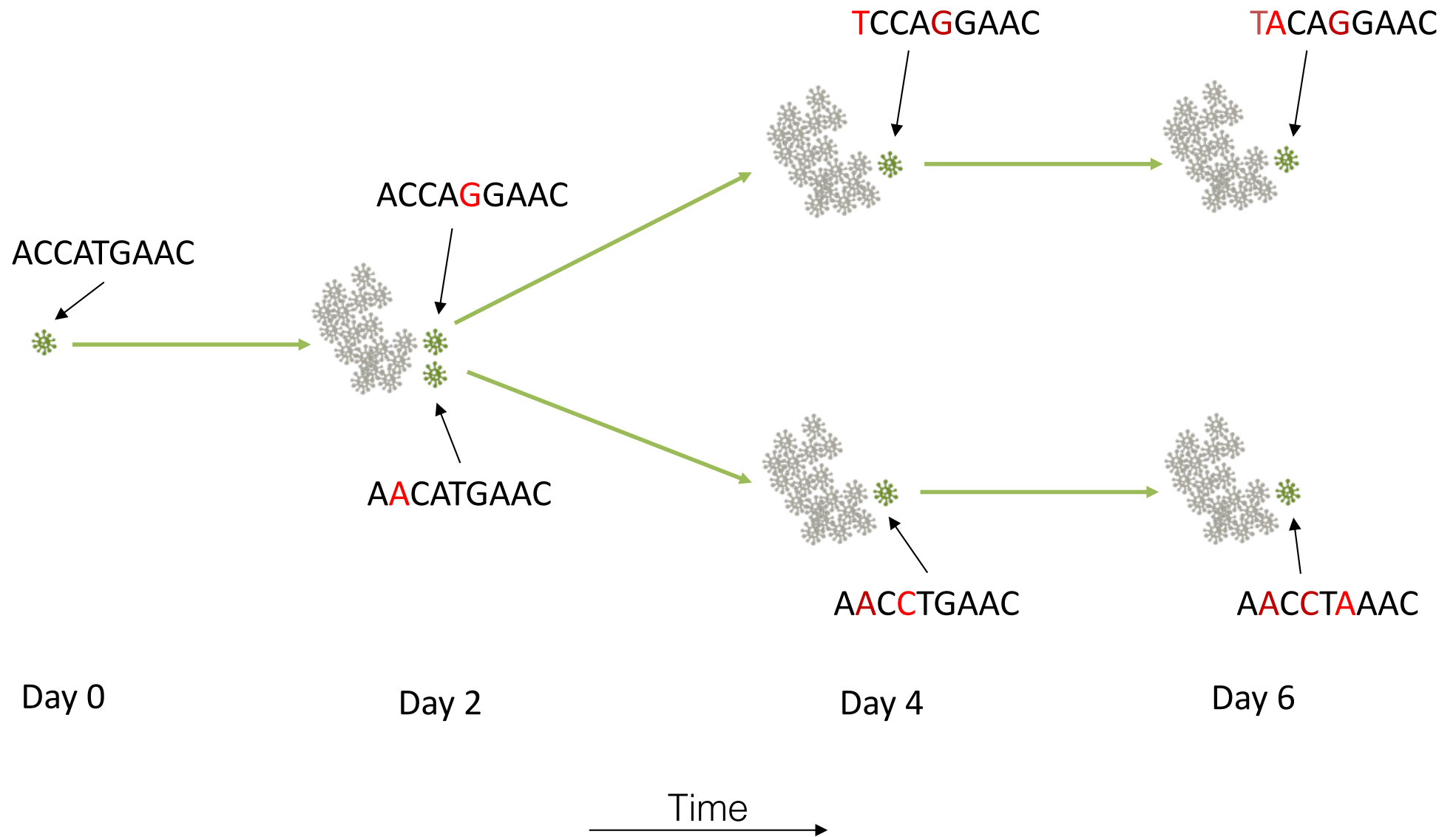


The scale bar is in units of substitutions per site

Phylogenies can provide information at very different scales

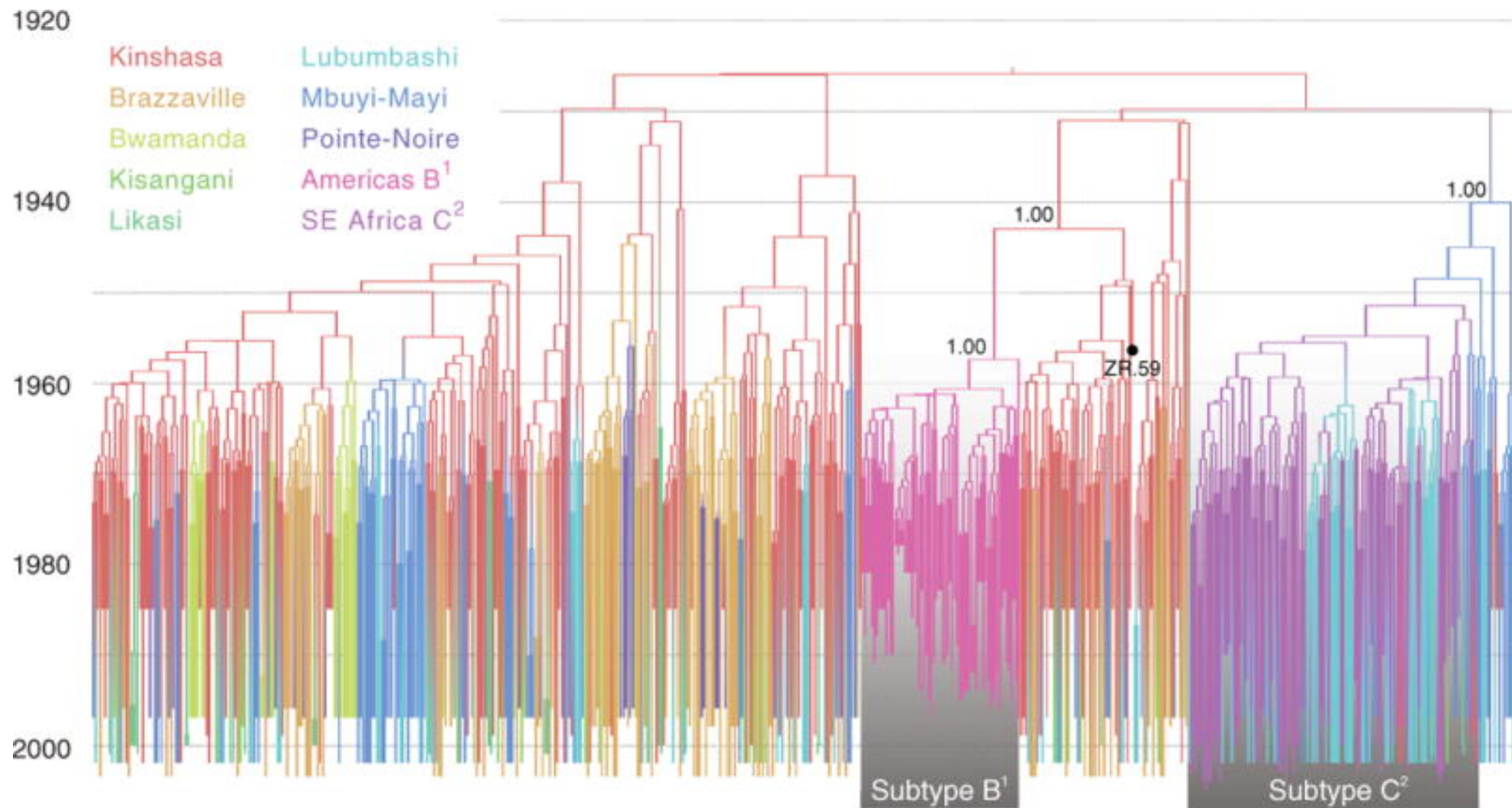


A tree can be dated if you have enough data to infer rates of substitution

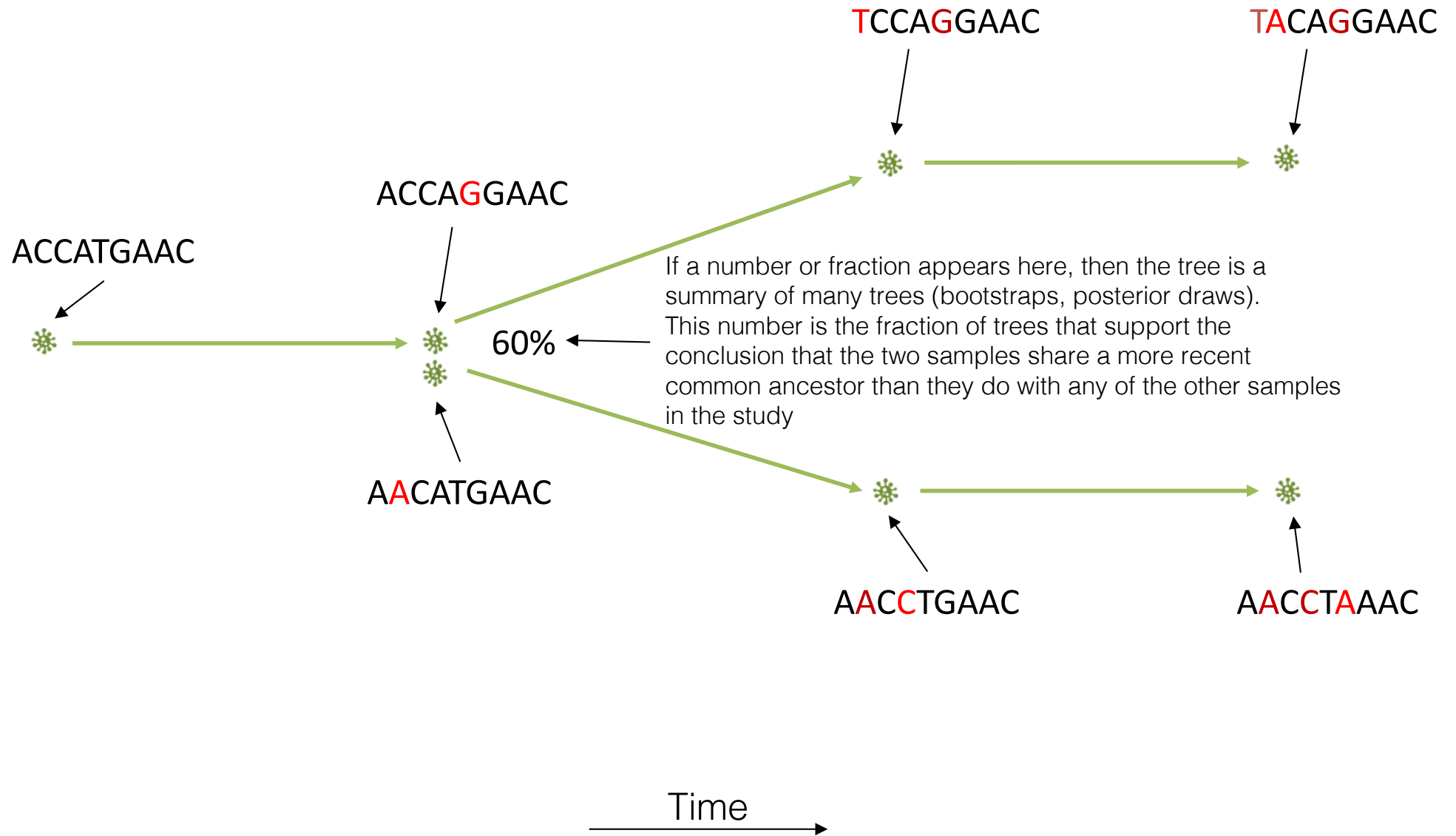


Example software include BEAST, TreeDater, TreeTime, Least Squares Dating (LSD)

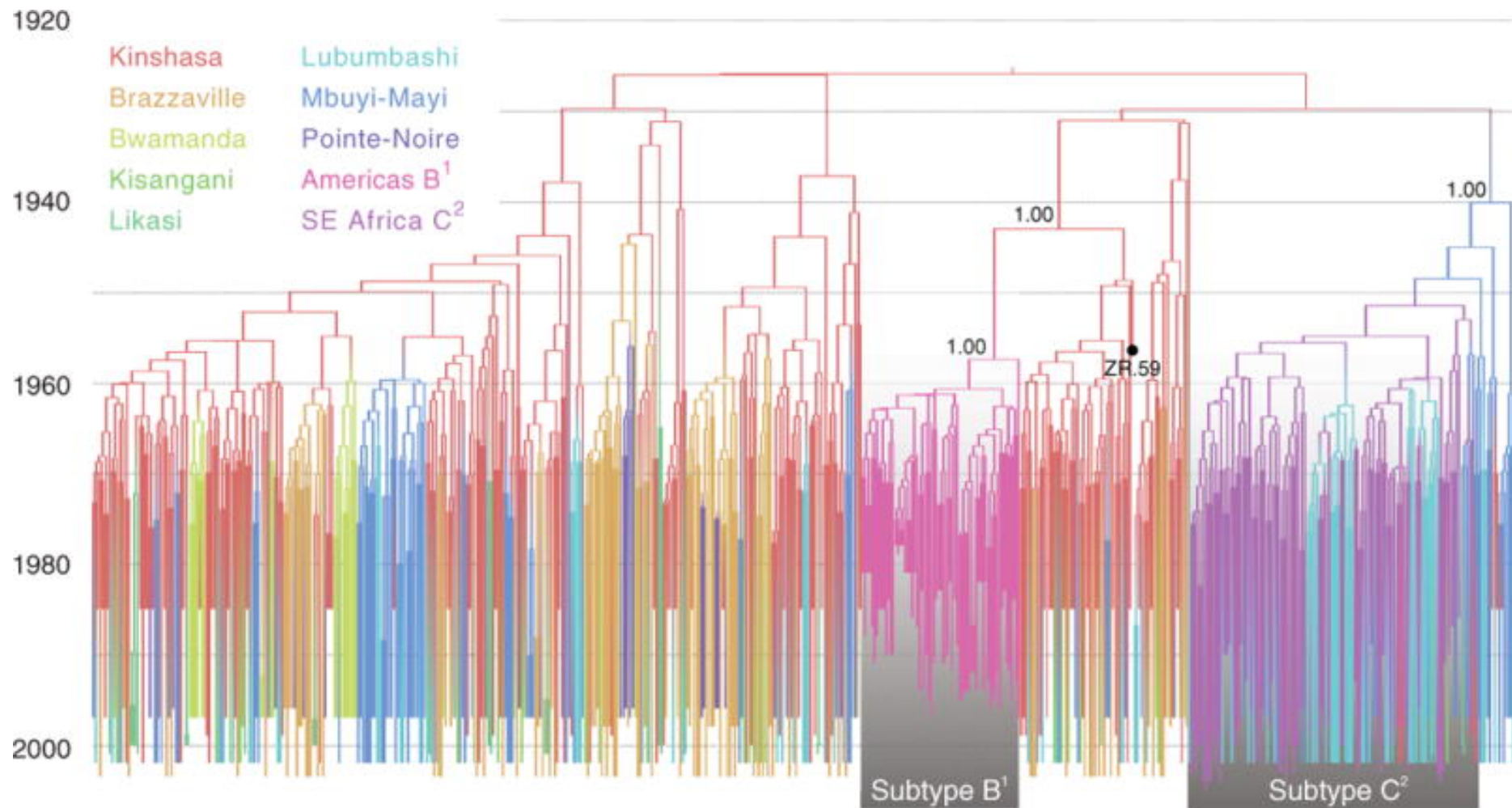
This phylogeny has been rotated for clarity. The past is at the top.



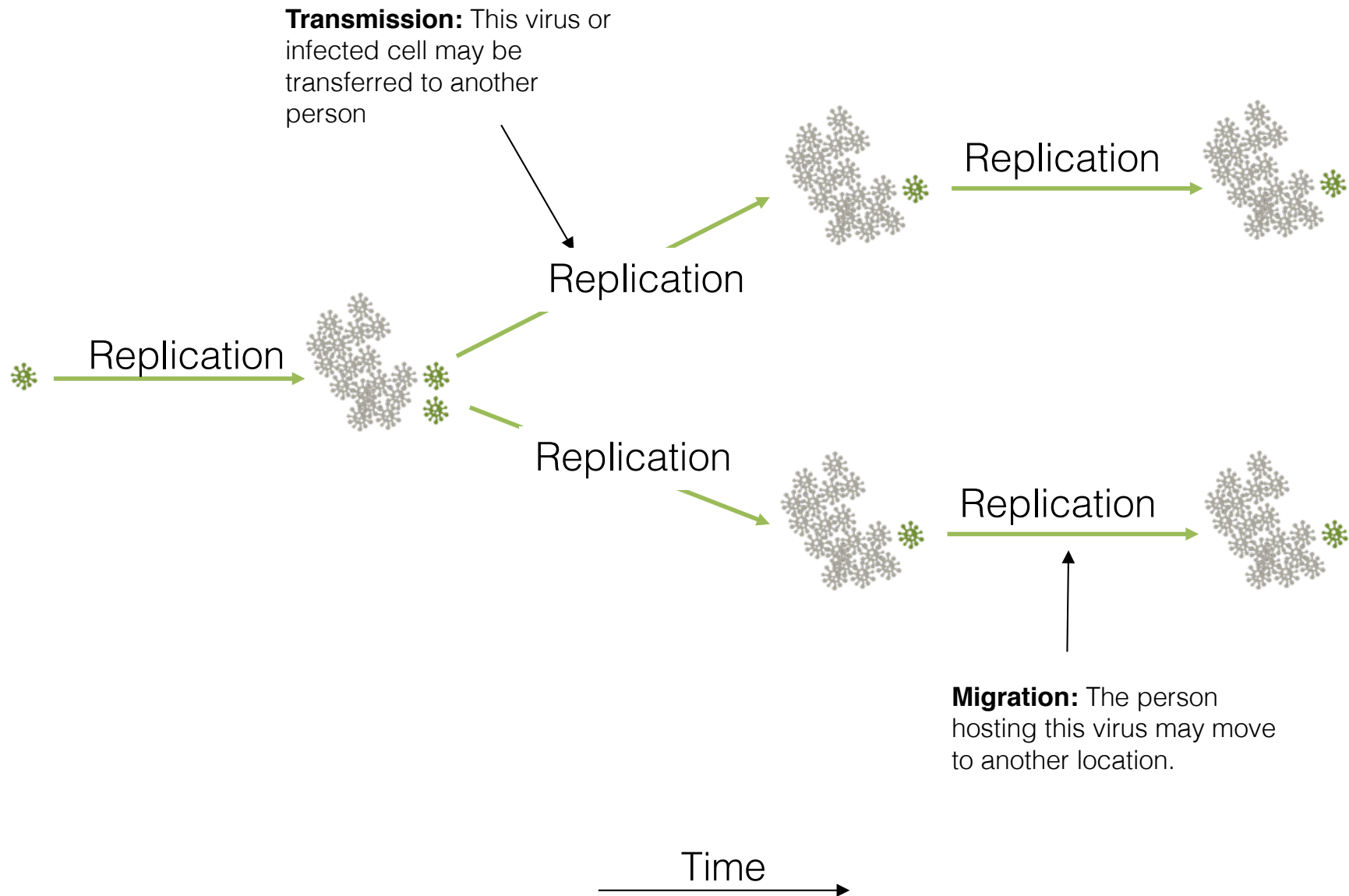
In practice, inferring phylogenies is full of uncertainty



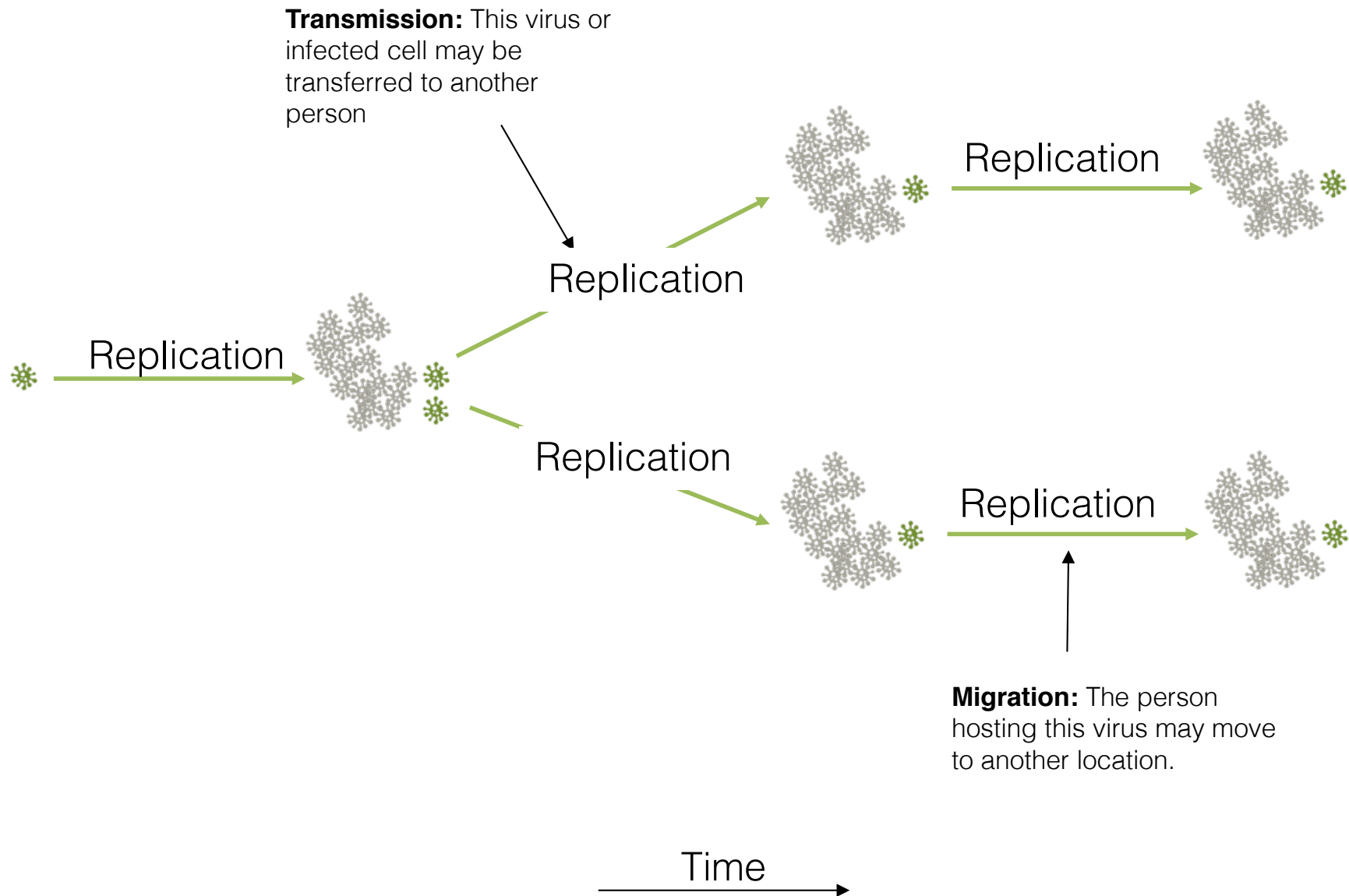
This phylogeny has been rotated for clarity. The past is at the top.



Interesting things may happen to viruses as they replicate:

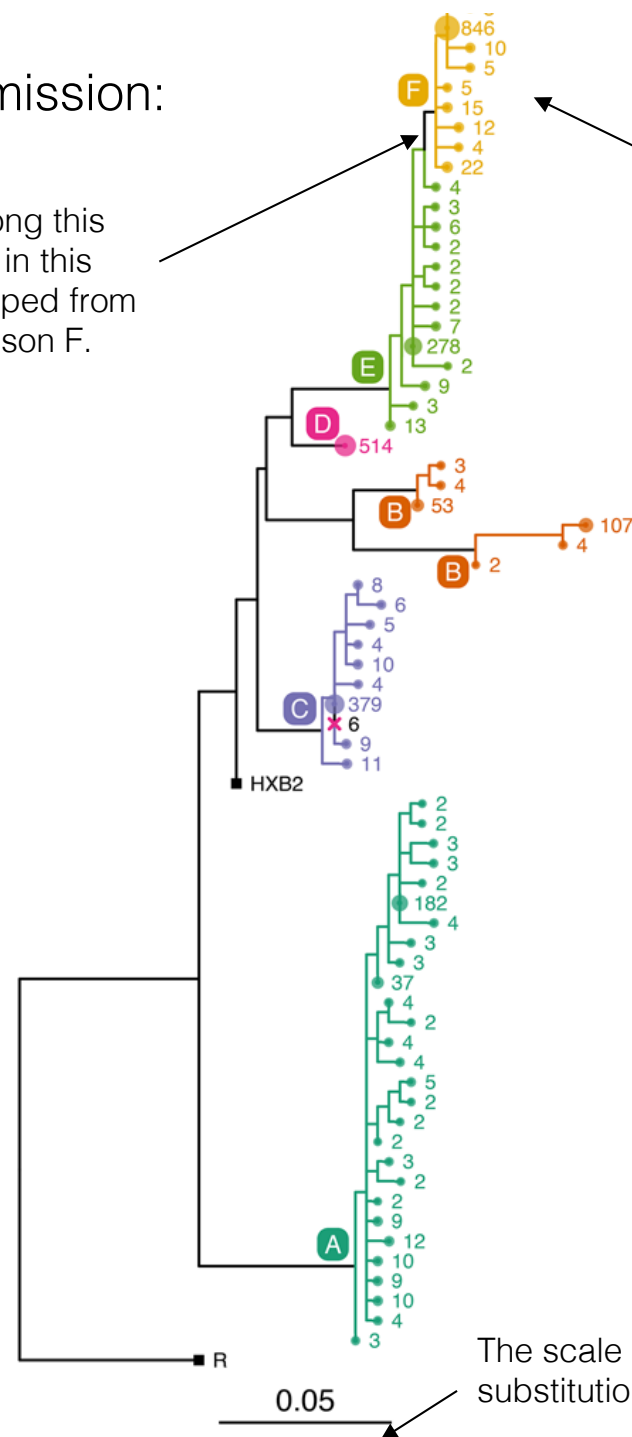


Interesting things may happen to viruses as they replicate:



A real example of transmission:

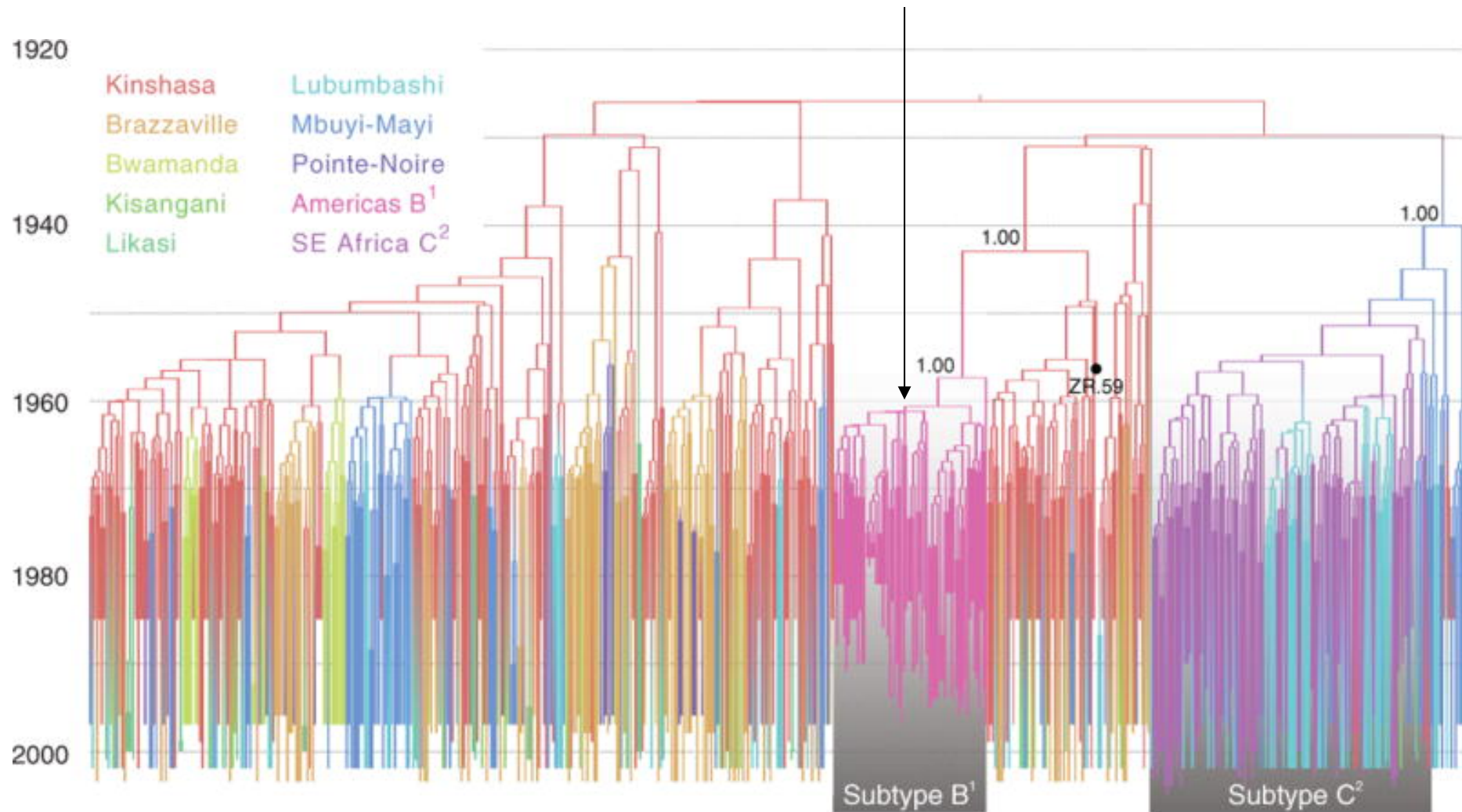
At some point along this branch, the virus in this lineage likely jumped from person E into person F.



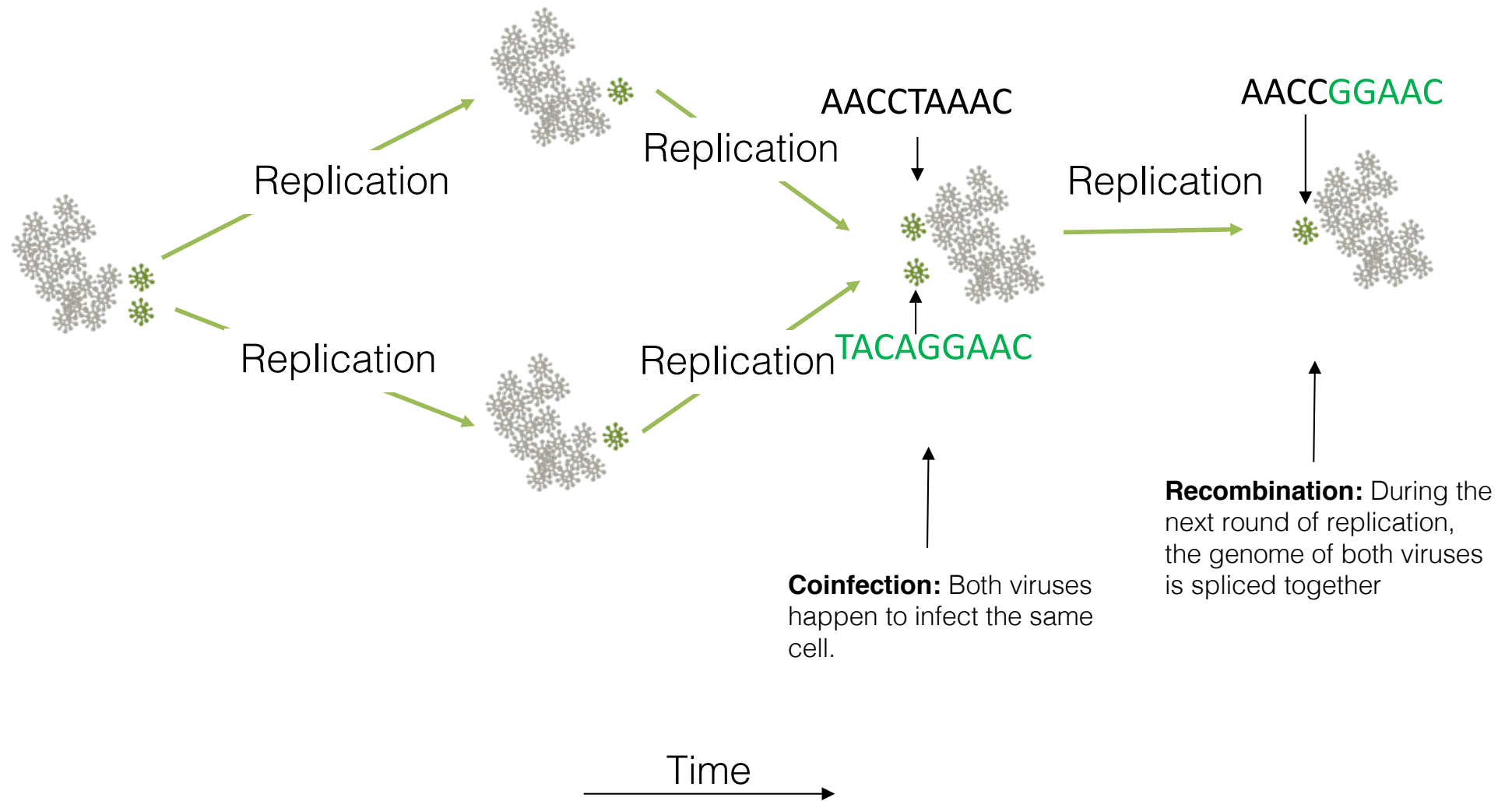
The scale bar is in units of substitutions per site

A real example of migration

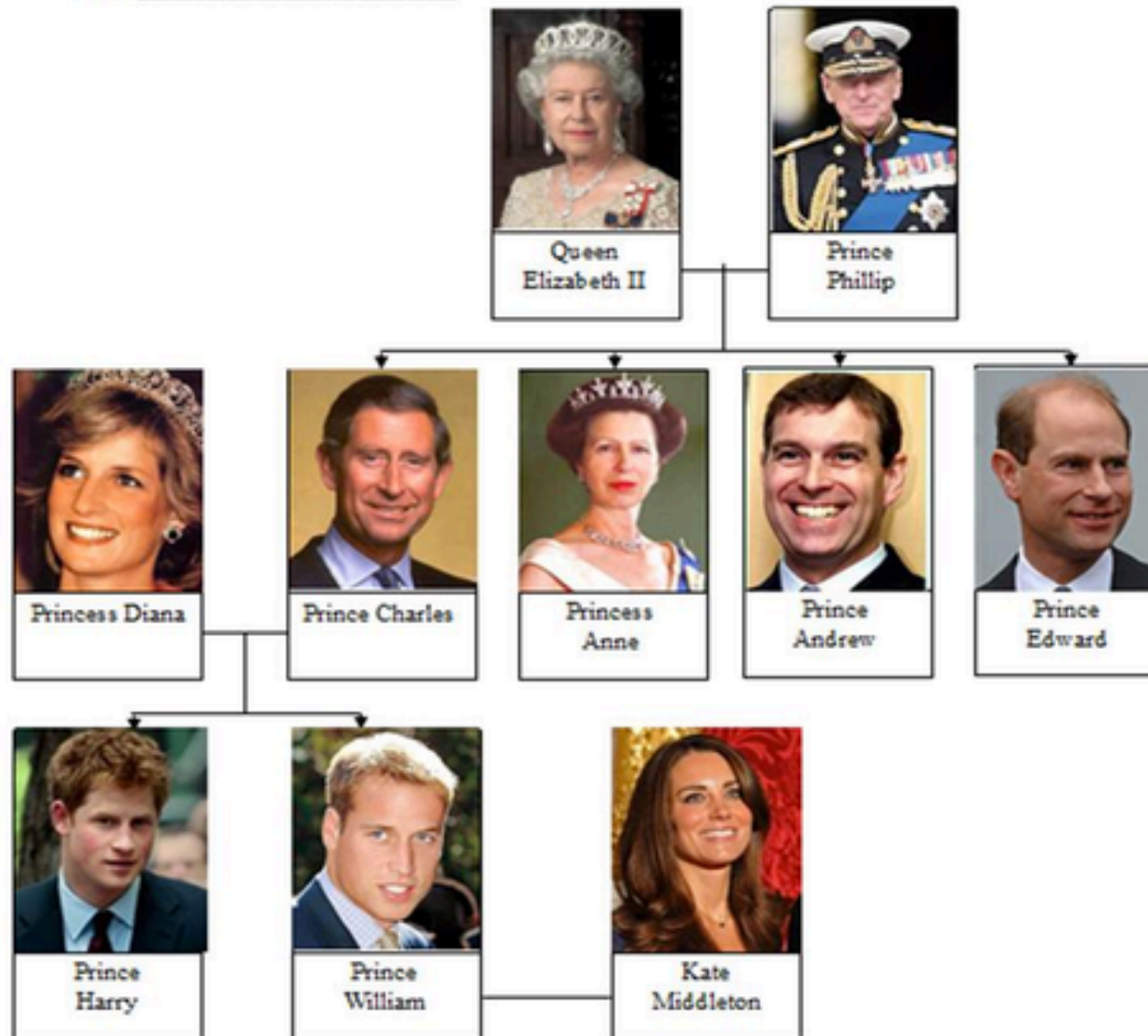
At some point along this branch, the person/people carrying this virus moved from Kinshasa to the Americas .



The final really complicated effect: recombination messes up tree-like structure, creating a 'ancestral graph instead'.



Luckily, HIV is not the only biological replicator that recombines.
Some have been well studied.



Summary

- Phylogenetics provides powerful insights into dynamics of virus spread at different scales.
- Trees are a natural way to describe ancestry (recombination still challenging).
- Ancestral state reconstruction is the key link to epidemiology.
- We have talked a lot about phyloscanner, but there are many other tools and methods we should use with PANGEA to obtain insights.

Thank you.

Questions?