Imperial College London

MACHINE LEARNING & GLOBAL HEALTH NETWORK

# Introduction to phylogenetics

Alexandra Blenkinsop
With thanks to Chris Wymant and Thibaut Jombart

February 24th 2024

# Overview

- Why phylogenetics?
- How to interpret a phylogeny
- How to run a phylogenetic pipeline
  - Data, steps and tools for inferring ancestral relationships
- Analysing phylogenetic trees

# What can phylogenetics tell us?

# Phylogenetics for understanding epidemics

**How can we target our interventions in an epidemic to have a greater impact?**

We can learn about the epidemic through analysis of pathogen sequence data.

Examples:

- characterising spatial transmission dynamics
- characterising population-level drug resistance
- detection of new variants
- estimating epidemiological parameters
- identifying population-level drivers of transmission

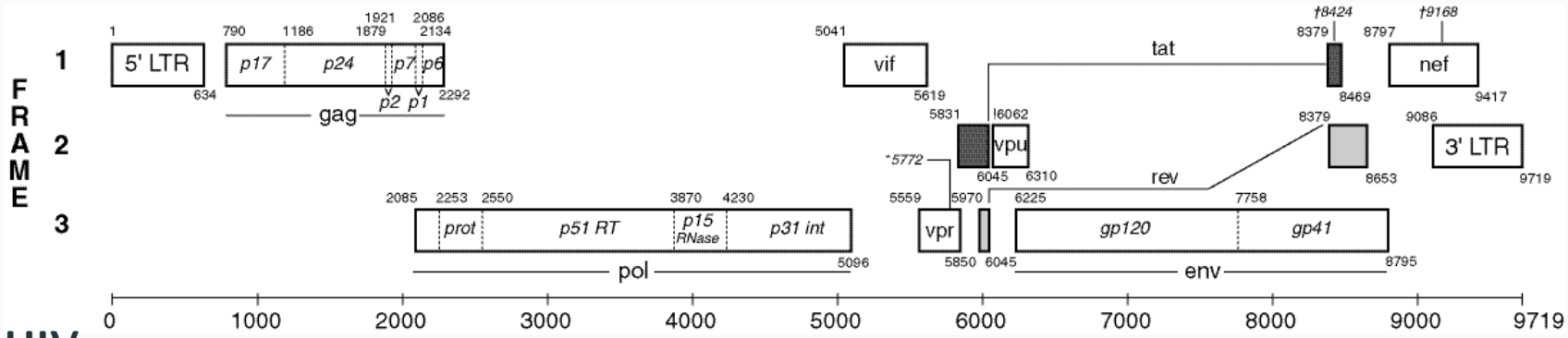
# Viral sequences

Nucleotides are the building blocks of nucleic acids - chains of these (sequences) encode information in DNA/RNA

- Consensus sequence
  - One representative sequence of most frequent nucleotide at each position
- Next generation sequencing (NGS)
  - Sequence thousands of reads simultaneously
  - Can be used for inferring direction of transmission
  - Supports robustness of conclusions

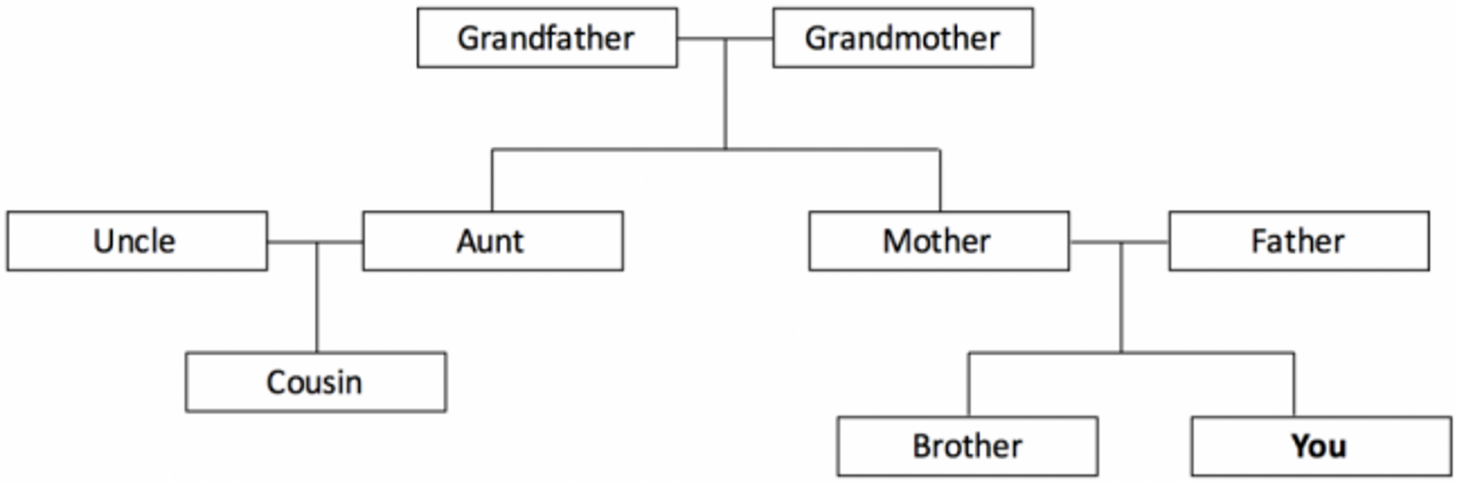
# Viral sequences

- Whole genome sequence covers all nucleotide positions
- Partial genome sequence covers part of the genome (e.g. *gag*, *pol*, *env* for HIV)



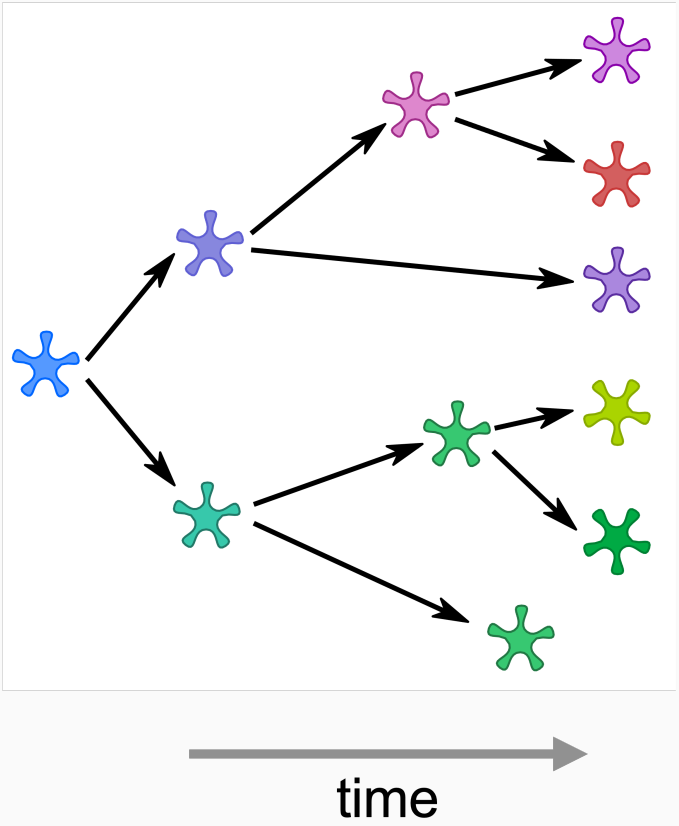

**HIV genome map**, `https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html`

Using sequence data to infer ancestral relationships

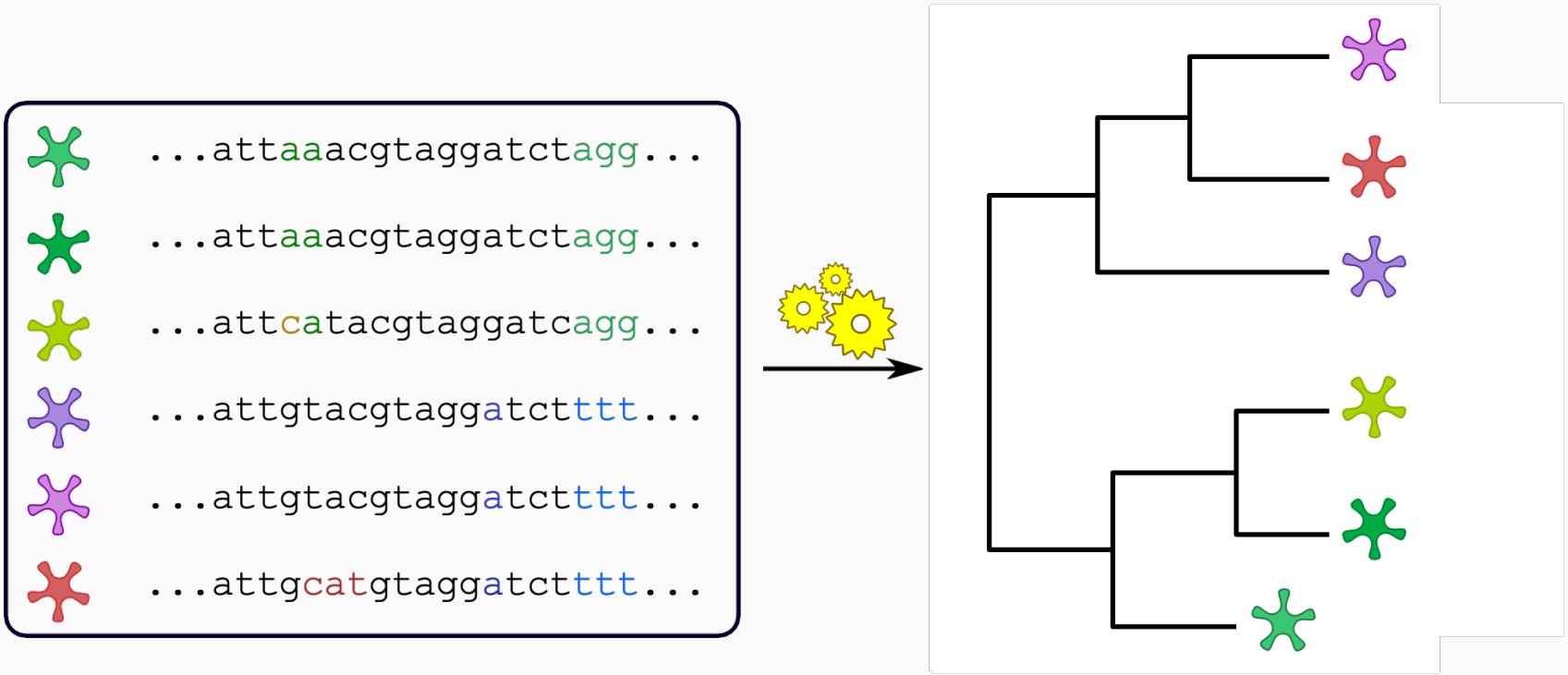


Genetic changes (mutations) accumulate over time

- Base substitution: replacement of a nucleotide (e.g. a→t)
- Insertions: extra nucleotide added (e.g. at →agt)
- Deletions: omission of a nucleotide during replication (e.g. agt →at)

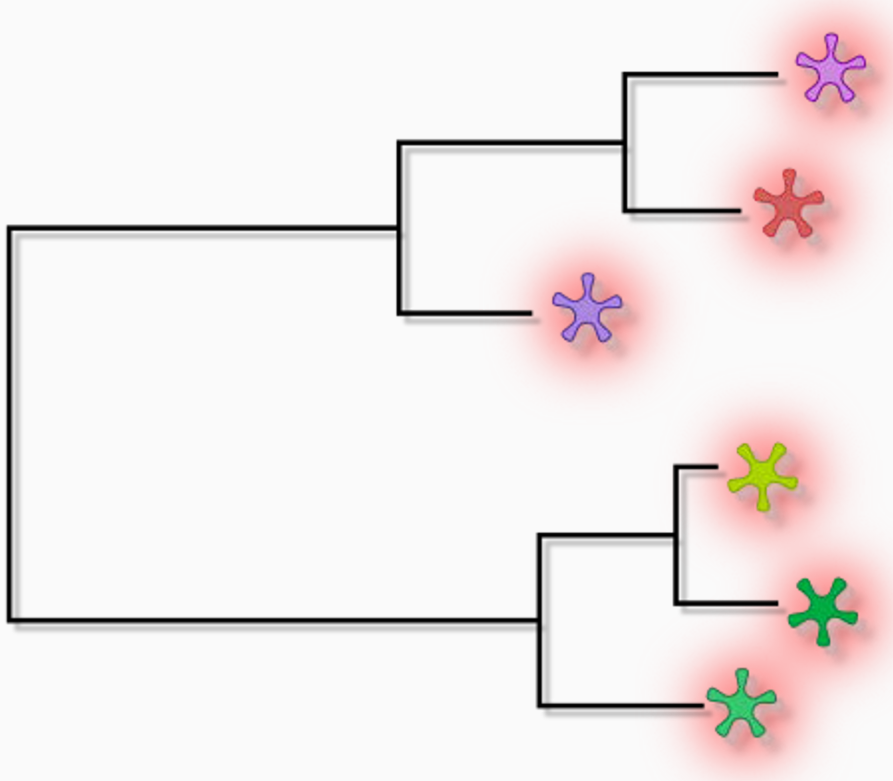


Using substitution patterns to reconstruct the evolutionary history



# Interpreting a phylogeny

# Features of a phylogeny

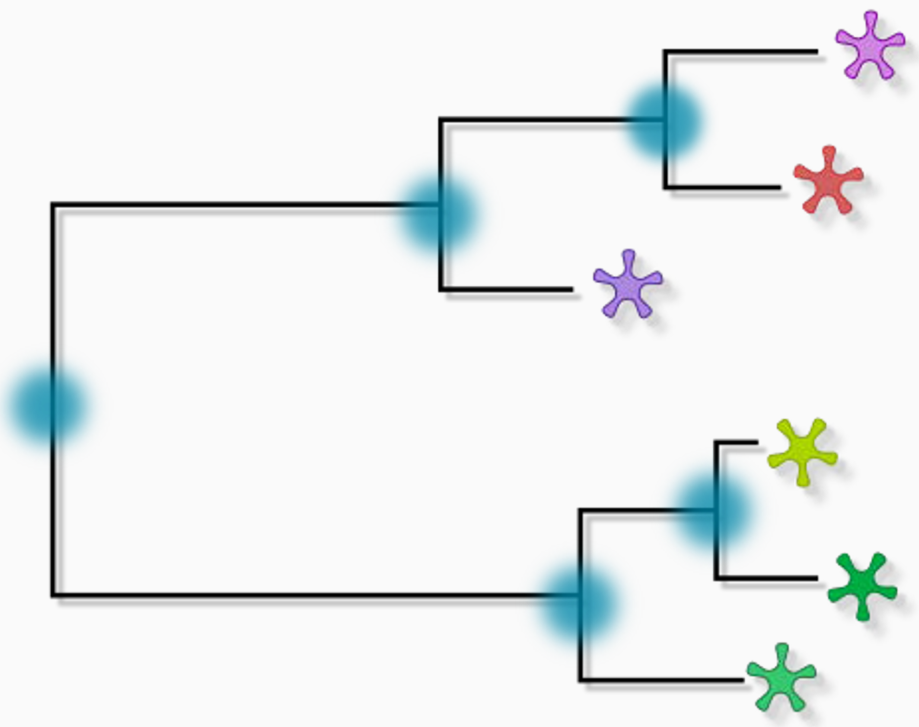


analysed sequences

the *tips* of the tree

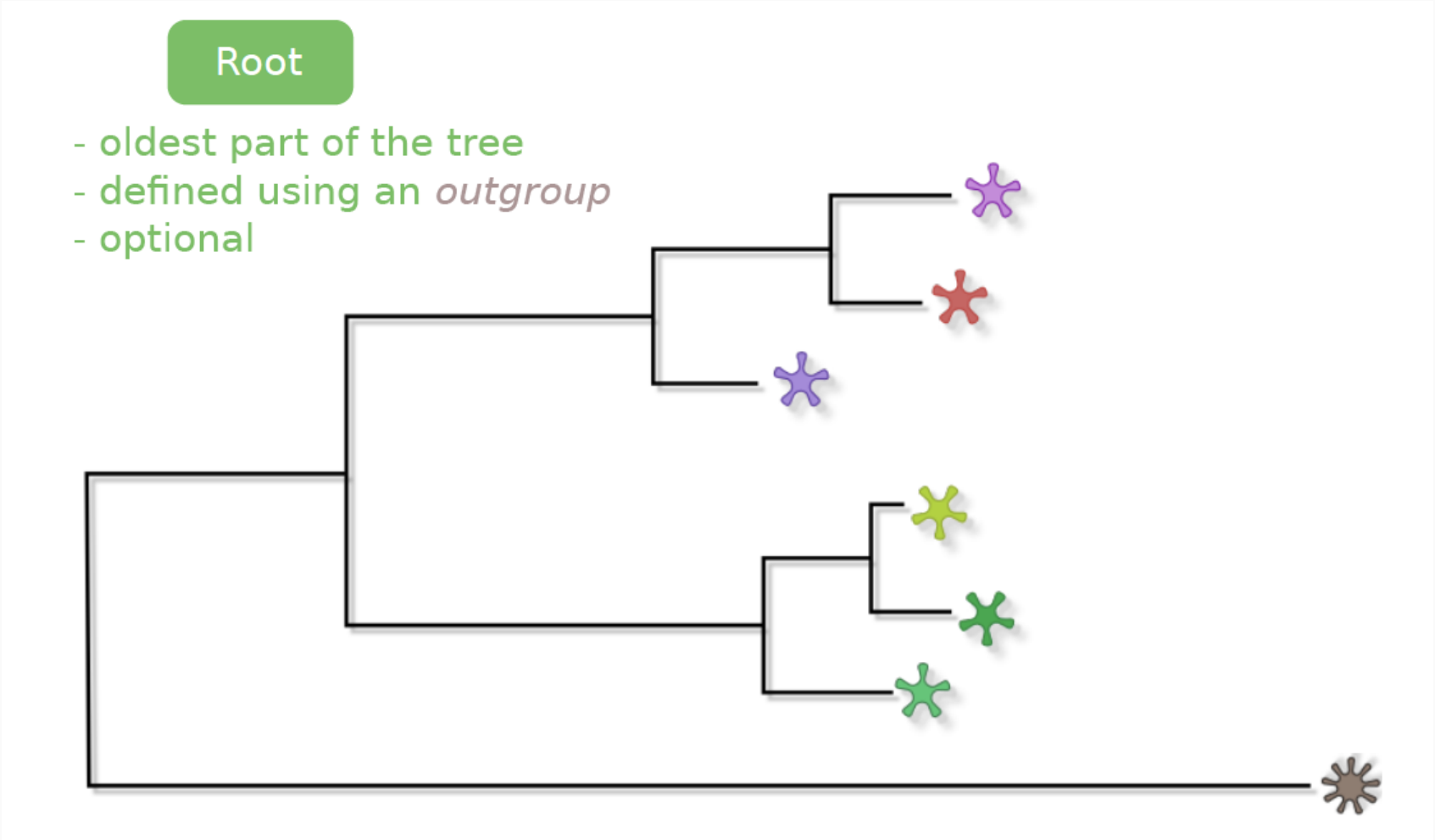
# Features of a phylogeny

**Most Recent Common Ancestors (MRCA)**

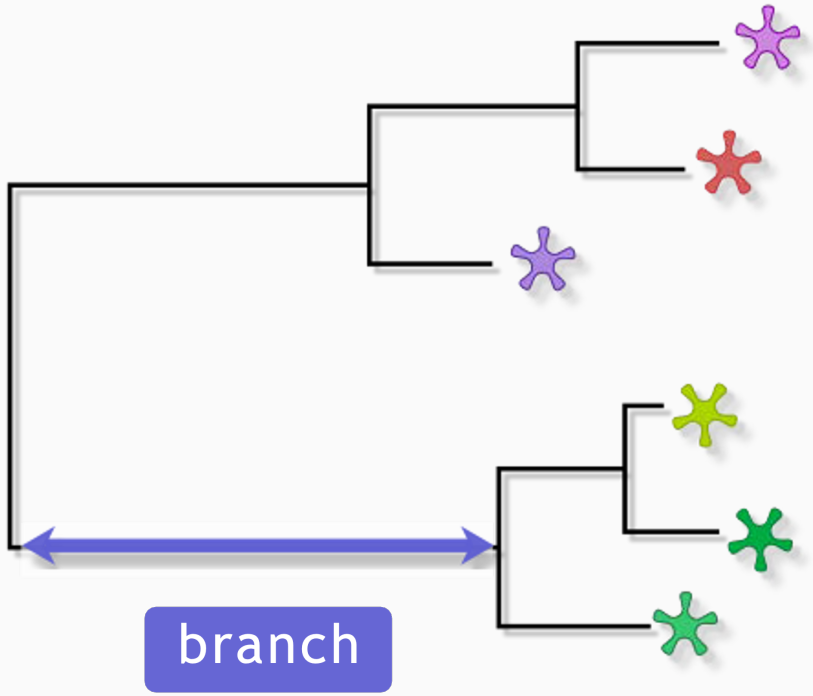


the *nodes* of the tree

# Features of a phylogeny



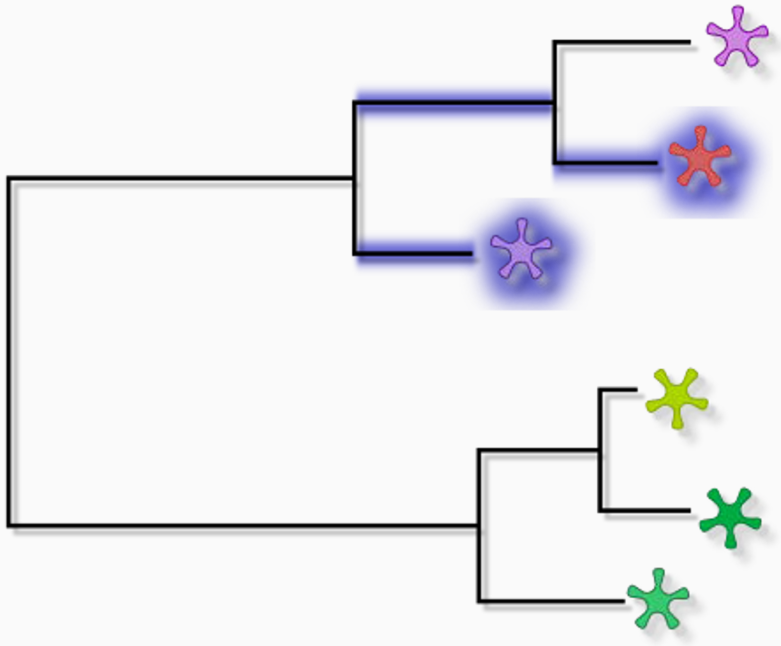
# Features of a phylogeny




- length = amount of evolution (**not time**, as a rule)

# Features of a phylogeny

Small distance ⇒ small amount of evolution, from which we infer epidemiological proximity of pathogens



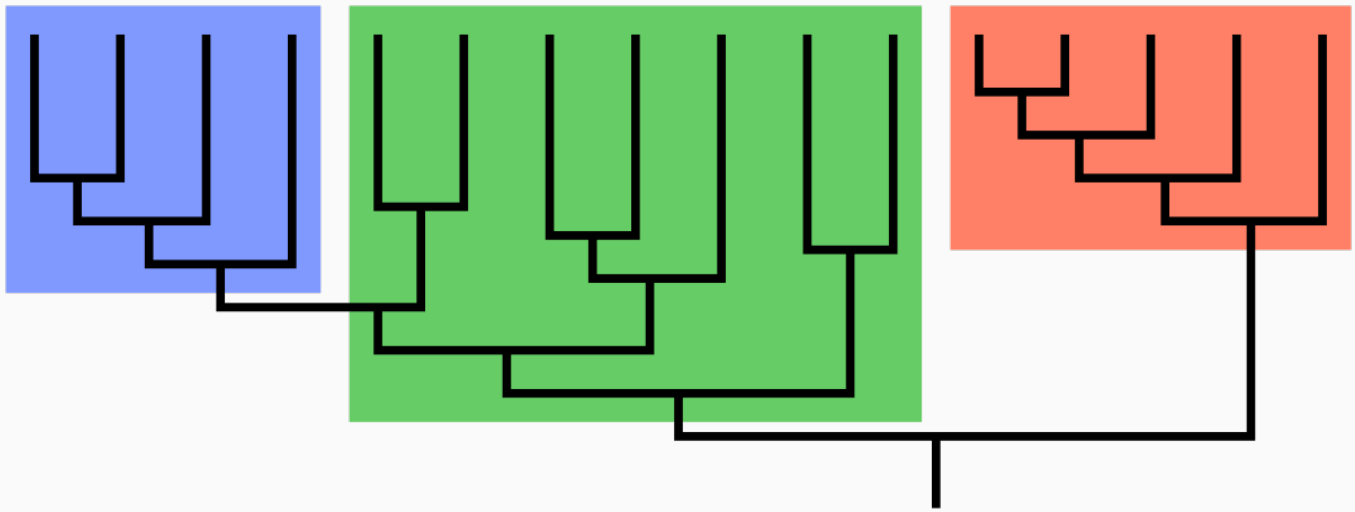
**distances between tips**

- "*patristic*" distance: sum of branch lengths

# Features of a phylogeny

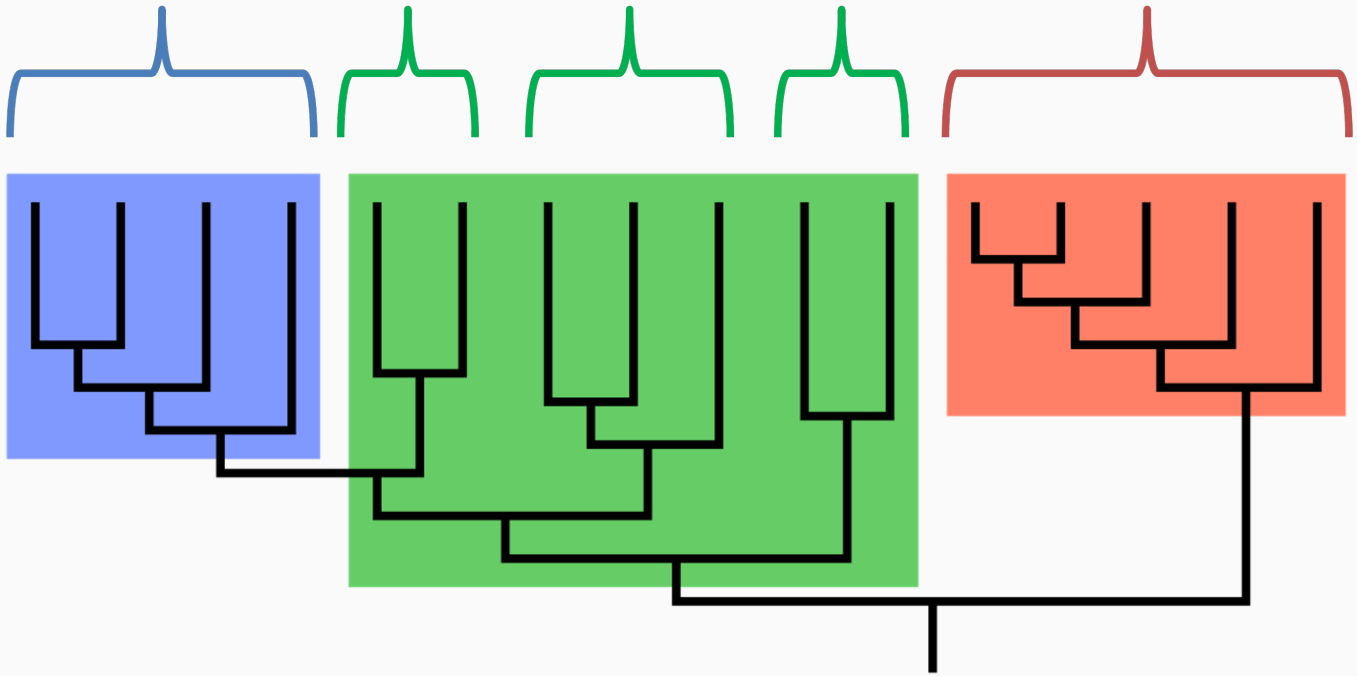
**Clade**: A common ancestor and all its descendents (monophyletic group)

Which of these groups are clades?



# Features of a phylogeny

**Clade**: A common ancestor and all its descendents (monophyletic group)



# Inferring a phylogeny

Phylogenetic algorithms infer the phylogeny using an assumed substitution model

Main approaches:

- Distance-based e.g. neighbour-joining
- Maximum parsimony
- Maximum-likelihood
- Bayesian

# Substitution models for phylogenetic inference

Parametric approaches are based on molecular models of how the pathogen accumulates substitutions (e.g. the relative rate of A→C versus T→G, etc.)

Non-exhaustive examples:

- Jukes and Cantor
  - equal base frequencies
  - equal rates of mutation
  - one parameter (substitution rate)
- Kimura and Nei
  - relaxes second assumption to reflect difference in rate of transition and transversion mutations
  - two parameters for rate
- Hasegawa–Kishono–Yano (HKY)
  - nucleotides occur at different frequencies
  - transitions and transversions occur at different rates
- General Time Reversible (GTR)
  - nucleotides occur at different frequencies
  - different rates of substitution for each pair of nucleotides

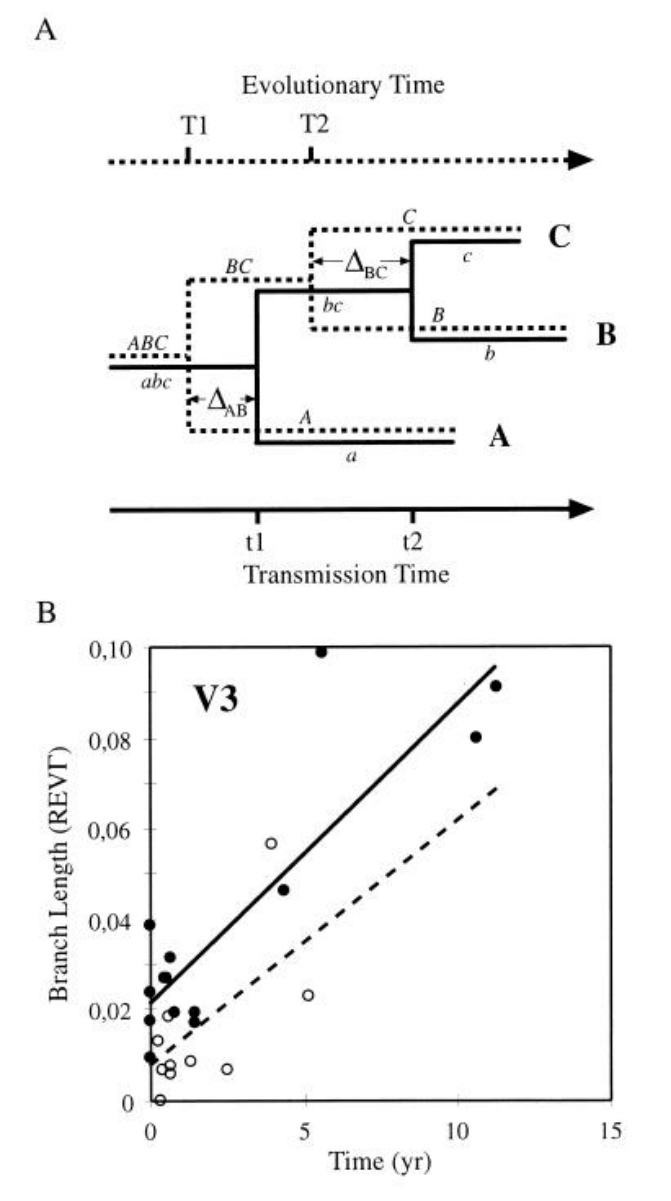
# Interpreting a phylogeny

Tree summary statistics can help characterise topological features and ancestral relationships, e.g.

- Max tree depth
- Maximum tree width
- Phylogenetic diversity
- Node to tip (bifurcation) ratio

# Interpreting a phylogeny

Genetic divergence (subst/site) = evolutionary rate (substs/site/year) × divergence time (years)

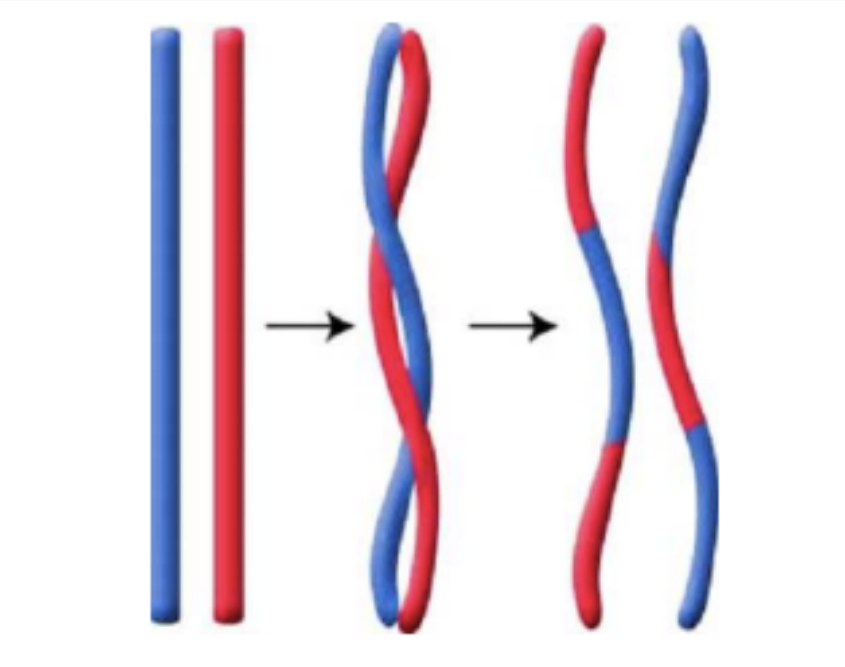



Leitner and Albert (1999) *Proc. Natl. Acad. Sci. USA*

# Processes which can affect phylogenetic inference

Recombination

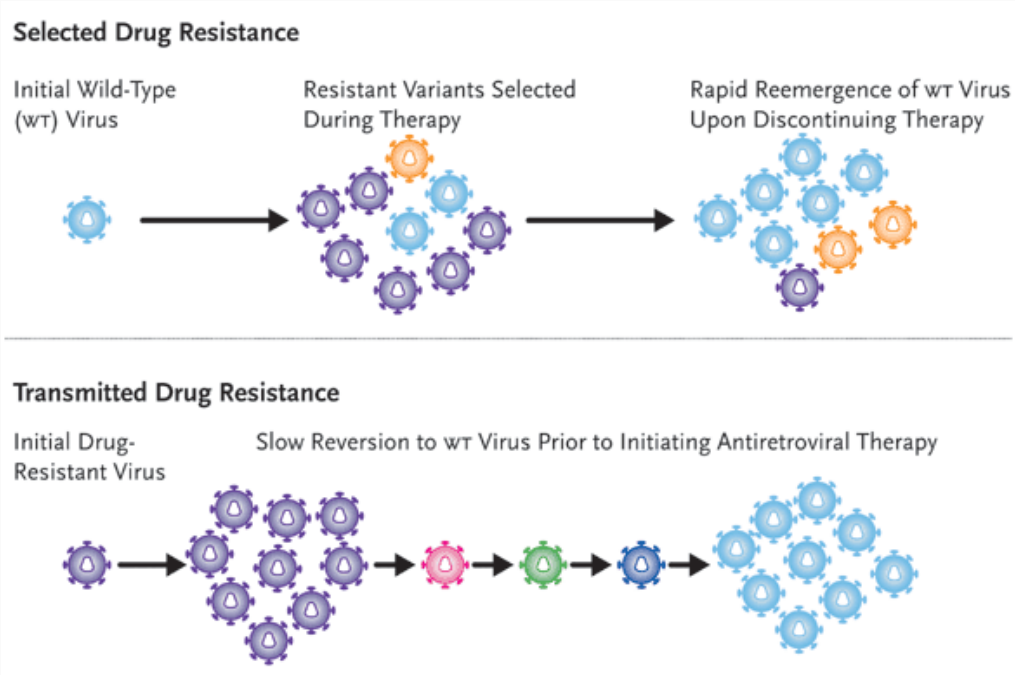
- When genome segments from different viruses are spliced together
- Genetic diversity no longer depends only on evolutionary rates and time



# Processes which can affect phylogenetic inference

Drug resistance

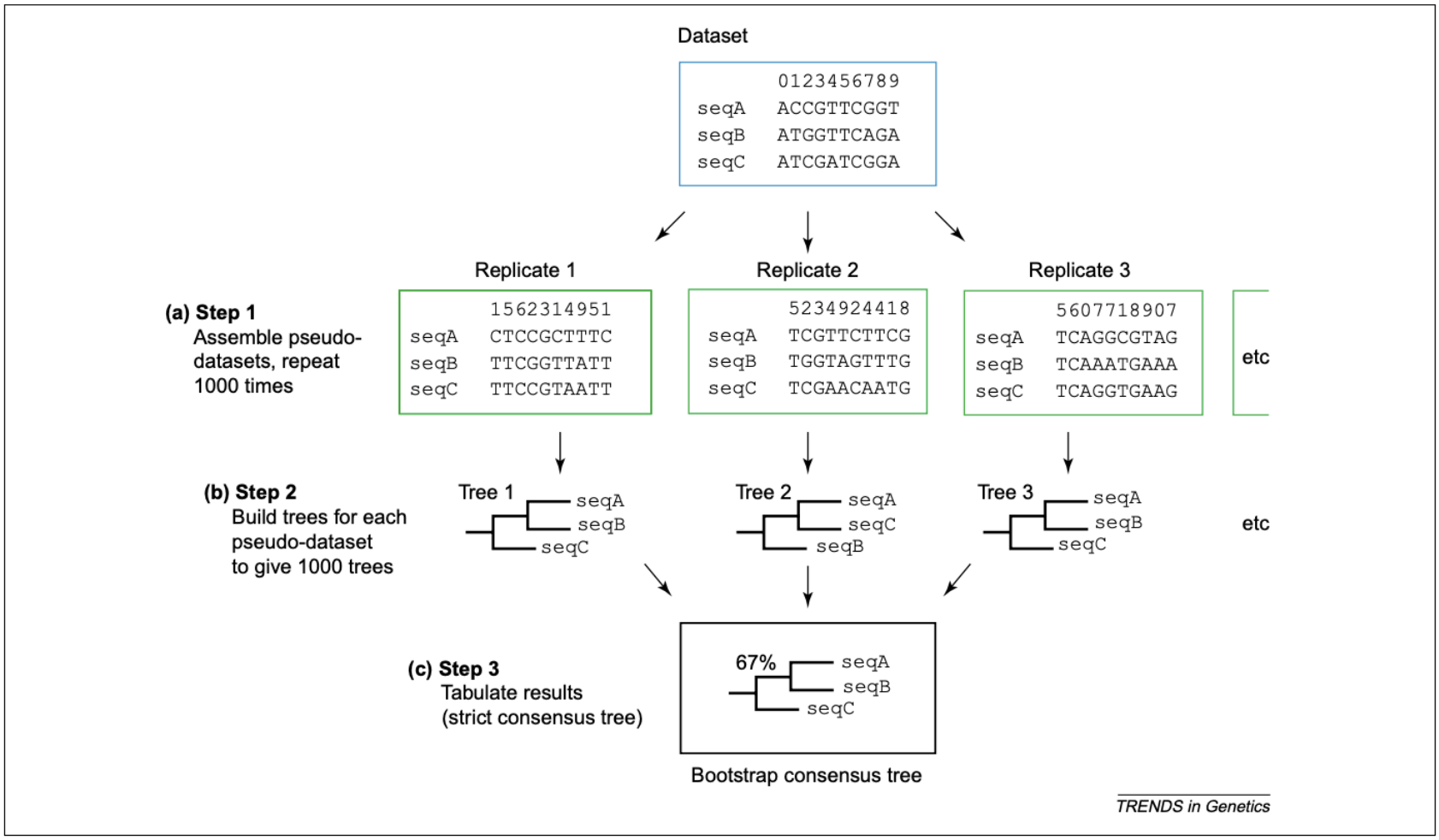
- Drug resistant mutations (DRMs) counteract treatment-mediated inhibition of viral replication
- Are the result of drug-selective pressure, either through treatment or can be transmitted
- Can bias phylogenetic inference




Kuritzkes, 2004

# Quantifying uncertainty

Estimating uncertainty in inferred tree topology via bootstrapping




Baldauf, 2003 *Trends in Genetics*

# Limitations of phylogenetic analyses

- Inferences are uncertain, e.g. in the case of pathogen spread, we are approximating the true unknown transmission tree
- Sampling bias

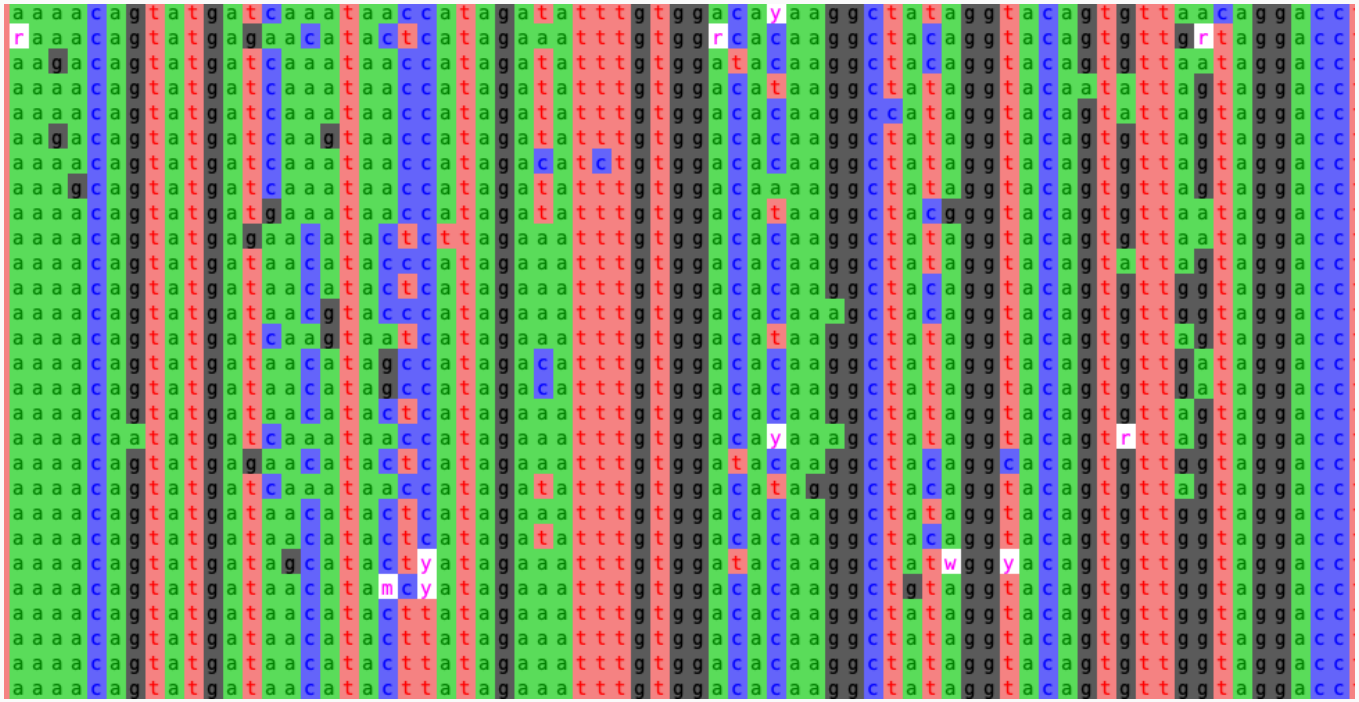
Statistical modelling can help by:

- Incorporating phylogenetic uncertainty
- Accounting for the unsampled population to make population-level inferences
- Amalgamating other data sources (e.g. clinical, mobility, contact data) to understand epidemiological trends

# Running a phylogenetic pipeline

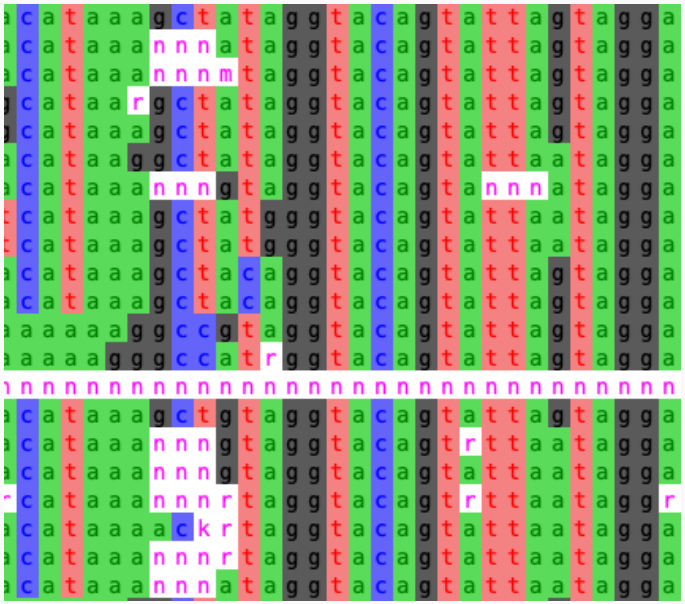
# Sequence alignment

- Align sequences using an algorithm which seeks to minimise some distance measure
- Pairwise vs multiple sequence alignment
- Software: *MAFFT*, *CLUSTALW*, *virulign*



# Mask drug resistant mutations (optional)

Mask known DRMs in the alignment to minimise bias in inferred ancestral relationships in phylogeny from common drug-resistant sites

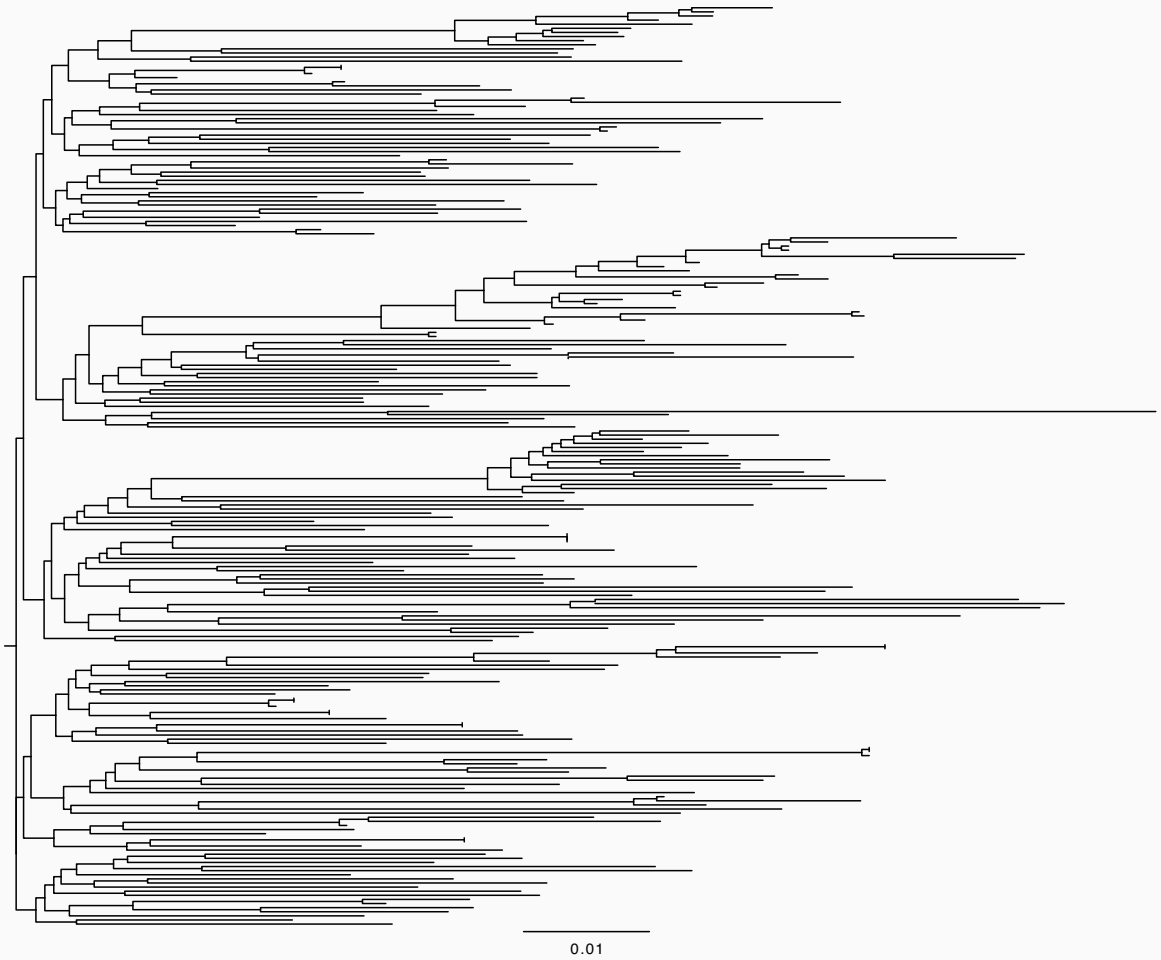


# Build trees

Software: *FastTree, RaxML, IQTree, PhyML, MrBayes, BEAST, ape (R package)*

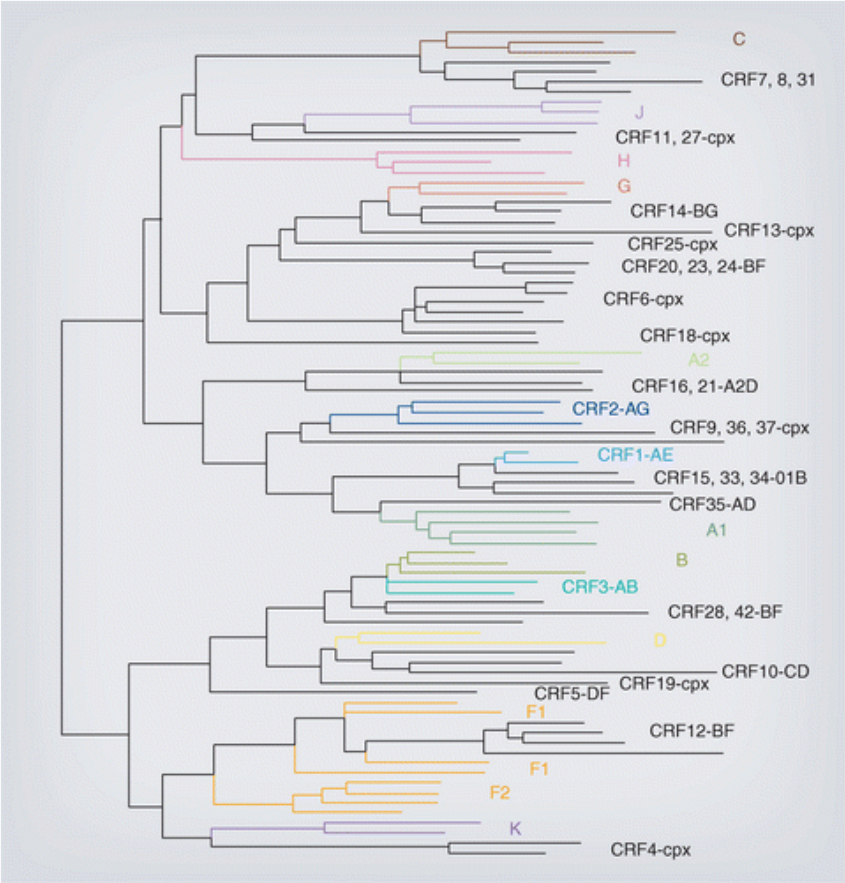
# Check trees (if feasible)

- Visually inspect trees
- Very long branches may indicate a problem with the alignment



# Select outgroups and root tree

Typically use reference sequences from a sequence database (e.g. GenBank)



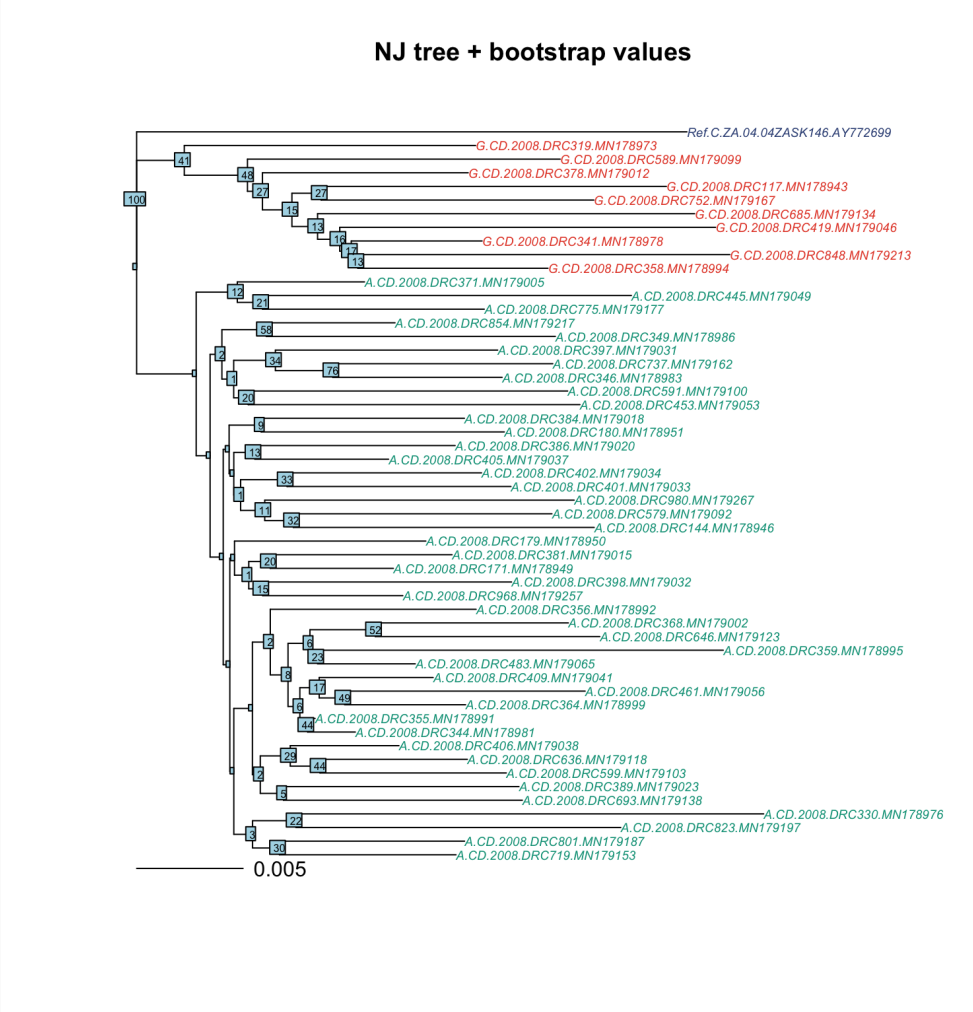

Castro-Nallar et al. (2012). *Future Virology*

# Quantify uncertainty

- Bootstrap resample alignment with replacement
- Build trees
- Label internal nodes of central alignment with bootstrap values

Tools:

- `boot.phylo` in ape package in R
- infer bootstrap trees in parallel using high-performance computing



# Analysing phylogenetic trees

# Analysing phylogenetic trees

- **Cluster analysis → how many, how big, importations**
- Ancestral state reconstruction
  - Characterising epidemiological transmission dynamics → source attribution
  - Phylogeography → spatial transmission dynamics
- Phylodynamics → estimating population-level parameters which shape phylogenies

# Identifying clusters

Distance-based clustering groups taxa with strong evidence of small genetic diversity, through user-specified bootstrap support and patristic distance thresholds

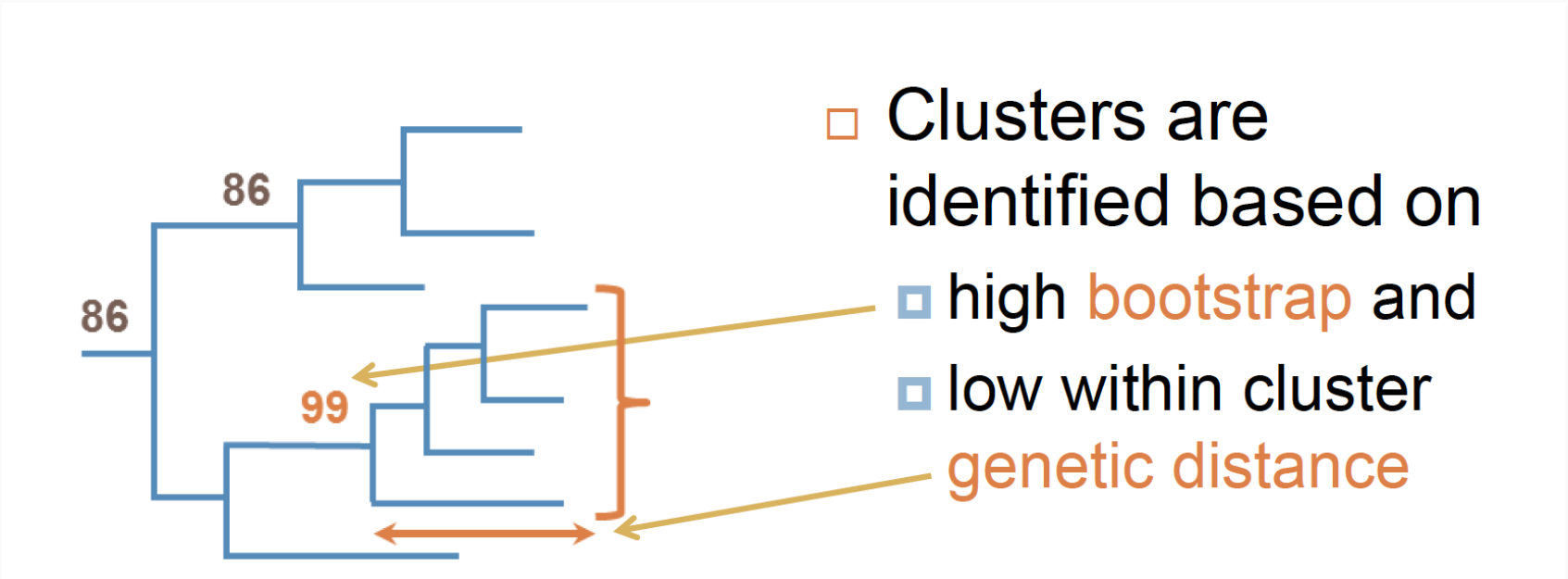


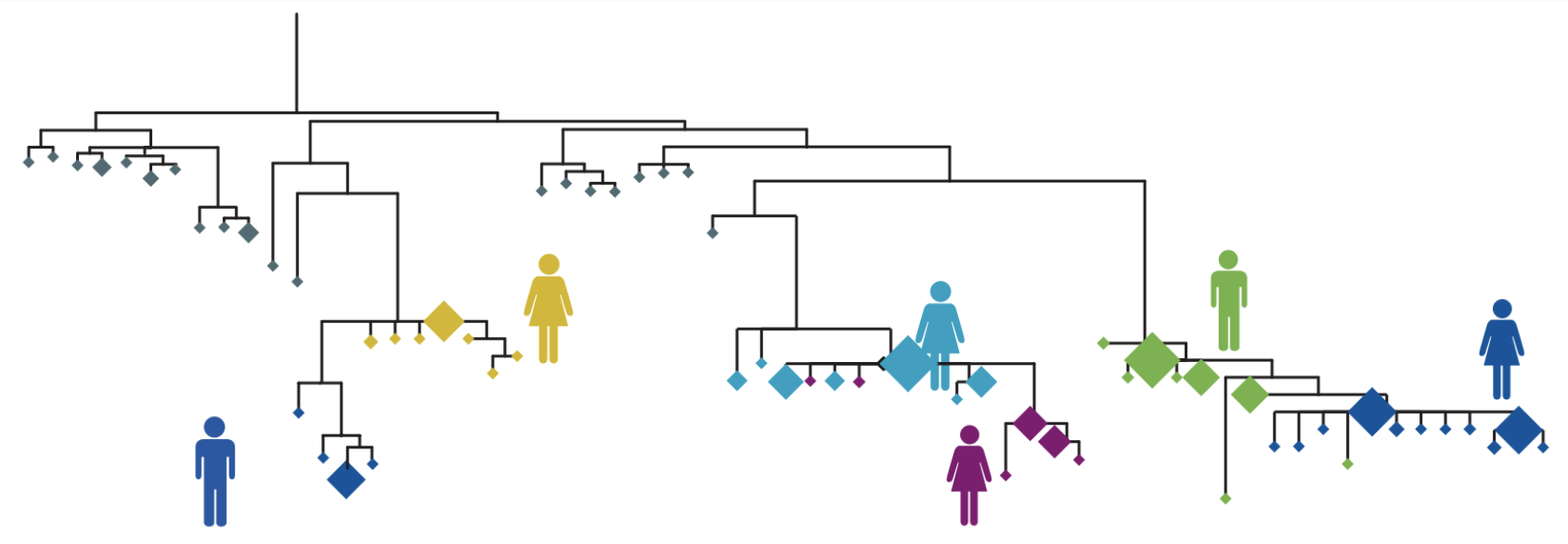

Figure from Cluster Picker (`https://hiv.bio.ed.ac.uk/software.html`)

# Analysing phylogenetic trees

- Cluster analysis $\rightarrow$ how many, how big, importations
- **Ancestral state reconstruction**
  - Characterising epidemiological transmission dynamics $\rightarrow$ source attribution
  - Phylogeography $\rightarrow$ spatial transmission dynamics
- Phylodynamics $\rightarrow$ estimating population-level parameters which shape phylogenies

# Ancestral state reconstruction

Identifying likely transmission pairs for source attribution



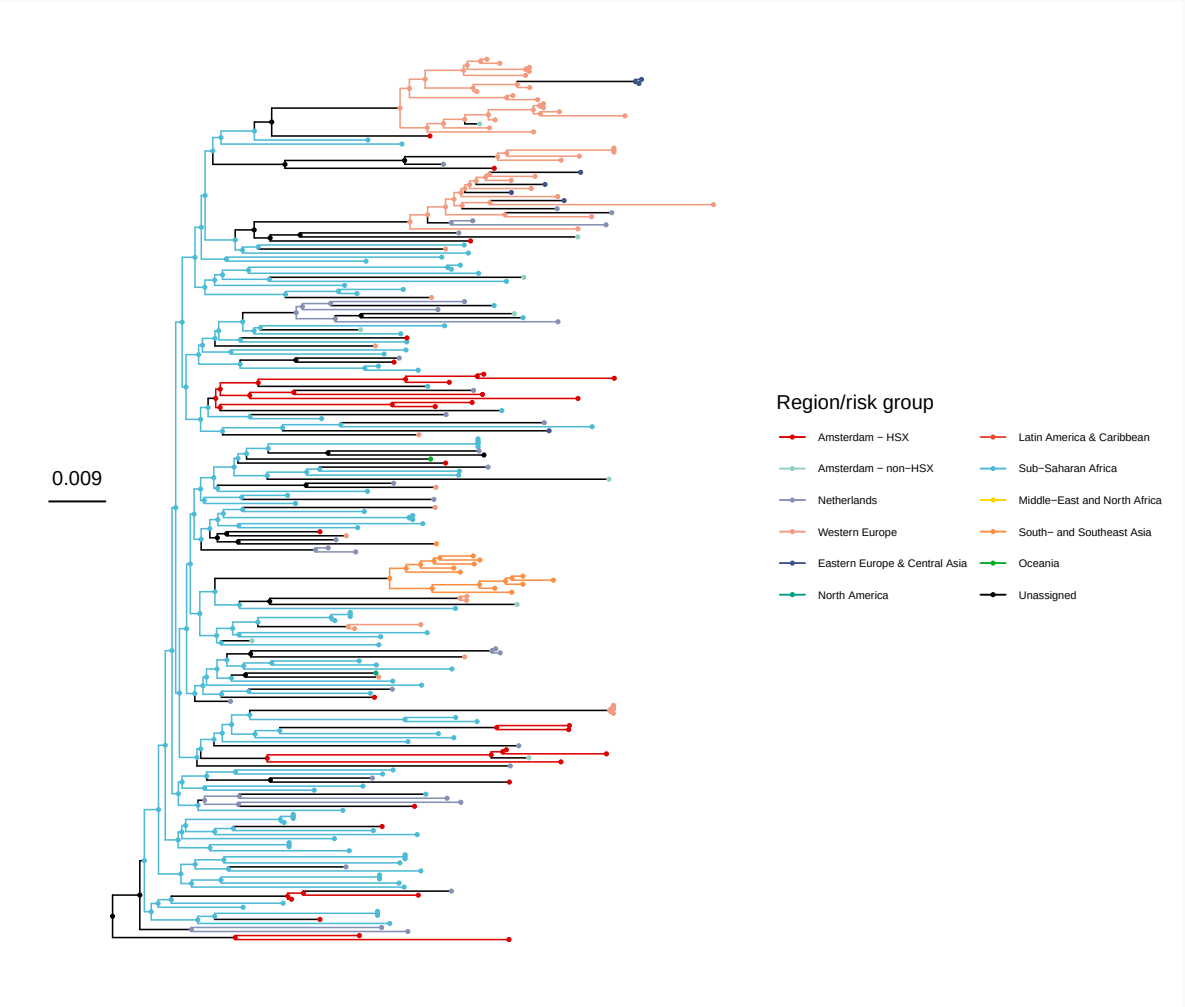
Ratmann et al (2020), *Lancet HIV*

# Ancestral state reconstruction

Separating phylogenetic transmission chains in study population

Software:

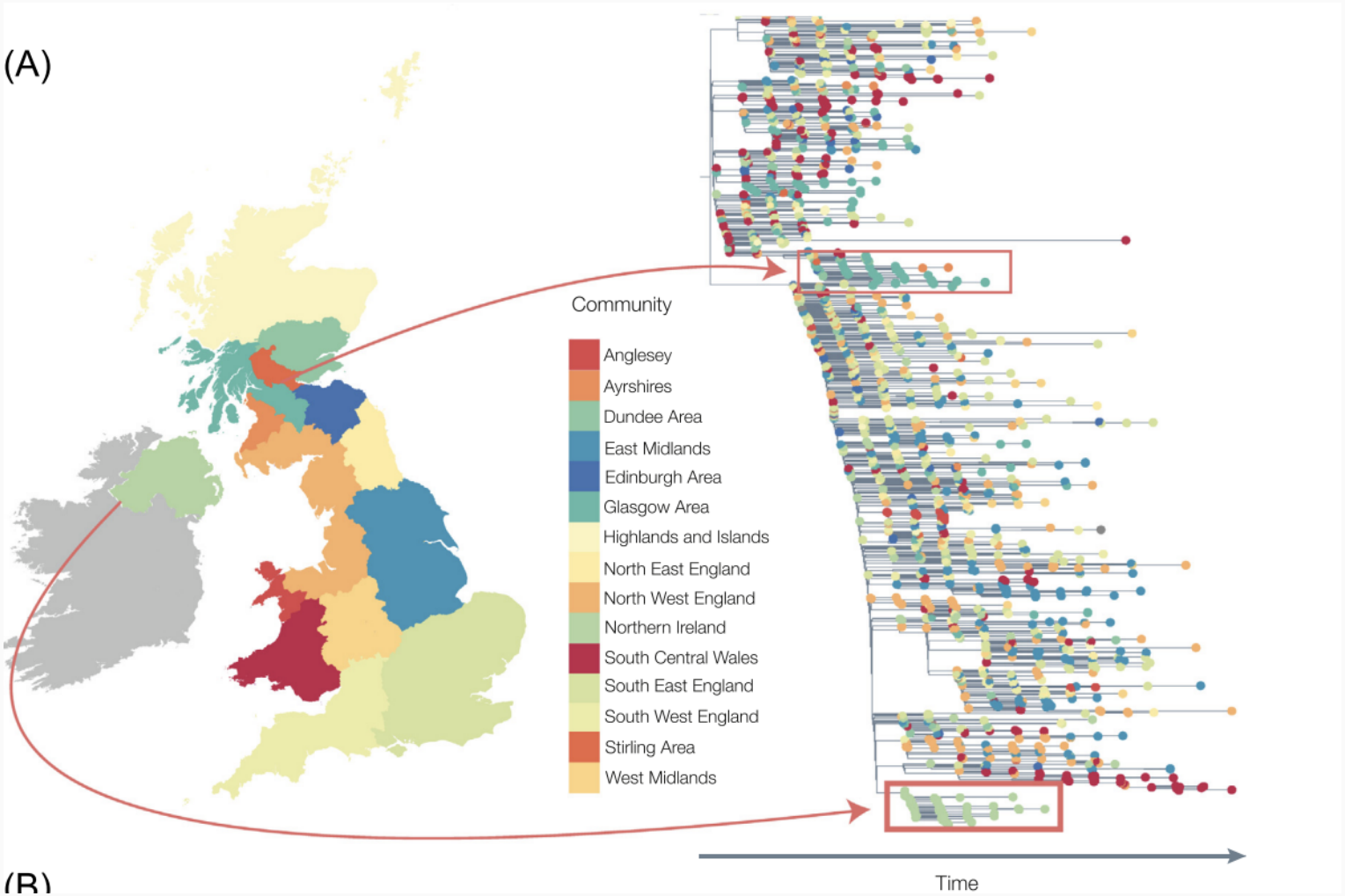
- *Phyloscanner*
- *BEAST*




Blenkinsop et al (2022), *eLife*

# Ancestral state reconstruction

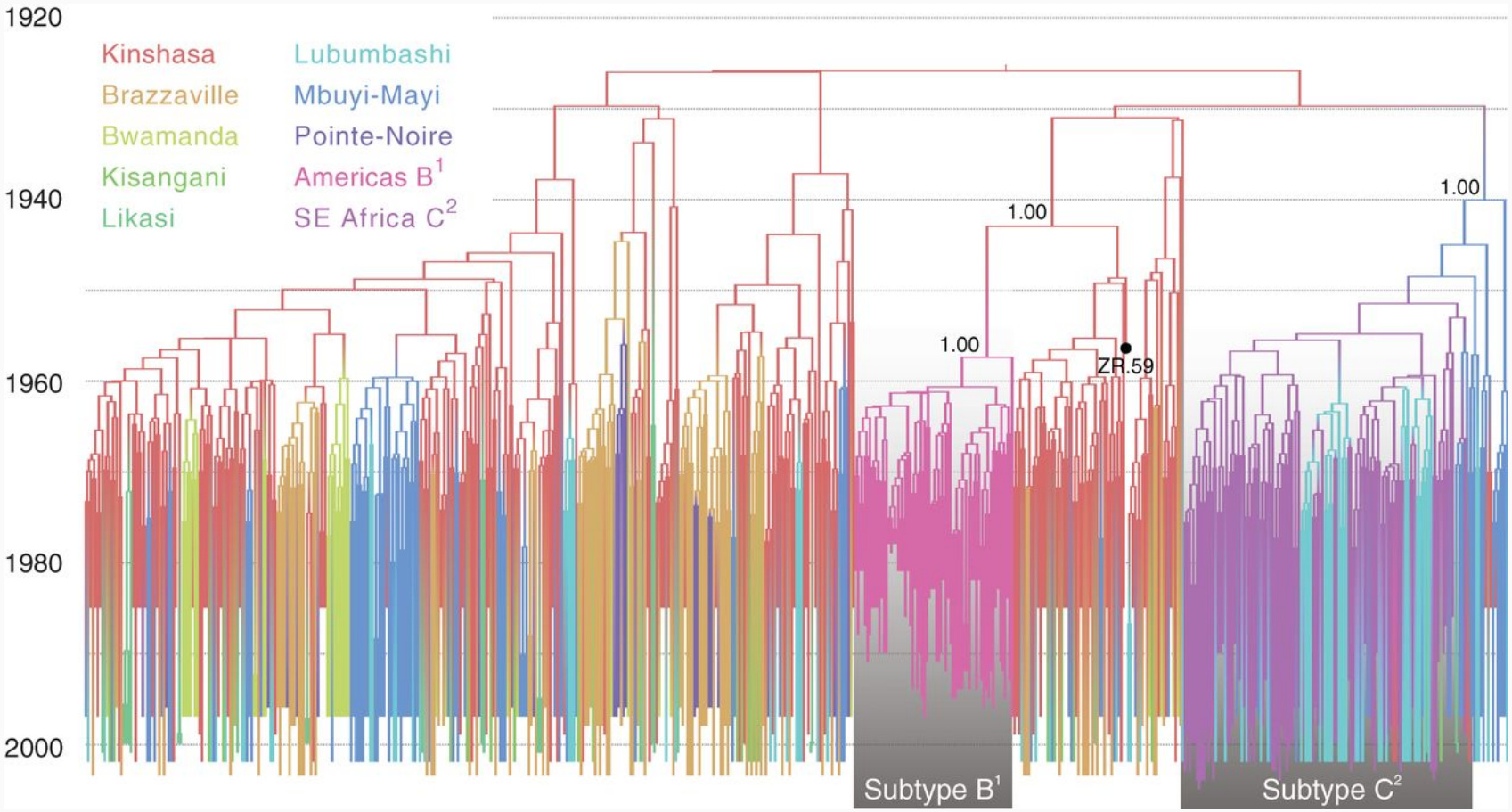
Tracking outbreaks across spatial domains




Hill et al (2021), *Trends in Parasitology*

# Ancestral state reconstruction

Inferring the spatial epidemiological history of a virus



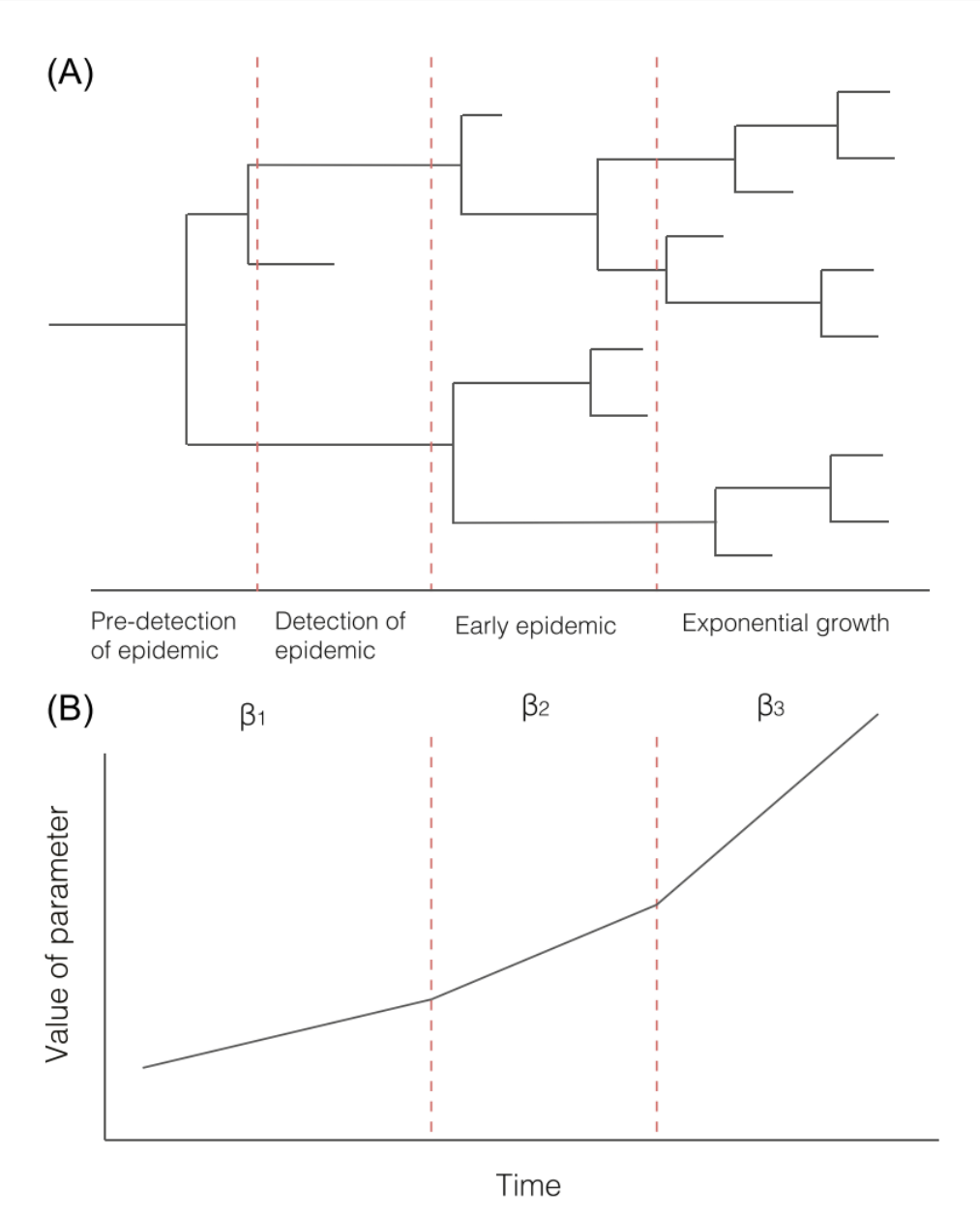

Faria et al. (2014), *Science*

# Analysing phylogenetic trees

- Cluster analysis $\rightarrow$ how many, how big, importations
- Ancestral state reconstruction
  - Characterising epidemiological transmission dynamics $\rightarrow$ source attribution
  - Phylogeography $\rightarrow$ spatial transmission dynamics
- **Phylodynamics $\rightarrow$ estimating population-level parameters which shape phylogenies**

# Viral phylodynamics

Characterising underlying processes which shape viral phylogenies - linking epidemiology with evolutionary dynamics




Hill (2021), *Trends in Parasitology*

# Summary

- Genomic data can provide insights into dynamics of virus spread at different scales
- Phylogenetic trees are a natural way to describe ancestry
- Reconstructing ancestral states helps to understand epidemiological patterns
- Utilising phylogenetic data in statistical epidemiological models enable us to make population-level inferences

# Acknowledgements

Thanks to Chris Wymant, Thibaut Jombart and PANGEA, whose slides I adapted.