Introduction to phylogenetics

Modern Statistics & Machine Learning for population health in Africa

Dr Alexandra Blenkinsop
28/3/2025

Overview

Motivation

What is pathogen phylogenetics?

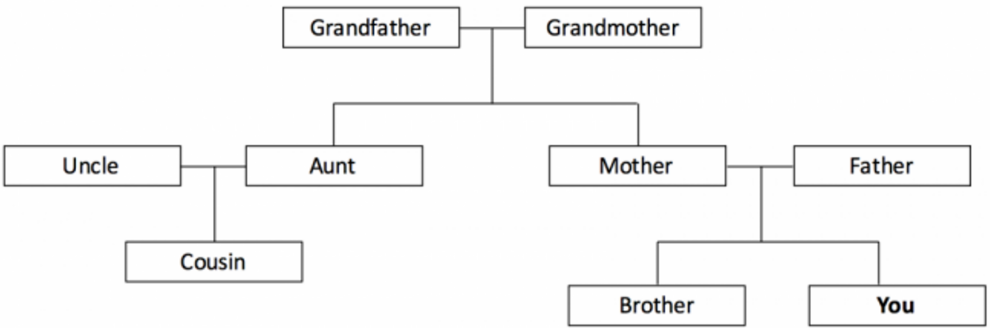
Definition: The study of evolutionary relationships among pathogen populations (e.g. viruses, bacteria, fungi).

Aim: Reconstruct a phylogenetic tree from genomic sequence data to understand how pathogens spread among a population.

Applications: HIV, Ebola, MPOX, COVID-19, Anti-microbial resistance,...

Phylogenetic trees

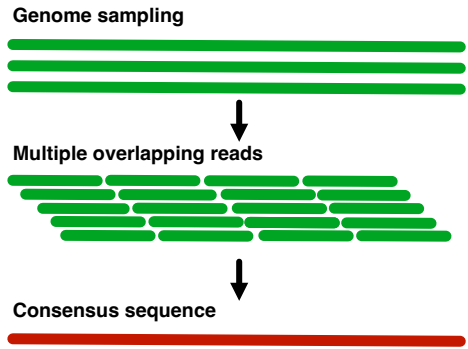
Think about a family tree - how does the topology correspond to how closely related you are to your family members?



Observed data
Pathogen sequences

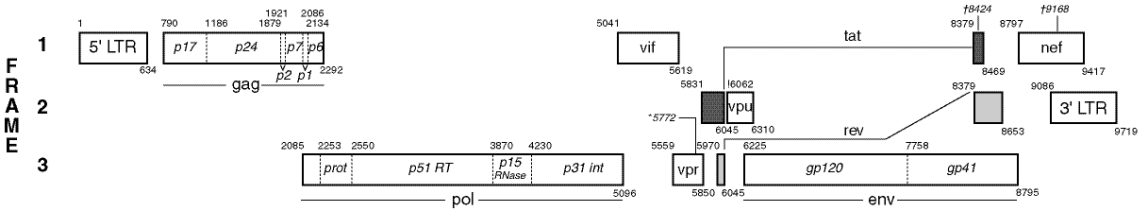
Nucleotides are the building blocks of nucleic acids - chains of these (sequences) encode information in DNA/RNA (adenine (A), cytosine (C), guanine (G) and thymine (T))

- Sanger sequencing (first-gen)
  - Long reads
  - Low throughput
  - Generates one sequence per sample (consensus)
- Next generation sequencing (NGS)
  - Fragment and sequence short reads
  - High throughout - sequence thousands of reads simultaneously
  - fast
  - Can be used for inferring direction of transmission
  - Robustness of conclusions



Observed data
Pathogen sequences

- Whole genome sequence covers all nucleotide positions
- Partial genome sequence covers part of the genome (e.g. *gag*, *pol*, *env* for HIV)
  - cheaper and more scalable
  - gene-specific signals
  - low quality samples

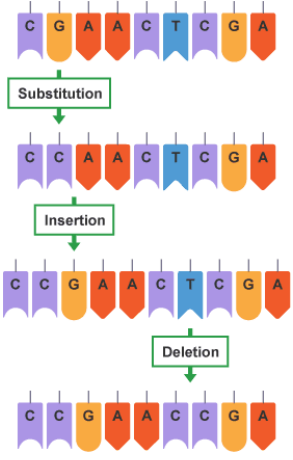


[1] https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html

The evolutionary process

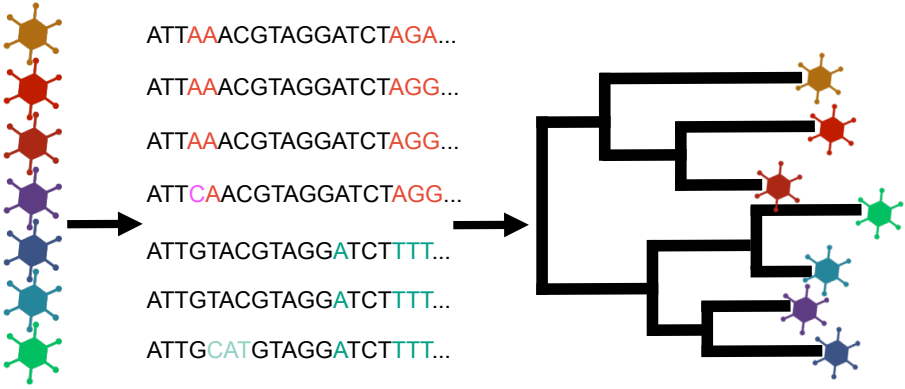
Genetic changes (mutations) accumulate over time due to evolutionary forces

- 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)



Phylogenetic inference

Using substitution patterns to reconstruct the evolutionary history



Phylogenetics for understanding epidemics

How to target interventions in an epidemic to optimise impact?

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

Examples:

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

Phylogenetics for understanding epidemics

Genomic surveillance in Africa

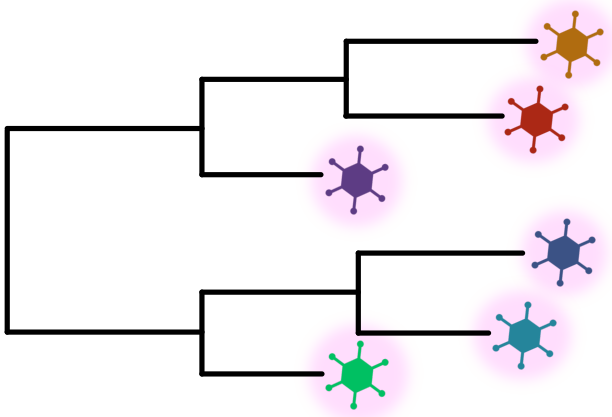
Africa Pathogen Genomics Surveillance Network

- Founded by the Africa Centres for Disease Control and Prevention and the World Health Organization in 2020
- A continental network to accelerate SARS-CoV-2 genomic sequencing and support other priority pathogens in Africa

Interpreting a phylogeny

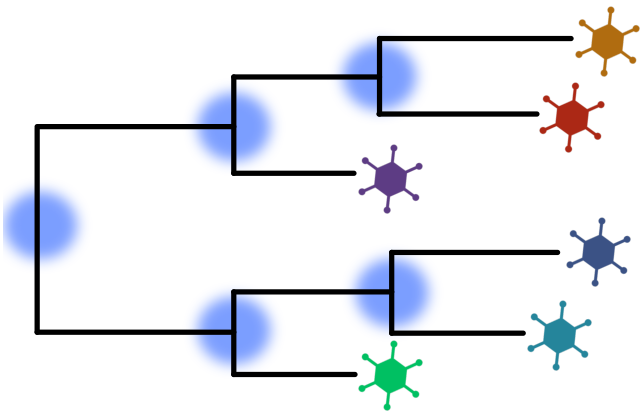
Features of a phylogeny
Tips of the tree

The analysed sequences



Features of a phylogeny
Most recent common ancestor (MRCA)

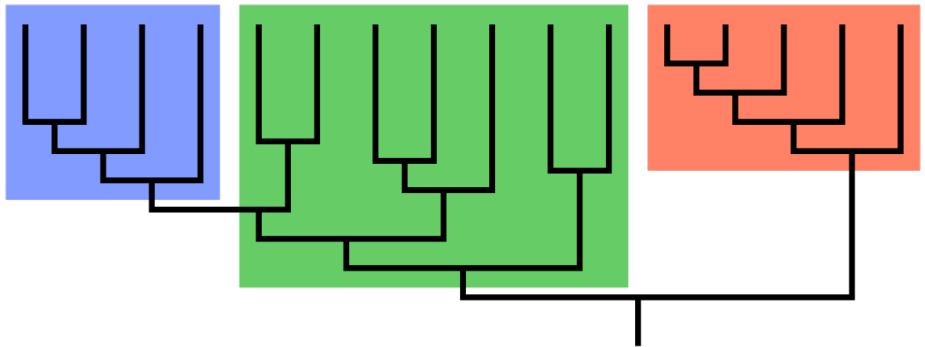
Internal node of the tree ancestral to a subset of taxa



Features of a phylogeny
Clade

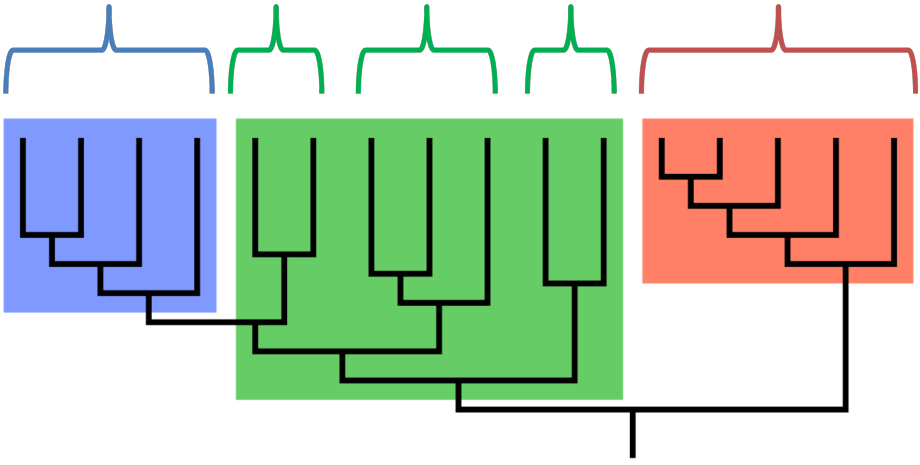
A common ancestor and all its descendents (monophyletic group)

Which of these groups are clades?



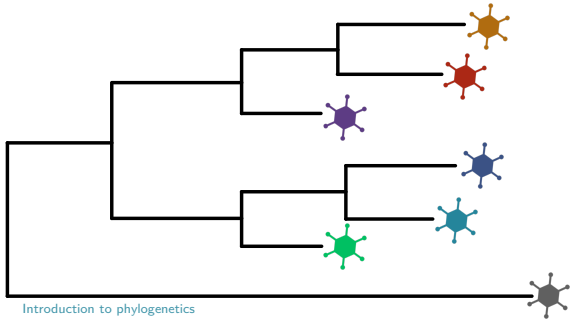
Features of a phylogeny

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



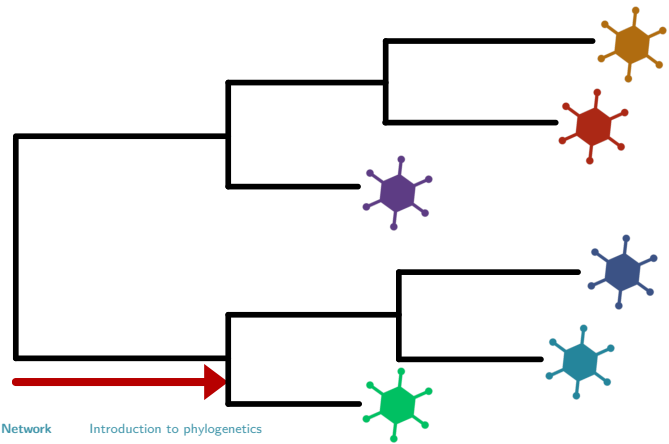
Features of a phylogeny
Root of the tree

- MRCA of all taxa
- Defined using an outgroup (reference sequence which is outside of the clade containing all population sequences)
- Gives directionality to evolution in the tree
- Optional



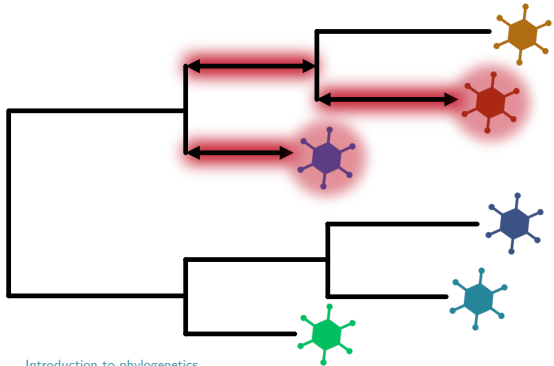
Features of a phylogeny
Branch

Branch length = amount of evolution (not time, in general)



Features of a phylogeny
Genetic distance

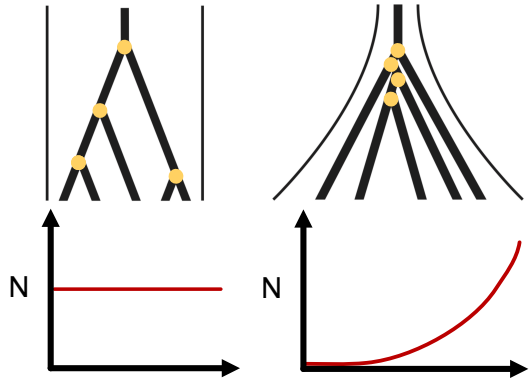
- Genetic distance represents the genetic similarity between taxa.
- Small distance $\Rightarrow$ small amount of evolution, from which we infer epidemiological proximity of pathogens



Coalescent theory

Coalescent theory

- A model which describes ancestral relationships between samples from a population
- **Key idea**: looking **backwards** in time
- **Motivation**: Allows us to focus on what we know (sampled nodes)
- Estimate time of evolutionary events depending on assumptions of underlying mathematical model (speed of pathogen spread)
- Do the data show signs of population structure, recombination, migration, age of a mutation?

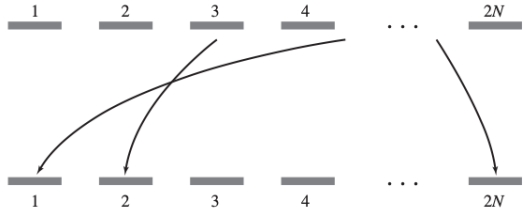


Population with a constant population size (N) and with exponential growth.

Coalescent theory
Wright–Fisher model

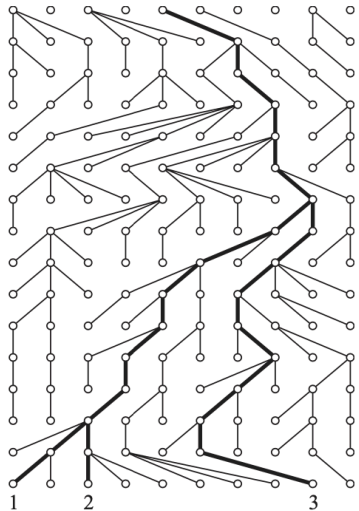
Simple Haploid reproduction model with the following assumptions

- Constant population size
- Population size of 2N genes (2N haploid individuals)
- Discrete, non-overlapping generations
- No positive selection
- Random selection of parents at each new generation (with replacement) with probability $\frac{1}{2N}$



Coalescent theory
Wright–Fisher model

- Population size: 10 genes
- Apply haploid model 16 times from top row
- Rearranging the overlapping lineages reveals tree-like structure



Coalescent theory

The number of descendents in one generation

Let v_i denote the number of descendants of gene i in generation t, $i = 1, 2, ..., 2N$
Probability of coalescence in generation $t = 1/(2\mathrm{N})$

Probability of k descendants

$P(v_i = k) = \binom{2N}{k}(\frac{1}{2N})^k(1 - \frac{1}{2N})^{2N-k}$

Binomial distribution
$\mathrm{Binom}(n, p) = \mathrm{Binom}(2N, \frac{1}{2N})$

Mean

$E[v_i] = np = 2N\frac{1}{2N} = 1$

Variance

$Var[v_i] = np(1 - p) = 2N\frac{1}{2N}(1 - \frac{1}{2N}) = 1$

Covariance of descendants from two genes i and j

$Cov(v_i, v_j) = E[v_i, v_j] - E[v_i]E[v_j] = -\frac{1}{2N}$

Correlation coefficient

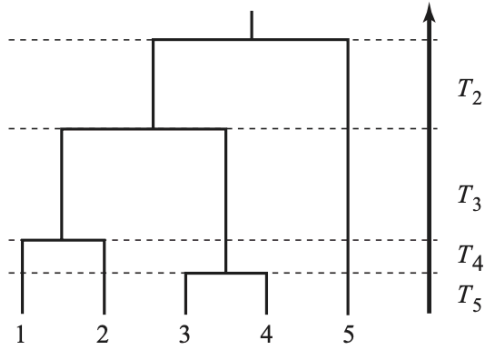
$Cor(v_i, v_j) = \frac{Cov(v_i, v_j)}{\sqrt{Var[v_i]Var[v_j]}} = -\frac{1}{2N-1}$

If 2N is large then $v_i \dot{\sim} Po(1)$

Coalescent theory
Discrete time coalescent

Shape and topology of the phylogenetic tree reflects historical population dynamics and structure.
Lineages coalesce more quickly among small populations.
For fixed mutation rate, more diverse populations have,

- longer coalescence times
- larger effective population sizes



Discrete time tree

Coalescent theory
Discrete time coalescent

Probability 2 genes coalesced T generations back

$$\Pr(\mathrm{T}{=}1) = \frac{1}{2N}$$
$$\Pr(\mathrm{T}{=}2) = \left(1 - \frac{1}{2N}\right)\frac{1}{2N}$$
$$\vdots$$
$$P(T = j) = \left(1 - \frac{1}{2N}\right)^{j-1}\frac{1}{2N}$$

Therefore, the coalescence time T for two genes to find a MRCA is distributed as,

$$T \sim \text{Geometric}\left(\frac{1}{2N}\right)$$

$$E[T] = \frac{1}{1/(2N)} = 2N$$

Expected time until MRCA = the number of genes in the population

Coalescent theory
Discrete time coalescent

Generalise: Probability n sampled genes coalesced T generations back, for $k < n$.
Probability k genes have k different ancestors in previous generation:

$$\frac{(2N-1)}{2N}\frac{(2N-2)}{2N}\cdots\frac{(2N-k+1)}{2N} = \prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right)$$
$$= 1-\sum_{i=1}^{k-1}\frac{j}{2N}+O\left(\frac{1}{N^2}\right)$$
$$= 1-\binom{k}{2}\frac{1}{2N}+O\left(\frac{1}{N^2}\right)$$

Probability of coalescence event in a single generation,

$$\binom{k}{2}\frac{1}{2N}$$

Probability that two out of k genes have a common ancestor $T_k = j$ generations ago $(j = 1, 2, \ldots)$ is,

$$P(T_k = j) = \left(1-\binom{k}{2}\frac{1}{2N}\right)^{j-1}\binom{k}{2}\frac{1}{2N}$$

So $T_k \dot{\sim}$ Geometric$(\binom{k}{2}/2N)$.

Kingman's n-coalescent
Continuous time tree

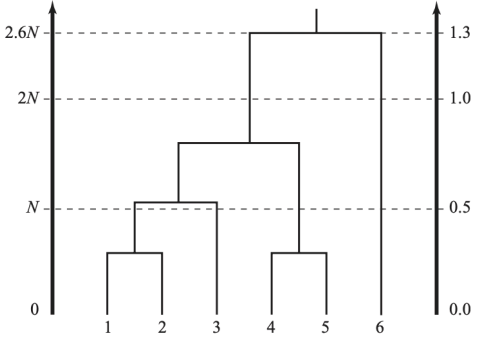
Expected time until the next coalescent event:

$$E[T_k] = \frac{2N}{\binom{k}{2}}$$

When sampling size k is small relative to population size N, as $N \to \infty$, coalescence events become rare and process converges to a continuous-time Markov chain.
If T_k^c is the waiting time for k genes to have $k-1$ ancestors,

$$T_k^c \sim \exp\left(\binom{k}{2}\right)$$

$$\mathrm{P}(T_k^c \leq t) = 1 - e^{\binom{k}{2}t}$$



Tree statistics

Height of a tree

Tree of sample size n.
Height H_n = sum of epochs, T_j, for $j = n, n-1, n-2, ..., 2$ ancestors

Distribution of H_n

$P(H_n \leq t) = \sum_{k=1}^{n} e^{-\binom{k}{2} t} \frac{(-1)^{k-1}(2k-1)n_{[k]}}{n_{(k)}}$

where $n_{[k]} = n(n-1)\cdots(n-k+1)$,and $n_{(k)} = n(n+1)\cdots(n+k-1)$.

Mean of H_n

$E(H_n) = \sum_{j=2}^{n} E[T_j] = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2(1-\frac{1}{n})$

Variance of H_n

$Var(H_n) = \sum_{j=2}^{n} Var(T_j) = 4\sum_{j=2}^{n} \frac{1}{j^2(j-1)^2}$

Tree statistics

Branch length

Distribution of L_n

$P(L_n \leq t) = (1-e^{-t/2})^{n-1}$

Mean of L_n

Weight coalescent times by number of lineages in that epoch - describes how much history shared by a sample of genes.

$$E(L_n) = \sum_{j=2}^{n} jE(T_j) = 2 \underbrace{\sum_{j=1}^{n-1} \frac{1}{j}}_{\approx log(n)}$$

Variance of L_n

$$Var(L_n) = \sum_{j=2}^{n} j^2 Var(T_j) = \underbrace{4 \sum_{j=1}^{n-1} \frac{1}{j^2}}_{\approx 2\pi^2/3 \text{ as } n \text{ increases}}$$

Effective population size

Limitations of Wright-Fisher model:

- Real-world populations are also determined by certain features (e.g. geography, social constraints), not captured by the model

For a real population, the effective population size, N_e is the population size of the haploid Wright–Fisher that best approximates the real population.

Extensions of Wright-Fisher model which relax constant population size assumption: Skyride model - non-parametric, uses a Gaussian Markov random field (GMRF) smoothing prior to estimate population size trajectories

Phylogenetic inference

Inferring a phylogeny

Popular approaches:

- Distance-based (fast, heuristic, doesn't model evolutionary process)
  - Neighbour-joining algorithm - iterative procedure joining similar sequences together
  - Unweighted Pair Group Method (UPGMA)
- Character-based (optimal but computationally expensive)
  - Maximum parsimony - finds tree with the fewest evolutionary changes
  - Maximum-likelihood - evaluates the probability of the observed data given a tree and a model of sequence evolution ($\mathcal{L}(T, \theta) = P(D \mid T, \theta)$)
  - Bayesian - estimate the posterior probability distribution of trees ($P(T \mid D) \propto P(D \mid T)P(T)$)

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

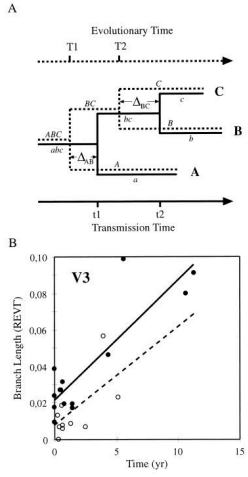
Characterising a phylogeny
Phylogenetic statistics

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

- Max tree depth - most divergent lineage
- Maximum tree width - identifies rapid population expansions
- Sum of all branch lengths - summarise phylogenetic diversity
- Node to tip (bifurcation) ratio - branching structure of tree

Characterising a phylogeny
Phylogenetic statistics

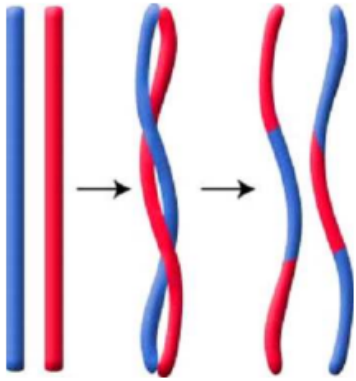
We can estimate the evolutionary rate of a virus, or the timing of infection from pairwise genetic distances and time since coalescence (or an assumed evolutionary rate), through the relationship,
Genetic divergence (subst/site) = evolutionary rate (substs/site/year) $\times$ divergence time (years)



Processes which can affect phylogenetic inference

Recombination

- When genome segments from different viruses are spliced together
- Increases genetic diversity, creating new variants
- Genetic diversity no longer depends only on evolutionary rates and time



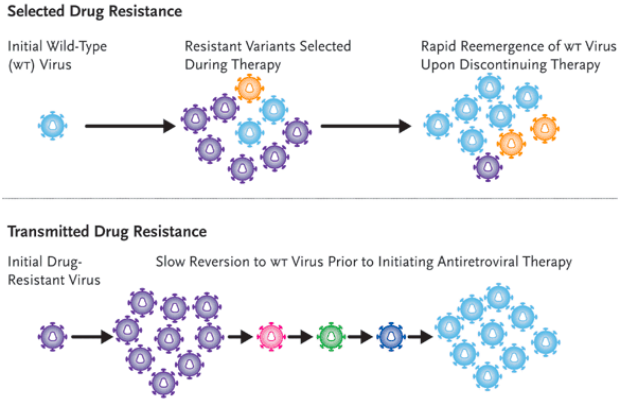
Consequences:

- May result in a tree topology which misleadingly suggests an exponentially growing population
- May over-estimate substitution rate heterogeneity
- May infer a biased molecular clock

Processes which can affect phylogenetic inference

Drug resistance

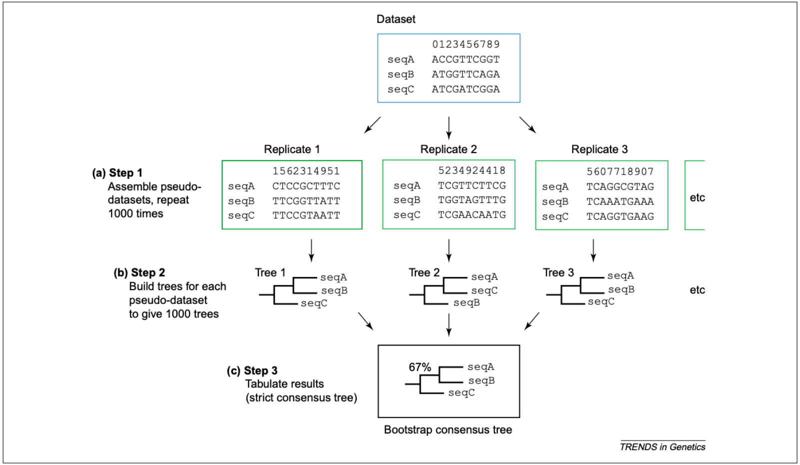
- Drug resistant mutations (DRMs) counteract treatment-mediated inhibition of viral replication
- Are the result of drug-selective pressure, either through treatment or can be transmitted
- Can bias phylogenetic inference through artificial clustering of unrelated strains



Kuritzkes, 2004

Quantifying uncertainty

Estimating uncertainty in inferred tree topology via bootstrapping



[2]

[2] 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

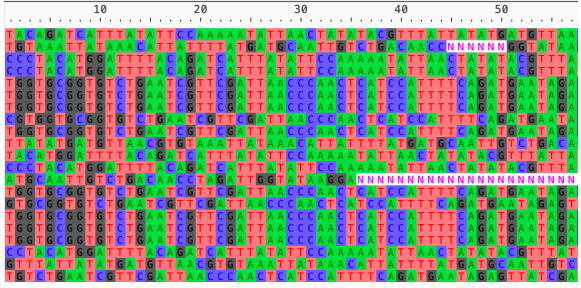
Statistical modelling can help by:

- 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

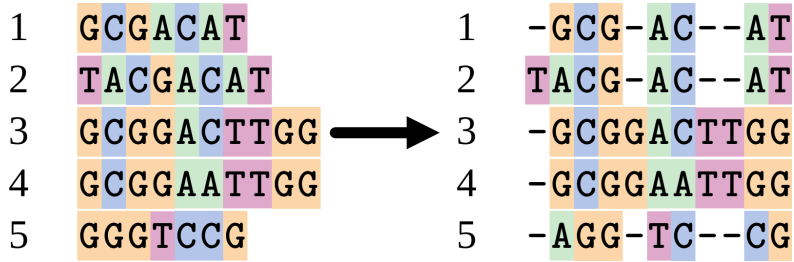
Sequence data

- Usually in .FASTA format
- First line of each sequence starts with > and includes an ID code and other information, e.g.: >TAXA_LABEL ttcctctt–tctcgac-tccatcttcgcggtag—ct
- Tools for visualising: *Aliview*, *Jalview*



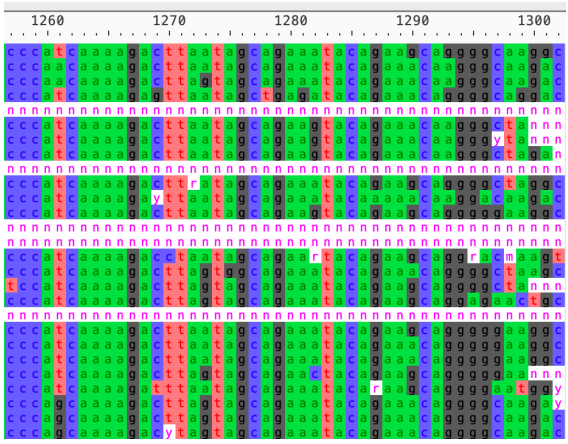
Sequence alignment

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



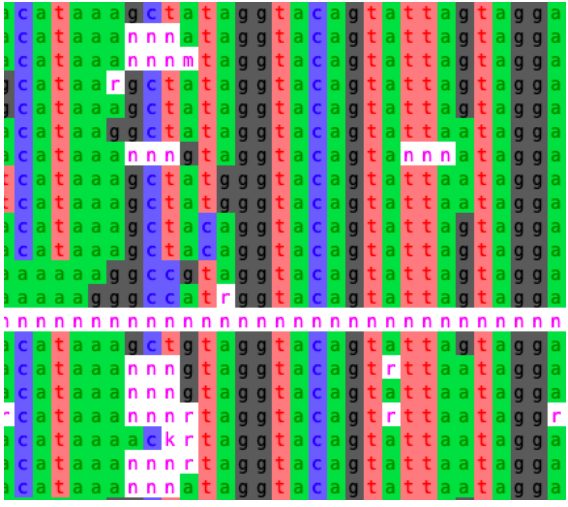
Fill gaps

- Sequences in alignment must all be the same length
- Trim the alignment and fill in any gaps



Mask drug resistant mutations (optional)

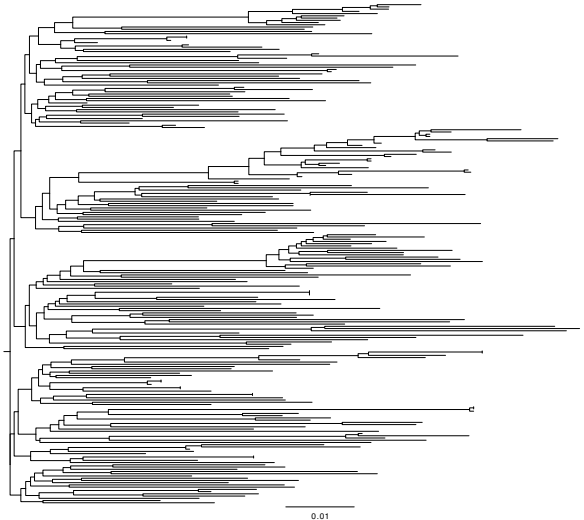
Mask known DRMs in the alignment to minimise bias in inferred ancestral relationships in phylogeny from common drug-resistant sites



Build trees

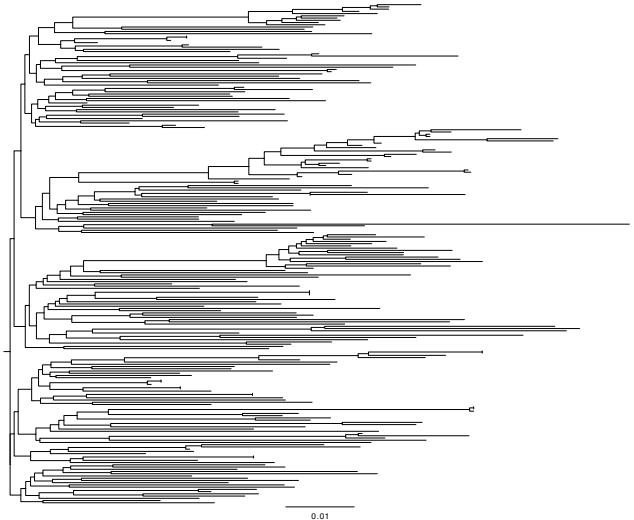
Software:

- *FastTree*
- *RaxML*
- *IQTree*
- *PhyML*
- *MrBayes*
- *BEAST*
- *phylostan*
- *ape (R package)*



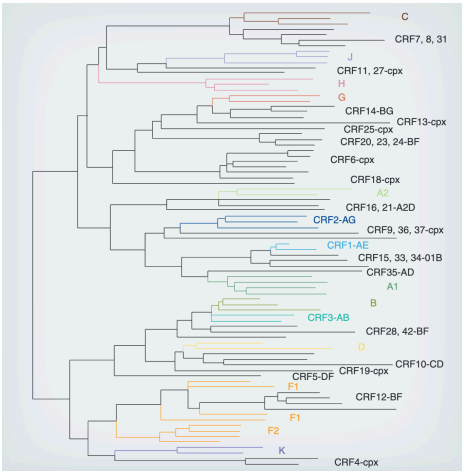
Check trees

- Visually inspect trees (if feasible), or compute summary statistics of branch lengths
- 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)
- Should be outside of study population clade, but not too genetically distant



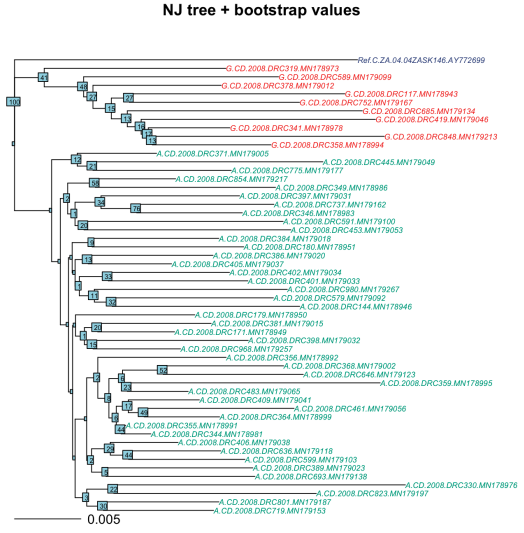
Castro-Nallar et al. (2012). *Future Virology*

Quantify uncertainty

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

Tools:

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



Tree dating

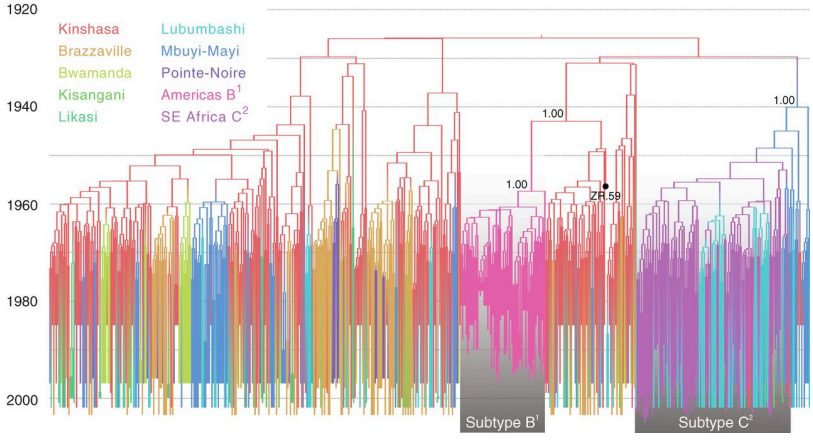
We can date a phylogenetic tree if we have dates of samples.
We can rescale the branch lengths of the phylogeny to obtain a dated tree, with ancestral nodes labelled in the temporal domain.

Tools

BEAST

IQTREE

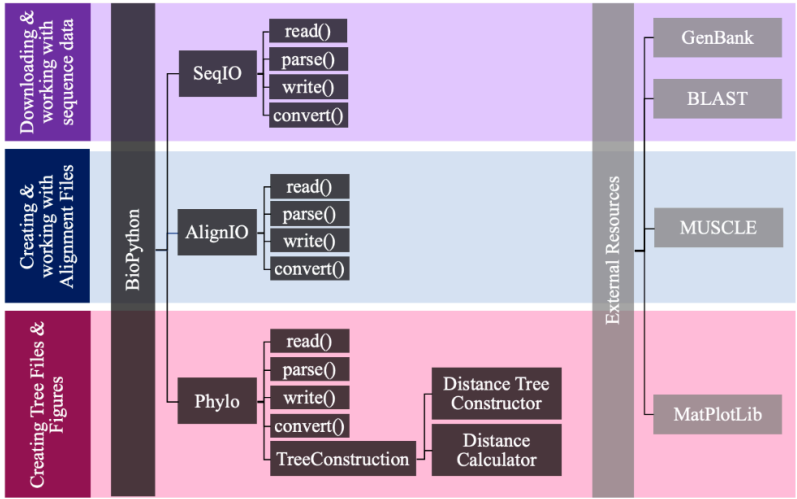
phylostan



Faria et al. (2014), *Science*

Tools for phylogenetic pipeline

Python and external



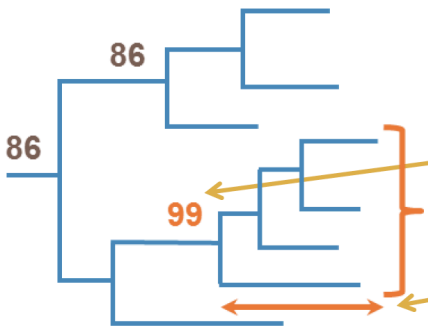
Analysing phylogenetic trees

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



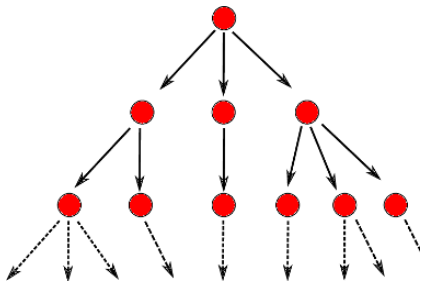
- Clusters are identified based on
  - high bootstrap and
  - low within cluster genetic distance

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

Analysing clusters

- Characterise clusters and estimate epidemiological parameters, e.g. using a branching process to model the final size of the epidemic from an initial case in each cluster
- Can incorporate a sampling mechanism to adjust for partially observed phylogeny (non-sequenced individuals)

Number of secondary infections of transmission degree n caused by k^{th} individual from preceding generation), $R_{k,n} \sim \text{Poisson}(R_0)$, where R_0 is the reproductive rate.

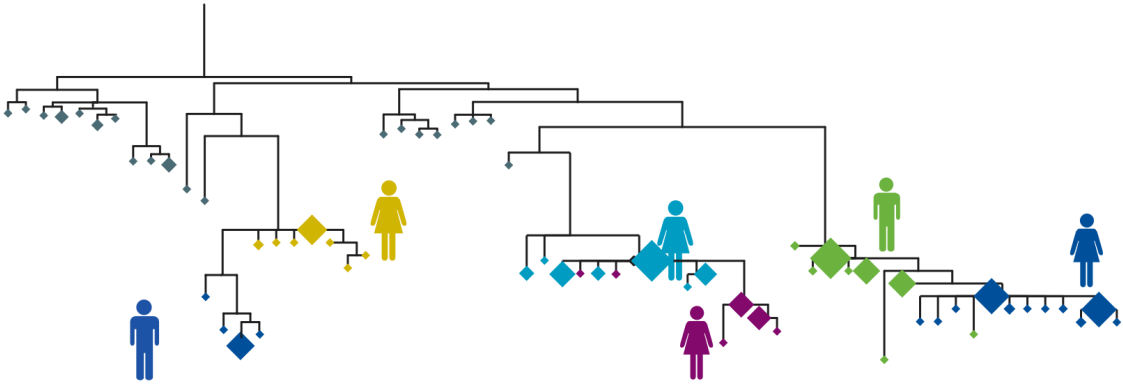


Analysing phylogenetic trees

- 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

Ancestral state reconstruction

Identifying likely transmission pairs for source attribution

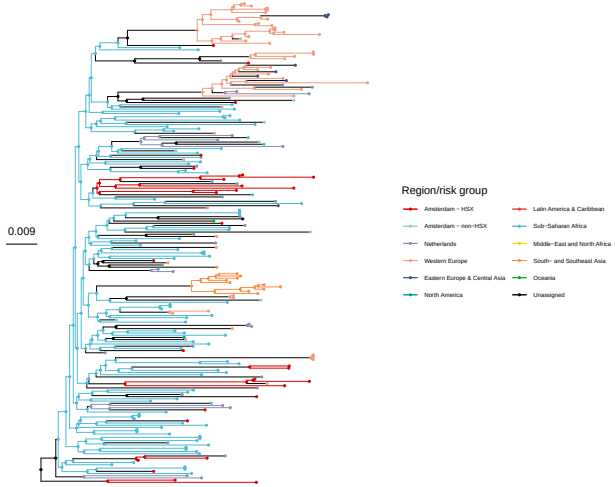


Ratmann et al (2020), *Lancet HIV*

Ancestral state reconstruction

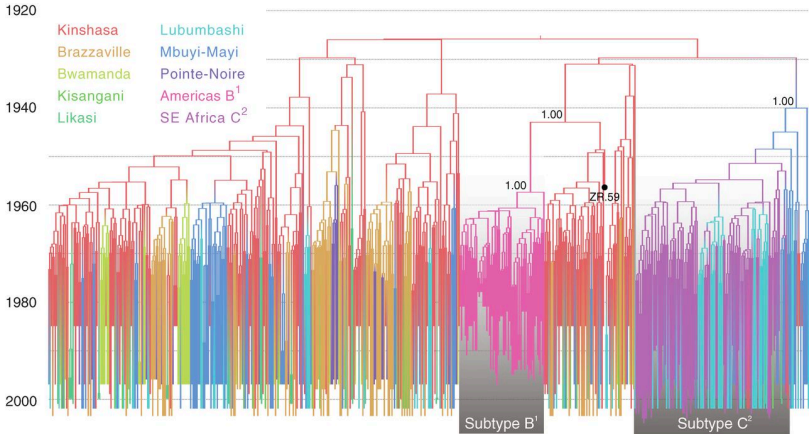
Separating phylogenetic transmission chains in study population
Software:

- *Phyloscanner*
- *BEAST*



Ancestral state reconstruction

Inferring the spatial epidemiological history of a virus (phylogeography)



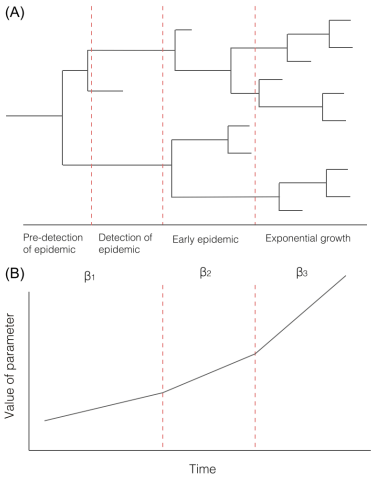
Faria et al. (2014), *Science*

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**

Viral phylodynamics

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



Hill (2021), *Trends in Parasitology*

Summary

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

References

[1] [2]


Hein, J., Schierup, M. H. & Wiuf, C.
Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory (Oxford University PressOxford, 2004).
URL http://dx.doi.org/10.1093/oso/9780198529958.001.0001.


Kingman, J. F. C.
On the genealogy of large populations.
Journal of Applied Probability **19**, 27–43 (1982).
URL http://dx.doi.org/10.2307/3213548.

Questions?