Introduction to Phylogenetics

Created by Dr Francesc Coll
Senior Staff Scientist
Wellcome Sanger Institute

Content

What are phylogenetic trees?

How are phylogenetic trees reconstructed?

Methods for phylogenetic inference

Homologous Recombination

Ancestral state reconstruction

How are phylogenetic trees interpreted?

Nomenclature, ancestry, topology, phylogeography

Applications of phylogenetic trees

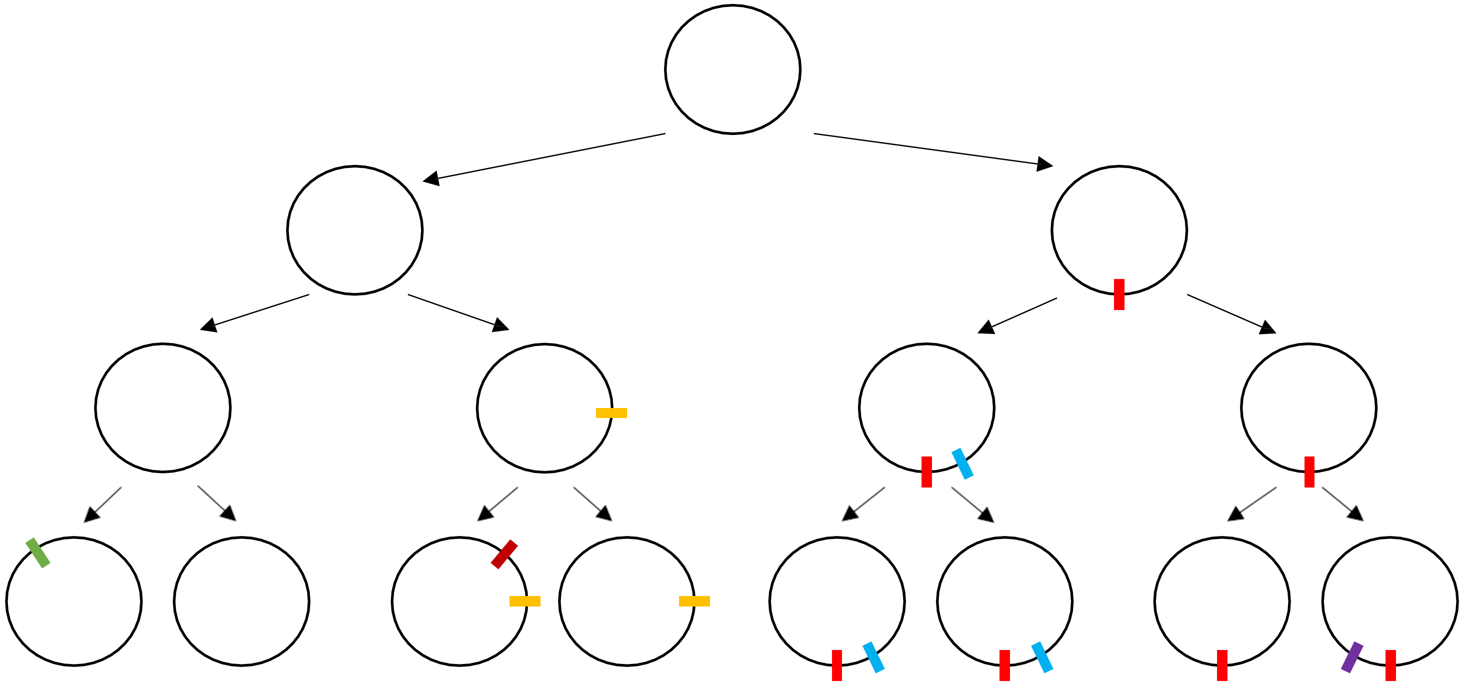
Phylogenetics applied to Genomic Surveillance

What are phylogenetic trees?

A phylogeny, also known as phylogenetic tree, depicts estimated evolutionary relationships between taxa - these can be species, strains or even genes.

Bacteria reproduce clonally replicating their DNA at high fidelity.

Random errors in DNA replication may still occur, resulting in a clonal progeny that will inherit these genetic replication 'errors' (i.e. mutations) in their DNA and may not be strictly identical to their progenitor cells.

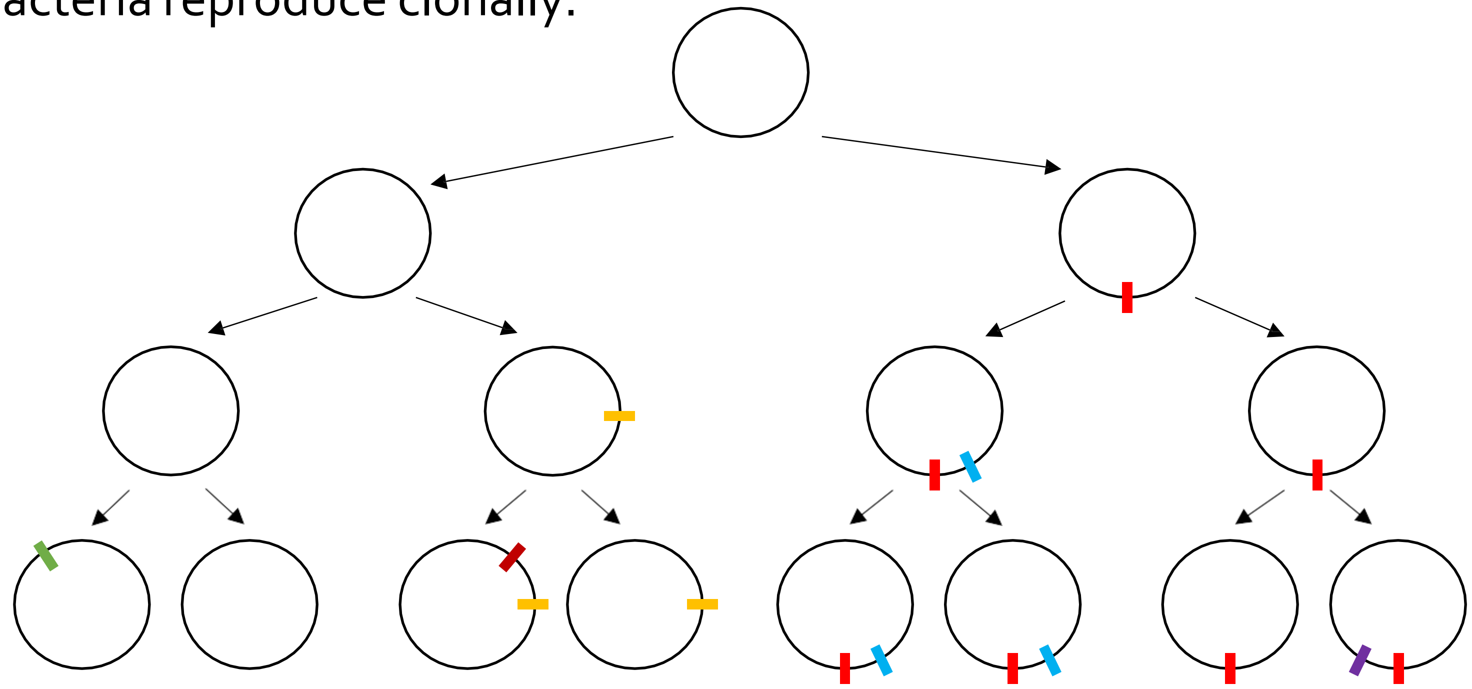


What are phylogenetic trees?

Bacterial strains that have recently originated from the same progenitor cell are thus expected to share identical genomes, or have diverged at most by only a few genetic differences (mutations).

The number and pattern of shared mutations between bacterial strains can be used to reconstruct their genealogical and evolutionary relationships.

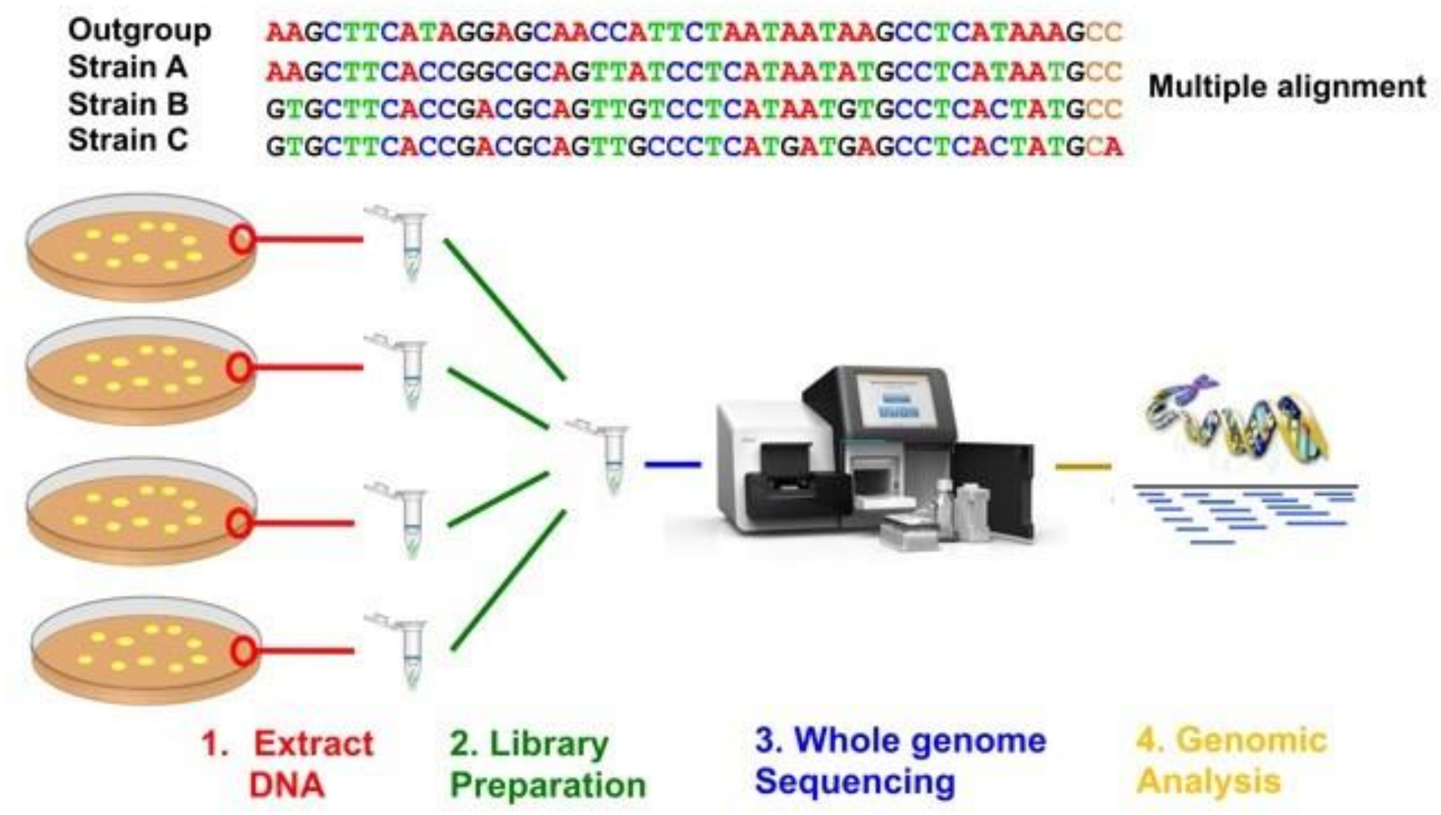
In the context of infectious diseases epidemiology, phylogenetic trees are commonly used to define evolutionary relationships between strains of the same bacterial species. This is possible because bacteria reproduce clonally.



How are phylogenetic trees reconstructed?

Today almost all phylogenetic trees are inferred from molecular sequence data, most often from DNA sequences.

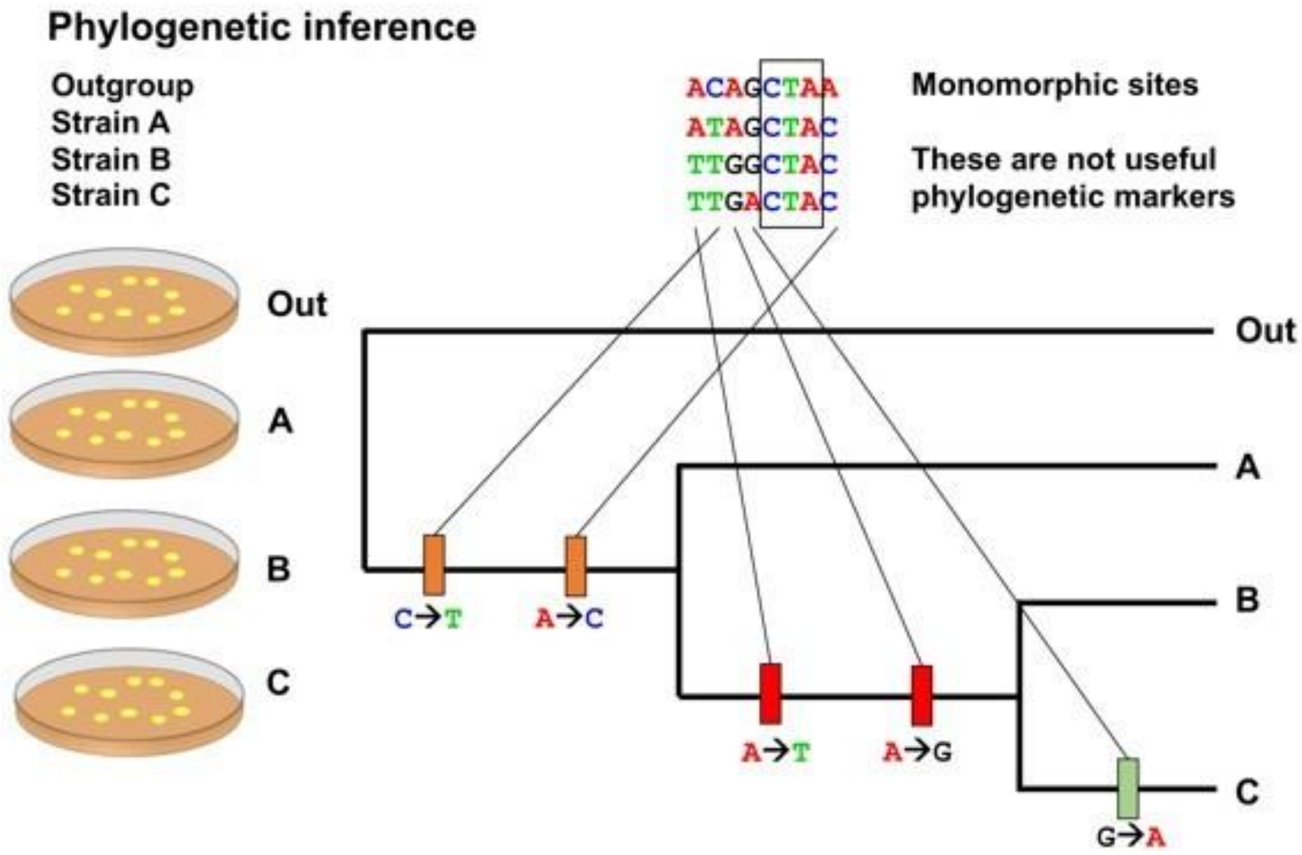
Whole-genome sequencing now makes it possible to ‘read’ the DNA sequence of the entire bacterial chromosome, which provides the ultimate level of resolution to discriminate between closely related strains.



```
Outgroup   AAGCTTCATAGGAGCAACCATTCTAATAATAAGCCTCATAAAGCC
Strain A   AAGCTTCACCGGCGCAGTTATCCTCATAATATGCCTCATAATGCC   Multiple alignment
Strain B   GTGCTTCACCGACGCAGTTGTCCTCATAATGTGCCTCACTATGCC
Strain C   GTGCTTCACCGACGCAGTTGCCCTCATGATGAGCCTCACTATGCA
```

How are phylogenetic trees reconstructed?

The identification of genetic changes (alleles) that are unique and common to multiple taxa (strains) are used to group them into monophyletic groups (clades) in a hierarchical manner (see example below) with the goal of constructing the most plausible genealogical relationships between strains and clades.



Methods for phylogenetic inference

Distance methods

Evolutionary distances are used to construct trees (UPGMA & Neighbor Joining).

Parsimony

Trees are created to minimize the number of changes that are needed
to explain the data.

Maximum Likelihood

Using a model for sequence evolution, create a tree that gives the highest likelihood of
occurring with the given data.

Bayesian methods

Like ML but can incorporate prior knowledge

Methods for phylogenetic inference

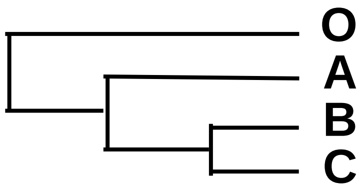
Distance-based methods use the amount of dissimilarity (the distance) between two aligned sequences to derive trees.

```
O  ACAGCTAA
A  ATAGCTAC
B  TTGGCTAC
C  TTGACTAC
```

DNA multiple alignment

	O	A	B	C
O	0	2	4	5
A		0	2	3
B			0	1
C				0

Distance matrix



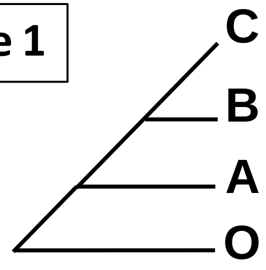
Phylogenetic tree

Methods for phylogenetic inference

Parsimony

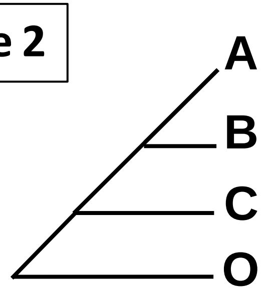
The most parsimonious tree, or shortest tree is one that requires the fewest total evolutionary changes.

Tree 1



TTGACTAC
TTGGCTAC
ATAGCTAC
ACAGCTAA

Tree 2

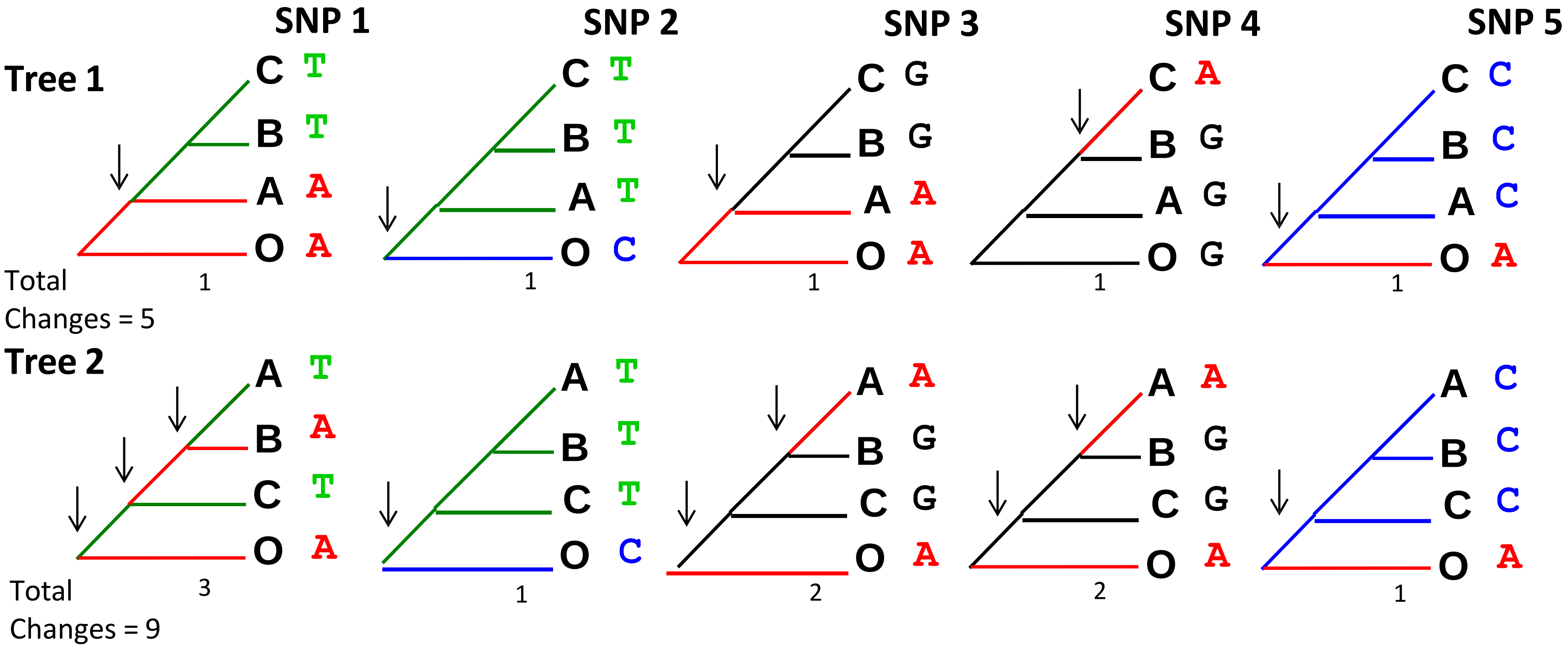


ATAGCTAC
TTGGCTAC
TTGACTAC
ACAGCTAA

Methods for phylogenetic inference

Parsimony

The most parsimonious tree, or shortest tree is one that requires the fewest total evolutionary changes.

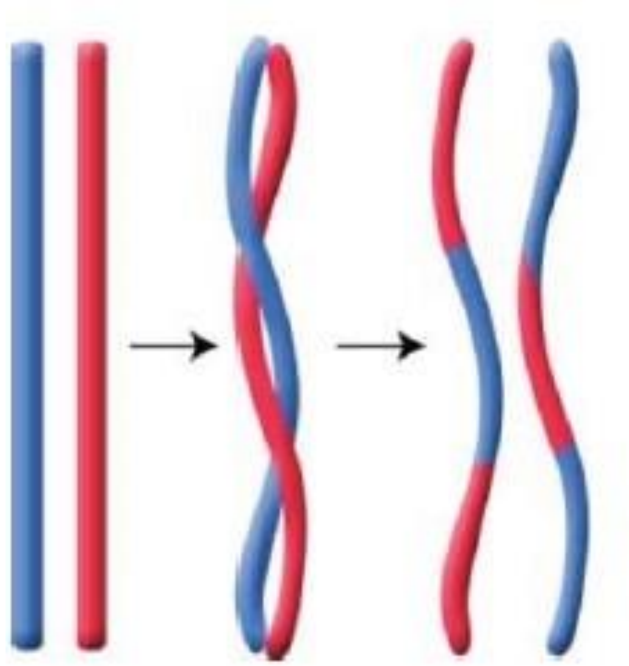


Homologous Recombination

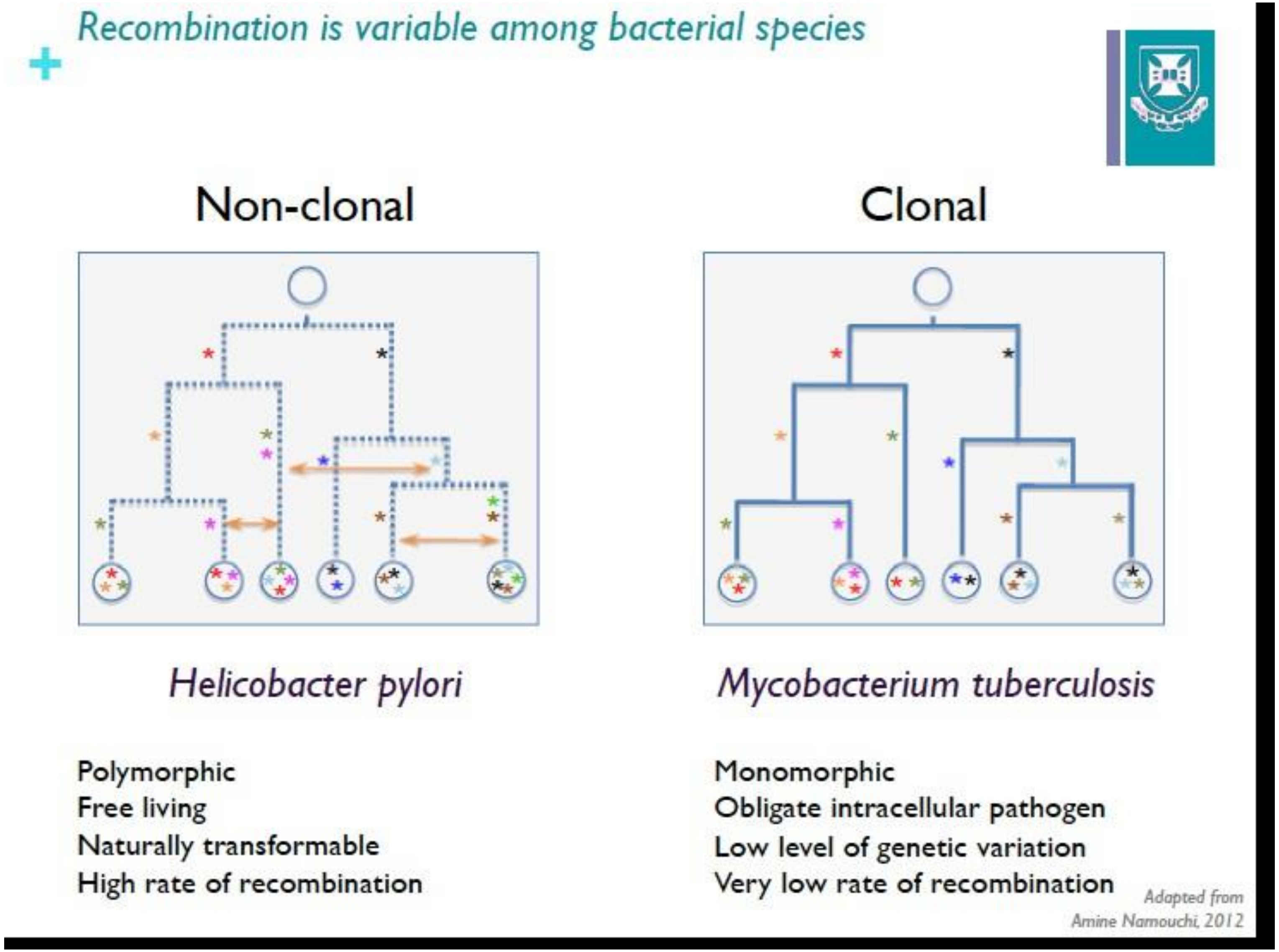
Bacteria reproduce clonally but sporadically exchange regions of their genomes by a process called homologous recombination which violates a fundamental assumption of phylogenetic methods.

Bacterial recombination event typically affects only a fraction of the genome.

It is recommended that removal of recombining sites to ameliorate their detrimental effect on phylogenetic analysis



Homologous Recombination



Recombination is variable among bacterial species

Non-clonal

Helicobacter pylori

Polymorphic
Free living
Naturally transformable
High rate of recombination

Clonal

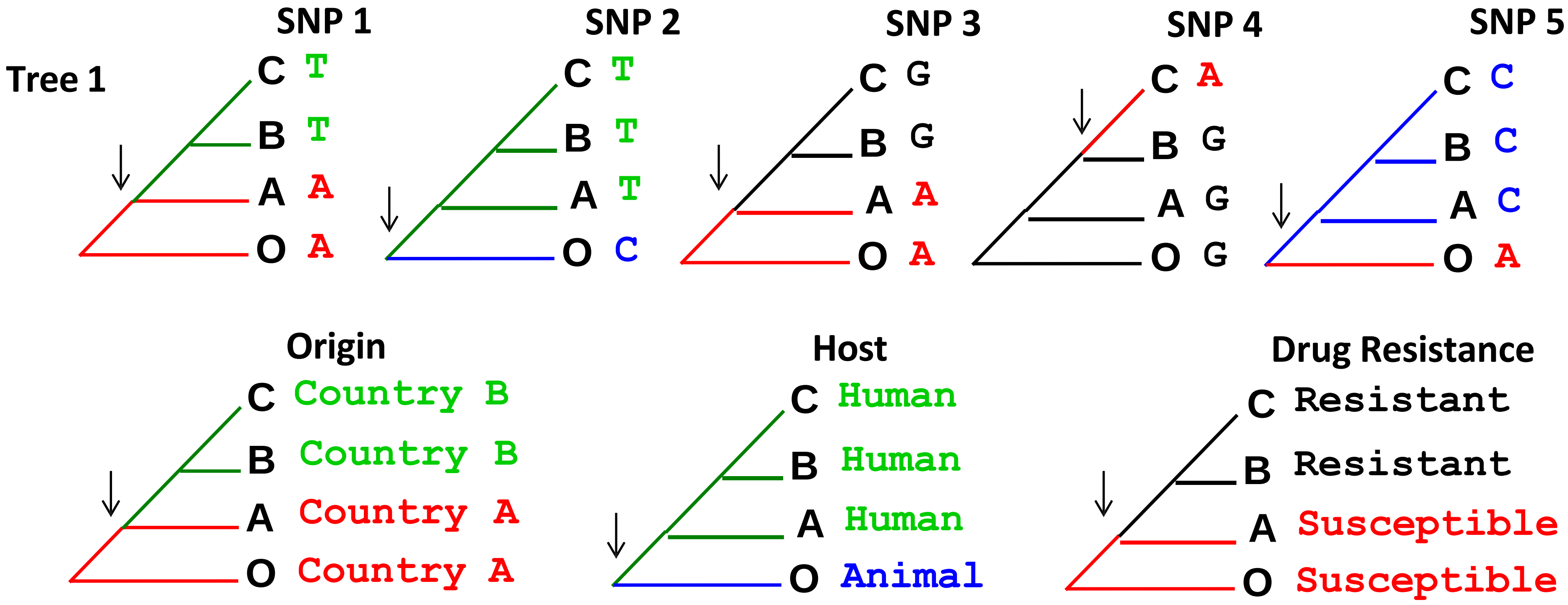
Mycobacterium tuberculosis

Monomorphic
Obligate intracellular pathogen
Low level of genetic variation
Very low rate of recombination

Adapted from
Amine Namouchi, 2012

Ancestral state reconstruction

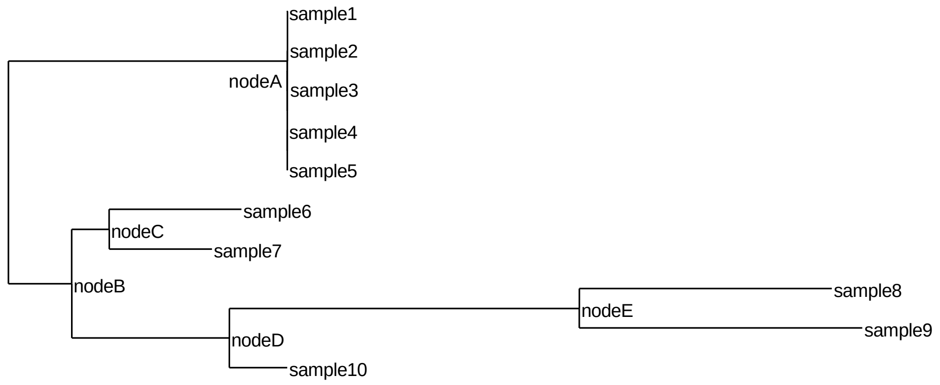
We can make use of a phylogeny and data strains to infer character states in ancestral taxa (map de evolution of traits)



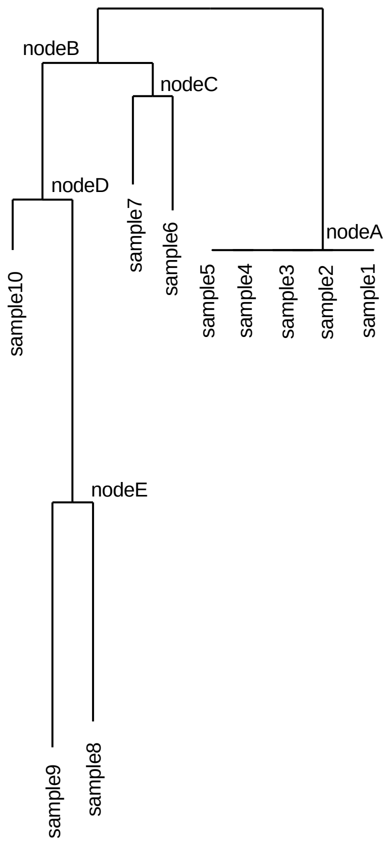
How are phylogenetic trees interpreted?

Same tree, different layouts

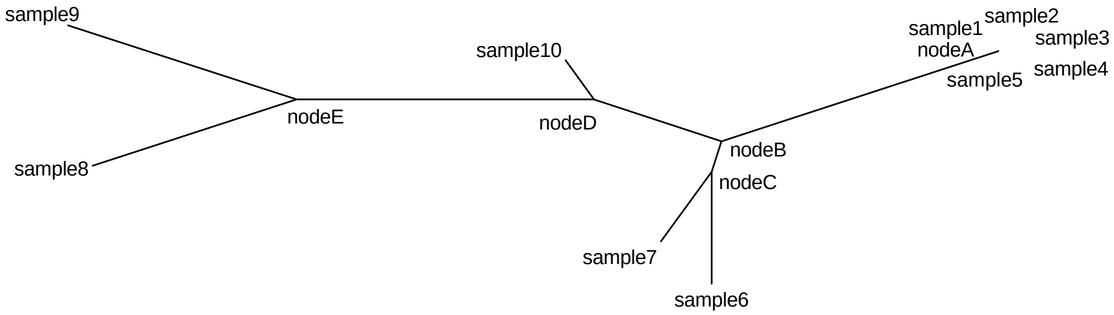
Rectangular phylogram (Horizontal)



Rectangular phylogram (Vertical)

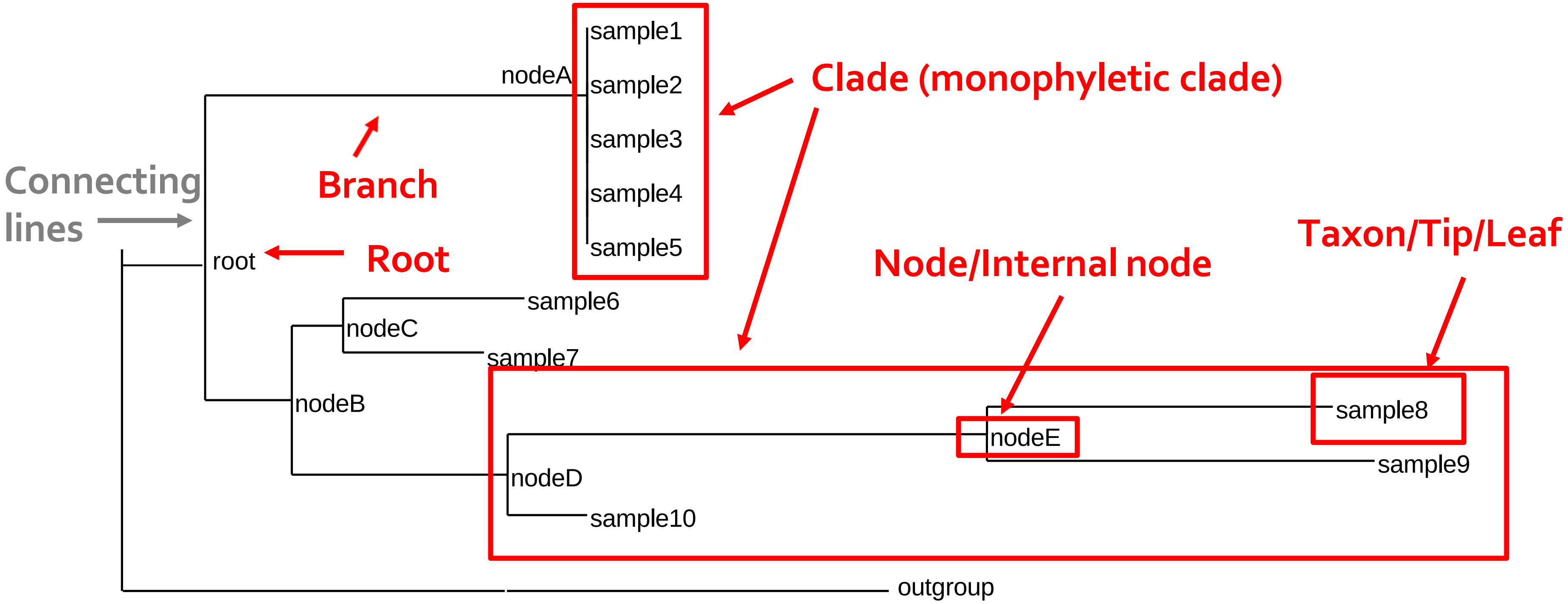


Radial phylogram



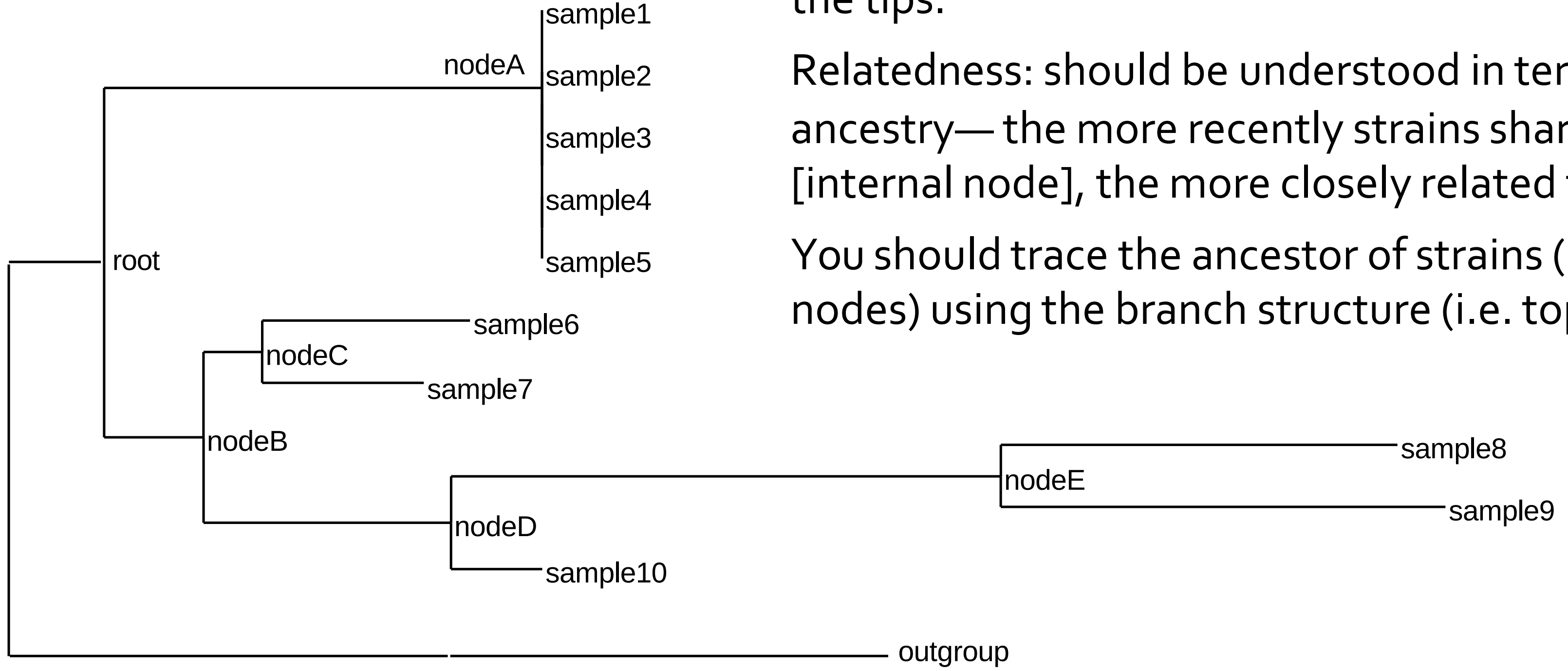
How are phylogenetic trees interpreted?

Nomenclature



How are phylogenetic trees interpreted?

Inferring relatedness from ancestry and topology



Phylogenies are commonly mis-interpreted when read along the tips.

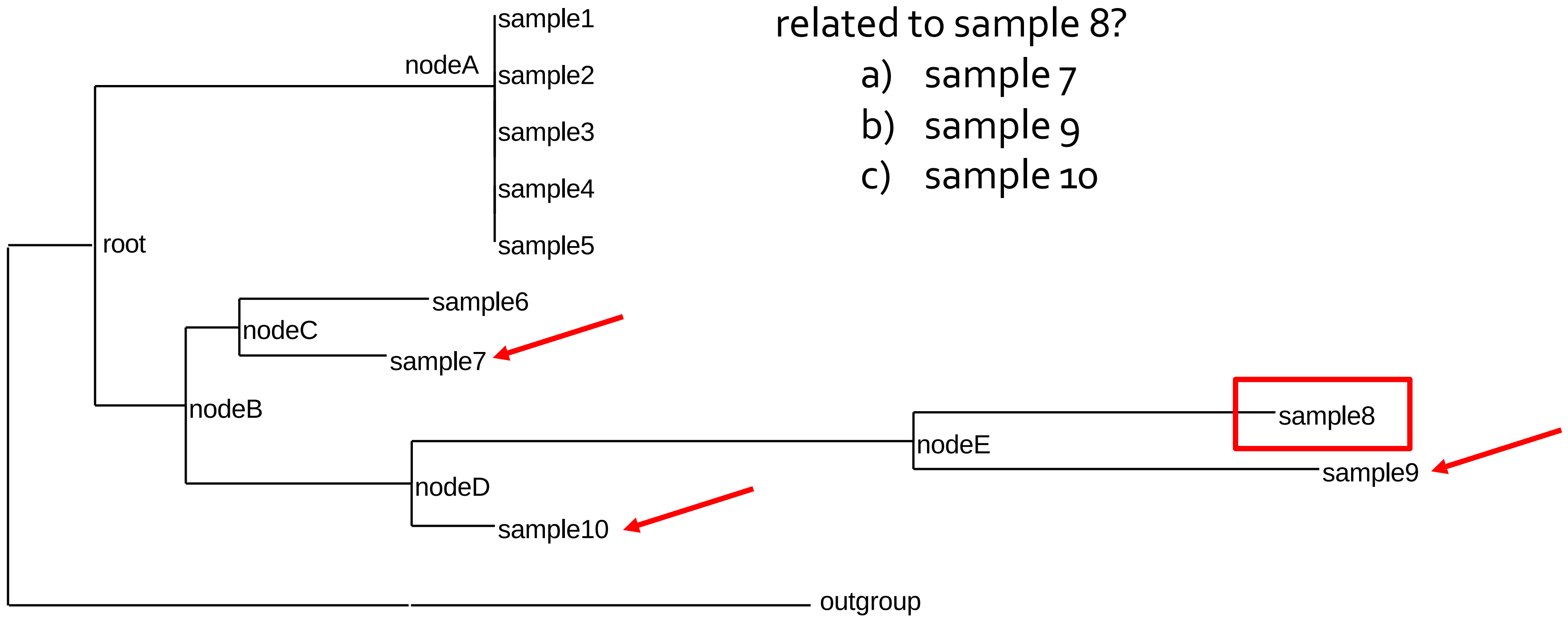
Relatedness: should be understood in terms of common ancestry— the more recently strains share a common ancestor [internal node], the more closely related they are.

You should trace the ancestor of strains (depicted as internal nodes) using the branch structure (i.e. topology) of the tree.

← More ancestral (closer to the root)   More recent (closer to the tips) →

How are phylogenetic trees interpreted?

Inferring relatedness from ancestry and topology



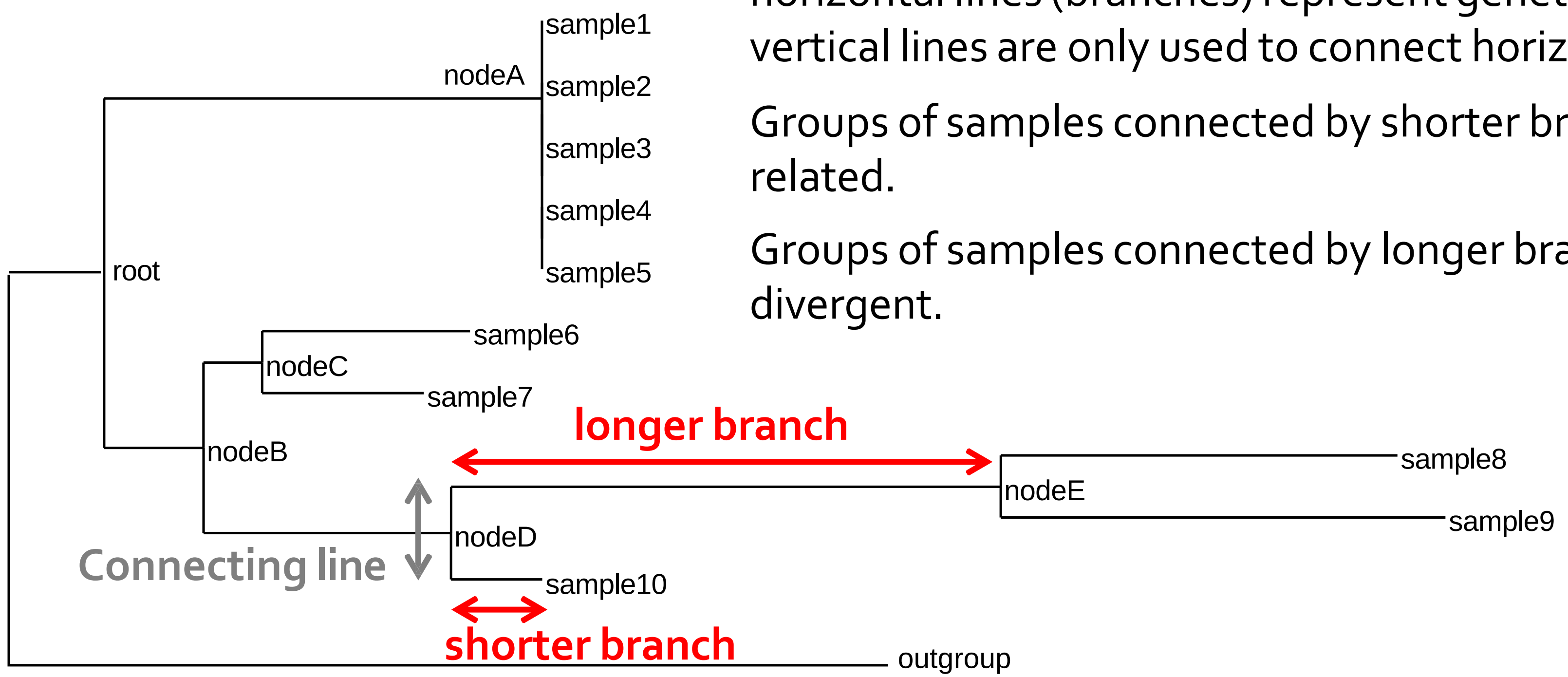
Question 1: Which sample is most closely related to sample 8?

a) sample 7
b) sample 9
c) sample 10

← More ancestral (closer to the root)   More recent (closer to the tips) →

How are phylogenetic trees interpreted?

Inferring relatedness from branch lengths



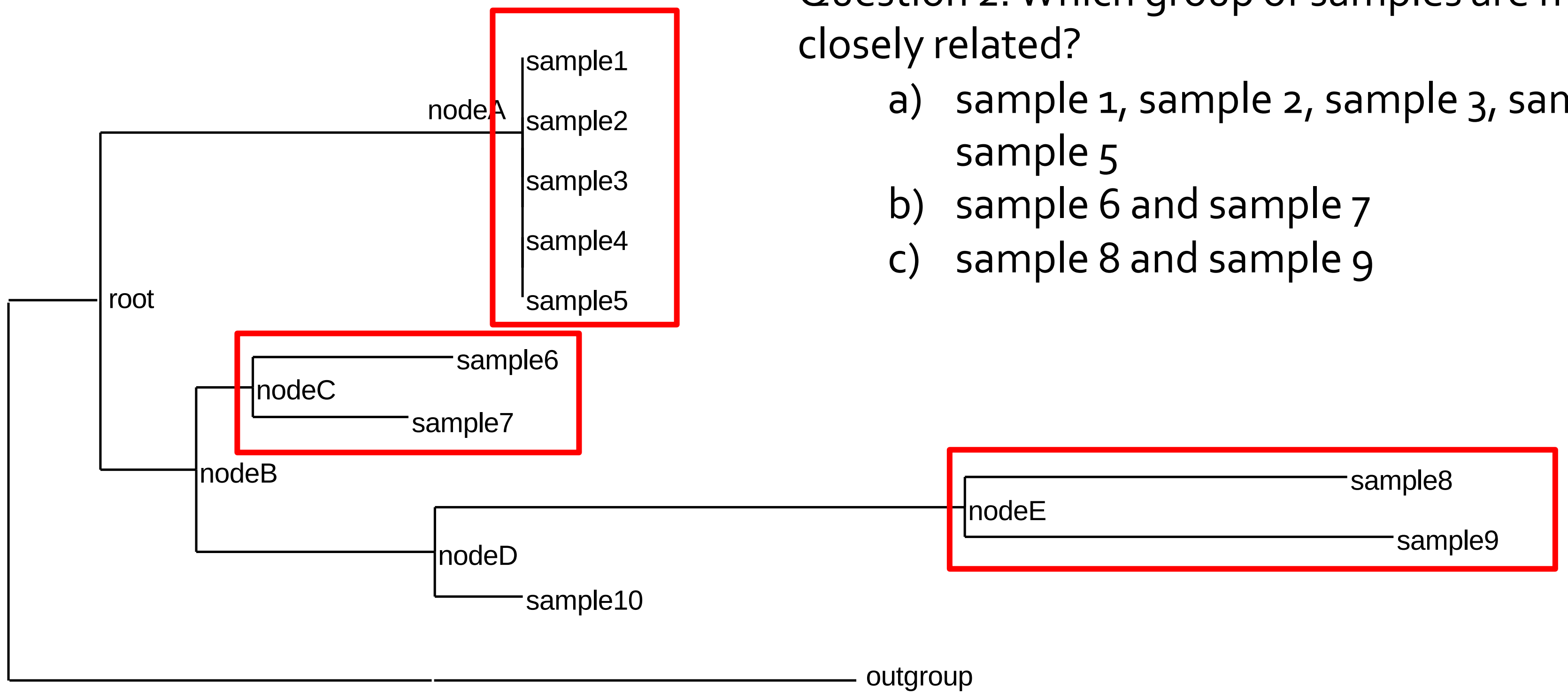
In a tree represented as a rectangular layout, the length of horizontal lines (branches) represent genetic distances whereas vertical lines are only used to connect horizontal lines.

Groups of samples connected by shorter branches are more closely related.

Groups of samples connected by longer branches are more divergent.

How are phylogenetic trees interpreted?

Inferring relatedness from branch lengths

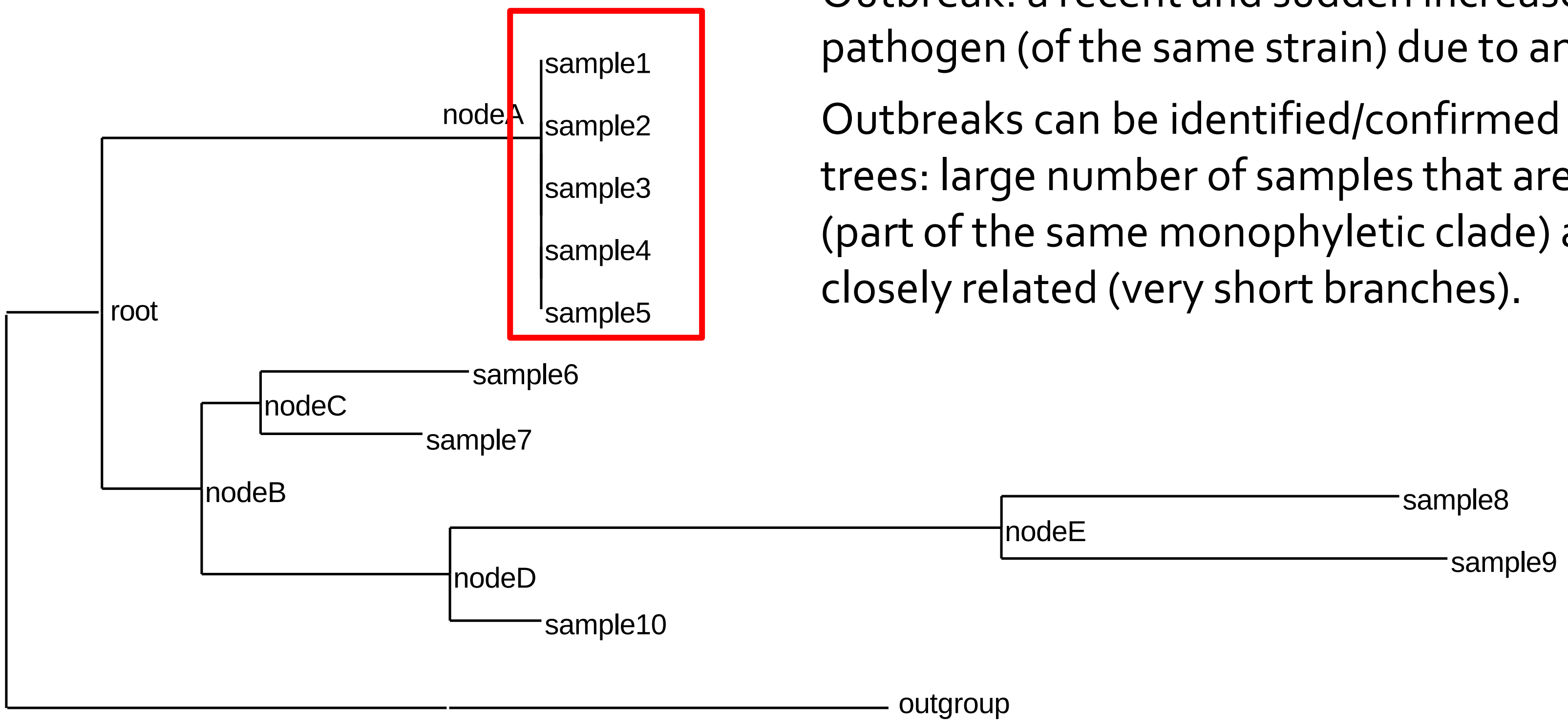


Question 2: Which group of samples are more closely related?

a) sample 1, sample 2, sample 3, sample 4, sample 5
b) sample 6 and sample 7
c) sample 8 and sample 9

How are phylogenetic trees interpreted?

Application: identifying outbreaks of pathogens

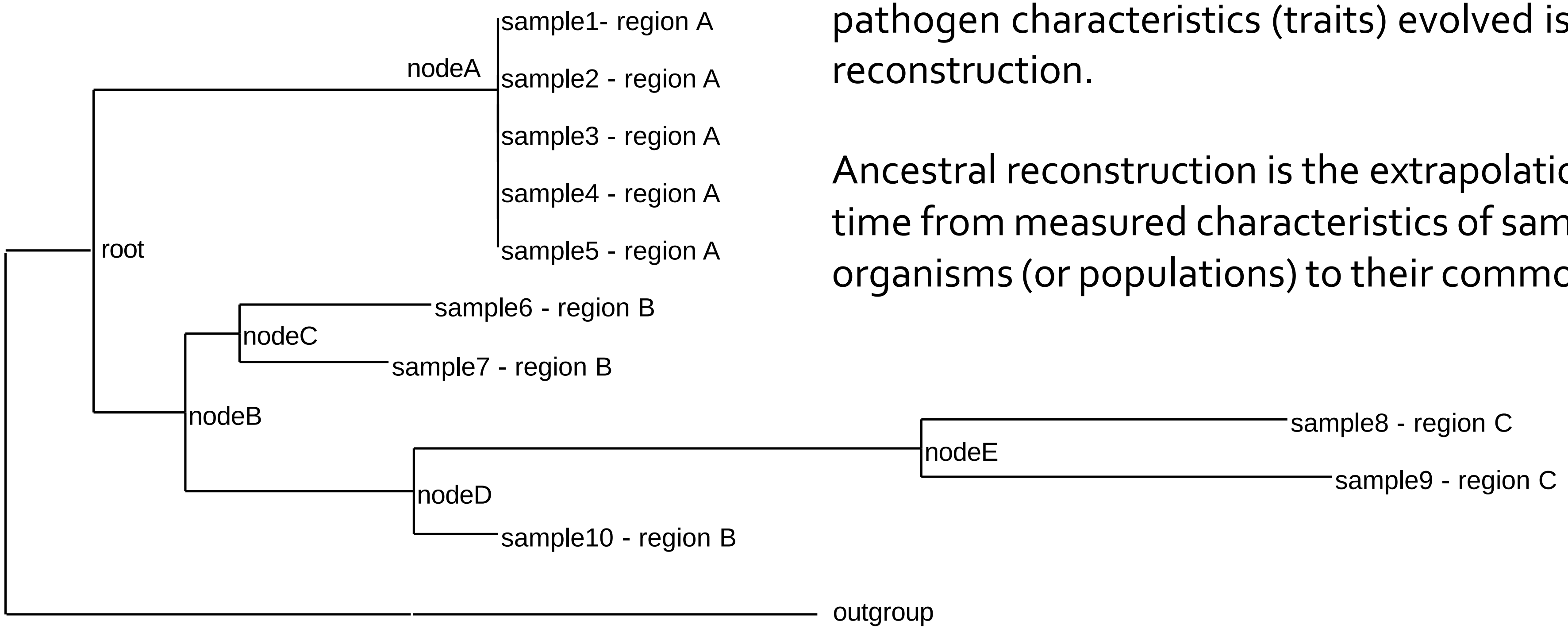


Outbreak: a recent and sudden increase in the prevalence of a pathogen (of the same strain) due to an increase transmission.

Outbreaks can be identified/confirmed using phylogenetic trees: large number of samples that are genetically clustered (part of the same monophyletic clade) and are also very closely related (very short branches).

How are phylogenetic trees interpreted?

Application: ancestral reconstruction & phylogeography

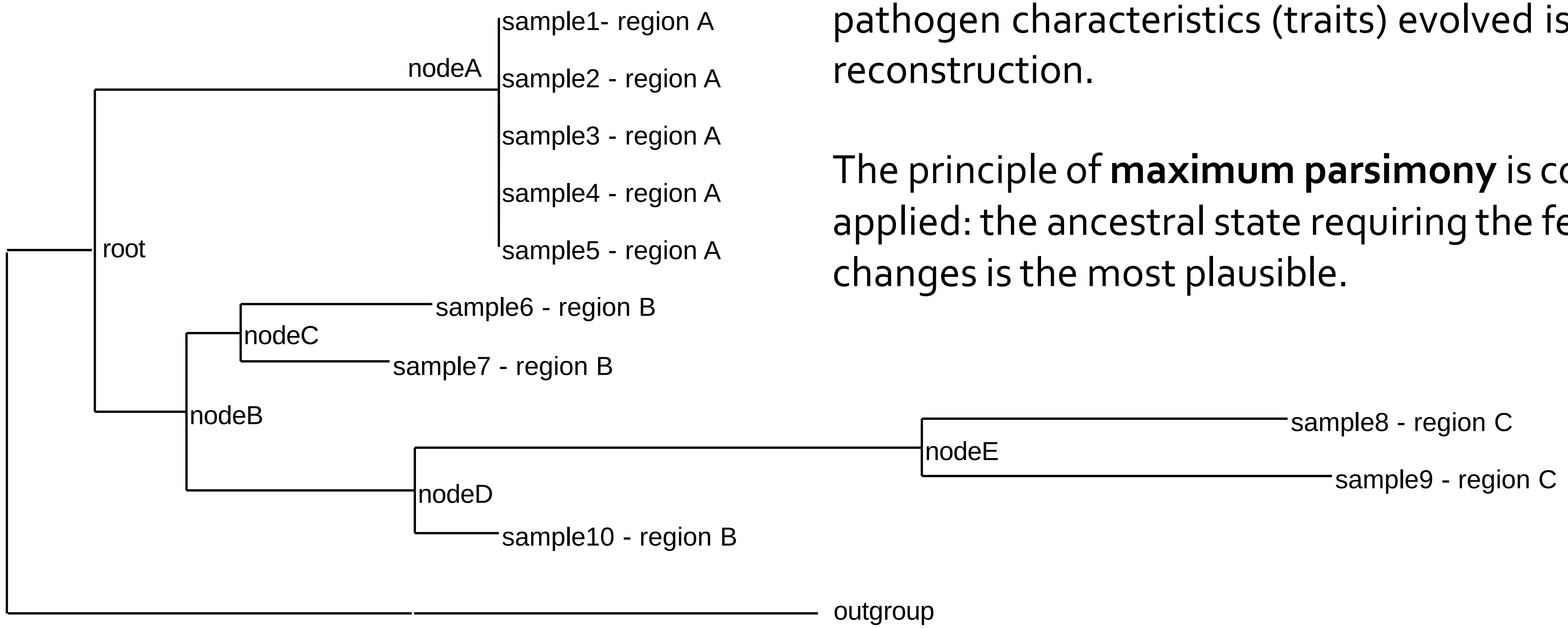


A common phylogenetic method used to study how pathogen characteristics (traits) evolved is ancestral reconstruction.

Ancestral reconstruction is the extrapolation back in time from measured characteristics of sampled organisms (or populations) to their common ancestors.

How are phylogenetic trees interpreted?

Application: ancestral reconstruction & phylogeography

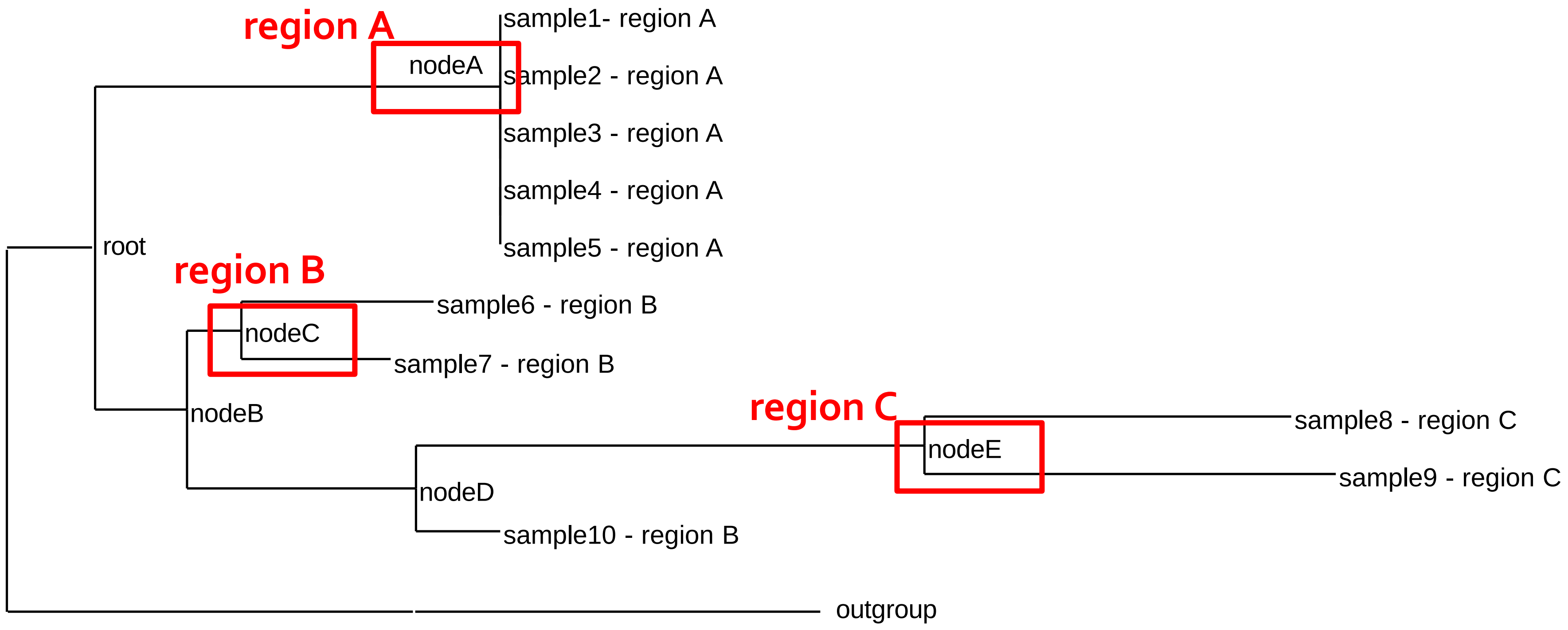


A common phylogenetic method used to study how pathogen characteristics (traits) evolved is ancestral reconstruction.

The principle of **maximum parsimony** is commonly applied: the ancestral state requiring the fewest changes is the most plausible.

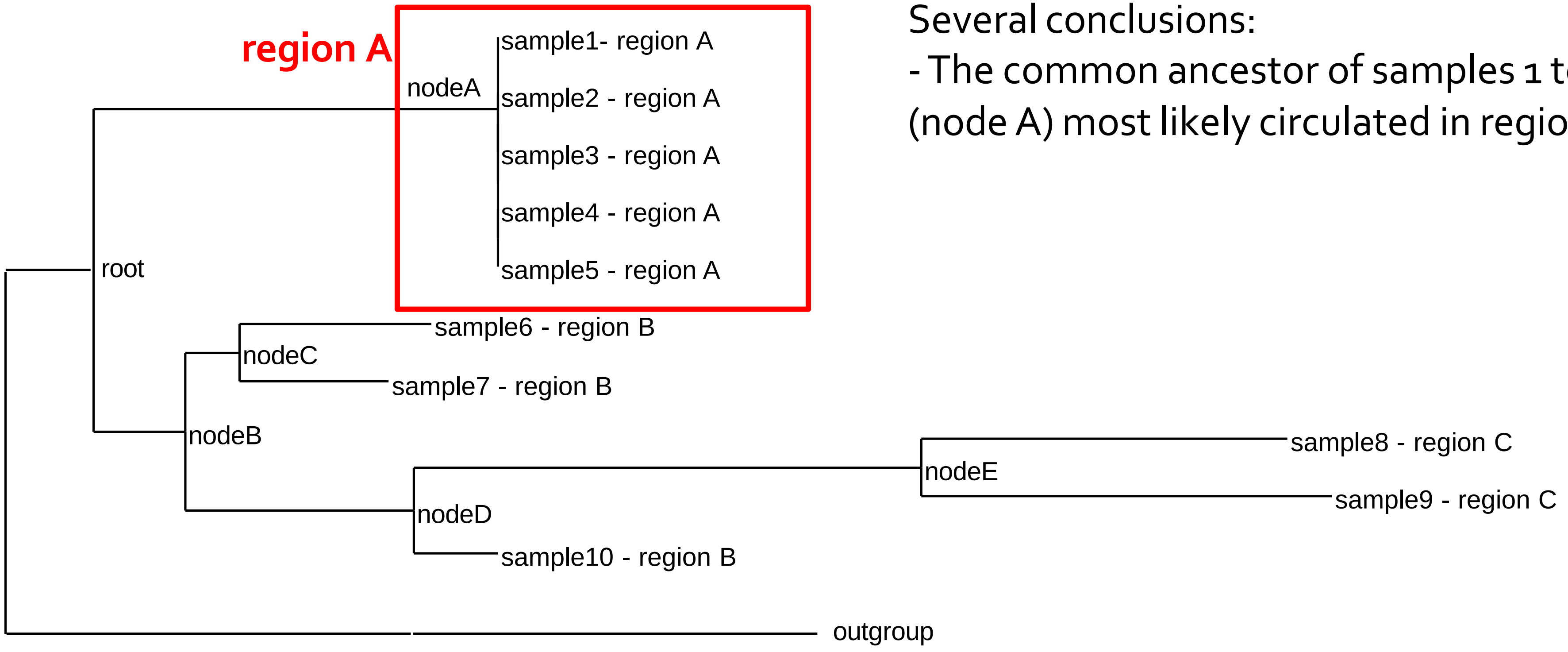
How are phylogenetic trees interpreted?

Application: ancestral reconstruction & phylogeography



How are phylogenetic trees interpreted?

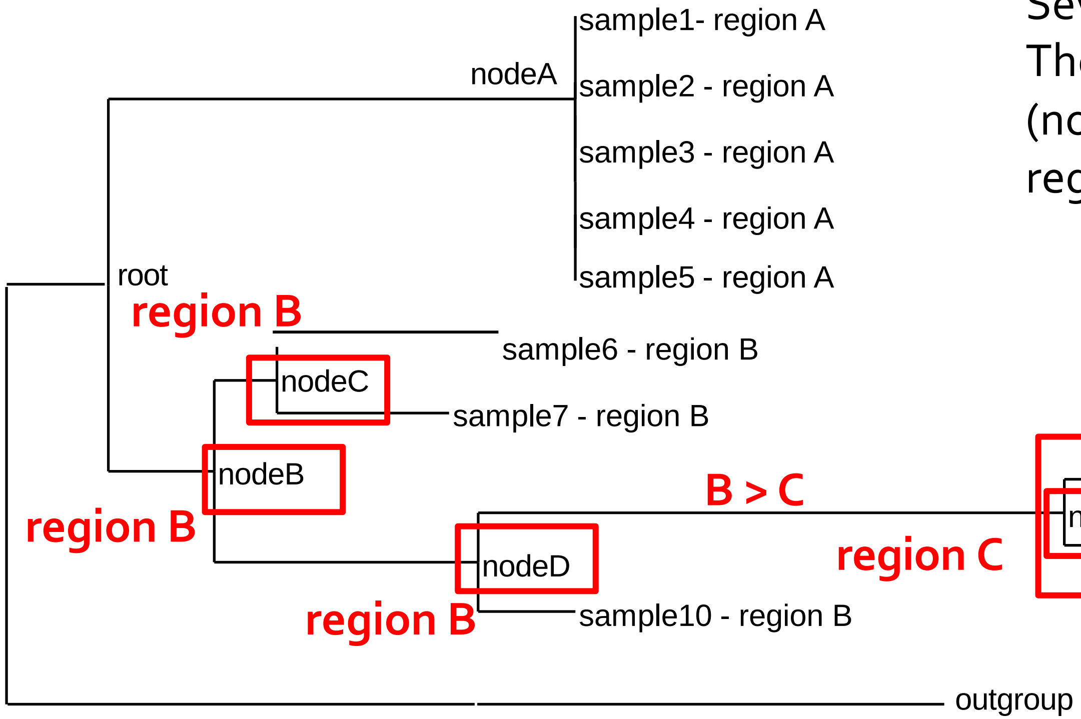
Application: ancestral reconstruction & phylogeography



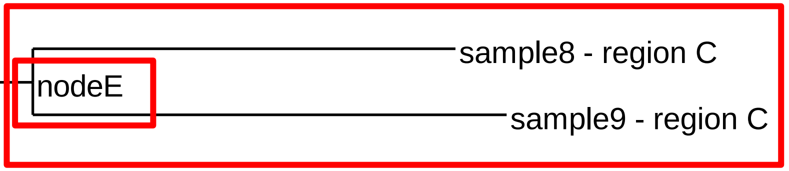
Several conclusions:
- The common ancestor of samples 1 to 5 (node A) most likely circulated in region A

How are phylogenetic trees interpreted?

Application: ancestral reconstruction & phylogeography



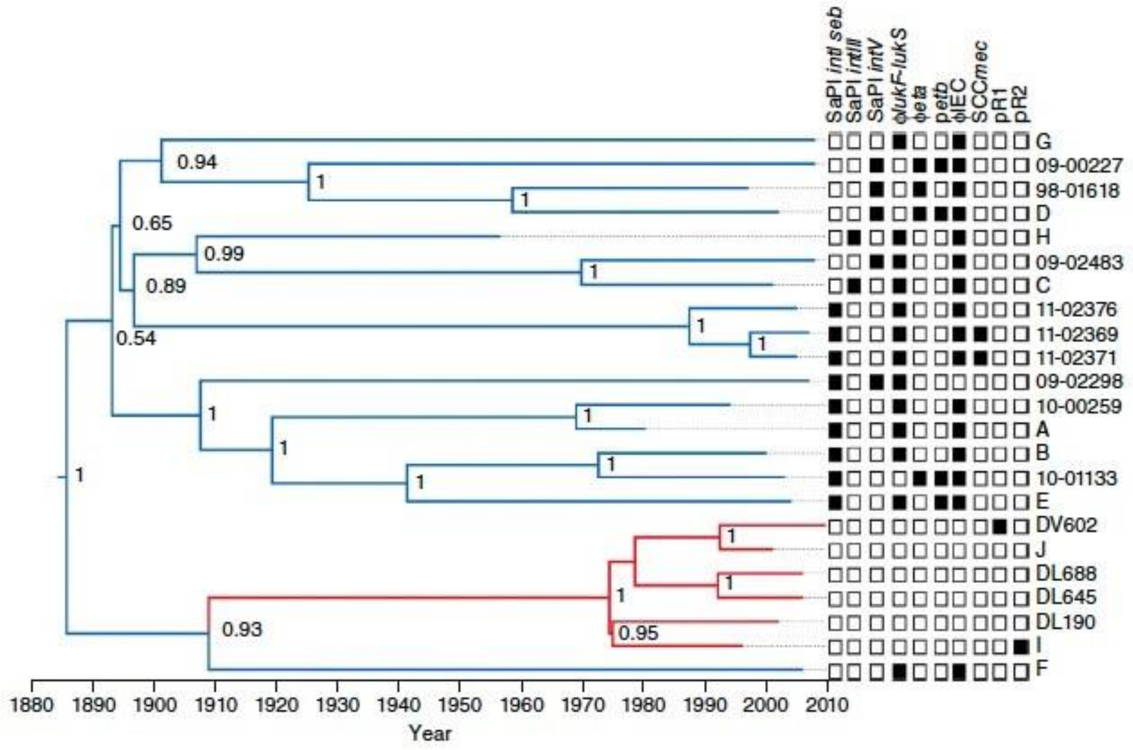
Several conclusions:
The common ancestor of samples 8 and 9 (node E) was most likely introduced from region B into region C.



Applications of phylogenetic trees

Trait evolution – example. host adaptation/tropism

Phylogeny ST121 *S. aureus*: human-to-rabbit host jump (blue, human; red, rabbit)



Human clades enclosing rabbit isolates (Source)

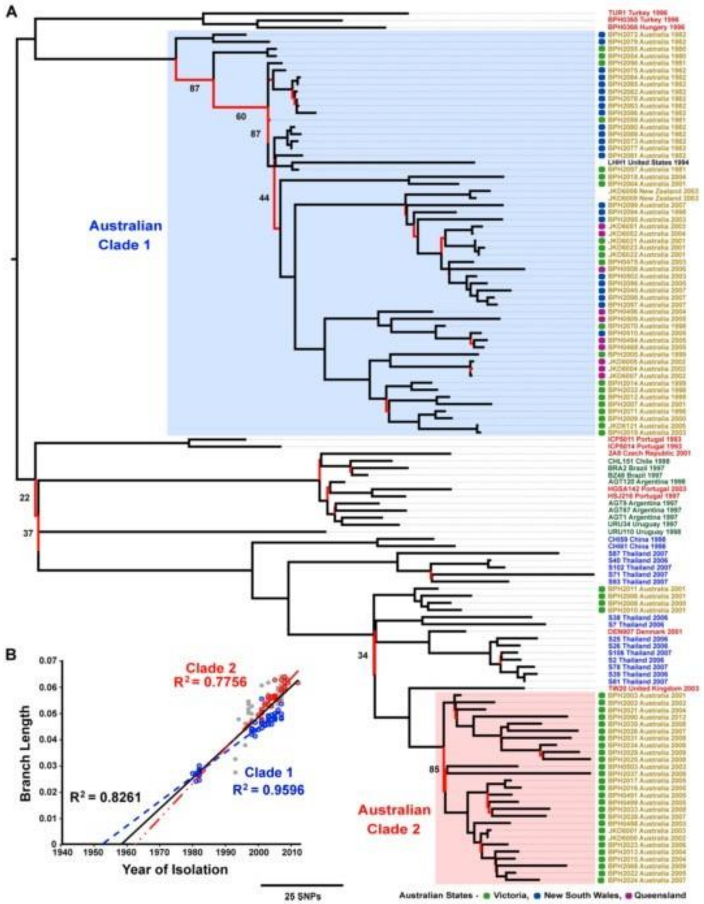
→

Rabbit clade nested within human isolates

Viana D *et al.* 2015. Nature genetics 47: 361–366.

Applications of phylogenetic trees

Geographic origins



Australian isolates

Asian isolates

Australian isolates

Asian isolates

Australian isolates

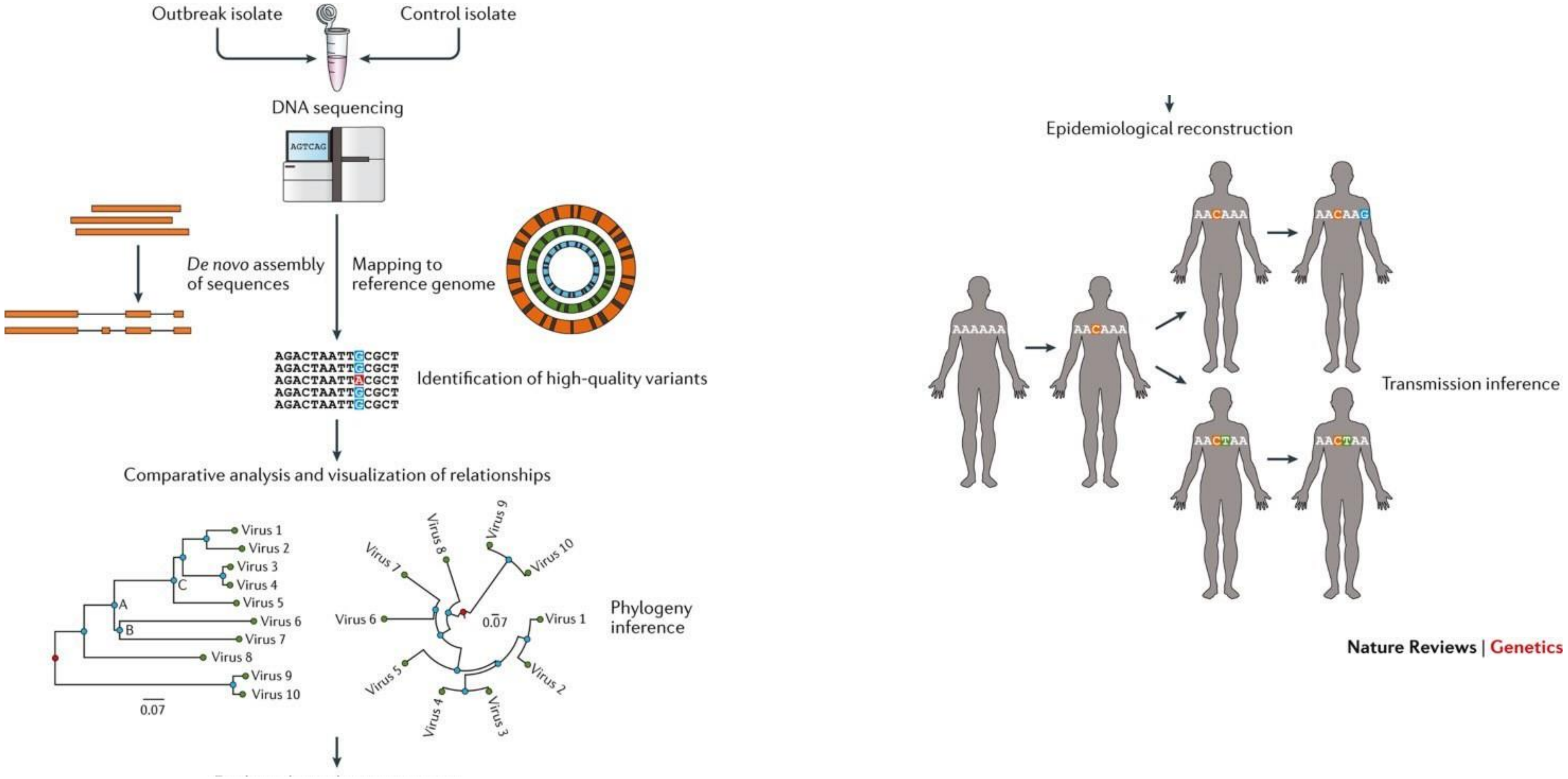
Baines SL et al. 2015 mBio

Clade 1 isolates represented all of the regions sampled and 27 of the 32 years in the temporal span of the collection (1980 to 2007) → more diverse → longer local circulation

Clade 1 being the original Australian ST239 clade and Clade 2 representing a more recent, previously unrecognized, reintroduction of ST239 into Australia from Asia.

Intercontinental transmission event between Asia and Australia, resulting in the local establishment of Clade 2.

Phylogenetics applied to Genomic Surveillance



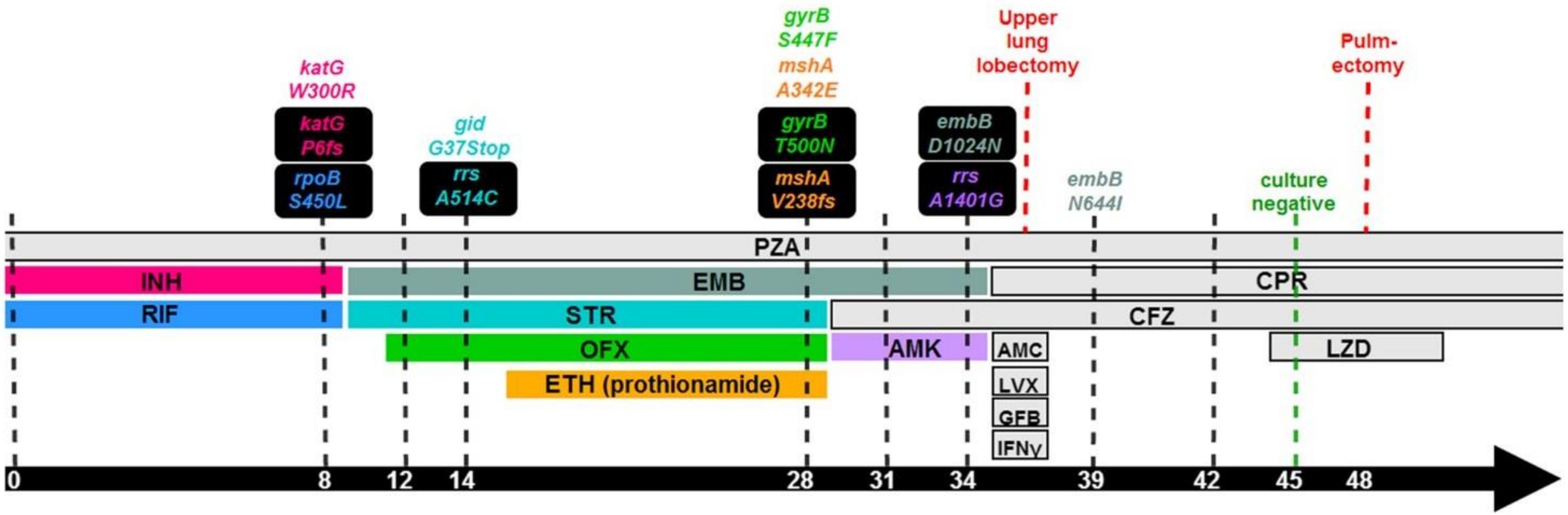
Source: Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics* **19**, 9–20 (2017).

Genomic surveillance of AMR

Emergence of resistance

WGS can be used to study the emergence of resistance within the same patient.

The first documented case of extensively drug-resistant tuberculosis evolved from a susceptible ancestor within a single patient. The vast majority of mutations identified over 3.5 years were either involved in drug resistance or hitchhiking in the genetic background of these.



Source: Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, Caugant DA, Mannsåker T, Mengshoel AT, Dyrhol-Riise AM, Balloux F. 2014. Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. *Genome Biology* **15**: 490.

Genomic surveillance of AMR

Emergence of resistance

WGS can be used to study the emergence of resistance within the same patient.

Other examples

- Acquisition of colistin resistance mutations in *Acinetobacter baumannii*
  Lim TP, Ong RT-H, Hon P-Y, Hawkey J, Holt KE, Koh TH, Leong ML-N, Teo JQ-M, Tan TY, Ng MM-L, et al. 2015. Multiple Genetic Mutations Associated with Polymyxin Resistance in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy* **59**: 7899–7902.

- Acquisition of vancomycin resistance mutations in *Staphylococcus aureus*
  Howden BP, Peleg AY, Stinear TP. 2014. The evolution of vancomycin intermediate Staphylococcus aureus (VISA) and heterogenous-VISA. *Infection, Genetics and Evolution* **21**: 575–582.

Source: Baker S, Thomson N, Weill F-X, Holt KE. 2018. Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. *Science* **360**: 733–738.

Genomic surveillance of AMR

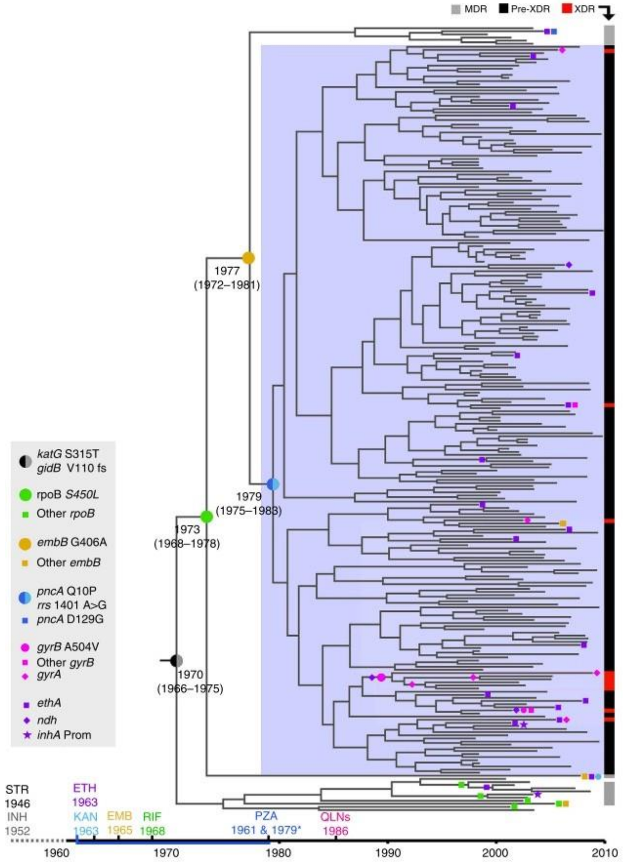
Emergence of resistance

WGS can be used to study the pattern (and order) of emergence of resistance in the same outbreak over time.

The timeline of the acquisition of antimicrobial resistance during a major ongoing outbreak of multidrug-resistant TB in Argentina was reconstructed.

The progenitor of the outbreak strain acquired resistance to isoniazid, streptomycin and rifampicin by around 1973, indicating continuous circulation of a multidrug-resistant TB strain for four decades.

Acquisition of resistance followed introduction of antibiotics (bottom of the figure).



Source: Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, Balloux F. 2015. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature communications* **6**: 7119.

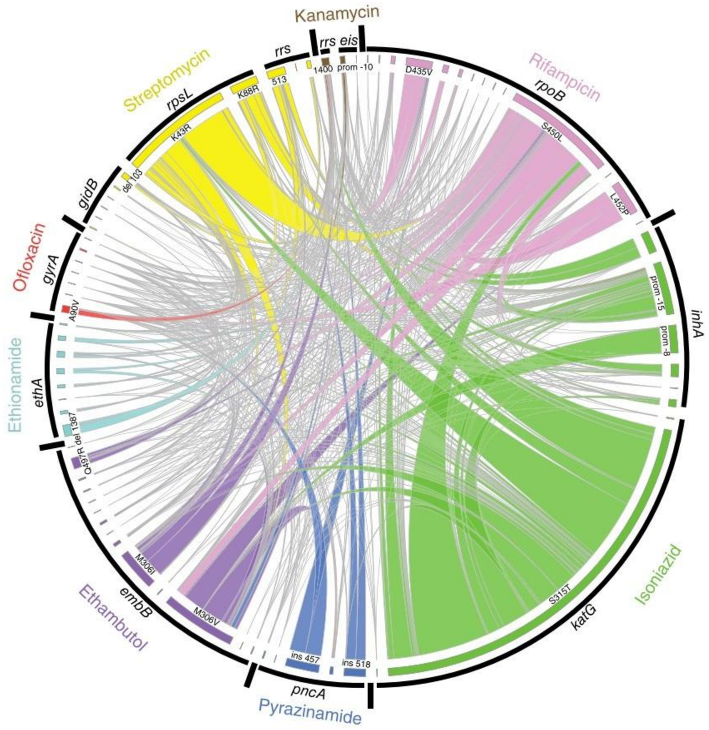
Genomic surveillance of AMR

Emergence of resistance

WGS can be used to study the pattern (and order) of emergence of resistance at the international level

WGS from 5,310 *M. tuberculosis* isolates from five continents. Despite diversity in geographical origin, genetic background and drug resistance, the patterns for the emergence of drug resistance were conserved globally.

Isoniazid resistance overwhelmingly arose before rifampicin resistance across all lineages, geographical regions and time. Earlier clinical introduction of isoniazid was not a major contributor. Ser315Thr mutation (the most common isoniazid-conferring resistance) is a well-tolerated mutation (low fitness cost).



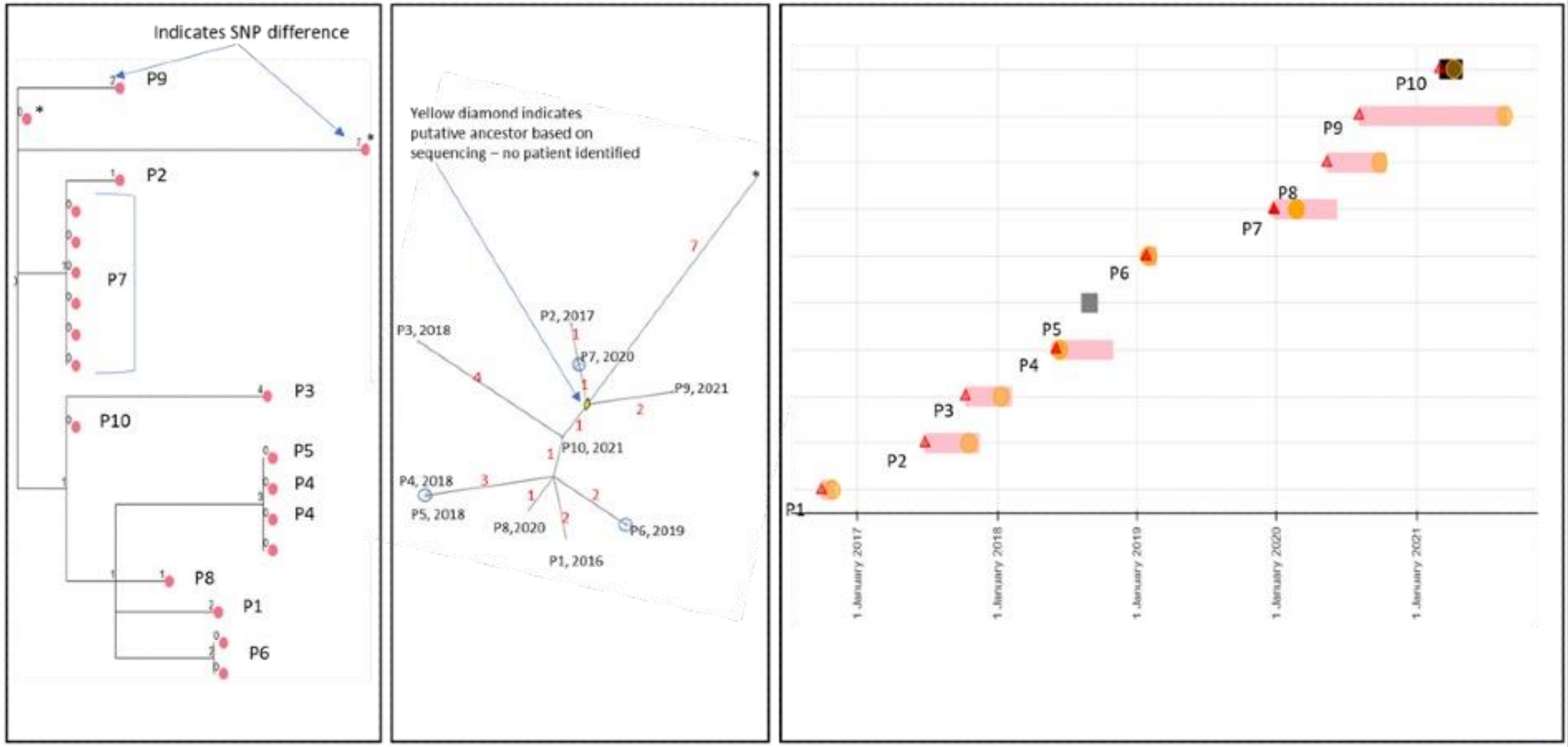
Source: Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, Brand J, Brand J, Jureen P, Malinga L, et al. 2017. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nature Genetics* **49**: 395–402.

Genomic surveillance: from proof-of-concept to routine use

Applied to *Mycobacterium tuberculosis* - Implementation of WGS by UKHSA

Recommended reading:

Mycobacterium tuberculosis whole-genome sequencing and cluster investigation handbook https://www.gov.uk/government/publications/tb-strain-typing-and-cluster-investigation-handbook/mycobacterium-tuberculosis-whole-genome-sequencing-and-cluster-investigation-handbook#appendix-1



Panel 1: Rooted ML tree. *non_UK sample

Panel 2: ML tree represented as blobogram (one sample non-UK). Red numbers=SNP distance

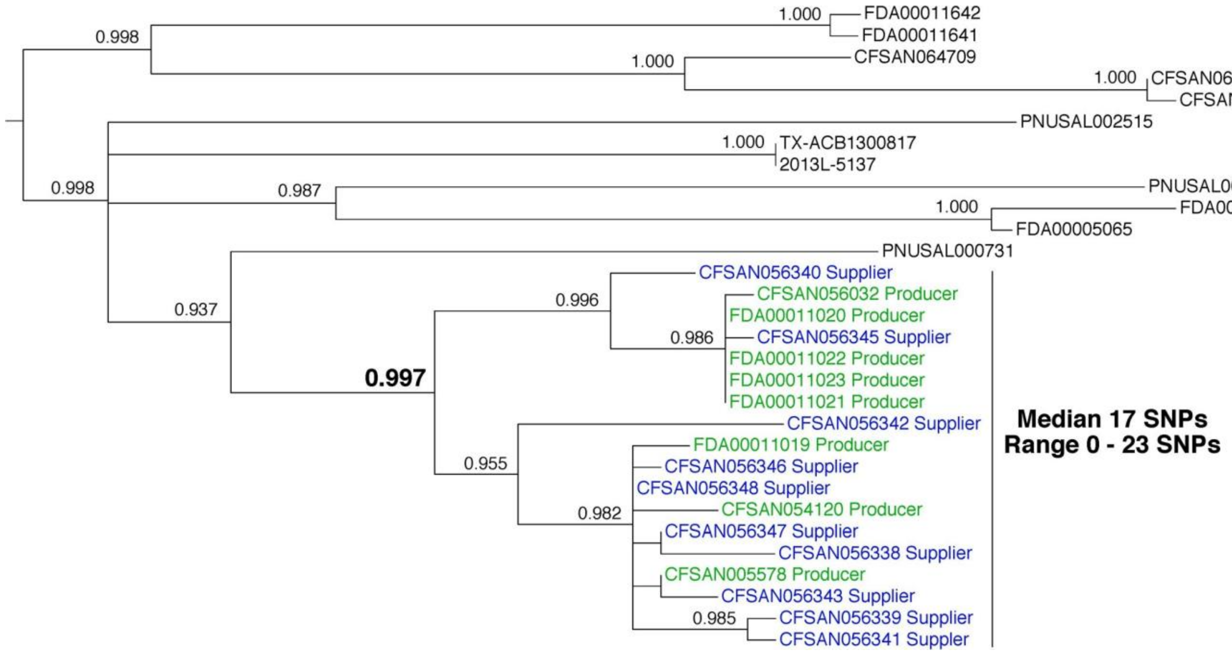
Panel 3: Timelines of patient: showing symptom start (▲)and diagnosis date (●). Pink bar =pulmonary disease, black= non pulmonary, grey= not known

Genomic surveillance: from proof-of-concept to routine use

Applied to foodborne outbreaks – Implementation of WGS by US FDA

Recommended reading:

Examples of How FDA Has Used Whole Genome Sequencing of Foodborne Pathogens For Regulatory Purposes https://www.fda.gov/food/whole-genome-sequencing-wgs-program/examples-how-fda-has-used-whole-genome-sequencing-foodborne-pathogens-regulatory-purposes



Phylogenetic analysis of genome sequences obtained from *Listeria monocytogenes* isolated from 2016 ice cream samples and the environment of a supplier.