Phylogeny: construction, visualization and interpretation

Introduction to Bioinformatics — Module 1

Public Health Alliance for Genomic Epidemiology

Erkison Ewomazino Odih

22 February 2021

Outline

- Phylogeny inference from SNPs
- Phylogenetic tree construction
- Phylogenetic tree interpretation
- Phylogenetic tree visualization – Microreact

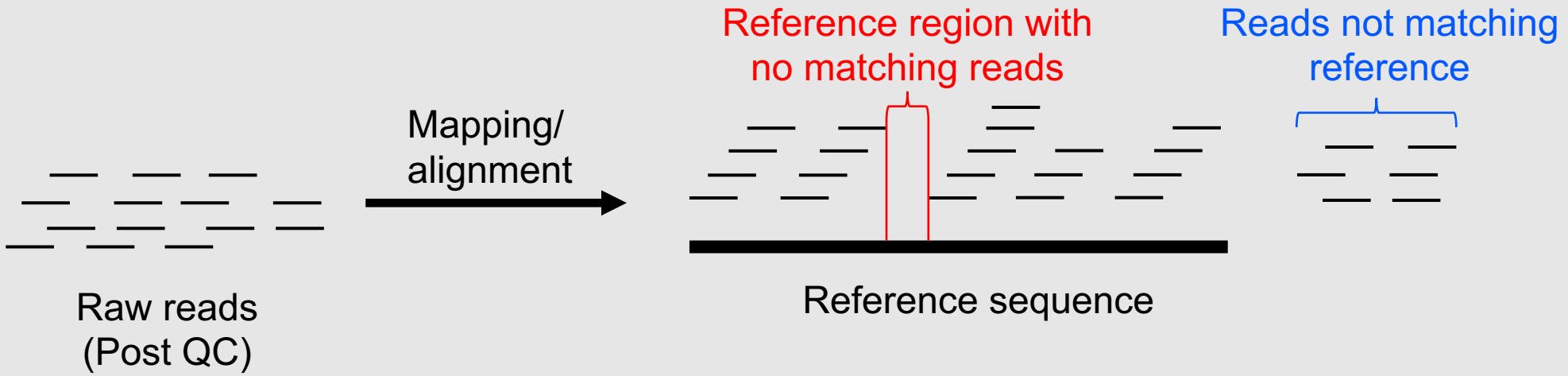
Phylogeny

- Phylogeny represents the evolutionary relationships and relatedness between a group of organisms.
- Inferred from single nucleotide polymorphisms (SNPs)
- SNPs are allelic nucleotide variants at given positions in the genome.
- For phylogeny analyses, SNPs are assumed to be:
  - Independent
  - Random

Public Health Alliance for Genomic Epidemiology

Inferring phylogeny from SNPs – Overview

1. Select reference and map reads to reference
   - The closer the reference to the sample data, the better.

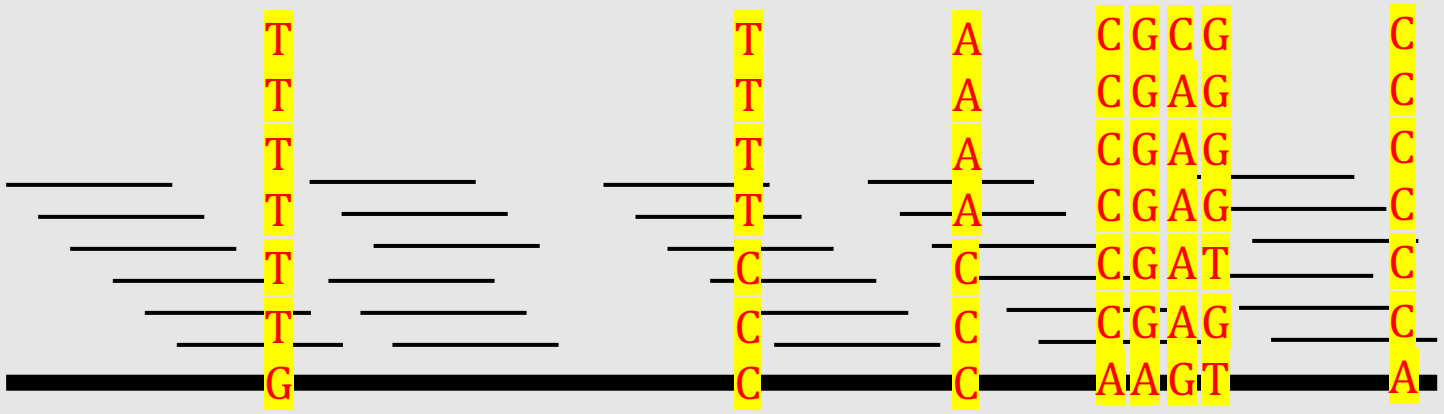


Software: BWA mem, samtools, Mauve…

Public Health Alliance for Genomic Epidemiology

Inferring phylogeny from SNPs – Overview

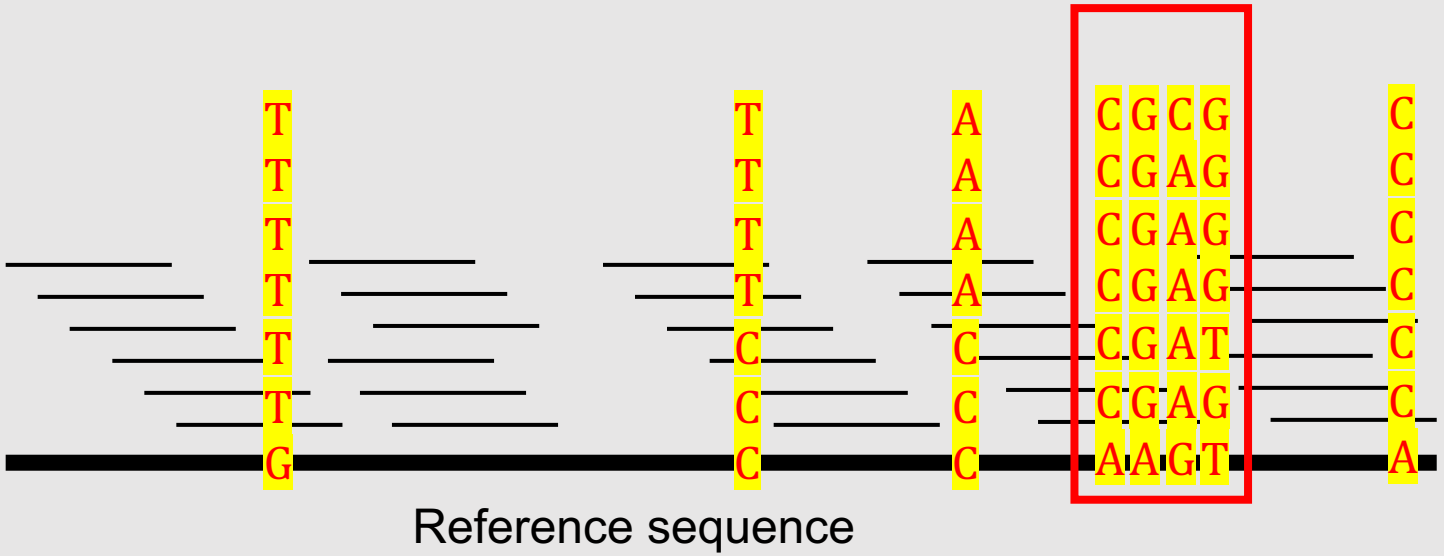
2. Call SNPs / variant calling and filtering of low quality SNPs



Software: Samtools, VarScan, bcftools…

Inferring phylogeny from SNPs – Overview

3. Filter recombination (optional)



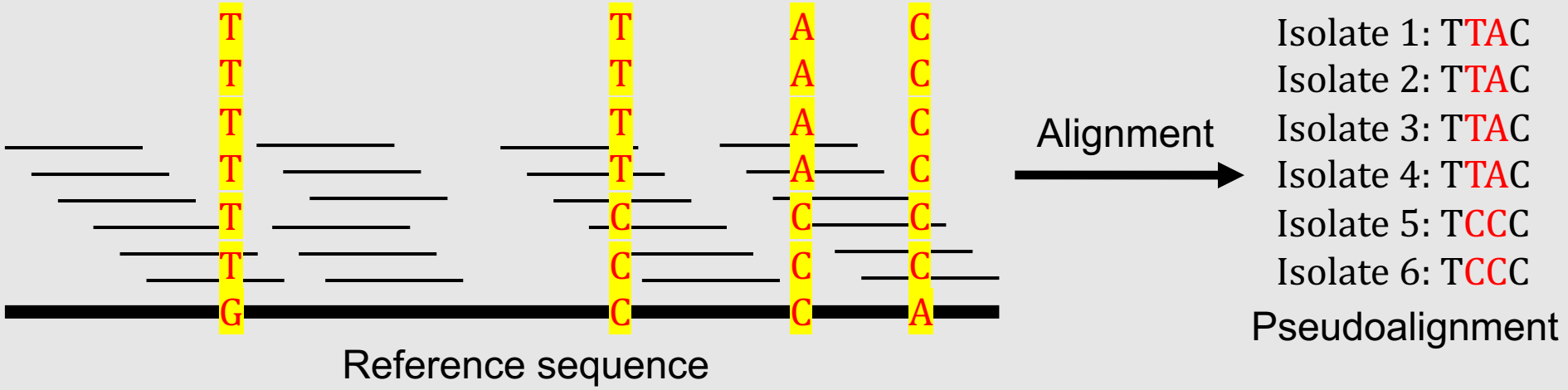
- Remove SNPs due to recent recombination events e.g mobile elements
- Often identified by relative SNP densities (clustering) along the alignment

Software: Gubbins…

Public Health Alliance for Genomic Epidemiology

Inferring phylogeny from SNPs – Overview

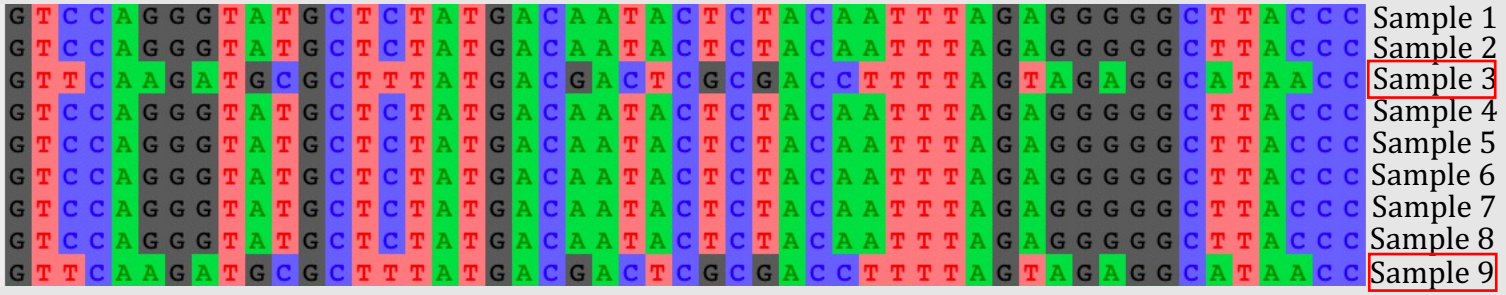
4. Concatenate filtered SNPs to create pseudogenomes and a pseudoalignment.
   - Constructed from only shared positions in all genomes and reference



Public Health Alliance for Genomic Epidemiology

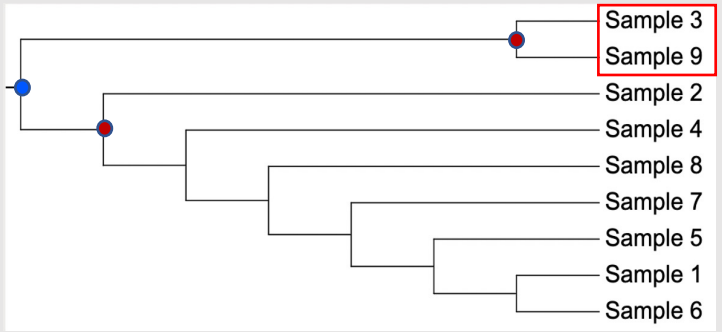
Inferring phylogeny from SNPs – Overview

5. Construct phylogenetic tree from pseudoalignment using chosen algorithm.



Pseudoalignment

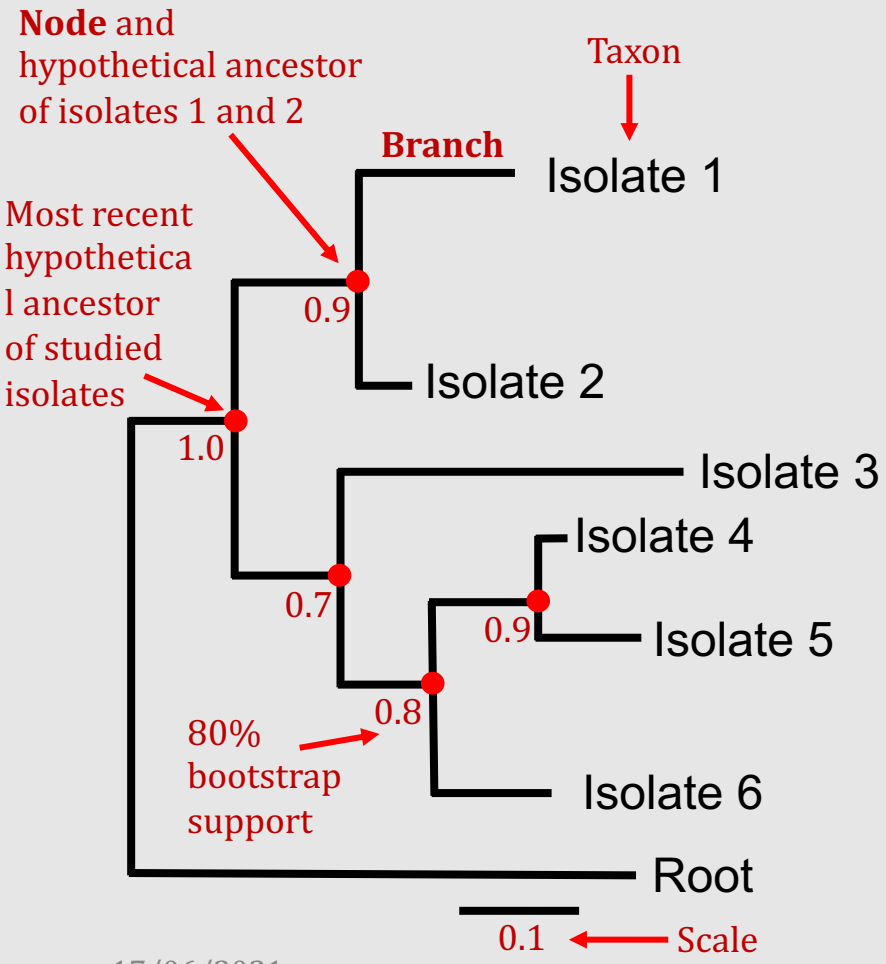
Tree construction



Software: IQ-Tree, FastTree

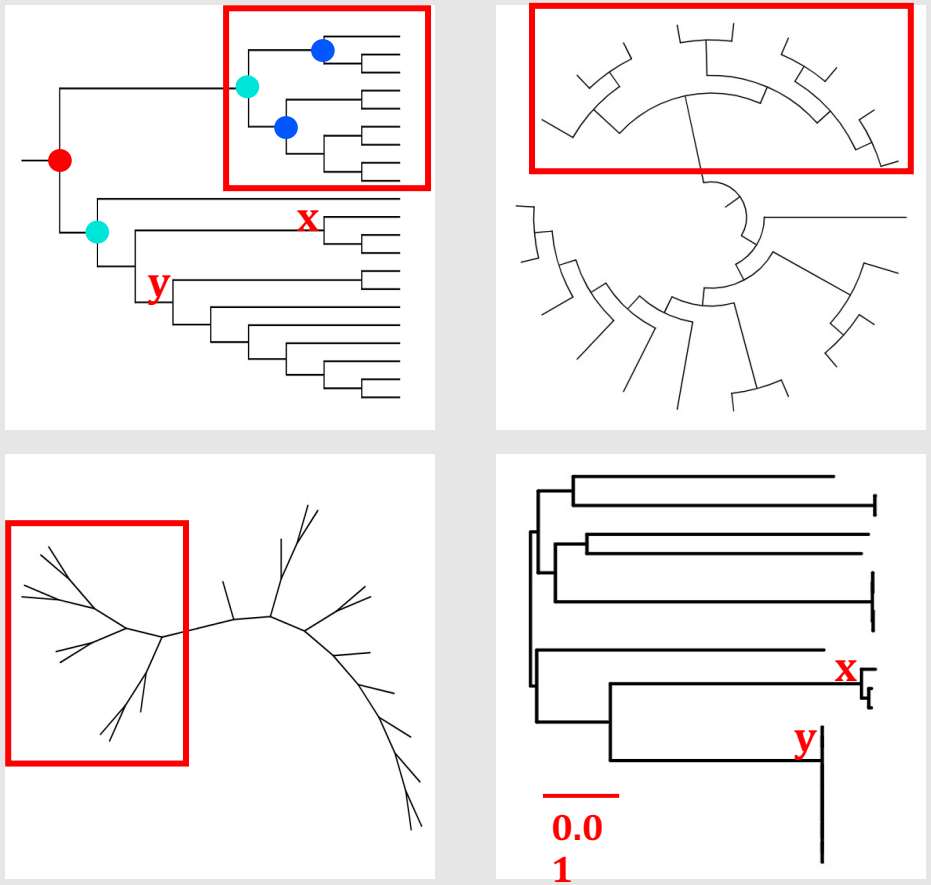
Public Health Alliance for Genomic Epidemiology

Phylogenetic tree interpretation



- **Taxa** are on the tree tips/leaves
- **Nodes:** hypothetical taxa ancestors
- **Branches** connect nodes and taxa
- **Topology:** branching structure of a tree
- **Clade:** taxa sharing a unique common ancestor
- **Bootstraps:** proportion of multiple replicate trees supporting each node.
- **Root:** distantly related to all taxa on the tree; tells direction of evolution

Phylogenetic tree interpretation

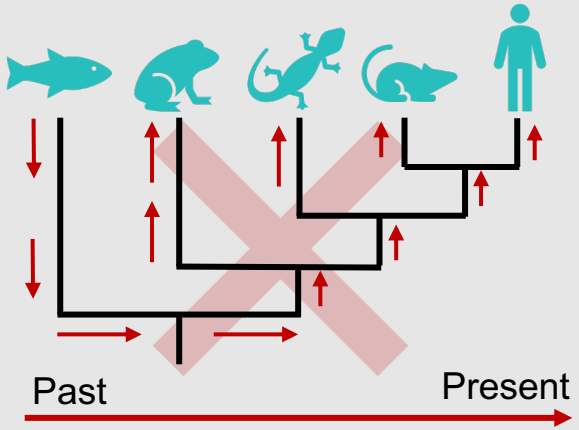


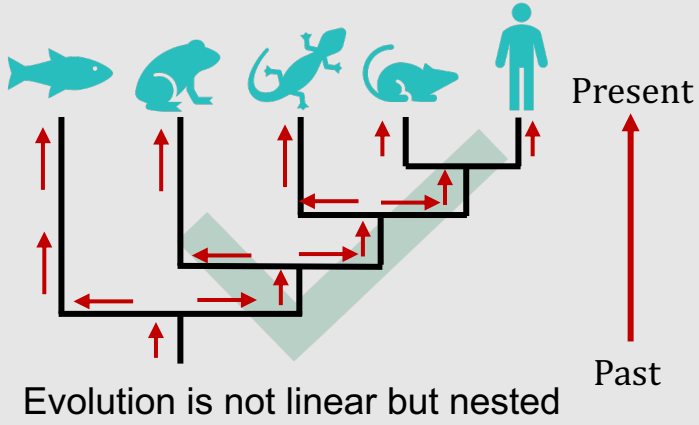
Reading phylogenetic trees

- Trees depict branching history; topology most important.
- Different orientations, same topology.
- Unless otherwise indicated, branch lengths are meaningless; avoid inferring temporal information that is not shown
- Branch lengths, when indicated, reflect the amount of evolutionary change as well as the passage of time.

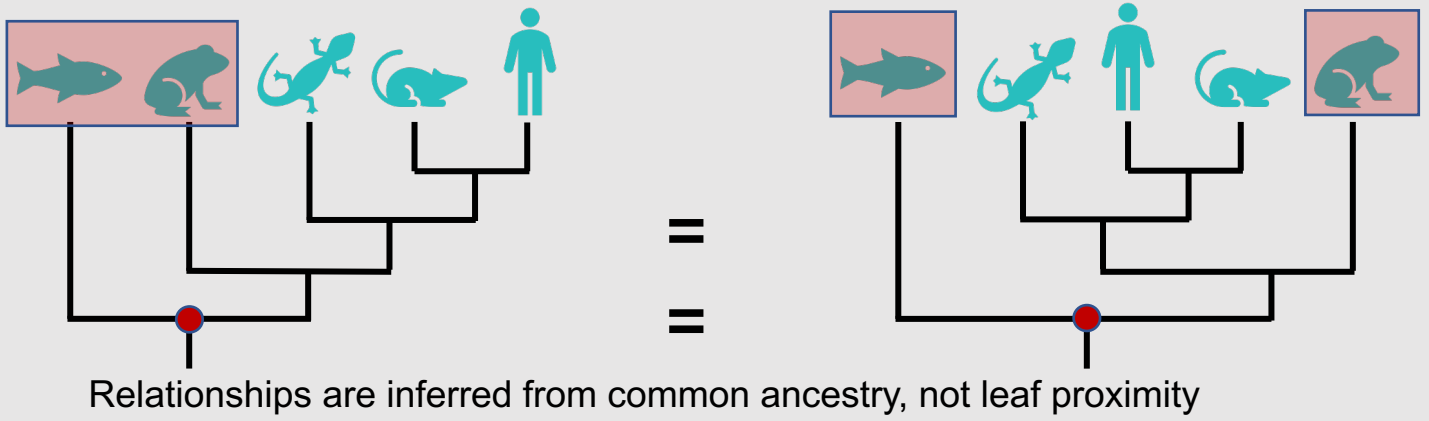
Phylogenetic tree interpretation: Misconceptions

1. Evolutionary timeline flows from tips/leaves





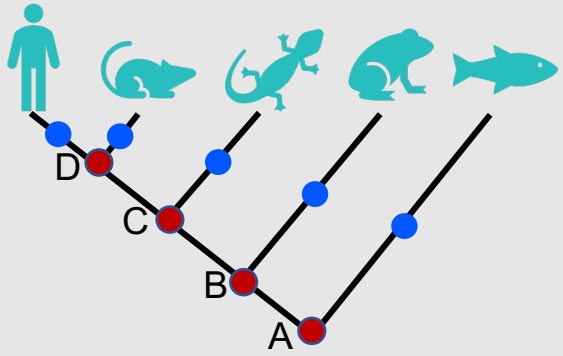
2. Related species have proximal leaves



Public Health Alliance for Genomic Epidemiology

Phylogenetic tree interpretation: Misconceptions

3. Genetic change occurs only at nodes



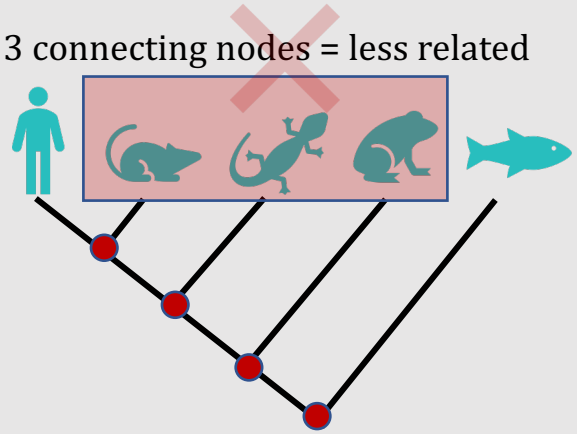
Fact:

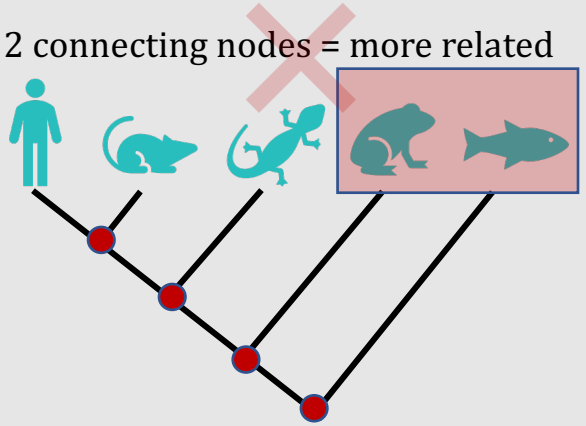
- Changes accumulate in taxa over time and are present along branches
- Nodes simply represent (hypothetical) common ancestors between taxa

● - Shared ancestor; divergence event

● - Unique ancestor

4. Related species have fewer connecting nodes





Public Health Alliance for Genomic Epidemiology

Phylogenetic tree visualisation: Microreact

- A flexible, interactive software/web application for easy visualization of datasets consisting of a combination of trees, maps, timelines, and associated metadata.
- Input:
  - Data file: Comma separated values (.csv) format; can contain a combination of textual metadata, locations and dates.
  - Optional tree file: Newick (.nwk) format.
- Output: interactive tree, map, timeline, and table.

Public Health Alliance for Genomic Epidemiology

Phylogenetic tree visualisation: Microreact

Sample data file

id	latitude	longitude	Country	Country__colour	Country__shape	Pedalism
Bovine	46.227638	2.213749	France	Red	square	Four
Gibbon	15.870032	100.992541	Thailand	Green	circle	Two
Orangutan	-0.589724	101.3431058	Sumatra	Blue	circle	Two
Gorilla	1.373333	32.290275	Uganda	#CC33FF	circle	Two
Chimp	-0.228021	15.827659	Congo	Orange	circle	Two
Human	55.378051	-3.435973	UK	#CCFF33	circle	Two
Mouse	40.463667	-3.74922	Spain	#00FFFF	square	Four

https://microreact.org/instructions

- **Mandatory column**

id – unique identifier for each data row

- **Optional columns**

latitude, longitude – geographic columns

year, month, day – temporal columns

<custom-name> – other metadata

<column-name>__colour – specify colour

<custom-name>__shape – specify shape

Public Health Alliance for Genomic Epidemiology

Phylogenetic tree visualisation: Microreact

Data file (.csv)

+/

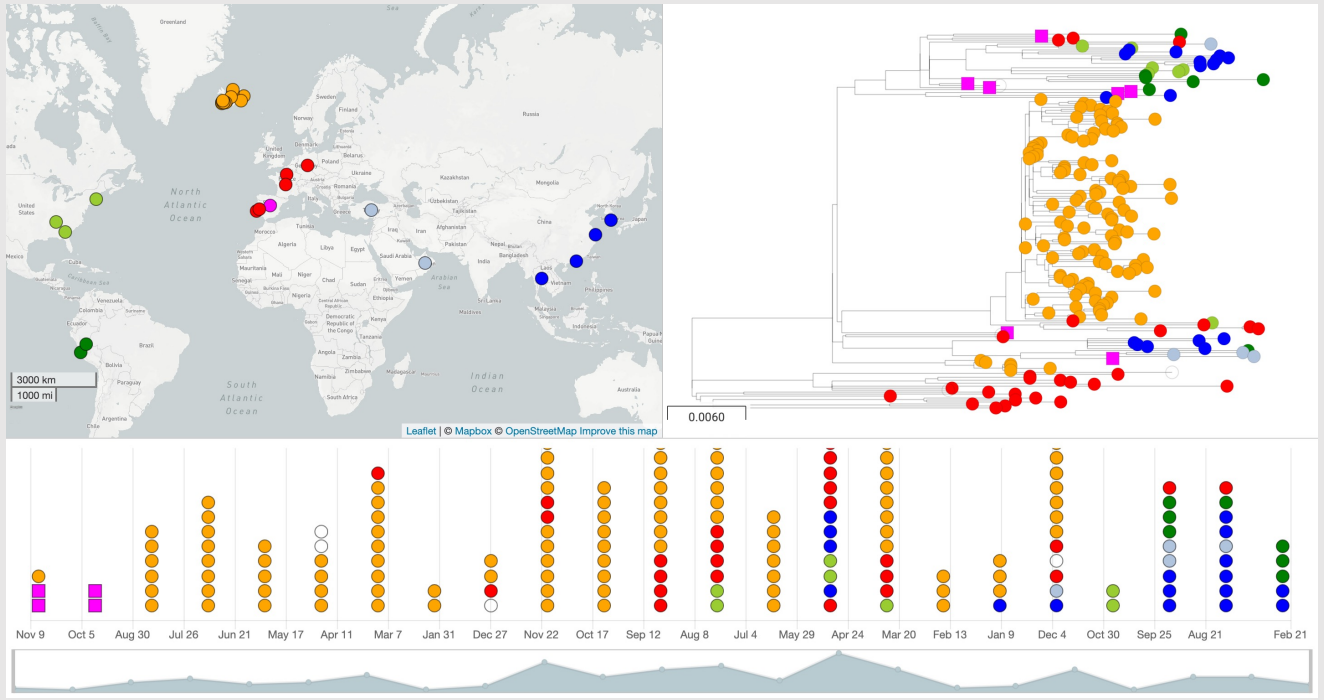
-

Tree file (.nwk)

Microreact

Web application

https://microreact.org



Source: https://microreact.org/project/N1TRn11L

Microreact benefits

- Visualization can be shared as a permanent web link among collaborators.
- The web link can be embedded within publications to enable readers to explore and download the data.
- Provides a simple, yet powerful, visualization method that aids research and discovery and facilitates the open sharing of datasets.

Public Health Alliance for Genomic Epidemiology

FUNDED BY
NIHR | National Institute for Health Research

Thank you

The Fleming Fund

Global Health Research Unit
Genomic Surveillance of AMR

RECTE SAPERE FONS