Introduction to Phylogenomics

Alan Christoffels

Phylogenetics

- Definition
  - The study of evolutionary relationships among organisms using molecular data:
    - DNA, protein sequences or other molecular markers
  - Evolutionary relationships include:
    - Reconstructing correct genealogical ties
    - Estimate time of divergence between organisms

Traditional methods for evo studies

- Morphology, anatomy
  - Used for taxonomy
- Physiology
- Paleontology
  - Time frame for evo studies

Molecular basis of evolution

Mechanisms of evolution

- By mutations of genes. Mutations spread through the population via:
- genetic drift: change in gene variant (or allele) in a population due to random chance.

  - Result in loss of genetic variation

  - Or a rare variant can become more frequent/fixed

Mechanisms of evolution

- By mutations of genes. Mutations spread through the population via:

- natural selection: ability of an organism to adapt to its surroundings; variation, inheritance, selection, time, adaption

Mechanisms of evolution

- By mutations of genes. Mutations spread through the population via:

- By gene duplication and recombination

Mutational changes of DNA sequences

1. Substitution.

Thr Tyr Leu Leu

ACC TA**A** TTG CTG

ACC T**C**T TTG CTG

Thr Tyr Leu Leu

3. Insertion.

Thr Tyr Leu Leu

ACC TAT TTG CTG

ACC TA**C** TTT GCT G—

Thr Tyr Phe Ala

2. Deletion.

Thr Tyr Leu Leu

ACC TAT T**T**G CTG

ACC TAT TGC TG-

Thr Tyr Cys

Types of Molecular data used for phylogenetic analysis

Molecular Data

- Characters
  - Provide info about individual OTUs

- Distances
  - Represent quantitative statement concerning dissimilarity between two OTUs

Characters

- Independent variables
  - Height, amino acid position

- Character state
  - Value of a character in a OTU
    - Alanine

  - Multi-state characters are **positions** in aligned sequences and
  - nucleotides are the character **states** at the positions

Distance Data

- Involve pairs of taxa
- Sequence data ----> character data
  - Transform into distances
    - Substitutions per site between two sequences
- % dis-similarity
- Multiple substitutions at a site

Measures of evolutionary distance between amino acid sequences.

$$p = n_d / n$$

n_d – number of amino acid differences between two sequences; n – number of aligned amino acids.

P-distance

Measures of evolutionary distance between amino acid sequences.

	P-distance	
Human/cow	**0.121**	
Human/kangaroo	**0.186**	
Human/carp	**0.486**	

* Non-linear relationship between distance and time???

Poisson correction for evolutionary distance.

2. PC-distance. Assumes rates of amino acid substitutions are equal among sites while correcting for multiple substitutions at the same site

PC-distance can be expressed through the p-distance:

$$d = -\ln(1 - p)$$

Estimation of evolutionary rates in hemoglobin alpha-chains.

	P-distance	PC-distance
Human/cow	**0.121**	**0.129**
Human/kangaroo	**0.186**	**0.205**
Human/carp	**0.486**	**0.665**

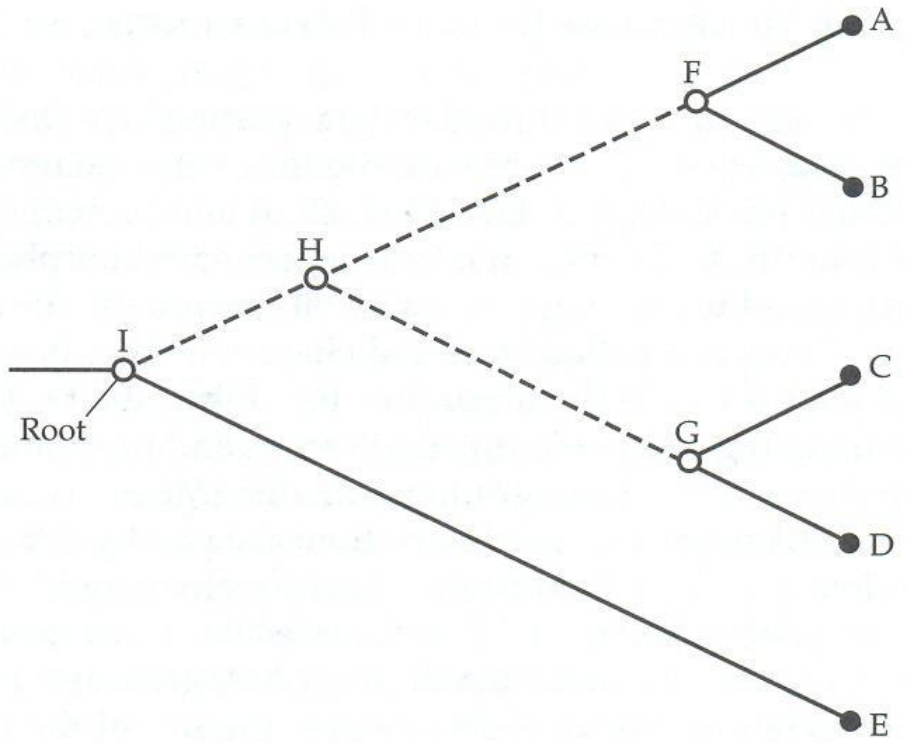
Phylogenetic tree

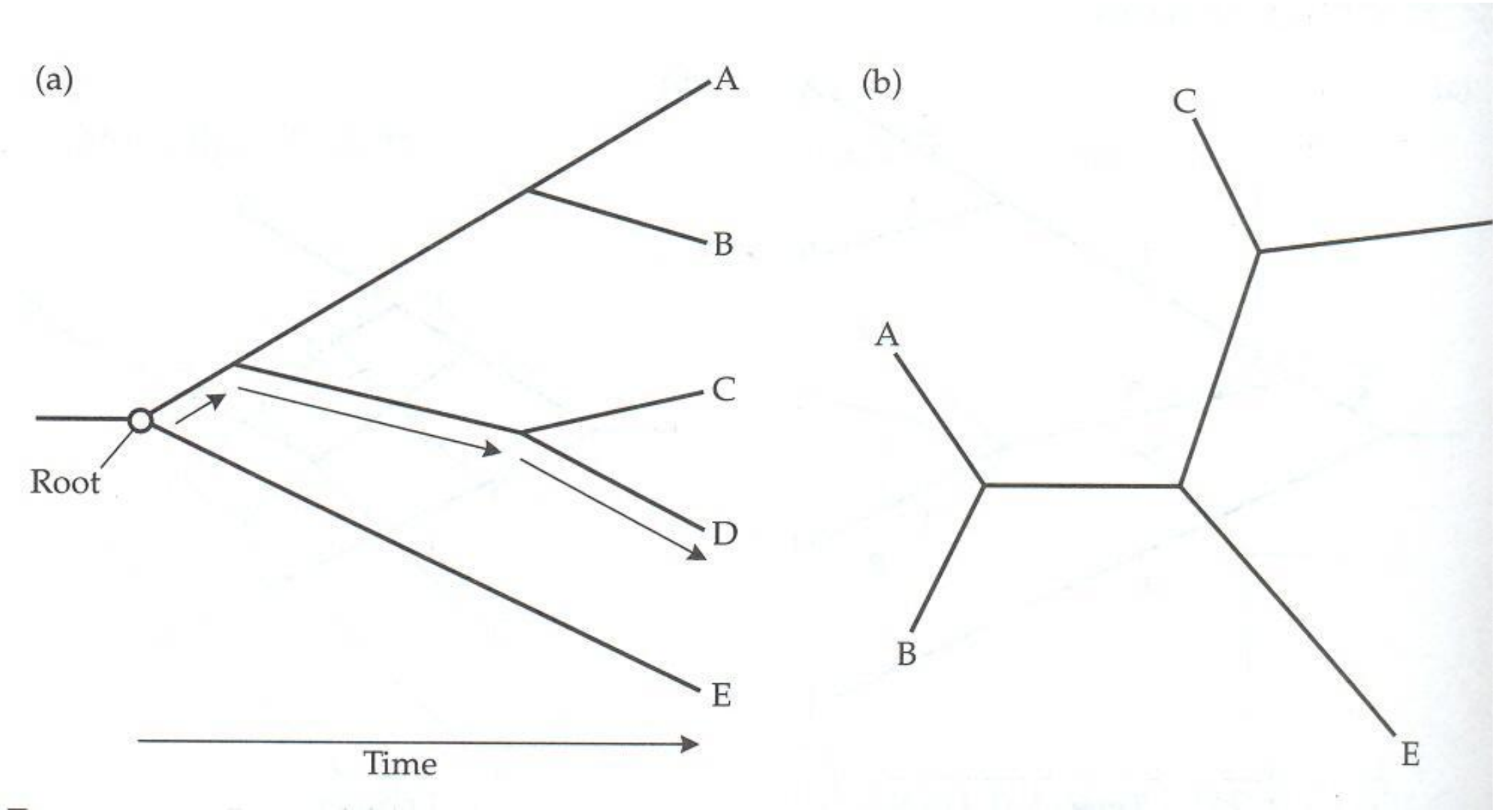
- Illustrates evolutionary relationship among a group of organisms

- Graph composed of nodes and branches
  - Nodes: taxonomic units (genes etc)
  - Branch connects two adjacent nodes
    - Defines relationships among TUs according to ancestry and descent
    - Branching pattern of tree -----topology

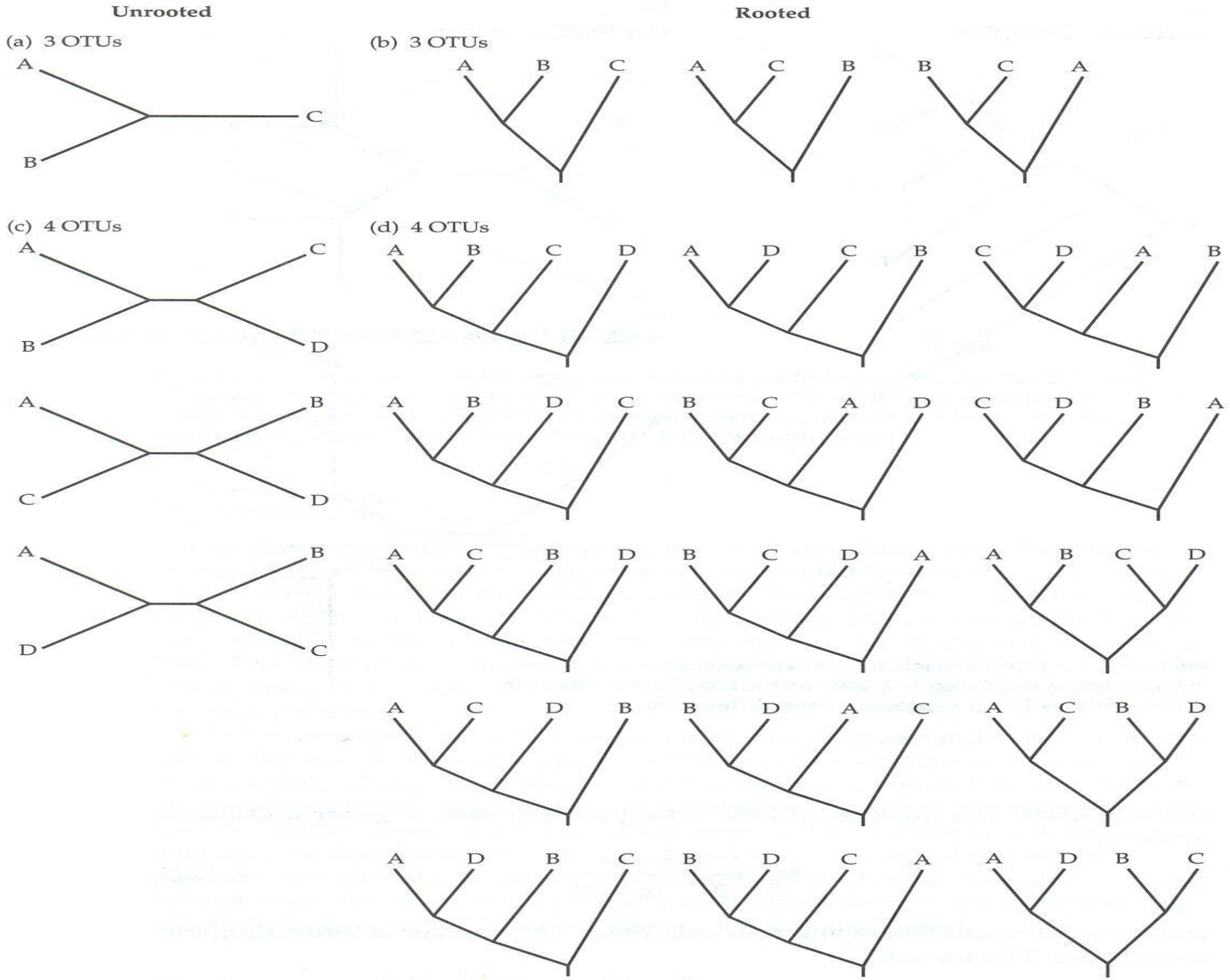
Phylogenetic tree

- Nodes
  - Terminal node:
    - Extant TU or operational taxonomic unit
  - Internal node:
    - Inferred ancestral unit
    - No empiral data




(a)
A
B
C
D
E
Root
Time
(b)
C
A
B
E

Counting number of trees?


Unrooted
Rooted
(a) 3 OTUs
(b) 3 OTUs
(c) 4 OTUs
(d) 4 OTUs
A
B
C
D

Number of Trees?

- Rooted trees for n OTUs (n >=2)

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Number of Trees?

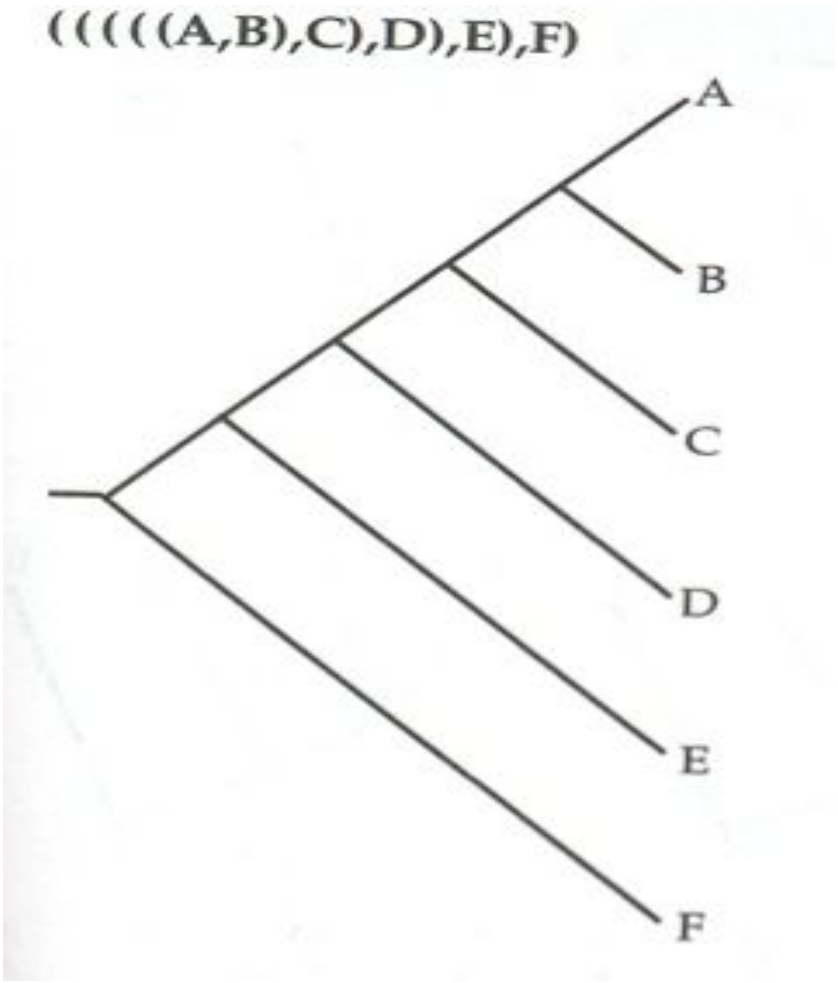
- Unrooted trees for n OTUs (n >=3)

$$\frac{(2n-5)!}{2^{n-3}(n-3)!}$$

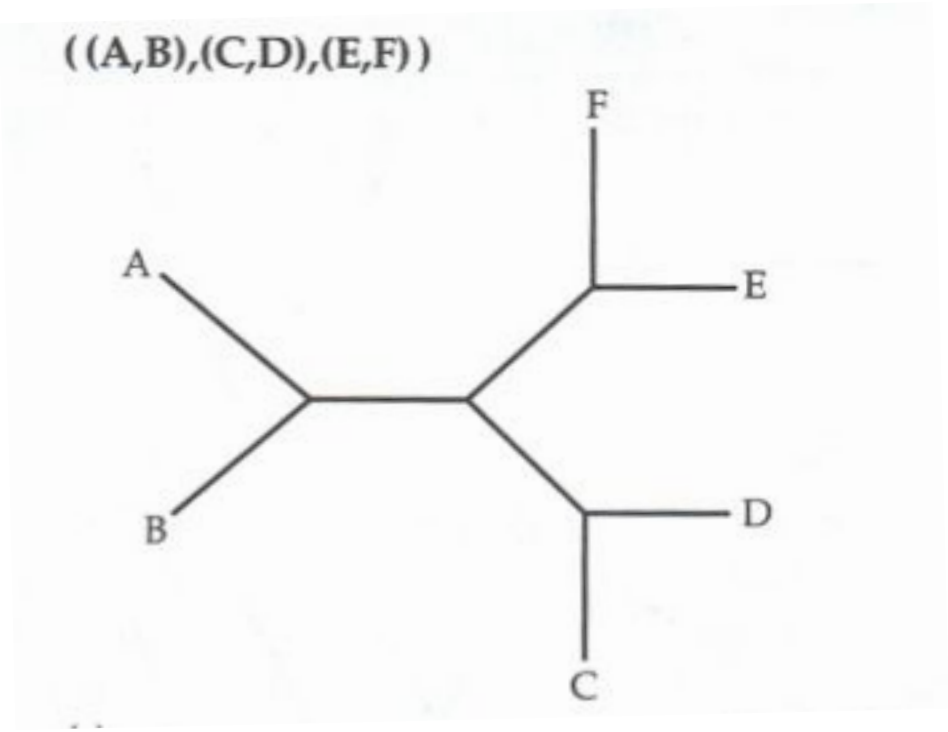
Newick Format

- Trees represented in linear form
  - Nested parentheses enclosing names and separated by commas
  - Used in computer programs
  - First used by Cayley (1857)
  - Adopted by informal standards committee of Society for Study of Evolution (1986)
  - Parentheses pattern ----> topology

Newick Format



Newick Format



Phylogenetics to trace disease outbreaks

- Outbreak
  - Occurs when a disease spreads through a population.
  - When it reach a certain size then it gets a new name (epidemic)
  - If the outbreak remains in a population for a long time then it becomes endemic
- When an epidemic spreads to multiple countries or continents then it is called a pandemic.

Phylogenetics to trace disease outbreaks

- When an outbreak is first detected
  - Rush by scientists to find out where it comes from to stop its spread
  - Build an epidemiological history of the detected cases and where they were found

- Trace and trace – track the new cases and limit their exposure with others

  - Works well in a small population but when the scale increases then we use genome sequencing

Phylogenetics to trace disease outbreaks

- Pathogens – genome changes over time is used to track and trace
- Place the genome of each isolate onto a phylogenetic tree
- Sequences that sit near to each on the tree are similar and share a common source of infection
- A new branch will be formed when sequences are dissimilar due to an increase in mutations in some of the sequences - ---> variants
- more sequences the better the resolution in tracing the outbreak

Phylogenetics to trace disease outbreaks

- Phylogenomics : intersection of genomics and phylogenetics. Analysis that involves genomic data and evolutionary reconstruction.

Evolutionary history of SARS-COV-2: Rate of evolution of time

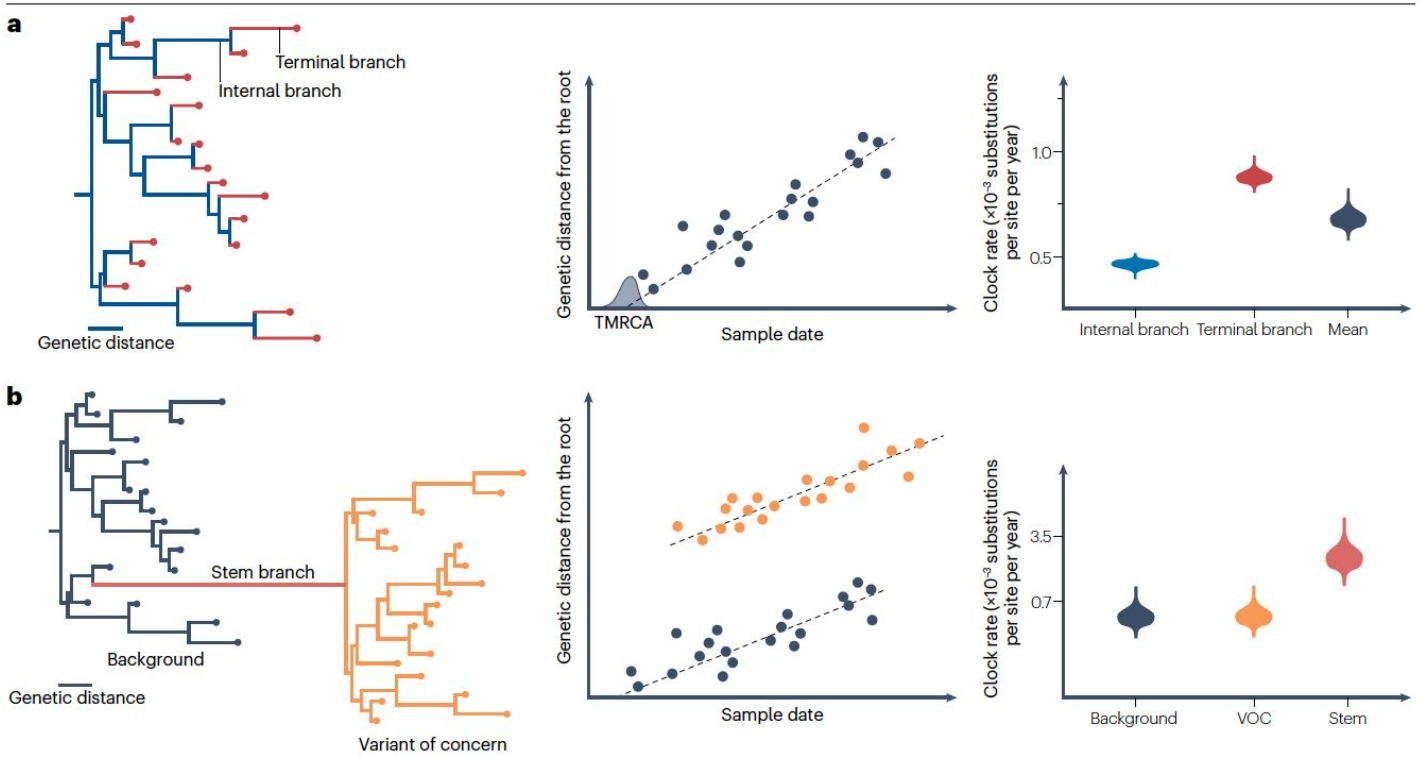


Fig. 1 | Reconstructing the evolutionary history of SARS-CoV-2 and its rate of evolution over time using viral sequence data. a, As the virus spreads and acquires further evolutionary changes, sequence data from different time points can be used to infer viral evolutionary rates as well as the time to the most recent common ancestor (TMRCA) of sampled viruses. The rate of evolution is elevated at terminal branches (shown in red) relative to internal branches (shown in blue). This is because the former includes more deleterious mutations that only persist in the population for a limited period of time before they are removed by purifying selection. **b**, If the genomic samples include both variant of concern (VOC) and non-VOC (background) sequences, the evolutionary rate at the stem branch (shown in red) connecting the clade of the VOC sequences to the rest of the tree is elevated in comparison with the rate on all other branches in the entire phylogeny. This is because of the large number of evolutionary changes present in all VOC samples relative to the background.

Nature review microbiology. 21, 361-379 (2023)

Evolution from Within hosts to between hosts

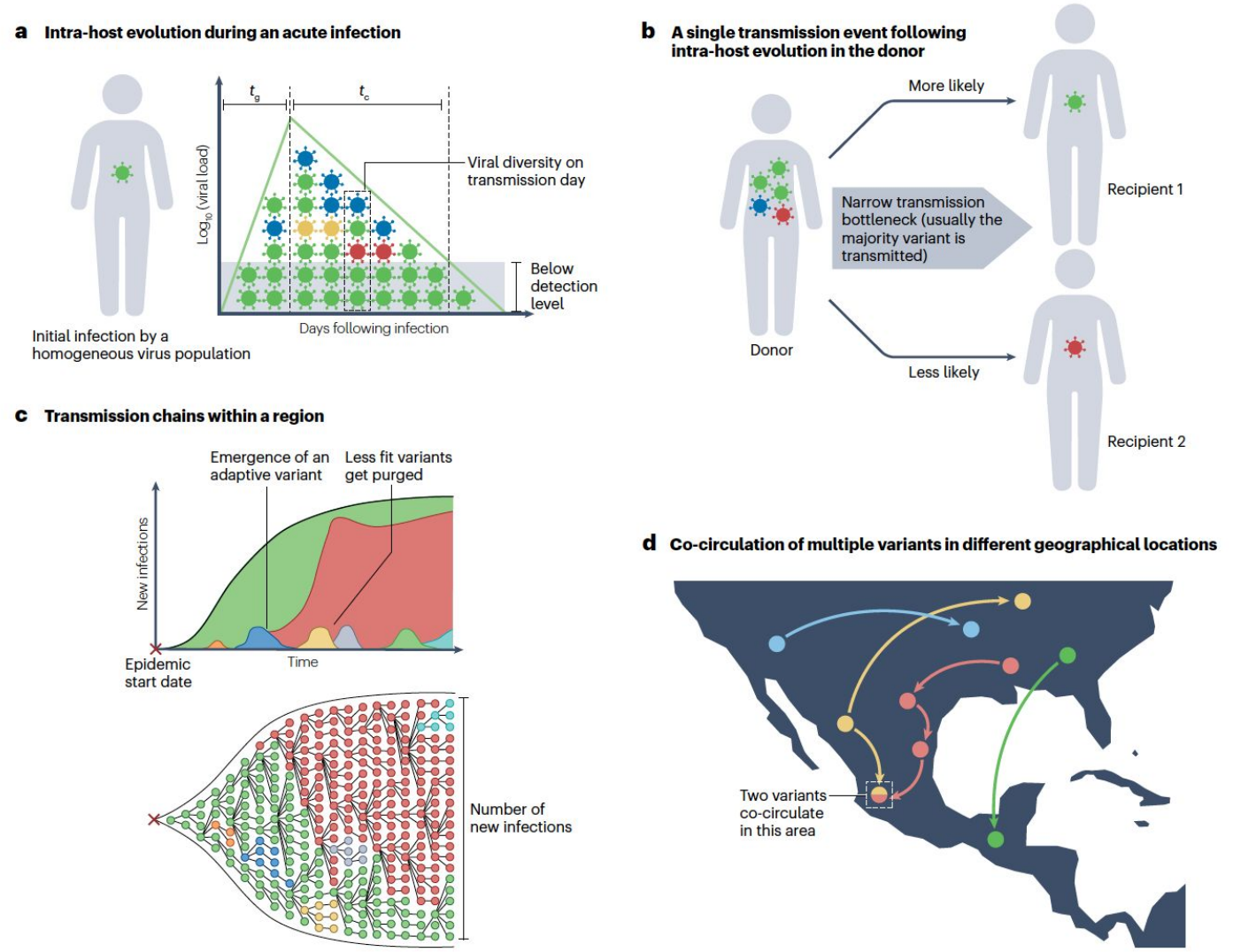


Fig. 2 | Levels of evolution from within hosts to between hosts. a, At the start of an acute infection, viral load increases rapidly for a period of time (t_g) before the immune system begins to clear the infection, at which point the viral load

population level. **c**, As the number of new infections grows over the course of an epidemic wave, new variants with selective advantage may emerge and reach high frequencies. Others, carrying deleterious mutations, get purged from the

Transmission bottleneck

- The infinitesimal number of viral particles that establish the viral population in a new host on transmission. Usually, these are a minuscule and often genetically unrepresentative sample from the virus population in the original host, which contributes to genetic drift.

Substitution Rate

- Also known as evolutionary rate. The rate at which new mutations accumulate in a viral population, usually measured per nucleotide site per year

Mutation rate

- Mutation rate: The probability of mutation, usually measured per nucleotide per replication cycle. Where a mutation refers to the change in the genome of a virus caused by errors during replication

- Mutations arise due to replication errors introduced by the virus' polymerase enzyme. Rate is lower than other RNA viruses (Hepatitis virus or HIV).

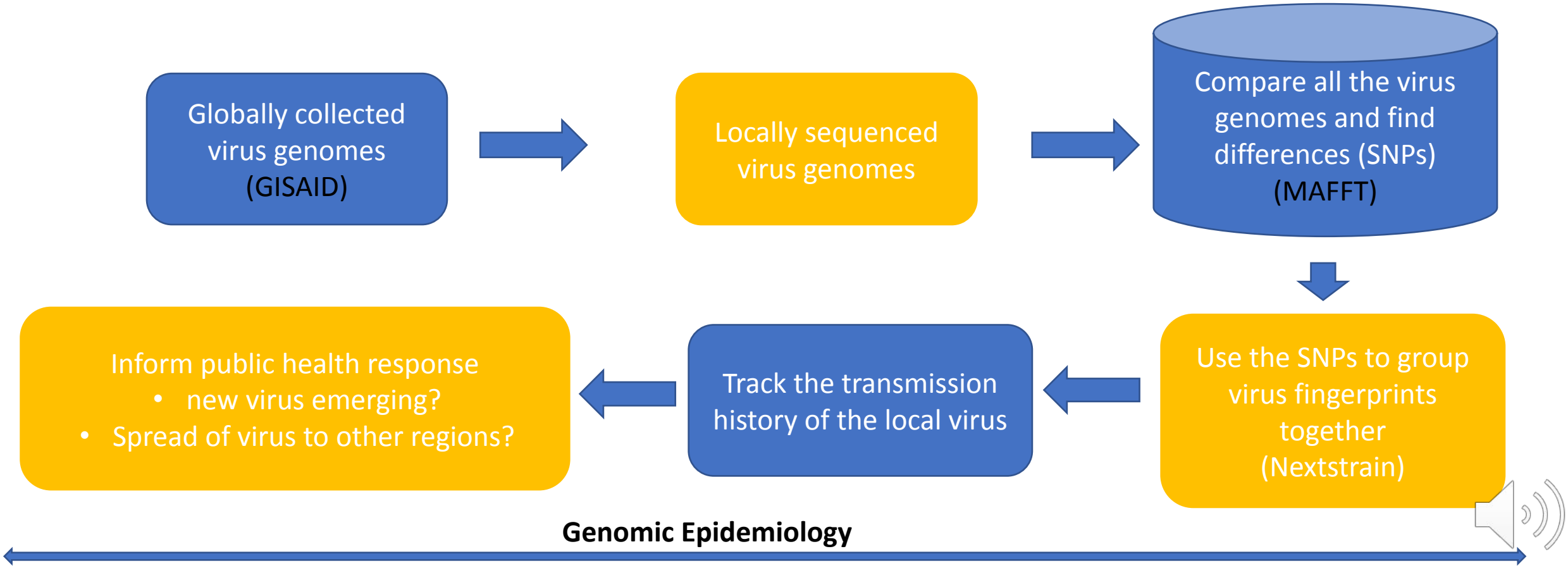
- **Replication errors** - source of genetic diversity (S-gene dropout: del in *Spike* gene used to detect Alpha variant)

Host mediated editing

- innate cell defence mechanisms may introduce substantial numbers of directed mutations into the SARS-CoV-2 genome
    - Alter evolutionary rate
- apolipoprotein B mRNA editing enzyme (excess of C → U transitions)
- SARS-CoV-2 genomes may also be edited by different cellular antiviral proteins
    - adenosine deaminases.. leading to A → G mutations

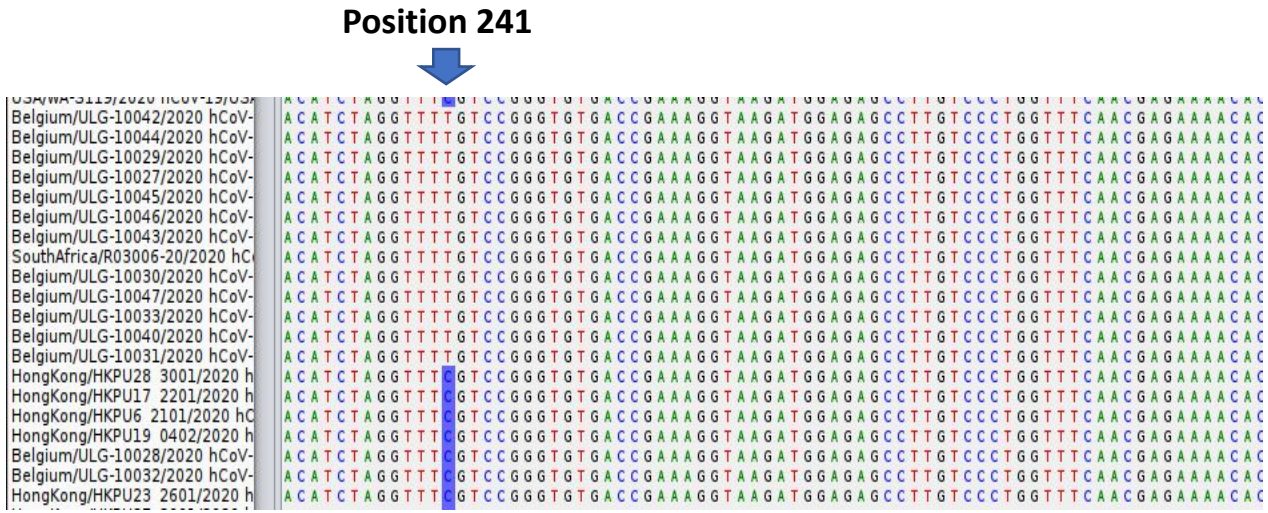
The role of pathogen genomics

How do I use this data to control COVID19 spread?



The role of pathogen genomics

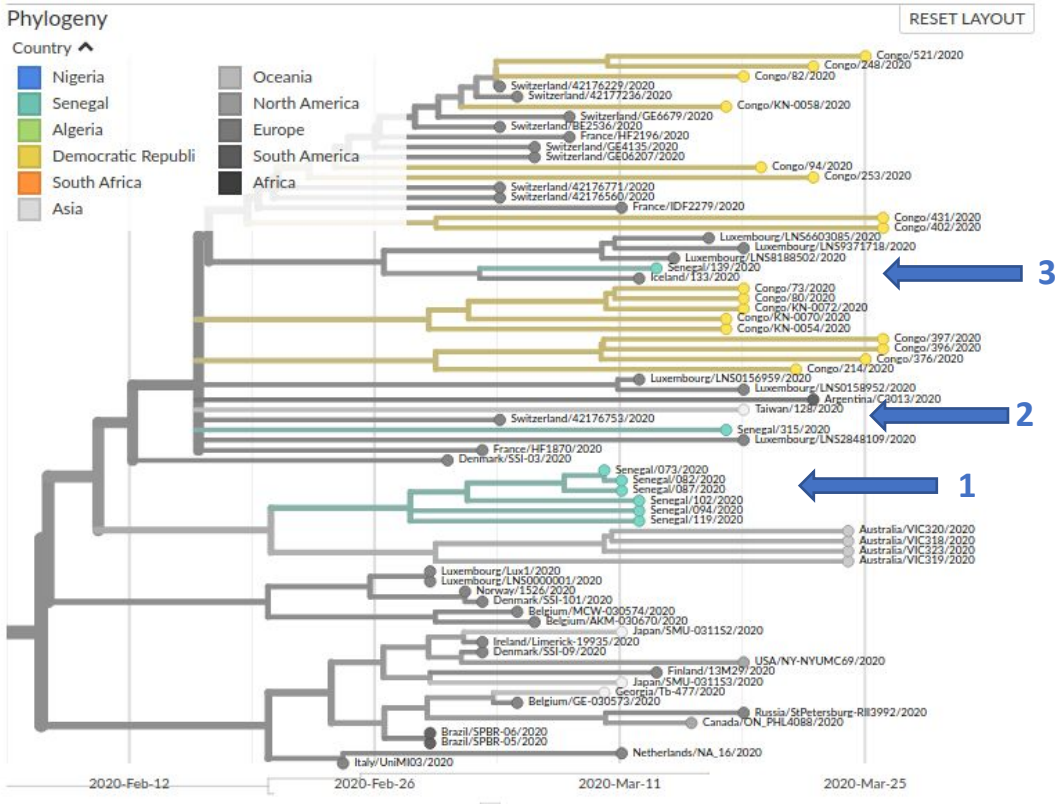
An example of using COVID19 data for genomic epidemiology?



Alignment of 3069 virus sequences
Belgium and SA sequences have a ”T” at position 241
Hongkong Sequences have a “C” at pos.241

The role of pathogen genomics

An example of using COVID19 data for genomic epidemiology?



Multiple clusters from Senegal – Suggests that the virus was introduced multiple times (pink arrows)