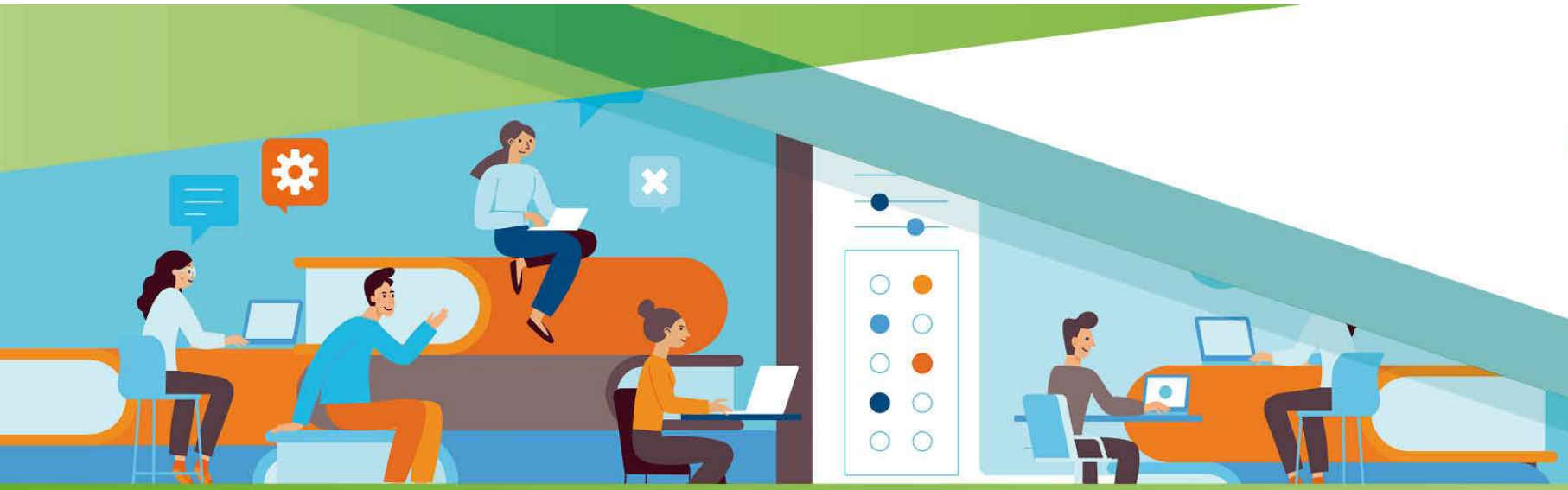


Phylogenetic analysis

GenEpi-BioTrain Virtual Training 29 - Genomic epidemiology of mpox virus (MPXV)



5 October 2026

Dr. David F. Nieuwenhuijse

Aim and learning objectives

Aim

Recognize the importance of reference selection and multiple sequence alignment curation and perform and describe the key concepts of phylogenetic analysis

Specific learning objectives

At the end of this session, you will be able to:

1. Find reference sequence data
2. Construct and evaluate a multiple sequence alignment
3. Use tools to perform phylogenetic tree construction
4. Interpret the patterns of a phylogenetic tree

Outline

This session consists of the following elements

1. Finding reference sequence data
2. Performing multiple sequence alignment
3. Key concepts of phylogeny
4. Performing phylogenetic analysis
5. Interpreting patterns in a phylogenetic tree

Selecting Reference Sequences

- Crucial for the result of the analysis
- What sequences to select depends on the research question
- More sequences is not always better
- Reference databases contain (many) low-quality sequences

Classified as ECDC NORMAL

(Open) viral genome reference data sources

INSDC (ENA/GenBank/DDBJ)	DDBJ National Institute of Genetics; INSDC; GenBank NCBI; ENA European Nucleotide Archive EMBL-EBI	Fully open access
GISAID	GISAID	Free-registration required (terms apply)
Pathoplexus (new)	Pathoplexus	No registration (terms apply)

Classified as ECDC NORMAL

Do not edit
How to change the design

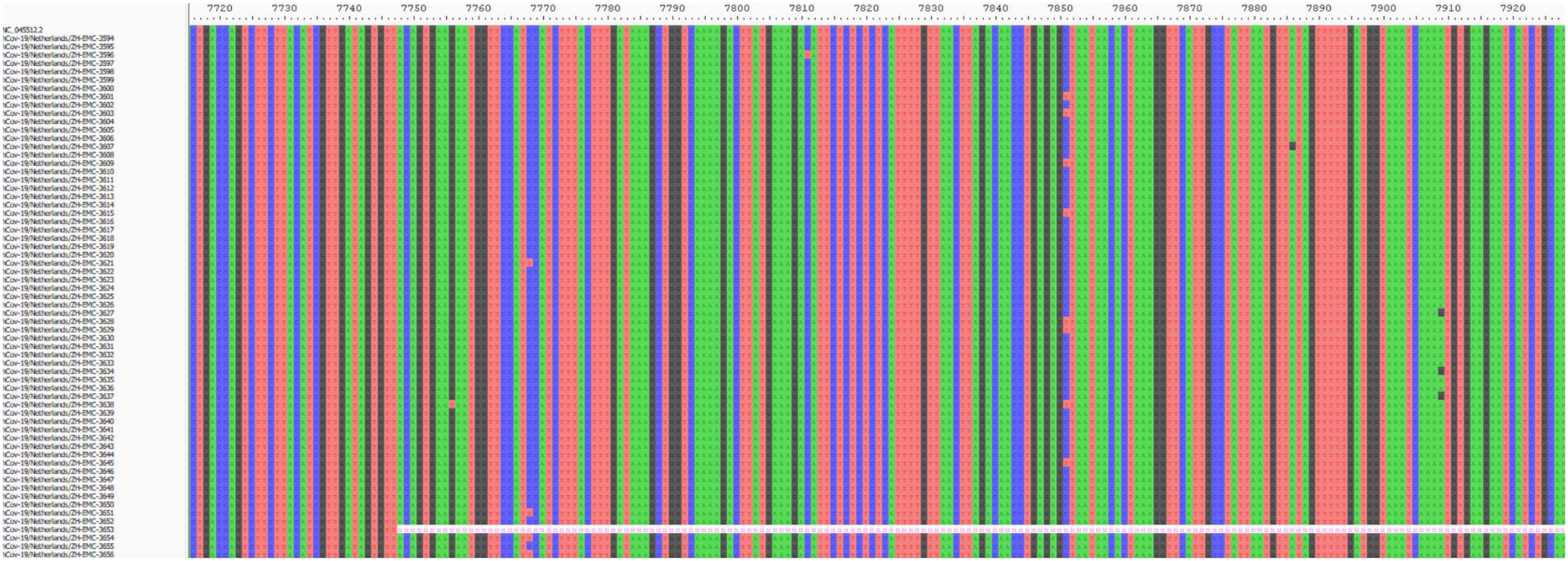


From which of these databases have you downloaded virus sequences before?

ⓘ The Slido app must be installed on every computer you're presenting from

slido

Multiple sequence alignment (MSA)



Classified as ECDC NORMAL

MSA software

- MAFFT, MUSCLE, Clustal Omega, MEGA
- Software makes best "scoring" alignment
- May not be biologically meaningful → manual inspection

Classified as ECDC NORMAL

Mpox specific: Reference guided MSA

- Large viral genomes are hard to align
- Variable size insertions and deletions make it even more complex
- Reference guided MSA fixes this
- Insertions compared to the reference are ignored
- No problem for phylogenetic analysis

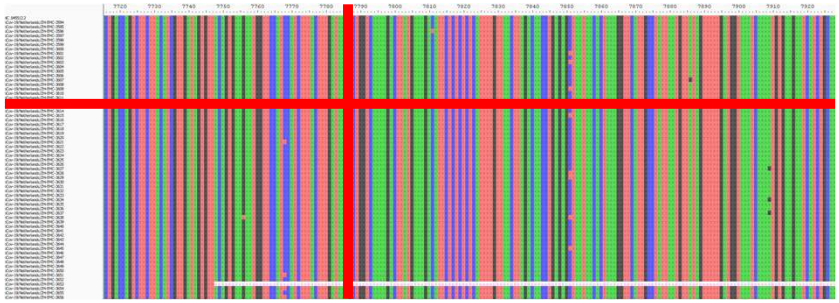


https://github.com/aineniamh/squirrel
https://github.com/virus-evolution/gofasta

Classified as ECDC NORMAL

Manual MSA curation

- Rule of thumb: remove complete sequences or mask complete columns (N)
- Careful with editing individual positions
- Things to look for:
  - SNPs next to positions with Ns
  - Clustered SNPs
  - Positions with non-ACTG nucleotides
  - Uncommon deletions/insertions (although they are ignored in phylogeny)



Classified as ECDC NORMAL

Mpox specific: Reproducible sequence masking

- Masking:
  - Repetitive regions
  - Inverted terminal repeat regions
  - Error rich regions
- Masking strategies:
  - Community driven
  - WHO official mask
  - Custom masking strategies
- Automatic masking software: Squirrel



https://github.com/aineniamh/squirrel

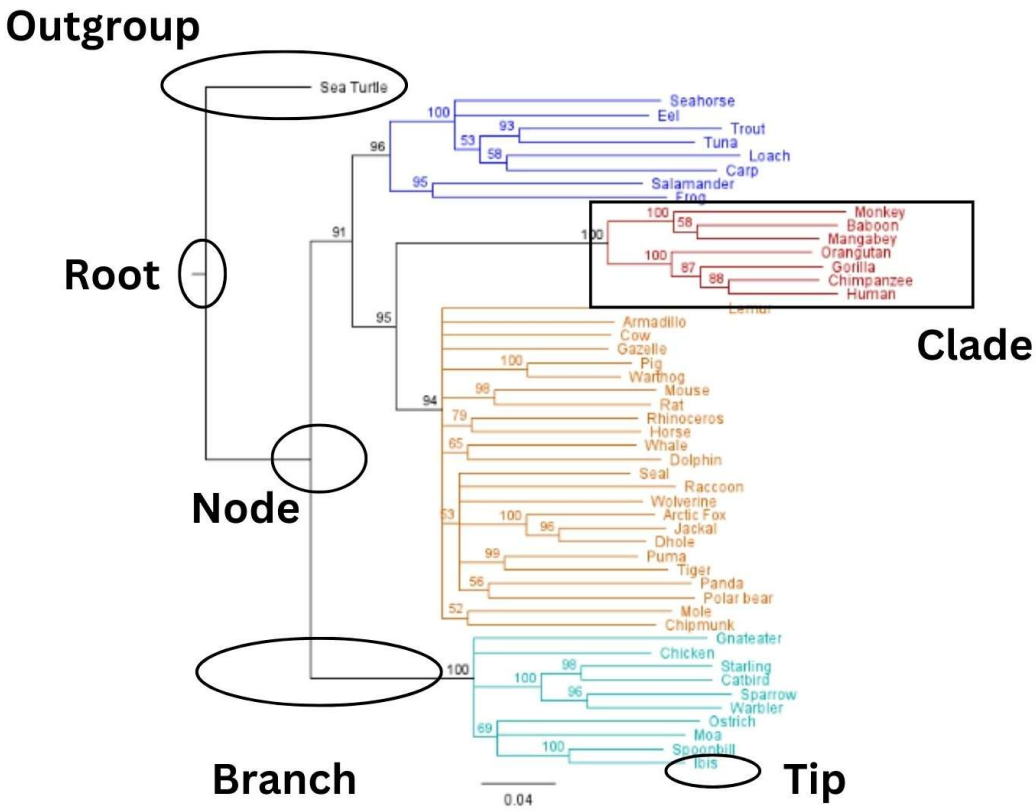
Classified as ECDC NORMAL

What is a phylogeny?

A phylogeny (or phylogenetic tree) is a diagram of evolutionary relationships

Phylogenetic trees are hypotheses; they model the real evolution based on (limited) sampling of evolutionary history

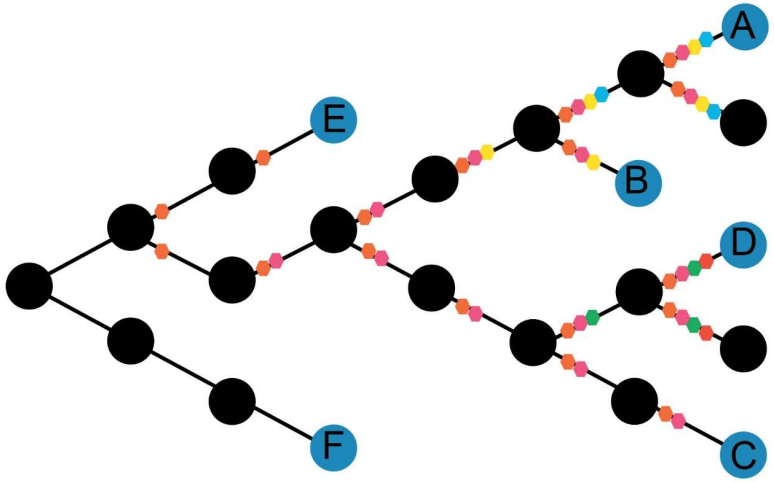
In genetic phylogeny the samples are genetic sequences



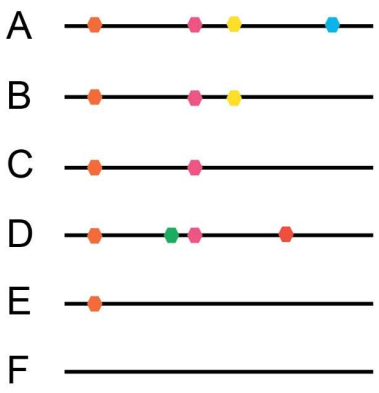
Classified as ECDC NORMAL

Infectious disease phylogenetic analysis

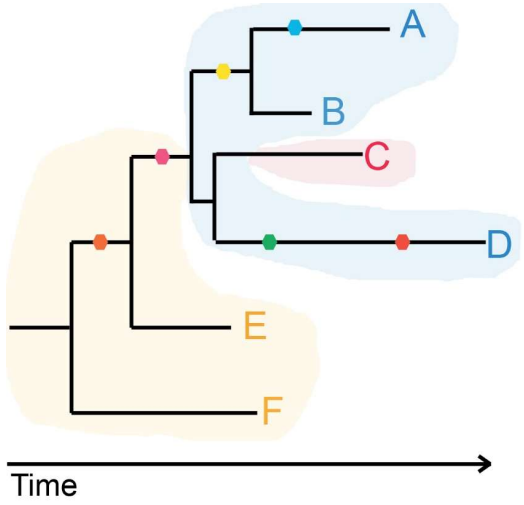
Transmission process



Sequences from samples



Reconstruction



⚠ Insertions and deletions are generally treated as missing data

Classified as ECDC NORMAL

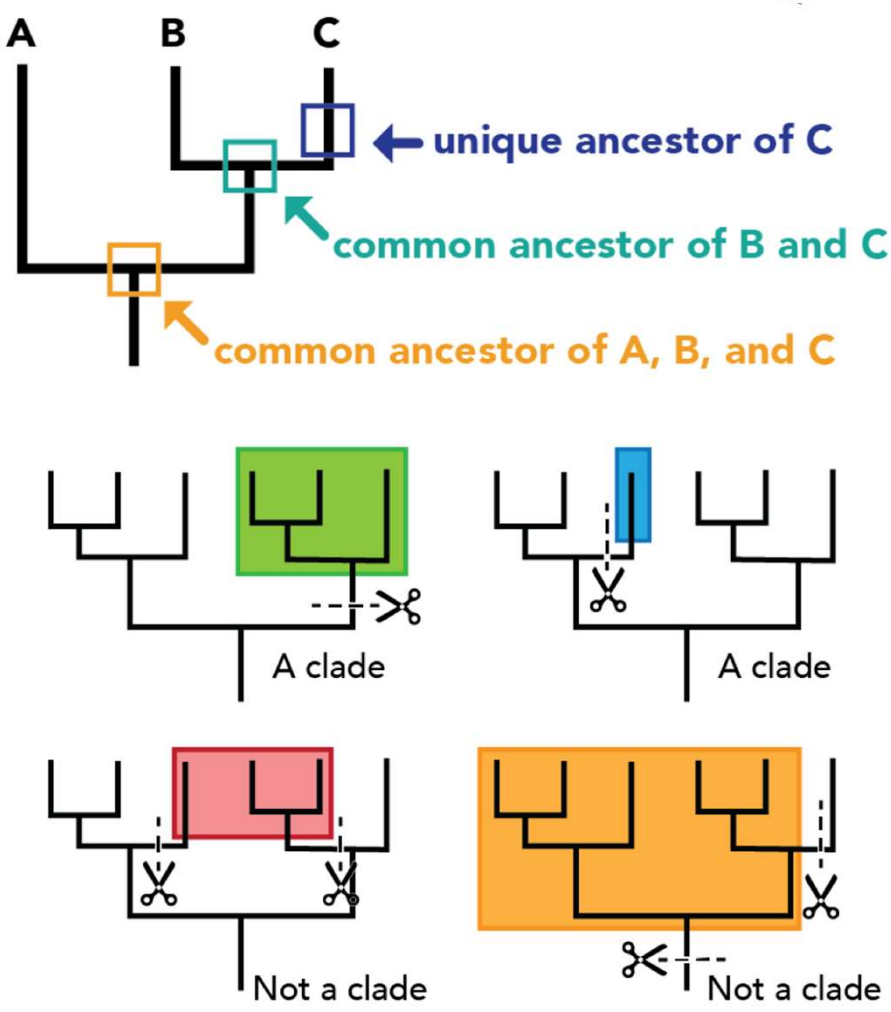
What is a clade?

A clade is a common ancestor plus all its descendants

Imagine clipping a single branch off the tree; everything on that pruned branch is a clade

Where the clade begins depends on clade definition decisions made by the community

A "cluster" is commonly a clade with highly similar sequences, indicating a single outbreak "cluster"



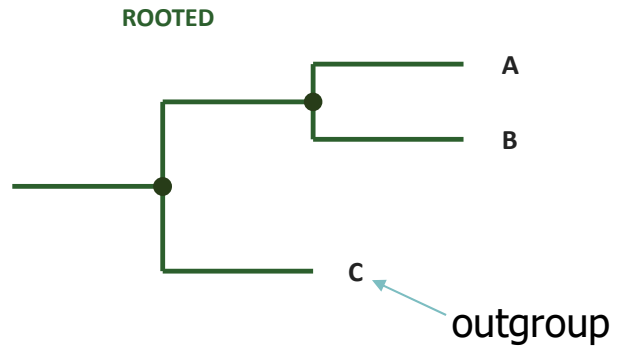
Classified as ECDC NORMAL

Rooted vs. unrooted trees

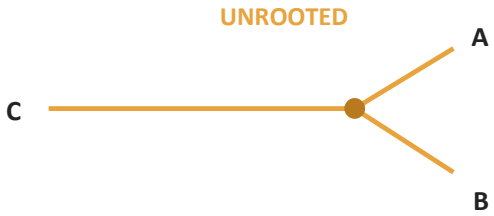
A rooted tree has a common ancestor (root) for all its tips

An unrooted tree shows the same relationships, but without a starting point

Rooting usually relies on an outgroup known to fall outside the group of interest, such as a known old sequence or the sequence of another species



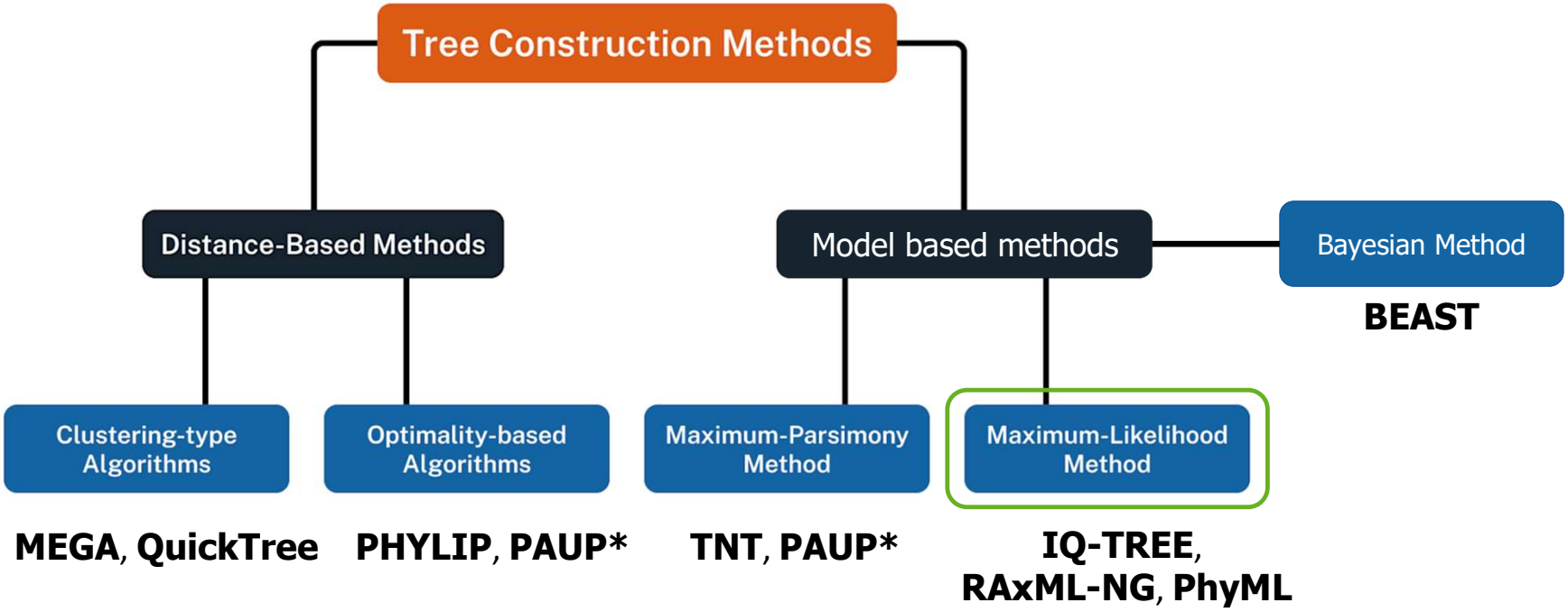
has a starting point — shows history



same relationships — no direction

Classified as ECDC NORMAL

Phylogenetic analysis



Classified as ECDC NORMAL

Maximum-Likelihood phylogeny

1. For a given tree topology, find optimal branch lengths (given the observed data) and substitution model parameters
2. Obtain a tree topology that maximizes the likelihood given branch lengths and substitution model parameters.

Classified as ECDC NORMAL

Evolutionary models

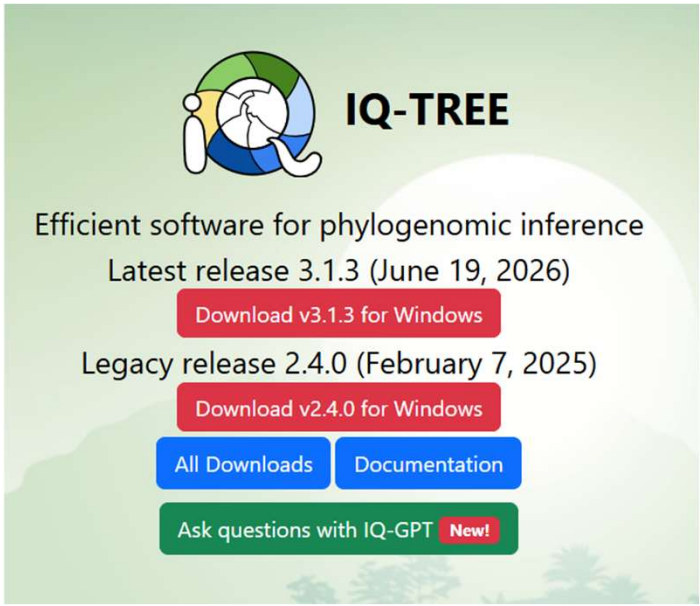
- Many models have been developed (https://iqtree.github.io/doc/Substitution-Models)
- Models describe substitution rates, frequencies and mutation rate heterogeneity
- There are nucleotide and amino acid models
- IQ-TREE can select the best fitting model automatically

Model	df	Explanation
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).
TNe	2	Like `TN` but equal base freq.
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).
K81u	5	Like `K81` but unequal base freq.
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.
TPM2u	5	Like `TPM2` but unequal base freq.
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.
TPM3u	5	Like `TPM3` but unequal base freq.
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.
TIMe	3	Like `TIM` but equal base freq.
TIM2	6	AC=AT, CG=GT and unequal base freq.
TIM2e	3	Like `TIM2` but equal base freq.
TIM3	6	AC=CG, AT=GT and unequal base freq.
TIM3e	3	Like `TIM3` but equal base freq.
TVM	7	Transversion model, AG=CT and unequal base freq.
TVMe	4	Like `TVM` but equal base freq.
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).

Classified as ECDC NORMAL

IQ-TREE

- Easy to run, however can take considerably long
- Can be run locally or using the available webserver
- Best model fit prediction system



Classified as ECDC NORMAL

Estimating statistical support for a phylogeny

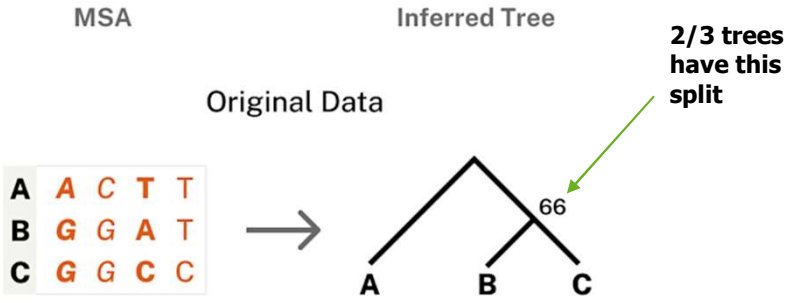
- We cannot re-run evolution to get another sample of a phylogeny
- Bootstrapping: "pull oneself up by one's bootstraps"
- Simply: resampling data with replacement
- For MSA: sample columns until you have as many columns as the original alignment
- Repeat 1000 times
- Generate a summary tree

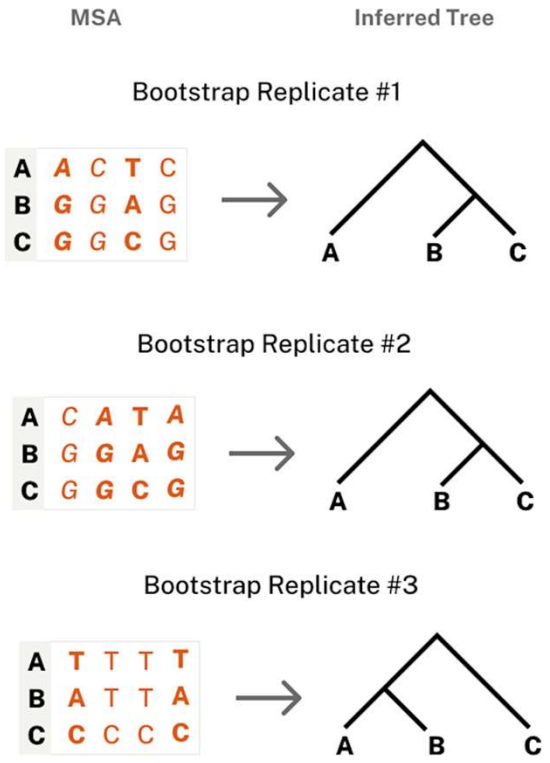


Rudolf Erich Raspe's
The Surprising Adventures of Baron Munchausen

Classified as ECDC NORMAL

Bootstrapping





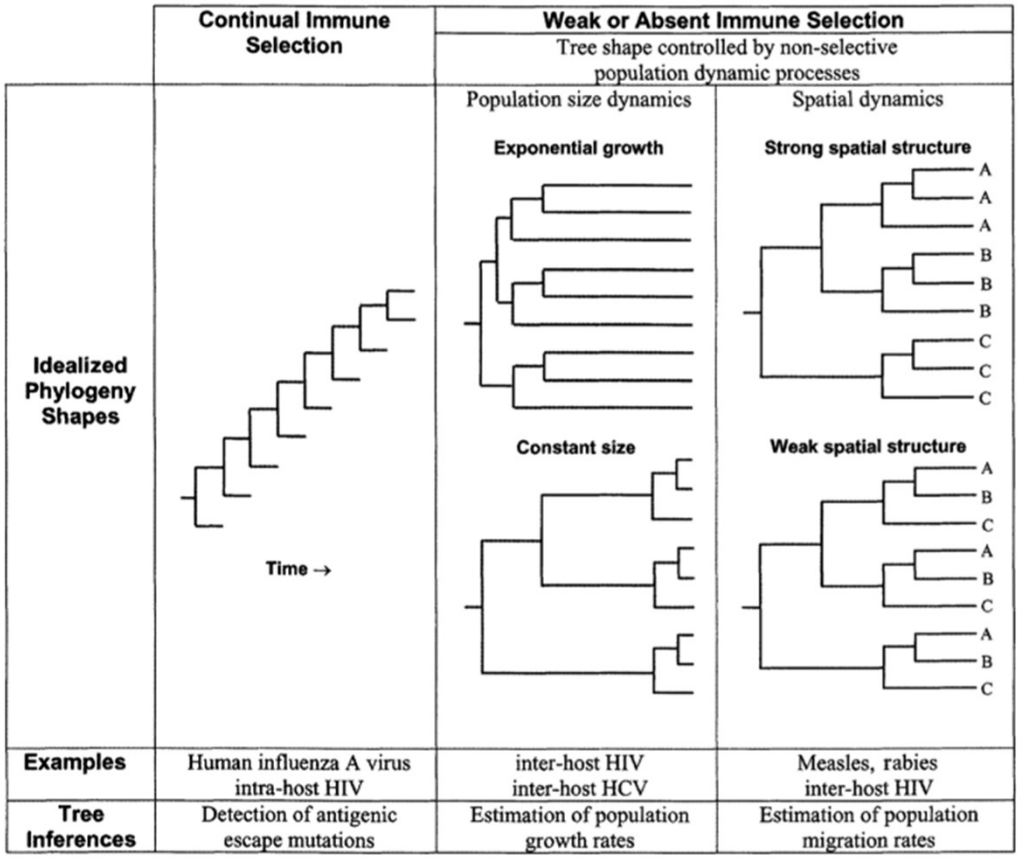
Same split every time -> Bootstrap support high
Different every time -> Bootstrap support low

Generally indicated with a percentage of how often the split occurred

Classified as ECDC NORMAL

Interpretation of tree structures

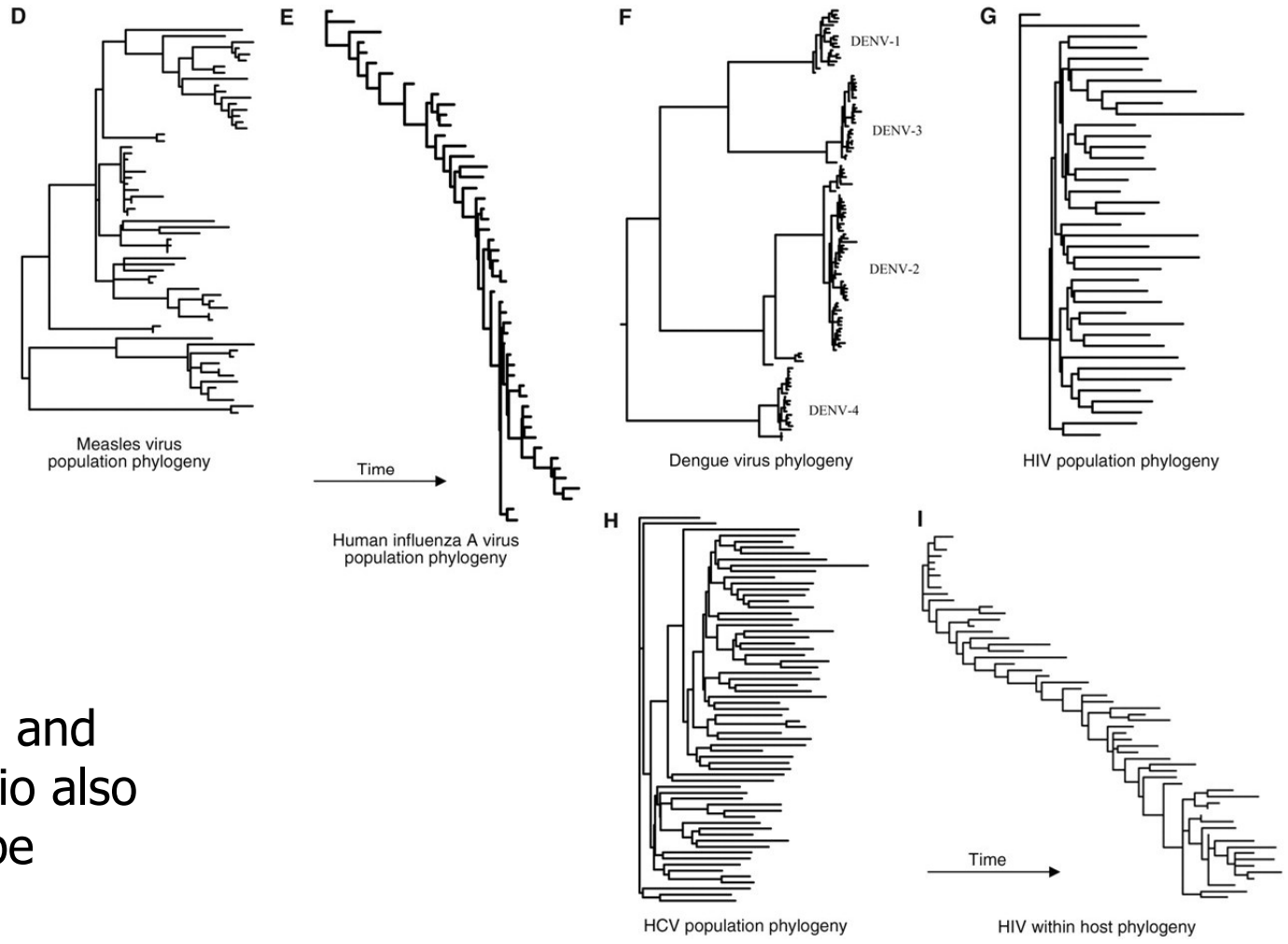
- Tree structure can tell a disease transmission story
- Further statistical modelling can help with structure interpretation (e.g. BEAST)



Bryan T. Grenfell et al. ,**Unifying the Epidemiological and Evolutionary Dynamics of Pathogens**.Science303,327-332(2004).DOI:10.1126/science.1090727

Classified as ECDC NORMAL

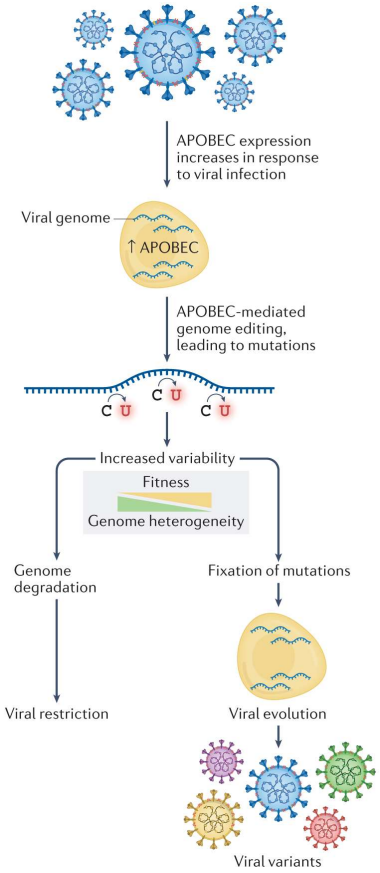
Interpretation of tree structures



Virus characteristics and transmission scenario also determine tree shape

Classified as ECDC NORMAL

Mpox specific: APOBEC3 mutations

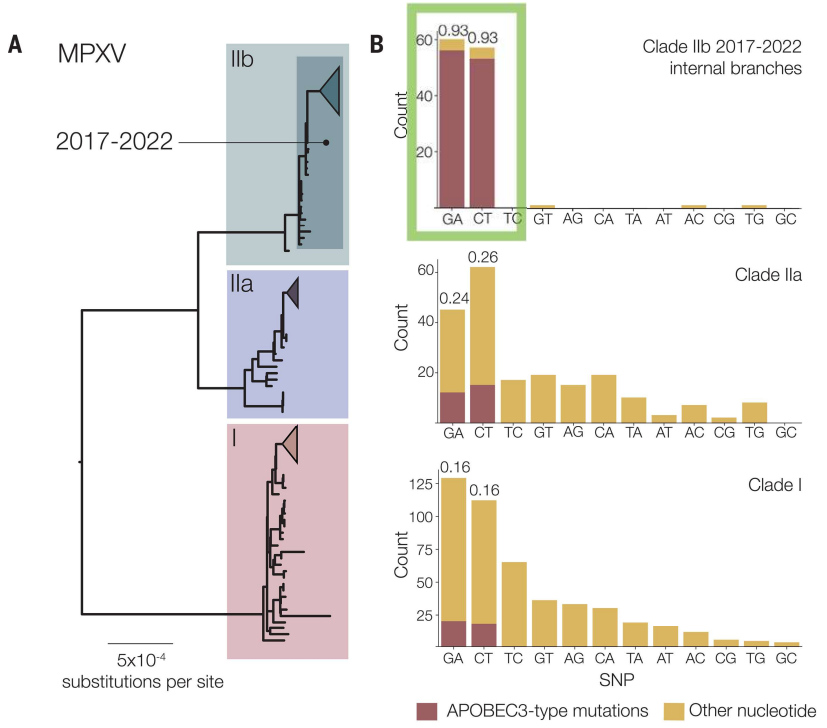


APOBEC3 signature mutations:

TC > TT

Reverse complement:

GA > AA

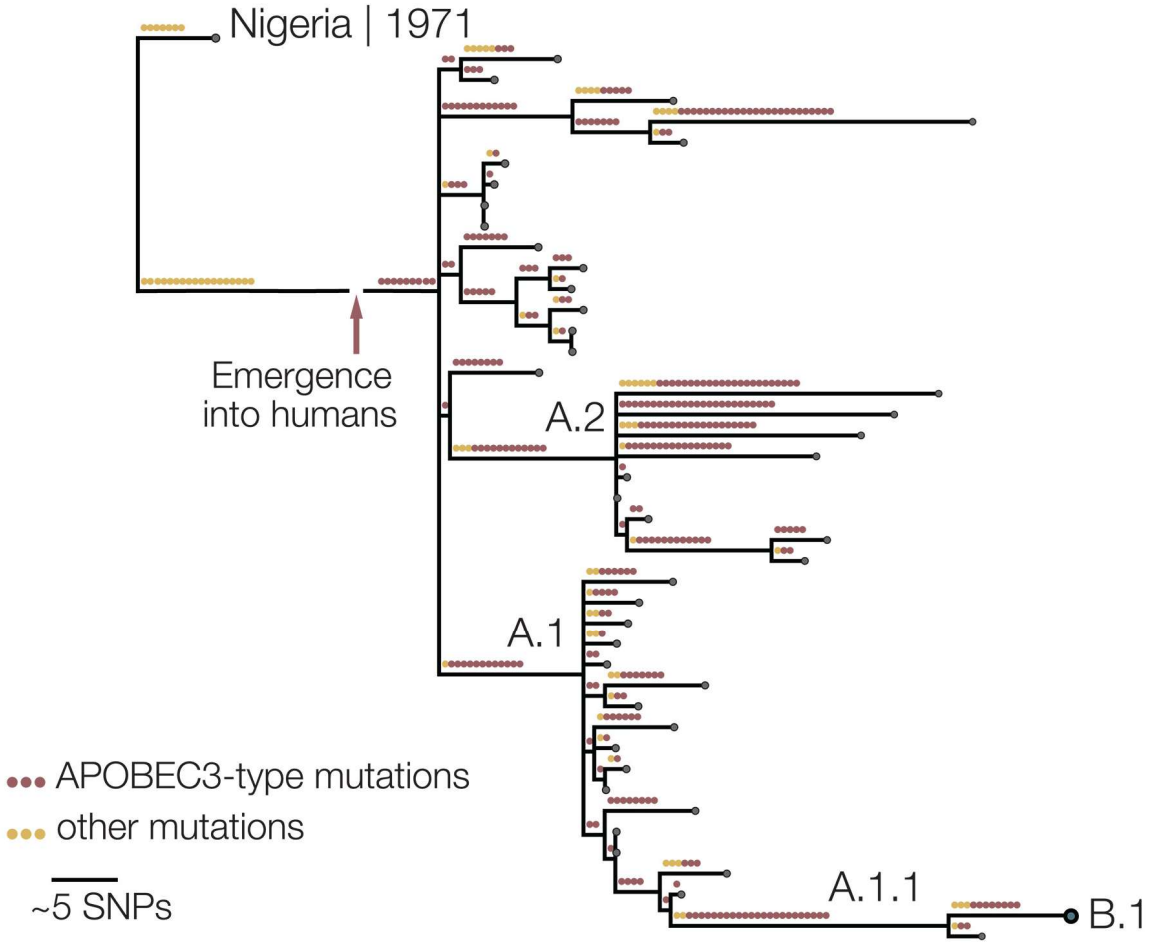


Áine O'Toole et al., **APOBEC3 deaminase editing in mpox virus as evidence for sustained human transmission since at least 2016**.Science382,595-600(2023).DOI:10.1126/science.adg8116

Pecori, R., Di Giorgio, S., Paulo Lorenzo, J. *et al.* **Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination**. *Nat Rev Genet* 23, 505–518 (2022). https://doi.org/10.1038/s41576-022-00459-8

Classified as ECDC NORMAL

Mpox specific: APOBEC3 mutation interpretation



Classified as ECDC NORMAL



Introduction to the exercise

GenEpi-BioTrain Virtual Training 29 - Genomic epidemiology of mpox virus (MPXV)



5 October 2026

Dr. David F. Nieuwenhuijse

Classified as ECDC NORMAL

Novel detections of Clade Ia mpox in previously unaffected region South-Kivu

- Genomic surveillance of hospitalized patients with suspected Mpox in South-Kivu – Democratic Republic of Congo (DRC)
- Lineage differentiating PCR detected common *Clade 1b* cases but also previously undetected *Clade Ia* cases
- Three Clade Ia samples are fit for sequencing
- Genomic epidemiology was deployed to investigate
- After successful sequencing you are tasked with the analysis of the data

Classified as ECDC NORMAL

Steps

1. Download the course material
2. Acquire mpox reference data
3. Align sequences using squirrel
4. Use provided clade Ia masking and compare the difference
5. Manually curate the alignment (if necessary)
6. Run IQ-TREE on the curated alignment using squirrel
7. Determine if there is one or multiple clusters
8. Use squirrel to determine the number of APOBEC3 mutations for the new strains

Classified as ECDC NORMAL

Command line suggestions

To just align your sequences (without masking):

squirrel --clade cladei --no-mask <your-sequences.fasta>

To add the additional masking strategy:

squirrel --clade cladei --additional-mask cladeIa_mask.csv <your-sequences.fasta>

To run IQ-TREE on the generated multiple sequence alignment:

squirrel --clade cladei --additional-mask cladeIa_mask.csv --run-phylo --outgroups KJ642617,KJ642615 <your-sequences.fasta>

To add APOBEC3 style mutation analysis to the workflow:

squirrel --clade cladei --additional-mask cladeIa_mask.csv --run-apobec3-phylo --outgroups KJ642617,KJ642615 --interactive-tree <your-sequences.fasta>

Classified as ECDC NORMAL

Reusing this content



This work is licensed under the Creative Commons (http://creativecommons.org) Attribution 4.0 Generic license

You are free:
to share – to copy, distribute and transmit the work
to remix – to adapt the work

Under the following conditions; you must attribute the work in the manner specified by the author or licensor but not:
- in any way that suggests that they endorse you or your use of the work
- in any way that suggests you are author of the work.

The authors of this slide/page encourage educators, trainers, and professionals to include this slide within their documents and presentations for rightful attribution of their works and thus also allow it to be easily shared. (ECV1-29/9/11)

Terms of usage (Learning Portal)