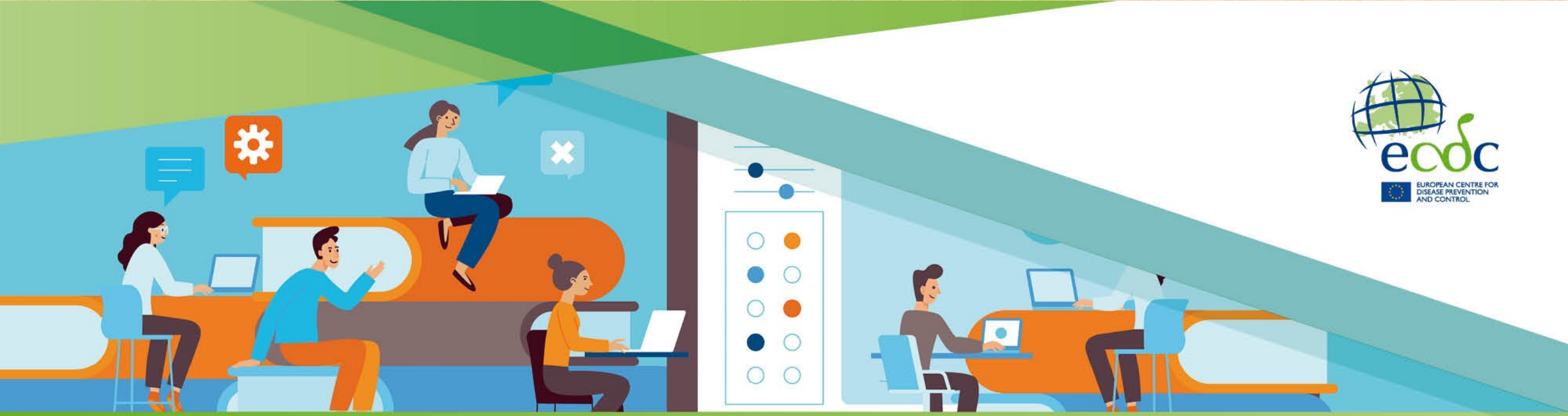


ecdc
EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL

Genomics and epidemiology of polioviruses: How a virus resists eradication

Guided correction of exercise

July 2027

Darlan da Silva Candido, PhD

Wellcome ECA Research Fellow, Imperial College London

IMPERIAL

Aim and learning objectives

Use genetic sequences to characterise and understand basic aspects of poliovirus spread during outbreaks

Specific learning objectives

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

1. **Understand** how to build your own background dataset.
2. **Generate** your own multiple sequence alignment and perform alignment quality control.
3. **Generate** a maximum likelihood phylogenetic tree.
4. **Root and interpret** a tree using FigTree.
5. **Perform** quality control of phylogenetic analysis.
6. **Build** a microreact project.

Outline

This session consists of the following elements

1. Building a multiple sequence alignment
2. Multiple sequence alignment
3. Maximum likelihood tree
4. Tree visualisation and interpretation
5. Exercise answers
6. Microreact

Building a background dataset

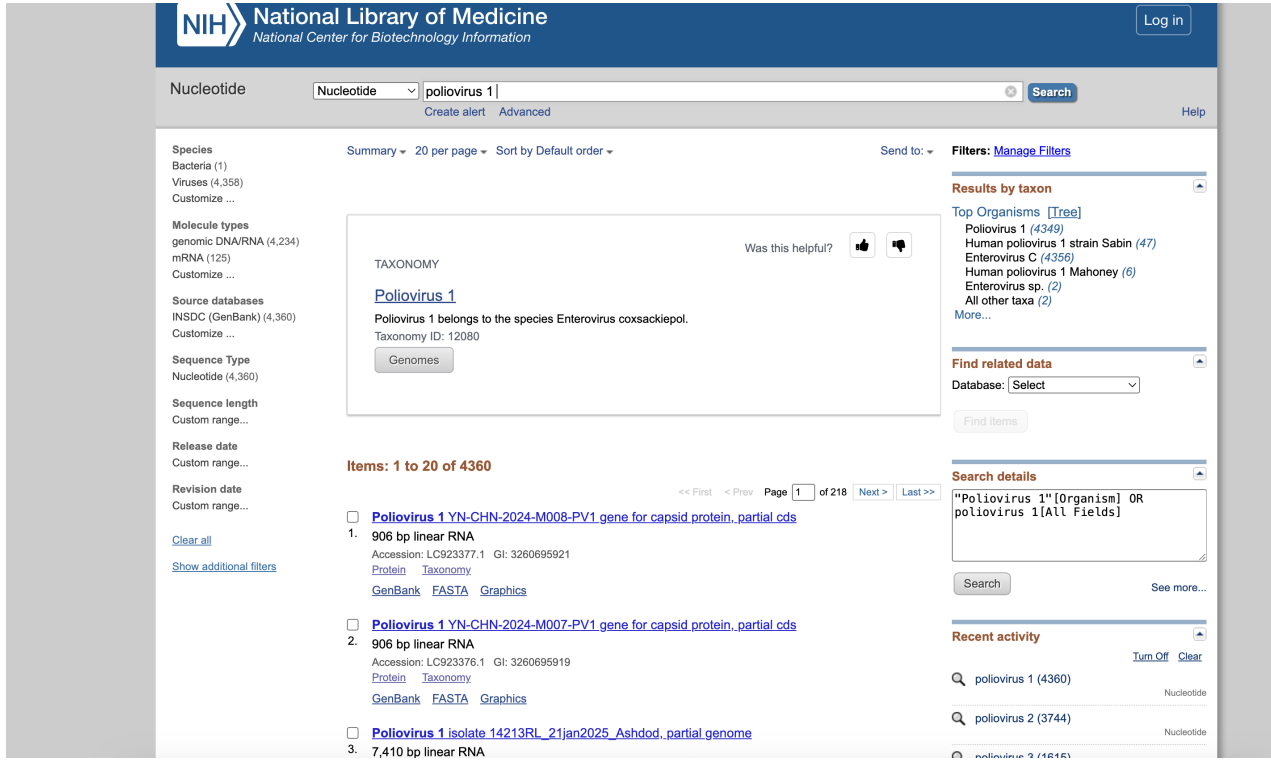
- Go to the main databases that keep sequences of the viruses you are investigating and download the information;
- Remove sequences which might belong to experiments, product development, reference genomes, duplicates or anything that might not represent the process you are trying to investigate.

What is a good background dataset?

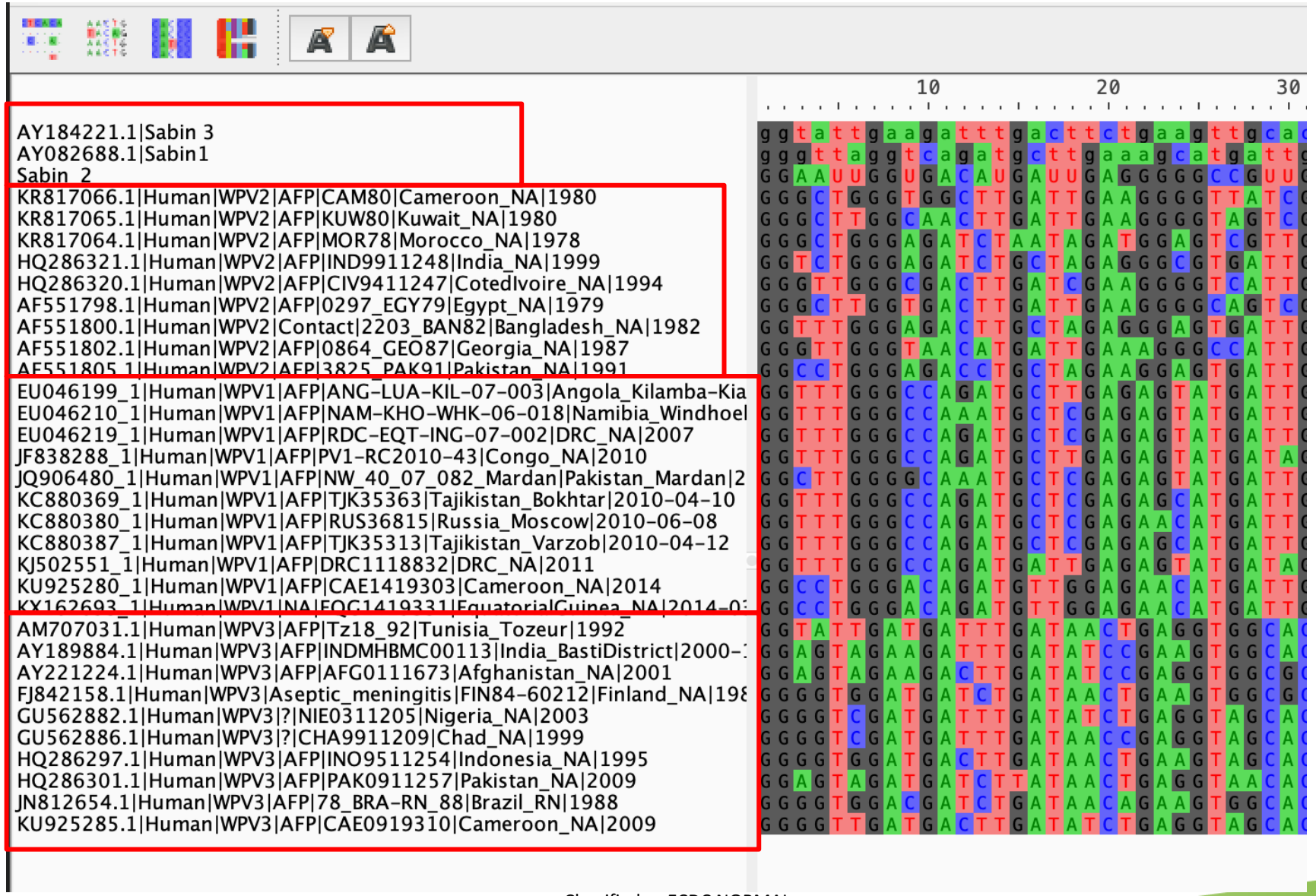
- It will depend on the question you are trying to answer
- Representative of the biology and epidemiology of the virus
- Evolutionary, geographically and temporally balanced
- Large enough (No need to be super large and repetitive)
- Good quality sequences (i.e., without lots of gaps, ambiguities, missing information or poor assembling)

Building a background dataset

- Sequences of poliovirus 1, 2 and 3, from different countries and years
- Sabin 1, 2 and 3
- Virus is extremely unlikely to be wild type poliovirus 2 and 3
- Wild type poliovirus 1 diversity more closely related to Pakistan and Afghanistan
- Vaccine viruses are more likely to be Sabin 2
- Your dataset will be limited to the sequences you have access to, i.e., publicly available or that might have been made available to you.



Building a background dataset



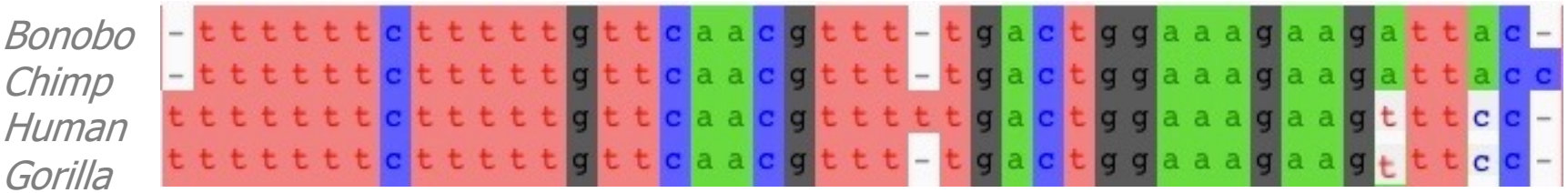
Why do we have to align?

Organize sites/positions that are homologues (if the organisms or genes or regions are homologues, so are their nucleotides)

To perform evolutionary/phylogenetics analysis – homology assumption (population history)

So…sequence alignments try to maximize similarities;

recover historical information



Multiple sequence alignment

Algorithms: Progressive /iterative; Hidden Markov Model

For multiple sequences

These methods differ in speed, accuracy, and their approach to minimizing errors caused by distant evolutionary relationships
Progressive: Build the alignment once by progressively adding sequences or groups of sequences
Iterative: Repeatedly improve an existing alignment (usually progressively generated)

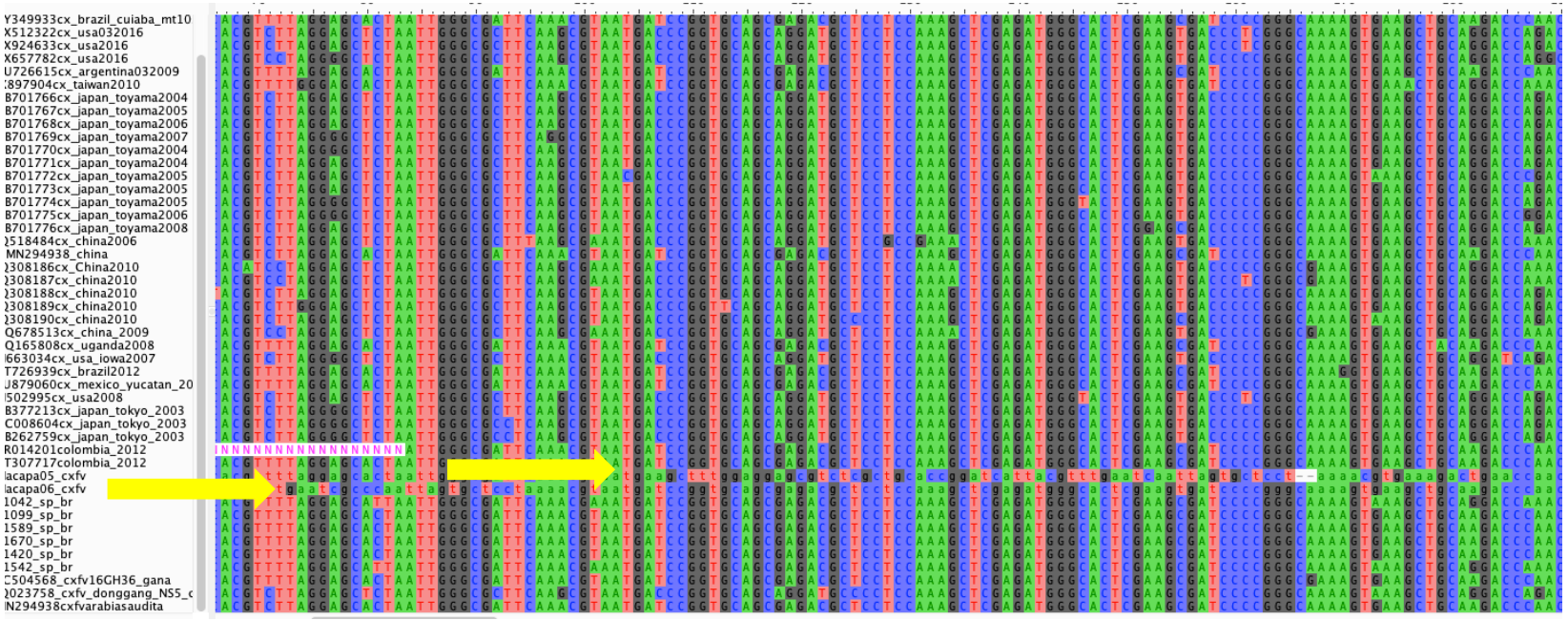
Feature	Progressive	Iterative
Speed	Very Fast	Moderate
Accuracy	Low to Moderate	High
Error Handling	Propagates errors	Corrects errors
Best Use Case	Large datasets	Refinement/Medium set
Example	ClustalW	MUSCLE, MAFFT

Multiple sequence alignment

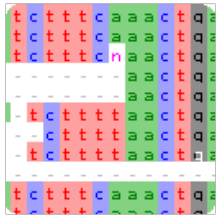
mafft –auto Background_dataset.fasta > align_Background_dataset.fasta

```
Last login: Tue Jul 14 16:12:20 on ttys001
(base) ddasilva@IC-P2PF9XL450 ~ % cd /Users/ddasilva/Dropbox/Mac/Desktop/ECDC_training/Class/Alignments
(base) ddasilva@IC-P2PF9XL450 Alignments % mafft --auto Background_dataset.fasta > align_Background_dataset.fasta
outputhat23=16
treein = 0
compacttree = 0
stacksize: 8176 kb
generating a scoring matrix for nucleotide (dist=200) ... done
All-to-all alignment.
tbfast-pair (nuc) Version 7.490
alg=L, model=DNA200 (2), 2.00 (6.00), -0.10 (-0.30), noshift, amax=0.0
0 thread(s)

outputhat23=16
Loading 'hat3.seed' ...
done.
Writing hat3 for iterative refinement
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00
tbutree = 1, compacttree = 0
Constructing a UPGMA tree ...
   30 / 33
done.

Progressive alignment ...
STEP    32 /32
done.
tbfast (nuc) Version 7.490
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
1 thread(s)
```

Multiple sequence alignment



ALIVIEW

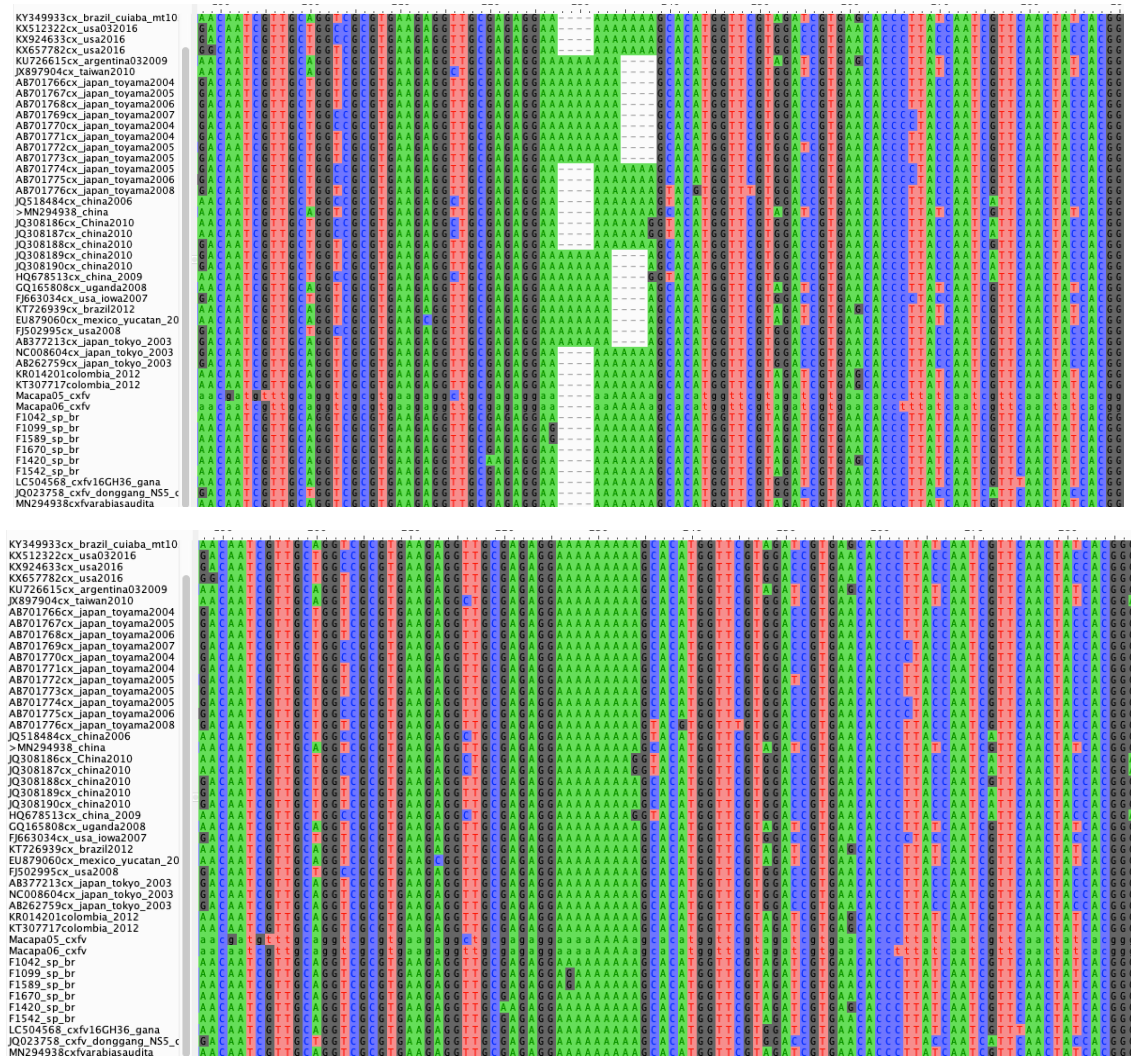


- ✓ Assembly issues (low coverage?)
- ✓ Reverse-complement genome
- ✓ Other region /same virus
- ✓ Same region/ Other virus

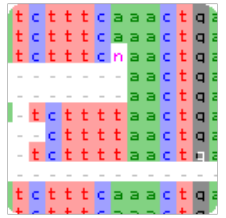
https://github.com/AliView/AliView

Slide shared by Dr Camila Romano – USP

Multiple sequence alignment



ALIVIEW

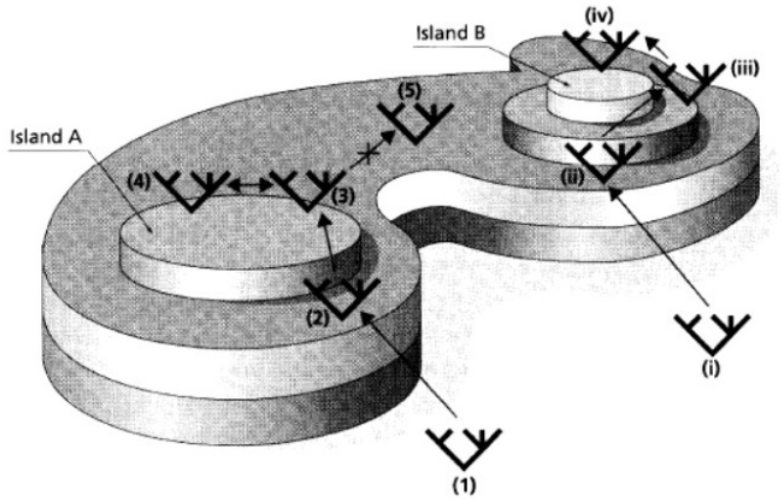


https://github.com/AliView/AliView

Slide shared by Dr Camila Romano – USP

Inferring the tree

	Tree evaluation	Step-wise clustering
Character state	Parsimony Maximum Likelihood (ML) Bayesian inference	
Distance matrix	Fitch-Margoliash	UPGMA Neigbour-joining (NJ)



Note: In step-wise addition, one can get stuck in a local sub-optimal tree peak.

Maximum parsimony

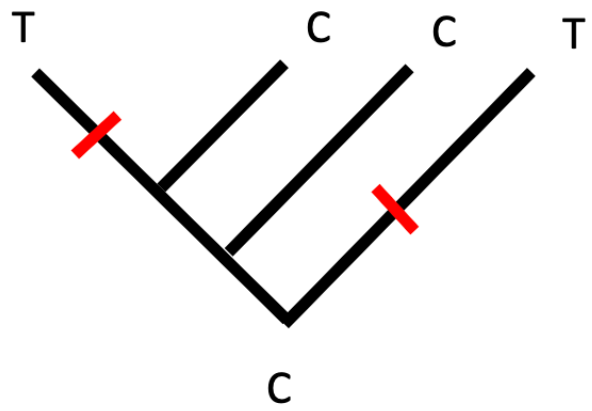
Key idea: *simplest explanation is the best*

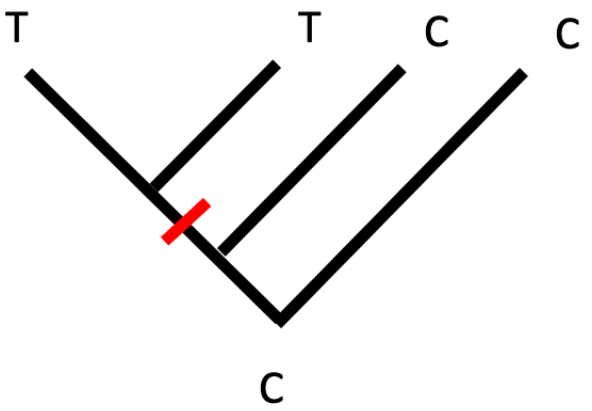
Optimizes algorithmically for identifying the tree that requires the least number of mutations

Limitations:

- It does not explicitly incorporate molecular evolution models
- No statistical consistency
- Can't estimate rates and dates

1	2	3
T	A	G
T	A	G
C	A	G
C	A	G





Superimposed mutations (multiple hits)

A given site may change more than once over evolutionary timescales

This is particularly the case for rapidly evolving organisms

AATCGCCCGTACGA

AAGCGCCCATACGA

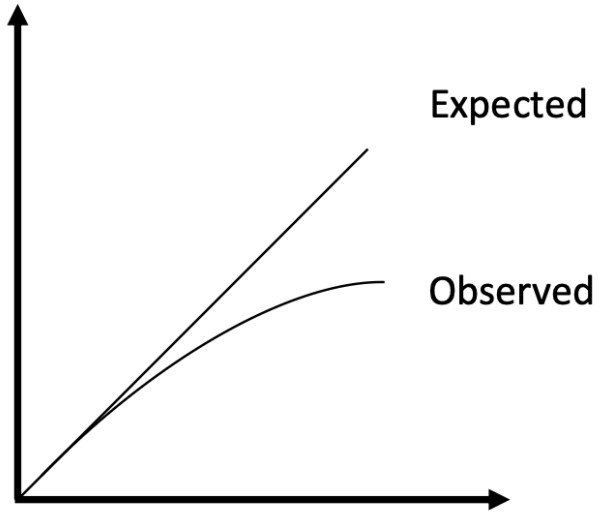
Observed mutations: 3

AATCGCCCGTACGA

A T T

AAGCGCCCATACGA

Actual mutations: 7

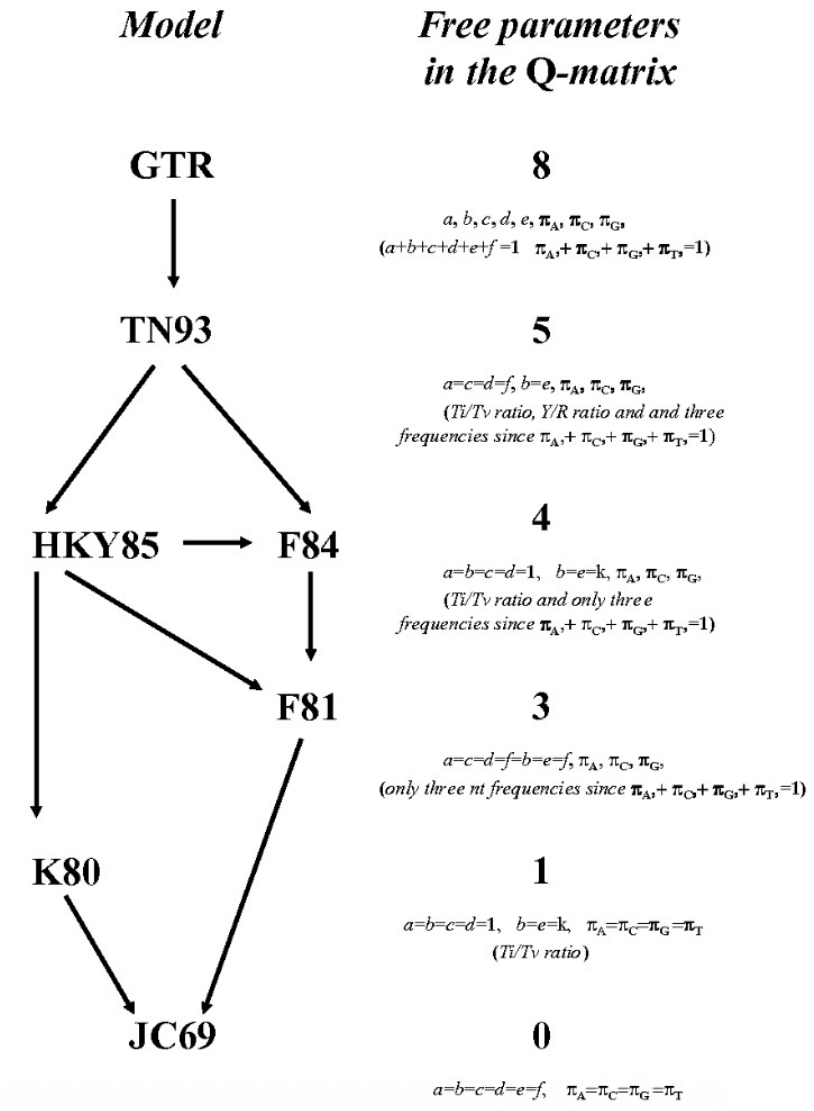


Nucleotide substitution models

Substituion models were proposed to account for unobserved mutations for phylogenetic inference

Multiple models exist, from the most simple to the most complex

Hierarchical nature

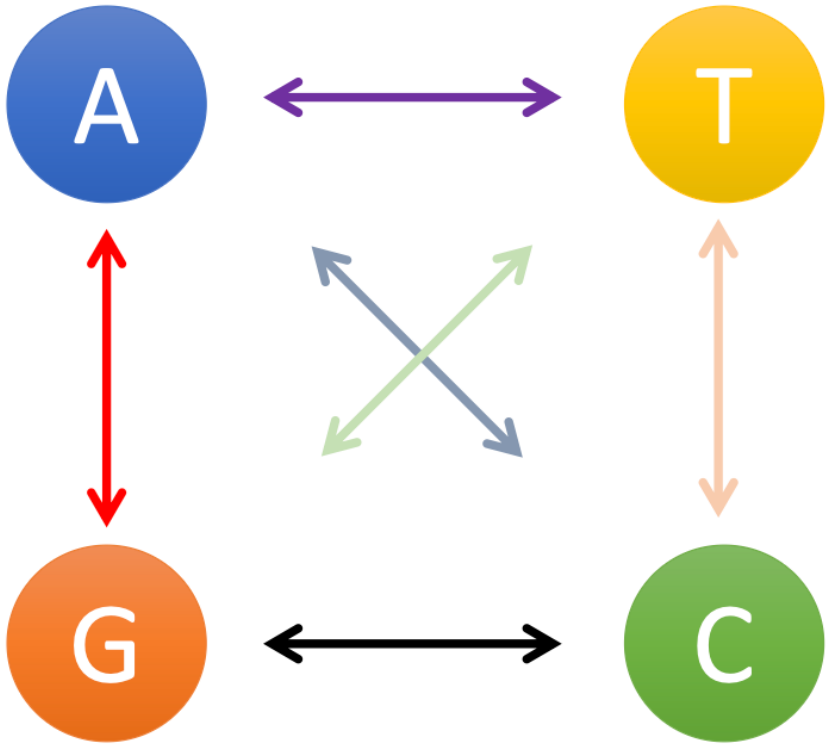


The phylogenetic handbook, 2012

Nucleotide substituion models

General time reversible (GTR)

- All rates are different
- Unequal base frequencies



Maximum-likelihood

A standard method in statistical inference

Key idea: *to calculate the probability of the data, given a hypothesis* (model and parameters)

ML conceptually:

- Compute the likelihood of possible outcomes given a set of parameters for a model
- The best estimate is the set of parameter values that maximize the likelihood (maximum likelihood estimate)

P (data | hypothesis)

In phylogenetics

- Data: multiple sequence alignment
- Probabilistic model: phylogenetic likelihood function
- Parameters: topology, branch lengths, nucleotide frequencies, substitution rates

P (alignment | tree)

Maximum-likelihood with IQ-TREE

IQ-TREE 2 edited_align_Background_dataset.fasta –B 1000 –alrt 1000 –m MFP

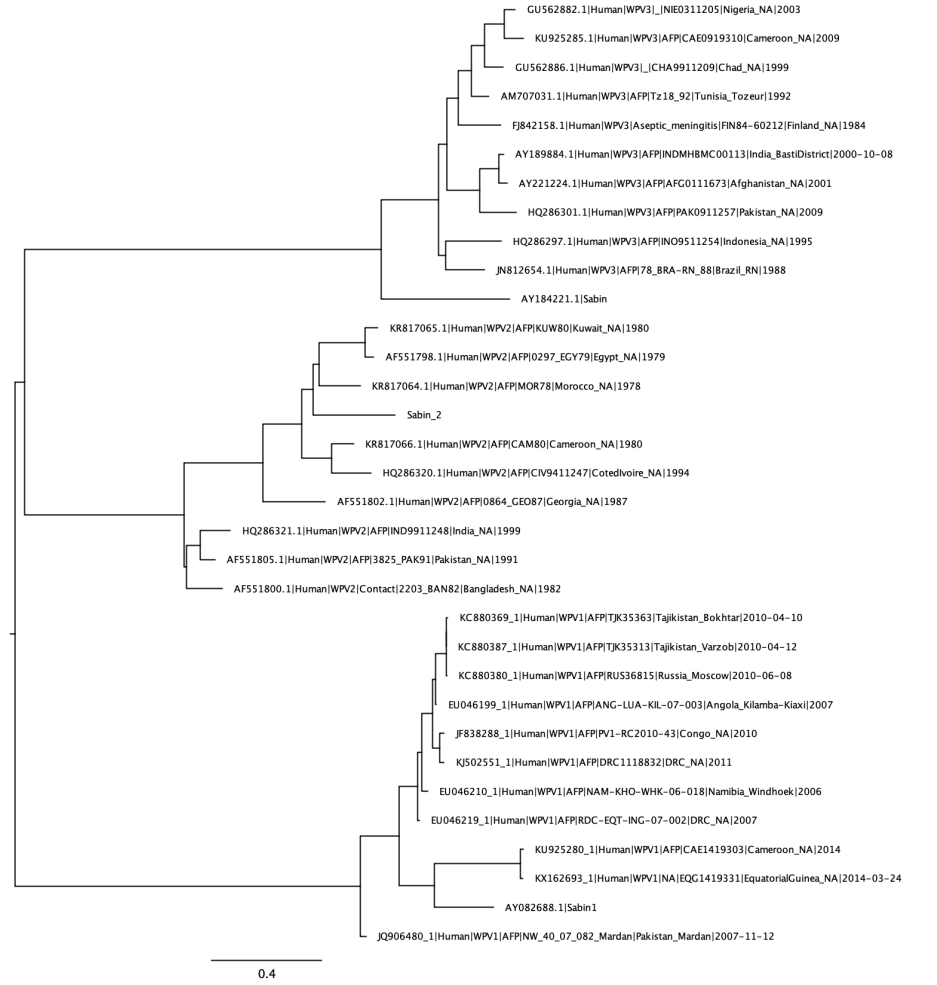
```
Last login: Sun Jul 19 11:57:06 on ttys001
cd %                                                                    (base) ddasilva@IC-P2PF9XL450 ~ % cd /Users/ddasilva/Dropbox/Mac/Desktop/ECDC_training/Class/Tree
(base) ddasilva@IC-P2PF9XL450 Tree % /Users/ddasilva/Dropbox/Mac/Downloads/iqtree-2.2.0-MacOSX/bin/iqtree2 -s edited_align_Background_dataset.fasta -B 1000 -alrt 1000 -m MFP
IQ-TREE multicore version 2.2.0 COVID-edition for Mac OS X 64-bit built Jun  1 2022
Developed by Bui Quang Minh, James Barbetti, Nguyen Lam Tung,
Olga Chernomor, Heiko Schmidt, Dominik Schrempf, Michael Woodhams, Ly Trong Nhan.

Host:    IC-P2PF9XL450 (SSE4.2, 64 GB RAM)
Command: /Users/ddasilva/Dropbox/Mac/Downloads/iqtree-2.2.0-MacOSX/bin/iqtree2 -s edited_align_Background_dataset.fasta -B 1000 -alrt 1000 -m MFP
Seed:    865943 (Using SPRNG - Scalable Parallel Random Number Generator)
Time:    Sun Jul 19 12:22:11 2026
Kernel:  SSE2 - 1 threads (10 CPU cores detected)

HINT: Use -nt option to specify number of threads because your CPU has 10 cores!
HINT: -nt AUTO will automatically determine the best number of threads to use.

Reading alignment file edited_align_Background_dataset.fasta ... Fasta format detected
Reading fasta file: done in 0.00196099 secs using 95.97% CPU
Alignment most likely contains DNA/RNA sequences
Constructing alignment: done in 0.000980139 secs using 76.42% CPU
Alignment has 33 sequences with 909 columns, 449 distinct patterns
449 parsimony-informative, 20 singleton sites, 440 constant sites
WARNING: Some sequence names are changed as follows:
GU562882.1|Human|WPV3|?|NIE0311205|Nigeria_NA|2003 -> GU562882.1|Human|WPV3|_|NIE0311205|Nigeria_NA|2003
GU562886.1|Human|WPV3|?|CHA9911209|Chad_NA|1999 -> GU562886.1|Human|WPV3|_|CHA9911209|Chad_NA|1999

                                                     Gap/Ambiguity  Composition  p-value
Analyzing sequences: done in 2.59876e-05 secs using 88.5% CPU
   1  AY184221.1|Sabin                                      0.99%    passed     71.72%
   2  AY082688.1|Sabin1                                     0.33%    passed     80.41%
   3  Sabin_2                                               0.66%    passed     69.14%
   4  KR817066.1|Human|WPV2|AFP|CAM80|Cameroon_NA|1980      0.66%    passed     98.21%
   5  KR817065.1|Human|WPV2|AFP|KUW80|Kuwait_NA|1980        0.66%    passed     90.95%
   6  KR817064.1|Human|WPV2|AFP|MOR78|Morocco_NA|1978       0.66%    passed     93.54%
   7  HQ286321.1|Human|WPV2|AFP|IND9911248|India_NA|1999    0.66%    passed     92.21%
   8  HQ286320.1|Human|WPV2|AFP|CIV9411247|CotedIvoire_NA|1994  0.66%  passed     95.14%
   9  AF551798.1|Human|WPV2|AFP|0297_EGY79|Egypt_NA|1979    0.66%    passed     83.15%
  10  AF551800.1|Human|WPV2|Contact|2203_BAN82|Bangladesh_NA|1982  0.66%  passed  77.59%
  11  AF551802.1|Human|WPV2|AFP|0864_GEO87|Georgia_NA|1987  0.66%    passed     99.42%
```

Maximum-likelihood with IQ-TREE

edited_align_Background_dataset.fasta.bionj
edited_align_Background_dataset.fasta.ckp.gz
edited_align_Background_dataset.fasta.contree
edited_align_Background_dataset.fasta.iqtree
edited_align_Background_dataset.fasta.log
edited_align_Background_dataset.fasta.mldist
edited_align_Background_dataset.fasta.model.gz
edited_align_Background_dataset.fasta.splits.nex
edited_align_Background_dataset.fasta.treefile

Open tree in FigTree

Does this tree make sense?



Add our sequences of interest

- Add your sequence of interest to the main alignment
- Perform the previous steps again

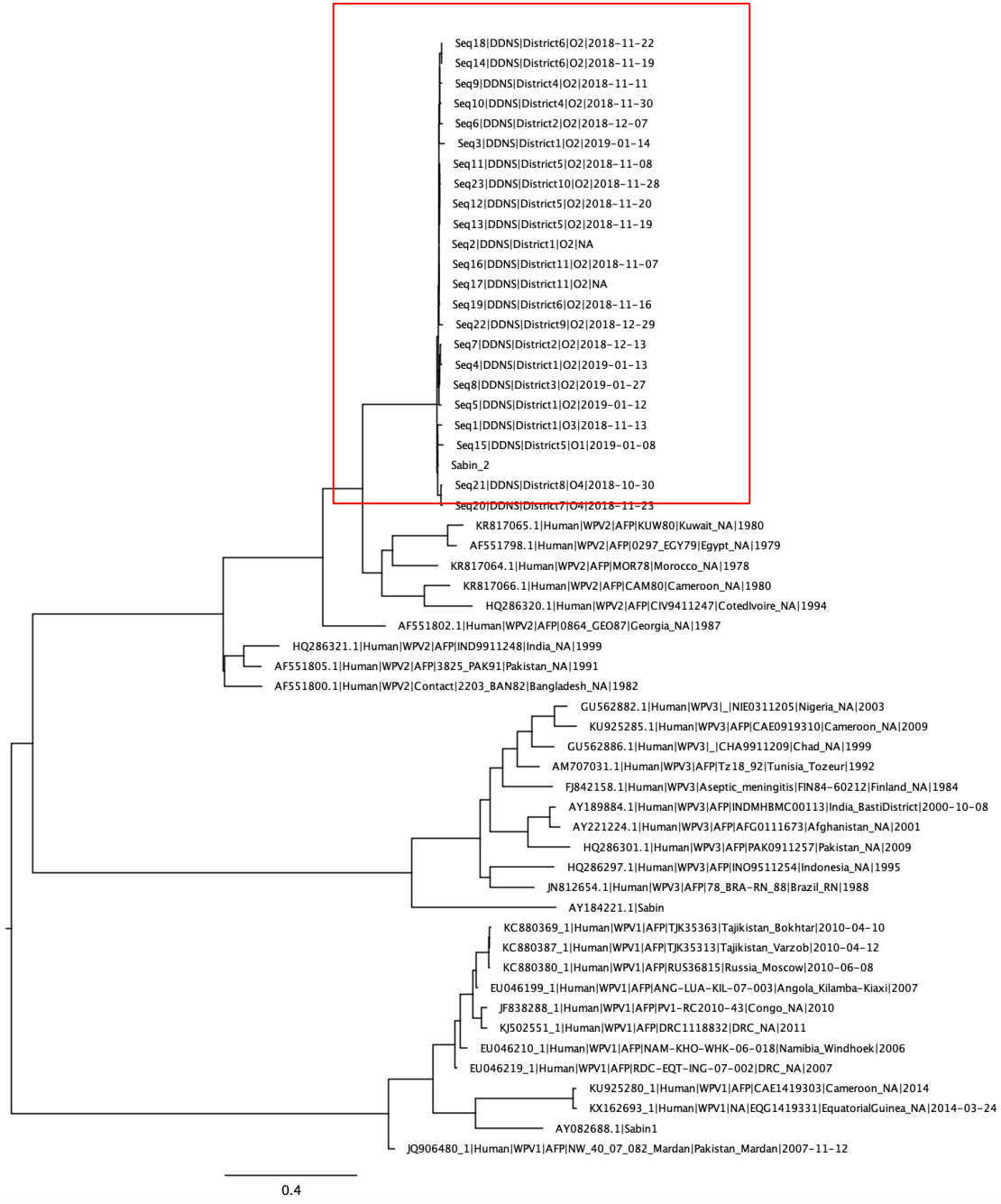
Answer the questions

Question 1. What is the serotype of the sequences of interest?

Answer: Poliovirus 2

Question 2: Are these sequences wild-type or vaccine-derived?

Answer: Vaccine



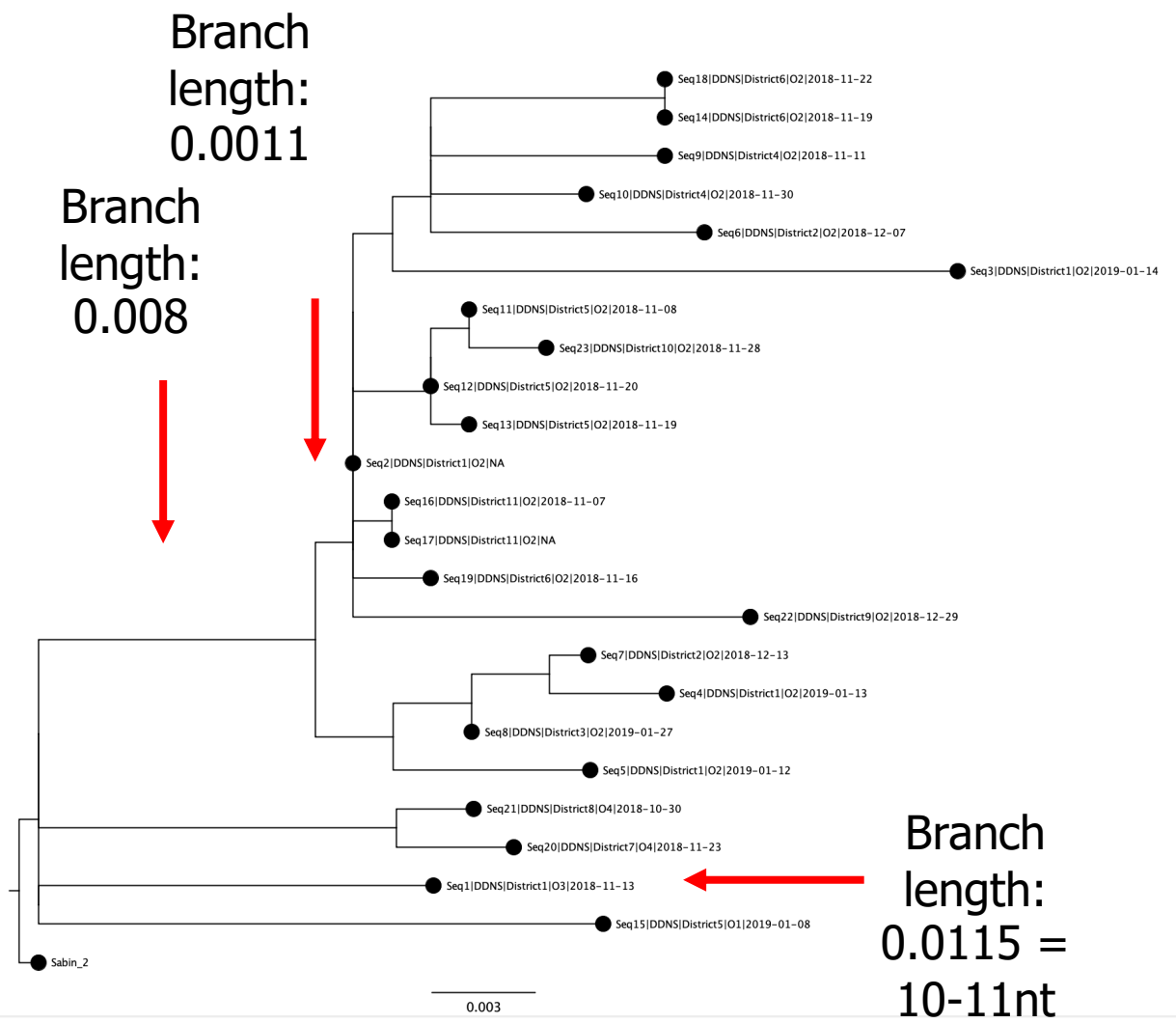
Answer the questions

Question 3: Are these sequences Sabin-like or VDPVs?

Answer: VDPVs (>= 6 nt from Sabin)

Question 4: How many different emergence groups do these sequences belong to?

Total = 0.0091 subst/site x ~900 = 8-9 nt



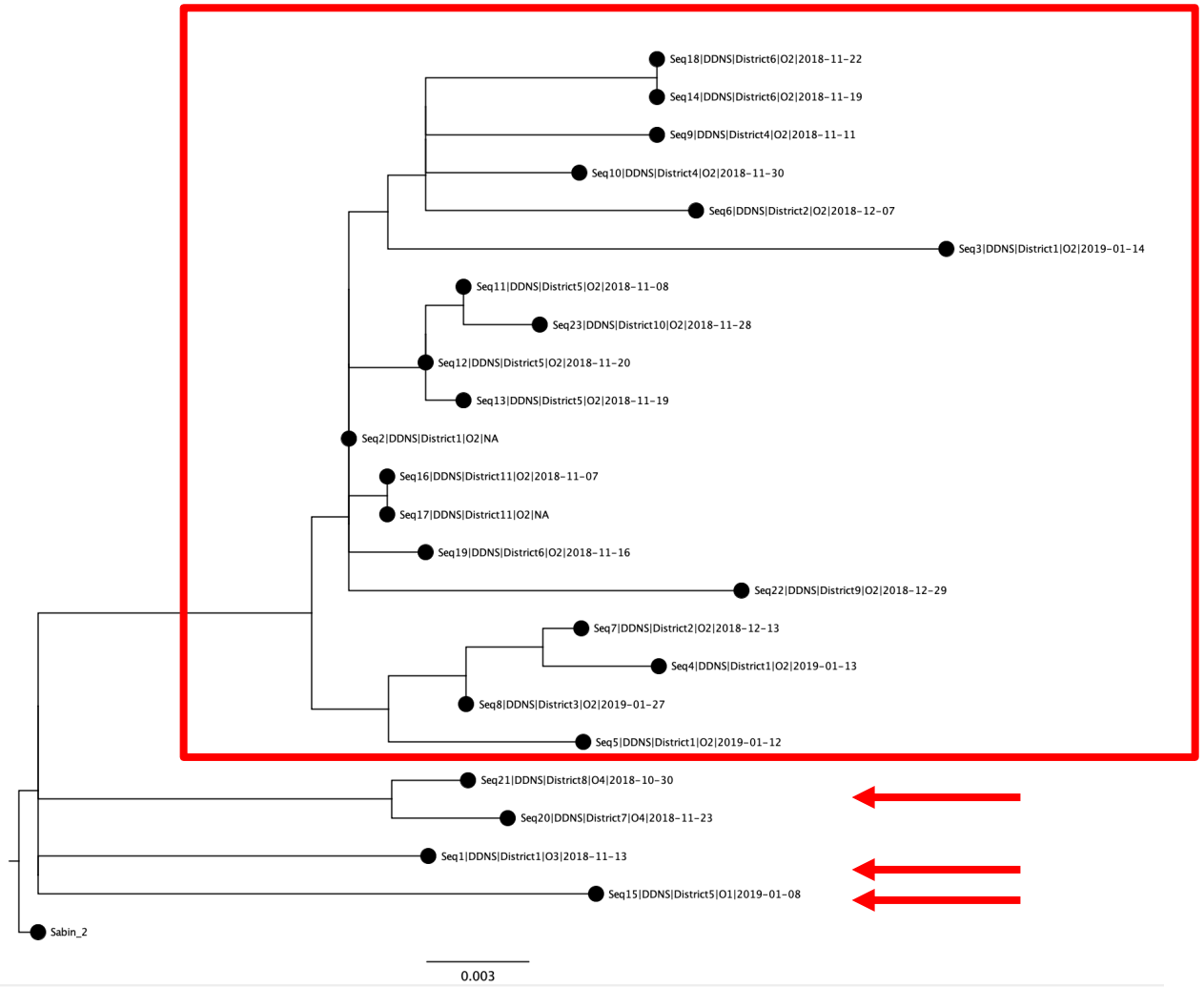
Answer the questions

Question 3: Are these sequences Sabin-like or VDPVs?

Answer: VDPVs (>= 6 nt from Sabin)

Question 4: How many different emergence groups do these sequences belong to?

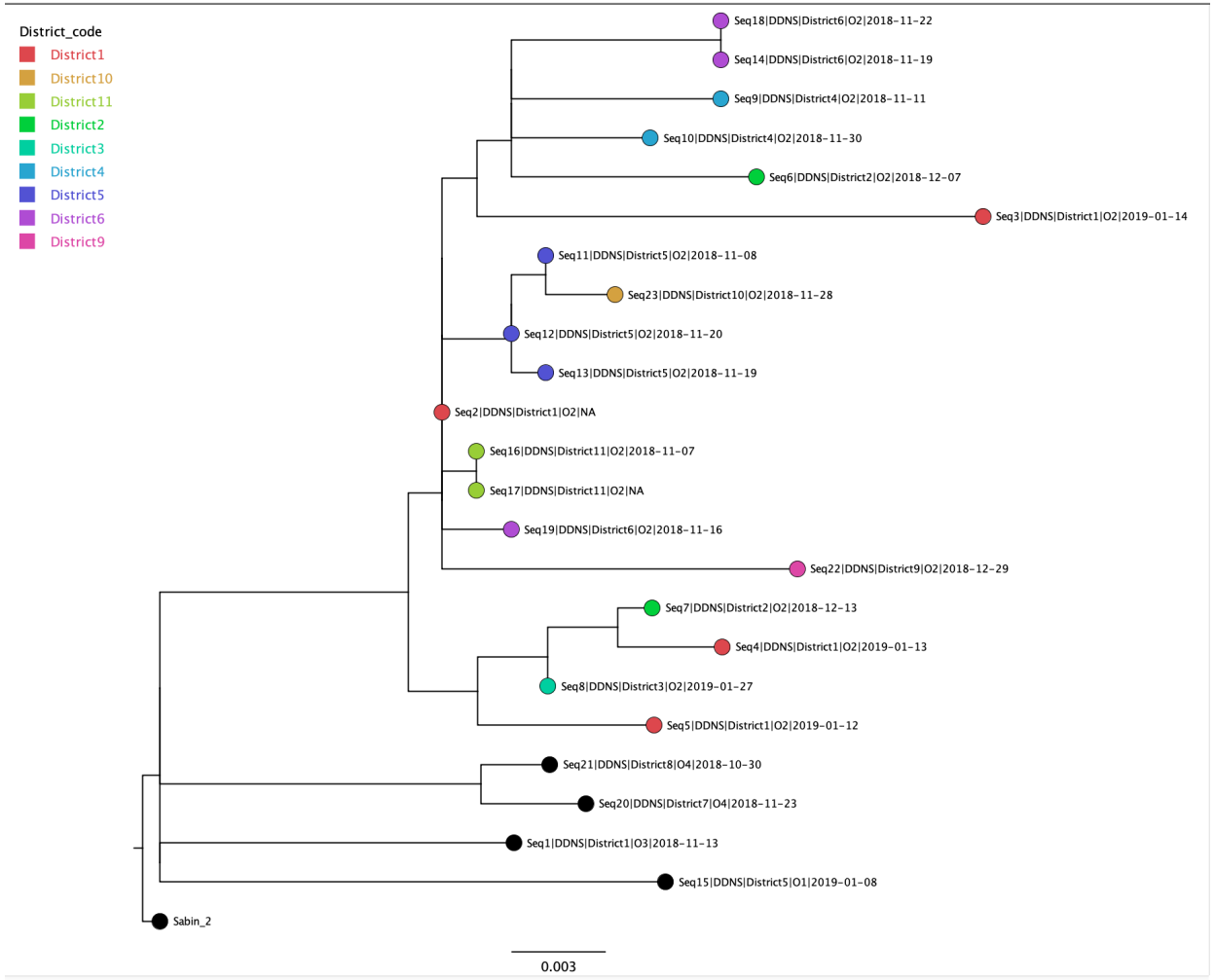
Answer: 4



Answer the questions

Question 5: How geographically intermixed was the observed spread?

Fairly good mixing with sequences from District 1 present in several clusters



Microreact

1. Drop the rooted tree saved as Newick and the CSV file with metadata information onto https://microreact.org/upload

Recommended material to explore further

- **Foundations**

Lemey P, Salemi M, Vandamme A-M (eds.). **The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.**

Page RDM, Holmes EC. **Molecular Evolution: A Phylogenetic Approach.**

- **Applications of genomic epidemiology**

Dudas G, Carvalho LM, Rambaut A, Bedford T. **Virus genomes reveal factors that spread and sustained the Ebola epidemic.** *Nature*. 2017.

Lemey P, Rambaut A, Bedford T, et al. **Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2.** *PLOS Medicine*. 2014.

Candido DS, Claro IM, de Jesus JG, et al. **Evolution and epidemic spread of SARS-CoV-2 in Brazil.** *Science*. 2020.

Dudas G, Carvalho LM, Rambaut A, Bedford T, et al. **MERS-CoV spillover at the camel–human interface.** *eLife*. 2018.

- **Poliovirus case studies**

Candido D, et al. **Genomic epidemiology of circulating vaccine-derived poliovirus transmission in the Democratic Republic of the Congo.** *Nature Microbiology*. 2026.

Jørgensen D, et al. **Genomic epidemiology of poliovirus transmission using environmental and clinical surveillance.** *Nature Communications*. 2025.

References

Genomic epidemiology

Candido DS, Claro IM, de Jesus JG, *et al.* **Evolution and epidemic spread of SARS-CoV-2 in Brazil.** *Science.* 2020.

Dudas G, Carvalho LM, Rambaut A, Bedford T, *et al.* **Virus genomes reveal factors that spread and sustained the Ebola epidemic.** *Nature.* 2017.

Dudas G, Carvalho LM, Rambaut A, Bedford T, *et al.* **MERS-CoV spillover at the camel–human interface.** *eLife.* 2018.

Faria NR, *et al.* **Genomic and epidemiological monitoring of yellow fever virus transmission.** *Science.* 2018.

Faria NR, *et al.* **Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil.** *Science.* 2021.

Grubaugh ND, Ladner JT, Lemey P, *et al.* **Tracking virus outbreaks in the twenty-first century.** *Nature Microbiology.* 2019.

Lemey P, Rambaut A, Bedford T, *et al.* **Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2.** *PLOS Medicine.* 2014.

Poliovirus

Candido D, *et al.* **Genomic epidemiology of circulating vaccine-derived poliovirus transmission in the Democratic Republic of the Congo.** *Nature Microbiology.* 2025.

Chopra K, *et al.* **Genomic epidemiology of vaccine-derived poliovirus outbreaks in the Democratic Republic of the Congo.** Under review.

Jorba J, Campagnoli R, De L, Kew O. **Calibration of multiple poliovirus molecular clocks covering an extended evolutionary range.** *Journal of Virology.* 2008.

Jørgensen D, *et al.* **Genomic epidemiology of poliovirus transmission using environmental and clinical surveillance.**

References

Jørgensen D, *et al.* **Genomic surveillance of poliovirus.** *Virus Evolution.* 2020.

Macklin GR, O'Reilly KM, Grassly NC, *et al.* **Evolving epidemiology of circulating vaccine-derived poliovirus.** *Science.* 2020.

World Health Organization (WHO). **Poliomyelitis factsheets and technical guidance.**

Global Polio Eradication Initiative (GPEI). **Polio surveillance and eradication resources.**

Phylogenetics

Grenfell BT, Pybus OG, Gog JR, *et al.* **Unifying the epidemiological and evolutionary dynamics of pathogens.** *Science.* 2004.

Holmes EC. **Evolutionary history and phylogeography of human viruses.** *Annual Review of Microbiology.* 2008.

Jhwueng DC, Liu FC. **Resolution of phylogenetic relationships.** *Diversity.* 2022.

Rambaut A, Holmes EC, O'Toole Á, *et al.* **A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology.** *Nature Microbiology.* 2020.

Sahin E, *et al.* **Polytomies in SARS-CoV-2 phylogenies.** *Journal of Medical Virology.* 2021.

Tian Y, Kubatko L. **Rooting phylogenetic trees under the coalescent model.** *BMC Ecology and Evolution.* 2017.

Books

Lemey P, Salemi M, Vandamme A-M (eds.). **The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.**

Reusing this content

Acknowledgements

The creation of this training material was commissioned by ECDC to Imperial College London with the direct involvement of Dr Darlan da Silva Candido