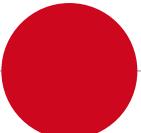


Phylogenetic analysis of RNA viruses

Alice Fusaro

Istituto Zooprofilattico Sperimentale delle Venezie (IZSVE), Padua, Italy



Introduction

Short generation time

Genome size: 3 to 30 kb
New RNA genome ~ 0,4 sec

Large populations

Several orders of magnitude larger than any population size for DNA organisms

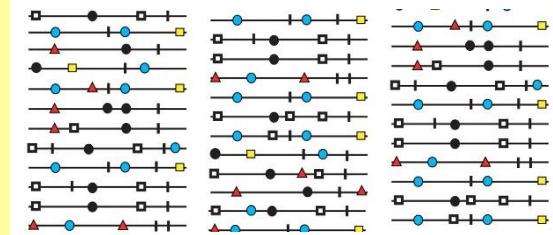
High error rate



Error-prone RNA synthesis
~ 10^{-5} to 10^{-2} sub/site/year

RNA VIRUS

Quasispecie



These characteristics are responsible for the high genetic variability and the enormous adaptive capacity of RNA viruses to changing environmental conditions such as immune pressure, antiviral treatments, host-switch that render these viruses a formidable challenge to animal and human health

How do virus evolve?

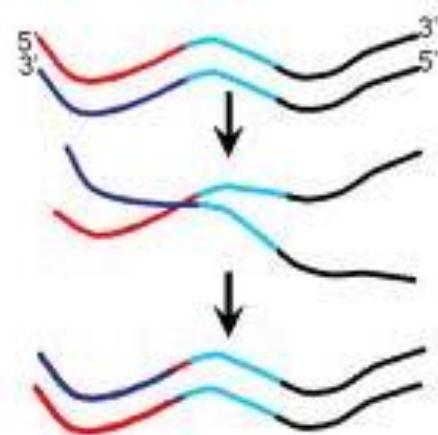
POINT MUTATIONS

ACTGTCA

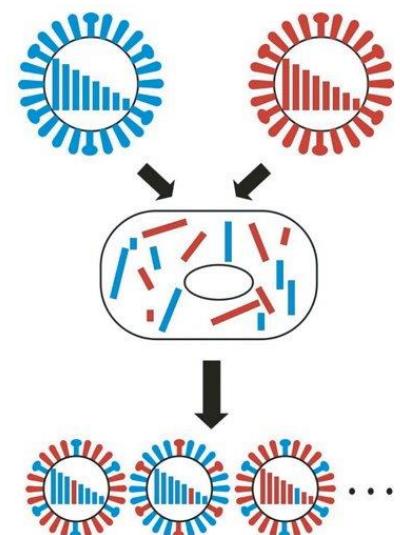


ACAGTCA

RECOMBINATION

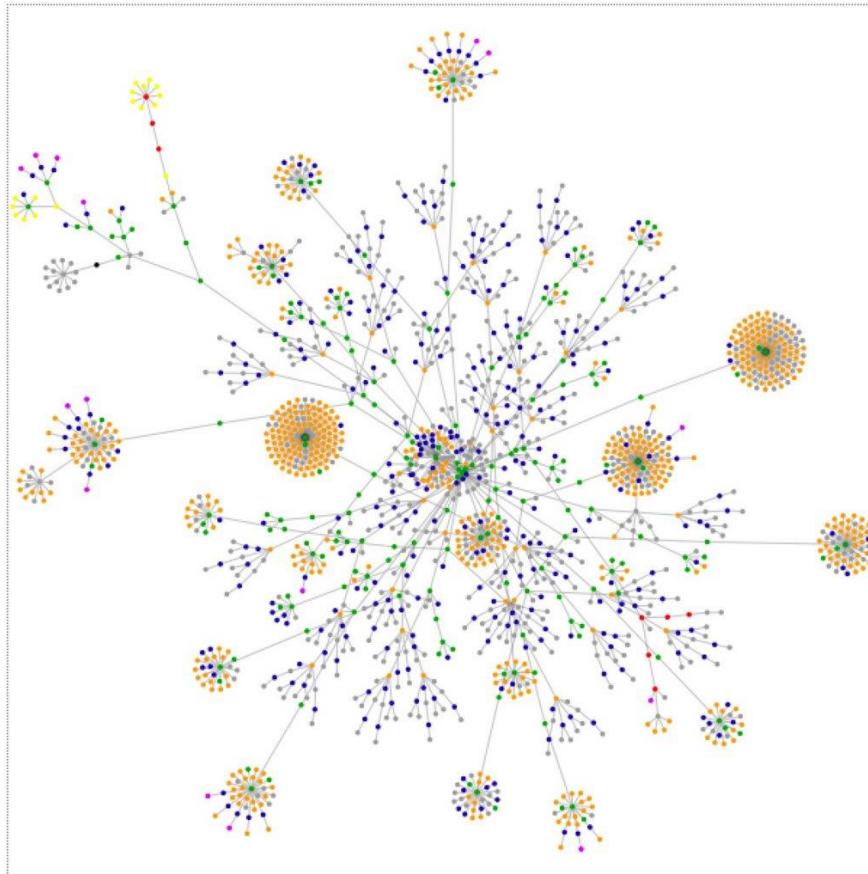


REASSORTMENT



RNA virus evolution

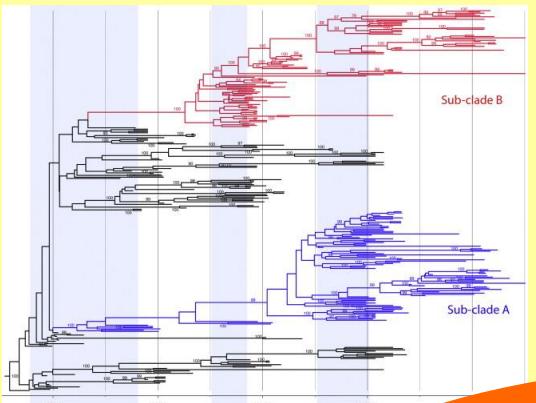
Genetic changes in 'real time'



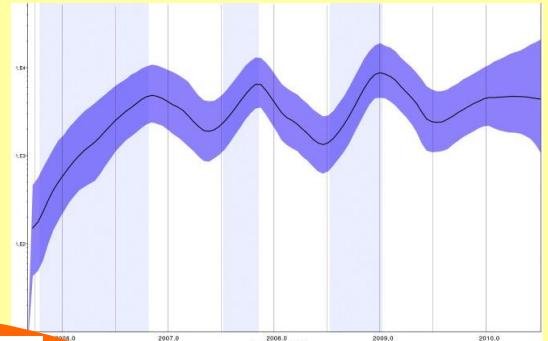
An understanding of the mechanisms of RNA virus sequence change is crucial to predict important aspects of their emergence and long-term evolution.

Moya et al., Nature 2004

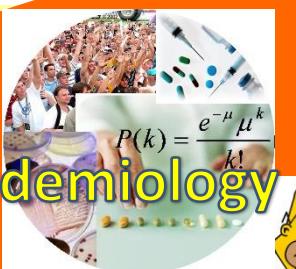
PHYLOGENY



EVOLUTIONARY DYNAMICS OF VIRAL POPULATIONS



Epidemiology

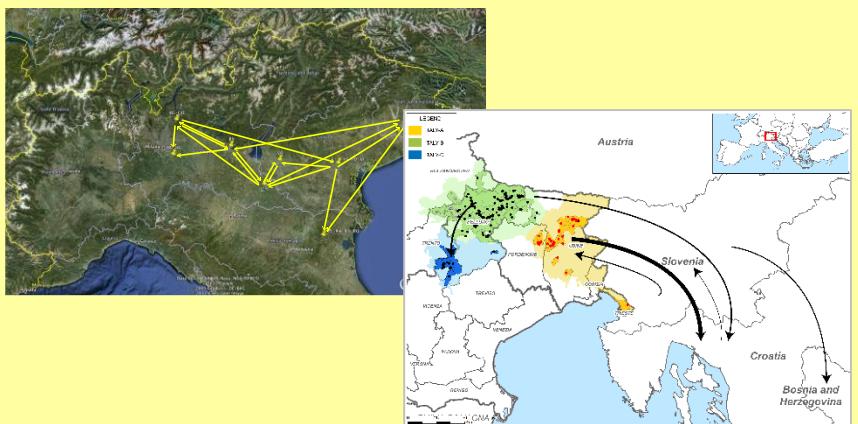


Sequencing

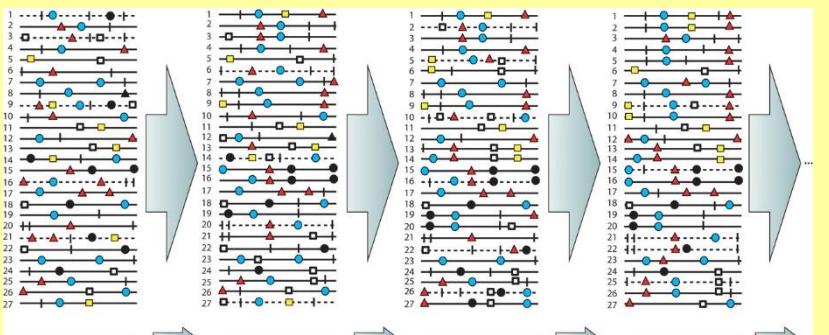


Bioinformatic

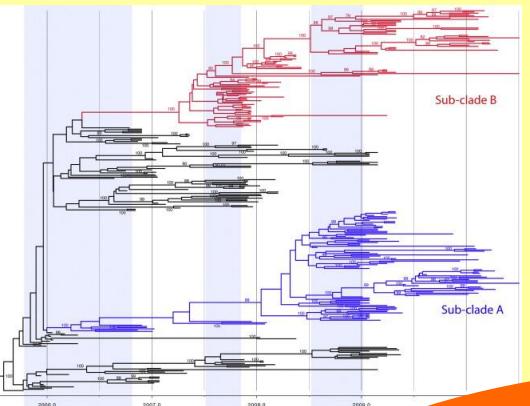
PHYLOGEOGRAPHY



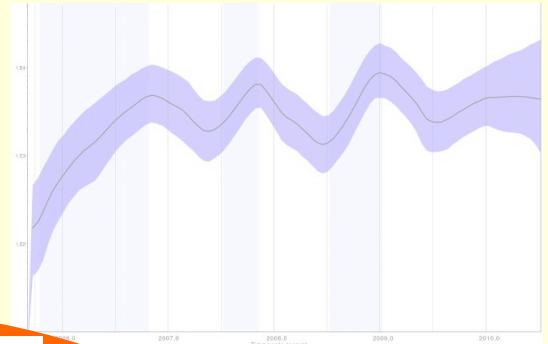
INTRA-HOST VARIABILITY NGS



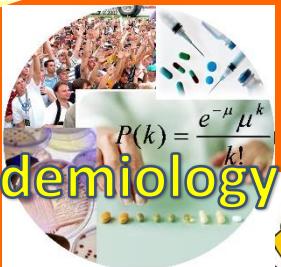
PHYLOGENY



EVOLUTIONARY DYNAMICS OF VIRAL POPULATIONS



Epidemiology

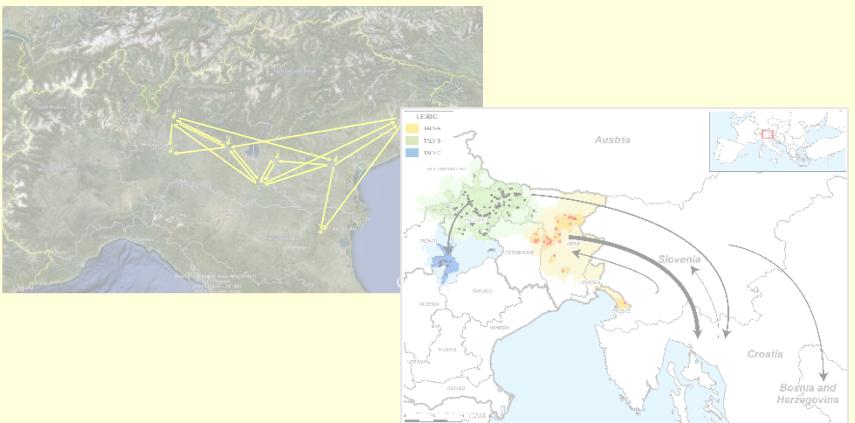


Sequencing

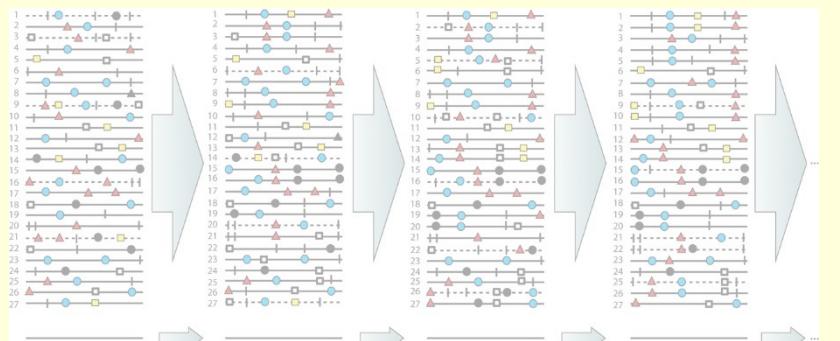


Bioinformatic

PHYLOGEOGRAPHY



INTRA-HOST VARIABILITY NGS



Which information can you gather from a phylogenetic analysis?

ORIGIN

GENETIC
RELATIONSHIP

NEW
INTRODUCTIONS

FOLLOW VIRAL
EVOLUTION

phylogeny

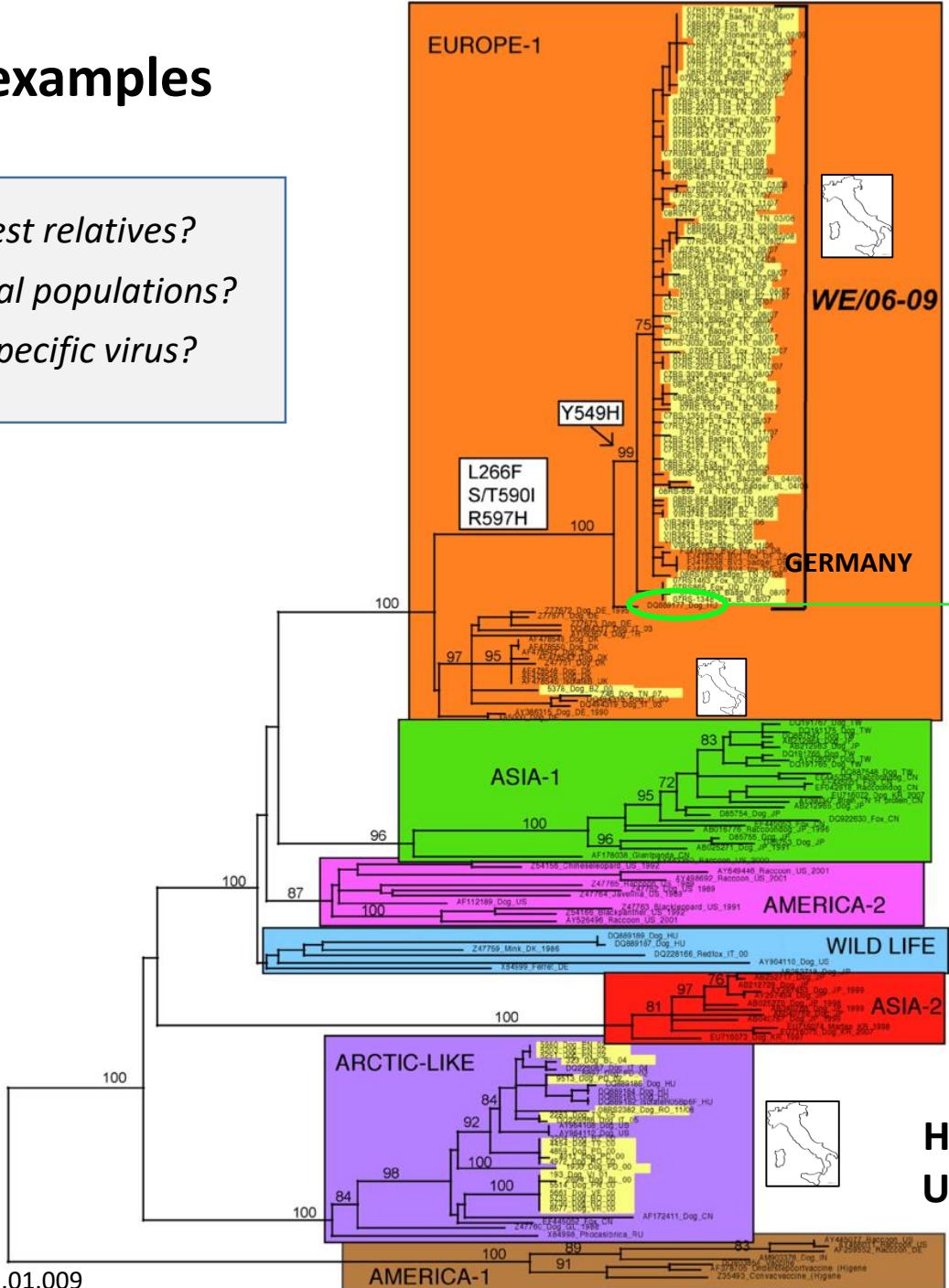
VIRAL GENOTYPING

REASSORTMENTS OR
RECOMBINATION

CO-CIRCULATION OF MULTIPLE
GENETIC GROUPS

Phylogeny- examples

- Which viruses are the closest relatives?
- Are circulating multiple viral populations?
- What was the origin of a specific virus?



ITALY
HUNGARY
GERMANY
DANIMARK

HUNGARY

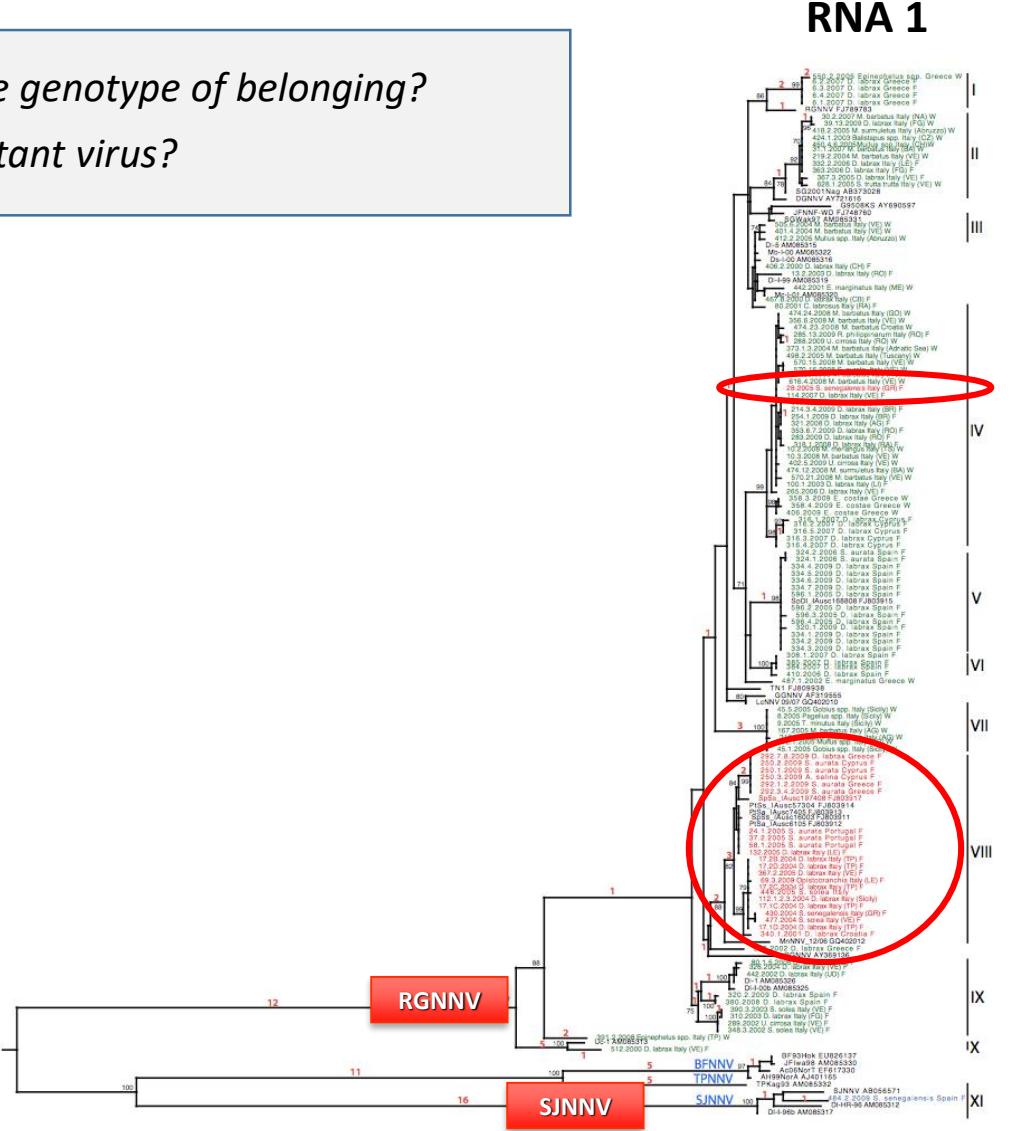


H gene 
Canine distemper virus
Family Paramyxoviridae

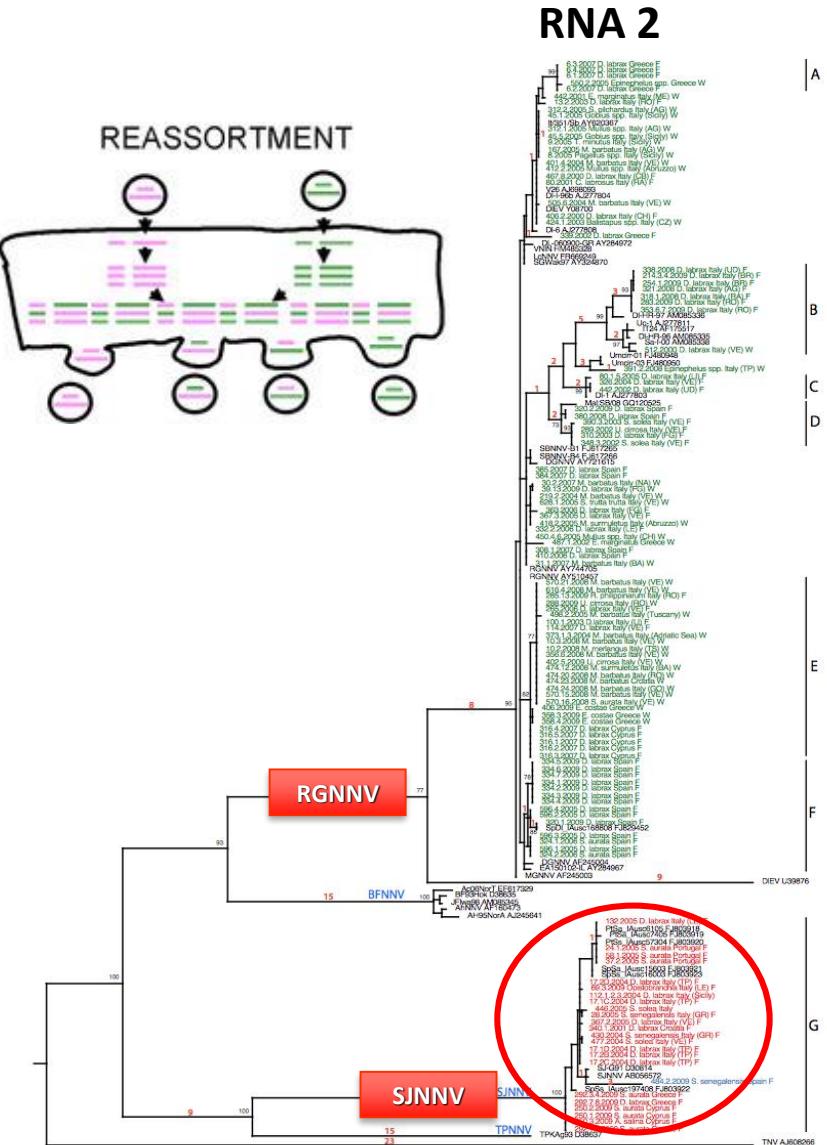
**ITALY
HUNGARY
US, CHINA**

Phylogeny- examples

*Which is the genotype of belonging?
Is a reassortant virus?*



RNA 1

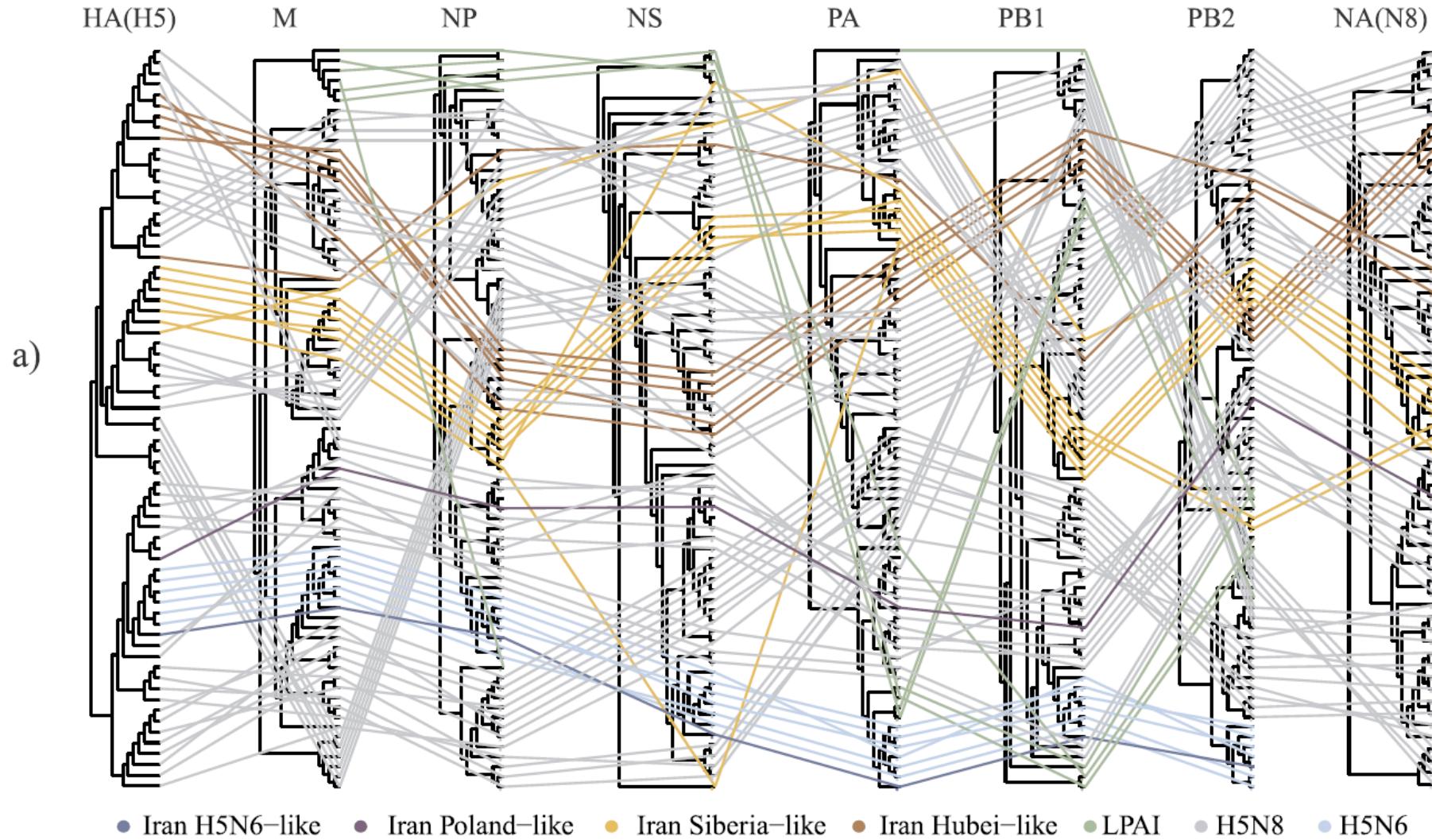


RNA 2

Phylogeny- examples

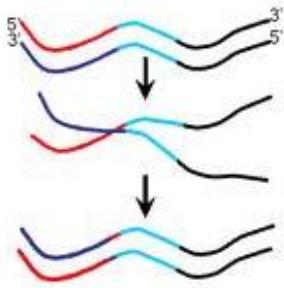
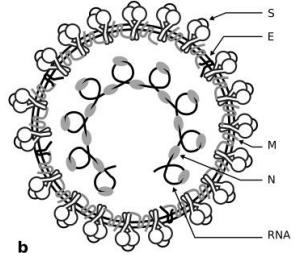
H. Abdollahi, et al.

Infection, Genetics and Evolution 83 (2020) 104342



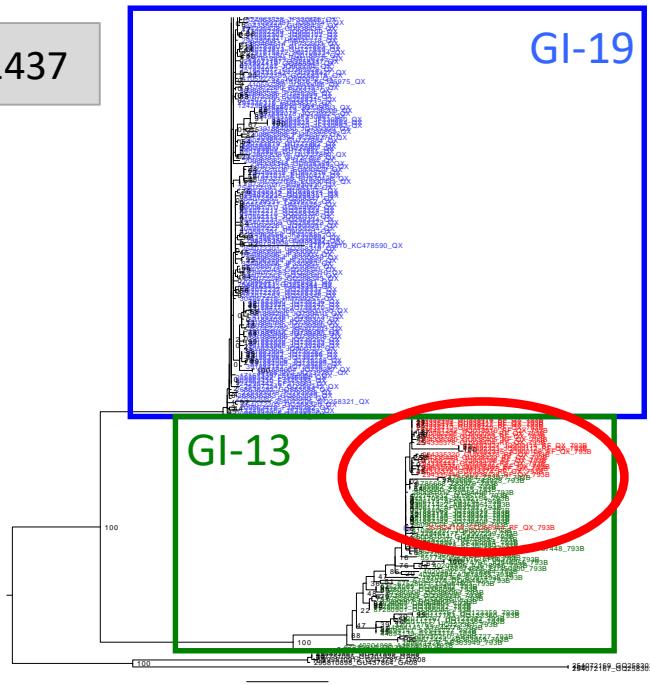
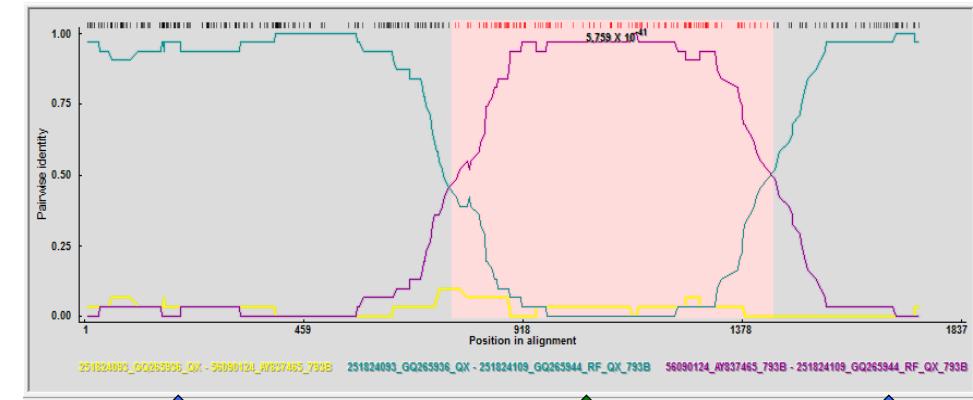
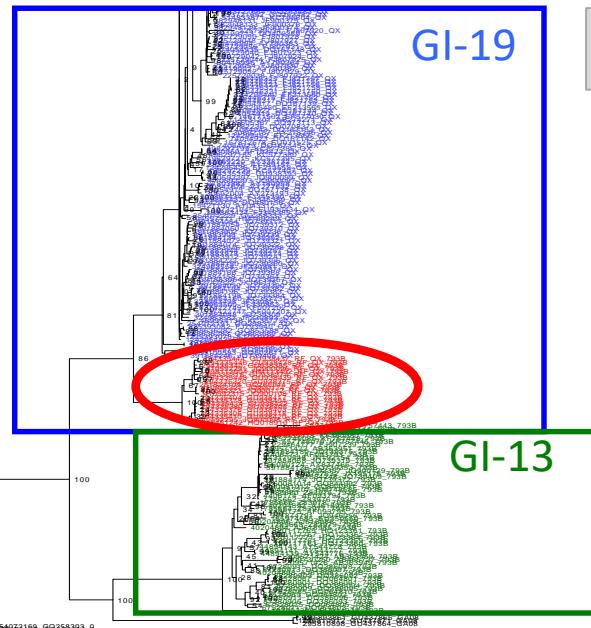
Phylogeny- examples

Is a recombinant virus?



nt START – nt 772
nt 1438 – nt END

IBV
Family: coronavirus
Non-segmented,
positive-sense single-
stranded RNA genome

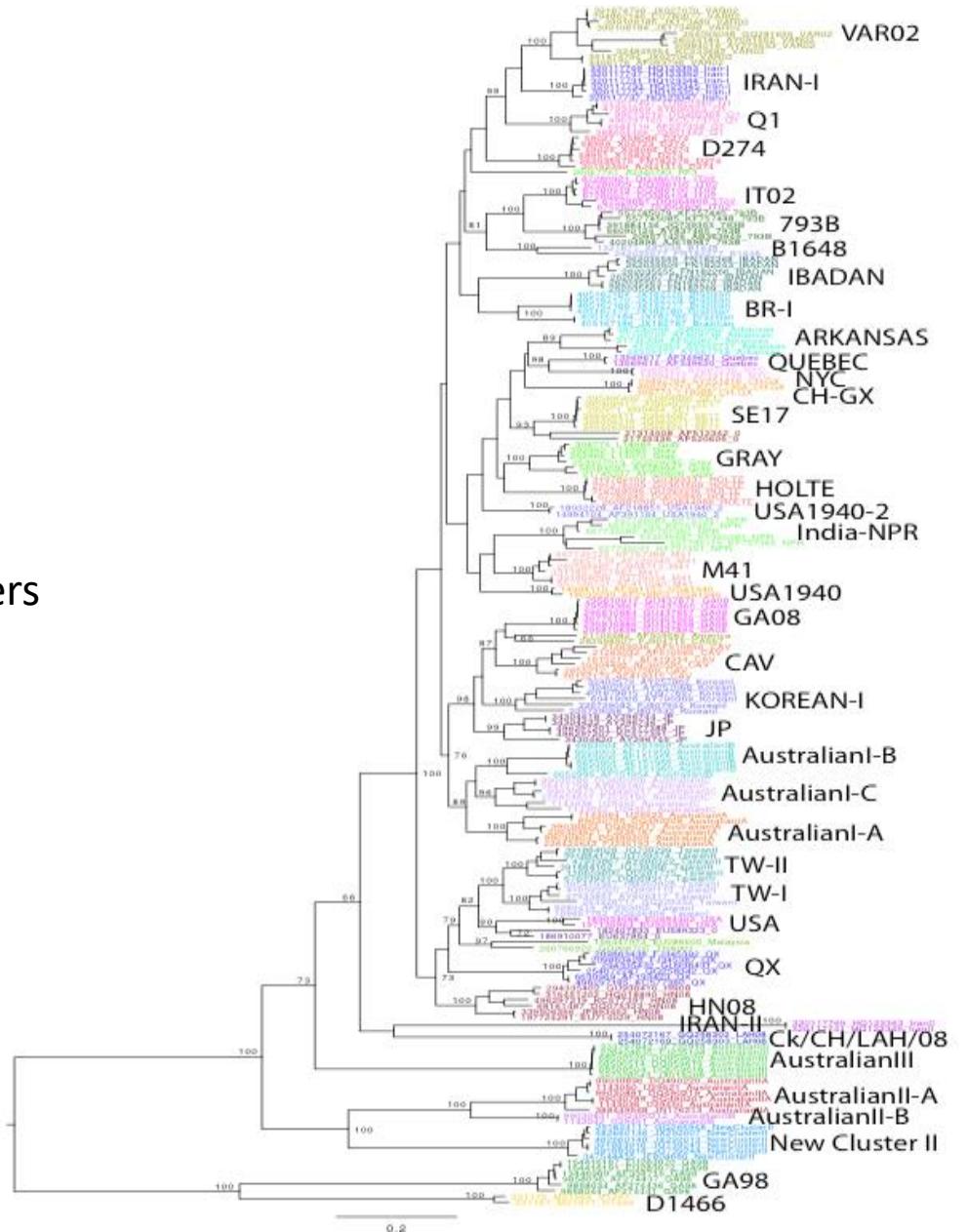
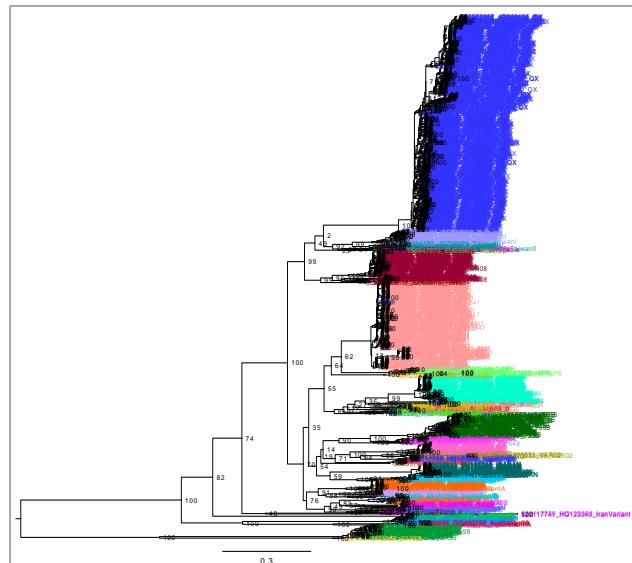


Phylogeny- examples

Classification system based on phylogenetic analysis

IBV
 Family: coronavirus
 Non-segmented,
 positive-sense single-
 stranded RNA genome

Identification of 40 different genetic clusters



Useful Textbooks & Software

Books:

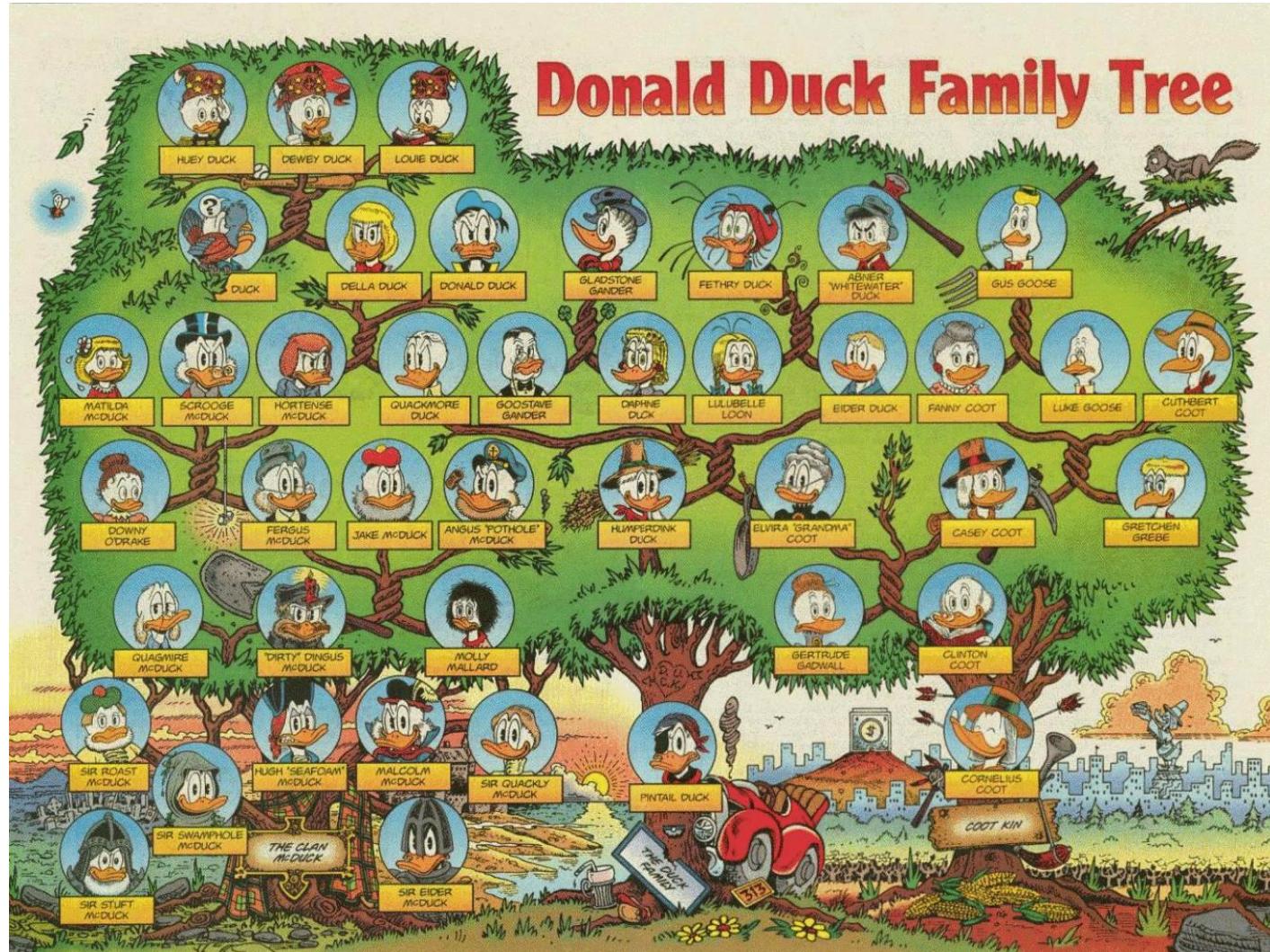
- Nei M & Kumar S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.
- Page RDM & Holmes EC. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science Ltd, Oxford.
- Lemey P, Salemi M & Vandamme A-M. (2009). *The Phylogenetic Handbook, 2nd Edition*. Cambridge University Press.

Computer Software

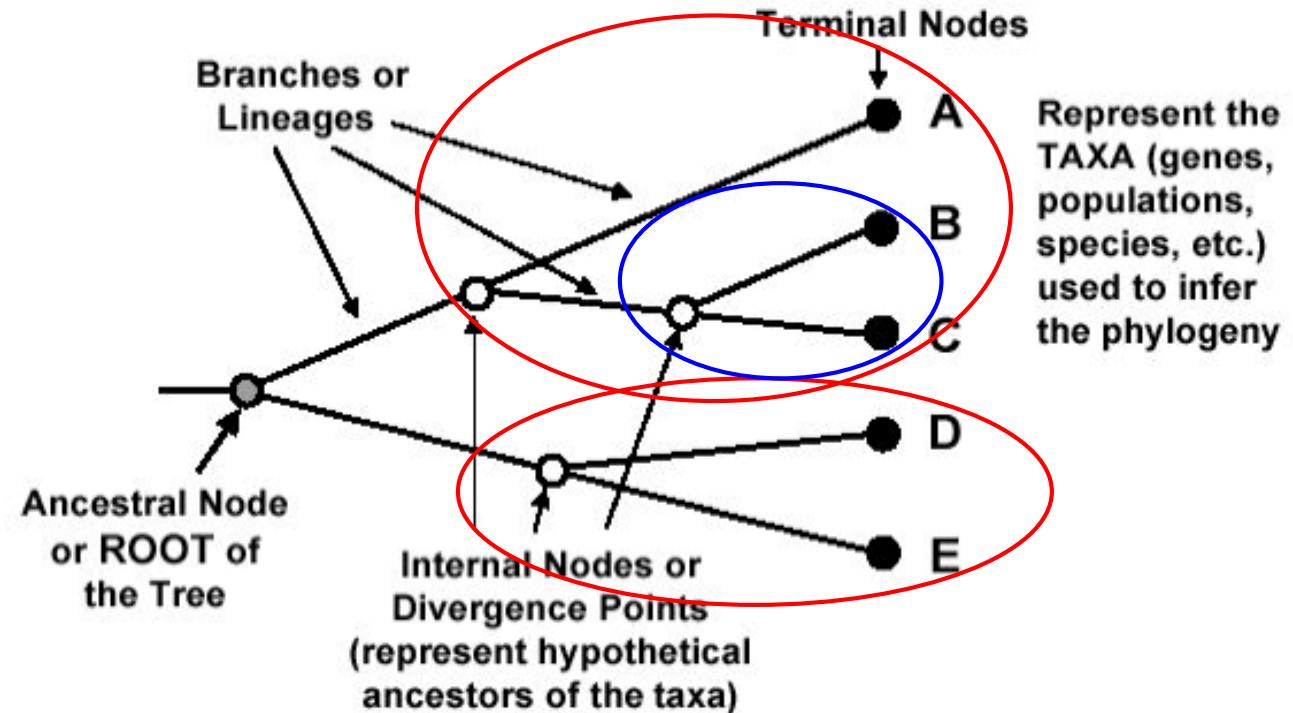
- **MEGA** (Molecular Evolutionary Genetics Analysis) <http://megasoftware.net/>
- **BEAST** (Bayesian Evolutionary Analysis Sampling Trees) <http://beast.bio.ed.ac.uk/>
- **MrBayes** (Bayesian inference of phylogeny) <http://mrbayes.csit.fsu.edu/>
- **PhyML** <http://www.atgc-montpellier.fr/phymil/binaries.php>
- **RaxML** <http://sco.h-its.org/exelixis/software.html>
- **IQTree** <http://www.iqtree.org>

What is a phylogenetic tree?

A phylogenetic tree is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor.



Phylogenetic tree - terminology



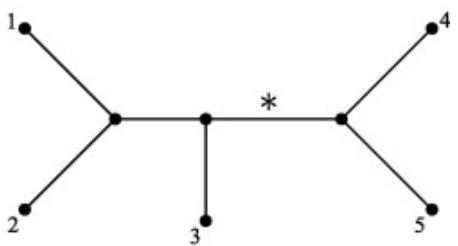
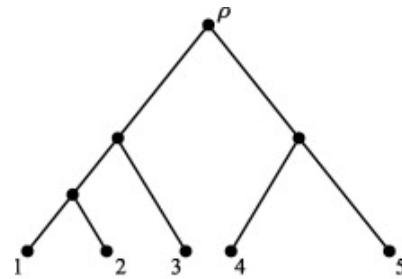
A tree consists of **nodes** connected by **branches**.

Terminal nodes (also called leaves, OTUs (Operational taxonomic Unit) or Terminal taxa) represent sequences or organisms for which we have data.

Internal nodes represent hypothetical ancestors. The ancestor of all the sequences that comprise the tree is the **root** of the tree

Phylogenetic tree - terminology

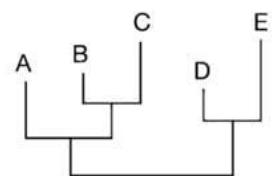
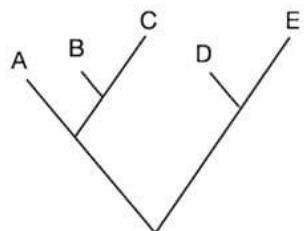
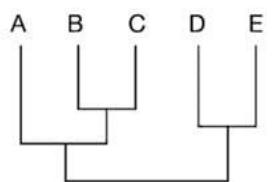
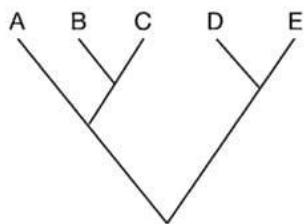
Rooted or unrooted?



Rooted trees are trees that have a specified root node, which represents the common ancestor of all the taxa in the tree.

Unrooted trees do not have a specified root node and show only the branching pattern of the evolutionary relationships among, without any information about their common ancestor.

Cladogram or phylogram?

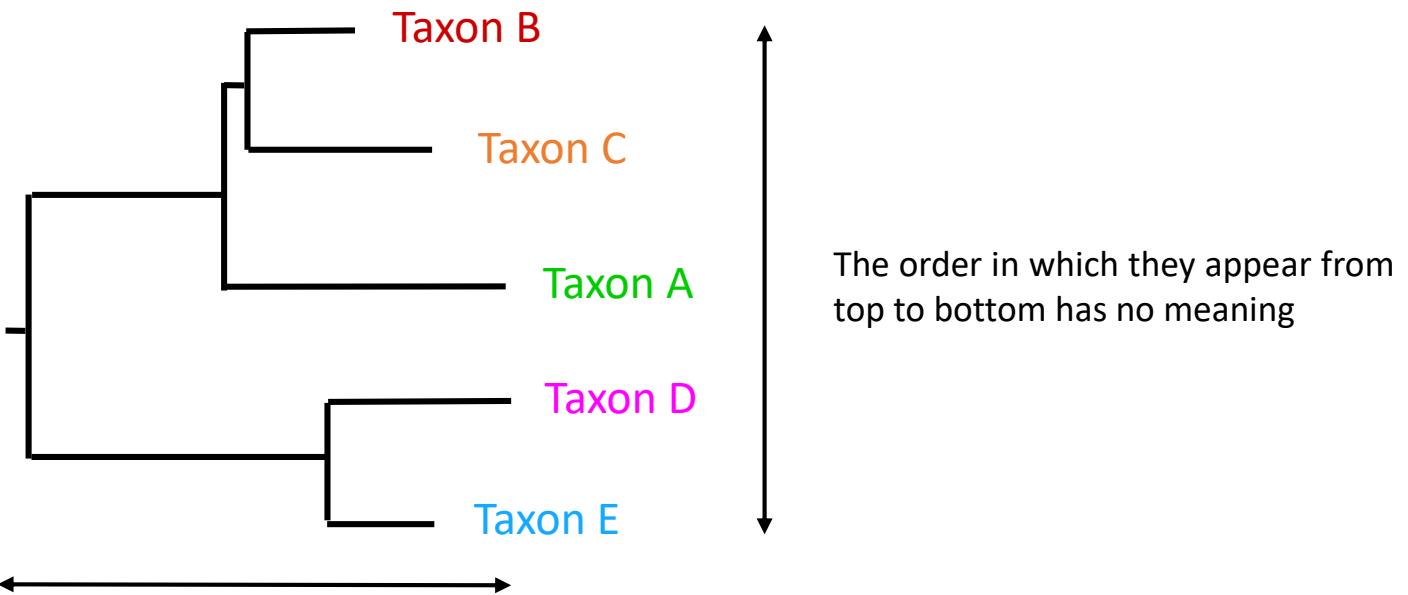


Cladogram is a type of phylogenetic tree that displays only the branching pattern of evolutionary relationships among taxa (unscaled tree: the branch lengths do not reflect the amount of evolutionary divergence between taxa).

Phylogram is a type of phylogenetic tree that represents the evolutionary relationships among taxa by showing both the branching pattern and the amount of evolutionary divergence (scaled tree: the branch lengths are proportional to the amount of evolutionary divergence).

Phylogenetic tree - interpretation

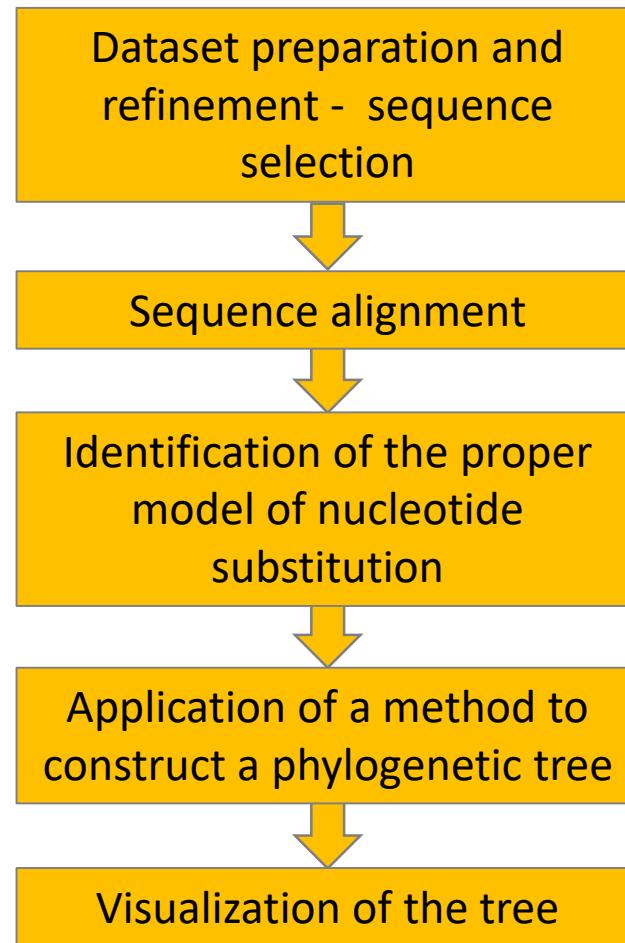
Phylogenetic trees diagram the evolutionary relationships between the taxa



This dimension either can have no scale (for 'cladograms'), can be proportional to **genetic distance or amount of change** (for 'phylogenograms'), or can be proportional to time (for 'time-scaled trees').

This analysis says that **B** and **C** are more closely related to each other than either is to **A**, and that **A**, **B**, and **C** form a clade that is a sister group to the clade composed of **D** and **E**. If the tree has a time scale, then **D** and **E** are the most closely related.

The workflow to construct a phylogenetic tree



Step 1 – sequence selection

Dataset preparation and refinement - sequence selection

GenBank (NCBI)

<http://www.ncbi.nlm.nih.gov/>



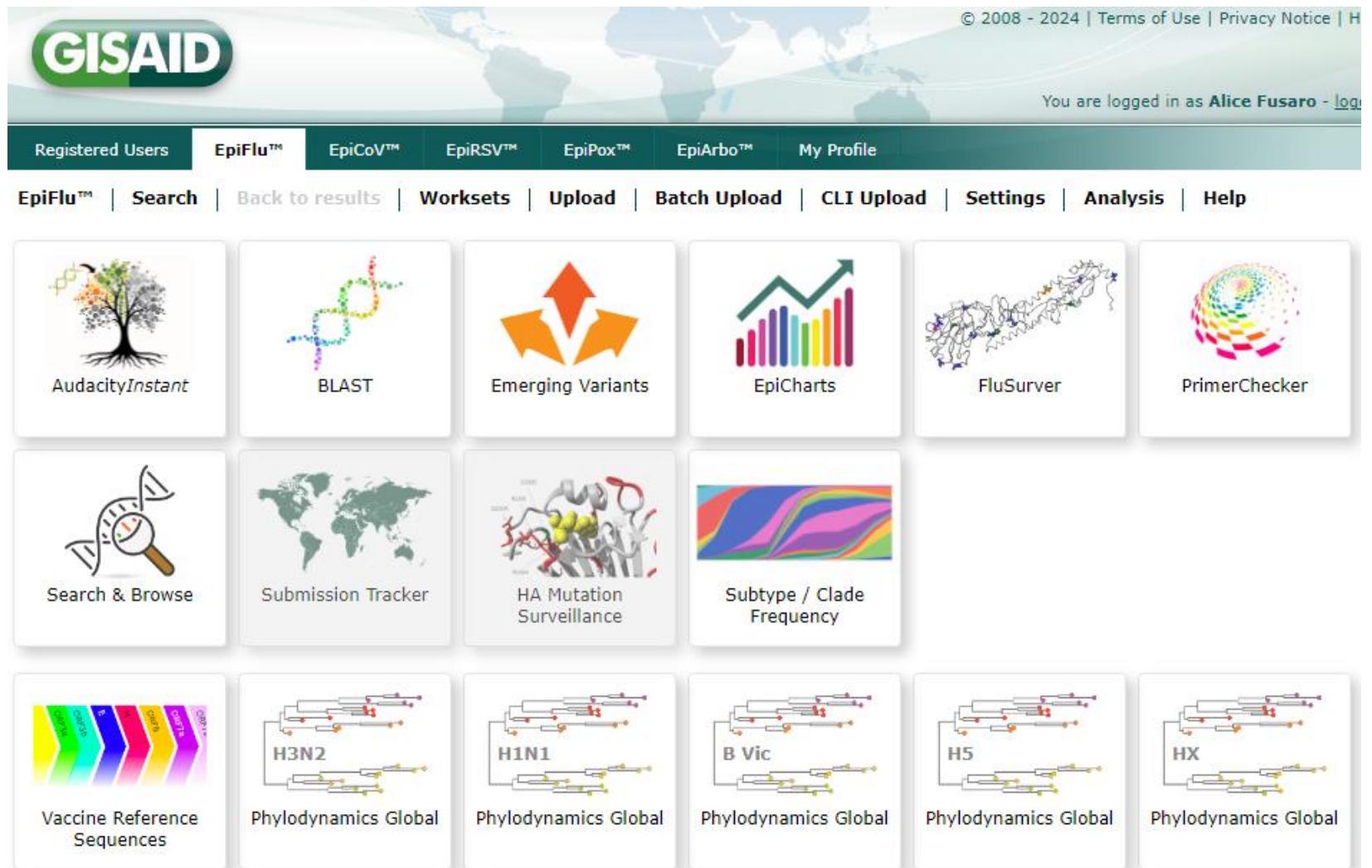
GISAID

<https://www.epicov.org/>



BLAST TOOL: to find related sequences

Step 1 – sequence selection



The screenshot shows the GISAID EpiFlu™ interface. At the top, there is a navigation bar with tabs for Registered Users, EpiFlu™ (which is selected), EpiCoV™, EpiRSV™, EpiPox™, EpiArbo™, and My Profile. Below the navigation bar, there are links for EpiFlu™, Search, Back to results, Worksets, Upload, Batch Upload, CLI Upload, Settings, Analysis, and Help. The main area displays a grid of 18 cards, each representing a different tool or feature:

- AudacityInstant: An icon of a tree with colored leaves.
- BLAST: An icon of a DNA helix.
- Emerging Variants: An icon of an orange upward-pointing arrow inside a V-shape.
- EpiCharts: An icon of a bar chart with a green arrow pointing upwards.
- FluSurver: An icon of a complex molecular structure.
- PrimerChecker: An icon of a colorful circular pattern.
- Search & Browse: An icon of a magnifying glass over a DNA helix.
- Submission Tracker: An icon of a world map.
- HA Mutation Surveillance: An icon of a protein structure with red and yellow segments.
- Subtype / Clade Frequency: An icon of colorful wavy lines.
- Vaccine Reference Sequences: An icon of a series of colored arrows pointing right.
- Phylogenetics Global: An icon of a phylogenetic tree for H3N2.
- Phylogenetics Global: An icon of a phylogenetic tree for H1N1.
- Phylogenetics Global: An icon of a phylogenetic tree for B Vic.
- Phylogenetics Global: An icon of a phylogenetic tree for H5.
- Phylogenetics Global: An icon of a phylogenetic tree for HX.

Step 1 – sequence selection

© 2008 - 2024 | Terms of Use | Privacy Notice | Help

You are logged in as Alice Fusaro - [logout](#)

[Registered Users](#) | **EpiFlu™** | [EpiCoV™](#) | [EpiRSV™](#) | [EpiPox™](#) | [EpiArbo™](#) | [My Profile](#)

[EpiFlu™](#) | [Search](#) | [Back to results](#) | [Worksets](#) | [Upload](#) | [Batch Upload](#) | [CLI Upload](#) | [Settings](#) | [Analysis](#) | [Help](#)

Basic filters

Predefined search

Search in Released files My released files My unreleased files Worksets

Search patterns

Type	H	N
A	1	1
B	2	2
C	3	3
	4	4
	5	5
	6	6
	7	7
	8	8
	9	9
	10	10

Lineage	
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10

Host	
all	-
Human	
Animal	
Avian	
Chicken	
Curlew	
Duck	
Eagle	
Falcon	
Goose	

Location	
all	-
Africa	
Antarctica	
Asia	
Europe	
North America	
Oceania	
South America	

Clades	
0	-
1	
1.1	
1.1.1	
1.1.2	
2.1.1	
2.1.2	
2.1.3	
2.1.3.1	
2.1.3.2	

Pathogenicity

Additional filters

Collection date (YYYY-MM-DD) From To

Submission date (YYYY-MM-DD) From To

Originating Laboratory

Submitting Laboratory

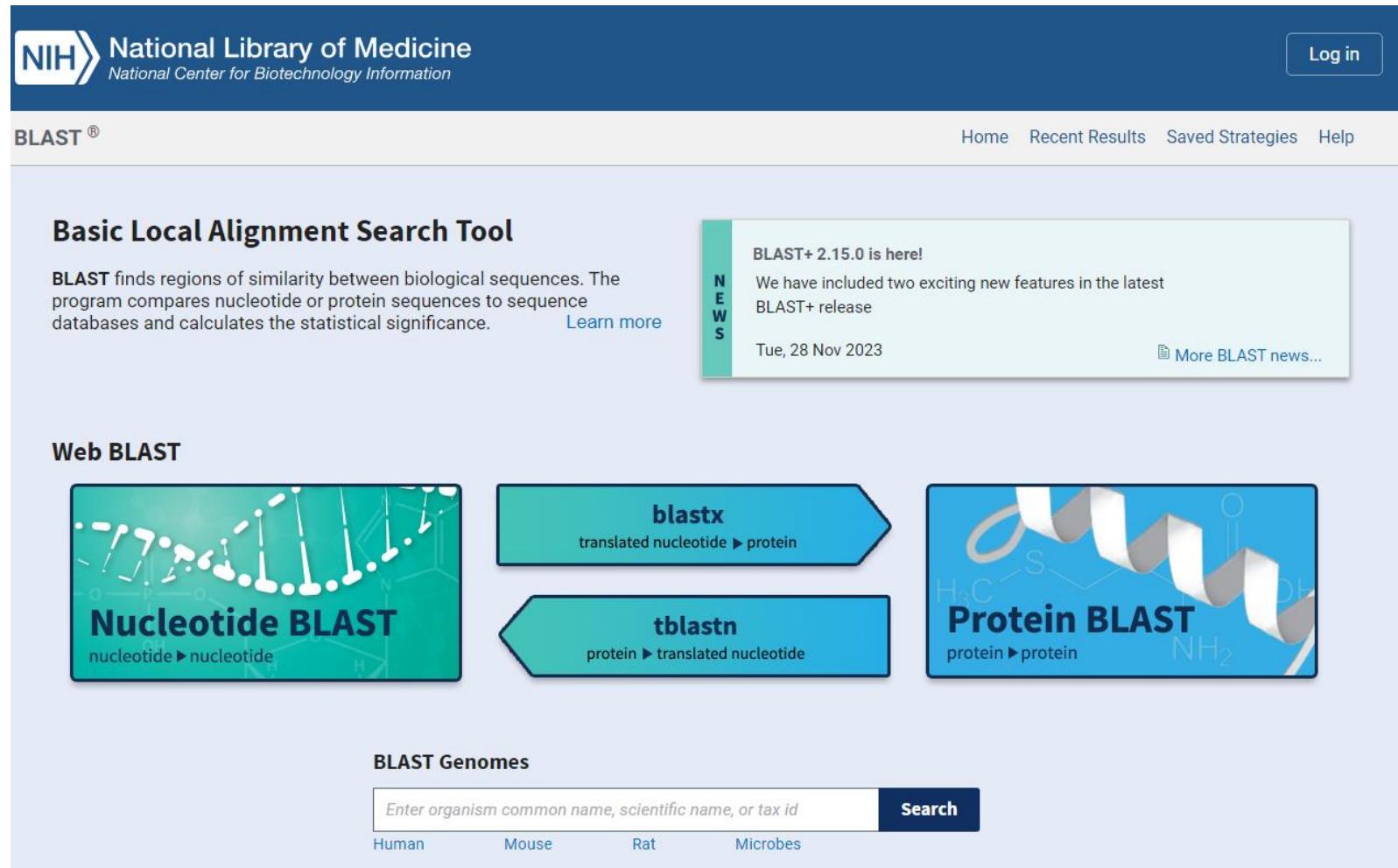
Required Segments PB2 PB1 PA HA NP NA MP NS HE P3 only complete Min Length

Passage details/history

[Help](#) Total: 31,516 viruses (191,630 sequences)

[Charts](#) [Reset](#) [Search](#)

Step 1 – sequence selection



The screenshot shows the official BLAST homepage from the National Library of Medicine. At the top, the NIH logo and "National Library of Medicine" are displayed, along with a "Log in" button. Below the header, the "BLAST®" logo is on the left, and a navigation bar includes "Home", "Recent Results", "Saved Strategies", and "Help".

The main content area features a "Basic Local Alignment Search Tool" section. It describes what BLAST does and links to "Learn more". To the right, a teal vertical bar labeled "NEWS" contains a news item about "BLAST+ 2.15.0 is here!" with a release date of "Tue, 28 Nov 2023" and a link to "More BLAST news...".

The "Web BLAST" section highlights three search tools: "Nucleotide BLAST" (nucleotide → nucleotide), "blastx" (translated nucleotide → protein), and "tblastn" (protein → translated nucleotide). To the right is the "Protein BLAST" tool (protein → protein), which features a 3D protein structure graphic.

The bottom section, "BLAST Genomes", includes a search bar with placeholder text "Enter organism common name, scientific name, or tax id", a "Search" button, and links for "Human", "Mouse", "Rat", and "Microbes".

Step 1 – sequence selection

BLASTN programs search nucleotide databases using a nucleotide query. mo

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

```
CGTGCATGCAGAGAAGAGCC
GGTTGGCATCAGAAGAACAAATCCTAAGGGCAGCCACGTGATCTACGG
GGCTCCAGGCAGGGCAGAGCC
ACCCCAGGCCCTCATAGACGAAGTCGCCAAAGTCTATGAAATTAAACQ
```

Or, upload file [Scegli file](#) Nessun file selezionato [?](#)

Job Title
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus
[New](#) Experimental databases Try experimental taxonomic nt databases [Download](#)
 For more info see [What are taxonomic nt databases?](#)

Organism Optional Nucleotide collection (nr/nt)
 Enter organism name or id—completions will be suggested exclude [Add organism](#)
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez Query Optional [YouTube](#) Create custom database
 Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn)
 Choose a BLAST algorithm [?](#)

BLAST Search database nt using Megablast (Optimize for highly similar sequences)
 Show results in a new window

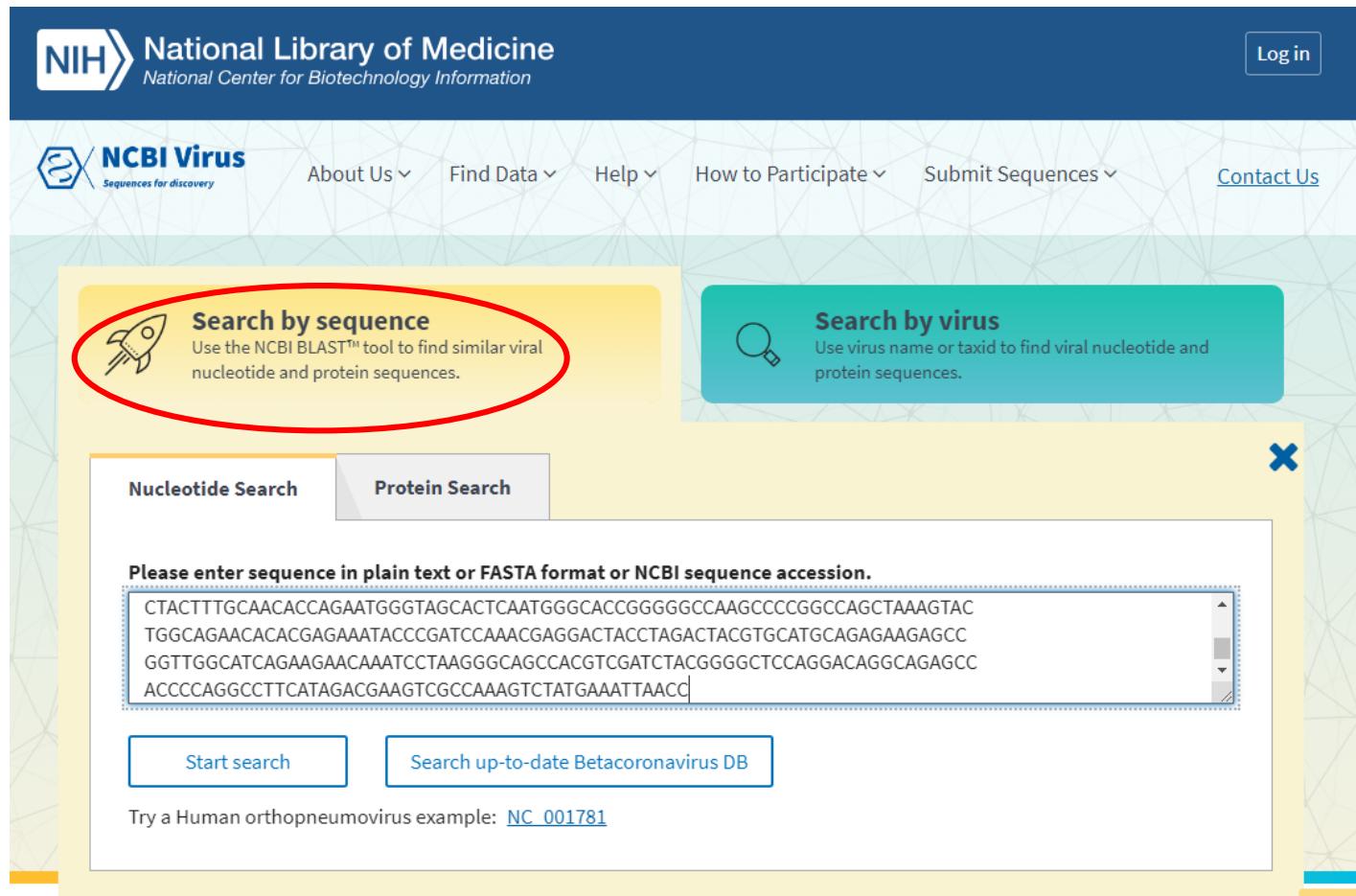
Step 1 – sequence selection

BLAST results

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Infectious bursal disease virus strain ks segment A polyprotein mRNA, complete cds	Infectious bursal...	4518	4518	100%	0.0	96.79%	3163	DQ927042.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus isolate Ventri-IBDV-Plus polyprotein mRNA, complete cds	Infectious bursal...	4495	4495	100%	0.0	96.64%	3040	KJ547670.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus isolate IBDV78/ABICvaccine VP5 and polyprotein genes, partial cds	Infectious bursal...	4495	4495	100%	0.0	96.64%	3080	JX424077.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus strain mb segment A polyprotein mRNA, complete cds	Infectious bursal...	4495	4495	100%	0.0	96.64%	3163	DQ927040.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus isolate A333D VP5 protein (VP5) and polyprotein genes, complete cds	Infectious bursal...	4484	4484	100%	0.0	96.56%	3073	OR206449.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus isolate A350A VP5 protein (VP5) and polyprotein genes, complete cds	Infectious bursal...	4479	4479	100%	0.0	96.53%	3073	OR206450.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus segment A viral protein 5 and polyprotein mRNA, complete cds	Infectious bursal...	4479	4479	100%	0.0	96.53%	3260	AF240686.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus isolate 89163 polyprotein gene, partial cds	Infectious bursal...	4457	4457	100%	0.0	96.38%	3130	ON100652.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus segment A, complete genome, strain 89163	Infectious bursal...	4457	4457	100%	0.0	96.38%	3198	HG974563.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus, segment A, isolate SH99	Infectious bursal...	4451	4451	100%	0.0	96.34%	3261	LM651365.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus segment A, complete sequence	Infectious bursal...	4451	4451	100%	0.0	96.34%	3260	AY598356.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus VP5 protein and structural polyprotein genes, complete cds	Infectious bursal...	4451	4451	100%	0.0	96.34%	3260	AY444873.3
<input checked="" type="checkbox"/>	Infectious bursal disease virus strain 3529/92 polyprotein mRNA, complete cds	Infectious bursal...	4440	4440	100%	0.0	96.27%	3039	KC189836.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus segment A VP5 and polyprotein genes, complete cds	Infectious bursal...	4440	4440	100%	0.0	96.27%	3260	EU595667.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus isolate 91168 polyprotein gene, partial cds	Infectious bursal...	4434	4434	100%	0.0	96.23%	3130	ON100653.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus isolate Harbin-1 segment A, complete sequence	Infectious bursal...	4434	4434	100%	0.0	96.23%	3261	EF517528.1
<input checked="" type="checkbox"/>	Infectious bursal disease virus strain SH/92 polyprotein mRNA, complete cds	Infectious bursal...	4433	4433	100%	0.0	96.20%	3039	AF533670.1

Step 1 – sequence selection

NCBI virus database



The screenshot shows the NCBI Virus database homepage. At the top, the NIH National Library of Medicine logo and the NCBI Virus logo are visible. The main navigation menu includes links for About Us, Find Data, Help, How to Participate, Submit Sequences, and Contact Us. Two prominent search options are displayed: "Search by sequence" (with a red oval highlighting the button) and "Search by virus". Below these, there are tabs for Nucleotide Search and Protein Search. A large input field is provided for entering a sequence in plain text or FASTA format, with a sample sequence (a long string of nucleotides) already entered. Two search buttons are present: "Start search" and "Search up-to-date Betacoronavirus DB". A note at the bottom suggests trying a Human orthopneumovirus example with accession number NC_001781. The footer of the page reads "NCBI Visual Data Dashboard".

Step 1 – sequence selection

NCBI virus database

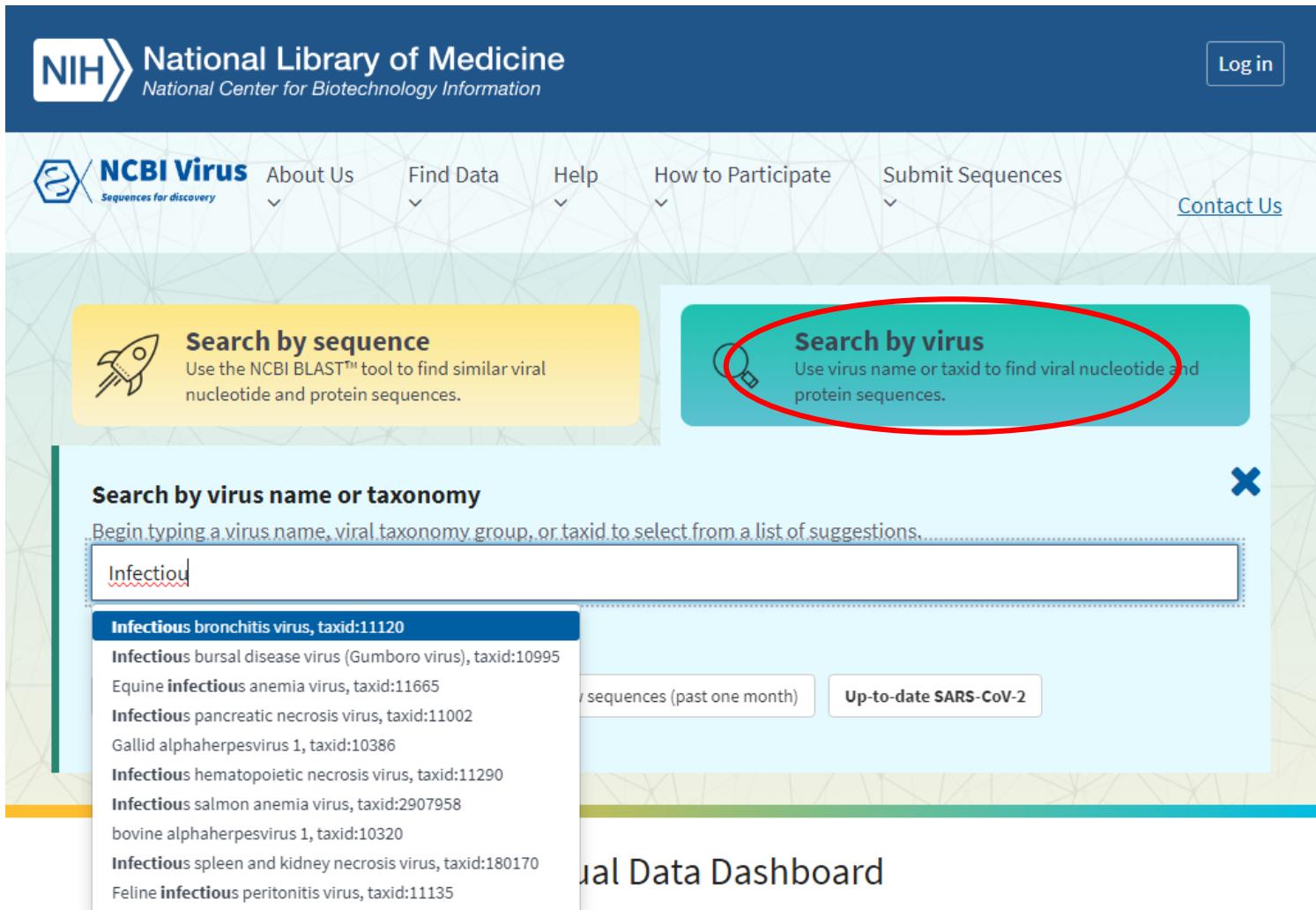
Refine Results		Reset
Virus	+ 	
Accession	+ 	
Sequence Length	+ 	
Ambiguous Characters	+ 	
Sequence Type	+ 	
RefSeq Genome Completeness	+ 	
Nucleotide Completeness	+ 	
Isolate	+ 	
Genotype <small>New!</small>	+ 	
Proteins	+ 	
Provirus	+ 	

Expand Table

Nucleotide (500)	Protein (0)	RefSeq Genome (0)				Select Columns
Accession 	Organism Name 	Submitters 	Organization 	Release Date 	Isolate 	
<input type="checkbox"/> NC_077769 <small>RefSeq</small>	Infectious bursal disease ...	Mundt,E., et al.	National Center for Biotech...	2023-05-06	1977, from Schobri	
<input type="checkbox"/> NC_004178 <small>RefSeq</small>	Infectious bursal disease ...	Brown,M.D., et al.	National Center for Biotech...	1996-05-03	UK661	
<input type="checkbox"/> PP108638	Infectious bursal disease ...	Le,T.H., et al.	Institute of Biotechnology...	2024-02-01		
<input type="checkbox"/> PP108640	Infectious bursal disease ...	Le,T.H., et al.	Institute of Biotechnology...	2024-02-01		
<input type="checkbox"/> PP108642	Infectious bursal disease ...	Le,T.H., et al.	Institute of Biotechnology...	2024-02-01		
<input type="checkbox"/> PP108644	Infectious bursal disease ...	Le,T.H., et al.	Institute of Biotechnology...	2024-02-01		
<input type="checkbox"/> OR960825	Infectious bursal disease ...	Chang,H., et al.	China Agricultural Univer...	2023-12-20	HB2023	
<input type="checkbox"/> OR523680	Infectious bursal disease ...	Wang,W., et al.	Guangxi University, Instit...	2023-12-18	GXYL211225	
<input type="checkbox"/> OR743230	Infectious bursal disease ...	Nie,J., et al.	Jiangxi Agricultural Unive...	2023-12-13		
<input type="checkbox"/> OM046608	Infectious bursal disease ...	Sultan,H., et al.	Faculty of Veterinary Medi...	2023-11-30	IBDV/Layer/Benisoi	

Step 1 – sequence selection

NCBI virus database



The screenshot shows the NCBI Virus database homepage. At the top, the NIH National Library of Medicine logo is displayed, along with a 'Log in' button. Below the header, the NCBI Virus logo ('Sequences for discovery') is shown, followed by navigation links: About Us, Find Data, Help, How to Participate, Submit Sequences, and Contact Us. Two main search options are presented: 'Search by sequence' (using BLAST) and 'Search by virus' (using virus name or taxid). The 'Search by virus' option is highlighted with a red oval. Below these, a search bar allows users to type in a virus name or taxid, with suggestions appearing as they type. The suggestions for 'Infectiou' include: Infectious bronchitis virus, taxid:11120; Infectious bursal disease virus (Gumboro virus), taxid:10995; Equine infectious anemia virus, taxid:11665; Infectious pancreatic necrosis virus, taxid:11002; Gallid alphaherpesvirus 1, taxid:10386; Infectious hematopoietic necrosis virus, taxid:11290; Infectious salmon anemia virus, taxid:2907958; bovine alphaherpesvirus 1, taxid:10320; Infectious spleen and kidney necrosis virus, taxid:180170; and Feline infectious peritonitis virus, taxid:11135.

NIH National Library of Medicine
National Center for Biotechnology Information

NCBI Virus Sequences for discovery

About Us Find Data Help How to Participate Submit Sequences Contact Us

Search by sequence Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.

Search by virus Use virus name or taxid to find viral nucleotide and protein sequences.

Search by virus name or taxonomy

Begin typing a virus name, viral taxonomy group, or taxid to select from a list of suggestions.

Infectiou

- Infectious bronchitis virus, taxid:11120
- Infectious bursal disease virus (Gumboro virus), taxid:10995
- Equine infectious anemia virus, taxid:11665
- Infectious pancreatic necrosis virus, taxid:11002
- Gallid alphaherpesvirus 1, taxid:10386
- Infectious hematopoietic necrosis virus, taxid:11290
- Infectious salmon anemia virus, taxid:2907958
- bovine alphaherpesvirus 1, taxid:10320
- Infectious spleen and kidney necrosis virus, taxid:180170
- Feline infectious peritonitis virus, taxid:11135

Search by sequence (past one month) Up-to-date SARS-CoV-2

Virtual Data Dashboard

Step 1 – sequence selection

NCBI virus database

Expand Table

	Nucleotide (15,513)	Protein (22,036)	RefSeq Genome (2)						Select Columns
Virus	+ Infectious bronchitis virus, taxid:11120								
Accession	+ NC_048213	Organism Name	Submitters	Organization	Release Date	Isolate	Species	Length	Nuc Completeness
Sequence Length	+ NC_001451	Infectious bronchitis virus	Senthil Kuma...	National Center for Biotech...	2020-05-15	Ind-TN92-03	Avian coronavirus	27464	compl
Ambiguous Characters	+ OQ791138	Infectious bronchitis virus	Ziebuhr,J., et al.	National Center for Biotech...	1993-06-12		Avian coronavirus	27608	compl
Sequence Type	+ PP729068	Infectious bronchitis virus	Ramiz,R.M., e...	UVAS, Lahore, Microbiolo...	2024-05-11	Chk/UVAS/PK/2022-R4	Avian coronavirus	1601	parti
RefSeq Genome Completeness	+ PP737794	Infectious bronchitis virus	Zhang,H.H., e...	Henan Agricultural Univ...	2024-05-11	C2023-1	Avian coronavirus	27229	compl
Nucleotide Completeness	+ OQ304463	Infectious bronchitis virus	Ramiz,R.M., e...	UVAS, Lahore, Microbiolo...	2024-05-07	IBV/QOL/UVAS/PK/2022/2	Avian coronavirus	566	parti
Isolate	+ OQ304464	Infectious bronchitis virus	Ramiz,R.M., e...	UVAS, Lahore, Microbiolo...	2024-05-07	IBV/QOL/UVAS/PK/2022/4	Avian coronavirus	562	parti
Genotype	+ PP374733	Infectious bronchitis virus	Le,H.D., et al.	Animal and Plant Quarant...	2024-05-05	IBV/Korea/AD04	Avian coronavirus	27647	compl
Proteins	+ PP374734	Infectious bronchitis virus	Le,H.D., et al.	Animal and Plant Quarant...	2024-05-05	IBV/Korea/AQ10	Avian coronavirus	27710	compl
Provirus	+ OQ790034	Infectious bronchitis virus	Yuan,W.	South China Agricultural ...	2024-05-01	210096GXNN	Avian coronavirus	1620	parti
Geographic Region	+ OQ790038	Infectious bronchitis virus	Yuan,W.	South China Agricultural ...	2024-05-01	210087GXYL	Avian coronavirus	1620	parti
Host	+ OQ790039	Infectious bronchitis virus	Yuan,W.	South China Agricultural ...	2024-05-01	210196GXYL	Avian coronavirus	1620	parti
Submitters	+ OQ790040	Infectious bronchitis virus	Yuan,W.	South China Agricultural ...	2024-05-01	210090GXYL	Avian coronavirus	1620	parti
Isolation Source	+ OQ790041	Infectious bronchitis virus	Yuan,W.	South China Agricultural ...	2024-05-01	210092GXNN	Avian coronavirus	1620	parti

Step 2 – alignment

Dataset preparation and refinement - sequence selection

Sequence alignment

NOT Aligned

A_migratoryduck_Jiangxi_2136_2005	CCCAGGGAAATTCAACGACTATGAA
A_Bar-headedGoose_Qinghai_60_05	TGTTACCCAGGGAAATTCAACGACT
A_domesticgoose_Pavlodar_1_2005	CAGGGAAATTCAACGACTATGAA
A_migratoryduck_Jiangxi_2295_2005	CTCTGTTACCCAGGGAAATTCAACG
A_chicken_India_NIV33491_06	TCTGTTACCCAGGGAAATTCAACGA
A_chicken_Nigeria_1047_30_2006	CCAGGCCAATGACCTCTGTTACCCAG
A_guineafowl_Nigeria_957_12_2006	CAGGGAAATTCAACGACTATGAAAGA

Aligned

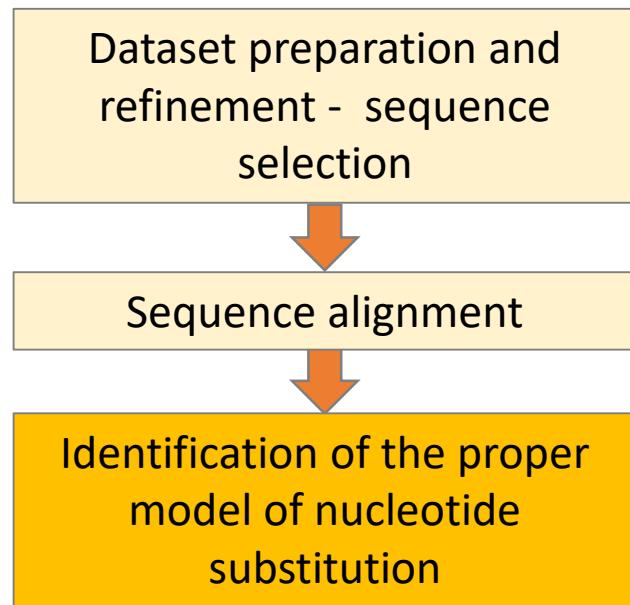
A_whooperswan_Aomori_1_2008	AAAAGATTGTAGTGTAGCAGGATGG
A_whooperswan_Aomori_2_2008	AAAAGATTGTAGTGTAGCAGGATGG
A_littleegret_HongKong_8863_2007	AAAAGATTGTAGTGTAGCAGGATGG
A_commonbuzzard_HongKong_9213_2007	AAAAGATTGTAGTGTAGCAGGATGG
A_blackcrownednightheron_HongKong	AAAAGATTGTAGTGTAGCAGGATGG
A_environment_DongtingLake_Hunan_5_4	AAAAGATTGTAGTGTAGCAGGATGG
A_duck_Angthong_NIAH8246_2004	GAGAGATTGTAGTGTAGCTGGATGG

An alignment program seeks an arrangement that **maximizes the net score**

Specific computer programs create the “the best” alignment of the sequences according to different algorithms

PROGRAMS: MAFFT, Clustal X, Muscle.

Step 3 – model of nucleotide substitution



Step 3 – model of nucleotide substitution

Alignment – kind of substitutions

Given two or more nt or amino acid sequences, usually the first goal is to calculate some measure of sequence dissimilarity.

The easiest way to estimate genetic distances: **p-distance** (number of nt differences between two sequences divided by the sequence length)

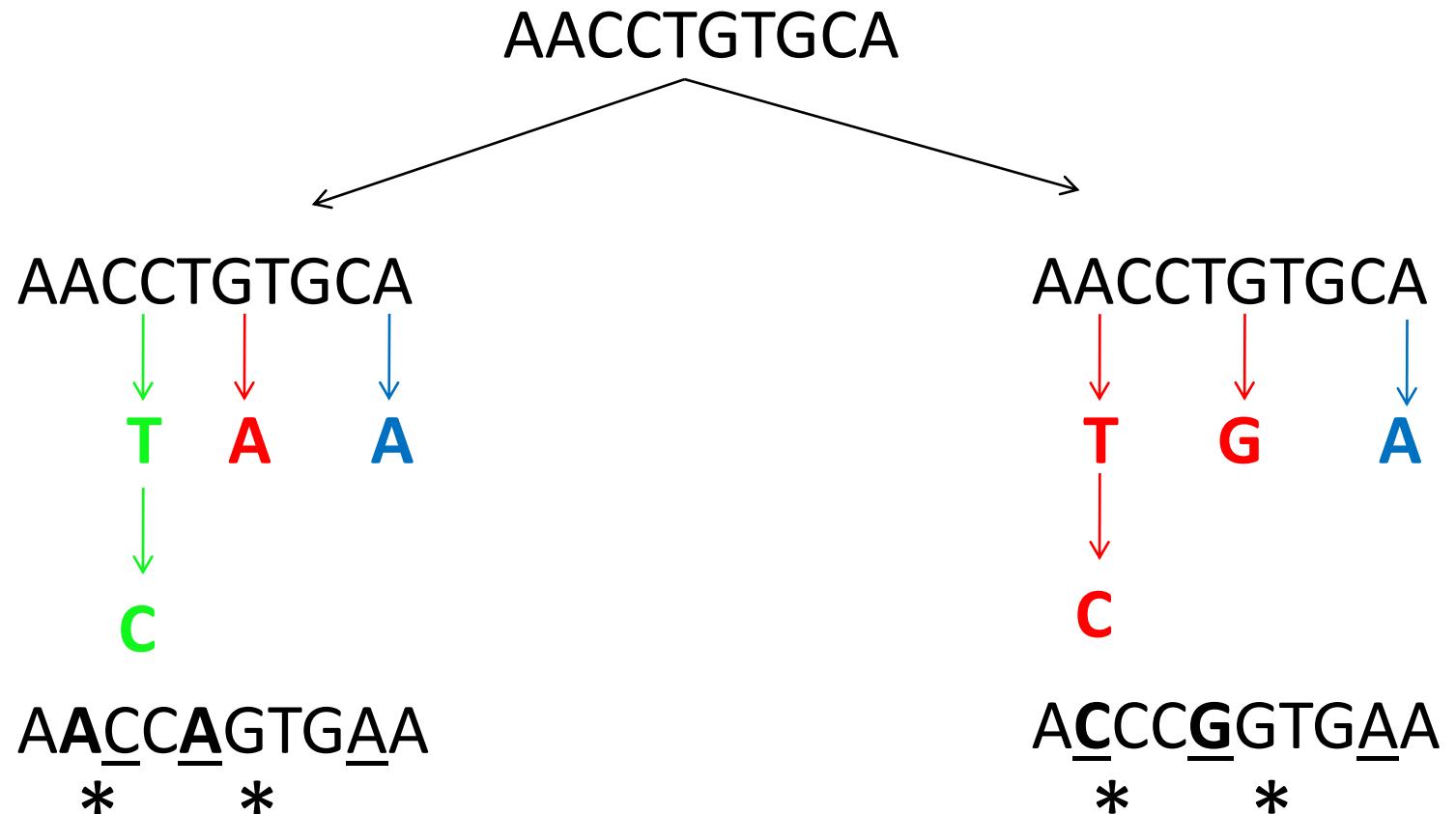
Seq1 AATCTGTGTA
Seq2 ATCCTGGGTT

P-distance=0.4

Usually underestimate the true distance: ***genetic (or evolutionary) distance d***

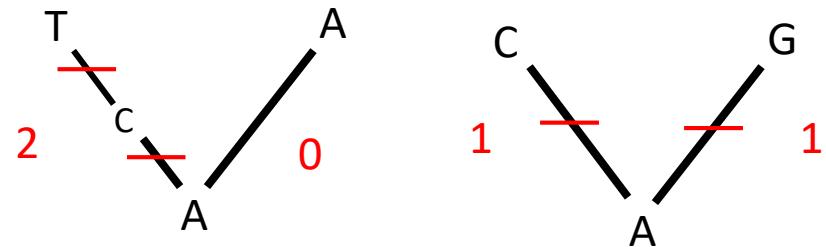
Step 3 – model of nucleotide substitution

Multiple, parallel, and back-substitutions



Step 3 – model of nucleotide substitution

Multiple substitutions can greatly obscure the actual evolutionary history of a pair of sequences.

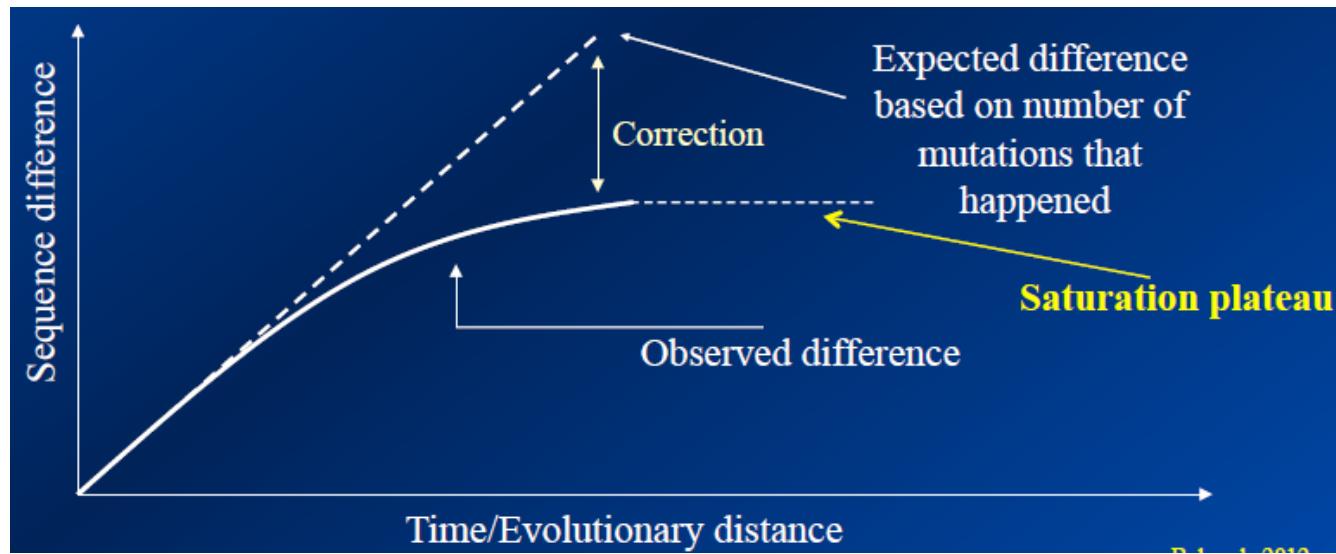


In both cases the 2 descendant sequences show only a single difference, while there have been 2 substitutions. Simply counting the number of differences between the 2 sequences underestimates the real amount of evolutionary changes

Models of DNA sequence evolution are required to recover the missing information through correcting for multiple substitutions.

Step 3 – model of nucleotide substitution

As the number of substitutions increases the probability that the same site may undergo more than one substitution become higher.



- When the difference between actual and observed divergence is low, observed distance (p) is a good estimator of genetic distance (d)
- When the difference between actual and observed divergence is high, p underestimates d and a “correction statistic” is required i.e. a model of DNA substitution

Step 3 – model of nucleotide substitution

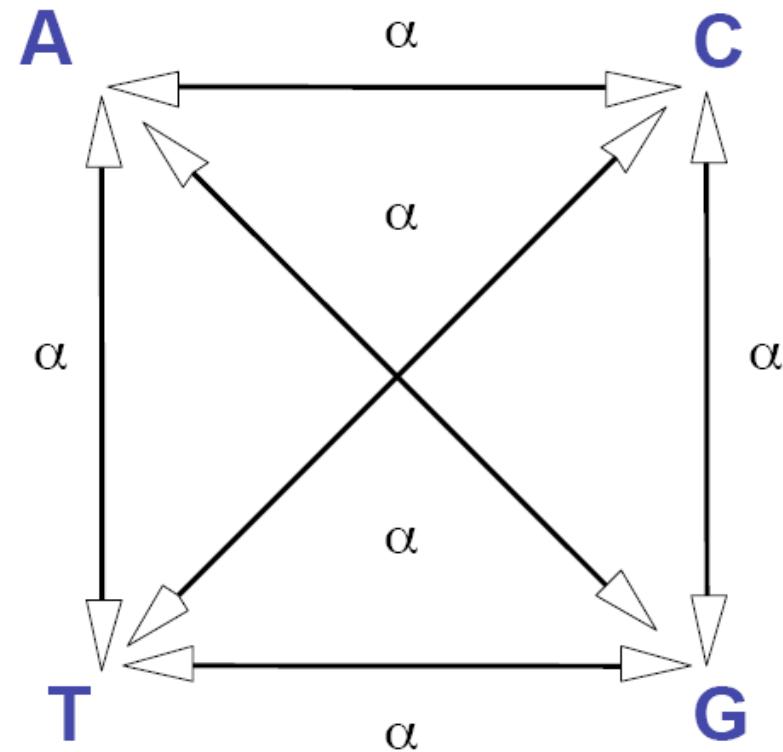
Jukes – Cantor (JC)

Assumptions:

- i. All bases evolve independently
- ii. All bases are at equal frequency
- iii. Each base can change with equal probability (α)

All substitution occur at the same rate (α)

Is this model too simple for real data?



Step 3 – model of nucleotide substitution

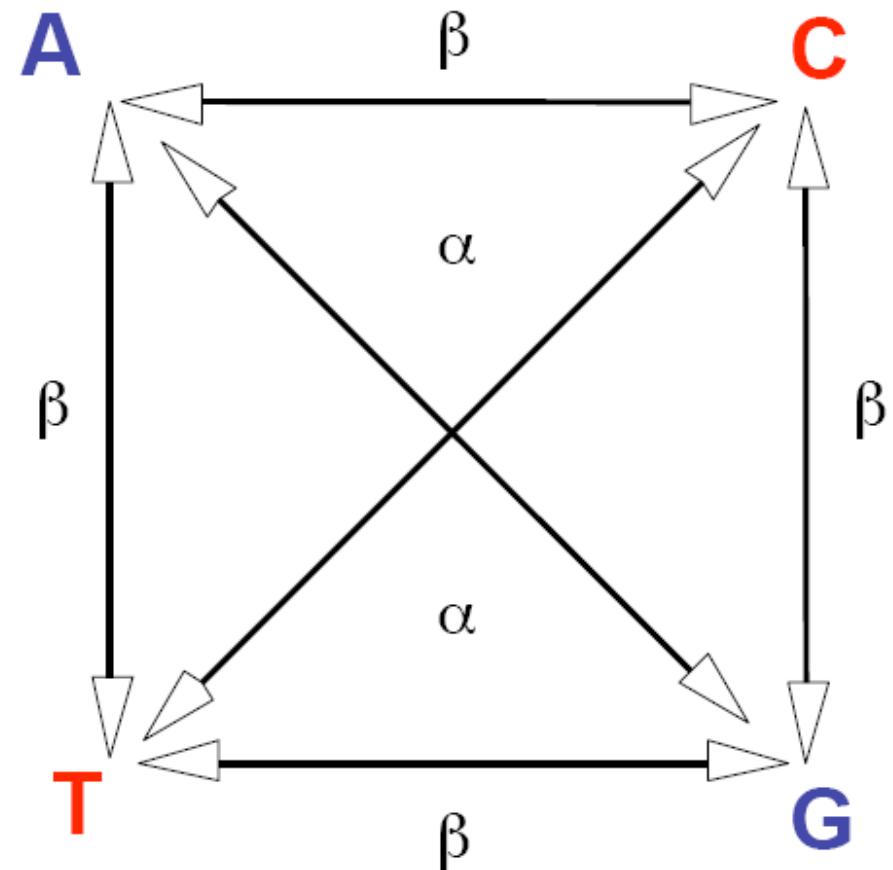
In Real Data, Transitions (α) and Transversions (β) occur at different rates

Transitions are generally more frequent than transversions

Kimura's 2 parameter (K2P)

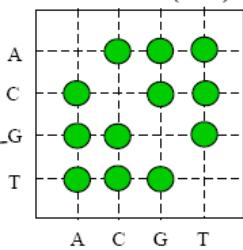
Assumptions:

- i. All bases evolve independently
- ii. All bases are at equal frequency
- iii. Transitions and transversions occur with different probabilities (α and β)
- iv. The Jukes-Cantor model is applied to transitions and transversions independently

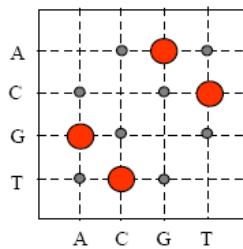


Step 3 – model of nucleotide substitution

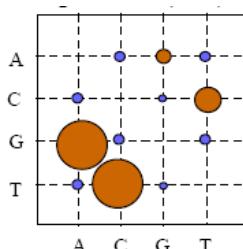
Simplest (few parameters)



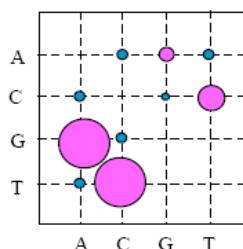
Base frequencies are equal and all substitutions are equally likely (Jukes-Cantor)



Base frequencies are equal but transitions and transversions occur at different rates (K2P)



Unequal base frequencies and transitions and transversions occur at different rates (Hasegawa-Kishino-Yano: HKY85)

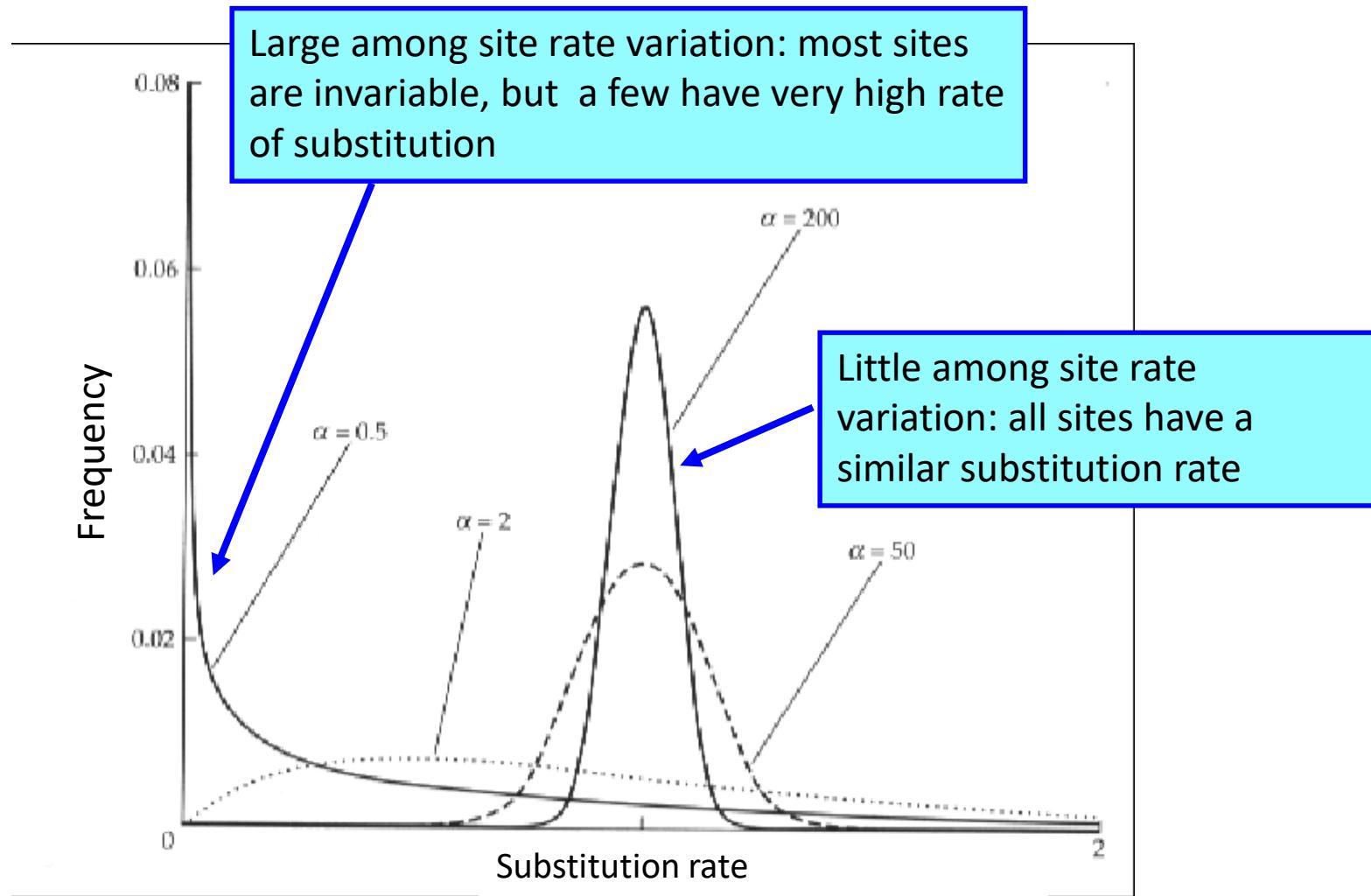


Unequal base frequencies and all substitution types occur at different rates (General Time Reversible Model: GTR)

Most complex (many parameters)



Step 3 – model of nucleotide substitution



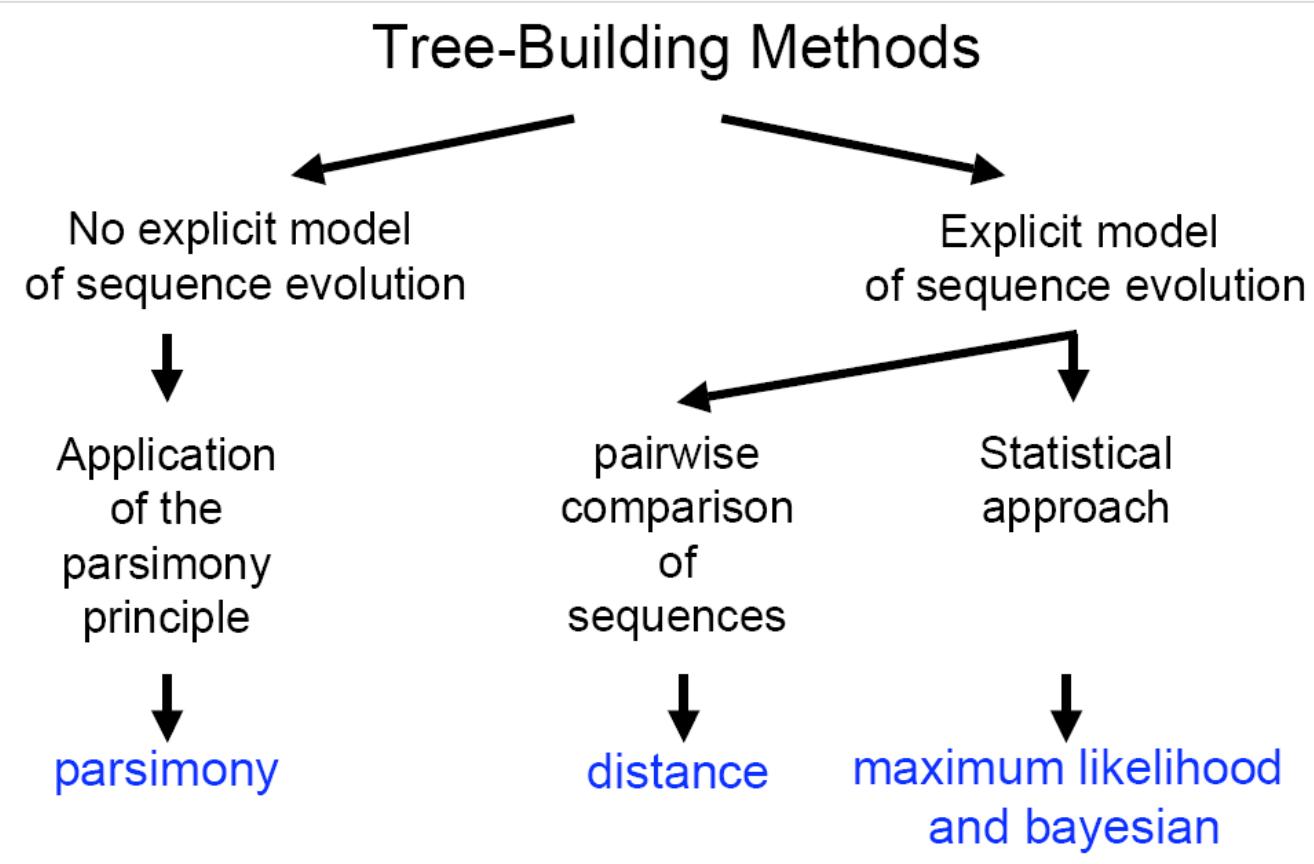
Step 4 – method selection

Dataset preparation and refinement - sequence selection

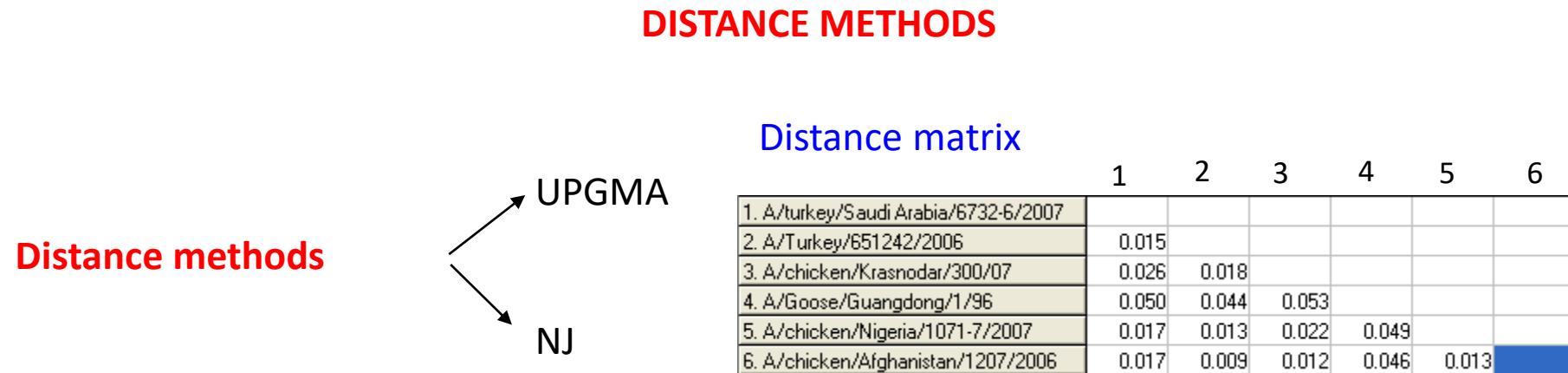
Sequence alignment

Identification of the proper model of nucleotide substitution

Application of a method to construct a phylogenetic tree



Step 4 – method selection



Distance methods are based on the idea that if we knew the actual evolutionary distance between all members of a set of sequences, than we could easily reconstruct the evolutionary history of those sequences

Advantages:

- allows the use of an explicit model of evolution
- very fast
- simple

Disadvantages:

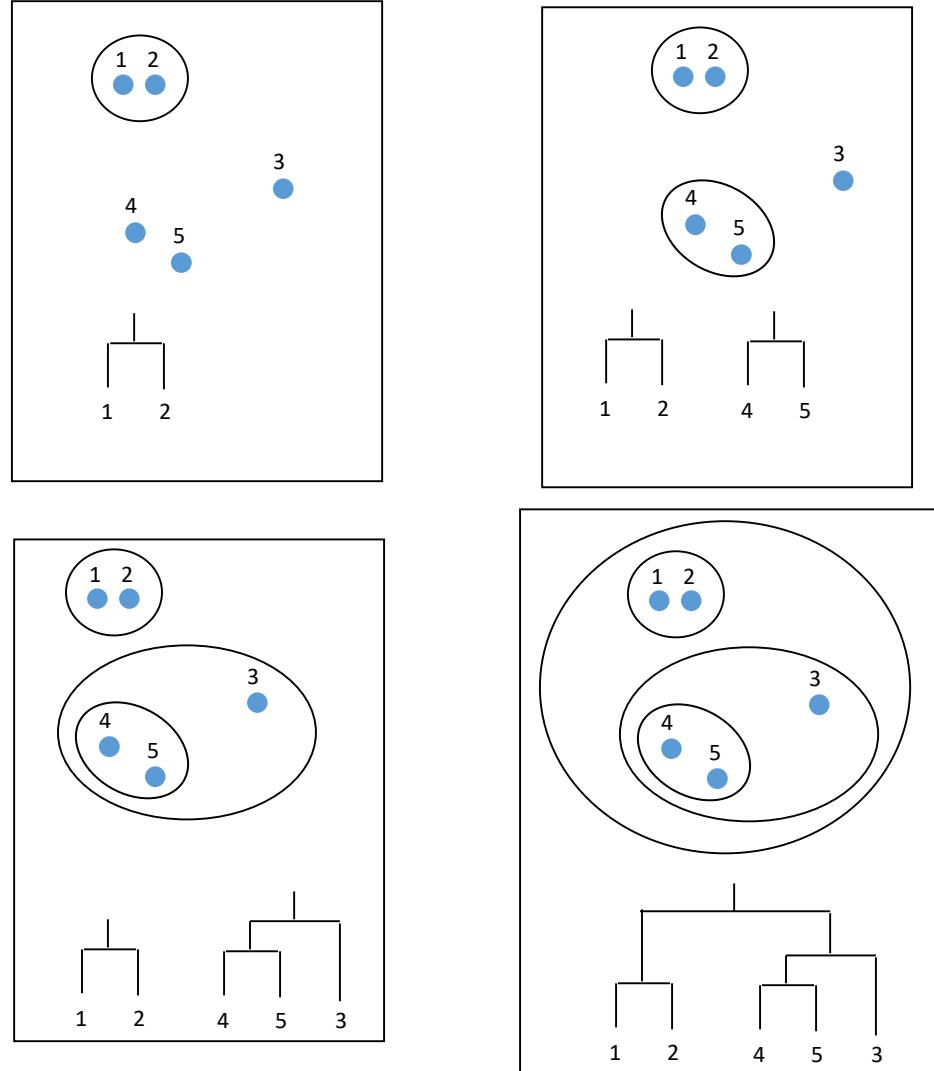
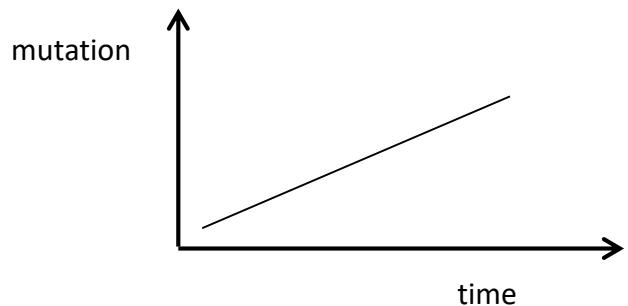
- only produces one tree with no indication of its quality
- reduces all sequence information into a single distance value
- dependent on the evolutionary model used (preferentially this model should be estimated from the data)

Step 4 – method selection

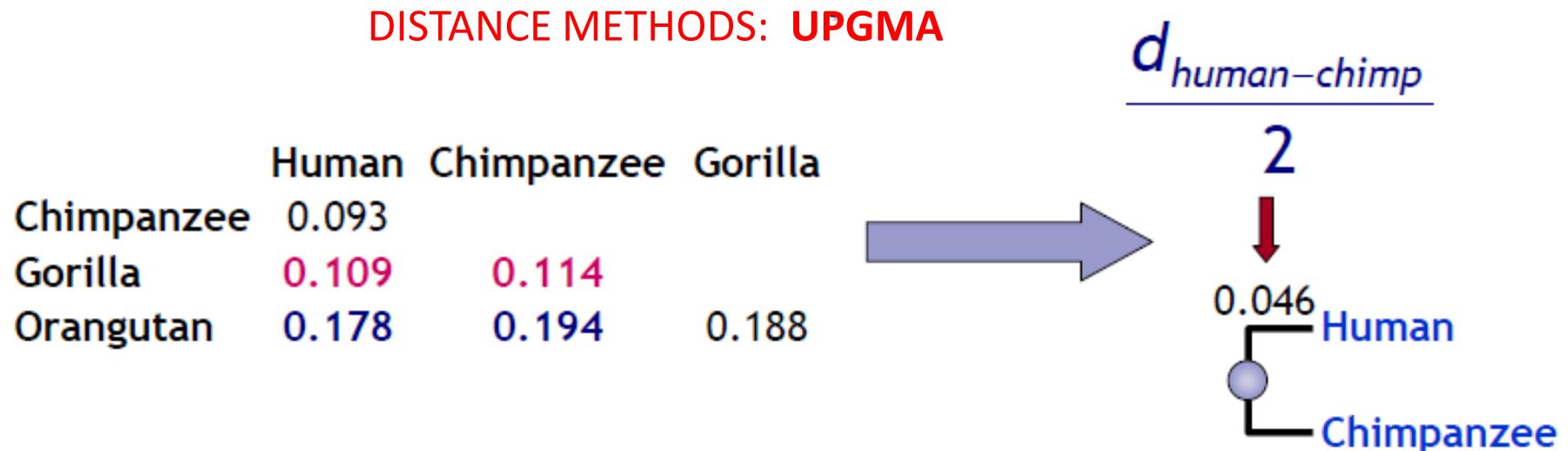
DISTANCE METHODS: UPGMA

UPGMA is a clustering algorithm that computes the distance between clusters using average pairwise distance

UPGMA assumes a constant molecular clock: all species represented by the leaves in the tree are assumed to accumulate mutations (and thus evolve) at the same rate. This is a major pitfalls of UPGMA.



Step 4 – method selection

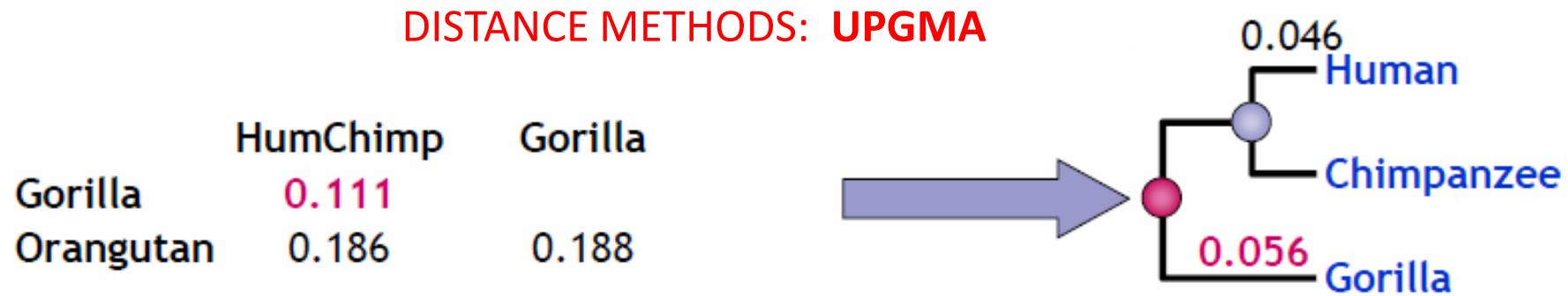


$$d_{\text{Gorilla}(\text{human+Chimp})} = \frac{0.109 + 0.114}{2} = 0.111$$

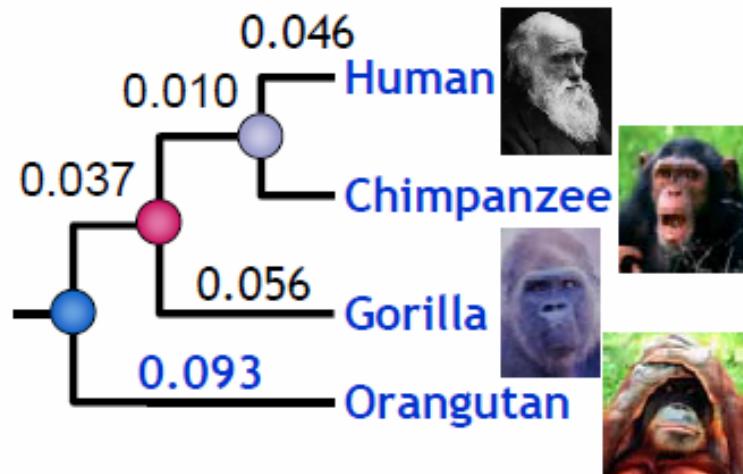
$$d_{(\text{human+chimp})\text{orangutan}} = \frac{0.178 + 0.194}{2} = 0.186$$



Step 4 – method selection



$$d_{(human+Chimp+Gorilla)Orango} = \frac{(0.178 + 0.194 + 0.188)}{3} = 0.187$$



Step 4 – method selection

DISTANCE METHODS: Neighbor Joining

The Neighbor Joining method is the most popular way to build trees from distance measurements

- Neighbor Joining corrects the UPGMA method for its (frequently invalid) assumption that the same rate of evolution applies to each branch of a tree.
- The distance matrix is adjusted for differences in the rate of evolution of each taxon (branch).
- Neighbor Joining has produced the best results in simulation studies and it is the most computationally efficient of the distance algorithms

Advantages

- is fast and thus suited for large datasets and for bootstrap analysis
- permit lineages with largely different branch lengths
- permits correction for multiple substitutions

Disadvantages

- reduces all sequence information into a single distance value
- gives only one possible tree
- strongly dependent on the model of evolution used.

	A	B	C	D
A	0	7	11	14
B	7	0	6	9
C	11	6	0	7
D	14	9	7	0

distance matrix

Step 1: We calculate the net divergence $r(i)$ for each OTU from all other OTUs

$$r(A) = 7+11+14=32$$

$$r(B) = 22$$

$$r(C) = 24$$

$$r(D) = 27$$

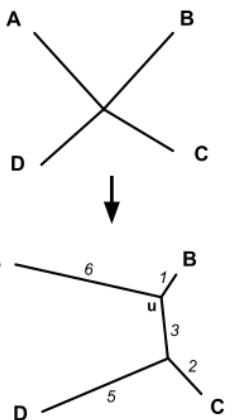
Step 2: Now we calculate a new distance matrix using for each pair of OUTs the formula:

$$M(ij)=(N-2) \times d(ij) - r(i) - r(j)$$

i.e. $M(A,B)= 2 \times 7 - 32 - 22 = -40$

Q matrix

	A	B	C	D
A	0	-40	-34	-34
B	-40	0	-34	-34
C	-34	-34	0	-40
D	-34	-34	-40	0



Step 4: Now we define new distances from U to each other terminal node:

$$d(CU) = [d(AC) + d(BC) - d(AB)] / 2 =$$

$$= (11 + 6 - 7)/2 = 5$$

With more taxa the entire procedure is repeated starting at step 1

Step 3: Now we choose as neighbors those two OTUs for which M_{ij} is the smallest (A,B and C,D). We calculate the distance of the pair members (ie. A,B) to the new node (u):

$$d(Au) = d(AB) / 2 + [r(A)-r(B)] / 2(N-2)$$

$$= 7/2 + (32-22)/4 = 6$$

$$d(Bu) = d(AB) - d(AU) = 1$$

Do the same for D,C.

Step 4 – method selection

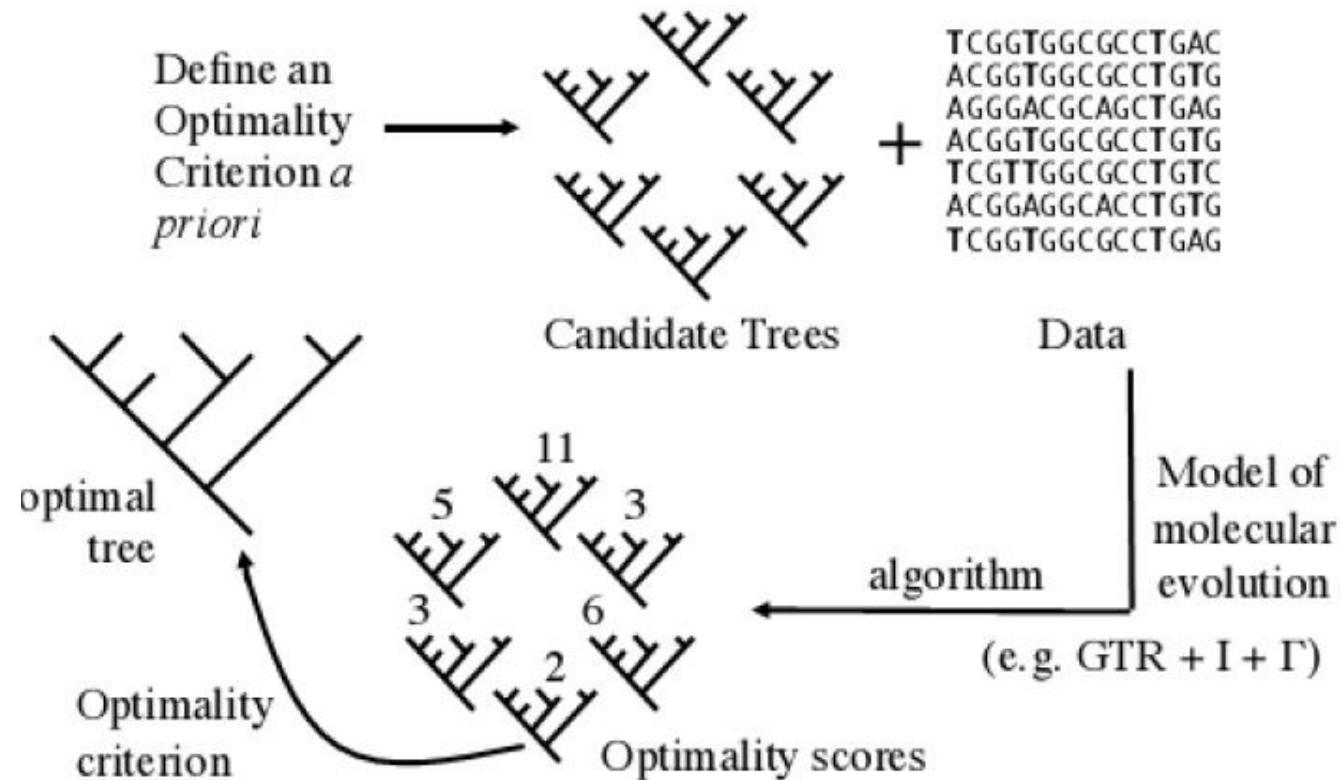
Statistical approach

MAXIMUM LIKELIHOOD

BAYESIAN METHODS

STATISTICAL METHODS: Maximum likelihood

- The most statistically valid approach to molecular phylogenetics along with the closely related Bayesian methods.
- Aims to find the tree topology that has the highest probability based on observed sequence characters. ML evaluates the possible tree topologies by calculating the probability of expecting each possible nucleotide position at every node for every sequence position, given the sequences at hand, and given an evolutionary model.
- The tree with the highest overall probability is chosen as the best one



Step 4 – method selection

STATISTICAL METHODS: Maximum likelihood

Advantages:

- allows the use of an explicit model of evolution
- evaluates different tree topologies
- allows unequal rates of evolution
- maintains all sequence information
- allows reconstruction of ancestral character states
- based on a statistical algorithm, thus allowing the statistical comparison of evolutionary models

Disadvantages:

- very computer intensive
- dependant on the evolutionary model used (preferentially this model should be estimated from the data)

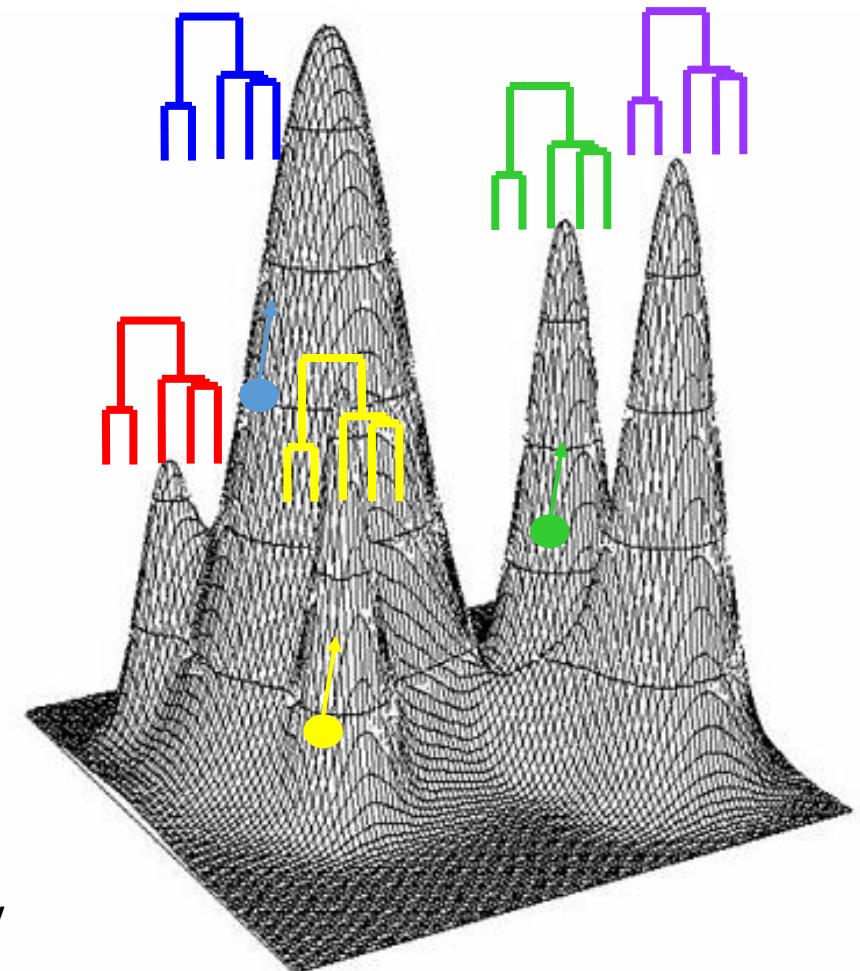
Step 4 – method selection

STATISTICAL METHODS: Bayesian method

Searching Through ‘Tree Space’

Bayesian Phylogenetics

- ❖ Using Bayesian statistics, you search for a set of *plausible trees* instead of a single best tree
- ❖ In this method, the “space” that you search in is limited by *prior* information
- ❖ generates a posterior distribution for a parameter, composed of a phylogenetic tree and a model of evolution, based on the prior for that parameter and the likelihood of the data, generated by a multiple alignment.



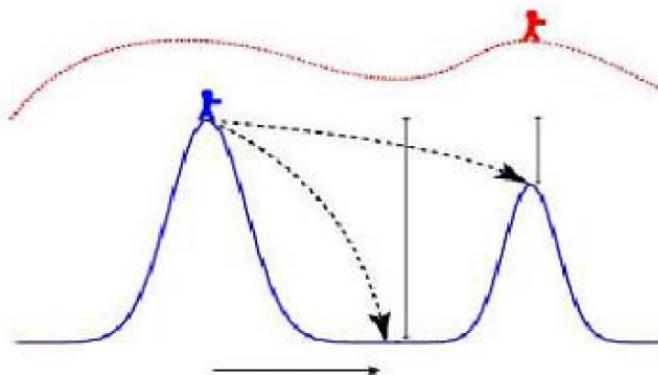
The highest hill represent the best tree topology

Step 4 – method selection

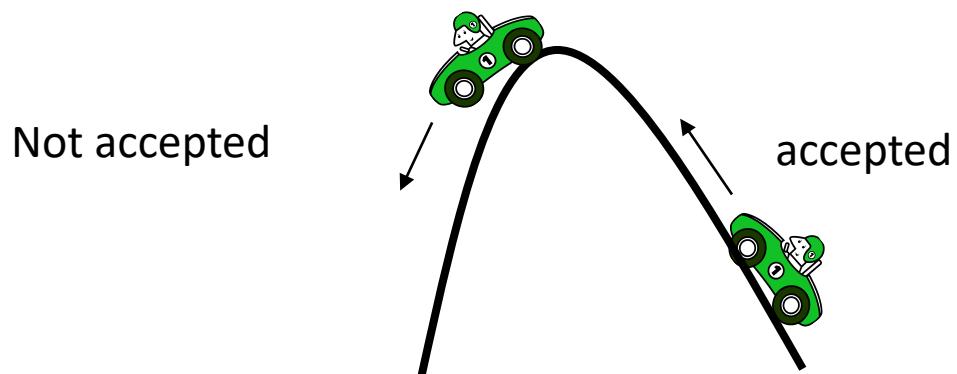
STATISTICAL METHODS: Bayesian method

There are two methods to explore the tree space:

- **HILL FINDING**: tries very different topology and allows to explore different parts of the tree space



- **HILL CLIMBING**: Allows to find the best tree in the hill



Bootstrap analyses

Original data set
with n
characters.

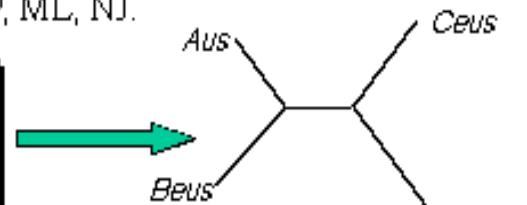
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Aus	C	G	A	C	G	T	G	G	T	C	T	A	T	A	C	A	C	G	A	
Ber	C	G	G	C	G	G	T	G	A	T	C	T	A	T	G	C	A	C	G	
Cer	T	G	G	C	G	G	C	G	T	C	T	C	A	T	A	C	A	A	T	
Deu	T	A	A	C	G	A	T	G	A	C	C	C	G	A	C	T	A	T	T	

Draw n characters
randomly with re-
placement.
Repeat m
times.

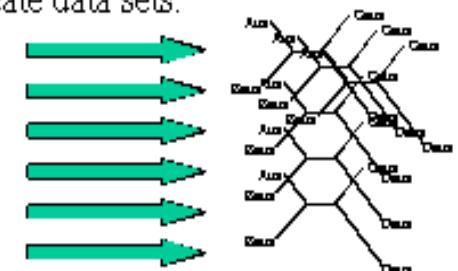
	1	3	13	8	3	19	14	6	20	20	7	1	9	11	17	10	6	14	8	16
Aus	G	A	A	G	A	G	T	G	A	A	T	C	G	C	A	T	G	T	G	C
Ber	G	G	A	G	G	G	T	G	G	G	T	C	A	C	A	T	G	T	G	C
Cer	G	G	A	G	G	T	T	G	A	A	C	T	T	T	A	C	G	T	G	C
Deu	A	A	G	G	A	T	A	A	G	G	T	T	A	C	A	C	A	A	G	T

m pseudo-replicates,
each with n characters.

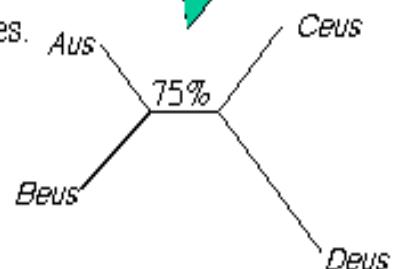
Original
analysis, e.g.
MP, ML, NJ.

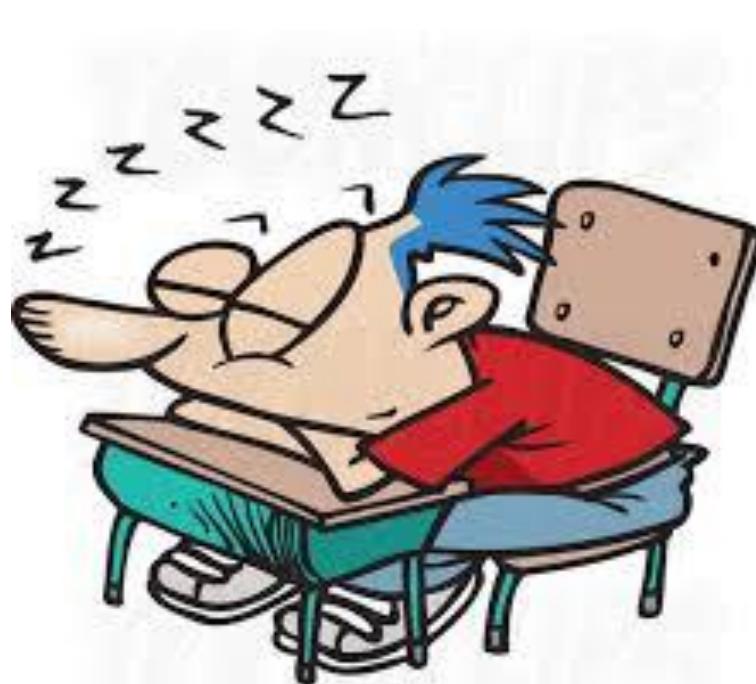


Repeat original analysis
on each of the pseudo-
replicate data sets.



Evaluate the
results from the
 m analyses.





Thank you for your attention!