Phylogenetic analysis of RNA viruses

Alice Fusaro

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

ESFLU WG2 Training School, Brescia, 16 October 2024

Introduction

Short generation time

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

Large populations

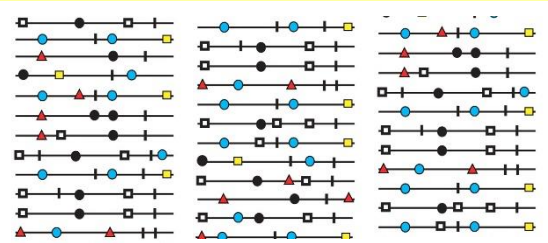
Several orders of magnitude larger than any population size for DNA organisms

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

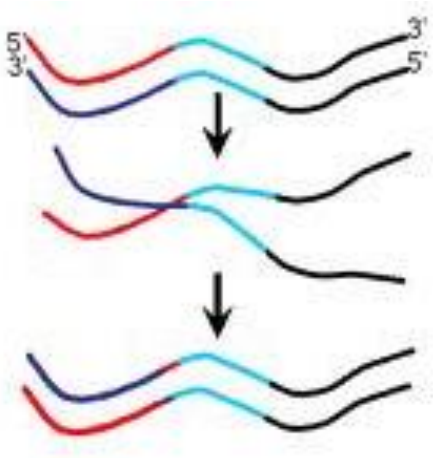
RECOMBINATION

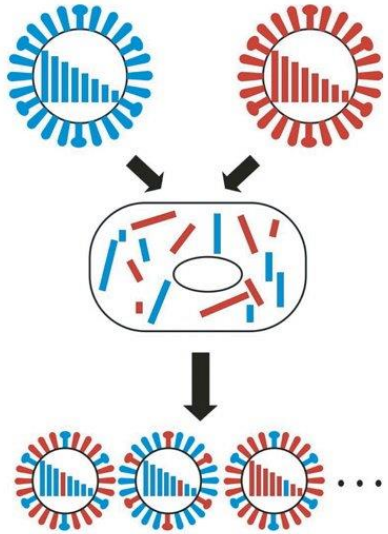
REASSORTMENT

ACTGTCA

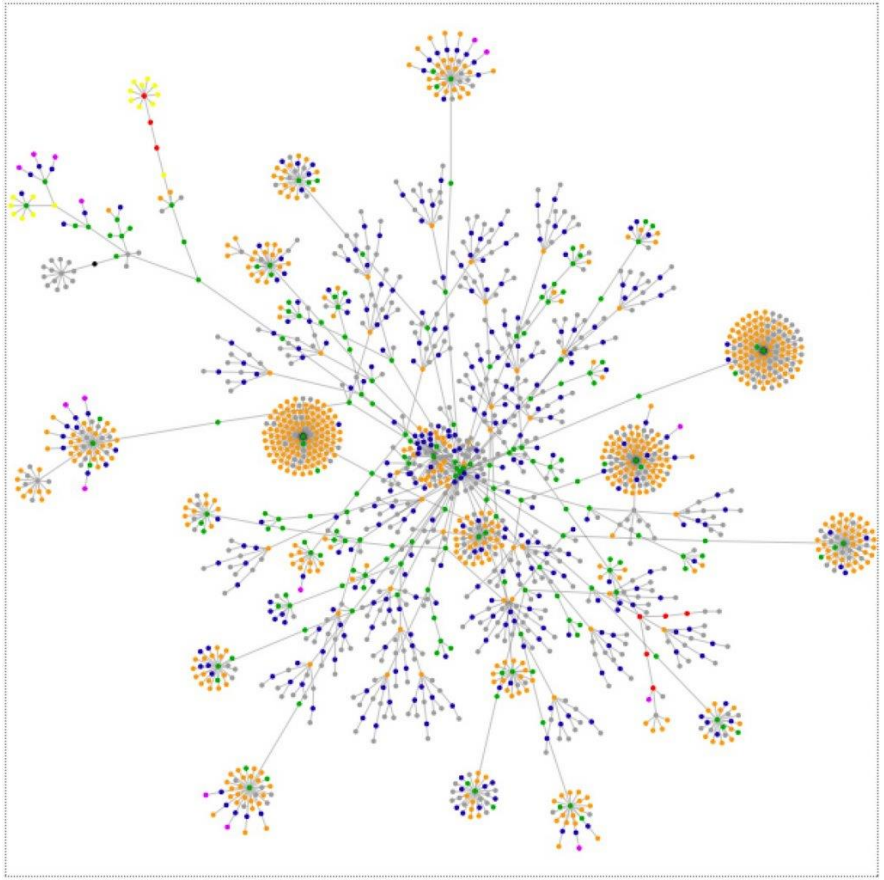
↓

ACAGTCA





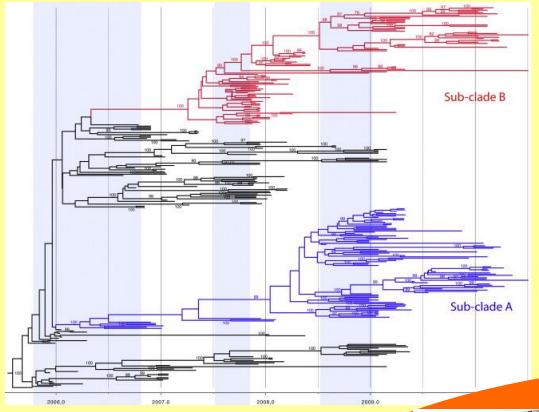
Genetic changes in 'real time’



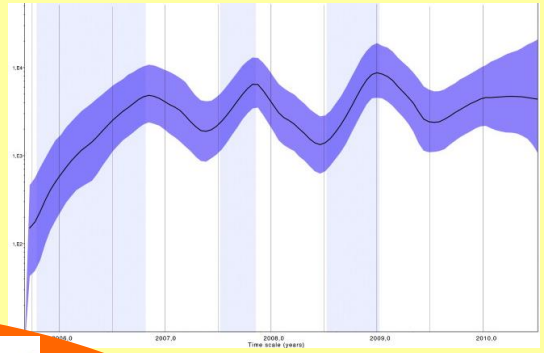
An understanding of the mechanisms of RNA virus sequence change is crucial to predict important aspects of their emergence and long-term evolution.

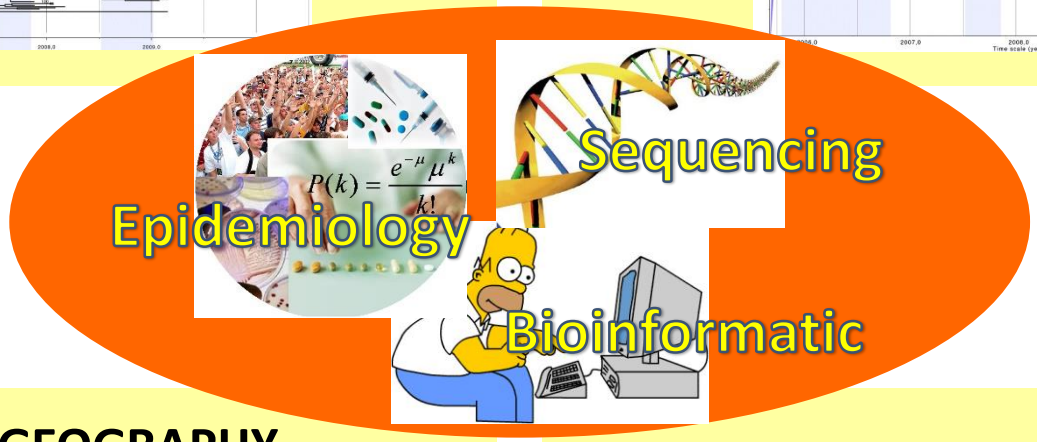
Moya et al., Nature 2004

PHYLOGENY

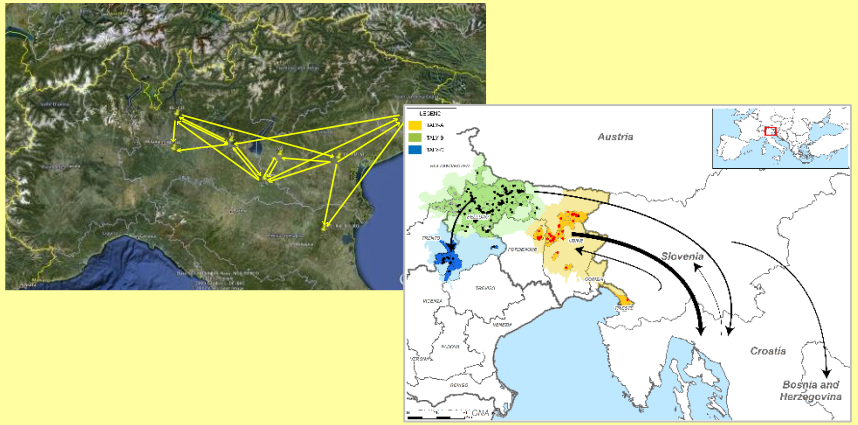


EVOLUTIONARY DYNAMICS OF VIRAL POPULATIONS

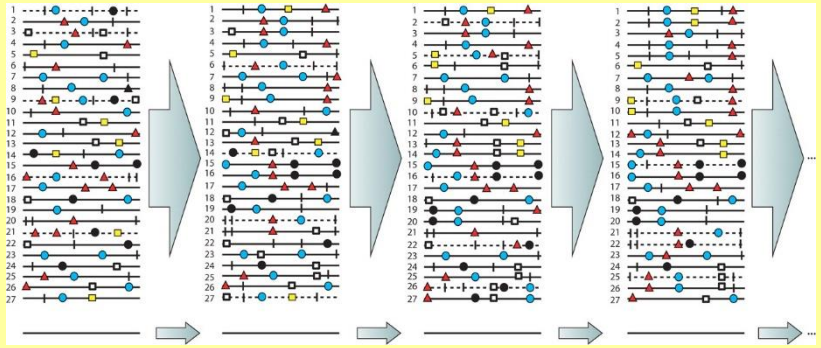




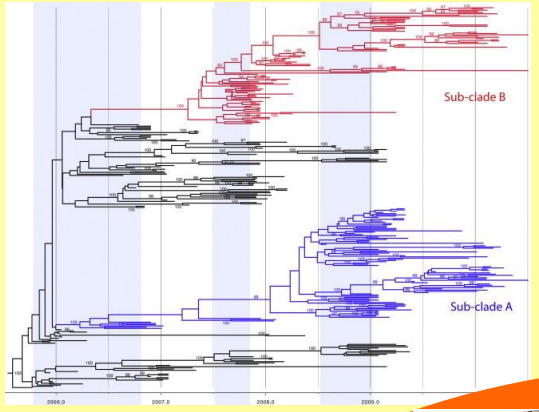
PHYLOGEOGRAPHY



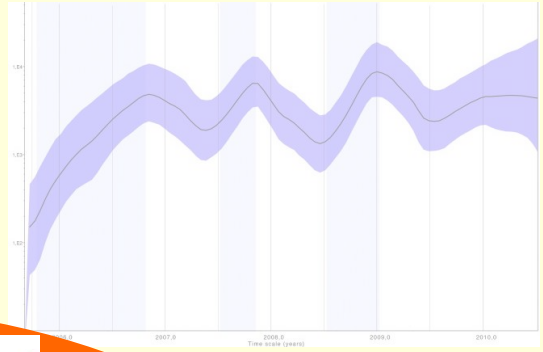
INTRA-HOST VARIABILITY NGS

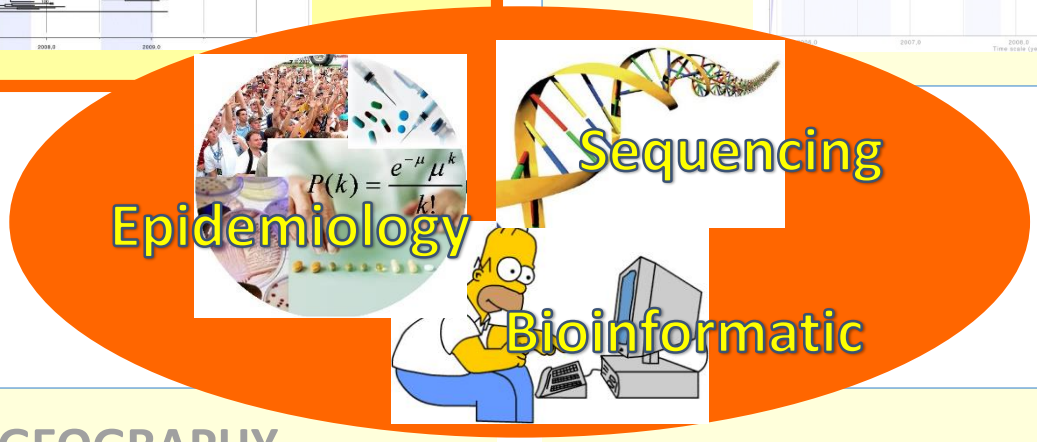


PHYLOGENY

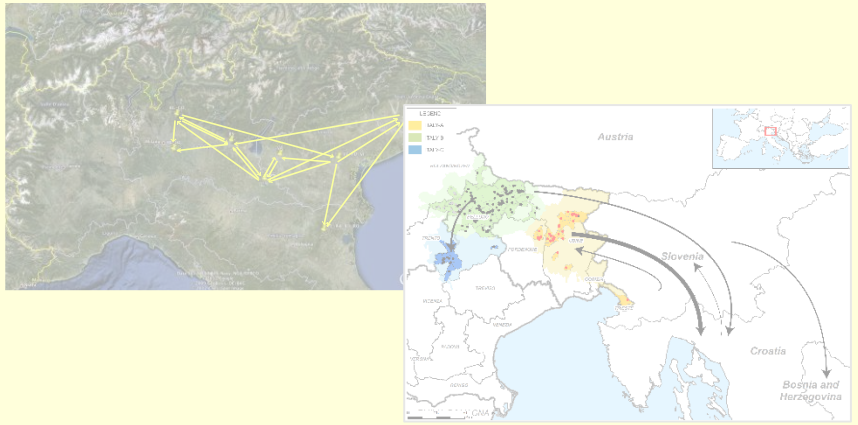


EVOLUTIONARY DYNAMICS OF VIRAL POPULATIONS

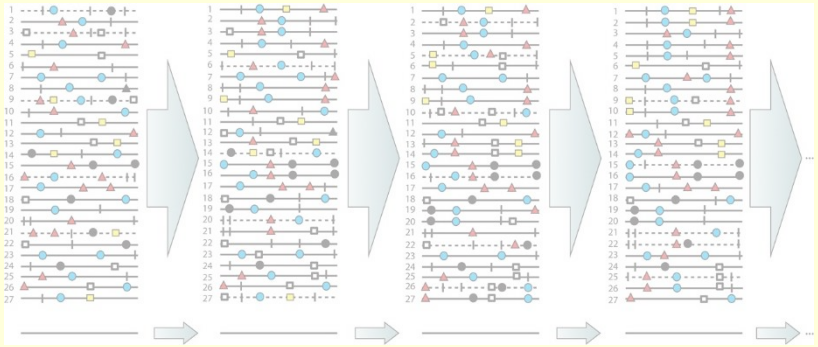




PHYLOGEOGRAPHY



INTRA-HOST VARIABILITY NGS



Which information can you gather from a phylogenetic analysis?

GENETIC RELATIONSHIP

ORIGIN

NEW INTRODUCTIONS

phylogeny

FOLLOW VIRAL EVOLUTION

VIRAL GENOTYPING

REASSORTMENTS OR RECOMBINATION

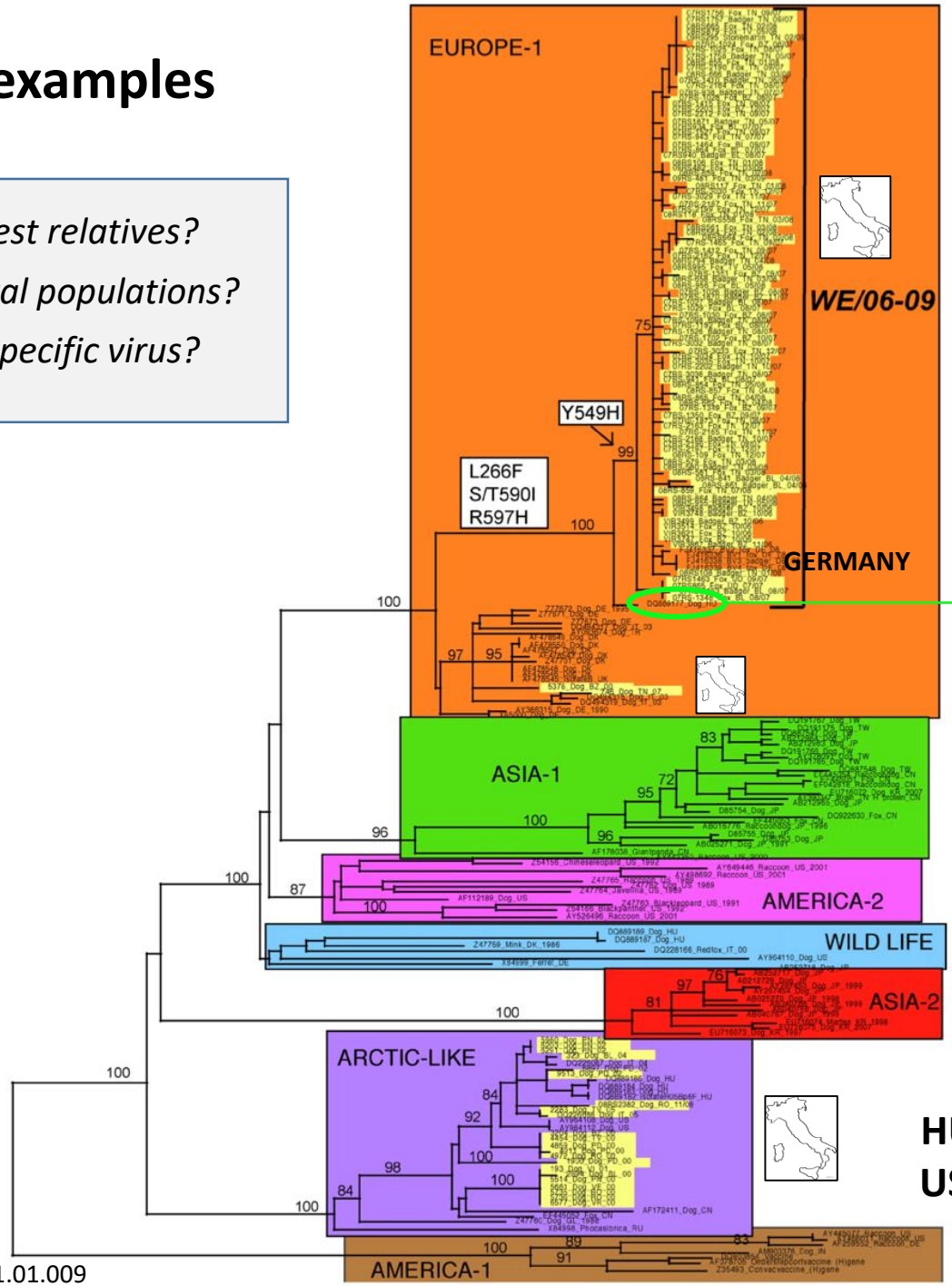
CO-CIRCULATION OF MULTIPLE GENETIC GROUPS

Phylogeny- examples

Which viruses are the closest relatives?

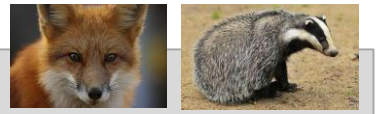
Are circulating multiple viral populations?

What was the origin of a specific virus?



ITALY
HUNGARY
GERMANY
DANIMARK

HUNGARY



H gene
Canine distemper virus
Family Paramyxoviridae

ITALY
HUNGARY
US, CHINA

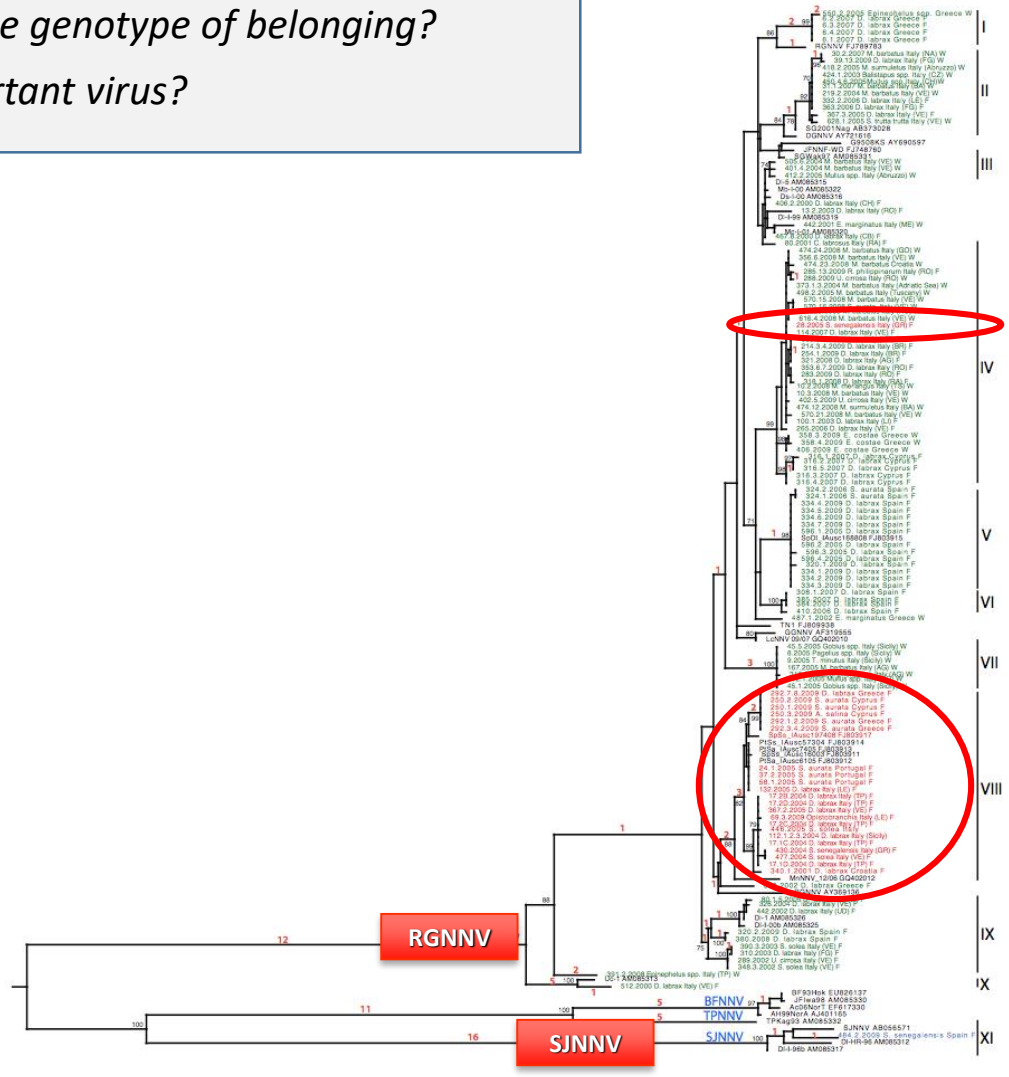
Monne et al., 2011: 10.1016/j.vetmic.2011.01.009

Phylogeny- examples

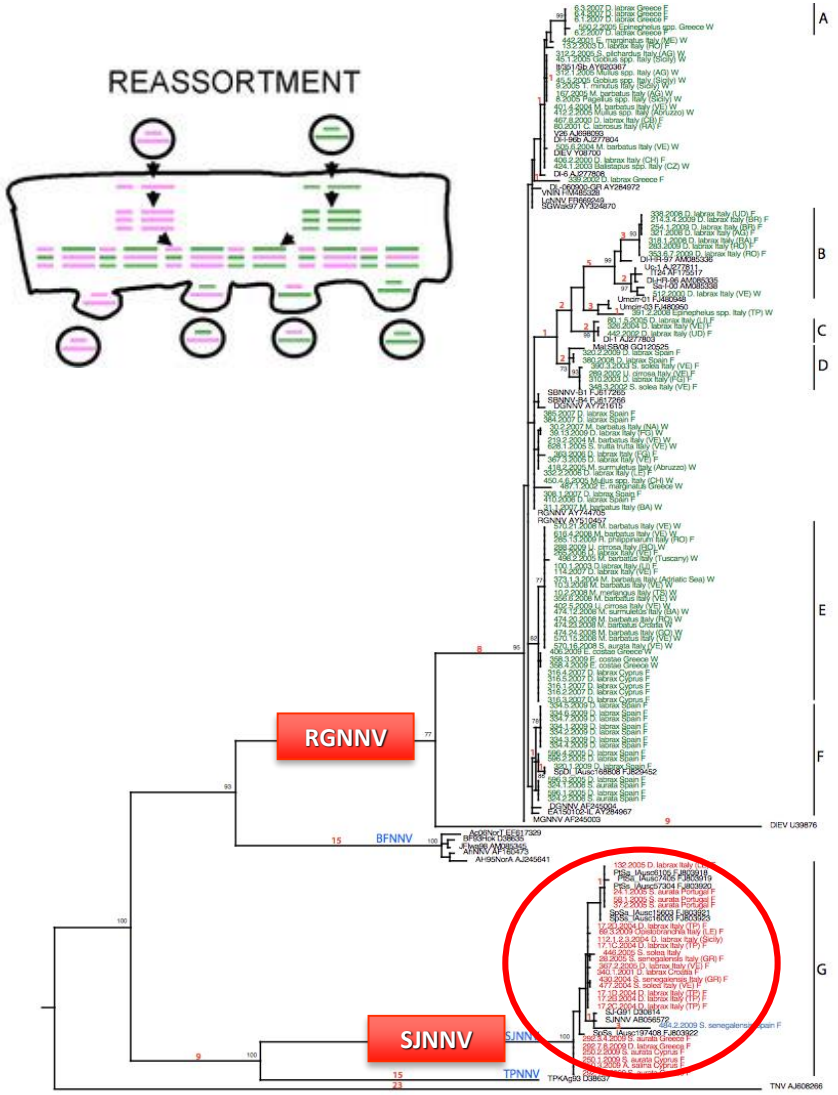
Which is the genotype of belonging?

Is a reassortant virus?

RNA 1



RNA 2



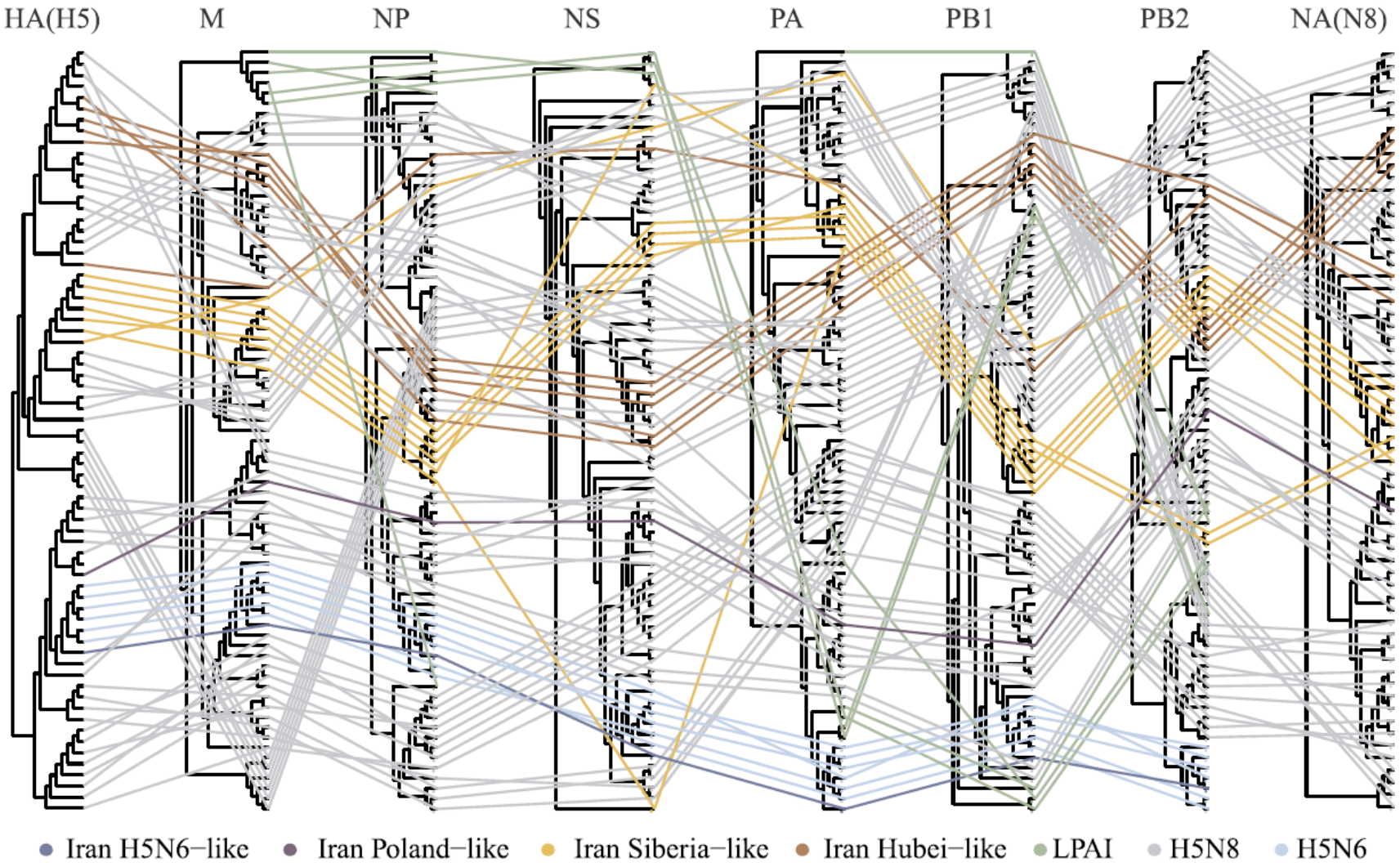
Panzarin et al., 2012: 10.1016/j.meegid.2011.10.007

Phylogeny- examples

H. Abdollahi, et al.

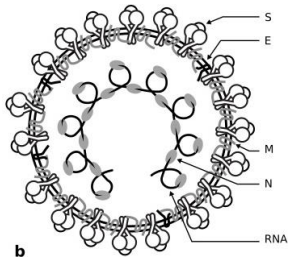
Infection, Genetics and Evolution 83 (2020) 104342

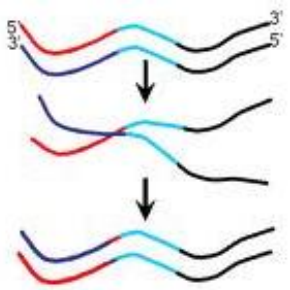
a)



Phylogeny- examples

Is a recombinant virus?

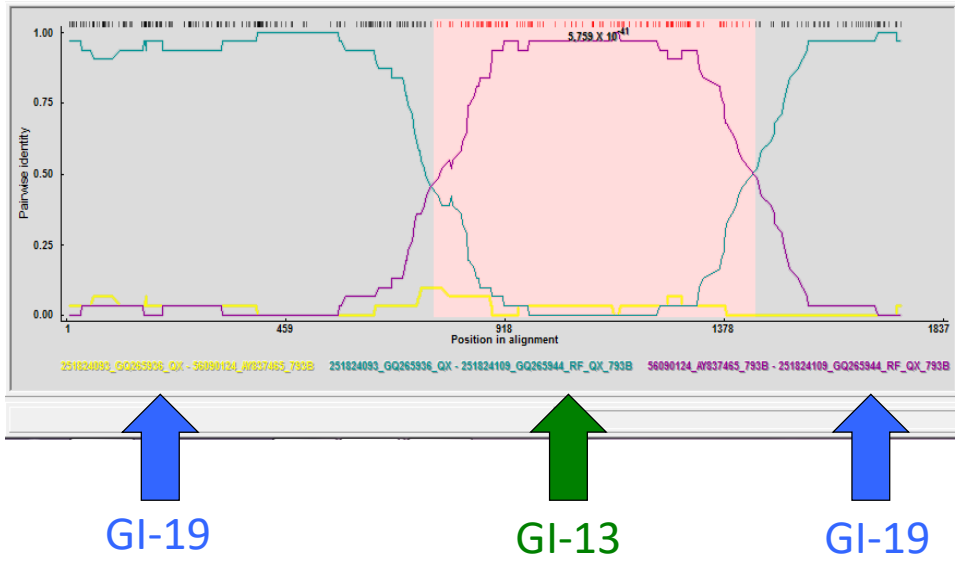




IBV

Family: coronavirus

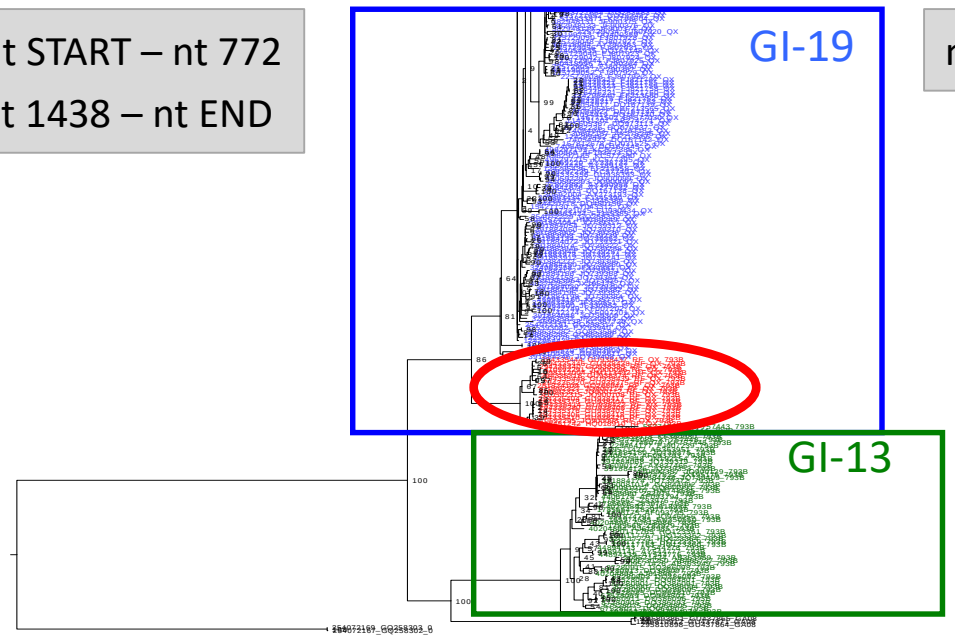
Non-segmented, positive-sense single-stranded RNA genome



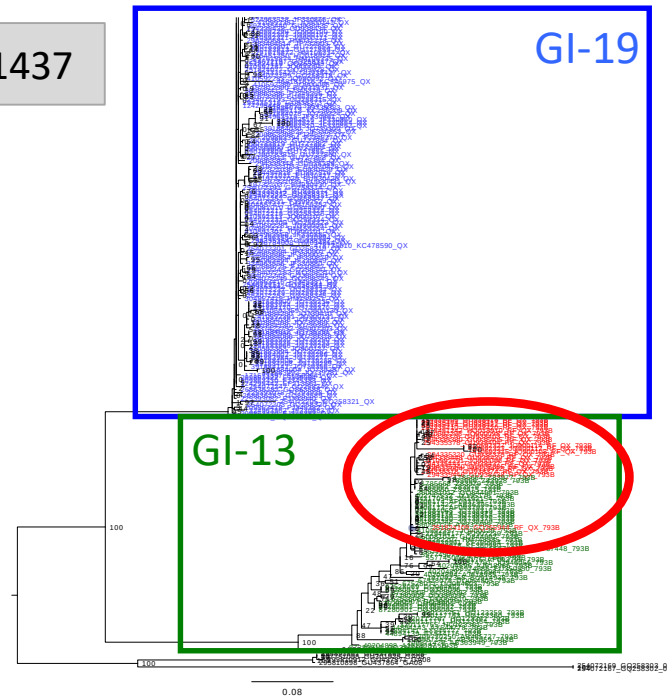
Simplot

nt START – nt 772

nt 1438 – nt END



nt 773 – nt 1437



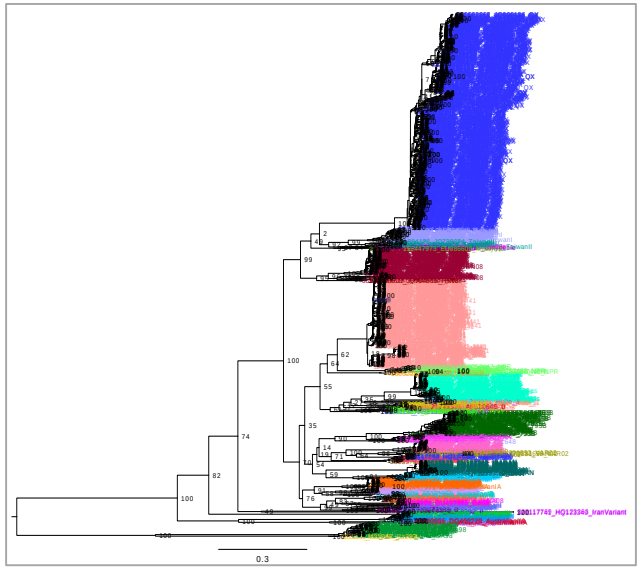
Valastro et al., 2016: http://dx.doi.org/10.1016/j.meegid.2016.02.015

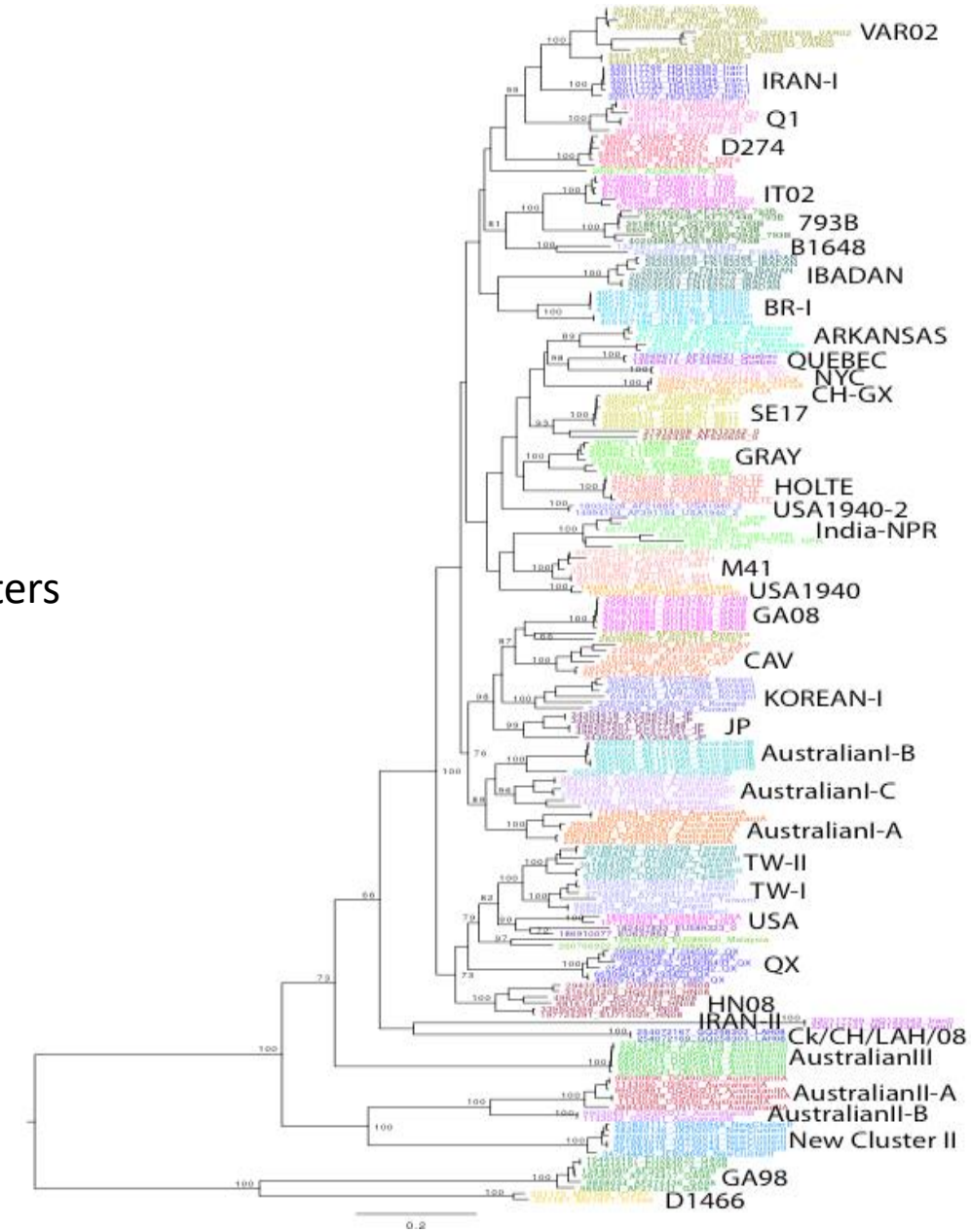
Phylogeny- examples

Classification system based on phylogenetic analysis

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

Identification of 40 different genetic clusters





Valastro et al., 2016: http://dx.doi.org/10.1016/j.meegid.2016.02.015

Useful Textbooks & Software

Books:

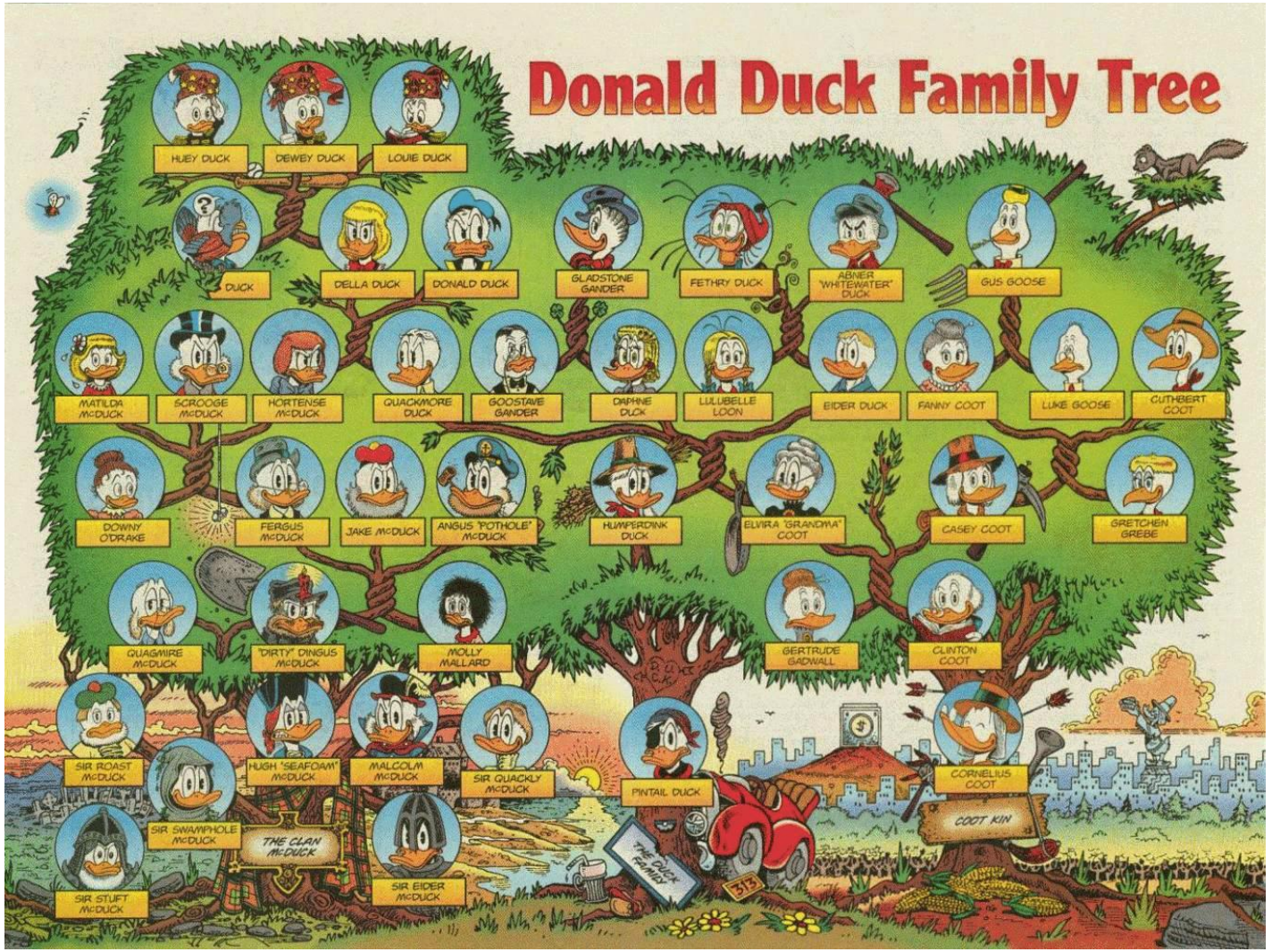
- Nei M & Kumar S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.
- Page RDM & Holmes EC. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science Ltd, Oxford.
- Lemey P, Salemi M & Vandamme A-M. (2009). *The Phylogenetic Handbook, 2nd Edition*. Cambridge University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

Computer Software

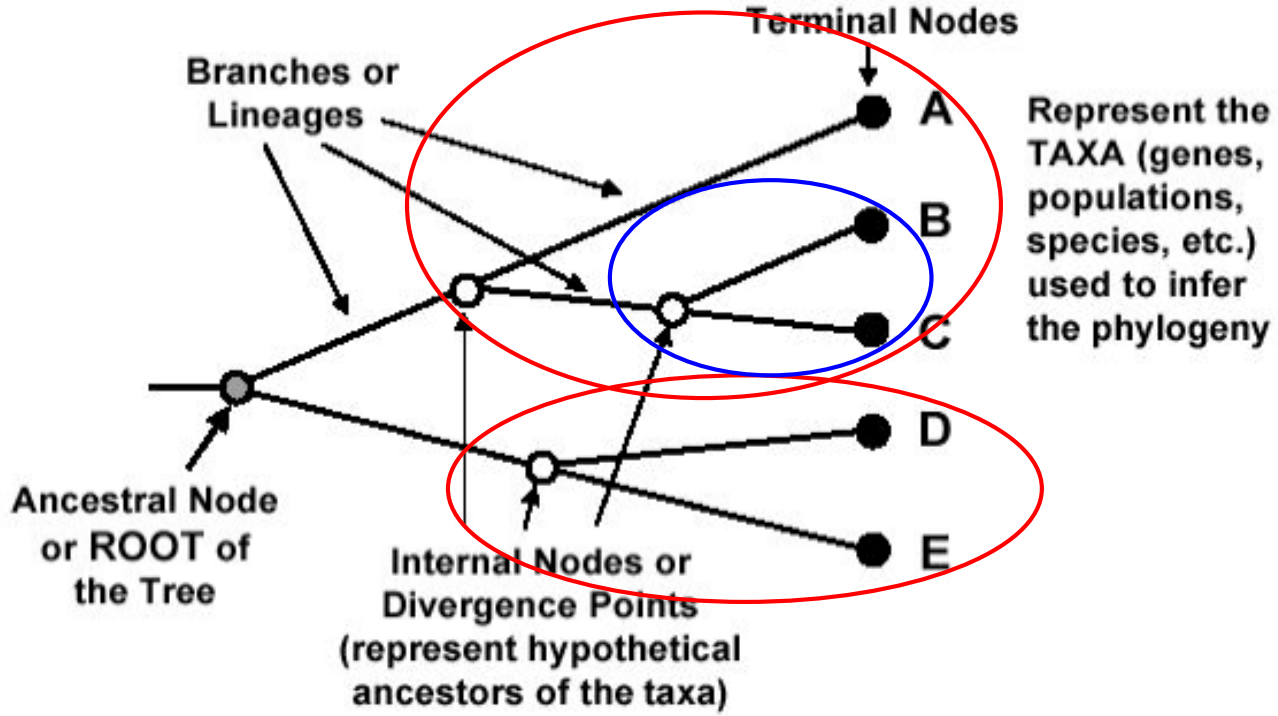
- **MEGA** (Molecular Evolutionary Genetics Analysis) http://megasoftware.net/
- **BEAST** (Bayesian Evolutionary Analysis Sampling Trees) http://beast.bio.ed.ac.uk/
- **MrBayes** (Bayesian inference of phylogeny) http://mrbayes.csit.fsu.edu/
- **PhyML** http://www.atgc-montpellier.fr/phyml/binaries.php
- **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



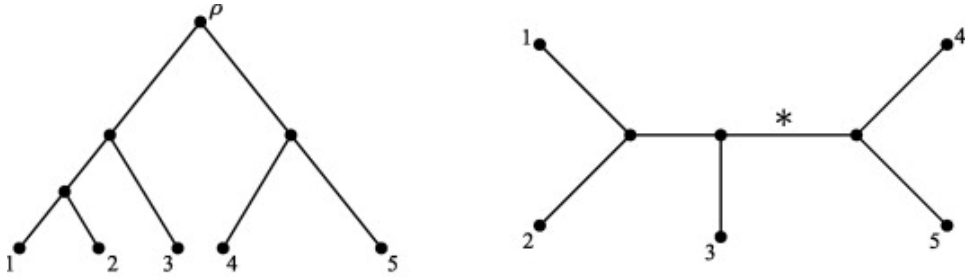
A tree consists of **nodes** connected by **branches**.

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

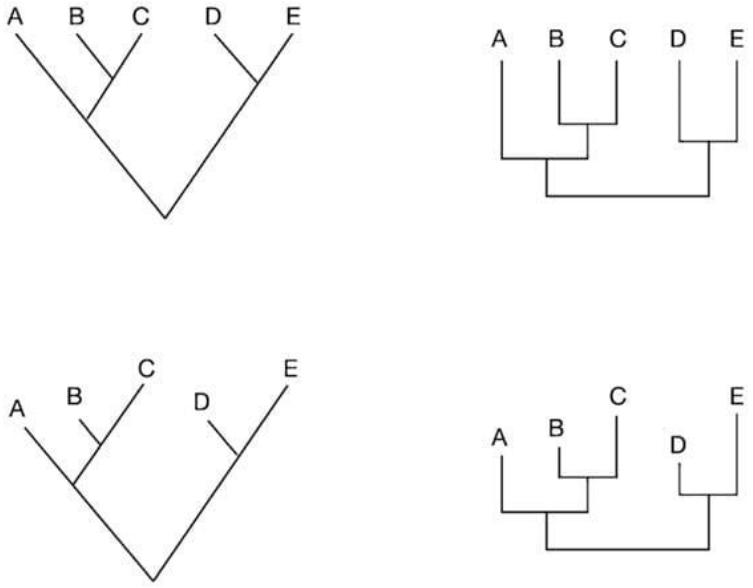
Rooted or unrooted?



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?

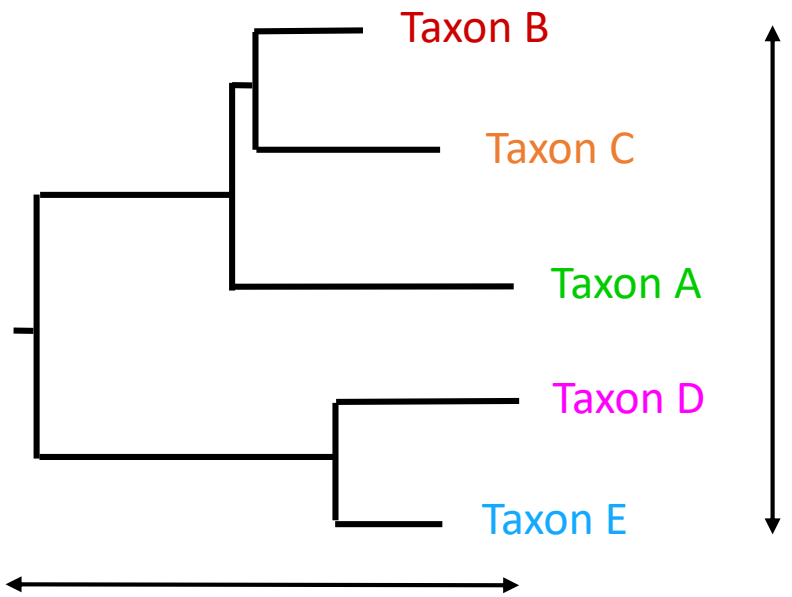


Cladogram is a type of phylogenetic tree that displays only the branching pattern of evolutionary relationships among taxa (unscaled tree: the branch lengths do not reflect the amount of evolutionary divergence between taxa).

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

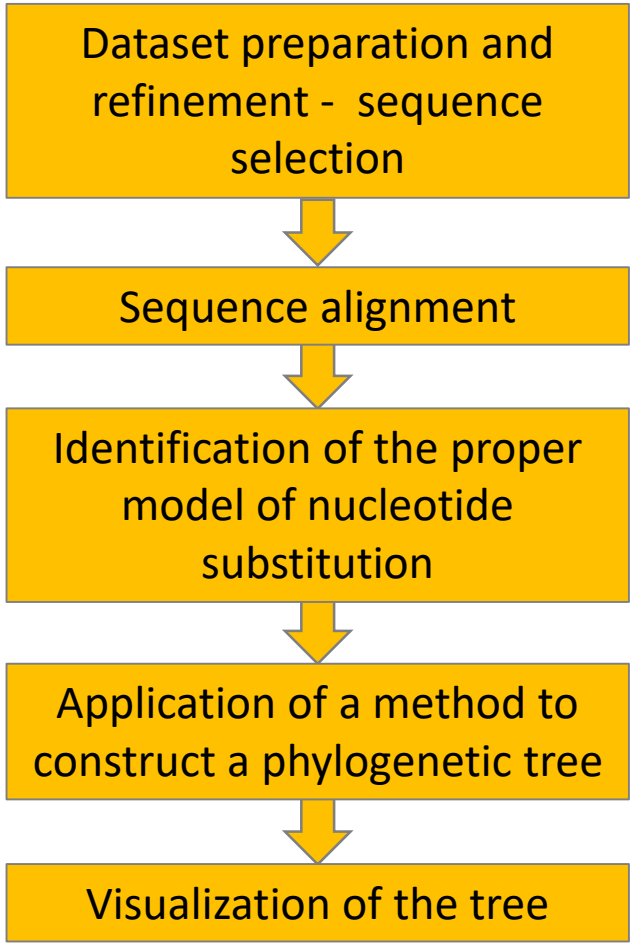


The order in which they appear from top to bottom has no meaning

This dimension either can have no scale (for 'cladograms'), can be proportional to **genetic distance or amount of change** (for 'phylograms'), 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)

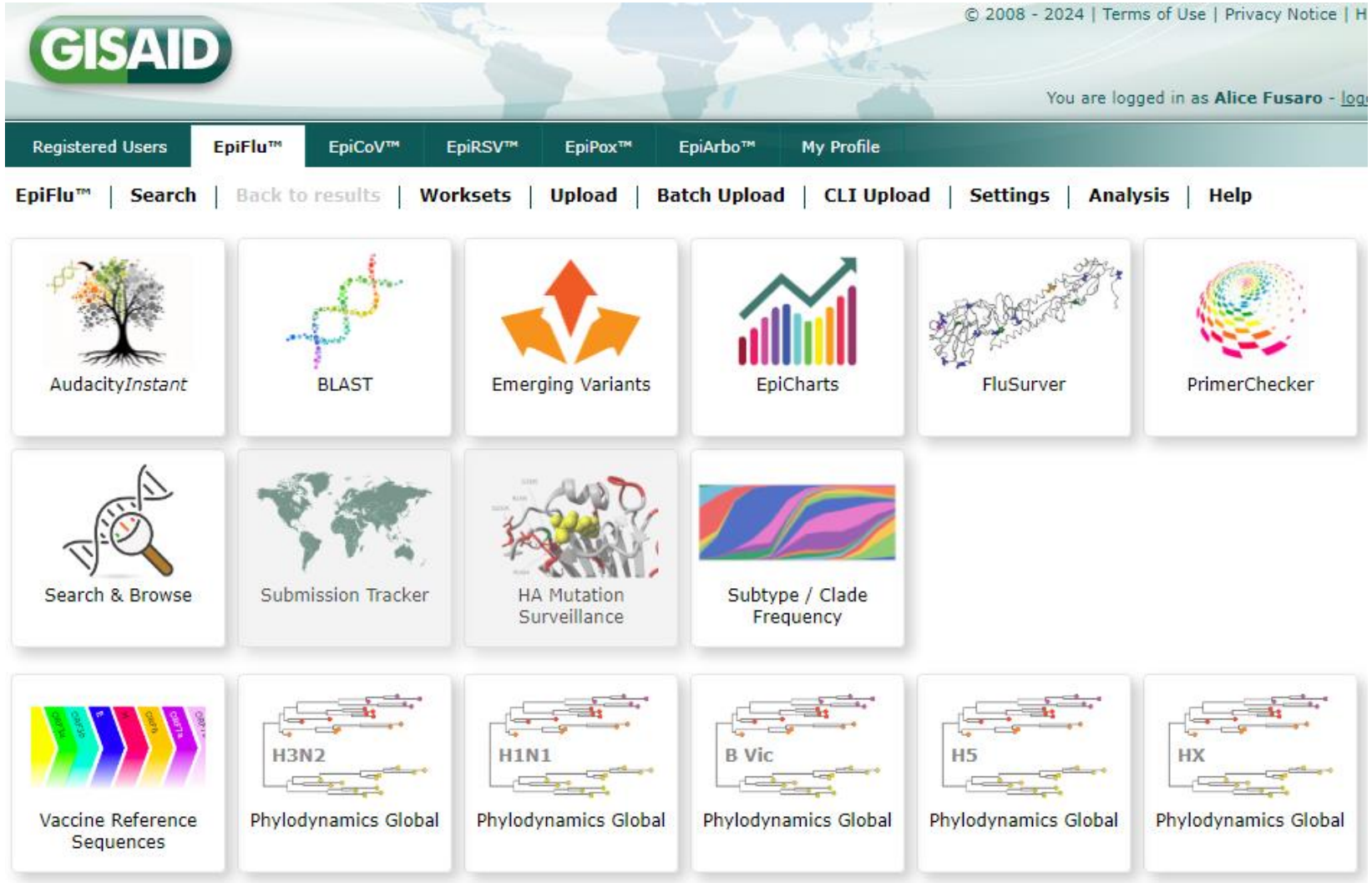
http://www.ncbi.nlm.nih.gov/

GISAID

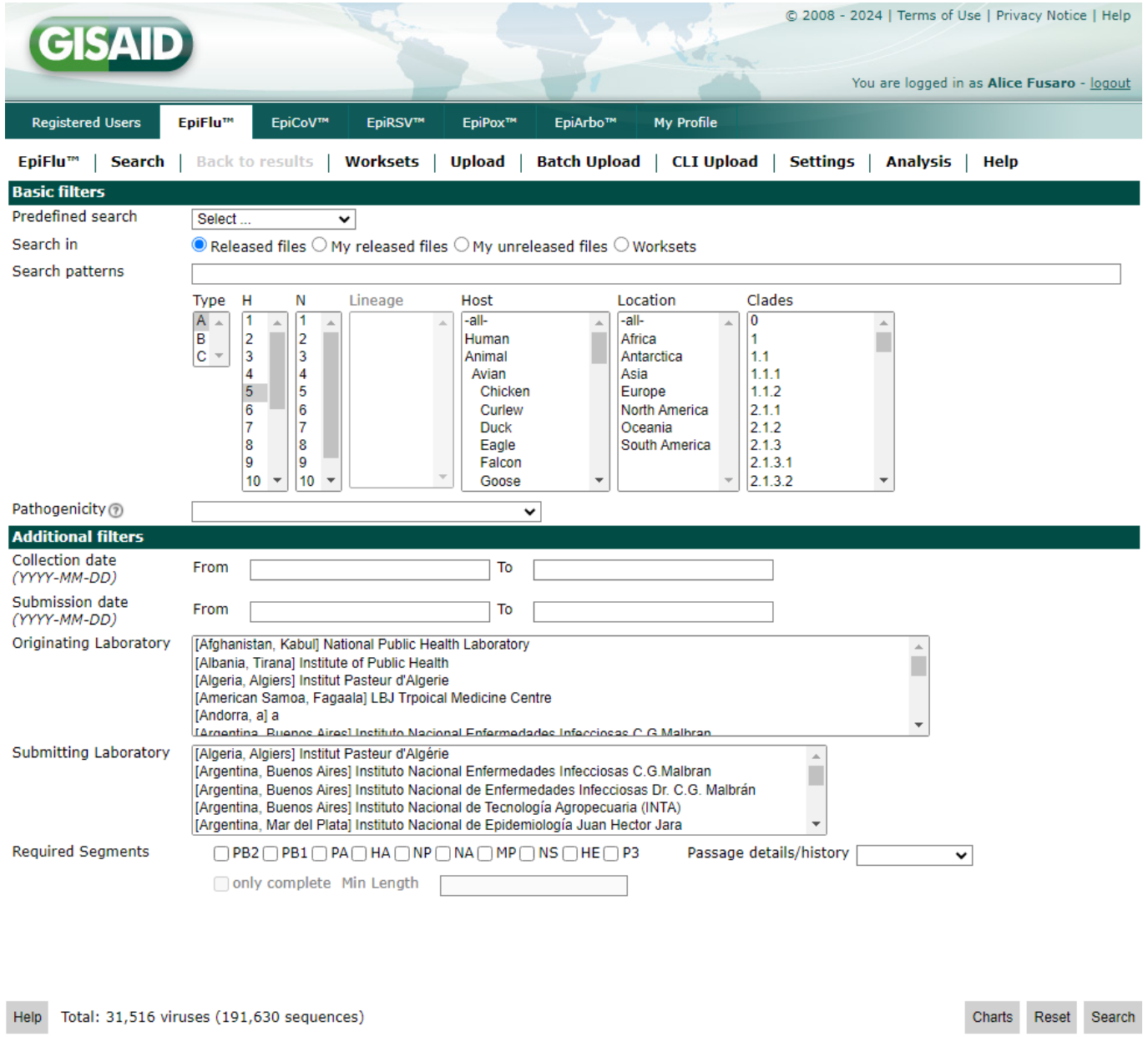
https://www.epicov.org/

BLAST TOOL: to find related sequences

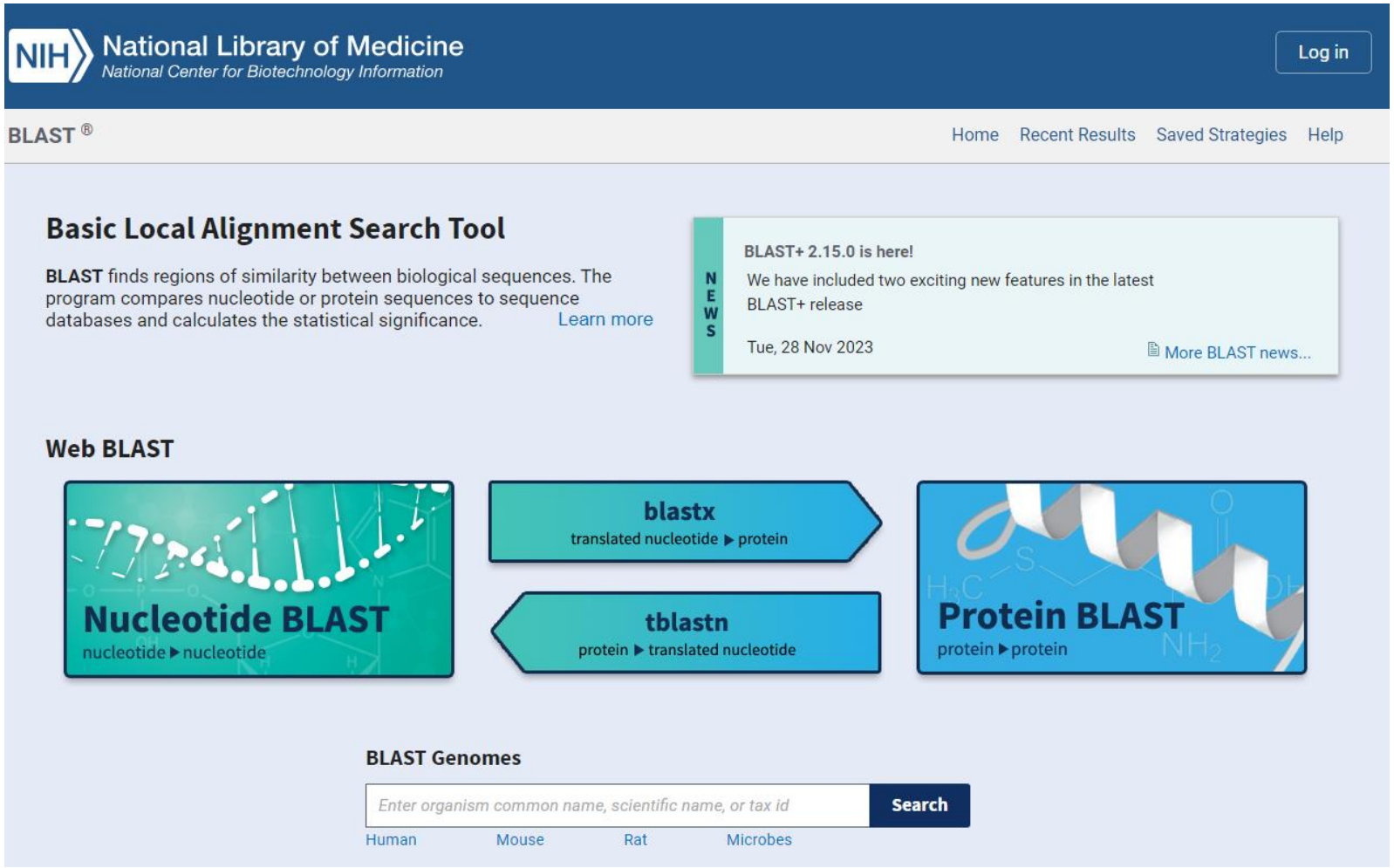
Step 1 – sequence selection



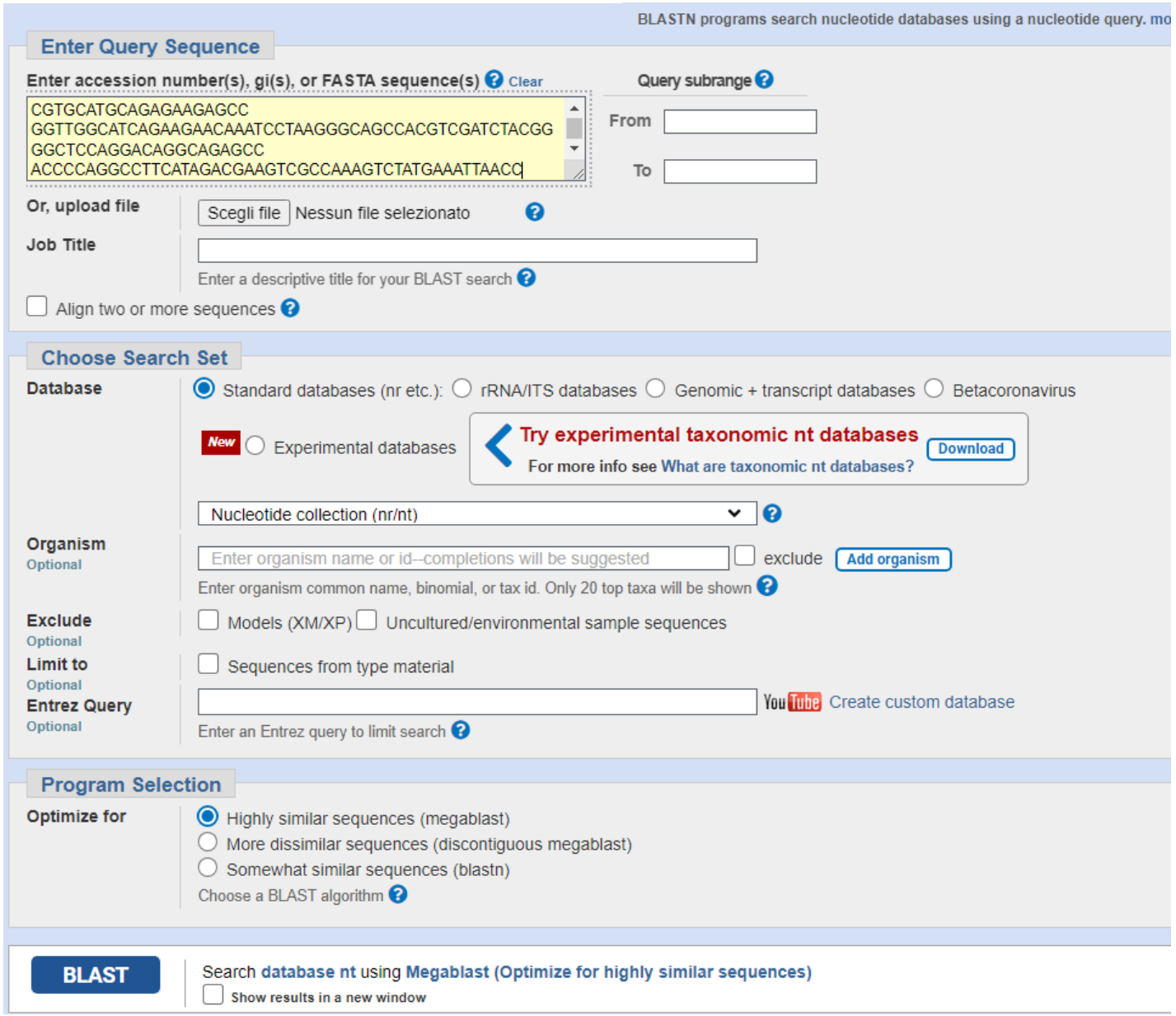
Step 1 – sequence selection



Step 1 – sequence selection



Step 1 – sequence selection



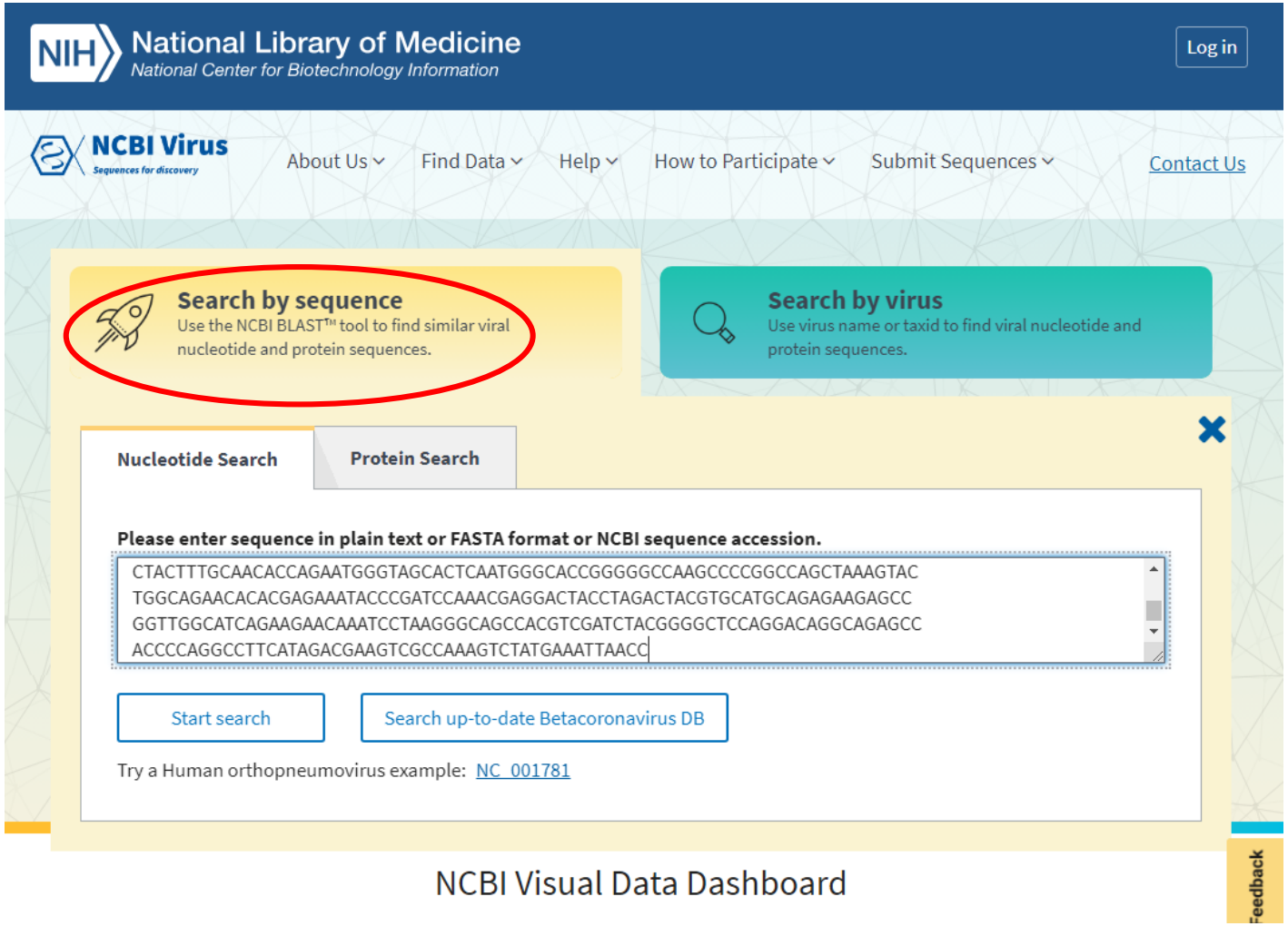
Step 1 – sequence selection

BLAST results

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Infectious bursal disease virus strain ks segment A polyprotein mRNA, complete cds	Infectious bursal ...	4518	4518	100%	0.0	96.79%	3163	DQ927042.1
Infectious bursal disease virus isolate Ventri-IBDV-Plus polyprotein mRNA, complete cds	Infectious bursal ...	4495	4495	100%	0.0	96.64%	3040	KJ547670.1
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
Infectious bursal disease virus strain mb segment A polyprotein mRNA, complete cds	Infectious bursal ...	4495	4495	100%	0.0	96.64%	3163	DQ927040.1
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
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
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
Infectious bursal disease virus isolate 89163 polyprotein gene, partial cds	Infectious bursal ...	4457	4457	100%	0.0	96.38%	3130	ON100652.1
Infectious bursal disease virus segment A, complete genome, strain 89163	Infectious bursal ...	4457	4457	100%	0.0	96.38%	3198	HG974563.1
Infectious bursal disease virus, segment A, isolate SH99	Infectious bursal ...	4451	4451	100%	0.0	96.34%	3261	LM651365.1
Infectious bursal disease virus segment A, complete sequence	Infectious bursal ...	4451	4451	100%	0.0	96.34%	3260	AY598356.1
Infectious bursal disease virus VP5 protein and structural polyprotein genes, complete cds	Infectious bursal ...	4451	4451	100%	0.0	96.34%	3260	AY444873.3
Infectious bursal disease virus strain 3529/92 polyprotein mRNA, complete cds	Infectious bursal ...	4440	4440	100%	0.0	96.27%	3039	KC189836.1
Infectious bursal disease virus segment A VP5 and polyprotein genes, complete cds	Infectious bursal ...	4440	4440	100%	0.0	96.27%	3260	EU595667.1
Infectious bursal disease virus isolate 91168 polyprotein gene, partial cds	Infectious bursal ...	4434	4434	100%	0.0	96.23%	3130	ON100653.1
Infectious bursal disease virus isolate Harbin-1 segment A, complete sequence	Infectious bursal ...	4434	4434	100%	0.0	96.23%	3261	EF517528.1
Infectious bursal disease virus strain SH/92 polyprotein mRNA, complete cds	Infectious bursal ...	4433	4433	100%	0.0	96.20%	3039	AF533670.1

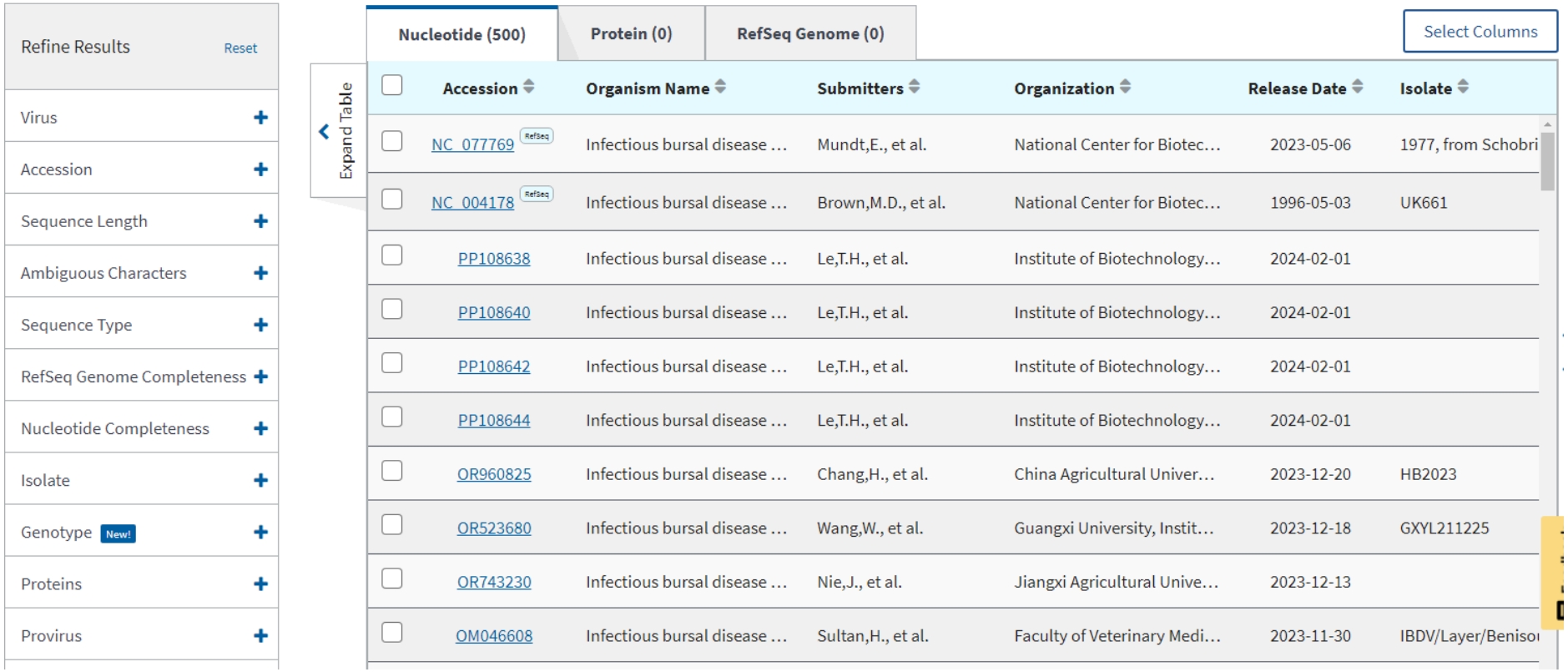
Step 1 – sequence selection

NCBI virus database



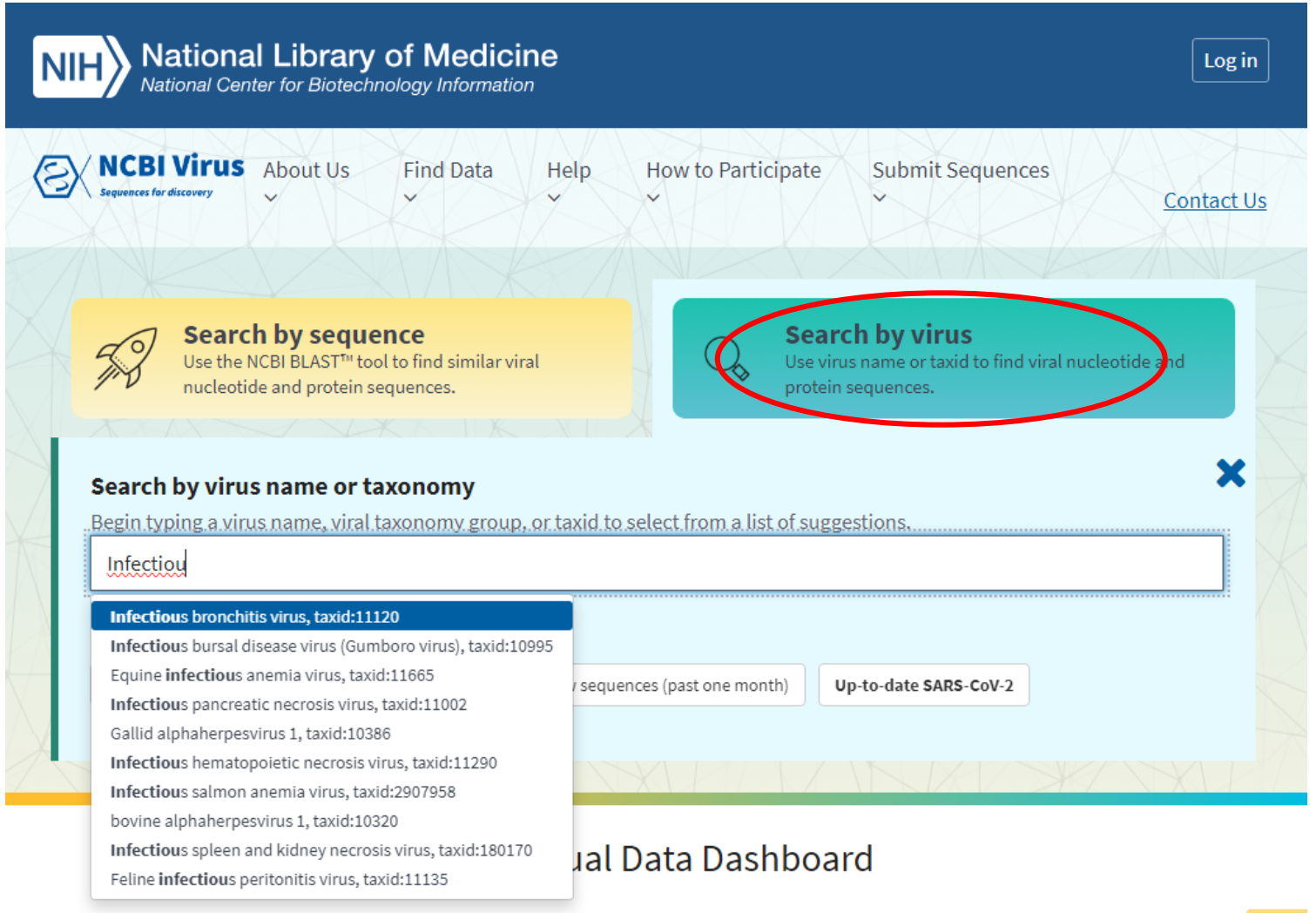
Step 1 – sequence selection

NCBI virus database



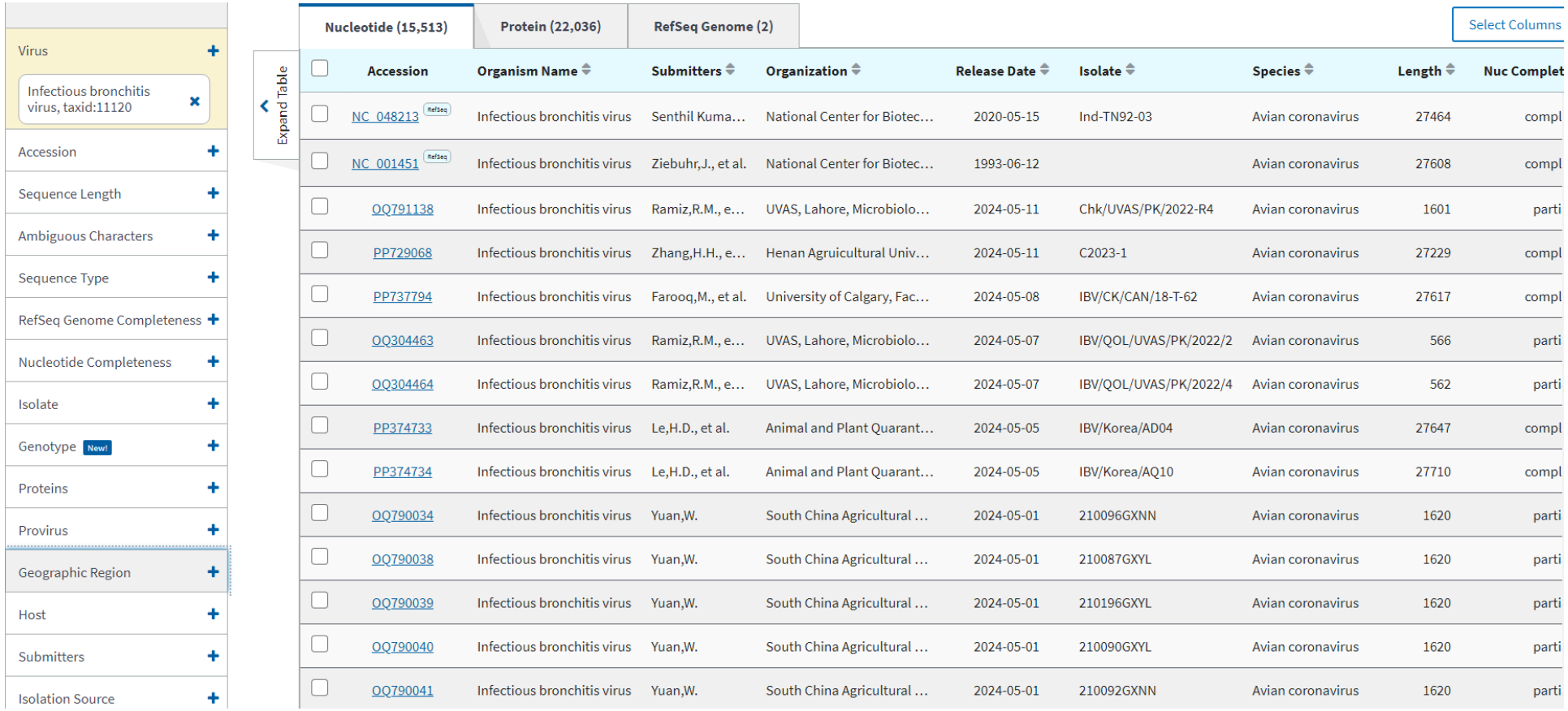
Step 1 – sequence selection

NCBI virus database



Step 1 – sequence selection

NCBI virus database



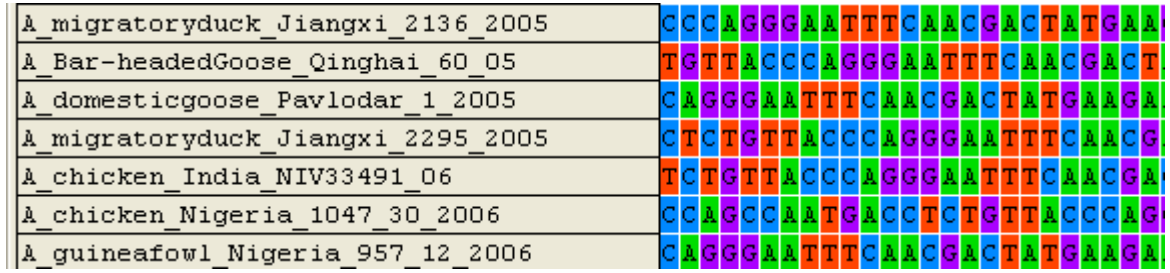
Accession	Organism Name	Submitters	Organization	Release Date	Isolate	Species	Length	Nuc Complet
NC_048213 RefSeq	Infectious bronchitis virus	Senthil Kuma…	National Center for Biotec…	2020-05-15	Ind-TN92-03	Avian coronavirus	27464	compl
NC_001451 RefSeq	Infectious bronchitis virus	Ziebuhr,J., et al.	National Center for Biotec…	1993-06-12		Avian coronavirus	27608	compl
OQ791138	Infectious bronchitis virus	Ramiz,R.M., e…	UVAS, Lahore, Microbiolo…	2024-05-11	Chk/UVAS/PK/2022-R4	Avian coronavirus	1601	parti
PP729068	Infectious bronchitis virus	Zhang,H.H., e…	Henan Agruicultural Univ…	2024-05-11	C2023-1	Avian coronavirus	27229	compl
PP737794	Infectious bronchitis virus	Farooq,M., et al.	University of Calgary, Fac…	2024-05-08	IBV/CK/CAN/18-T-62	Avian coronavirus	27617	compl
OQ304463	Infectious bronchitis virus	Ramiz,R.M., e…	UVAS, Lahore, Microbiolo…	2024-05-07	IBV/QOL/UVAS/PK/2022/2	Avian coronavirus	566	parti
OQ304464	Infectious bronchitis virus	Ramiz,R.M., e…	UVAS, Lahore, Microbiolo…	2024-05-07	IBV/QOL/UVAS/PK/2022/4	Avian coronavirus	562	parti
PP374733	Infectious bronchitis virus	Le,H.D., et al.	Animal and Plant Quarant…	2024-05-05	IBV/Korea/AD04	Avian coronavirus	27647	compl
PP374734	Infectious bronchitis virus	Le,H.D., et al.	Animal and Plant Quarant…	2024-05-05	IBV/Korea/AQ10	Avian coronavirus	27710	compl
OQ790034	Infectious bronchitis virus	Yuan,W.	South China Agricultural …	2024-05-01	210096GXNN	Avian coronavirus	1620	parti
OQ790038	Infectious bronchitis virus	Yuan,W.	South China Agricultural …	2024-05-01	210087GXYL	Avian coronavirus	1620	parti
OQ790039	Infectious bronchitis virus	Yuan,W.	South China Agricultural …	2024-05-01	210196GXYL	Avian coronavirus	1620	parti
OQ790040	Infectious bronchitis virus	Yuan,W.	South China Agricultural …	2024-05-01	210090GXYL	Avian coronavirus	1620	parti
OQ790041	Infectious bronchitis virus	Yuan,W.	South China Agricultural …	2024-05-01	210092GXNN	Avian coronavirus	1620	parti

Step 2 – alignment

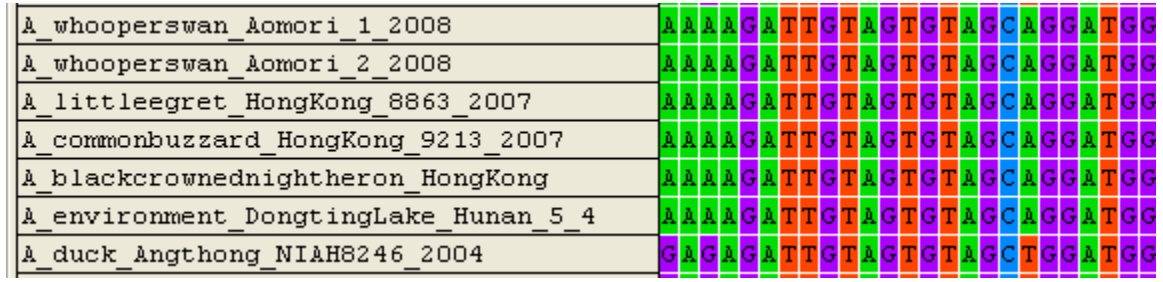
Dataset preparation and refinement - sequence selection

Sequence alignment

NOT Aligned



Aligned

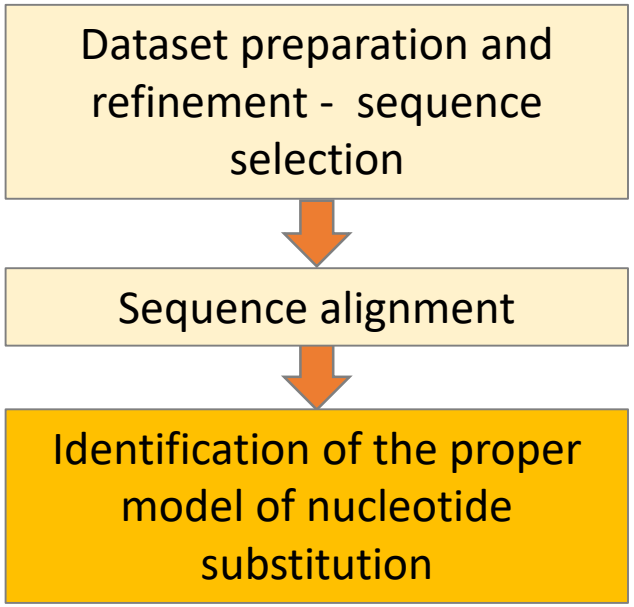


An alignment program seeks an arrangement that maximizes the net score

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

Aligment – 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)

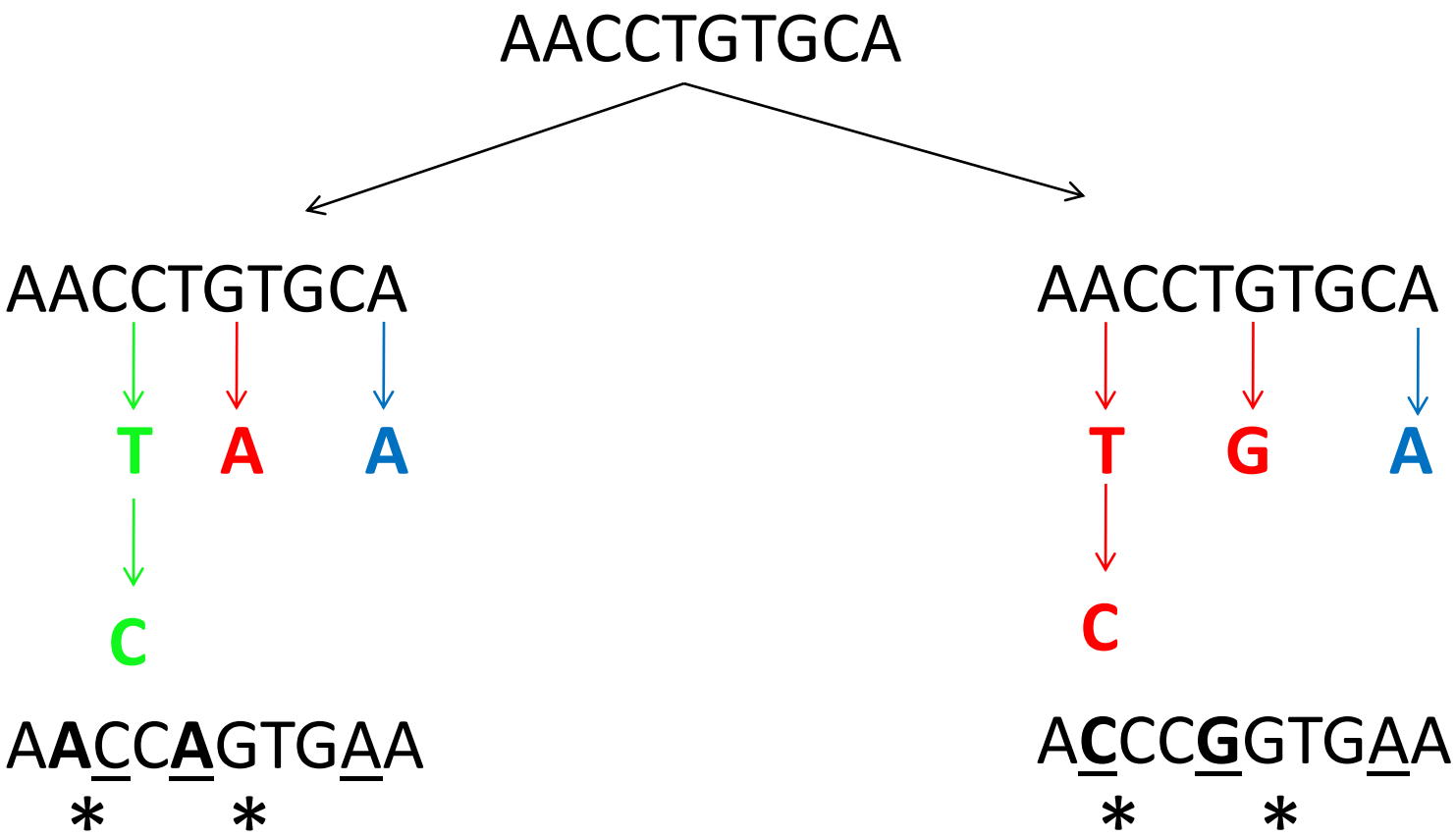
Seq1 AATCTGTGTA
Seq2 ATCCTGGGTT

P-distance=0.4

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

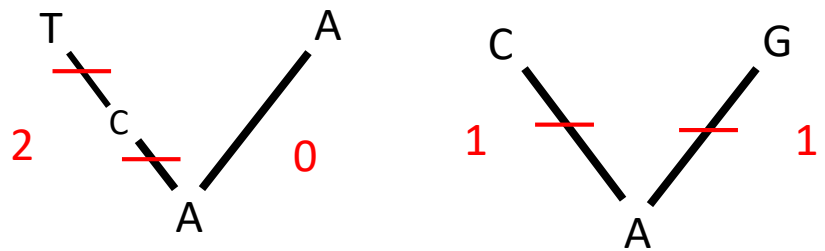
Step 3 – model of nucleotide substitution

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.

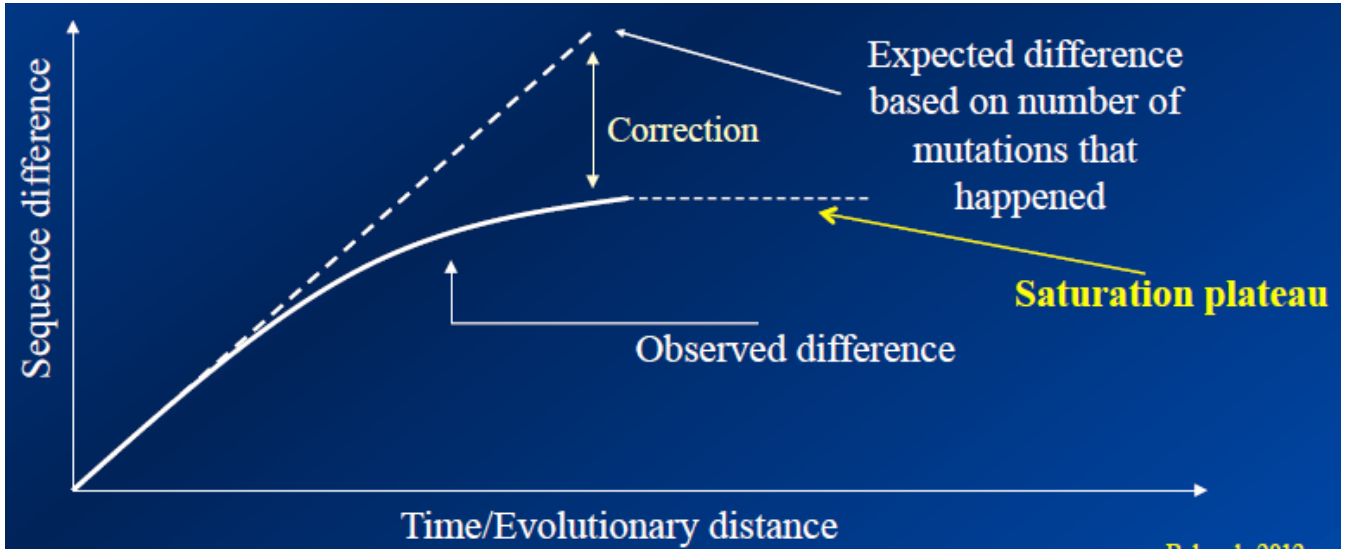


In both cases the 2 descendent sequences show only a single difference, while there have been 2 substitutions. Simply counting the number of differences between the 2 sequences underestimates the real amount of evolutionary changes

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:

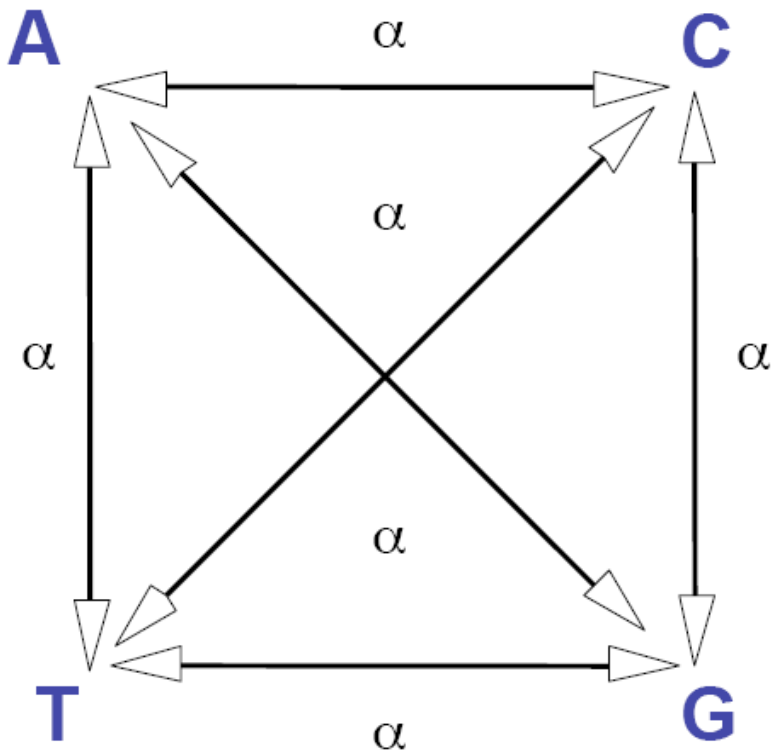
i. All bases evolve independently

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

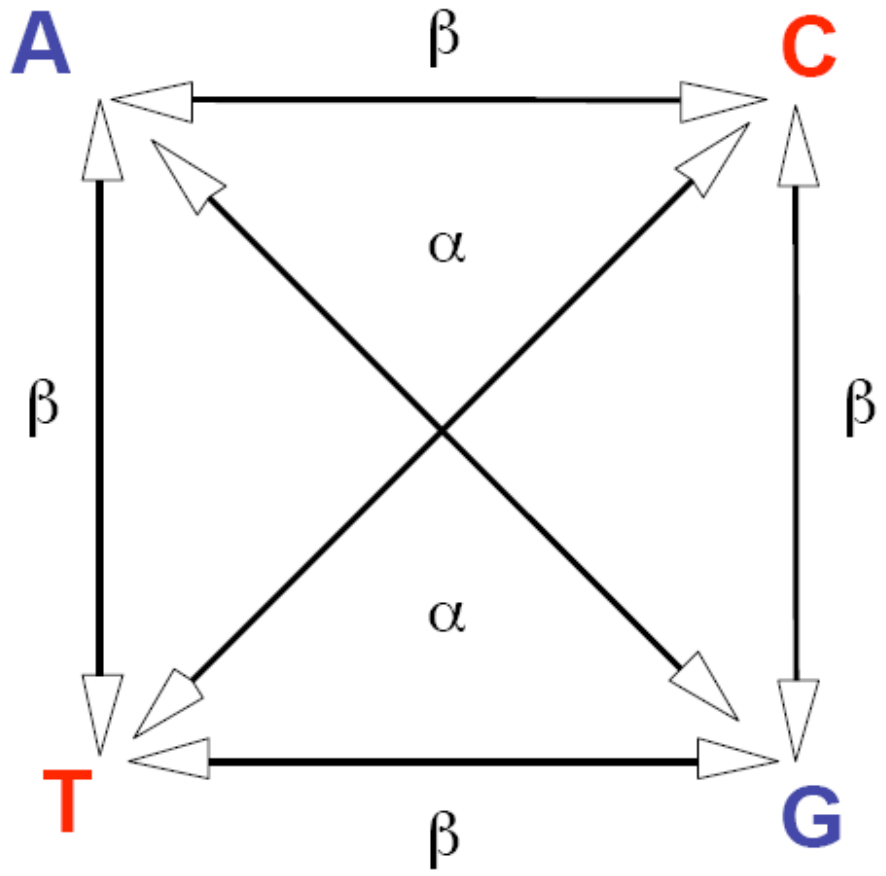
In Real Data, Transitions (α) and Transversions (β) occur at different rates

Transitions are generally more frequent than transversions

Kimura's 2 parameter (K2P)

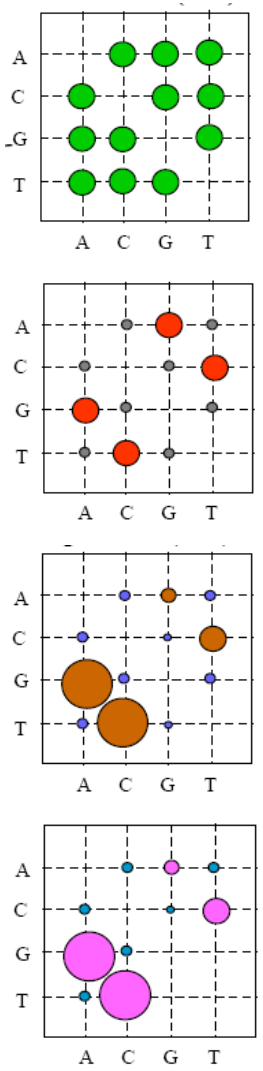
Assumptions:

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)



Most complex (many parameters)

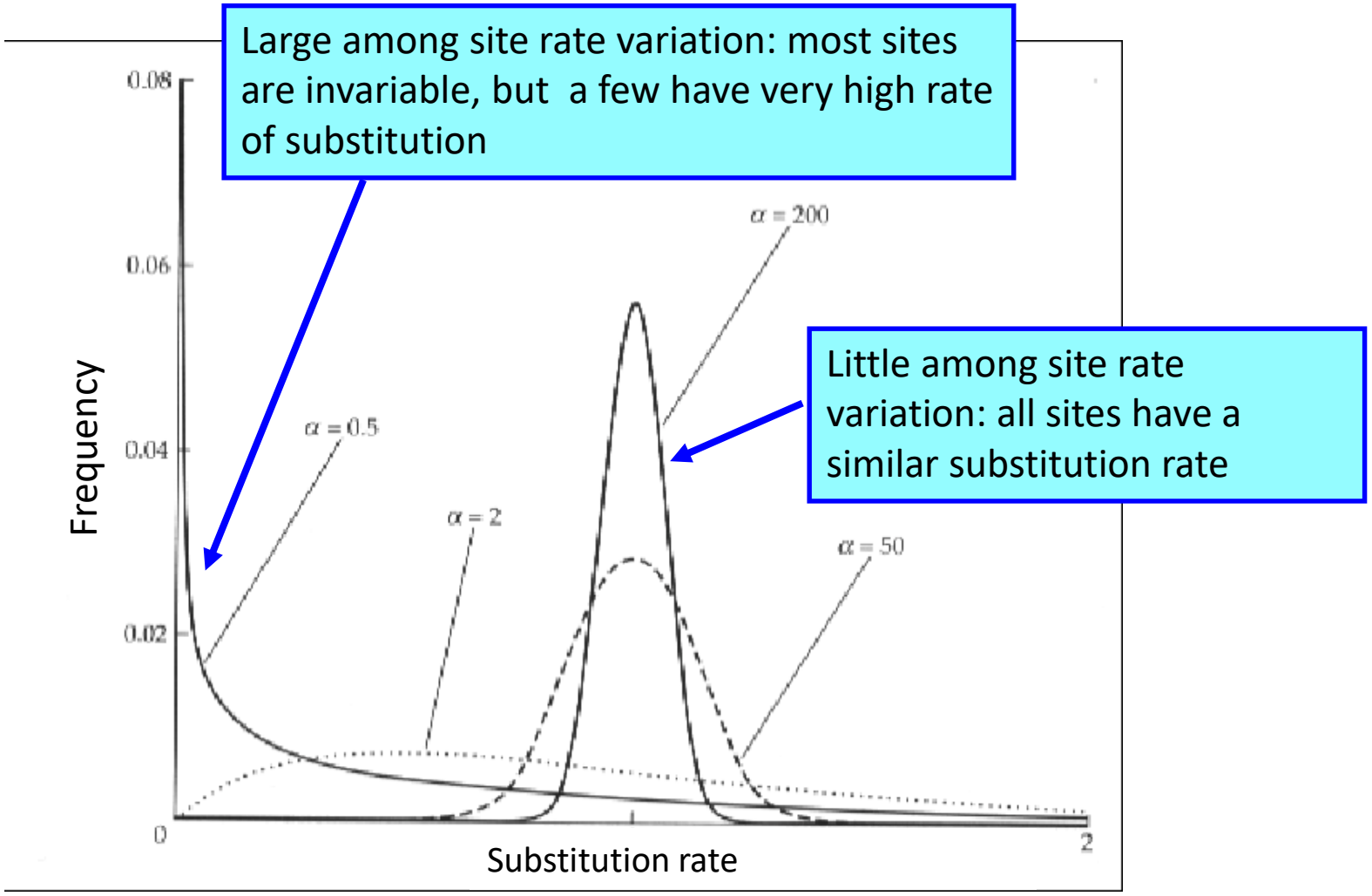
Base frequencies are equal and all substitutions are equally likely (Jukes-Cantor)

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)

Step 3 – model of nucleotide substitution



Step 4 – method selection

Dataset preparation and refinement - sequence selection

↓

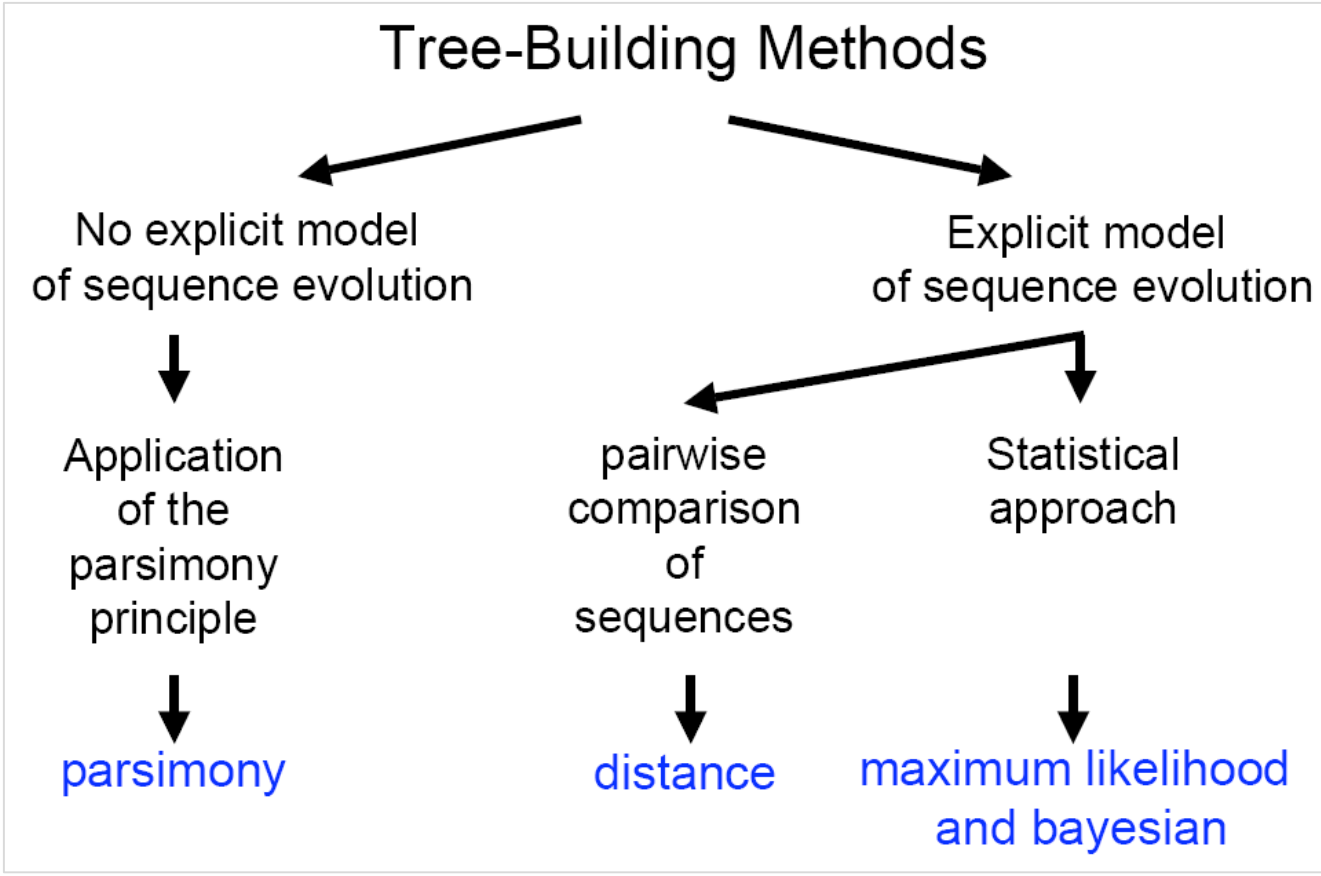
Sequence alignment

↓

Identification of the proper model of nucleotide substitution

↓

Application of a method to construct a phylogenetic tree



Step 4 – method selection

DISTANCE METHODS

Distance methods

- UPGMA
- NJ

Distance matrix

	1	2	3	4	5	6
1. A/turkey/Saudi Arabia/6732-6/2007						
2. A/Turkey/651242/2006	0.015					
3. A/chicken/Krasnodar/300/07	0.026	0.018				
4. A/Goose/Guangdong/1/96	0.050	0.044	0.053			
5. A/chicken/Nigeria/1071-7/2007	0.017	0.013	0.022	0.049		
6. A/chicken/Afghanistan/1207/2006	0.017	0.009	0.012	0.046	0.013	

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 easly reconstruct the evolutionary history of those sequences

Advantages:

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

Disadvantages:

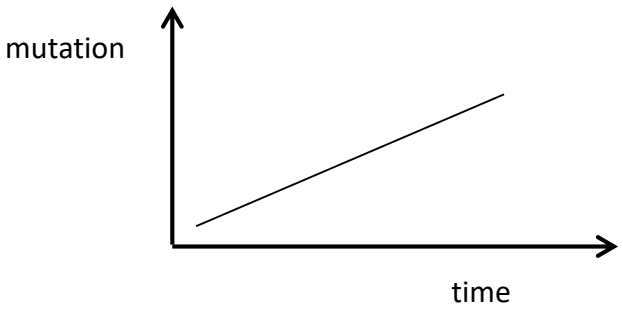
- only produces one tree with no indication of its quality
- reduces all sequence information into a single distance value
- dependent on the evolutionary model used (preferentially this model should be estimated from the data)

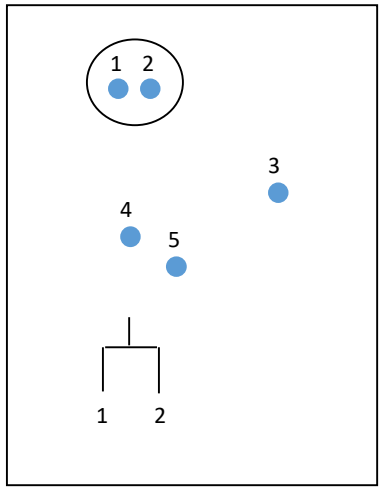
Step 4 – method selection

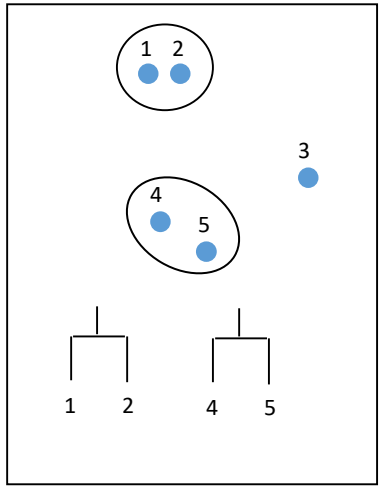
DISTANCE METHODS: **UPGMA**

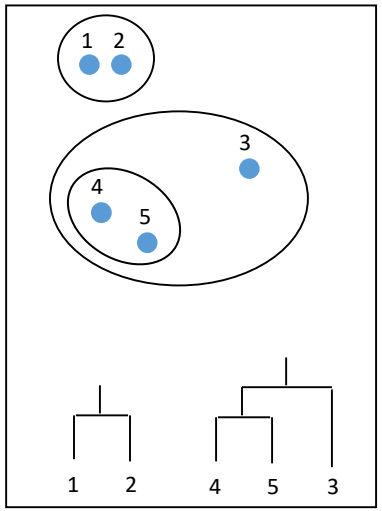
UPGMA is a clustering algorithm that computes the distance between clusters using average pairwise distance

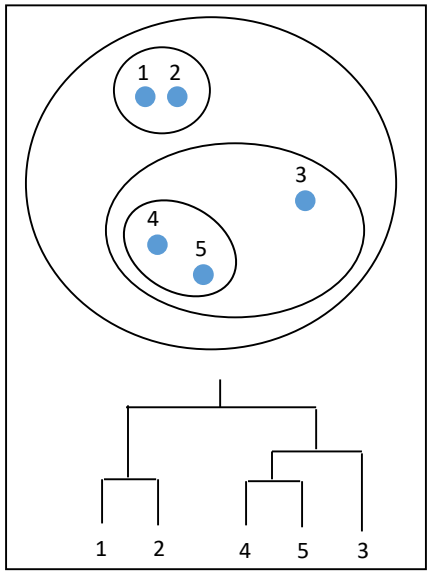
UPGMA assumes a constant molecular clock: all species represented by the leaves in the tree are assumed to accumulate mutations (and thus evolve) at the same rate. This is a major pitfalls of UPGMA.









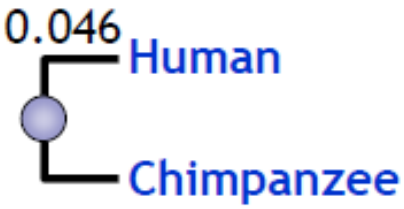


Step 4 – method selection

DISTANCE METHODS: **UPGMA**

	Human	Chimpanzee	Gorilla
Chimpanzee	0.093		
Gorilla	0.109	0.114	
Orangutan	0.178	0.194	0.188

$$\frac{d_{human-chimp}}{2}$$



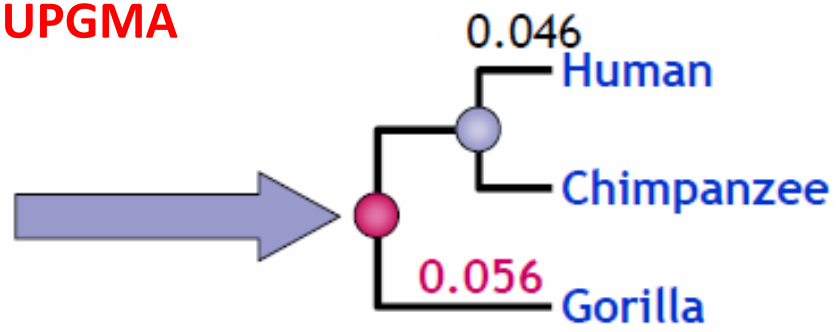
$$d_{Gorilla(human+Chimp)} = \frac{0.109 + 0.114}{2} = 0.111$$

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

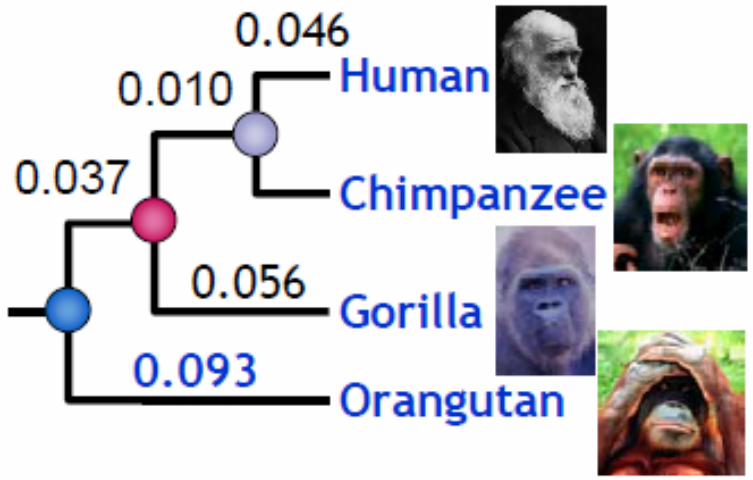
Step 4 – method selection

DISTANCE METHODS: **UPGMA**

	HumChimp	Gorilla
Gorilla	0.111	
Orangutan	0.186	0.188



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



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
- permist 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

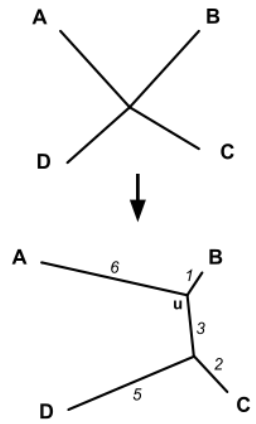
distance matrix

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) x d(ij) - r(i) - r(j)
i.e. M(A,B)= 2 x 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 3: Now we choose as neighbors those two OTUs for which Mij 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: 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 4 – method selection

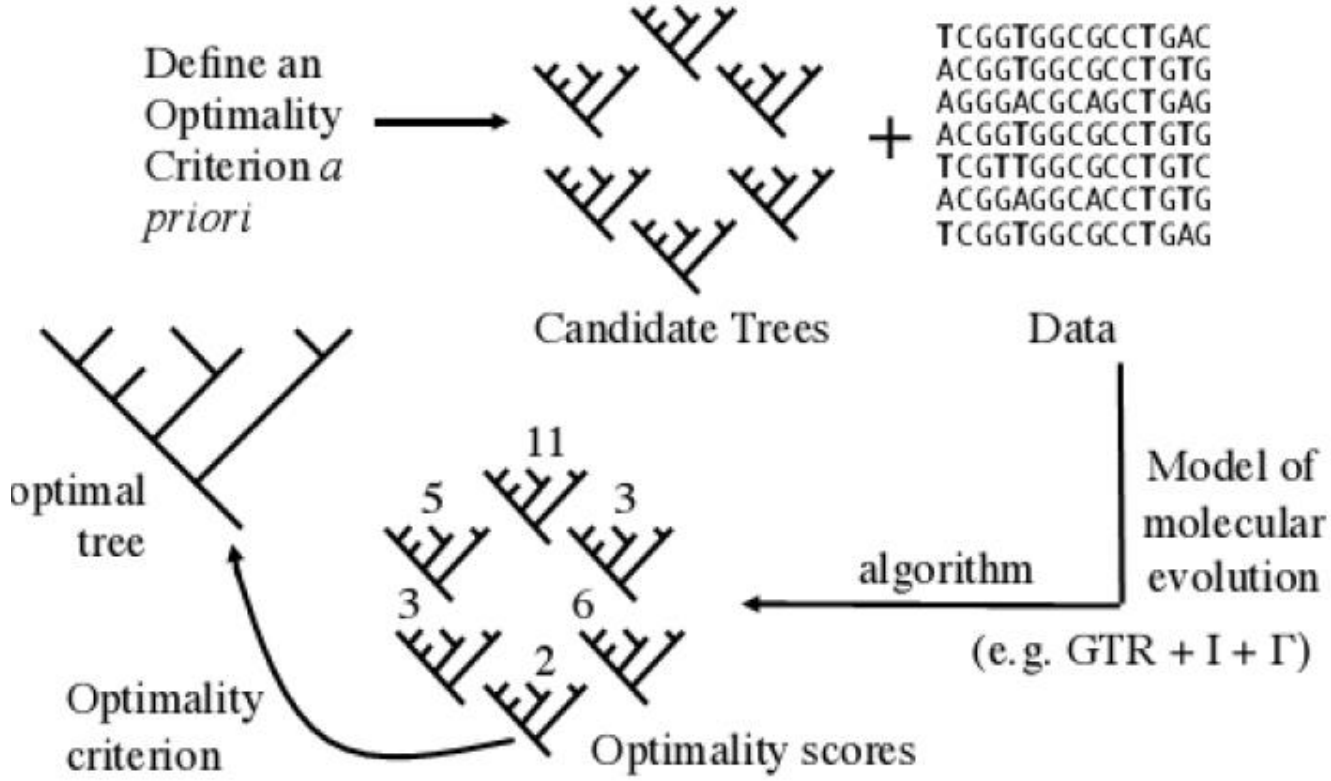
Statistical approach

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)

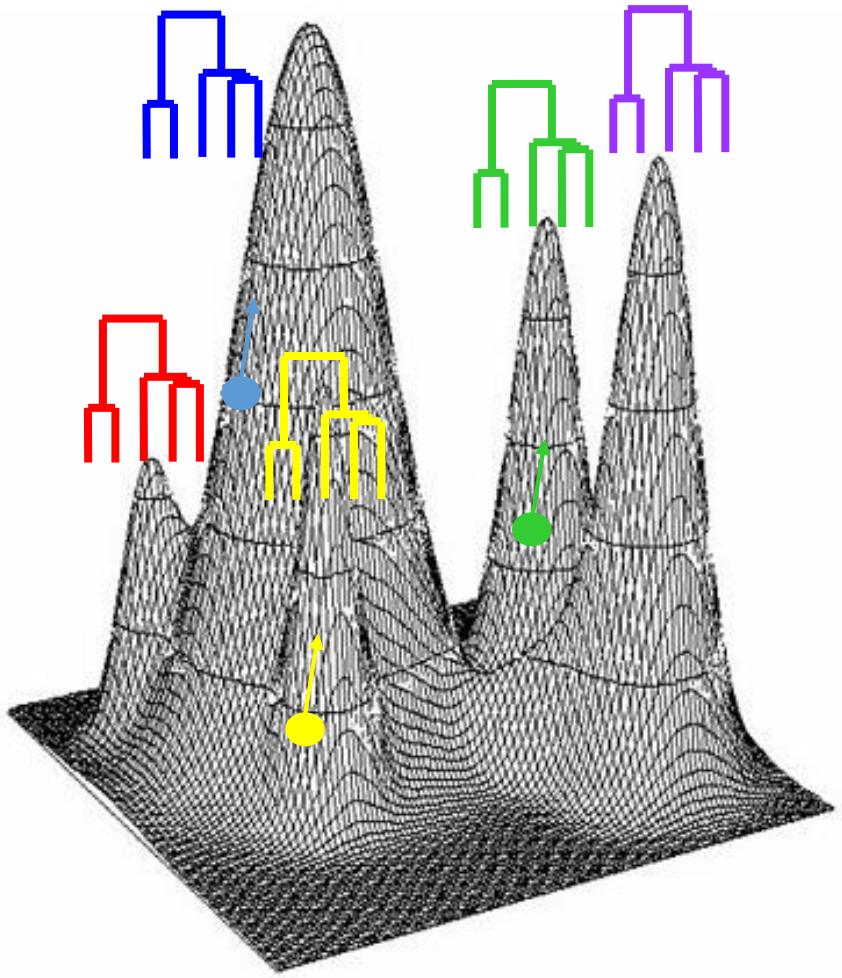
Step 4 – method selection

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.



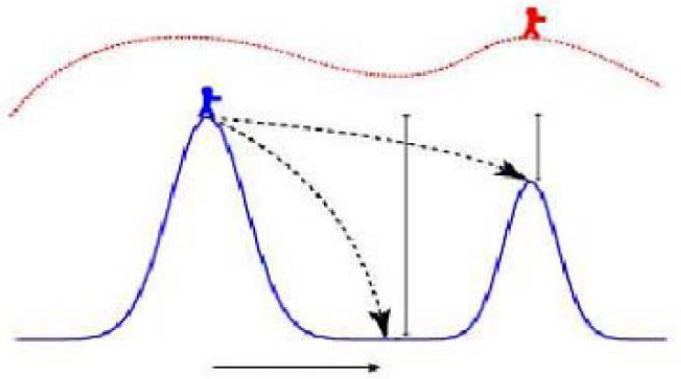
The highest hill represent the best tree topology

Step 4 – method selection

STATISTICAL METHODS: **Bayesian method**

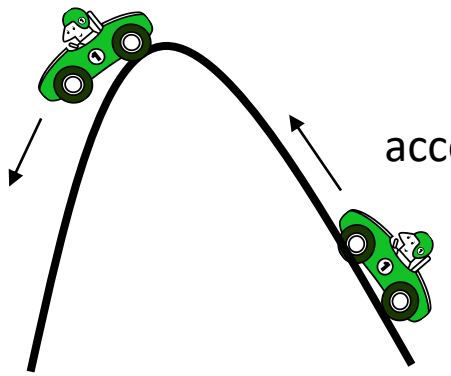
There are two methods to explore the tree space:

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



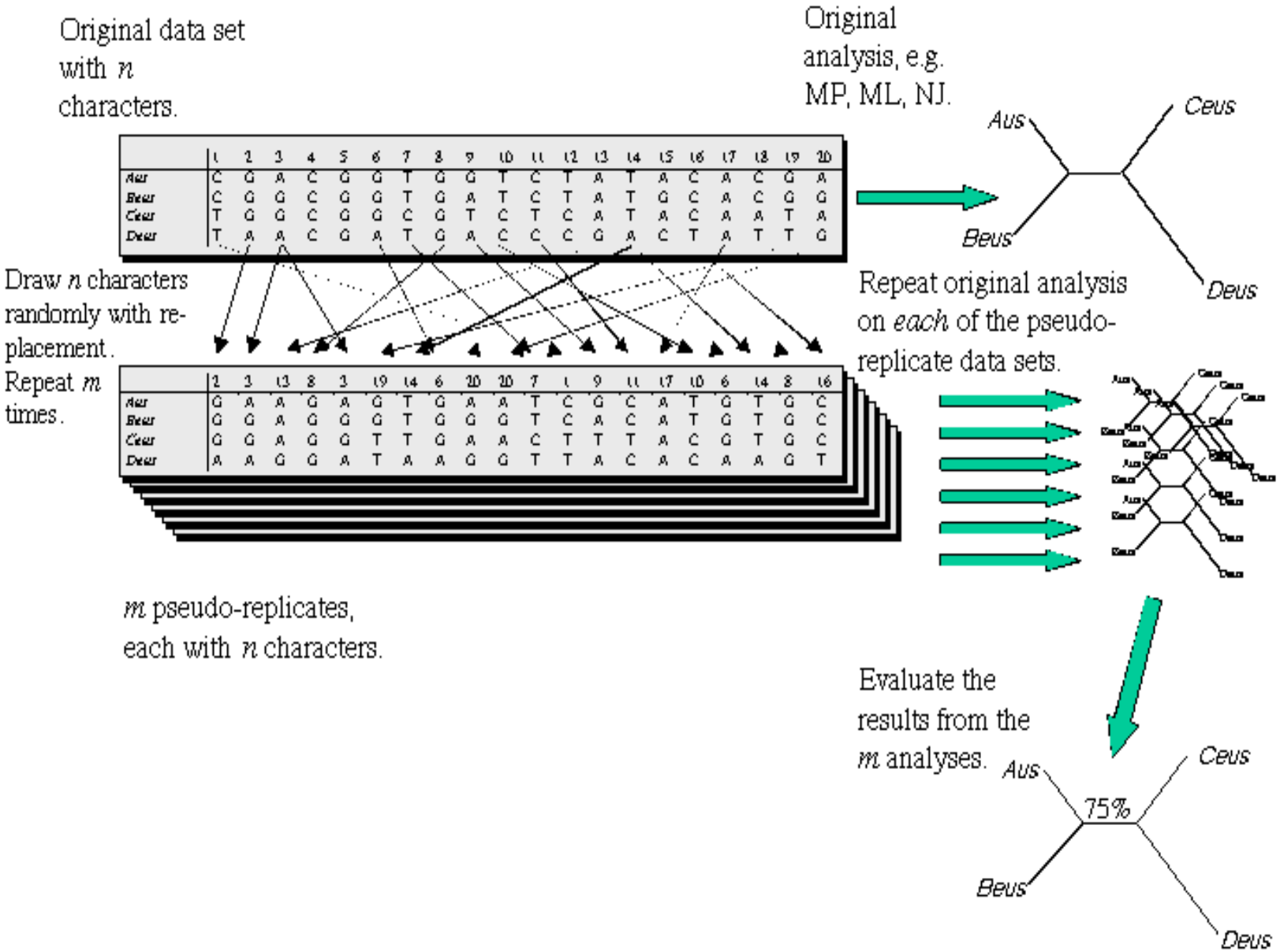
- **HILL CLIMBING**: Allows to find the best tree in the hill

Not accepted



accepted

Bootstrap analyses



Thank you for your attention!