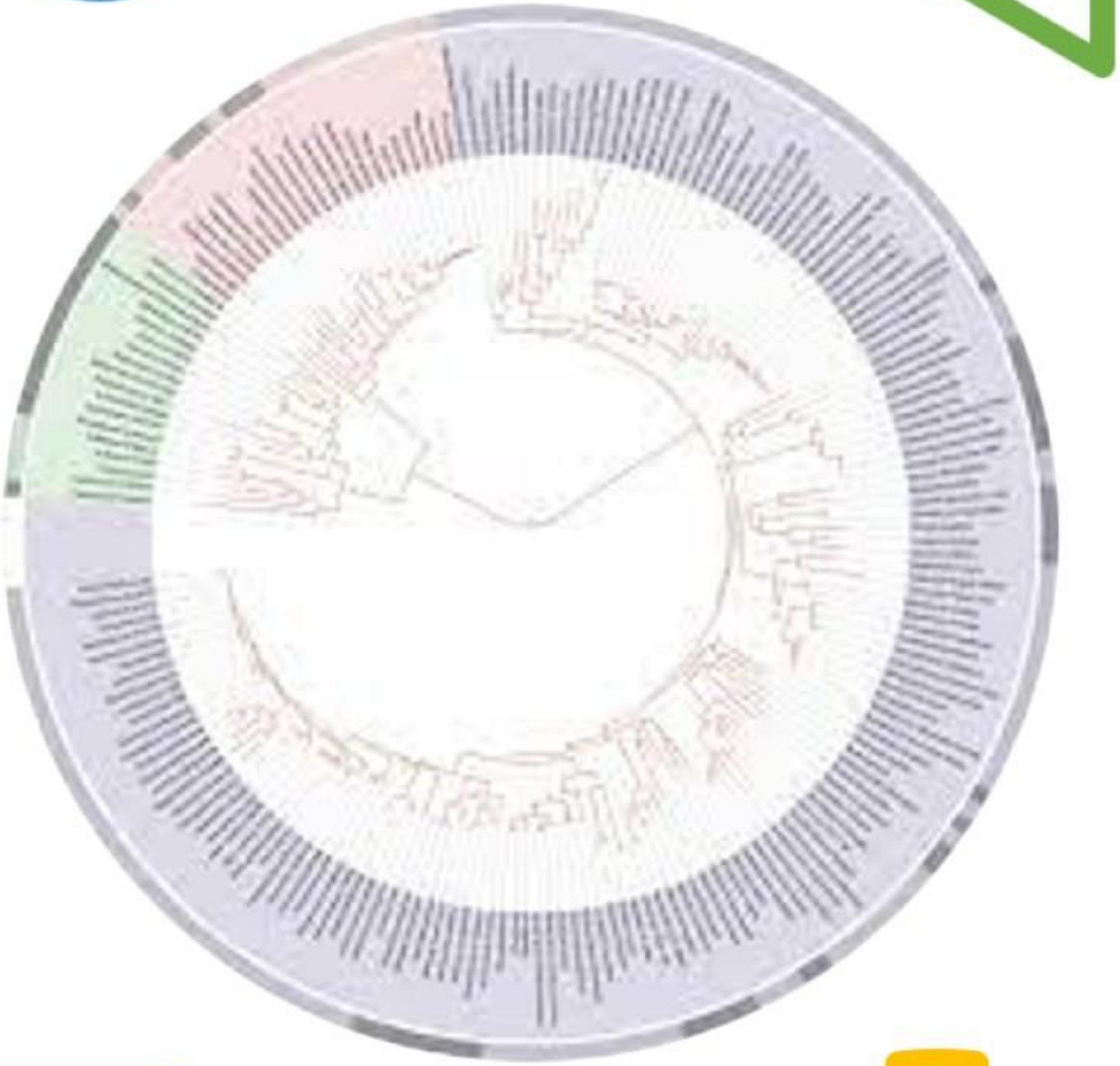


Phylogenetic Tree (part 2)

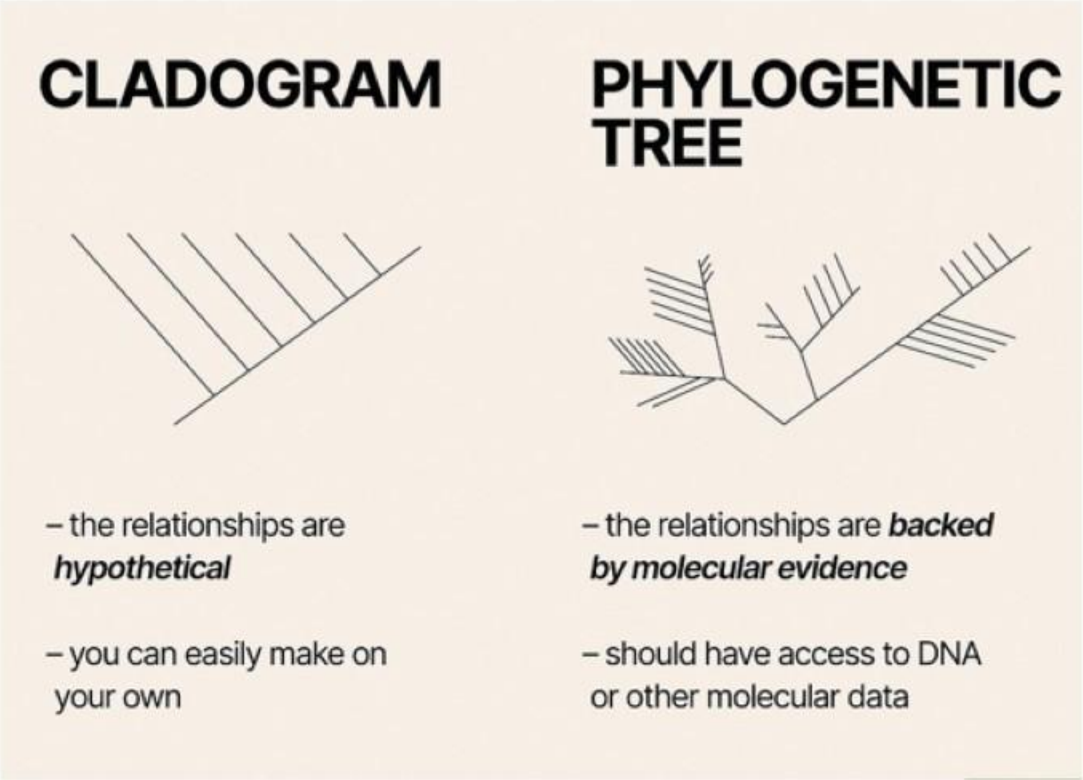
Bioinformatics Lec 8/2025

By

Dr Delveen R. Ibrahim

How to interpret the results of a Phylogenetic Tree

- Remember, Phylogram is different from cladogram and different comparing to a dendrogram .
- Phylogram is a scaled tree based on molecular evidence while cladogram is not.
- Dendrogram has a scale of similarity or distances between the samples, and clade / cluster are grouped according to that percentages of similarity or distance.



CLADOGRAM

– the relationships are ***hypothetical***

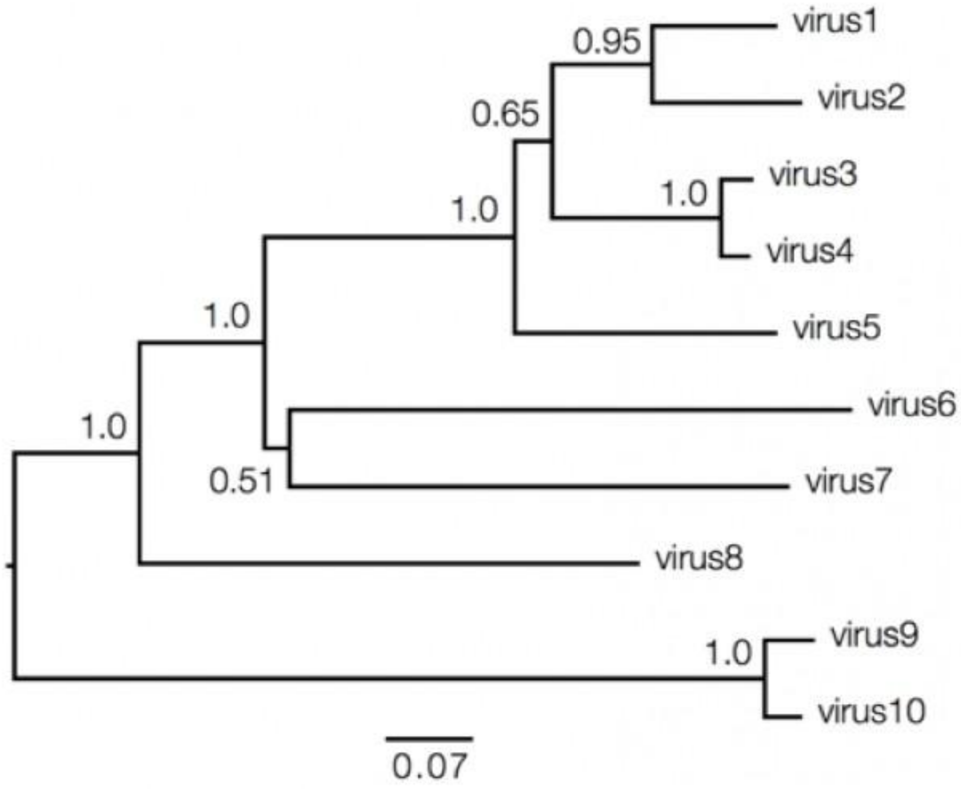
– you can easily make on your own

PHYLOGENETIC TREE

– the relationships are ***backed by molecular evidence***

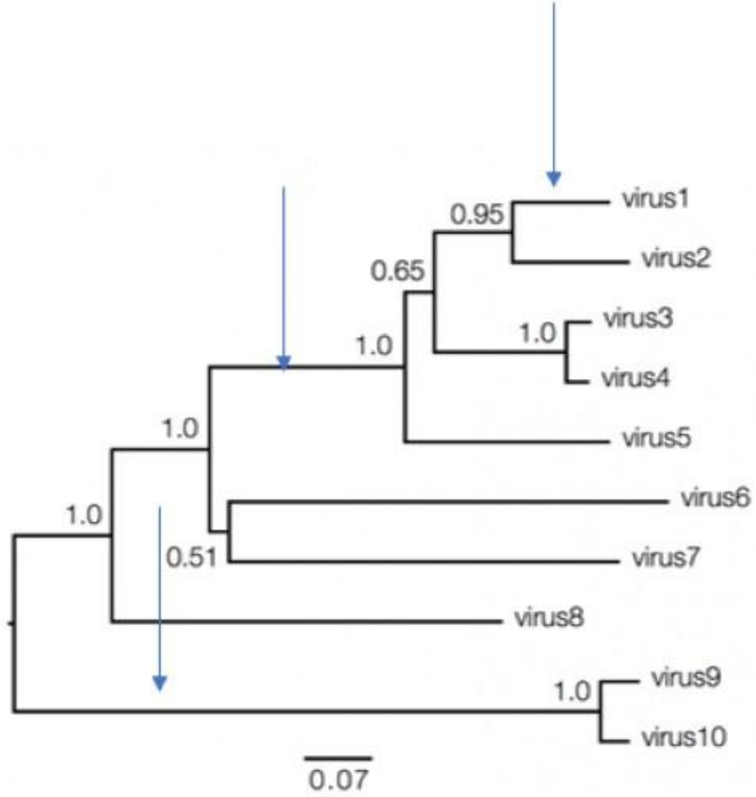
– should have access to DNA or other molecular data

A scaled Phylogenetic tree

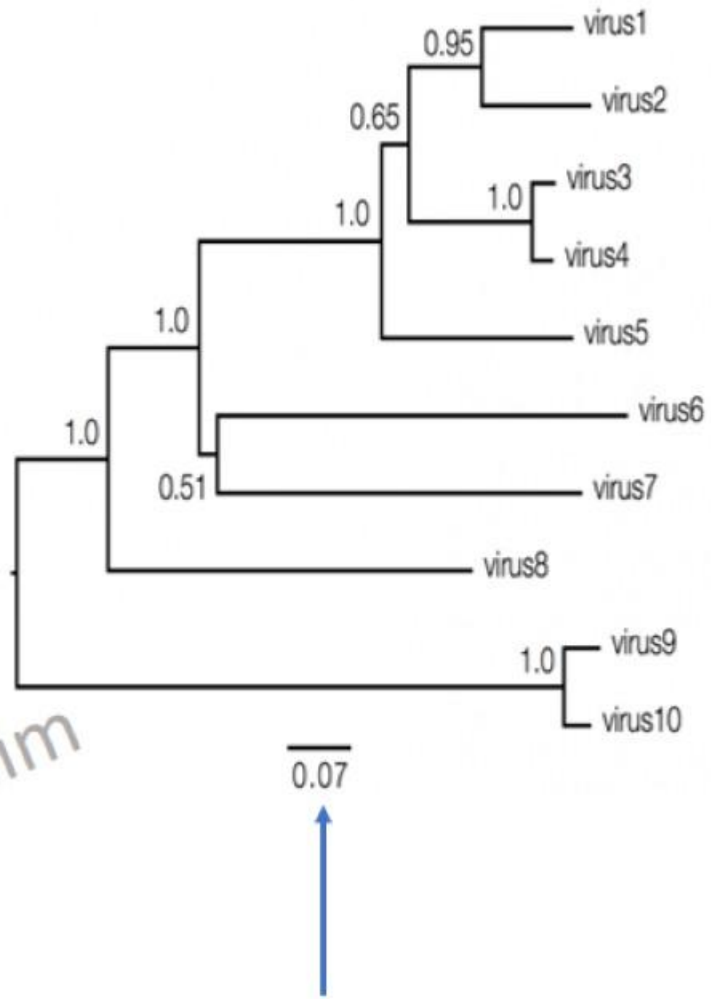


What information does the tree contain?

- We can start with the dimensions of the figure. In this figure the horizonal dimension gives the amount of genetic change.
- The horizonal lines are branches (indicated by arrow) and represent evolutionary lineages changing over time. The longer the branch in the horizonal dimension, the larger the amount of change.

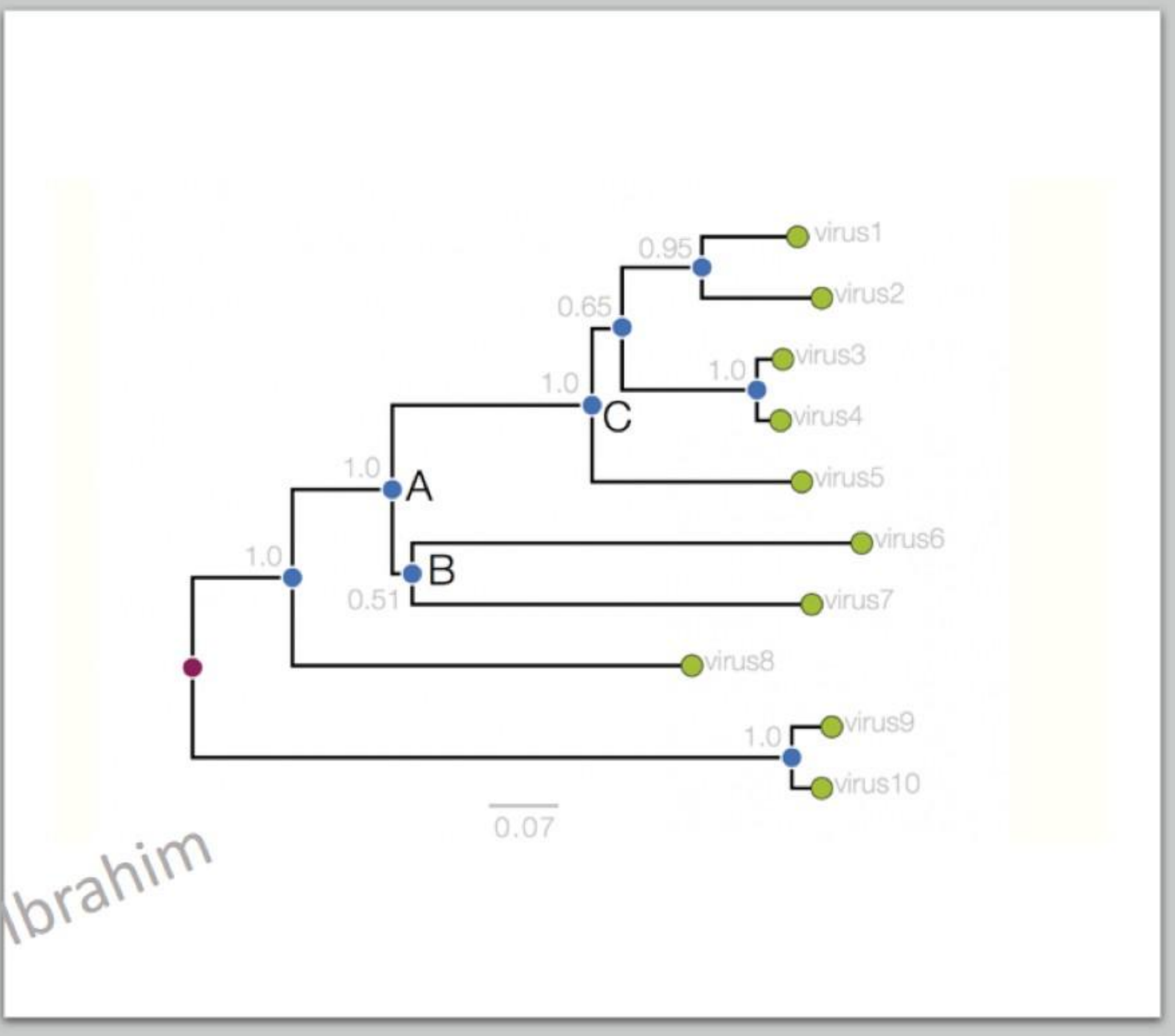


- The bar at the bottom of the figure provides a scale for this. In this case the line segment with the number ‘0.07’ shows the length of branch that represents an amount genetic change of 0.07.
- The units of branch length are usually nucleotide substitutions per site – that is the number of changes or ‘substitutions’ divided by the length of the sequence (although they may be given as % change, i.e., the number of changes per 100 nucleotide sites).
- The vertical lines simply tell you which horizontal line connects to which and how long they are.



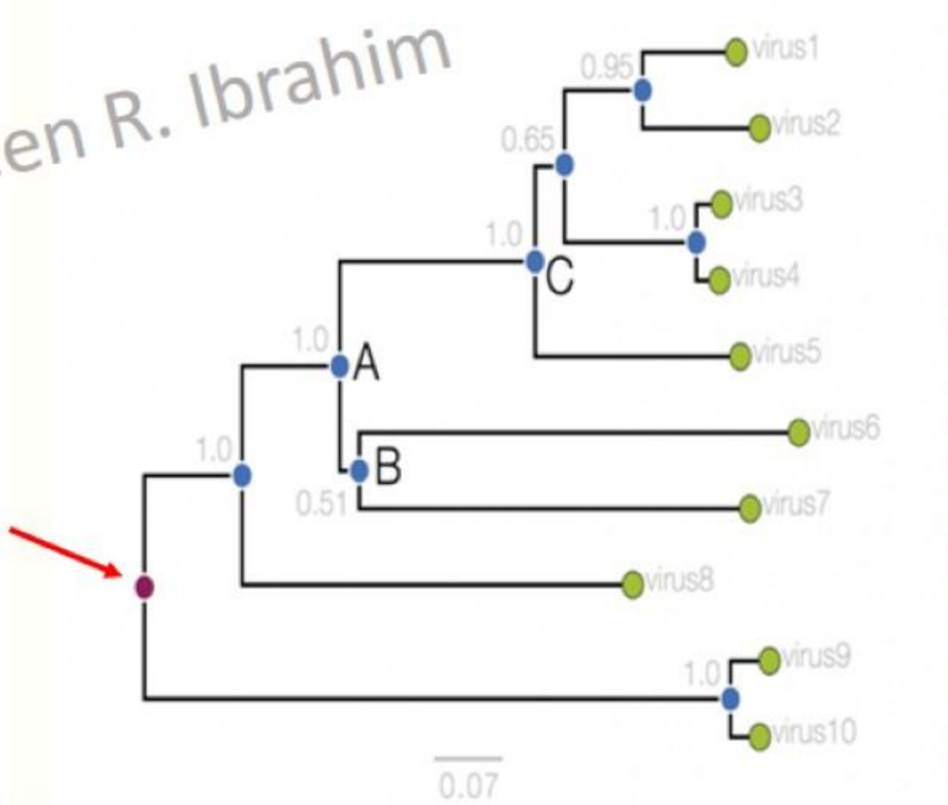
Tree structure

- As explained previously, the tree has different types of nodes.
- This can be broken down into nodes (external and internal)
- The tips or external nodes are shown here with green circles, and these represent the actual viruses sampled and sequences (in this example).
- These are our data, and we usually know information about these, beyond the actual sequence, such as when they were collected, what host they were in, where that host was found, clinical features of the disease.



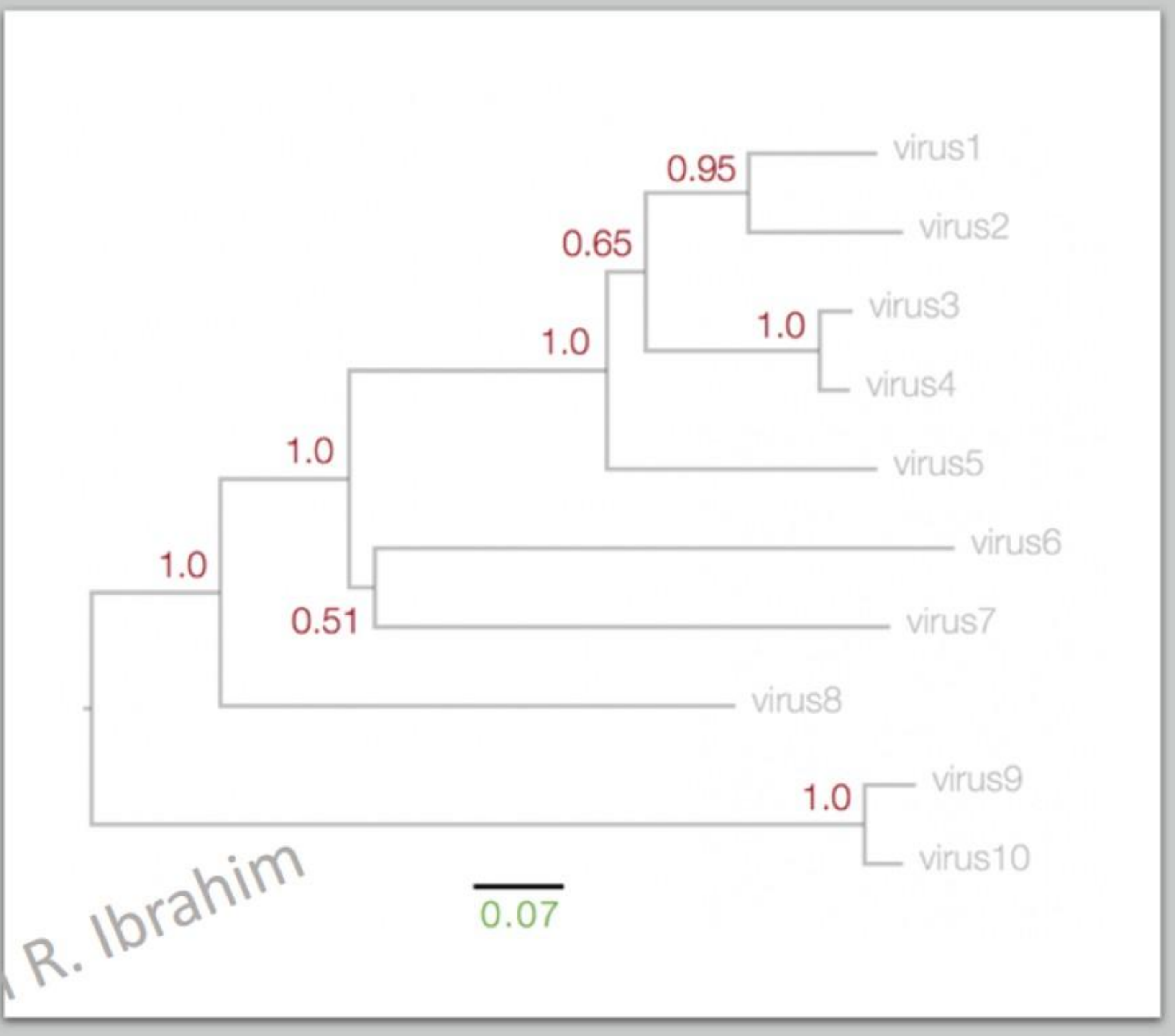
The internal nodes are represented by blue circles, and these represent putative ancestors for the sampled viruses

- An ancestor in this context is an infected host at sometime in the past that in turn infected 2 or more new hosts producing chains of infections that lead to the sampled viruses. The branches then represent this chain of infections.
- This tree is rooted which suggests we know where the ultimate common ancestor of all the sampled viruses was (the red circle).
- Knowing this gives the tree an order of branching events in the horizonal dimension:
- Ancestor ‘A’ exists prior to ancestor's ‘B’ and ‘C’ and time is approximately flowing from left to right.

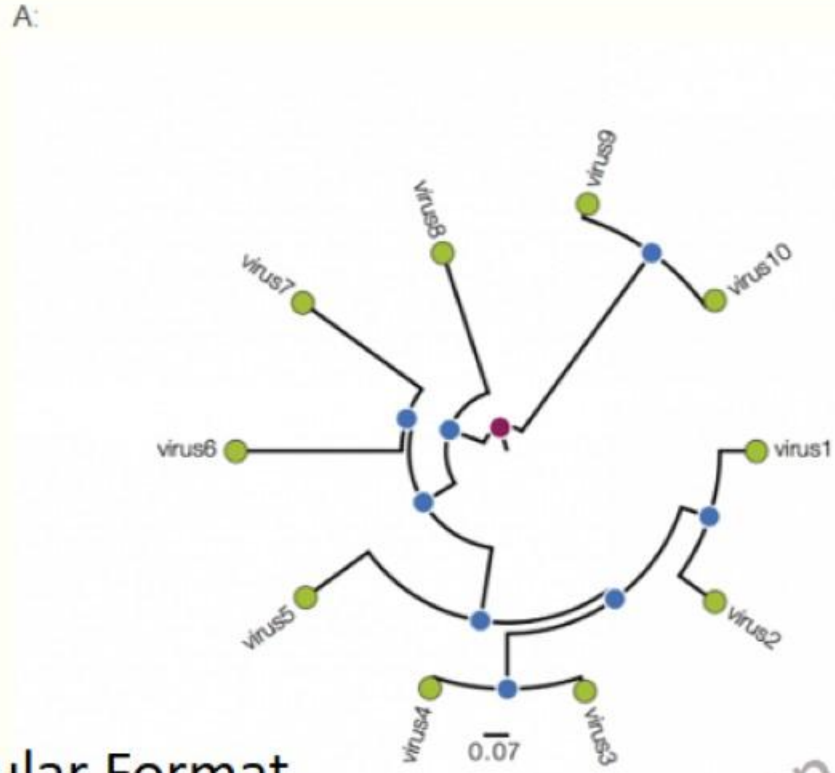


Phylogenetic tree reliability

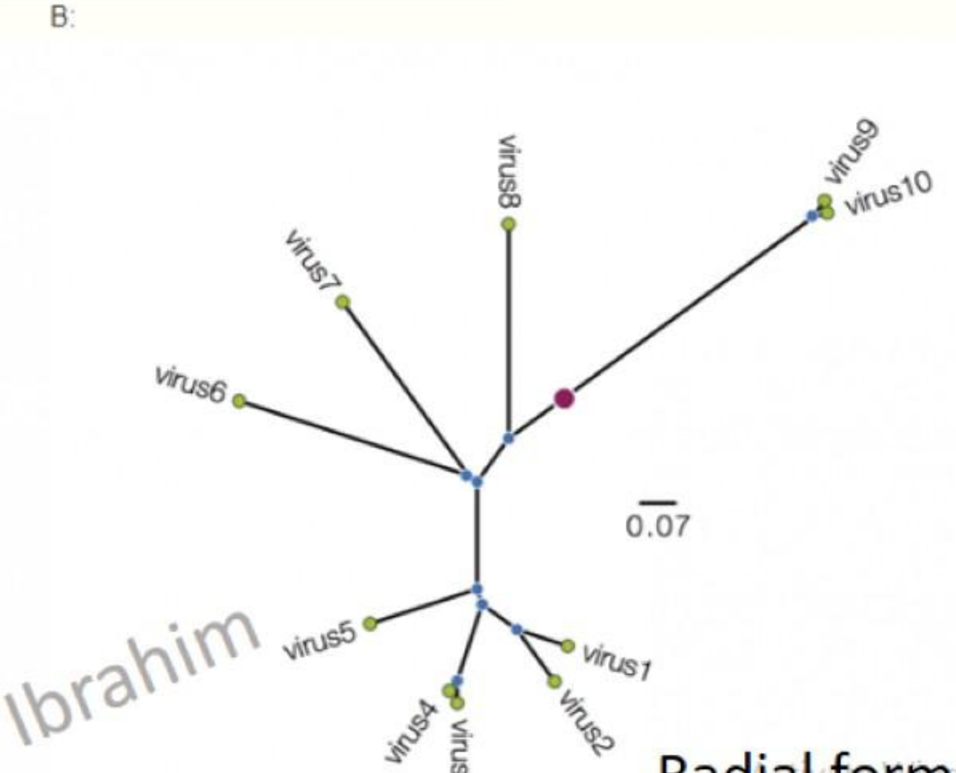
- The numbers next to each node, in red, above, represent a measure of support for the node.
- These are generally numbers between 0 and 1 (but may be given as percentages or range from 0 to 100) where 1(or 100) represents maximal support.
- These can be computed by a range of statistical approaches including **bootstrapping**.
- The details of what technique was used usually will be in the figure legend.
- A high value means that there is strong evidence that the sequences to the right of the node cluster together to the exclusion of any other.



Trees are sometimes drawn in other ways. Both these figures are representations of the same underlying tree as above:



Circular Format tree



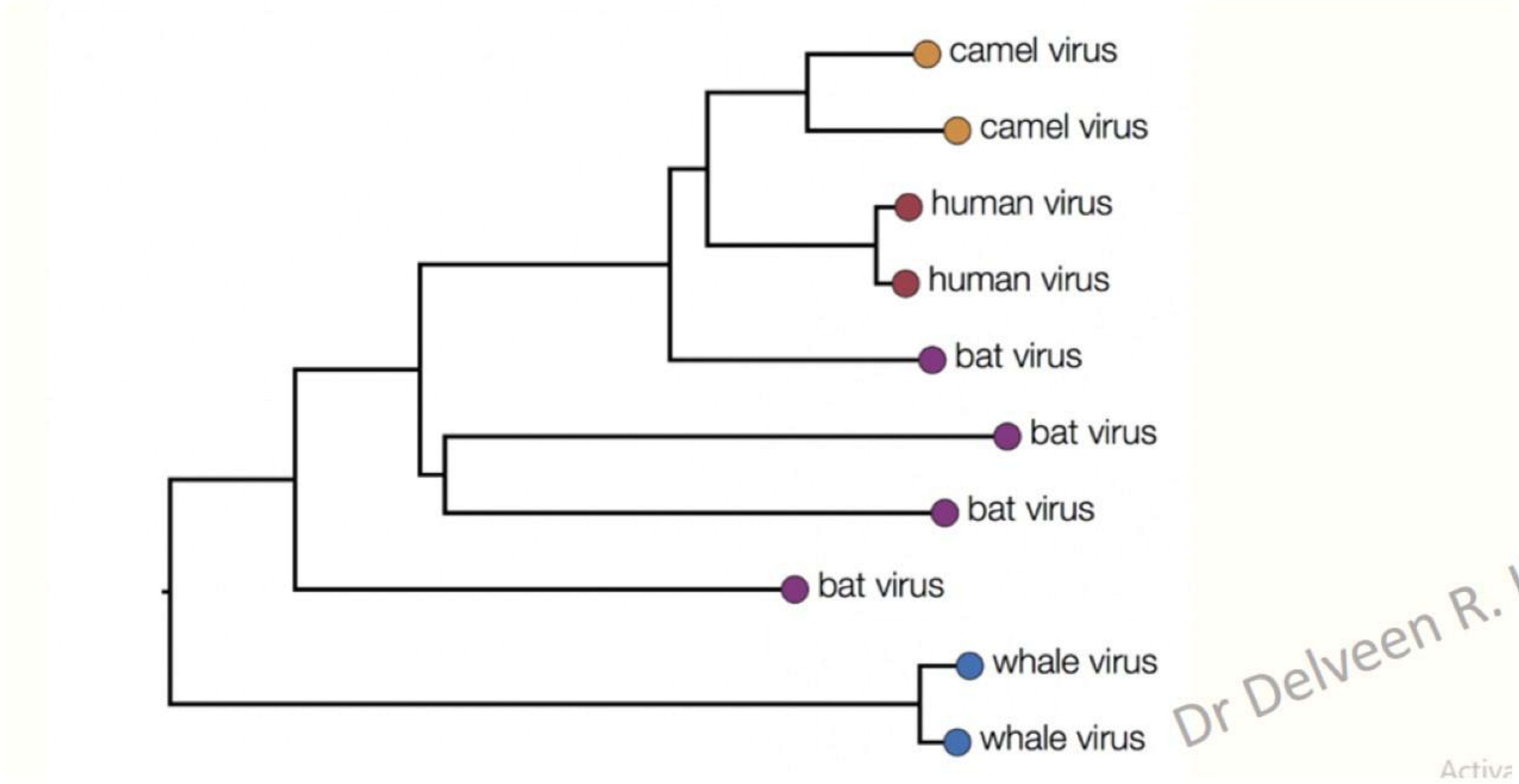
Radial format tree

Circular format vs Radial format tree

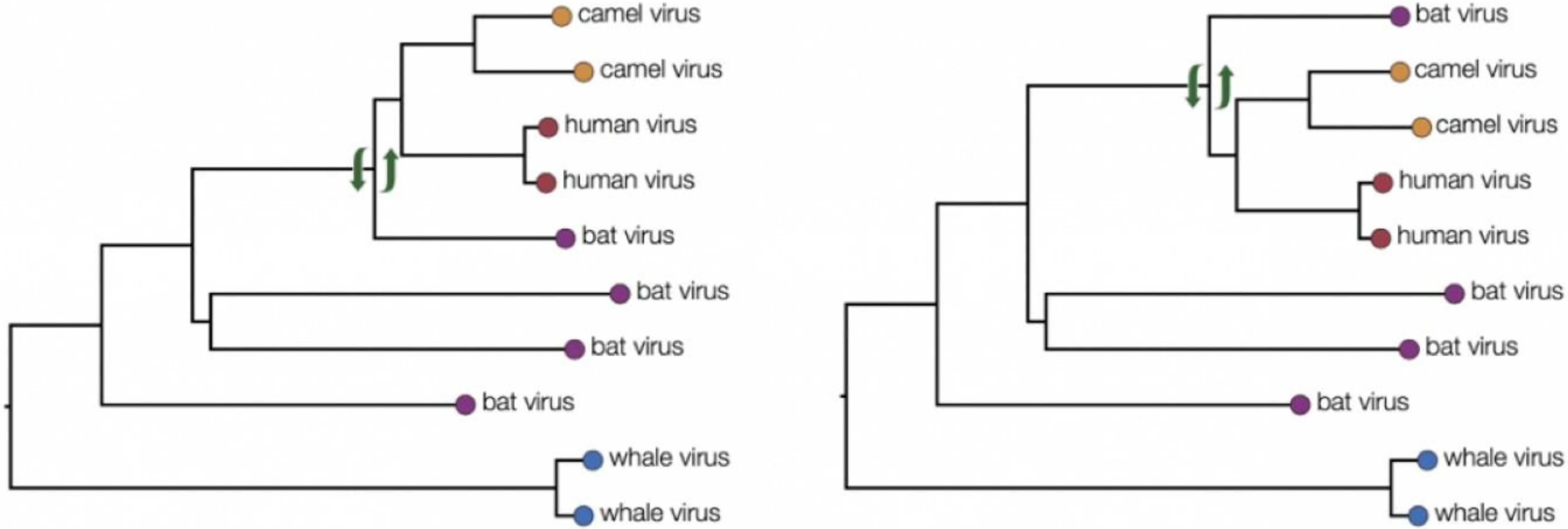
- In the pervious slide:
- **Tree A** is in polar format (often called a **circle tree**). This is basically the same as the trees above but in polar coordinates. These tree formats are often used to make a big visual impact in papers but generally have reduced readability - it is difficult to compare how far nodes are from the centre.
- **Tree B** is a **radial** format tree. This is often used when the rooting of the tree is not known (although here the root is colored with red). This format tends to clump closely related sequences together making their precise relationships difficult to see.

Reconstructing epidemiology

Here is the same tree as above but with the tips labeled by the type of host they were isolated from:



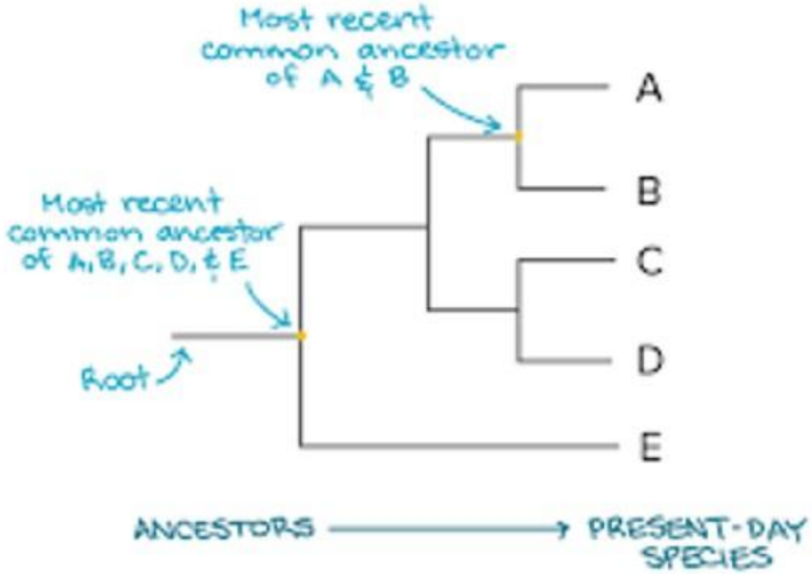
- You can immediately see that there is some structure there with viruses grouping by host.
- For example, the two viruses from humans have a closer common ancestor with each other than they do with any other virus.
- At first glance it may seem that human viruses are more closely related to bat viruses than camel viruses because they sit next to each other but remember that the vertical dimension is meaningless.
- In fact, the viruses can be swapped round at any internal node and the tree is the same, see the next figure:


camel virus
camel virus
human virus
human virus
bat virus
bat virus
bat virus
bat virus
whale virus
whale virus
bat virus
camel virus
camel virus
human virus
human virus
bat virus
bat virus
bat virus
whale virus
whale virus

So, what is the source of the human virues in this example??

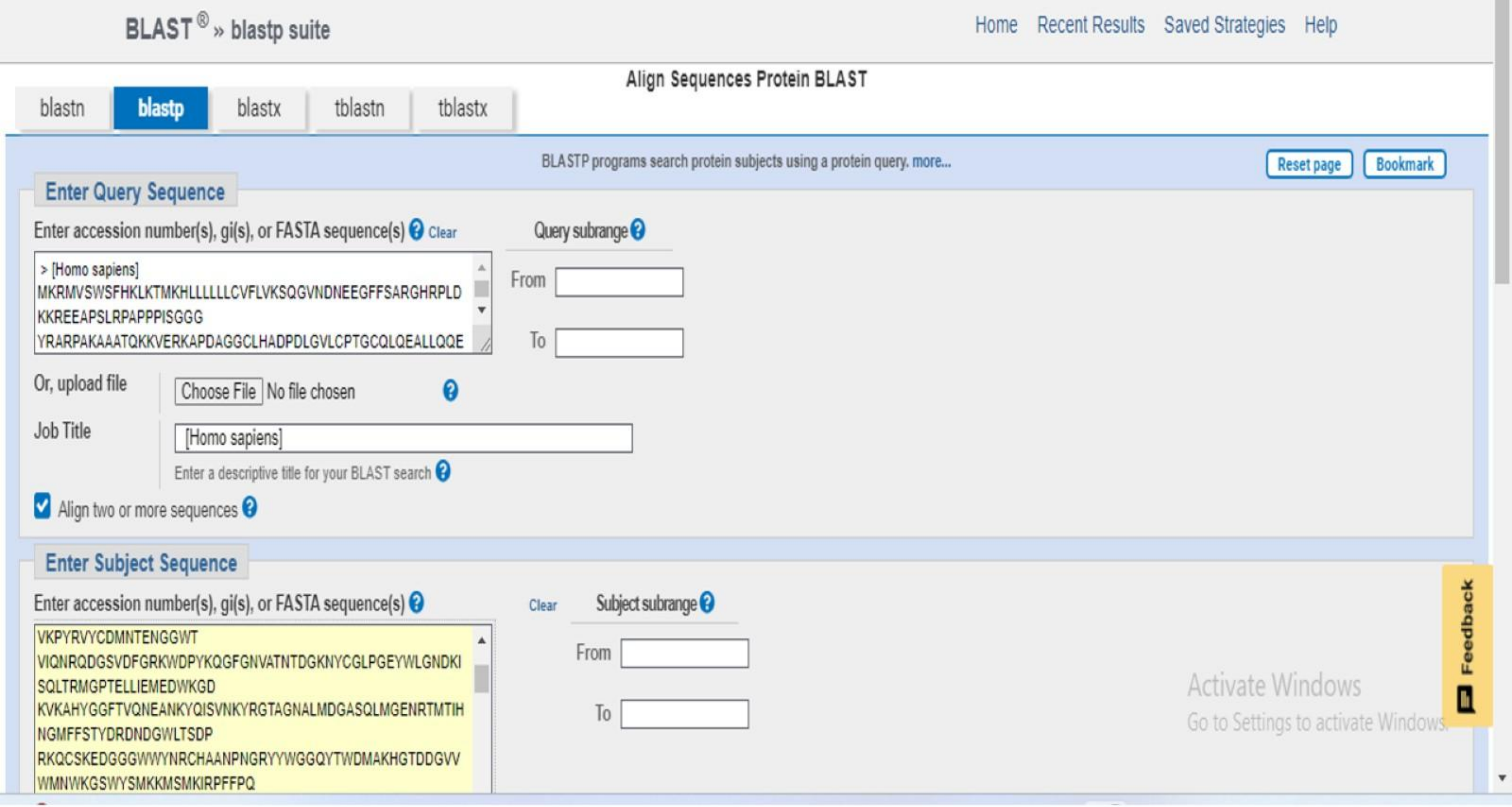
- The human and camel viruses are more closely related to each other and equally related to the bat viruses.
- This means we can't say from this tree if camels are the source of the human viruses or vice-versa, or just as likely, bats are independently the source of both human and camel outbreaks.
- We can however suggest that bats were the ultimate source of both camel and human viruses because of the much greater diversity of bat viruses.
- Another way to look at this is that the common ancestors of the human and camel viruses lie within the diversity of all the bat viruses.

Lab 8 practical tasks


Most recent common ancestor of A & B
Most recent common ancestor of A, B, C, D, & E
Root
A
B
C
D
E
ANCESTORS
PRESENT-DAY SPECIES

Task 1:

- Use the file provided in Moodle to answer the following:
- Find out what is the query sequence (the first sequence in the file) represents (belong to which gene)?
- Find the distance tree comparing the first sequence (query) and all the other sequences listed after.
- Read the distance tree and describe your results.



Align Sequences Protein BLAST

BLASTP programs search protein subjects using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
> [Homo sapiens]
MKRMVSWSFHKLKTMKHLLLLLLCVFLVKSQGVNDNEEGFFSARGHRPLD
KKREEAPSLRPAPPPISGGG
YRARPAKAAATQKKVERKAPDAGGCLHADPDLGVLCPTGCQLQEALLQQE
```

Query subrange

From

To

Or, upload file: No file chosen

Job Title: [Homo sapiens]

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

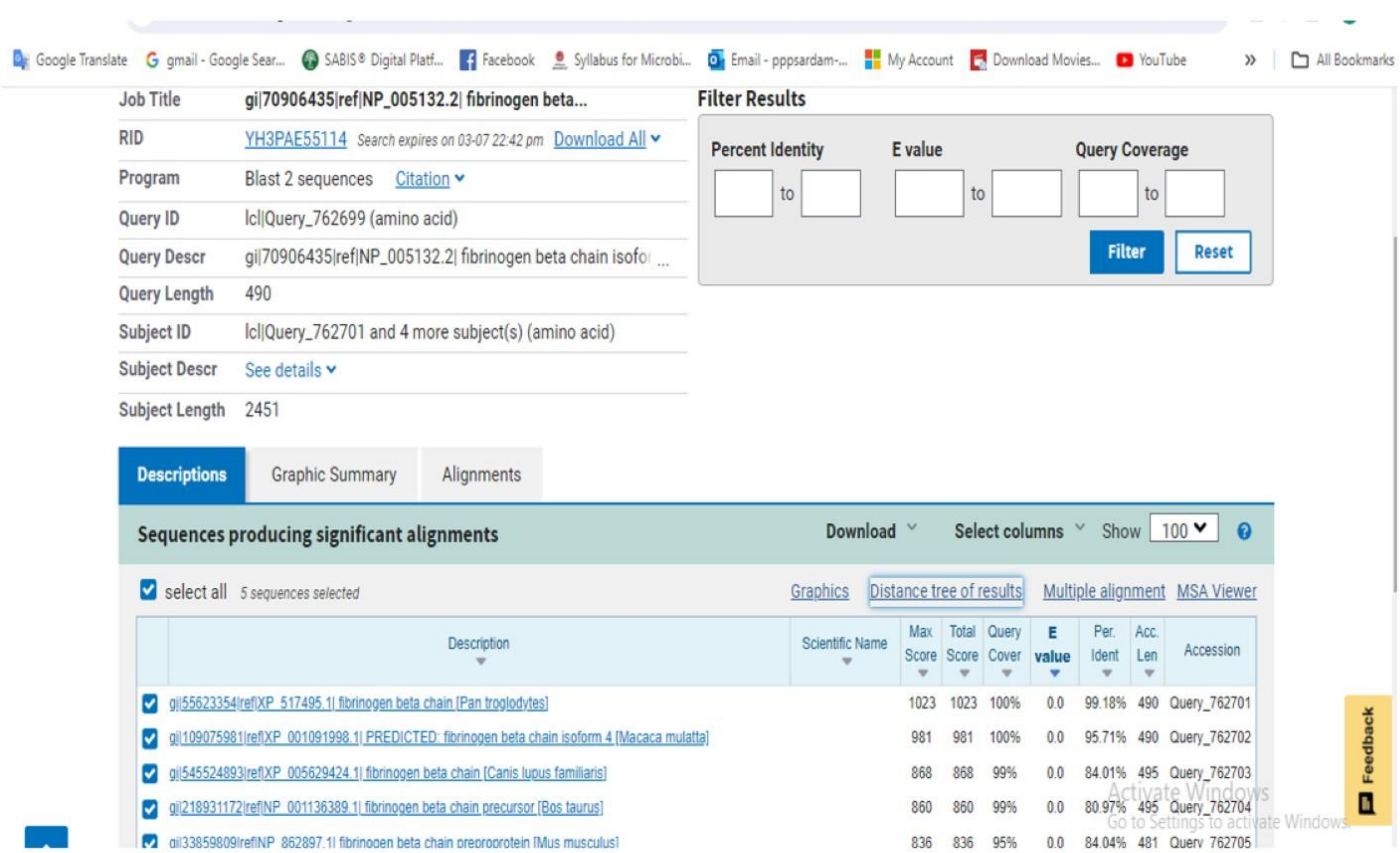
Enter accession number(s), gi(s), or FASTA sequence(s)

```
VKPYRVYCDMNTENGGWT
VIQNRQDGSVDFGRKWDPYKQGFGNVATNTDGKNYCGLPGEYWLGNDKI
SQLTRMGPTELLIEMEDWKGD
KVKAHYGGFTVQNEANKYQISVNKYRGTAGNALMDGASQLMGENRTMTIH
NGMFFSTYDRDNDGWLTSDP
RKQCSKEDGGGWWYNRCHAANPNGRYYWGGQYTWDMAKHGTDDGVV
WMNWKGSWYSMKKMSMKIRPFFPQ
```

Subject subrange

From

To

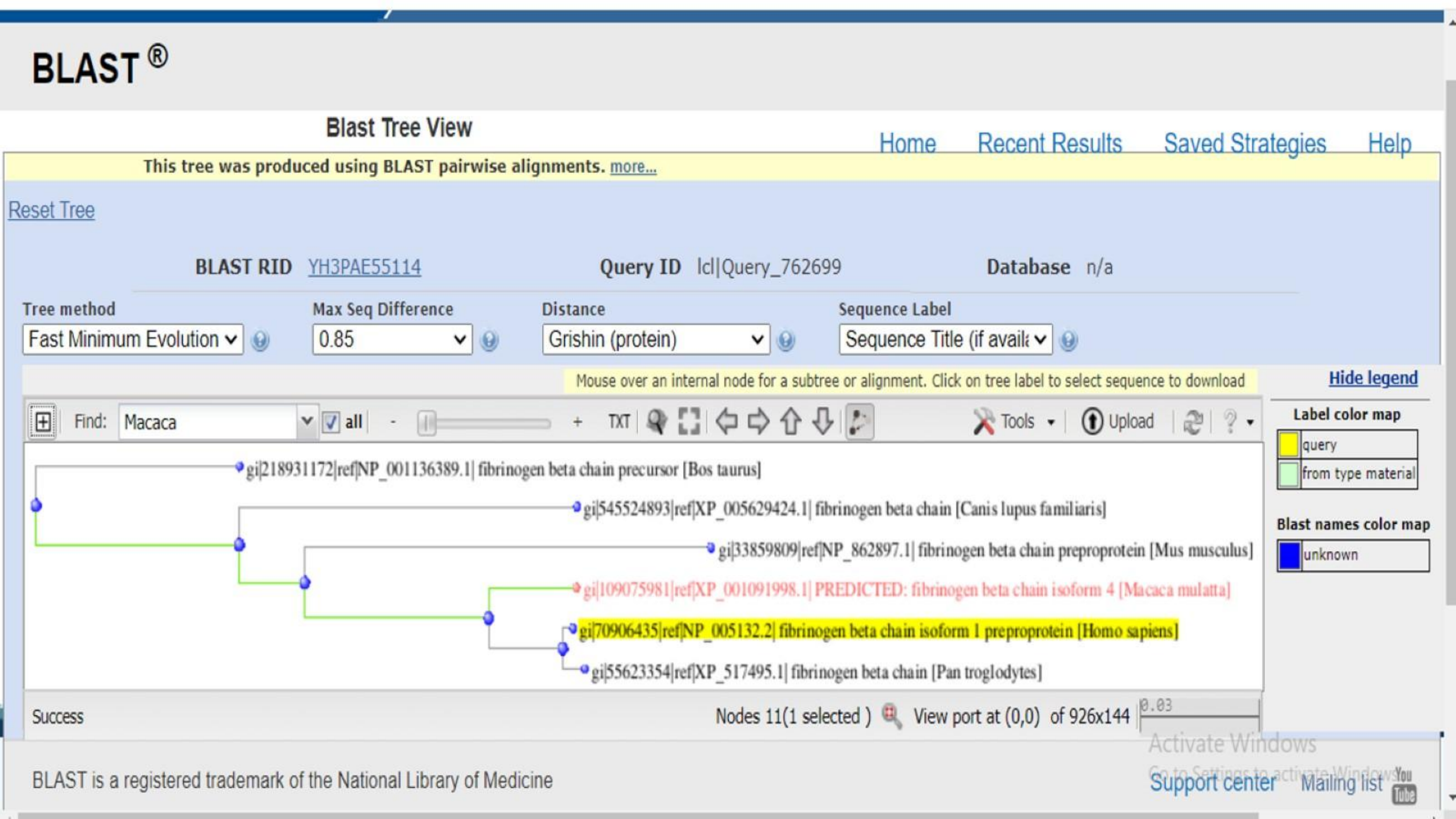


Job Title	gi|70906435|ref|NP_005132.2| fibrinogen beta...
RID	YH3PAE55114 *Search expires on 03-07 22:42 pm* Download All
Program	Blast 2 sequences Citation
Query ID	lcl|Query_762699 (amino acid)
Query Descr	gi|70906435|ref|NP_005132.2| fibrinogen beta chain isofor ...
Query Length	490
Subject ID	lcl|Query_762701 and 4 more subject(s) (amino acid)
Subject Descr	See details
Subject Length	2451

Sequences producing significant alignments

☑ select all *5 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	gi|55623354|ref|XP_517495.1| fibrinogen beta chain [Pan troglodytes]		1023	1023	100%	0.0	99.18%	490	Query_762701
☑	gi|109075981|ref|XP_001091998.1| PREDICTED: fibrinogen beta chain isoform 4 [Macaca mulatta]		981	981	100%	0.0	95.71%	490	Query_762702
☑	gi|545524893|ref|XP_005629424.1| fibrinogen beta chain [Canis lupus familiaris]		868	868	99%	0.0	84.01%	495	Query_762703
☑	gi|218931172|ref|NP_001136389.1| fibrinogen beta chain precursor [Bos taurus]		860	860	99%	0.0	80.97%	495	Query_762704
☑	gi|33859809|ref|NP_862897.1| fibrinogen beta chain preproprotein [Mus musculus]		836	836	95%	0.0	84.04%	481	Query_762705



BLAST ®

Blast Tree View

This tree was produced using BLAST pairwise alignments. more...

BLAST RID YH3PAE55114 Query ID lcl|Query_762699 Database n/a

Tree method: Fast Minimum Evolution

Max Seq Difference: 0.85

Distance: Grishin (protein)

Sequence Label: Sequence Title (if avail:

Mouse over an internal node for a subtree or alignment. Click on tree label to select sequence to download

Find: Macaca

gi|218931172|ref|NP_001136389.1| fibrinogen beta chain precursor [Bos taurus]

gi|545524893|ref|XP_005629424.1| fibrinogen beta chain [Canis lupus familiaris]

gi|33859809|ref|NP_862897.1| fibrinogen beta chain preproprotein [Mus musculus]

gi|109075981|ref|XP_001091998.1| PREDICTED: fibrinogen beta chain isoform 4 [Macaca mulatta]

gi|70906435|ref|NP_005132.2| fibrinogen beta chain isoform 1 preproprotein [Homo sapiens]

gi|55623354|ref|XP_517495.1| fibrinogen beta chain [Pan troglodytes]

Label color map

🟨	query
🟩	from type material

Blast names color map

🟦	unknown

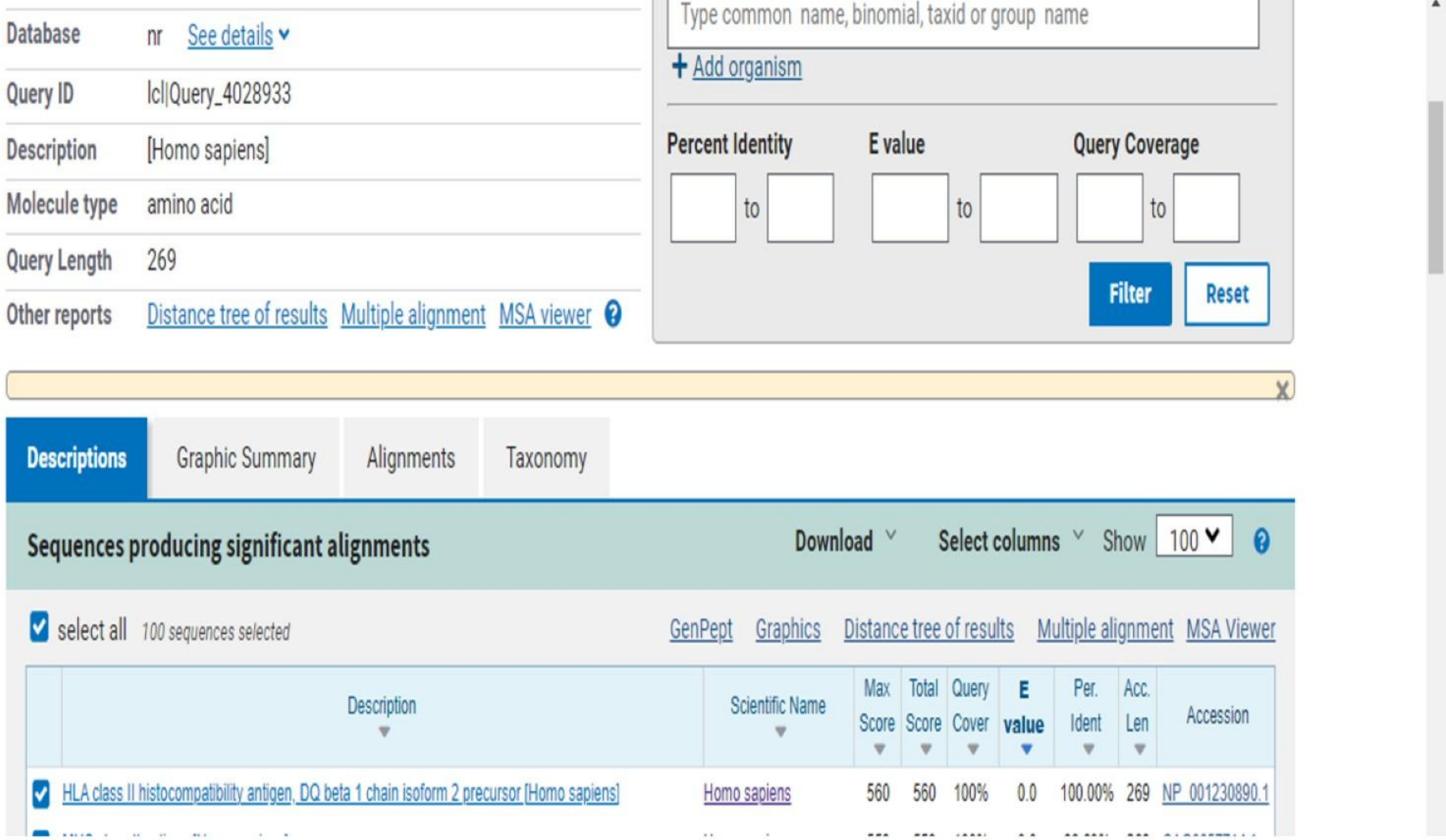
Success Nodes 11(1 selected) View port at (0,0) of 926x144 0.03

BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list

Task 2:

- Use the file provided in Moodle to do the following task:
- Use the first sequence in the file as your query and perform multiple alignment with the other sequences in the same file.
- To which gene does the query sequence belongs?
- **Note:** edit the description line for each organism , add the informal names for them
- Describe the distance tree in general, describe your results in term of finding the out group, the main clades , which one evolved first, and determine the sister taxa.
- Save the tree using circular and radial format
- Save your tree using rectangular format with midpoint root option.



Database	nr See details
Query ID	lcl|Query_4028933
Description	[Homo sapiens]
Molecule type	amino acid
Query Length	269
Other reports	Distance tree of results Multiple alignment MSA viewer

Type common name, binomial, taxid or group name

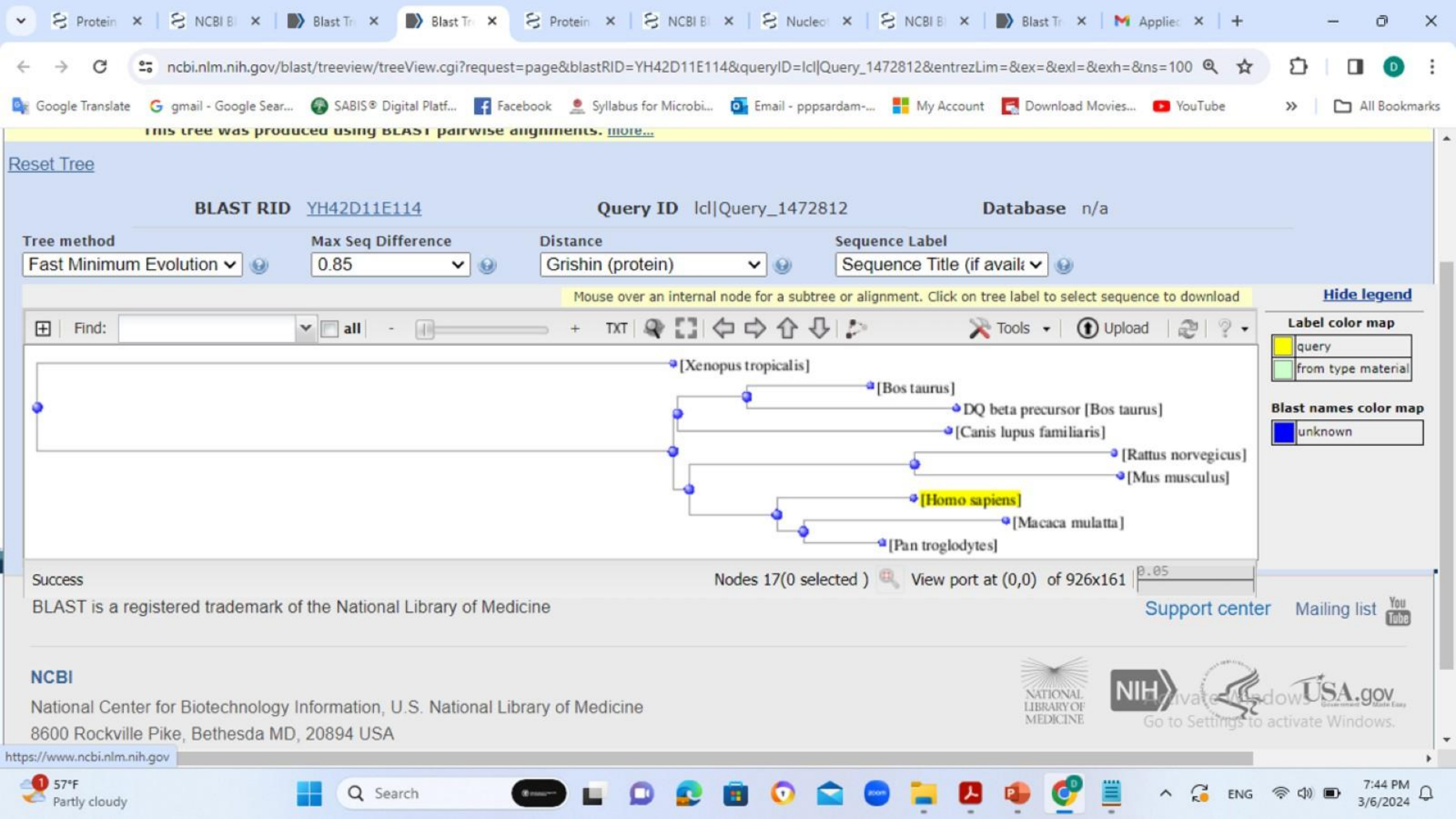
+ Add organism

Percent Identity			E value			Query Coverage		
	to			to			to	

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	HLA class II histocompatibility antigen, DQ beta 1 chain isoform 2 precursor [Homo sapiens]	Homo sapiens	560	560	100%	0.0	100.00%	269	NP_001230890.1



This tree was produced using BLAST pairwise alignments. more...

Reset Tree

BLAST RID YH42D11E114 **Query ID** lcl|Query_1472812 **Database** n/a

Tree method Fast Minimum Evolution

Max Seq Difference 0.85

Distance Grishin (protein)

Sequence Label Sequence Title (if avail

Mouse over an internal node for a subtree or alignment. Click on tree label to select sequence to download

Hide legend

Label color map

🟨	query
🟩	from type material

Blast names color map

🟦	unknown

Success Nodes 17(0 selected) View port at (0,0) of 926x161

BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

NATIONAL LIBRARY OF MEDICINE
NIH
USA.gov
Activate Windows
Go to Settings to activate Windows.
https://www.ncbi.nlm.nih.gov
57°F
Partly cloudy
Search
ENG
7:44 PM
3/6/2024