Phylogenetic Trees - part II

Mark Stenglein, MIP 280A4

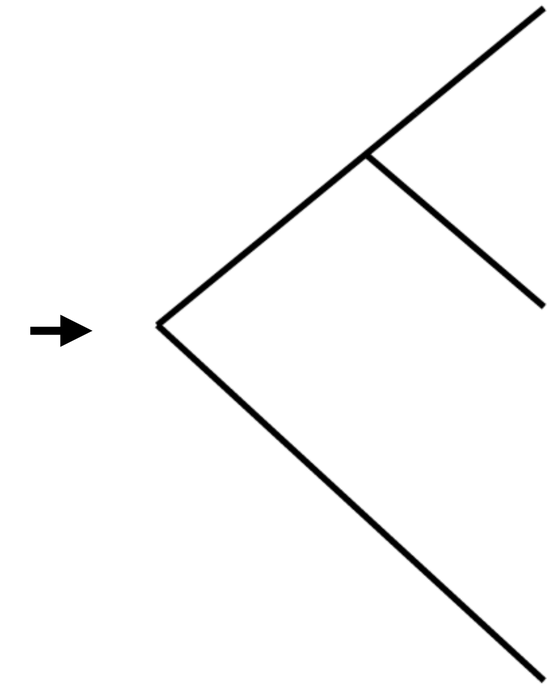
Last class: how do you interpret phylogenetic trees?

Today:

- How do you make a tree?
- How do you root a tree? What's a root good for?
- How would you know if your tree is any good?

Multiple sequence alignments are the input to tree-building software

```
GAGGTCACACGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
GAGGTCACGCGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
GAGGTCACACGCATGGTCGTCATCATGGTCATCGCATTCCTGATTTGCTGGTTGCCCT
```



ring-tailed cat

raccoon

kinkajou

Multiple sequence alignments can be represented in fasta format

```
>Kamiti_River_virus_NP_937780
---LEKSTAVGLGIKWKTLLNSLDKDTFDKYKVRGVNETER----------GDYVSRGGLKLDEVIRKYNWKPNGRVVDLGCGRGGWSQRVVMEESVSSVYGYTVG-GQEK
>Cell_fusing_agent_virus_YP_009259301
--ALEKSTTIGLGMKWKMTLNALDGDAFTKYKSRGVNETER----------GDYVSRGGLKLNEIISKYEWRPSGRVVDLGCGRGGWSQRAVMEETVSSALGFTIG-GAEK
>Aedes_flavivirus_YP_009259331
FRALEKSTTIGLGIKWKFMLNALDKDAFERYKVRGVNETGK----------GDYVSRGGLKLDEIIRKYQWTPMGHVVDLGCGRGGWSQRVVMEETVTSVHGFTIG-GNDK
>Mercadeo_virus_YP_009259288
--SINKAPVGGMGYRWKRELNKKSLSEFNDYRSRGVNETER----------GDYVSRGGLKMDELITKFGWEPKGRVVDLGCGRGGWAQRLVADRRVTKVNAYTLG-GAEF
>Quang_Binh_virus_YP_009259365
--SLVKTDTCGIGYRWKEILNSLDKNAFDQYRSRGVNETDK----------GDYVSRGGLKMDELLRKYQWEPKGVAVDLGCGRGGWSQRLVMDSRISAVHGFTLG-GNNF
>Hanko_virus_YP_009268628
--SLEKSATGGLGHRWKKILNAMSQDDFNSYRLYGVDETDK----------GDYVSRGGLKLRELTLKYGWKPEGICVDLGCGRGGWSQHLAMDPRVTRIEAYTLG-GSTF
>Montana_myotis_leukoencephalitis_virus_NP_775653
---GLSLSHLTLGEDWKLKLNKMTKSDFLEYRTRLITEVDRGEAVYQLGRGKTNTGHAVSRGTSKLAWMHERSLVRLEGCVVDLGCGRGGWSYYSAAQNPVRKVDAYTLG-YGGH
>Rio_Bravo_virus_NP_776080
---GIVTSHHTLGEKWKLELNQLSHKEFLSYRKVGILEVDRPDAVMNLNKGKTNTGHAVSRGTSKLAWMHERGFIPLSGHVVDLGCGRGGWSYYCAAQTPVRKVNAYTLG-TGAH
>Modoc_virus_NP_740267
---GICSSAPTLGEIWKRKLNQLDAKEFMAYRRRFVVEVDRNEAREALAKGKTNTGHAVSRGTAKLAWIDERGGVELKGSVVDLGCGRGGWSYYAASQPNVREVKAYTLG-TSGH
>Apoi_virus_NP_775688
---GVSSSYITYGEQWKRELNKLNAQAFFLYKSRLVHEIDRAEAVSNLSKGRTNTGHAVSRGTSKLAWMHERGYVPLKGVVVDLGSGRGGWSYYAAAQERVRKVNAYTLATTKGH
>Dengue_virus_1_NP_722465
---GTGAQGETLGEKWKRQLNQLSKSEFNTYKRSGIIEVDRSEAKEGLKRGETT-KHAVSRGTAKLRWFVERNLVKPEGKVIDLGCGRGGWSYYCAGLKKVTEVKGYTKG-GPGH
>Dengue_virus_3_YP_001531176
---GTGSQGETLGEKWKKKLNQLSRKEFDLYKKSGITEVDRTEAKEGLKRGETT-HHAVSRGSAKLQWFVERNMVVPEGRVIDLGCGRGGWSYYCAGLKKVTEVRGYTKG-GPGH
>Dengue_virus_2_NP_739590
---GTGNIGETLGEKWKSRLNALGKSEFQIYKKSGIQEVDRTLAKEGIKRGETD-HHAVSRGSAKLRWFVERNMVTPEGKVVDLGCGRGGWSYYCGGLKNVREVKGLTKG-GPGH
>Dengue_virus_4_NP_740325
---GTGTTGETLGEKWKRQLNSLDRKEFEEYKRSGILEVDRTEAKSALKDGSKI-KHAVSRGSSKIRWIVERGMVKPKGKVVDLGCGRGGWSYYMATLKNVTEVKGYTKG-GPGH
>Zika_virus_YP_009227205
----GGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALKDGVATGGHAVSRGSAKIRWLEERGYLQPYGKVVDLGCGRGGWSYYAATIRKVQEVRGYTKG-GPGH
>Zika_virus_YP_009430308
----GGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALKDGVATGGHAVSRGSAKLRWLVERGYLQPYGKVIDLGCGRGGWSYYAATIRKVQEVKGYTKG-GPGH
>Spondweni_virus_YP_009227195
---RGGGNGETVGEKWKERLNRMTALEFYAYKRSGITEVCREPARRALKDGVVTGGHAVSRGSAKLRWMVERGHVNLVGRVVDLGCGRGGWSYYAASQKQVLEVRGYTKG-GAGH
>Murray_Valley_encephalitis_virus_NP_722539
----GRAGGRTLGEQWKEKLNAMGKEEFFSYRKEAILEVDRTEARRARREGNKVGGHPVSRGTAKLRWLVERRFVQPIGKVVDLGCGRGGWSYYAATMKNVQEVRGYTKG-GPGH
>Japanese_encephalitis_virus_NP_775674
----GRPGGRTLGEQWKEKLNAMSREEFFKYRREAIIEVDRTEARRARRENNIVGGHPVSRGSAKLRWLVEKGFVSPIGKVIDLGCGRGGWSYYAATLKKVQEVRGYTKG-GAGH
>Usutu_virus_YP_164818
----GRPGGRTLGEQWKEKLNGLSKEDFLKYRKEAITEVDRSAARKARRDGNKTGGHPVSRGSAKLRWMVERQFVKPIGKVVDLGCGRGGWSYYAATLKGVQEVRGYTKG-GPGH
>West_Nile_virus_NP_776022
----GGAKGRTLGEVWKERLNHMTKEEFTRYRKEAITEVDRSAAKHARREGNITGGHPVSRGTAKLRWLVERRFLEPVGKVVDLGCGRGGWCYYMATQKRVQEVKGYTKG-GPGH
```

Note gaps in the sequences

Some tree building software requires alignments in weird custom formats
For example Mr Bayes, a program for Bayesian inference of trees

2.2.1 Getting Data into MrBayes

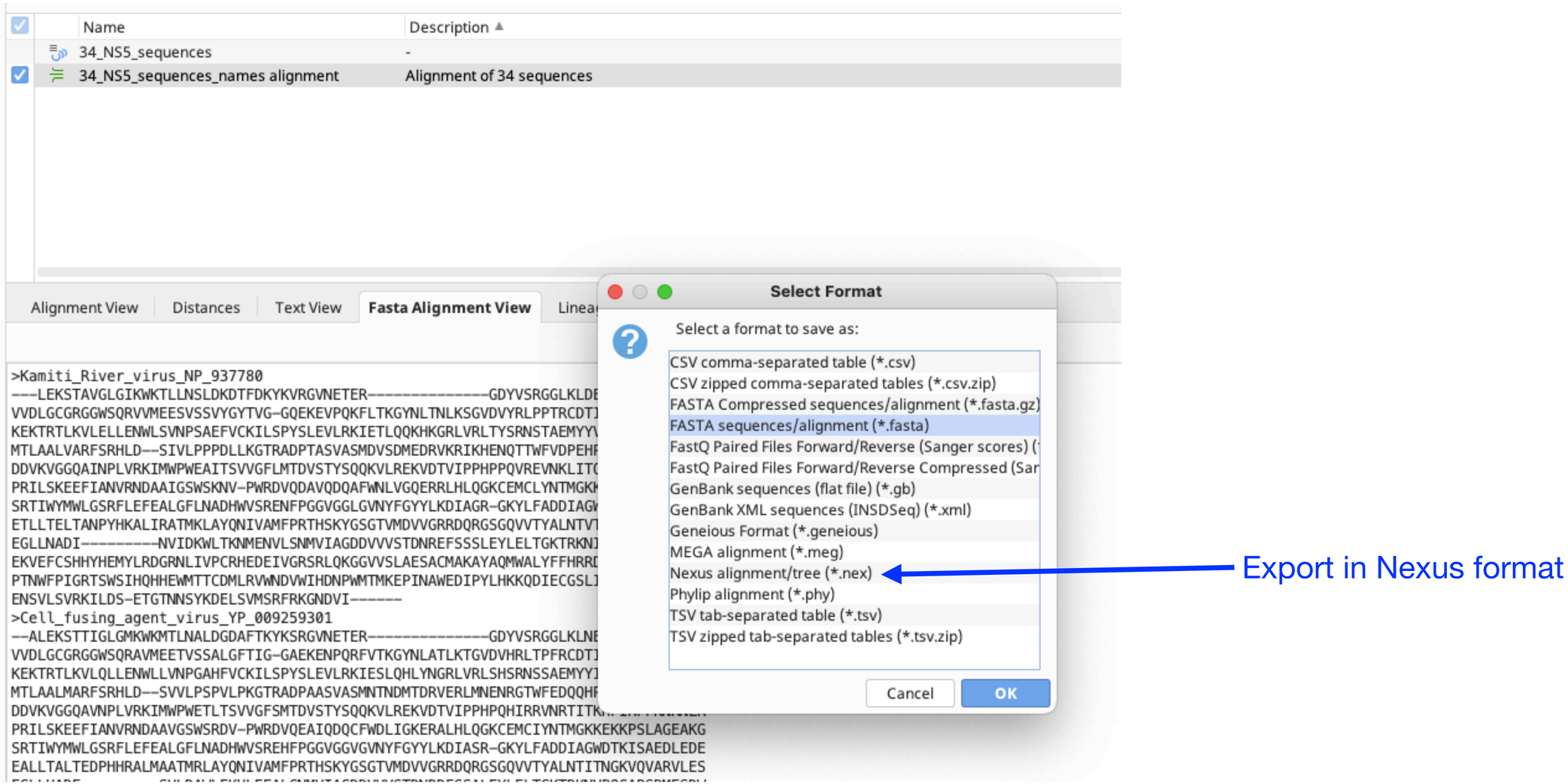
To get data into MrBayes, you need a so-called Nexus file that contains aligned nucleotide or amino acid sequences, morphological ("standard") data, restriction site (binary) data, or any mix of these four data types. The Nexus data file is often generated by another program, such as Mesquite (Maddison and Maddison, 2006). Note, however, that MrBayes version 3 does not support the full Nexus standard, so you may have to do a little editing of the file for MrBayes to process it properly. In particular, MrBayes uses a fixed set of symbols [1] for each data type and does not support user-defined symbols.

In addition to the standard one-letter ambiguity symbols for DNA and RNA listed above, ambiguity can also be expressed using the Nexus parenthesis notation.

This is the kind of stuff that makes some people hate bioinformatics!

Ronquist et al (2020)

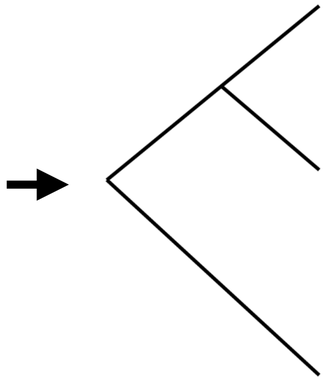
If you do have to inter-convert between formats, try exporting from Geneious



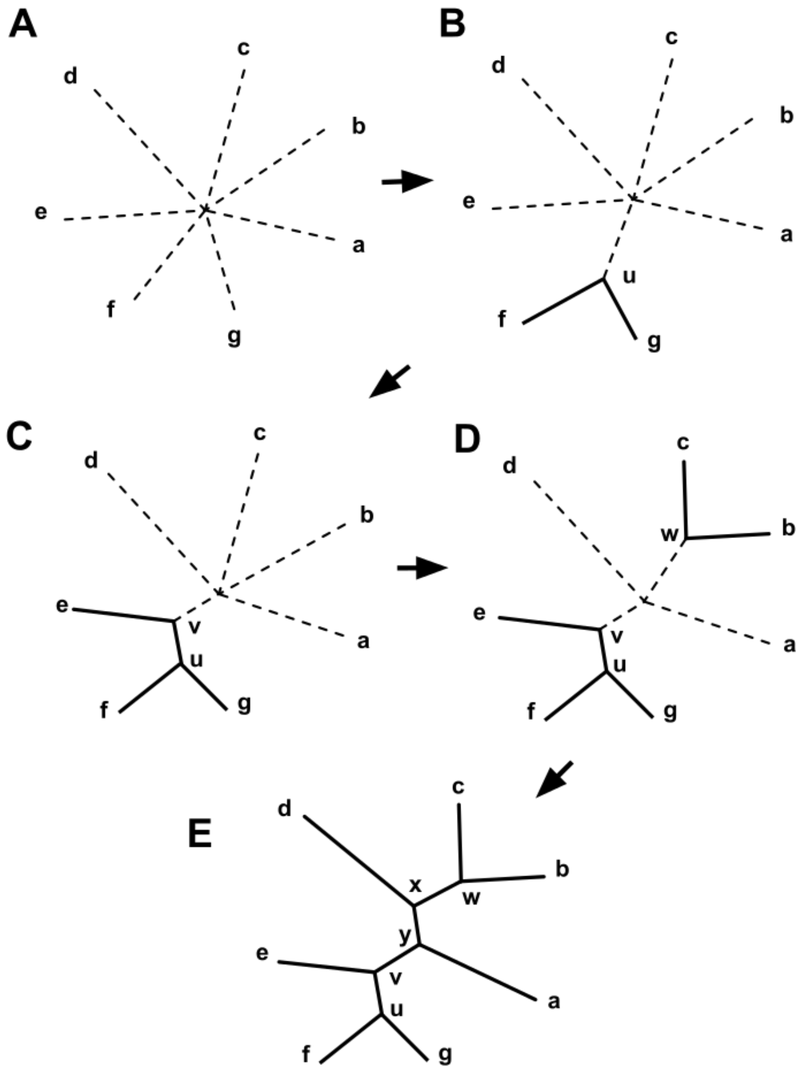
Tree building option 1: quick and dirty NJ tree

Multiple sequence alignment

Neighbor joining tree



Neighbor-joining tree-building algorithm



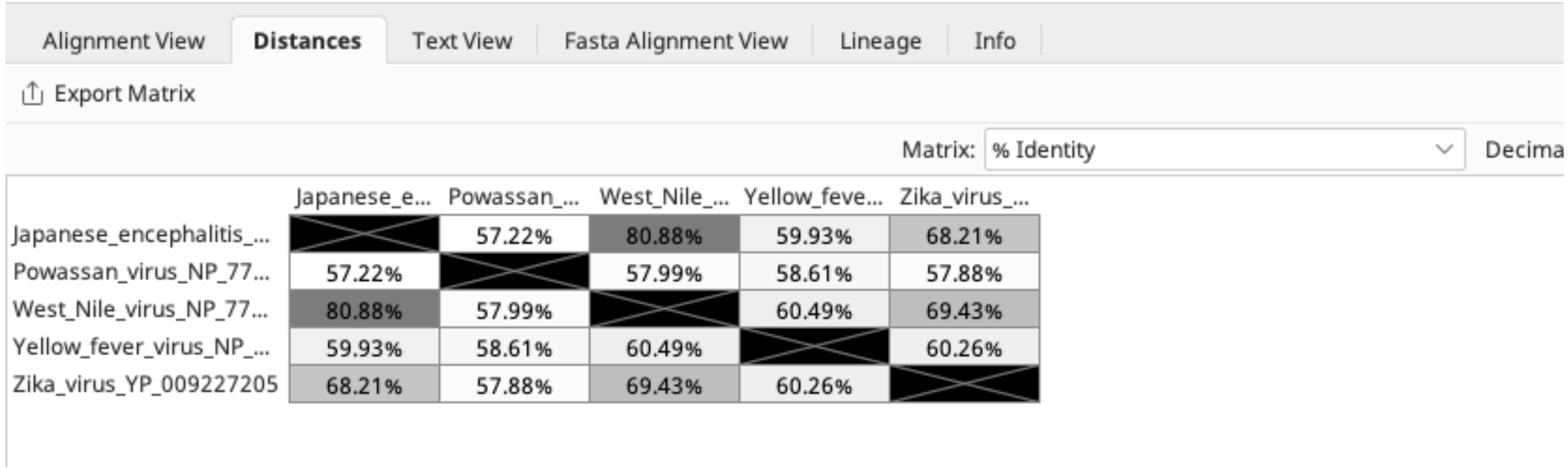
1. Create multiple sequence alignment of homologous sequences. This process will calculate distances between sequences.
2. Arrange all the taxa on a star-shaped tree (A)
3. Find the 2 most closely related taxa (in this example: f & g)
4. Join these to create a new node (u in this example, B)
5. Calculate the distances between all remaining taxa plus this new node
6. Repeat steps 2-4 until done.

1. Advantages: fast, often correct

2. Disadvantages: no measure of confidence (without bootstrapping). May not work as well as fancier methods.

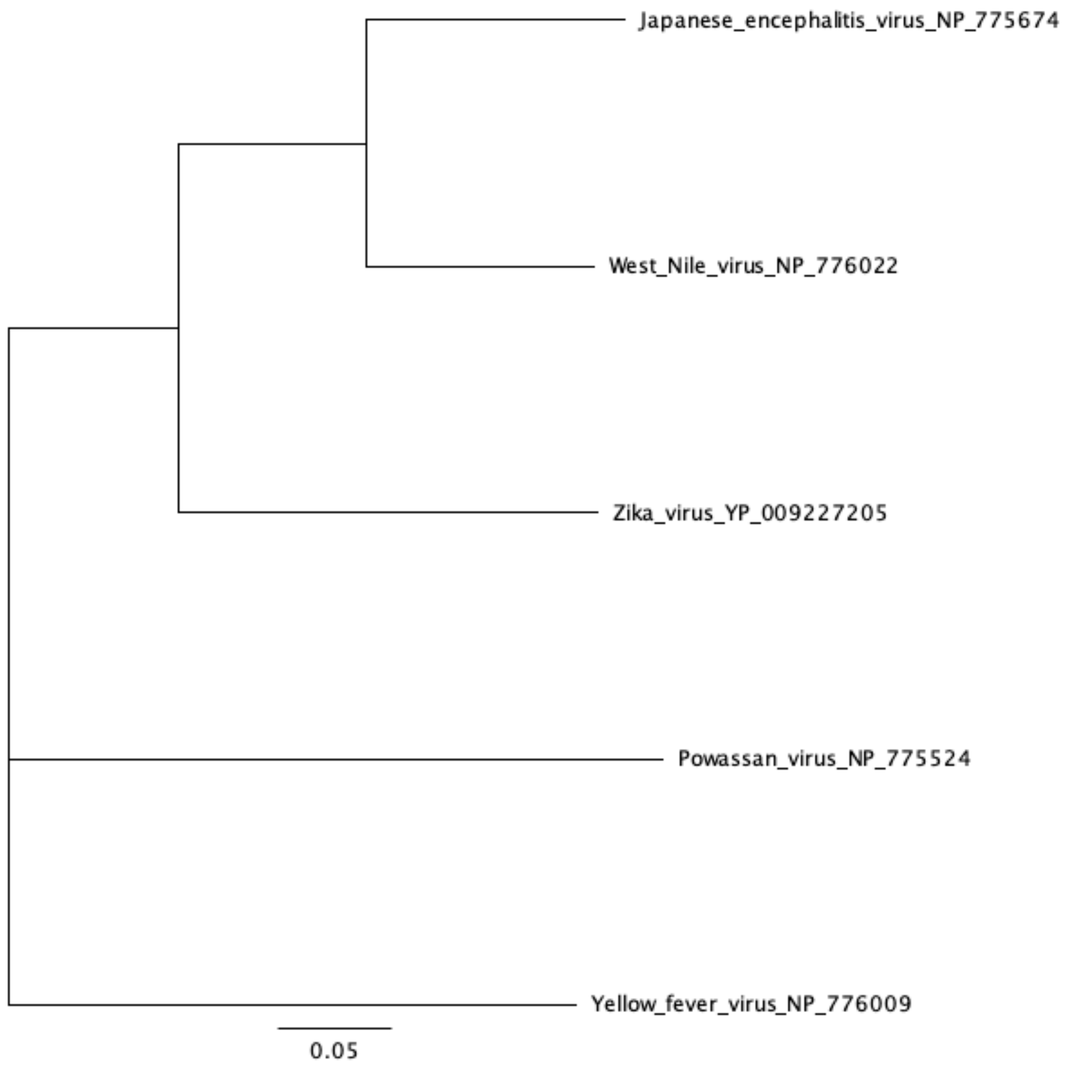
Which of these viruses would be joined first in a NJ tree-building algorithm?

MAFFT Alignment of 5 flavivirus NS5 sequences



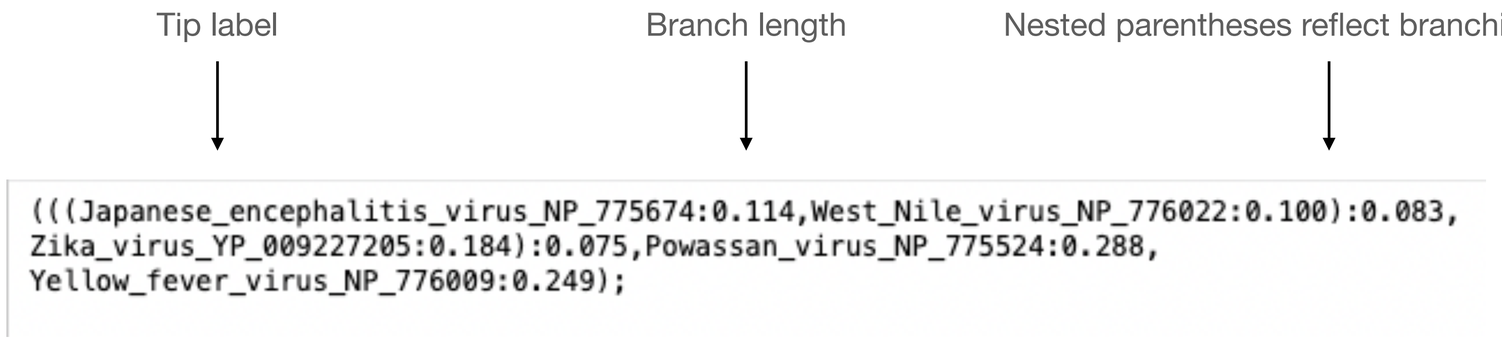
	Japanese_e...	Powassan_...	West_Nile_...	Yellow_feve...	Zika_virus_...
Japanese_encephalitis_...		57.22%	80.88%	59.93%	68.21%
Powassan_virus_NP_77...	57.22%		57.99%	58.61%	57.88%
West_Nile_virus_NP_77...	80.88%	57.99%		60.49%	69.43%
Yellow_fever_virus_NP_...	59.93%	58.61%	60.49%		60.26%
Zika_virus_YP_009227205	68.21%	57.88%	69.43%	60.26%	

NJ tree from this alignment of 5 NS5 sequences

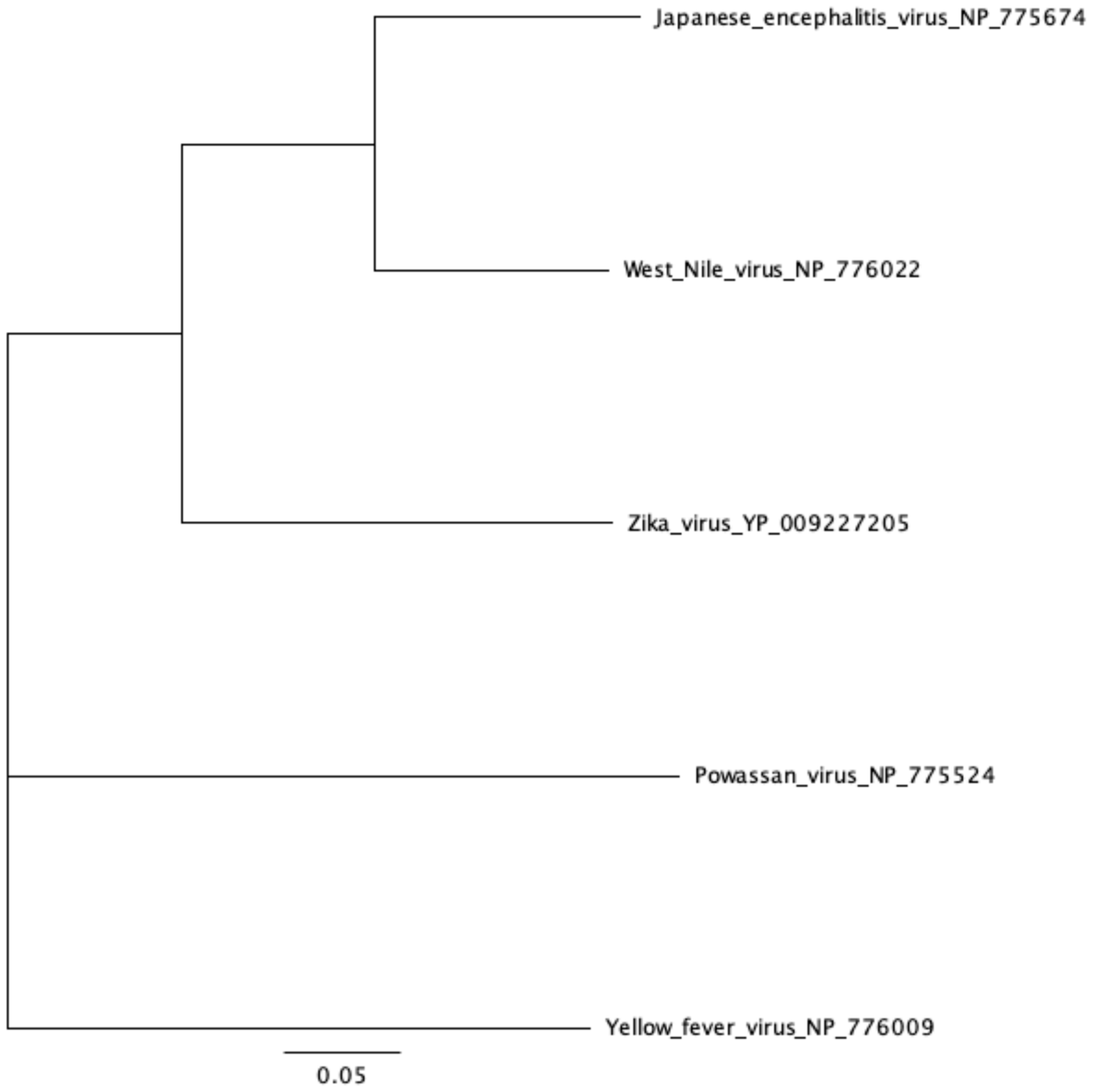


Scale bar represents (genetic) distance between taxa in units of substitutions per site

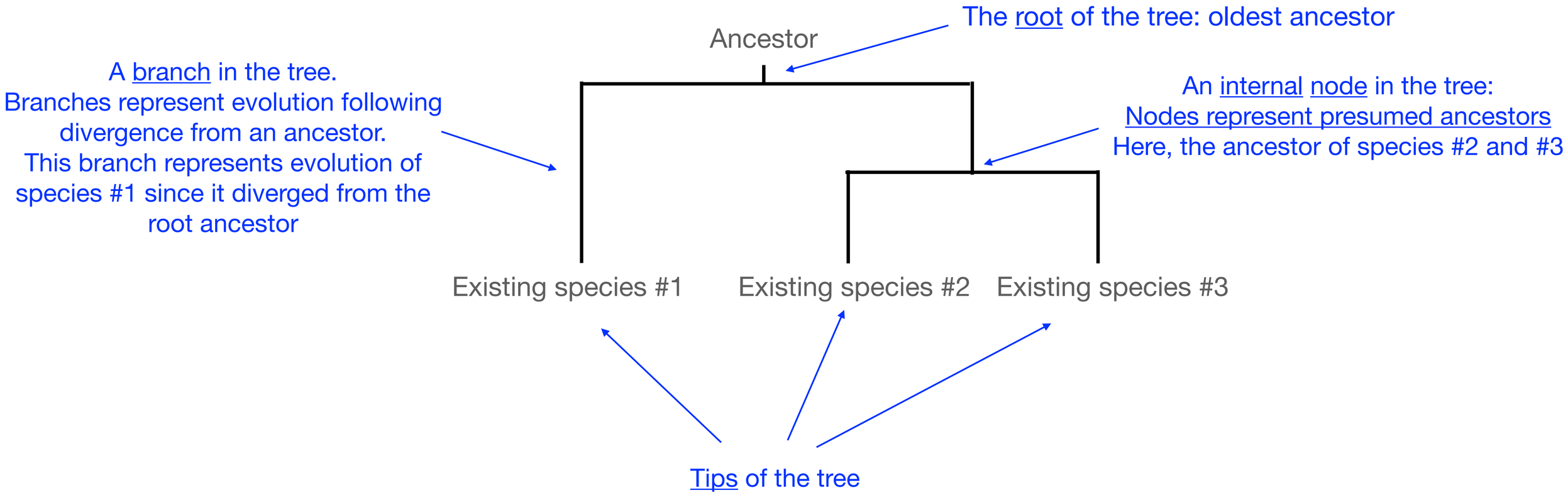
Trees are commonly represented in newick format



Tree building software usually produce unrooted trees (like this one)

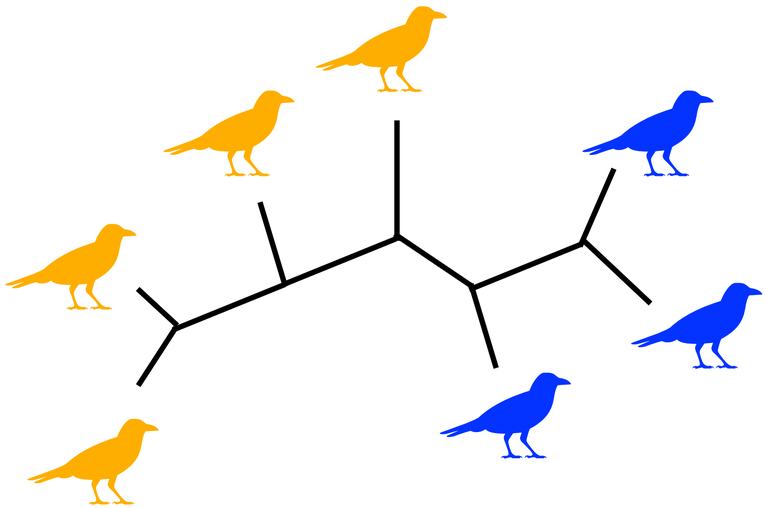


Parts of a phylogenetic tree



The root of the tree represents the common ancestor from which all the taxa in the tree derive. Rooted trees are generally more useful and easier to interpret than unrooted trees.

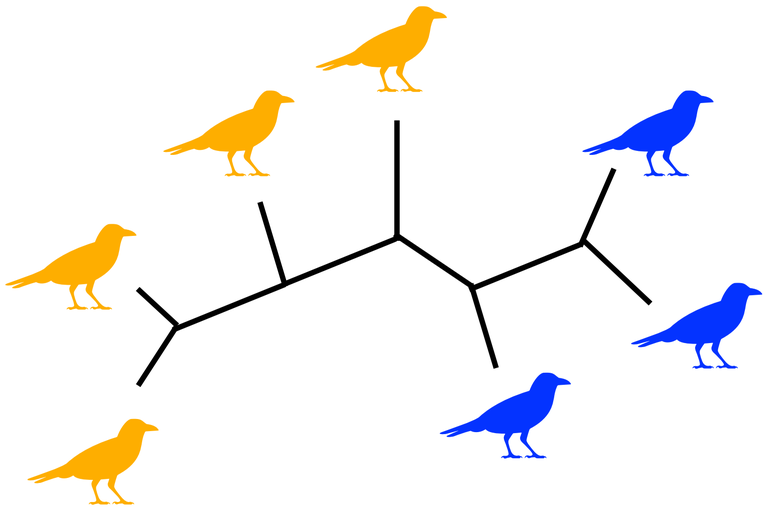
Unrooted tree



Was the ancestor of these birds blue or golden?

The root of the tree represents the common ancestor from which all the taxa in the tree derive. Rooted trees are generally more useful and easier to interpret than unrooted trees.

Unrooted tree



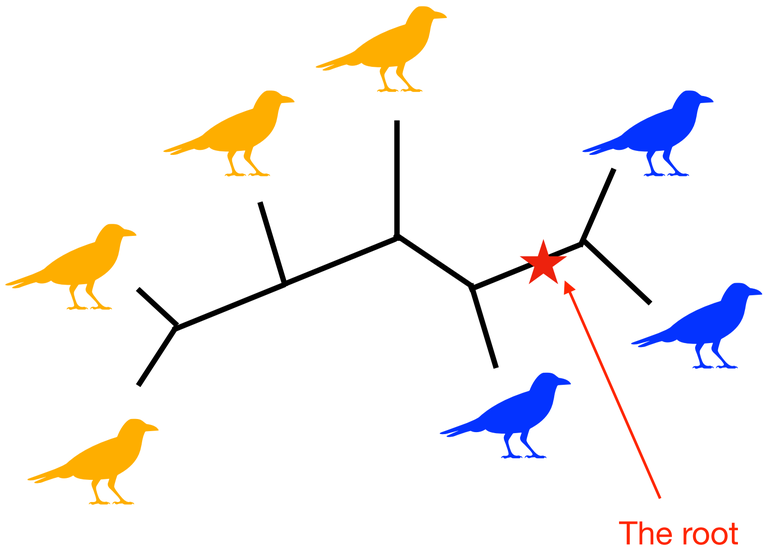
Was the ancestor of these birds blue or golden?

The problem is we don't know where to put the root of this tree

It could go anywhere! (The root could be placed on any branch)

The root of the tree represents the common ancestor from which all the taxa in the tree derive. Rooted trees are generally more useful and easier to interpret than unrooted trees.

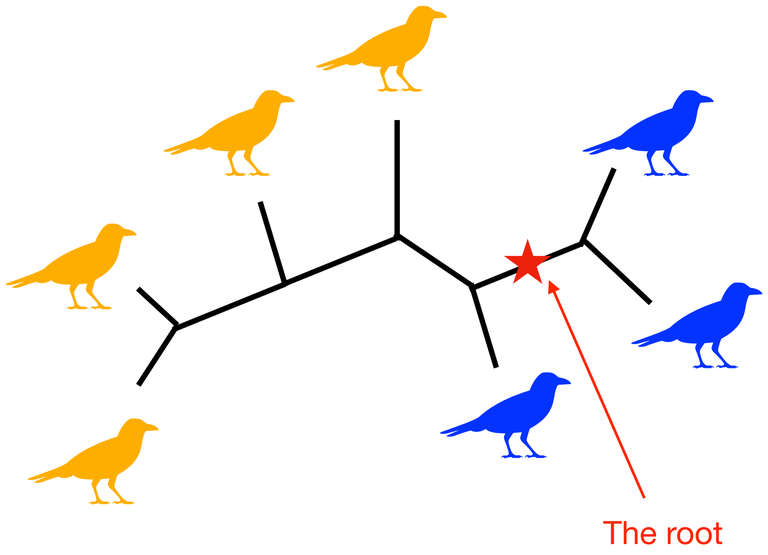
Unrooted tree

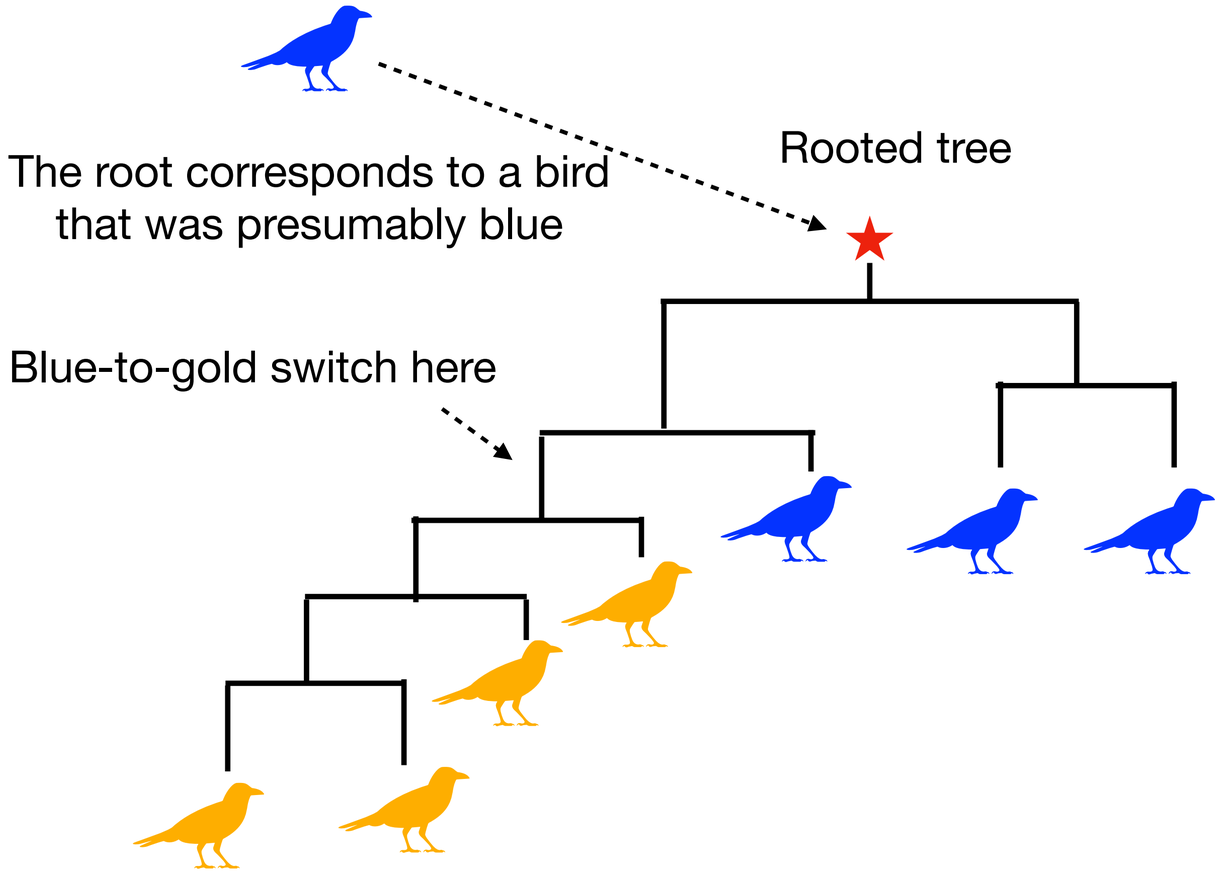


Was the ancestor of these birds blue or golden?

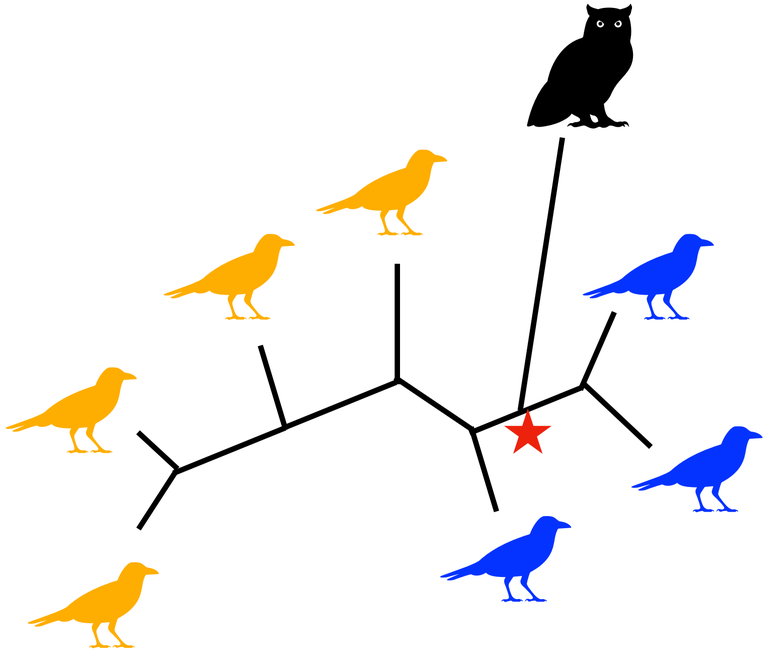
Rooted trees give you more information than unrooted trees.

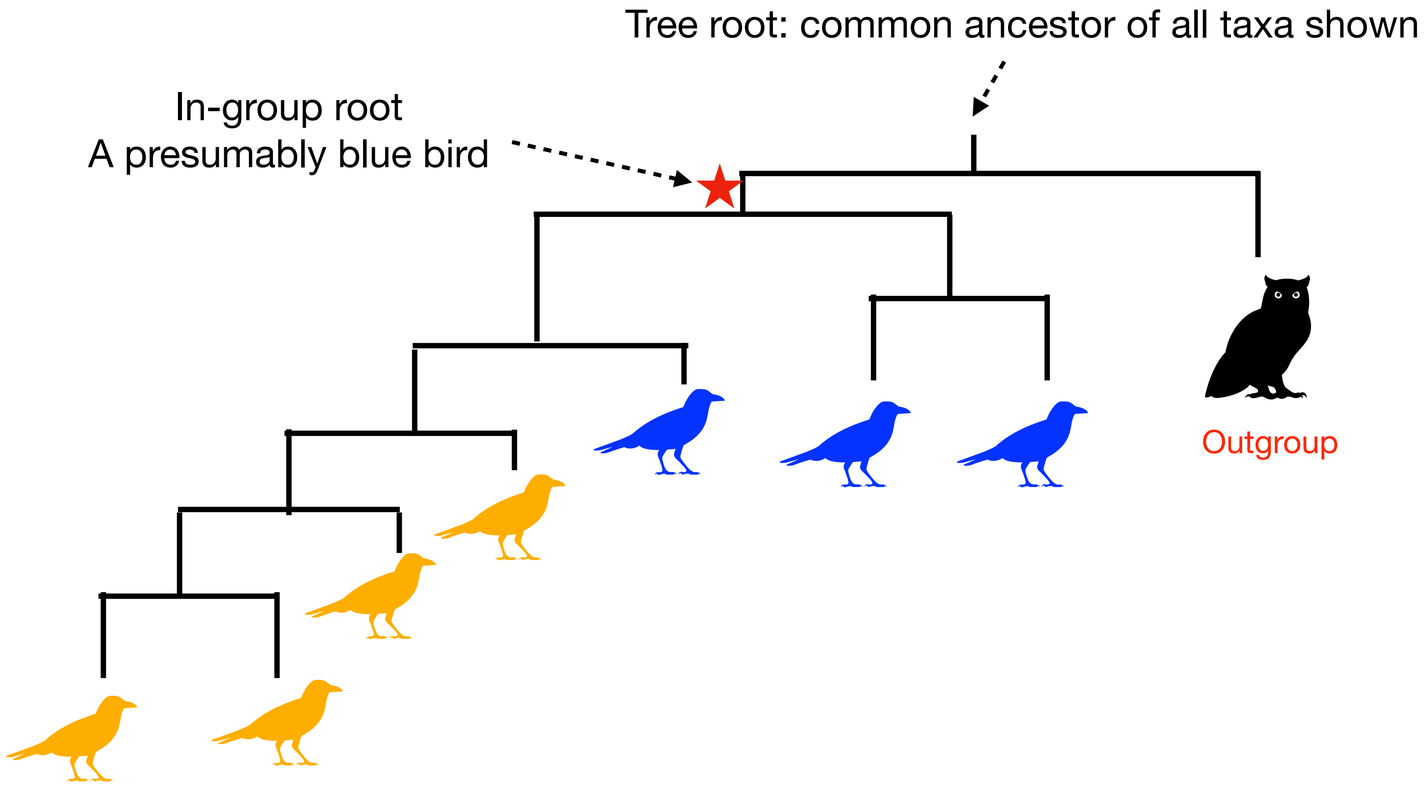
Unrooted tree



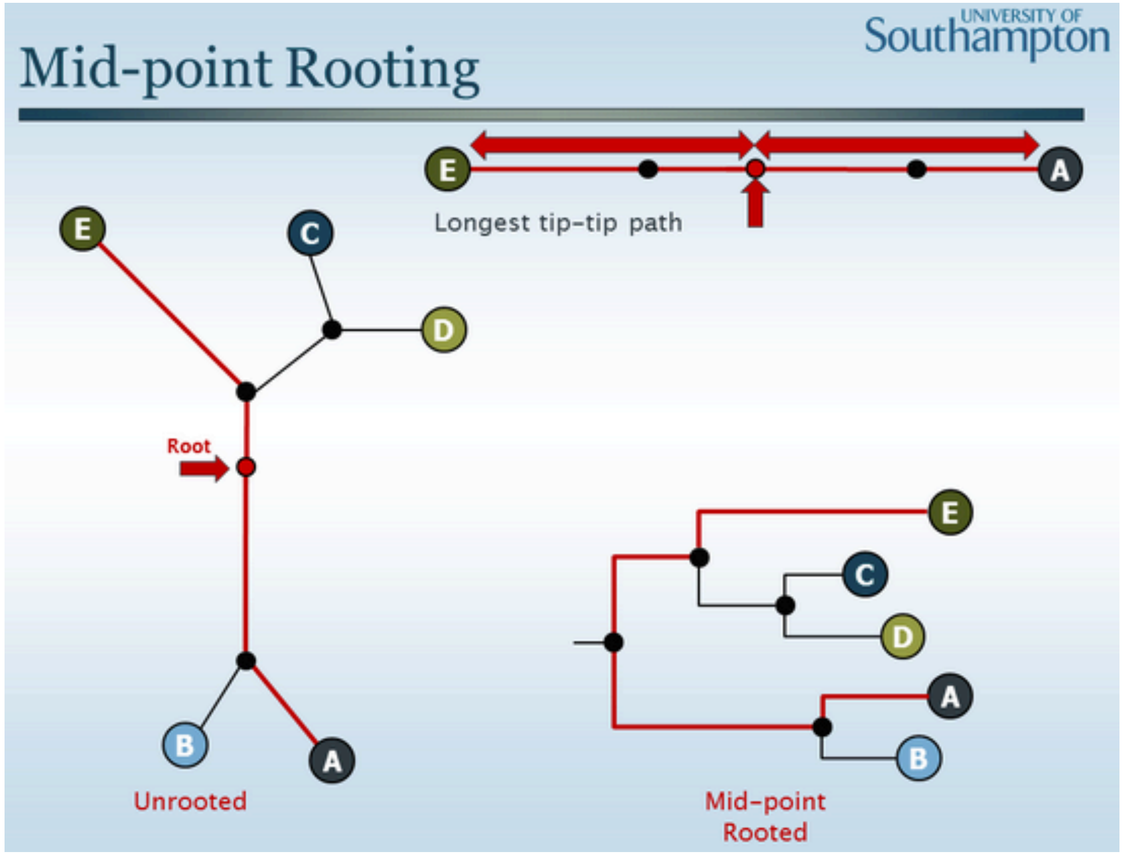


One way to root a tree is to use an outgroup.
To do this you have to know that the “in group” taxa share a more recent common ancestor than they do with the outgroup.





Another way to root a tree is to ‘midpoint root’ it by placing the root halfway between the two most distant tips

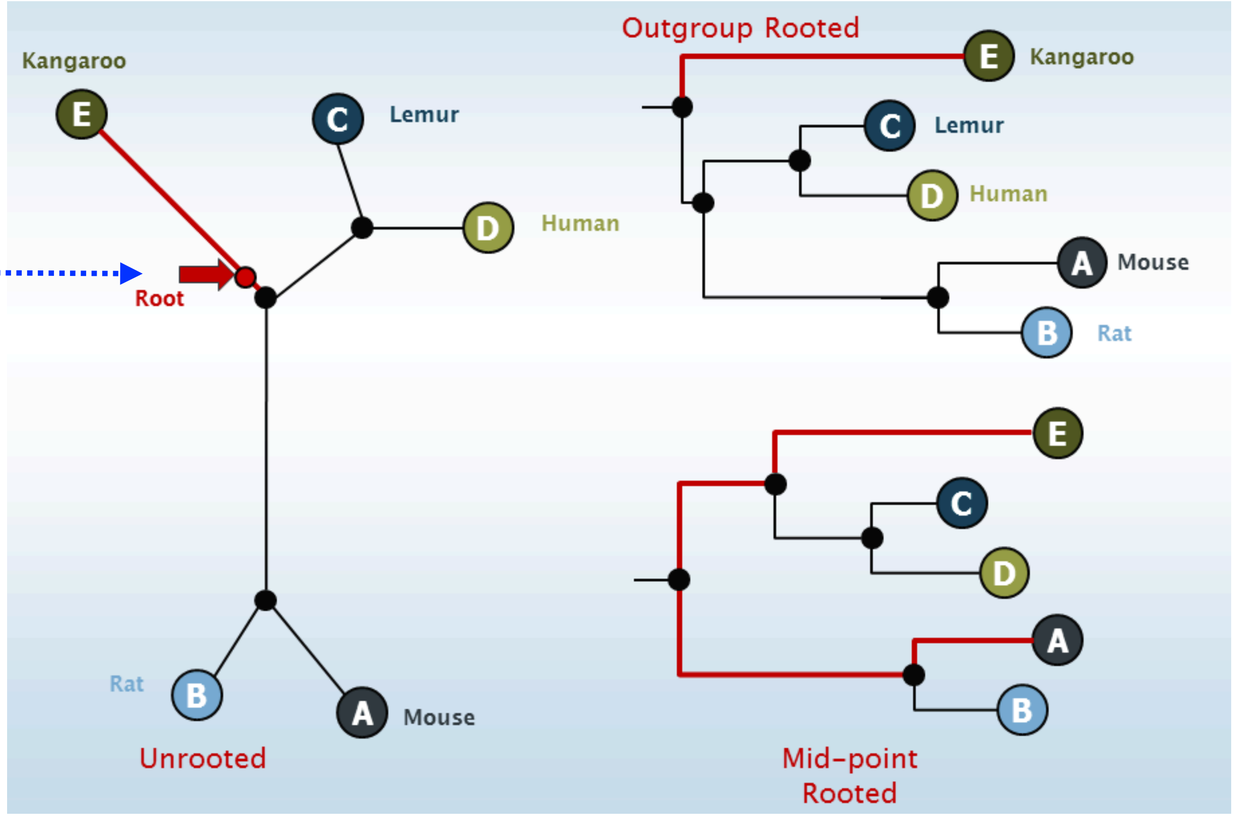


Richard Edwards

Midpoint rooting is less reliable than outgroup rooting

Actual root on branch to marsupials

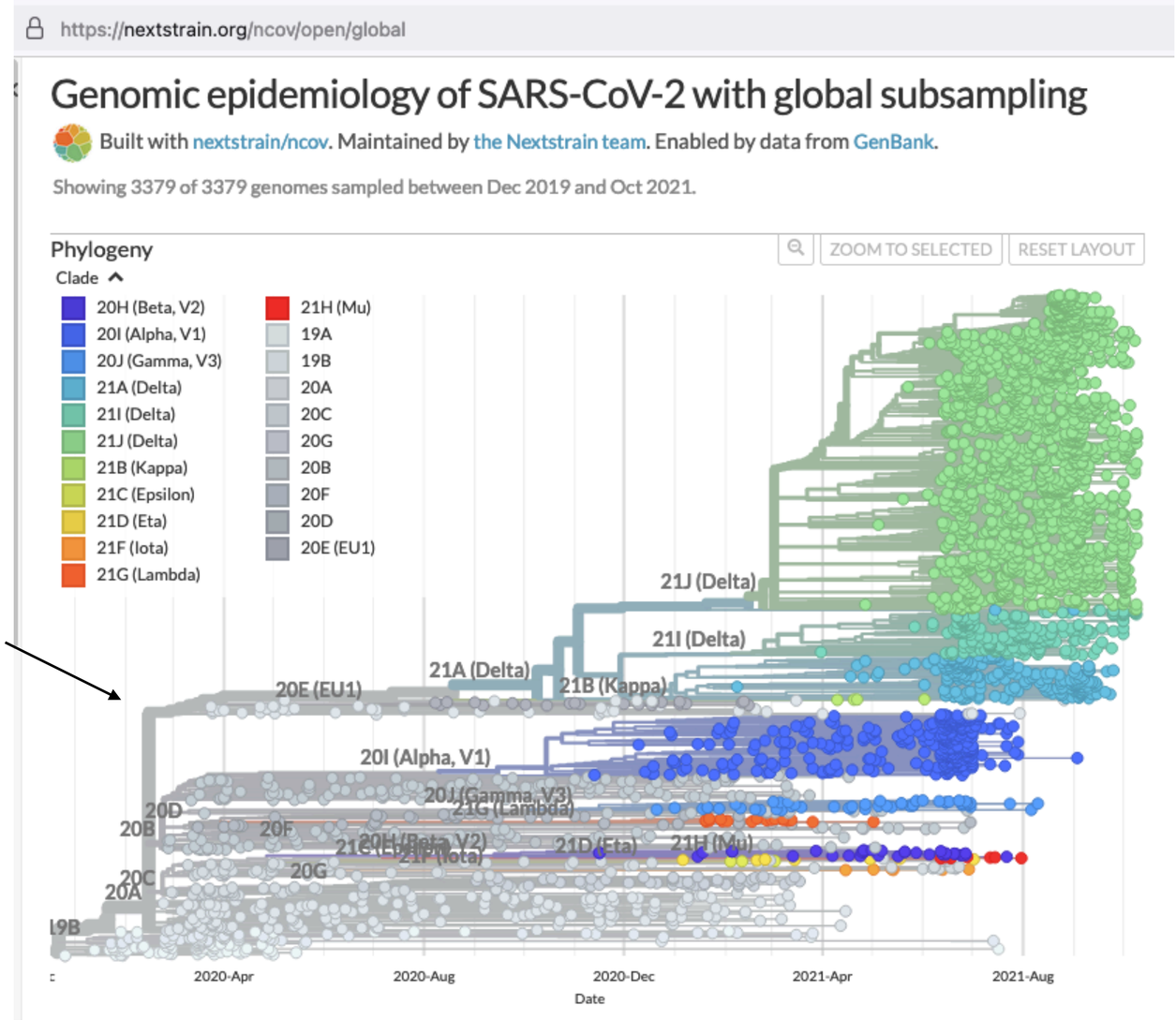
(Marsupials are an outgroup to other mammals)



This tree gives the false impression that kangaroos are more closely related to primates than primates are to rodents

Richard Edwards

You can also root using time-calibrated trees



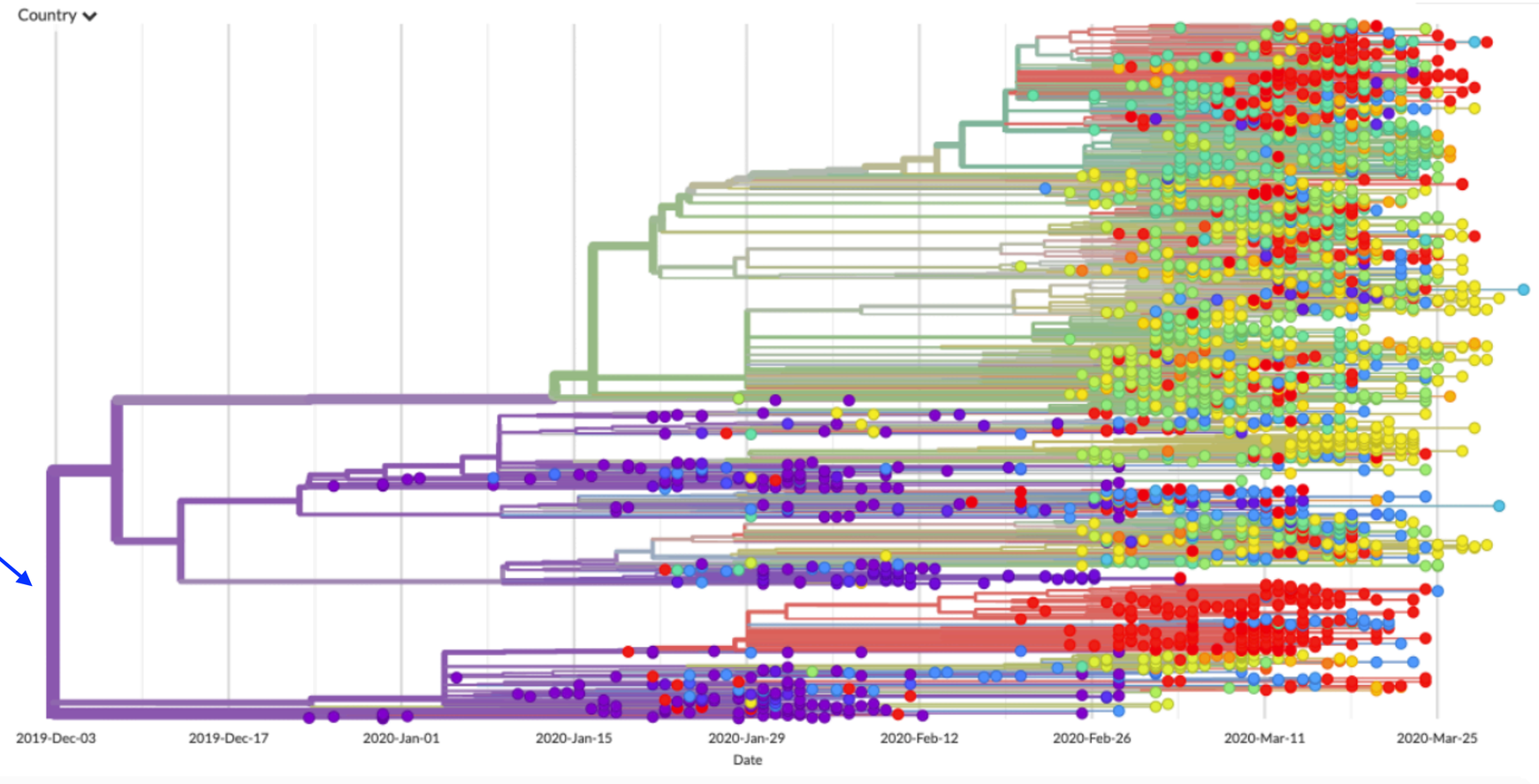
The date of internal nodes (presumed ancestors) are positioned based on an assumed rate of evolution

The tips are positioned on the x axis (time) according to their actual sampling date

https://nextstrain.org/ncov/open/global

You can also root using time-calibrated phylogenies

A nextstrain SARS-CoV-2 tree from Mar 2020

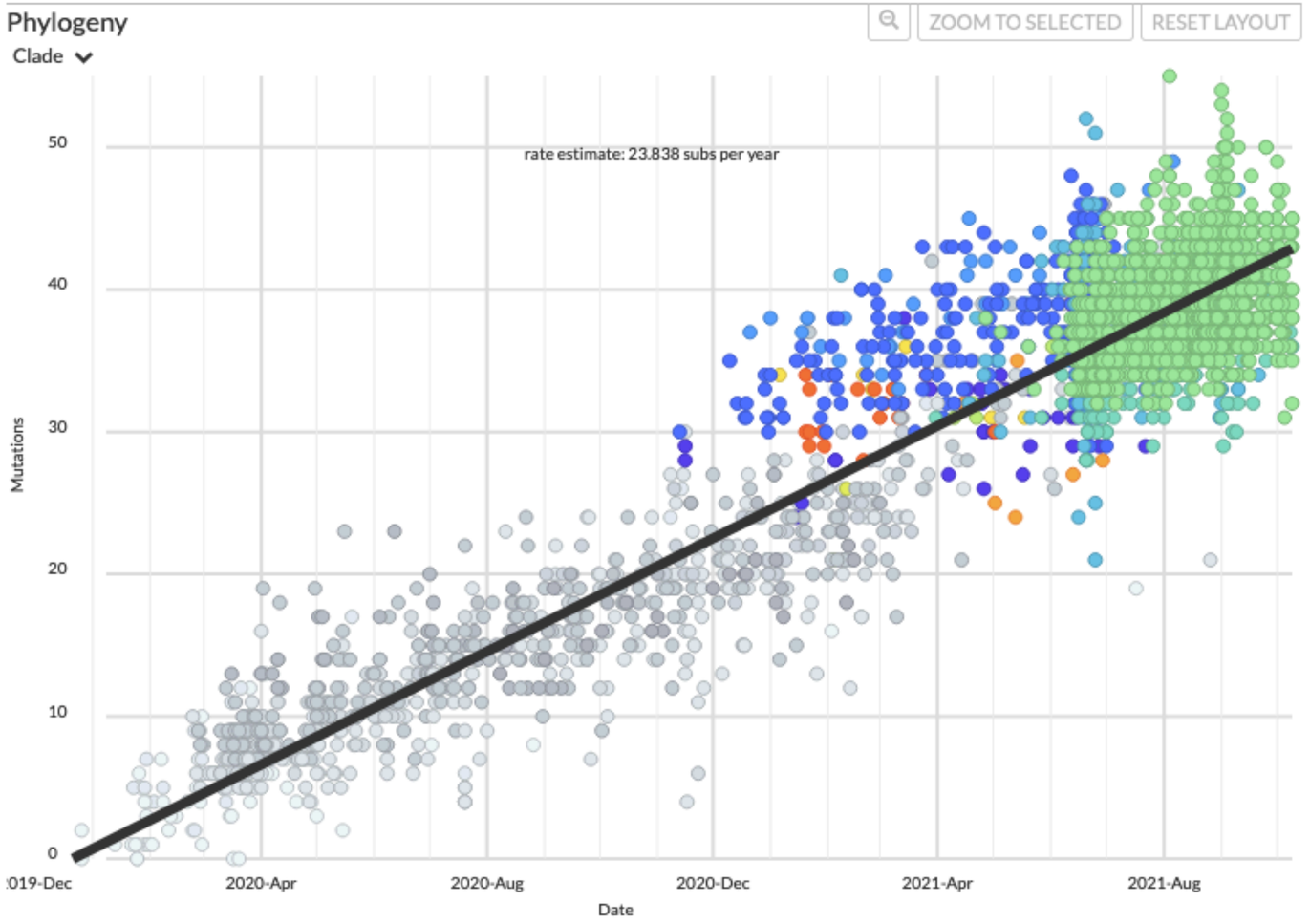


Date and position of root

When did SARS-CoV-2 first infect humans?

https://nextstrain.org/ncov/open/global

Estimate rate of SARS-CoV-2 evolution: ~24 mutations in the genome per year

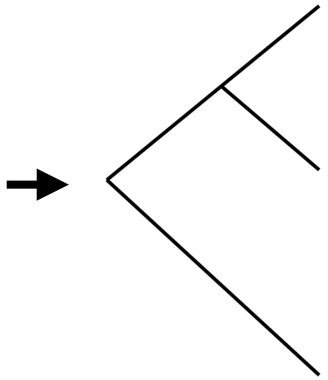


https://nextstrain.org/ncov/open/global

Back to the question: how do you know if your tree is any good?

Multiple sequence alignment

Neighbor joining tree

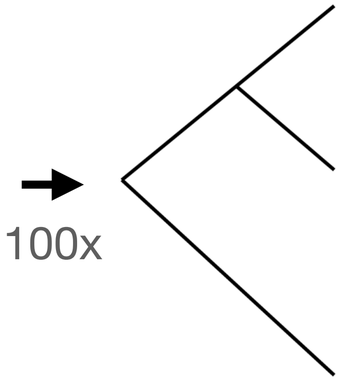


Is this tree reliable?

Tree building option 2: NJ tree with bootstrapping

Multiple sequence alignment

Neighbor joining tree**s**



100 bootstraps: 100 trees

Bootstrap replicate trees are created by shuffling the columns of the alignment a bunch of times and creating trees from the shuffled alignments

Constructing bootstrap data sets. The original data set of 4 taxa (A–D) each with 10 nucleotide characters is bootstrapped across characters (with replacement) to produce bootstrap pseudoreplicates. Each pseudoreplicate contains each of the 4 original taxa, but some original characters are represented more than once and some not at all.

Original Data Set

Taxa	Characters
	1 2 3 4 5 6 7 8 9 10
A	C G A A C C A C T T
B	C G A A C C G G T T
C	G G T A C C G G A T
D	G C T A G C G C A T

Bootstrap Data Sets

Bootstrap Pseudoreplicate 1:

Taxa	Characters
	8 10 7 4 1 10 2 8 5 3
A	C T A A C T G C C A
B	G T G A C T G G C A
C	G T G A G T G G C T
D	C T G A G T C C G T

Bootstrap Pseudoreplicate 2:

Taxa	Characters
	1 8 10 4 2 9 2 8 5 6
A	C C T A G T G C C C
B	C G T A G T G G C C
C	G G T A G A G G C C
D	G C T A C A C C G C

Bootstrap Pseudoreplicate 3:

Taxa	Characters
	3 2 5 7 1 6 9 4 4 10
A	A G C A C C T A A T
B	A G C G C C T A A T
C	T G C G G C A A A T
D	T C G G G C A A A T

Bootstrap Pseudoreplicate 4:

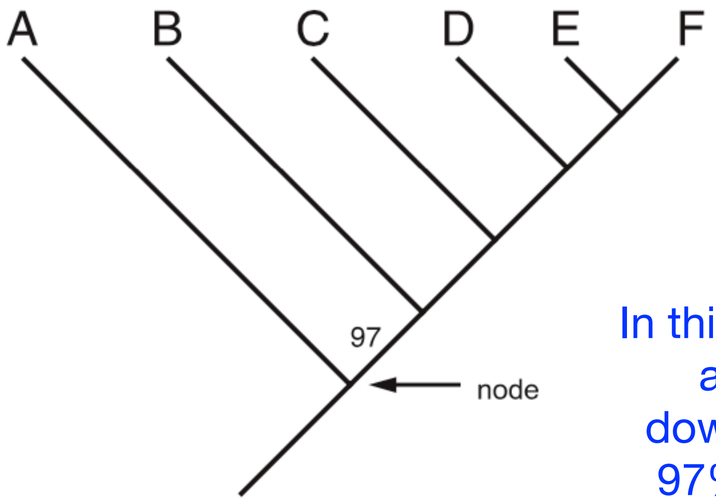
Taxa	Characters
	7 8 5 8 9 6 4 10 1 5
A	A C C C T C A T C C
B	G G C G T C A T C C
C	G C G A C C A T G C
D	G C G C A C A T G C

Create a new tree from each of these “pseudoreplicate” alignments

Columns can repeat in pseudoreplicate alignments

Soltis and Soltis (2003) Statistical Science

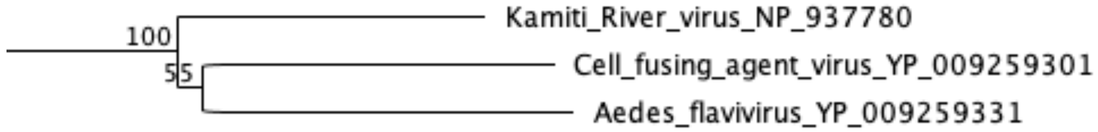
Bootstrap values represent how often the taxa downstream of the branch appeared together in the bootstrap trees



In this example, taxa B, C, D, E, and F grouped together downstream of this branch in 97% of bootstrap replicates

Soltis and Soltis (2003) Statistical Science

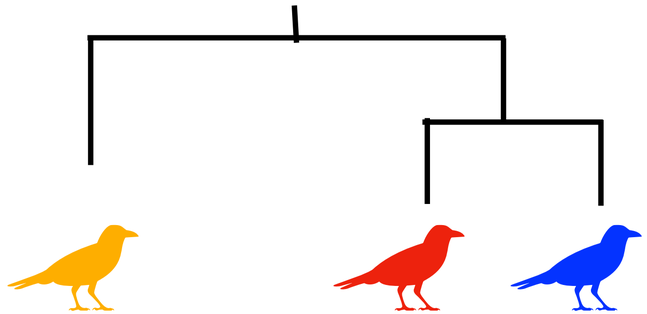
Bootstrap values represent how often the taxa downstream of the branch appeared together in the bootstrap trees



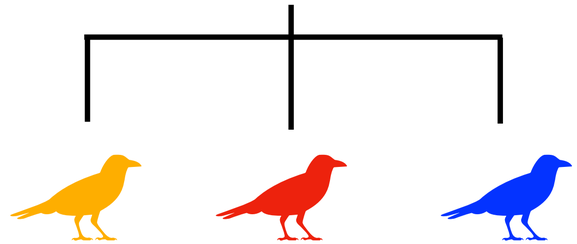
These 3 viruses group together in 100% of trees. But cell fusing agent virus and Aedes flavivirus only group together in 55% of bootstrap replicate trees.

Polytomies represent situations where there is not enough information to resolve the topology of branches downstream of a node

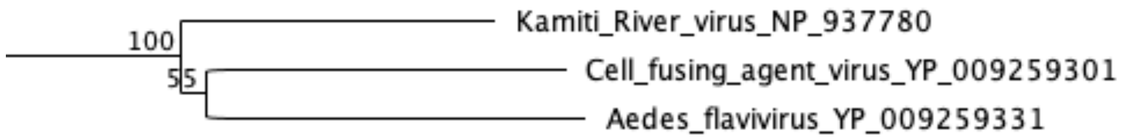
Resolved topology



Polytomy



What are different ways you could re-draw this group of 3 viruses?



These 3 viruses group together in 100% of trees. But cell fusing agent virus and Aedes flavivirus only group together in 55% of bootstrap replicate trees.

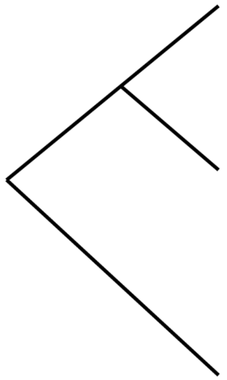
Tree building option 3: maximum likelihood (ML) or Bayesian tree

Multiple sequence alignment

Model selection

MODELTEST

ML or Bayesian tree



These methods usually give you branch support values

Models of sequence evolution

Can vary in whether different types of mutations occur more frequently than others
And whether different positions in sequences change more frequently than others

Model where all types of mutations equally likely

vs

Model where mutations not equally likely

Jukes - Cantor model

All nucleotides undergo changes at the same rate

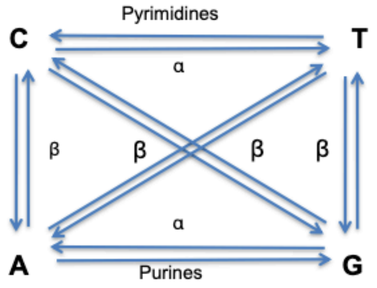
Nucleotide frequencies are the same

$$q_A = q_C = q_G = q_T = ¼$$

	A	T	C	G
A	–	α	α	α
T	α	–	α	α
C	α	α	–	α
G	α	α	α	–

Kimura 2-parameter model

Transitions (α) (purine to purine or pyrimidine to pyrimidine subsitutions) are more common than transversions (β)



	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	β	-	β
G	α	β	β	-

Models of sequence evolution

Can vary in whether different types of mutations occur more frequently than others
And whether different positions in sequences change more frequently than others

Model where all positions in a sequence evolve at equal rates

vs

Model where third bases in codons evolve faster

RNA codon table

1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Amino Acids

Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

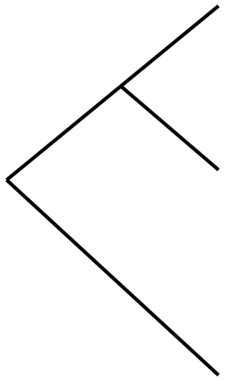
Tree building option 3: maximum likelihood or Bayesian tree

Multiple sequence alignment

Model selection

MODELTEST

ML or Bayesian tree



These methods find the tree that, *given the model,* is most likely to produce the observed sequences

A good tree building option: try multiple methods and report the extent to which they agree

The branches are labeled with ML and Bayesian support values
(Trees created with two methods)