

From DNA sequence to phylogenetic tree, using the GULO gene for Vitamin C synthesis

by Nick Matzke¹ and Wilda Laux²

¹School of Biological Sciences, Rutherford Discovery Fellow, University of Auckland

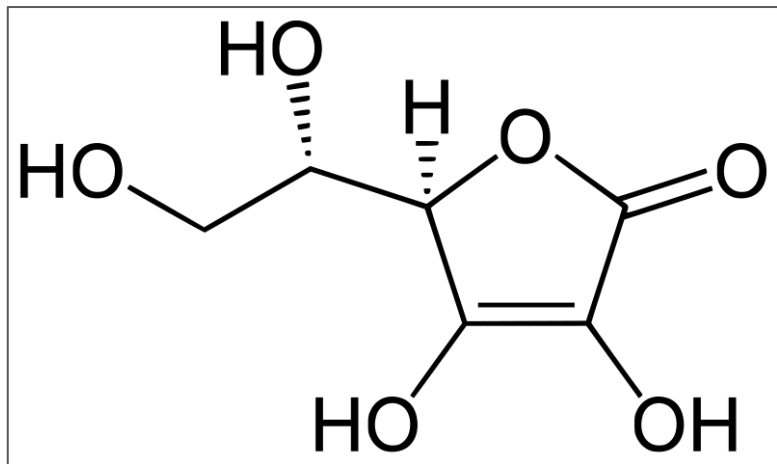
²Department of Molecular Medicine and Pathology, University of Auckland

Link for viewing: (File -> Make a Copy to get your own to edit)

Short: <https://tinyurl.com/gulolab1p0>

Long: <https://docs.google.com/presentation/d/1kWkLm49umMvwmvB0sfjsz0wV8J-iyFAW5WhaMIRfrOc/edit#slide=id.p>

What is Vitamin C?



Ascorbic acid = Vitamin C

In the body,
Vitamin C helps
sythesize...

Dopamine
(neurotransmitter)

Carnitine (helps
convert fat to energy)

Collagen
(structural protein)

Effects if missing:

lassitude (low
mental energy)

lassitude (low
physical energy)

Breakdown of bones,
cartilage, tissue
(bleeding gums)

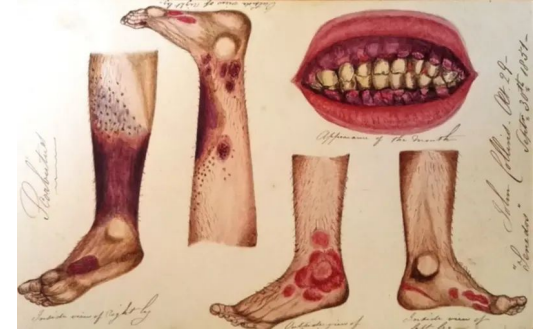
Scurvy
pirates:

*Pirates of the
Caribbean (2003)*

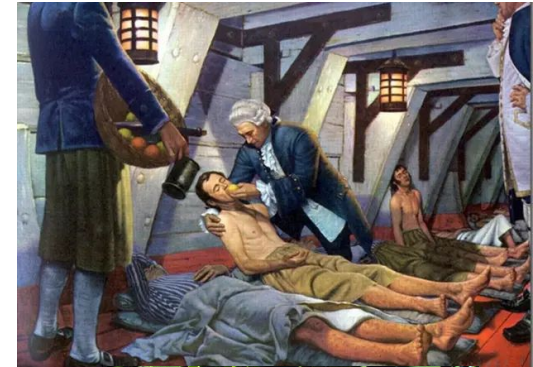


Scurvy (Latin: "scorbutus")

- "Ascorbic acid": name comes from "anti-scorbutic"
- Scurvy becomes a problem after several months without Vitamin C
 - Long winters in grain-fed cultures
 - European sailing voyages, which often lasted months/years
 - Scurvy caused more deaths at sea than shipwrecks, combat, other diseases, etc. combined
 - First medical trial, conducted in 1747 by James Lind (1716-1794), testing 6 remedies for scurvy. Oranges+lemons worked
 - Captain James Cook: ordered sailors take citrus extract etc. to prevent scurvy
 - Royal Navy eventually implemented a "Lime Ration", leading to the term "Limeys"
- Still can be a modern problem, with poor diet



<https://www.bbc.com/news/uk-england-37320399>

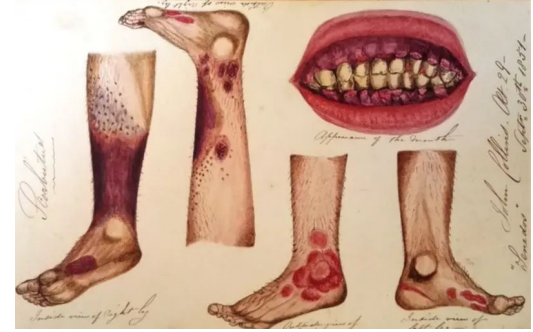


Scurvy (Latin: "scorbutus")

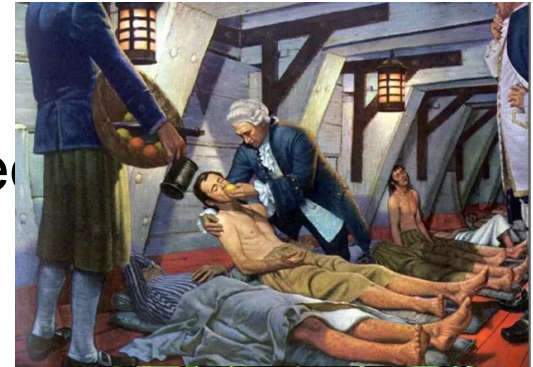
- "Ascorbic acid": name comes from "anti-scorbutic"

Scurvy becomes a problem after several months without Vitamin C

- Long winters in grain-fed cultures
- European sailing voyages, which often lasted months/years
- Scurvy caused more deaths at sea than shipwrecks, combat, other diseases, etc. combined

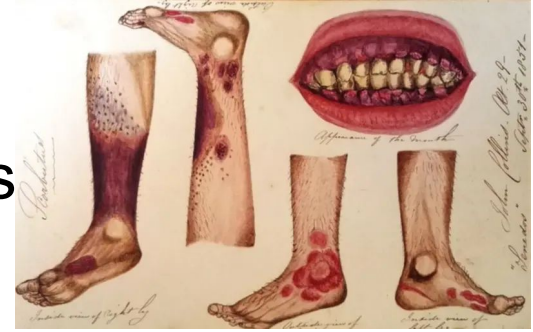


<https://www.bbc.com/news/uk-england-37320399>

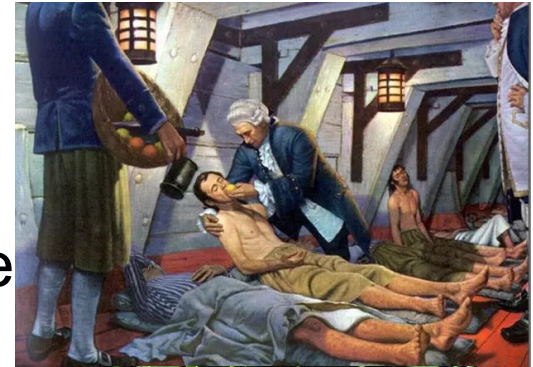


Scurvy (Latin: "scorbutus")

- First medical trial, conducted in 1747 by James Lind (1716-1794), testing 6 remedies for scurvy. Oranges+lemons worked
- Captain James Cook: ordered sailors take citrus extract etc. to prevent scurvy
- Royal Navy eventually implemented a "Lime Ration", leading to the term "Limeys"
- Still can be a modern problem, with poor diet

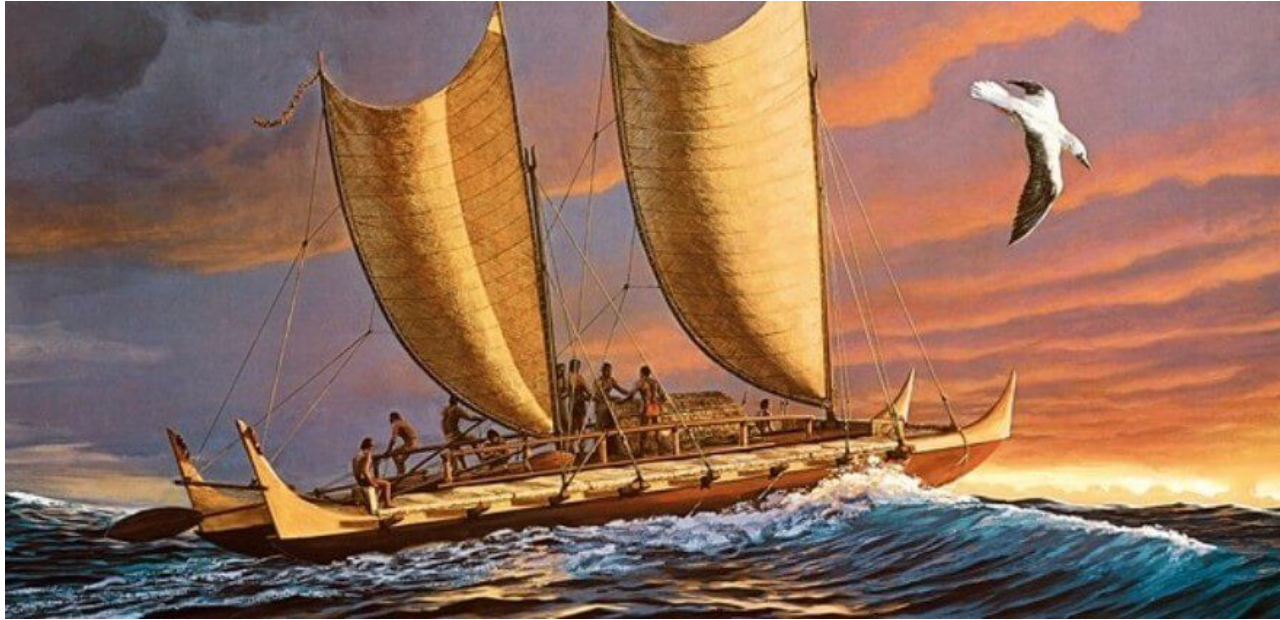


<https://www.bbc.com/news/uk-england-37320399>



Scurvy (Latin: "scorbutus")

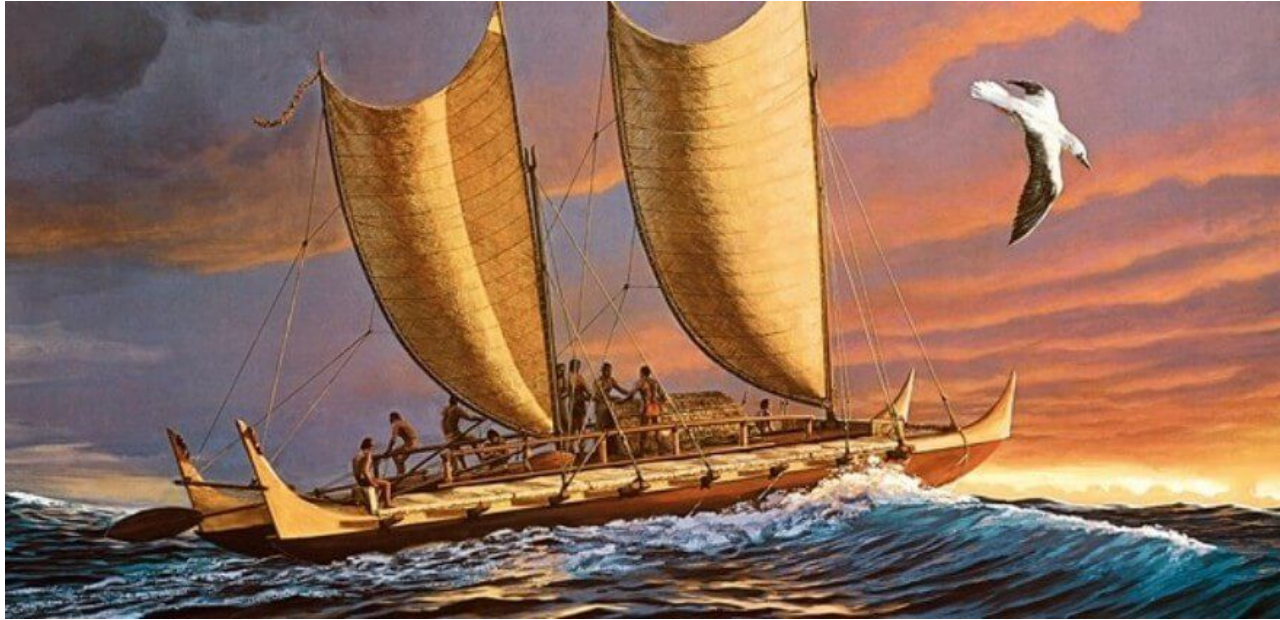
- Thought question: why was scurvy apparently not such an issue for Polynesian voyagers?



<http://www.waihekelocal.co.nz/about/history-of-waiheke/>

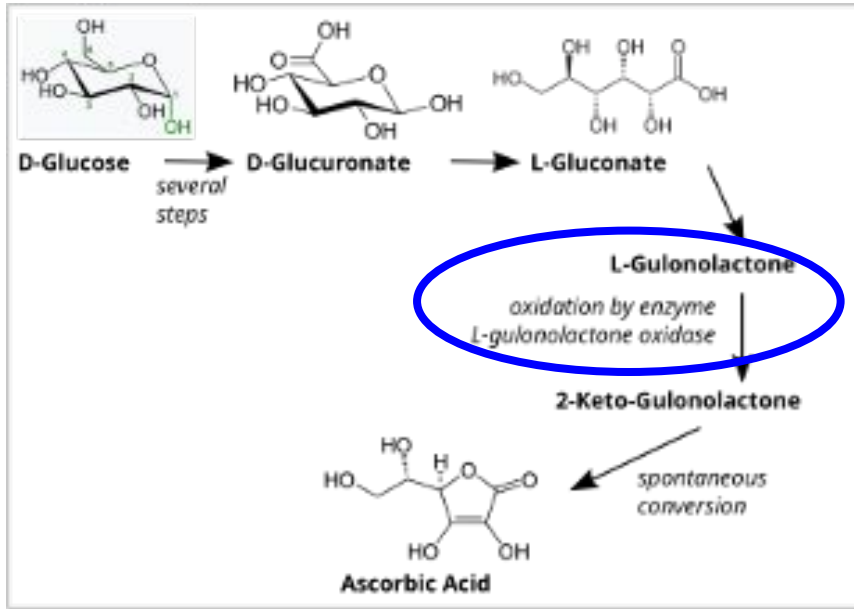
Scurvy (Latin: "scorbutus")

- Thought question: why was scurvy apparently not such an issue for Polynesian voyagers?

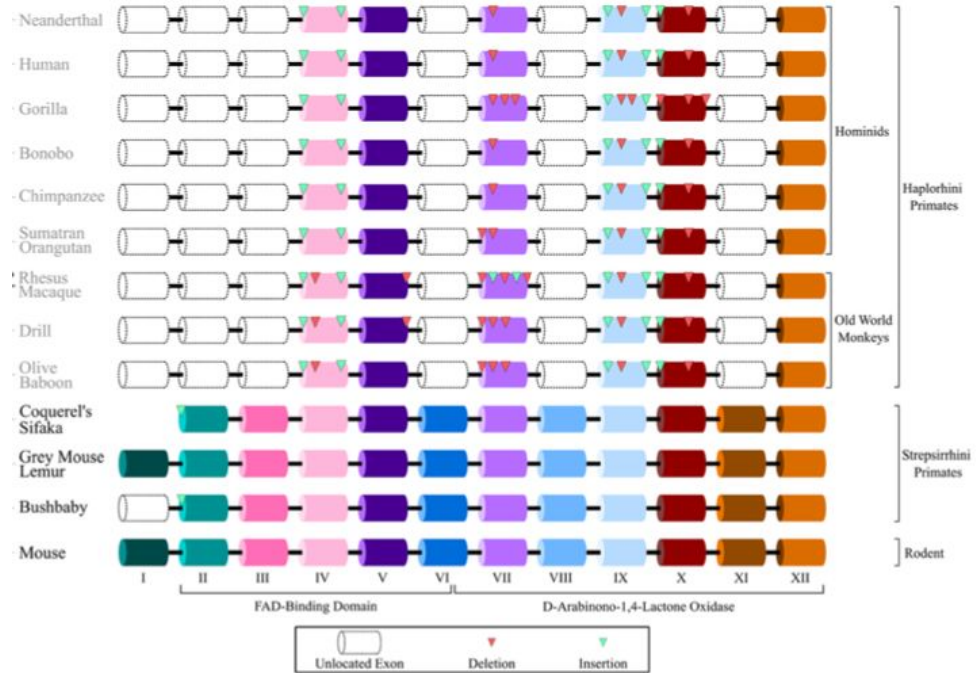


- TOTA (2024) suggests: breadfruit, kelp etc. as Vitamin C sources
- Also consider: travel times. Cook (1769), Tahiti-NZ: 2 months; voyages with Polynesian outrigger canoes: ~3 weeks

Most mammals don't get scurvy, even carnivores. Why not?



They have the L-gulonolactone oxidase protein (an enzyme), which is translated from a functional *GULO* gene



Most primates have a *GULO* pseudogene (*GULOP*), with 6 of 12 exons missing.

(in functional genes, exons are expressed; introns are noncoding DNA between the exons, which is snipped out of the mRNA during transcription)

Question: what can the GULO/GULOP DNA sequence teach us about evolution?

- We are providing the DNA sequence of *GULO/GULOP* exon 12, for various mammals
- The goal here is for students to *discover* what might be in the data, not to be led through the procedure
 - (Hints will be provided, though)
- Phylogenetic inference methods are an advanced topic (upper-level university courses). We are not trying to teach those. We are trying to get across the basic logic of getting from DNA to an evolutionary tree/phylogeny

Your data: GULO/GULOP exon 12, unaligned



Your job: try to line them up, by hand!

Be creative. Use tape!

Your data: GULO/GULOP exon 12, unaligned



Hints:

- keep them in numeric order (not required, but easier)
- start from the back (we ended most sequences in the same place)

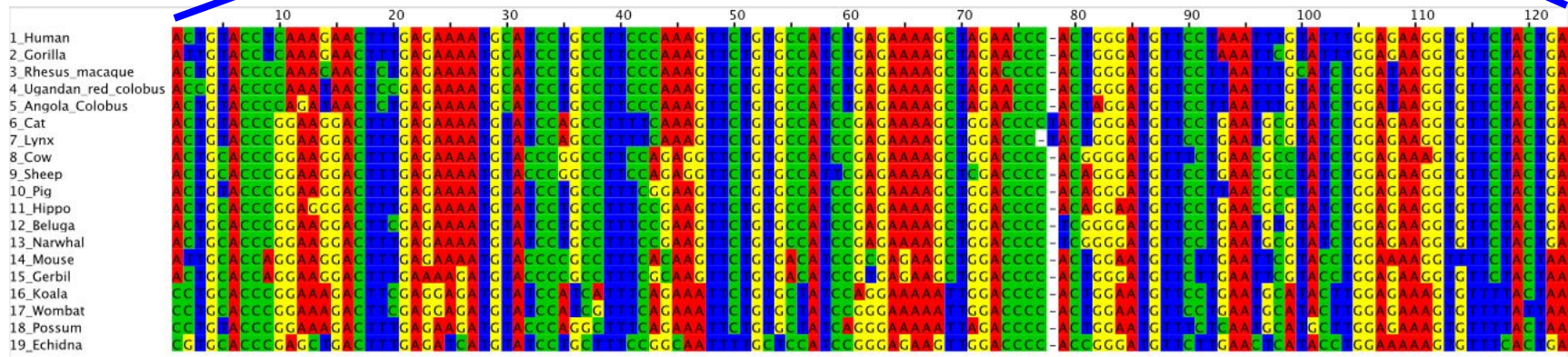
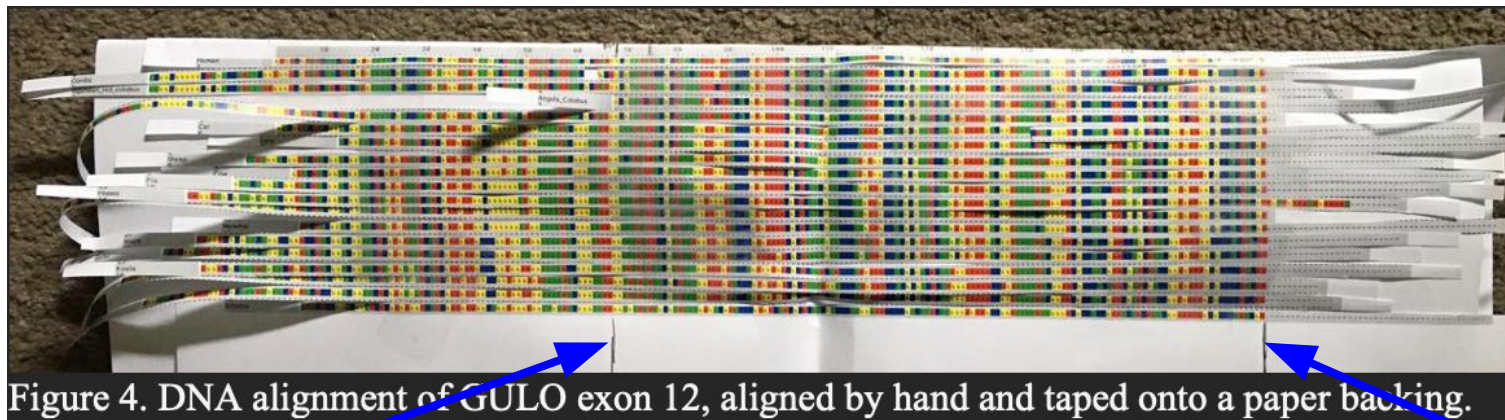
Example hand alignment



Figure 4. DNA alignment of GULO exon 12, aligned by hand and taped onto a paper backing.

- Thought questions: What patterns do you see? (see worksheet, pp. 2-3)

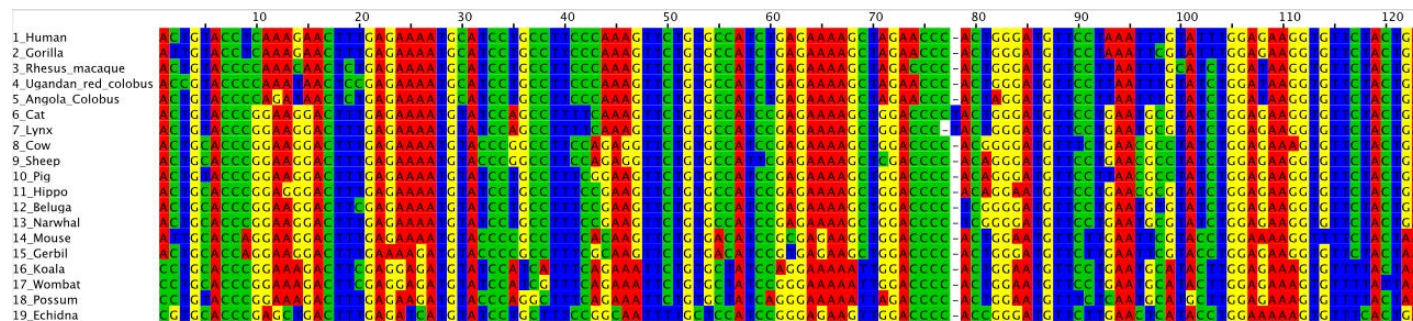
Crop to core alignment



Thought questions: What patterns do you see?

Count DNA differences between pairs of species

(substitutions, which are not merely mutations; substitutions are mutations that have spread to fixation, i.e. 100%, in the population)

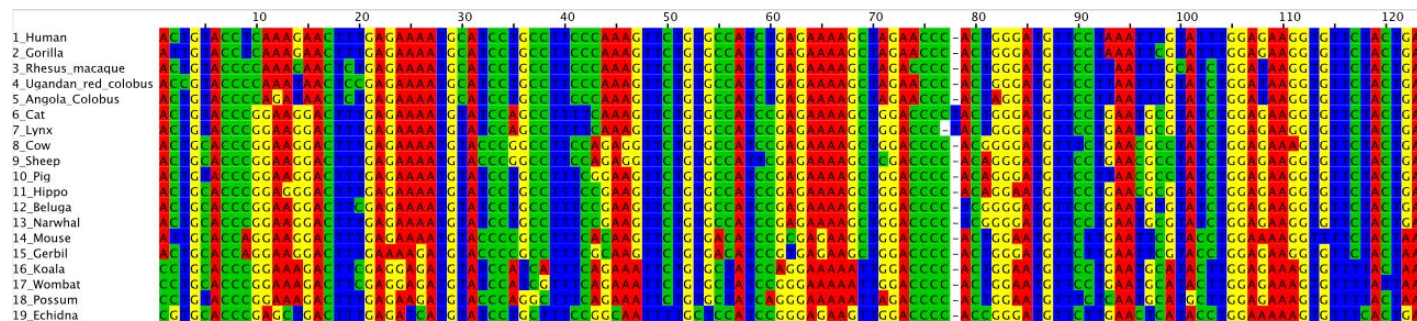


Each group:
take a box &
fill it in

	Human	Gorilla	Rhesus_macaque	Ugandan_red_colobus	Angola_colobus	Cat	Lynx	Pig	Cow	Sheep	Hippo	Beluga	Narwhal	Mouse	Mongolian_gerbil	Koala	Wombat	Possum	Echidna
Human			8	8	8	16	16	18	23	22	19	17	17	27	25	35	35	32	38
Gorilla			10	10	10	16	16	18	23	22	19	19	17	25	25	35	35	32	37
Rhesus_macaque	8	10				17	17	18	24	23	20	18	18	28	27	37	37	34	37
Ugandan_red_colobus	8	10				19	19	20	26	25	22	18	20	30	29	37	37	36	38
Angola_colobus	8	10				19	19	20	26	25	22	20	20	30	29	39	39	36	36
Cat	16	16	17	19	19			9	14	14	9	9	7	19	17	25	25	26	34
Lynx	16	16	17	19	19			9	14	14	9	9	7	19	17	25	25	26	34
Pig	18	18	18	20	20	9	9				6	9	7	21	18	26	26	26	29
Cow	23	23	24	26	26	14	14				11	12	10	21	20	26	26	25	30
Sheep	22	22	23	25	25	14	14				10	13	11	21	20	28	28	28	32
Hippo	19	19	20	22	22	9	9	6	11	10				18	18	24	24	27	29
Beluga	17	19	18	18	20	9	9	9	12	13				20	18	24	24	29	32
Narwhal	17	17	18	20	20	7	7	7	10	11				18	16	24	24	27	30
Mouse	27	25	28	30	30	19	19	21	21	21	18	20	18			28	28	29	32
Mongolian_gerbil	25	25	27	29	29	17	17	18	20	20	18	18	16			27	27	28	31
Koala	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27				31
Wombat	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27				31
Possum	32	32	34	36	36	26	26	26	25	28	27	29	27	29	28				35
Echidna	38	37	37	38	36	34	34	29	30	32	29	32	30	32	31	31	31	35	0

Count DNA differences between pairs of species

(substitutions, which are not merely mutations; substitutions are mutations that have spread to fixation, i.e. 100%, in the population)



Each group:
take a box &
fill it in

	Human	Gorilla	Rhesus_macaque	Ugandan_red_colobus	Angola_colobus	Cat	Lynx	Pig	Cow	Sheep	Hippo	Beluga	Narwhal	Mouse	Mongolian_gerbil	Koala	Wombat	Possum	Echidna
Human	0	2	8	8	8	16	16	18	23	22	19	17	17	27	25	35	35	32	38
Gorilla	2	0	10	10	10	16	16	18	23	22	19	19	17	25	25	35	35	32	37
Rhesus_macaque	8	10	0	5	5	17	17	18	24	23	20	18	18	28	27	37	37	34	37
Ugandan_red_colobus	8	10	5	0	4	19	19	20	26	25	22	18	20	30	29	37	37	36	38
Angola_colobus	8	10	5	4	0	19	19	20	26	25	22	20	20	30	29	39	39	36	36
Cat	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Lynx	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Pig	18	18	18	20	20	9	9	0	10	9	6	9	7	21	18	26	26	26	29
Cow	23	23	24	26	26	14	14	10	0	5	11	12	10	21	20	26	26	25	30
Sheep	22	22	23	25	25	14	14	9	5	0	10	13	11	21	20	28	28	28	32
Hippo	19	19	20	22	22	9	9	6	11	10	0	7	5	18	18	24	24	27	29
Beluga	17	19	18	18	20	9	9	9	12	13	7	0	2	20	18	24	24	29	32
Narwhal	17	17	18	20	20	7	7	7	10	11	5	2	0	18	16	24	24	27	30
Mouse	27	25	28	30	30	19	19	21	21	21	18	20	18	0	8	28	28	29	32
Mongolian_gerbil	25	25	27	29	29	17	17	18	20	20	18	18	16	8	0	27	27	28	31
Koala	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	0	3	13	31
Wombat	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	3	0	13	31
Possum	32	32	34	36	36	26	26	26	25	28	27	29	27	29	28	13	13	0	35
Echidna	38	37	37	38	36	34	34	29	30	32	29	32	30	32	31	31	31	35	0

General idea for evolutionary relationships:

Group by similarity

This could be by tree,
Venn Diagram, etc.

	Human	Gorilla	Rhesus_macaque	Ugandan_red_colobus	Angola_colobus	Cat	Lynx	Pig	Cow	Sheep	Hippo	Beluga	Narwhal	Mouse	Mongolian_gerbil	Koala	Wombat	Possum	Echidna
Human	0	2	8	8	8	16	16	18	23	22	19	17	17	27	25	35	35	32	38
Gorilla	2	0	10	10	10	16	16	18	23	22	19	19	17	25	25	35	35	32	37
Rhesus_macaque	8	10	0	5	5	17	17	18	24	23	20	18	18	28	27	37	37	34	37
Ugandan_red_colobus	8	10	5	0	4	19	19	20	26	25	22	18	20	30	29	37	37	36	38
Angola_colobus	8	10	5	4	0	19	19	20	26	25	22	20	20	30	29	39	39	36	36
Cat	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Lynx	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Pig	18	18	18	20	20	9	9	0	10	9	6	9	7	21	18	26	26	26	29
Cow	23	23	24	26	26	14	14	10	0	5	11	12	10	21	20	26	26	25	30
Sheep	22	22	23	25	25	14	14	9	5	0	10	13	11	21	20	28	28	28	32
Hippo	19	19	20	22	22	9	9	6	11	10	0	7	5	18	18	24	24	27	29
Beluga	17	19	18	18	20	9	9	9	12	13	7	0	2	20	18	24	24	29	32
Narwhal	17	17	18	20	20	7	7	7	10	11	5	2	0	18	16	24	24	27	30
Mouse	27	25	28	30	30	19	19	21	21	21	18	20	18	0	8	28	28	29	32
Mongolian_gerbil	25	25	27	29	29	17	17	18	20	20	18	18	16	8	0	27	27	28	31
Koala	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	0	3	13	31
Wombat	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	3	0	13	31
Possum	32	32	34	36	36	26	26	26	25	28	27	29	27	29	28	13	13	0	35
Echidna	38	37	37	38	36	34	34	29	30	32	29	32	30	32	31	31	31	35	0

General idea for evolutionary relationships:

Group by similarity

This could be by tree,
Venn Diagram, etc.

Be creative!

	Human	Gorilla	Rhesus_macaque	Ugandan_red_colobus	Angola_colobus	Cat	Lynx	Pig	Cow	Sheep	Hippo	Beluga	Narwhal	Mouse	Mongolian_gerbil	Koala	Wombat	Possum	Echidna
Human	0	2	8	8	8	16	16	18	23	22	19	17	17	27	25	35	35	32	38
Gorilla	2	0	10	10	10	16	16	18	23	22	19	19	17	25	25	35	35	32	37
Rhesus_macaque	8	10	0	5	5	17	17	18	24	23	20	18	18	28	27	37	37	34	37
Ugandan_red_colobus	8	10	5	0	4	19	19	20	26	25	22	18	20	30	29	37	37	36	38
Angola_colobus	8	10	5	4	0	19	19	20	26	25	22	20	20	30	29	39	39	36	36
Cat	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Lynx	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Pig	18	18	18	20	20	9	9	0	10	9	6	9	7	21	18	26	26	26	29
Cow	23	23	24	26	26	14	14	10	0	5	11	12	10	21	20	26	26	25	30
Sheep	22	22	23	25	25	14	14	9	5	0	10	13	11	21	20	28	28	28	32
Hippo	19	19	20	22	22	9	9	6	11	10	0	7	5	18	18	24	24	27	29
Beluga	17	19	18	18	20	9	9	9	12	13	7	0	2	20	18	24	24	29	32
Narwhal	17	17	18	20	20	7	7	7	10	11	5	2	0	18	16	24	24	27	30
Mouse	27	25	28	30	30	19	19	21	21	21	18	20	18	0	8	28	28	29	32
Mongolian_gerbil	25	25	27	29	29	17	17	18	20	20	18	18	16	8	0	27	27	28	31
Koala	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	0	3	13	31
Wombat	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	3	0	13	31
Possum	32	32	34	36	36	26	26	26	25	28	27	29	27	29	28	13	13	0	35
Echidna	38	37	37	38	36	34	34	29	30	32	29	32	30	32	31	31	31	35	0

Doing a "neighbor-joining" (NJ) *algorithm* by hand

Algorithm: arabic word for a procedure that manipulates data, does a calculation, etc.

This is not the exact NJ algorithm, but gives the general idea)

- Which two species are most similar genetically?
- Which two species are next most similar genetically?
- Which two species are next most similar?
- Which two species are next most similar?
- What species is most similar to the colobus monkey group?
- Which group/species seems most genetically similar to the human/gorilla/monkey group?
- Use similar grouping logic to group possum, koala, wombat.
- Use similar grouping logic to group pig, cow, sheep, and hippo, beluga, narwhal.
- Do marsupial mammals and placental mammals seem to form genetic similarity groups as well?

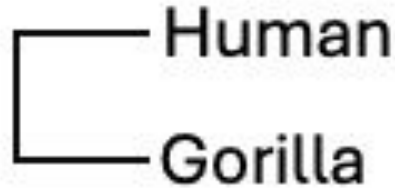
	Human	Gorilla	Rhesus_macaque	Ugandan_red_colobus	Angola_colobus	Cat	Lynx	Pig	Cow	Sheep	Hippo	Beluga	Narwhal	Mouse	Mongolian_gerbil	Koala	Wombat	Possum	Echidna
Human	0	2	8	8	8	16	16	18	23	22	19	17	17	27	25	35	35	32	38
Gorilla	2	0	10	10	10	16	16	18	23	22	19	19	17	25	25	35	35	32	37
Rhesus_macaque	8	10	0	5	5	17	17	18	24	23	20	18	18	28	27	37	37	34	37
Ugandan_red_colobus	8	10	5	0	4	19	19	20	26	25	22	18	20	30	29	37	37	36	38
Angola_colobus	8	10	5	4	0	19	19	20	26	25	22	20	20	30	29	39	39	36	36
Cat	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Lynx	16	16	17	19	19	0	0	9	14	14	9	9	7	19	17	25	25	26	34
Pig	18	18	18	20	20	9	9	0	10	9	6	9	7	21	18	26	26	26	29
Cow	23	23	24	26	26	14	14	10	0	5	11	12	10	21	20	26	26	25	30
Sheep	22	22	23	25	25	14	14	9	5	0	10	13	11	21	20	28	28	28	32
Hippo	19	19	20	22	22	9	9	6	11	10	0	7	5	18	18	24	24	27	29
Beluga	17	19	18	18	20	9	9	9	12	13	7	0	2	20	18	24	24	29	32
Narwhal	17	17	18	20	20	7	7	7	10	11	5	2	0	18	16	24	24	27	30
Mouse	27	25	28	30	30	19	19	21	21	18	20	18	0	8	28	28	29	32	
Mongolian_gerbil	25	25	27	29	29	17	17	18	20	20	18	18	8	0	27	27	28	31	
Koala	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	0	3	13	31
Wombat	35	35	37	37	39	25	25	26	26	28	24	24	24	28	27	3	0	13	31
Possum	32	32	34	36	36	26	26	26	25	28	27	29	27	29	28	13	13	0	35
Echidna	38	37	37	38	36	34	34	29	30	32	29	32	30	32	31	31	31	35	0

Tree drawing, quick example:

1. If 2 species had 2 total differences:

Tree drawing, quick example (not from *GULO*):

1. If human and gorilla are closer to each other than to anything else (2 total differences in DNA) connect them like this:



2. The colobus monkeys have 4 differences, and colobus vs. macaque is 5 differences. Connect them like this:



Using these principles, try to draw a tree that shows the groups-within-groups structure of the DNA data

Use a pencil so you can erase if you need to.

Start with the smallest groups (linking the species with the most similar DNA).

Then group those small groups into bigger groups.

It does not have to be perfect, you are exploring the structure in the DNA data.

Use the questions to guide you.

Draw your phylogenetic tree here

Human
Gorilla
Rhesus_macaque
Ugandan_red_colobus
Angola_colobus
Cat
Lynx
Pig
Cow
Sheep
Hippo
Beluga
Narwhal
Mouse
Mongolian_gerbil
Koala
Wombat
Possum
Echidna

The phylogenetic tree *inferred* by NJ on GULO exon 12 sequence (see R script in Appendix)

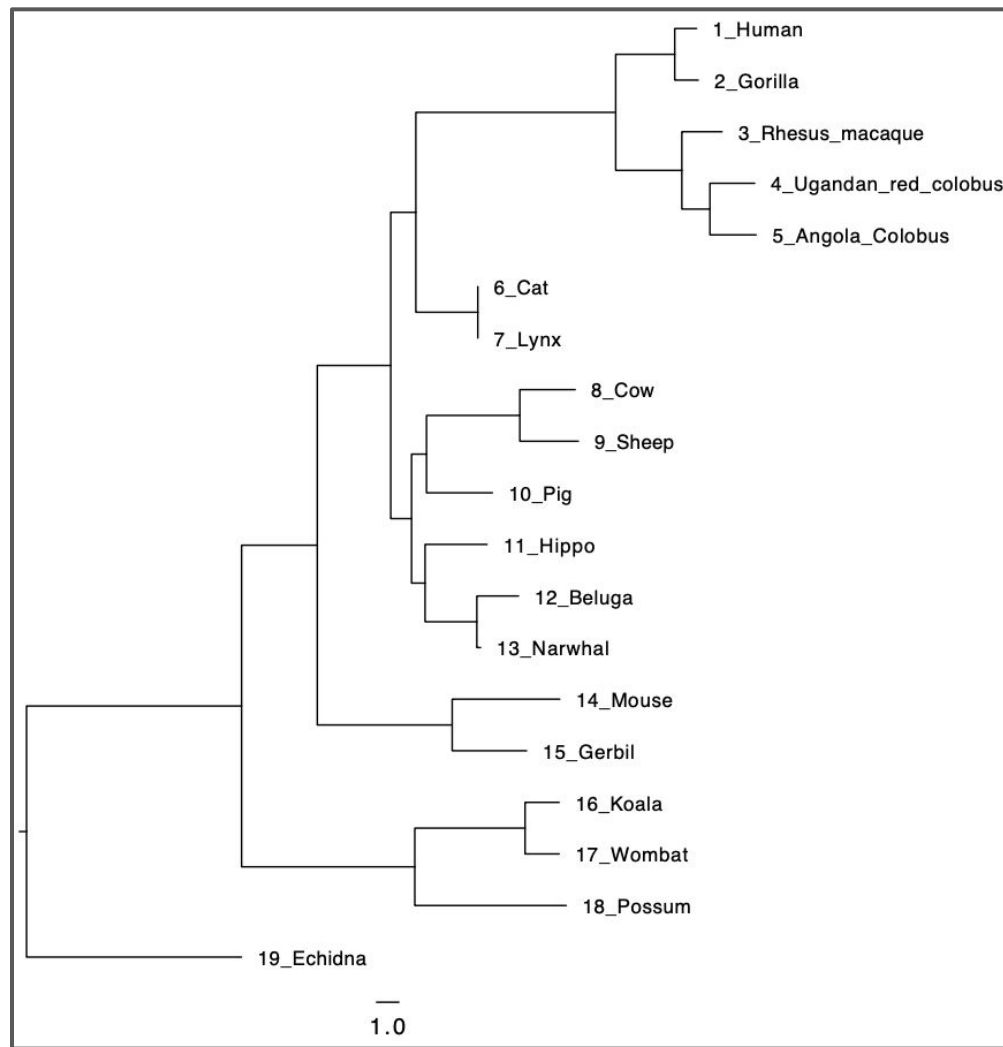
Compare your tree to the computer-estimated tree

Notes: This is not merely "tree-drawing" or even "reconstruction".

The modern terminology is *inference* or *estimation*. We are attempting to learn/infer the phylogenetic history from the DNA data.

Phylogenetics is now a science of *statistical inference*, not just a graphical exercise.

Phylogenies represent *hypotheses* that can be *tested*, refined, revised with new data and newer models & inference methods



Reading a phylogeny

Internal nodes: represent speciation events (population splitting events) millions of years ago

Tip node (also terminal node, terminal taxon, leaf, tip, or just species) - represents the observed DNA sequence, usually (not always) from a living species

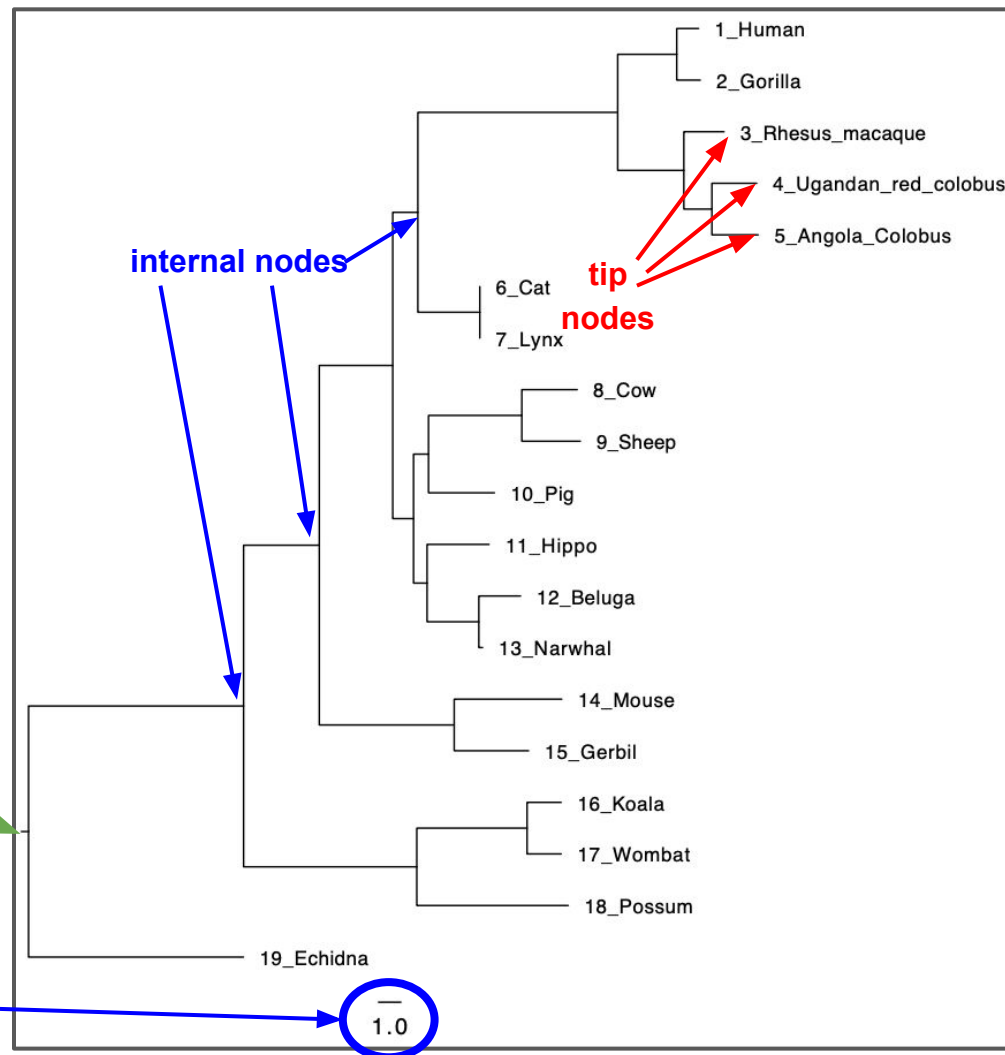
Root node: Earliest node in the tree (computers give you unrooted trees; the human decides which branch determines the root, usually using an "outgroup", a distant relative of the other species. Here we use echidna, an egg-laying mammal).

the "root" node

internal nodes

tip
nodes

Scale bar represents ~1 DNA
change (1 substitution)



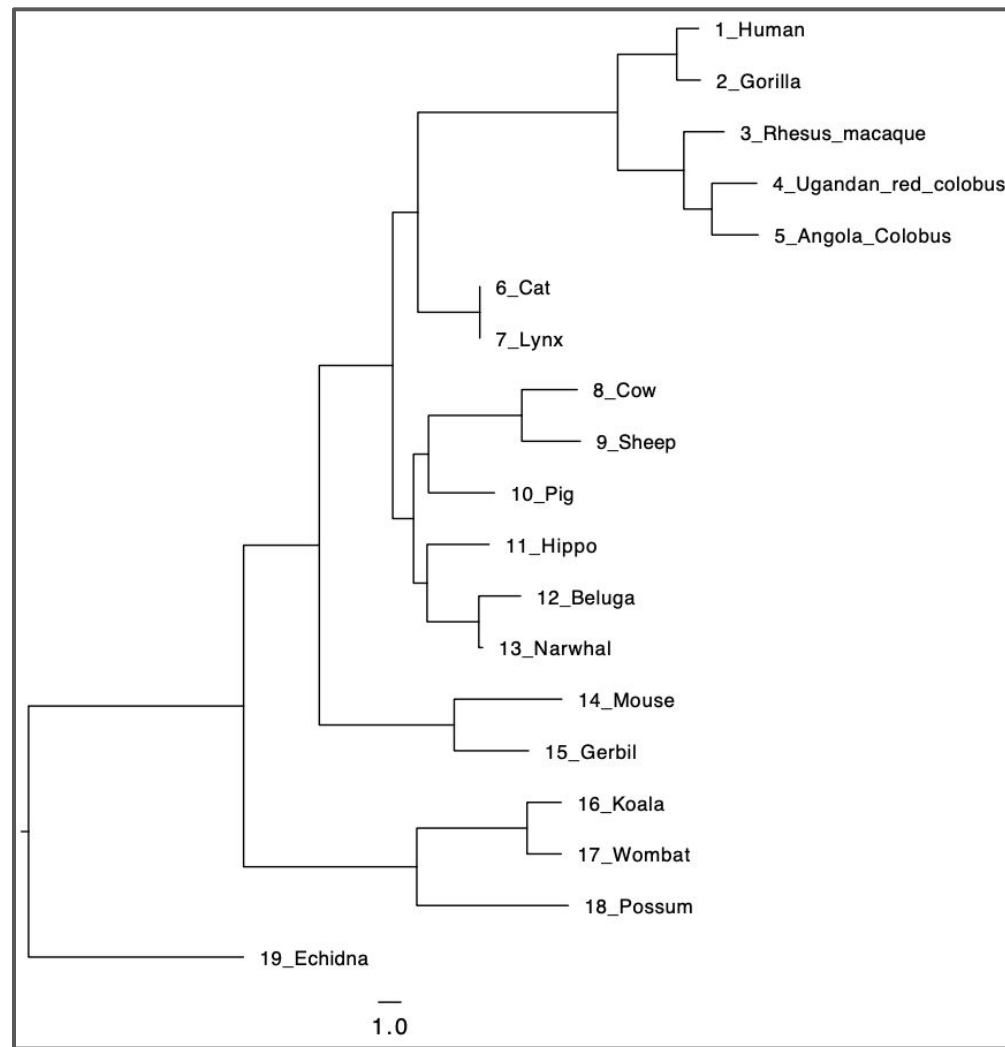
Thought questions; points to consider

This tree is surprisingly good, for a tree based on just 1 exon with 122 DNA bases

Look for:

- great apes
- primates
- cetartiodactyls
- placentals
- marsupials
- monotremes

Why are the primates on an extra-long branch?



Conclusions

Similar logic can be used on other DNA markers:

- mitochondrial DNA
- amino acids of protein sequences
- morphological characters
- texts that have been copied and modified
- languages
- behaviors (e.g. bird song)
- geographic ranges (Matzke's research)
- protein structure (Matzke's research, see right)

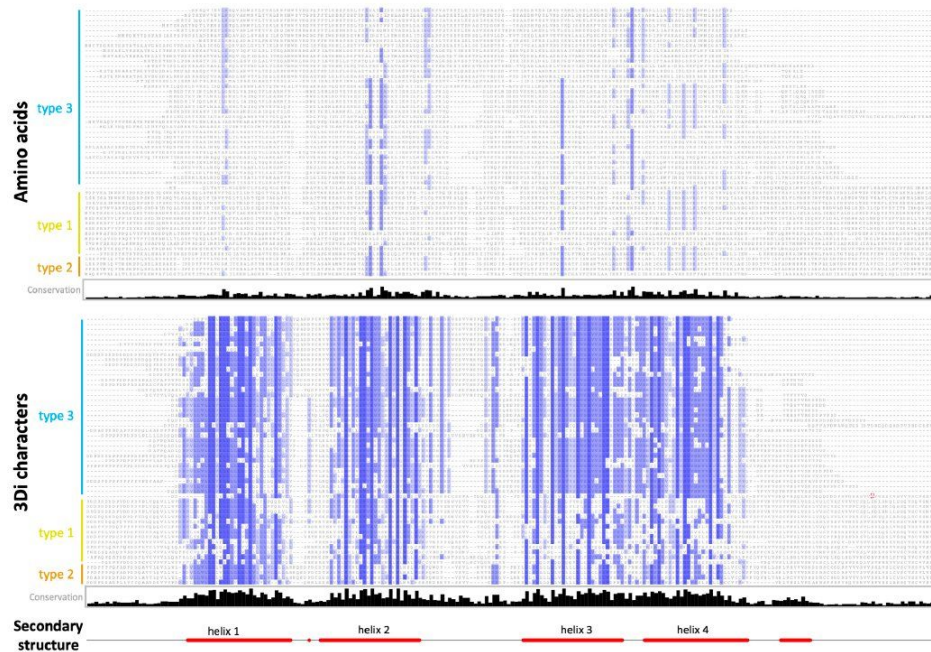
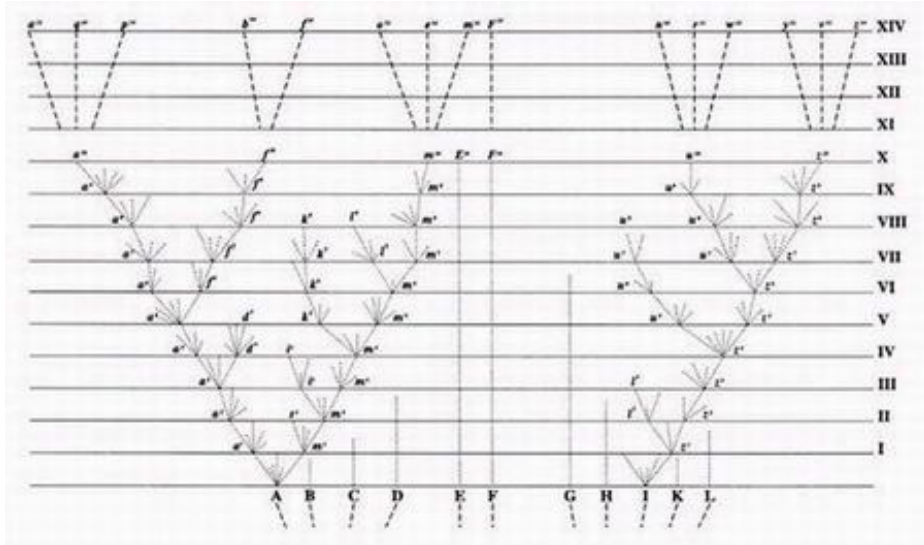


Figure 2. AA and 3Di alignments for 53 AlphaFold structures of the ferritin-like superfamily, displayed with Jalview's pairwise identity color scheme, where characters matching the majority character (if any) are colored, and darker colors are closer to 100% agreement. The 3Di alignment was generated with famsa3di; the AA alignment was generated by one-to-one replacement of 3Di states with AAs. Sites with <35% data were trimmed. FASTA files are available in SI.

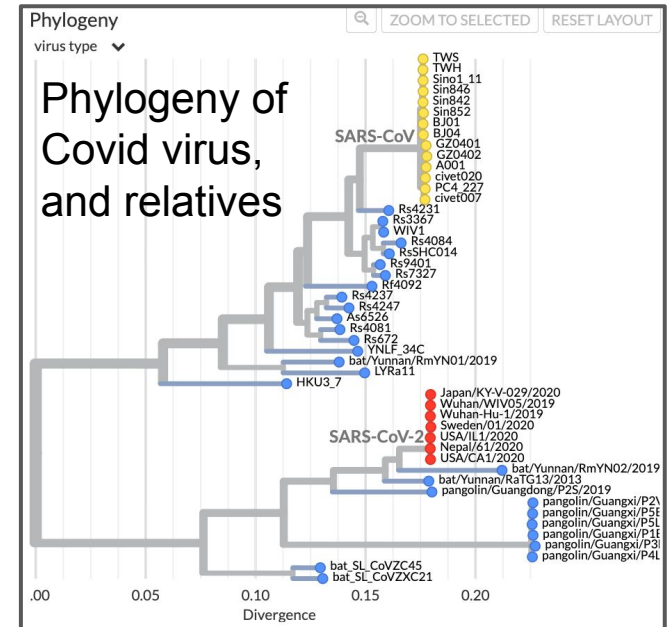
Conclusions

Charles Darwin's only figure in *On the Origin of Species* (1859) was an imagined phylogenetic tree, representing evolution, meaning: "descent with modification"



With DNA sequencing and computer algorithms for statistically inferring trees, phylogenetics is now a robust statistical science

(come study it at the University of Auckland!)



<https://nextstrain.org/groups/bla/sars-like-cov>