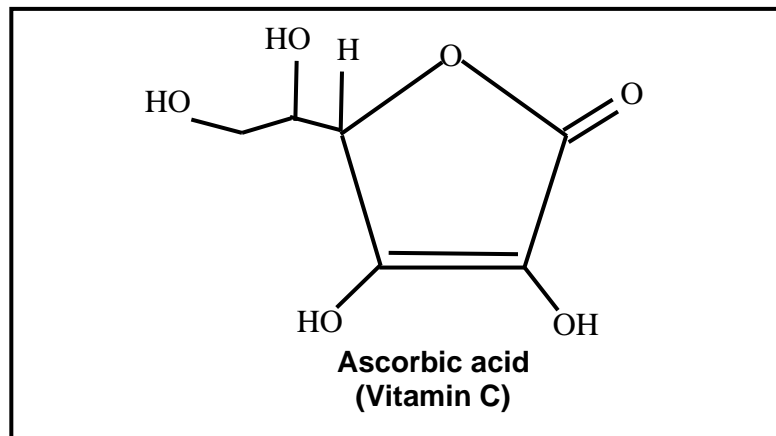


This resource is for educational purposes. For any other use, please consult the writers for permission

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

Teacher resource



A resource developed by Dr Nick Matzke¹ in collaboration with Dr Wilda Laux²

¹School of Biological Sciences, University of Auckland, Email: n.matzke@auckland.ac.nz

²Department of Molecular Medicine and Pathology, University of Auckland, Email: wilda.laux@auckland.ac.nz

Nick Matzke & Wilda Laux

Purpose: Students will be able to use genetic information about the *GULO* gene/pseudogene to analyse the evolutionary relationships between a selection of mammal species.

Background on scurvy: In vertebrates, glucose (sugar) is converted into Vitamin C through an enzyme-controlled biosynthetic pathway as shown below.

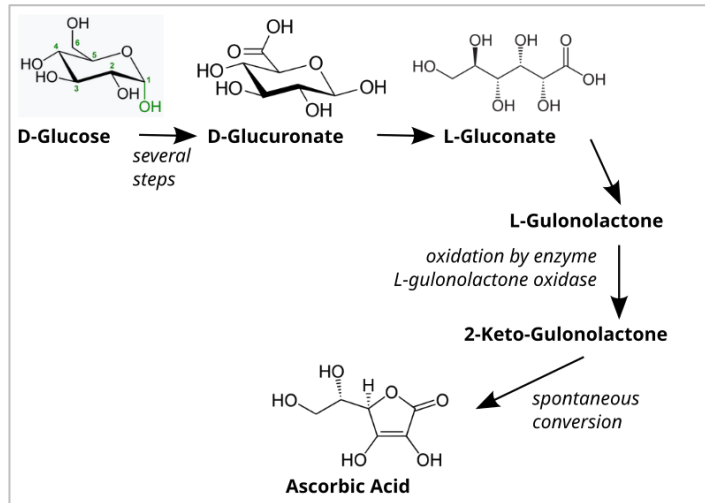


Figure 1. Vitamin C Biosynthesis in Vertebrates.

(Source: Ian Alexander, Wikipedia, 22 November 2017

https://en.m.wikipedia.org/wiki/File:Vitamin_C_Biosynthesis_in_Vertebrates.svg)

One enzyme involved in the Vitamin C's pathway is L-gulonolactone oxidase (*GULO*) which is coded by the *GULO* gene/pseudogene.

Vitamin C is a co-substrate in various other biological pathways, including those involved in the formation of collagen, dopamine, and carnitine (De Tullio, 2010). Severe lack of Vitamin C causes the disease scurvy, most famously associated with pirates and other long-term sailors. Symptoms of scurvy include bleeding gums, loose teeth, and weak bones and cartilage (all byproducts of defective collagen), and "lassitude," a lack of mental and physical energy (byproducts of lack of dopamine, a neurotransmitter, and carnitine, which helps the body convert fat to energy).

The very term "ascorbic acid" derives from the term "anti-scorbutic," derived from the Latin term for scurvy, "scorbutus." Abundant online resources discuss the history of scurvy, which was a particular risk for long sailing voyages, as scurvy becomes a risk after several months without vitamin C. Other situations where scurvy can be found include long winters in grain-fed cultures, or modern people with a poor diet severely lacking in fresh fruits and vegetables. Scurvy was a significant problem for the British navy, which eventually required a "lime ration" for sailors. Lime is high in vitamin C, and stores well. This led to the term "limey."

It may be interesting to consider how other sailing cultures avoided scurvy; one source (TOTA, 2024) notes that Polynesian voyagers obtained vitamin C from dried breadfruit, kelp, and sweet potatoes; although it should also be noted that Polynesian double-hulled canoes were much faster than European sailing ships, so voyages were often weeks rather than months.

Humans need vitamin C, but most vertebrates, including most mammals, can synthesize vitamin C without getting vitamin C in their diet. Lions, dolphins, etc. are not eating a lot of fresh fruit. So, why do humans need vitamin C in their diet?

Nick Matzke & Wilda Laux

It turns out that humans, along with most other primates, have a *GULO* gene which is “broken” – it is a pseudogene known as *GULOP* (*GULO* pseudogene). Probably the tree-dwelling, monkey-like ancestors of primates had abundant sources of fruit in their diet, and thus the presence or absence of a functional *GULO* did not matter for fitness. Mutations accumulated and eventually parts of the *GULO* sequence were lost. The functional *GULO* protein consists of about 440 amino acids, encoded by 1320 nucleotides spread across 12 exons. However, in Haplorhini primates (monkeys and apes) only 6 of these exons remain detectable in the genome (exons 4, 5, 7, 9, 10, 12).

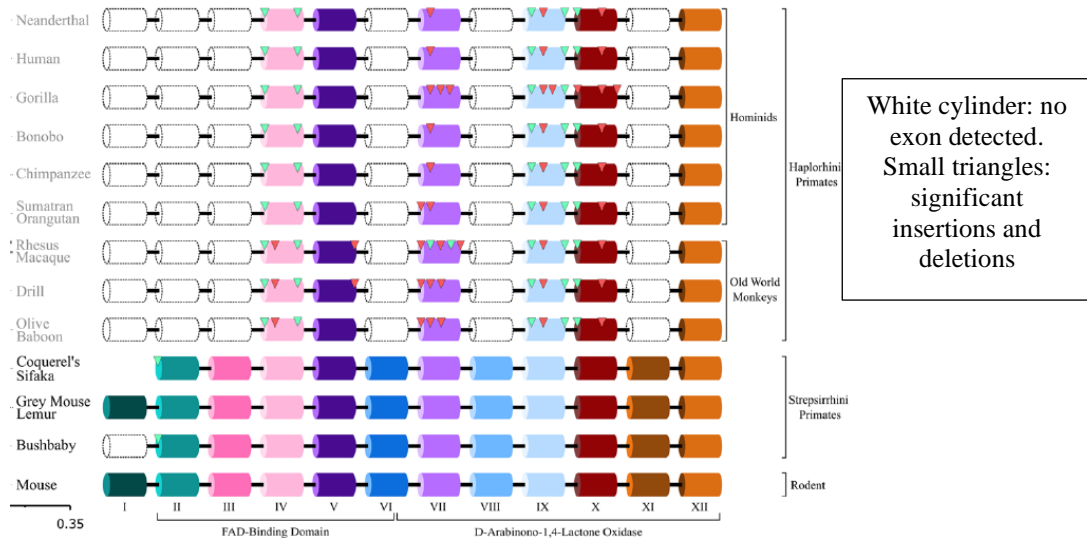


Figure 2. Exon structure of functional *GULO* genes (typically 12 exons), compared to the *GULOP* pseudogenes of Haplorhine primates (6 exons detectable). (Source: Figure 2 of Mansueto & Good 2024).

***GULO* and arguments for evolution.** *GULO* became slightly famous in internet debates over evolution, an issue much in the news in the USA in the 2000s, for the following reason. A common, “person on the street” intuition about biology might be that species have genetic similarities because the species have similar functional needs. So, perhaps genetic similarities are not necessarily evidence for the common ancestry, but for common function. Whatever the virtues of this argument (there are weighty philosophical or even theological debates here, out of scope for a basic science class), it is clear that the intuition would not work to explain genetic similarities in *nonfunctional* genetic material.

The evolutionary explanation for genetic similarities is, simply, *copying*. The DNA sequence is copied every generation, and if this process continues for millions of years, even very different-looking species will share genetic similarities inherited from their common ancestral species.

Mutations vs. substitutions. Very rarely, errors occur in the DNA copying process, called mutations. Even more rarely, those mutations will spread in the population over many generations, either by chance (genetic drift) or by differential reproduction/survival (natural selection). Mutations that spread throughout an entire population have become “fixed” in the population (as opposed to being variable in the population). Fixed mutations are known as *substitutions*.

Because substitutions happen so rarely, we can use them to measure how related different species are. Species that share common ancestry more recently will share more genetic similarity, as less time has been available for substitutions to accumulate. We can use the pattern of similarities and differences to attempt to estimate a *phylogenetic tree* – the history of speciation events that produced the species we see today.

Nick Matzke & Wilda Laux

This is evolutionary process of “descent with modification” that Darwin hypothesized in his 1859 book *On the Origin of Species*.

Background on sequence alignment. *Sequence alignment* is the process whereby a biologist decides which DNA bases are homologous. “Homology” is the assertion that two traits descend from a common ancestor by copying. Originally, the term “homology” was used to refer to e.g. the similarities in arm structure across vertebrates, but now it is commonly used for molecular traits as well.

Methods.

Aim: To give students a concrete experience in the basics of phylogenetics, using real genetic data. The intent is to avoid all abstractions (computers, phylogenetic inference algorithms), and produce a hands-on exercise. Students will manually align paper strips containing the GULO exon 12’s gene sequence from different mammal species to produce a phylogenetic tree that shows the evolutionary relationships between these species.

1. Distribute materials.

Each student or group (suggestion: make groups of 3-4 students) needs, minimally:

- A copy of the student instructions.
- A copy of the computer-derived tree (GULO_exon12_aligned_19sp_TREE_PRINT_TO_A4.pdf) to use as model answer at the end)
- A paper copy of 19 unaligned DNA sequences (GULO_exon12_unaligned_19sp_PRINT_TO_A3.pdf), printed out in a large enough form to be readable
 - Suggestion: print the unaligned sequences in colour, to A3 paper, with landscape view (not portrait)
 - Either the teacher or the students will need to cut the unaligned sequences into 19 separate paper strips. Cut along the dashed lines in-between each of the 19 sequences.
 - *Hint: Have a bunch of backup copies of the sequences*, in case DNA strips get ripped, cut, students want to re-do a section, etc.
- A clear A3 sheet of paper to tape the sequences onto as the students align them.
- Plenty of clear tape
- Optional: A paper copy of the “core alignment,” as backup if the student-made alignments are too messy. This is *GULO_exon12_aligned_19s_PRINT_TO_A4.pdf*, and can be printed (colour, landscape) on A4 paper.
- Optional: A copy of the core and non-core alignment of all 19 strips to compare with what they get after they align the DNA strips in step 2.
(This is Figure 3. *GULO_exon12_aligned_core_and_noncore_19sp.png*)

2. Align the DNA strips.

Students should:

- By hand, slide two strips of DNA sequences back and forth against each other until they find matching/similar sequence. When they do, they should then get another DNA strip and align it against those already aligned.
- This soon gets hard to do without disturbing other paper strips, so students should use tape where convenient.

Nick Matzke & Wilda Laux

Tips to give students as they struggle with doing the manual alignment:

- While assembling the alignment, students may find it convenient or intuitive to place sequences near each other when the sequences are more similar.
- Students may also find it convenient to first try aligning species that students think are more similar overall (e.g. cat and lynx; primates; whales, etc.)
- It turns out that the species names are numbered 1 to 19. The assignment can be done without this ordering, but having them in order makes it easier to check correctness, and to line them up with the eventual phylogenetic tree.
 - If you want to make it easier for students, have them order the DNA strips by number from the beginning
- There are a few “-” characters inside individual DNA sequences. **Leave these as-is.**
 - The “-” characters are called “gaps.” These represent *indels* – places where DNA bases have been *inserted* or *deleted* over evolutionary history (more rarely, they might be due to a DNA sequencing error, or human computational error).
 - In real-life science, computer algorithms insert gaps to help line up DNA. A more complex version of the exercise would have students do this by hand. But this involves cutting many strips into pieces, which rapidly gets chaotic.
- The real “trick” with this dataset is that the *end* of the sequences mostly all line up – it’s just the front part of each sequence that contains different amounts of DNA sequence. So, if students get stuck, ask them to try lining sequences up from the end first.

The result should look something like this:

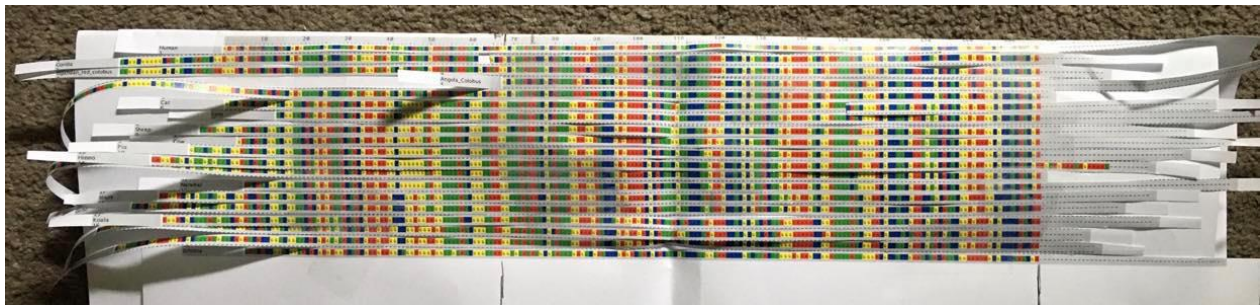


Figure 3. DNA alignment of GULO exon 12, aligned by hand and taped onto a paper backing. A printed version of this is *GULO_exon12_aligned_core_and_noncore_19sp.png*

Question for students: Is there a pattern in the DNA data? Describe what you see.

3. Highlight the core alignment. Aligned, these sequences have some leading and trailing sequence to be ignored – the phylogenetic information is in the shared, clearly homologous sequence.

- The “core alignment” corresponds to positions 75 through 197 only in the *GULO_exon12_aligned_core_and_noncore_19sp.png*. On the Human sequence, which is first in the PDF, this starts with ACTGTACCTCAAAGAA..., and ends with the end of the Human sequence (...CTACTGA).
- Once this region is identified, have students delineate the core alignment (e.g., indicate where it starts and stops by drawing arrows or a box)
- Optional: Students may cut their alignment down to the core alignment
If the student alignment is messy, provide the core alignment printout, which will be much neater than the student product (*GULO_exon12_aligned_19sp_PRINT_TO_A4.pdf*)

Nick Matzke & Wilda Laux

4. Analyse the core alignment. Leaving out extra sequence at the beginning/end produces the “core alignment” of exon 12 found in Figure 4 below.

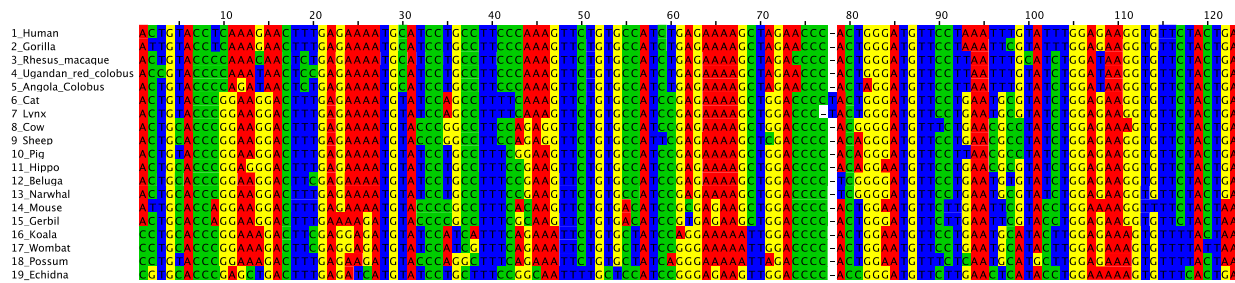


Figure 4. Aligned DNA sequences including *GULO/GULOP* exon 12, for 19 mammal species. This is *GULO_exon12_aligned_19sp_PRINT_TO_A4.pdf*

Questions for students:

Which sequence seems most different from the others? echidna

Which pair of sequences seem most similar? Cat and lynx, identical except for a missing base

Do the differences in each column appear randomly, or do they seem to have some structure? they have structure- similar species seem to share substitutions

5. In depth analysis: Counting differences between pairs of sequences. Nowadays, scientists use a computer to count the number of differences between sequences. However, this can be done by hand. This information can be displayed in what we call a *genetic distances matrix*.

However, with 19 species, this would be 171 pairwise distances, with many sequences having 20-40 differences. Instead of doing them all, students will fill in the blanks for the closely related species shown in Figure 5 only.

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

Figure 5. Fill in the blanks by counting the number of differences between pairs of aligned sequences. Students may do this individually, or in groups, depending on the numbers of students.

The results should be as follows (minor differences will not matter):

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

Figure 6. Answers for the fill-in-the-blanks exercise from step 5.

6. Hypothesise a phylogeny based on the GULO core alignment distances. Students will attempt to *draw* a phylogeny to demonstrate the evolutionary relationships between the 19 mammal species used above. This can be exploratory and does not have to be perfect. Students use the following questions to guide them.

- Which two species are most similar genetically? (cats and lynx, with 0 differences)
- Which two species are next most similar genetically? (humans and gorillas, with 2 differences)
- Which two species are next most similar? (wombats and koalas, with 3 differences)
- Which two species are next most similar? (the colobus monkeys, 4 differences)
- What species is most similar to the colobus monkey group? (Rhesus macaque, 5 differences)
- Which group/species seems most genetically similar to the human/gorilla/monkey group? (cat/lynx)
- Use similar grouping logic to group possum, koala, wombat (marsupials)
- 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? (yes)

7. Comparing the manually generated phylogenetic tree to a computer-derived phylogenetic tree.

The above process is a rough approximation of a computer algorithm called “neighbour-joining” (NJ), a fast and easy computer algorithm. NJ gives a decent first approximation of a phylogenetic tree. Although more advanced statistical methods are now standard for research, they are all based on the same data (a DNA alignment) and basic logic (shared similarity suggests closer relationship).

Students can now compare the NJ-generated tree below (Figure 7) with their by-hand efforts. The scale bar indicates “number of DNA changes”.

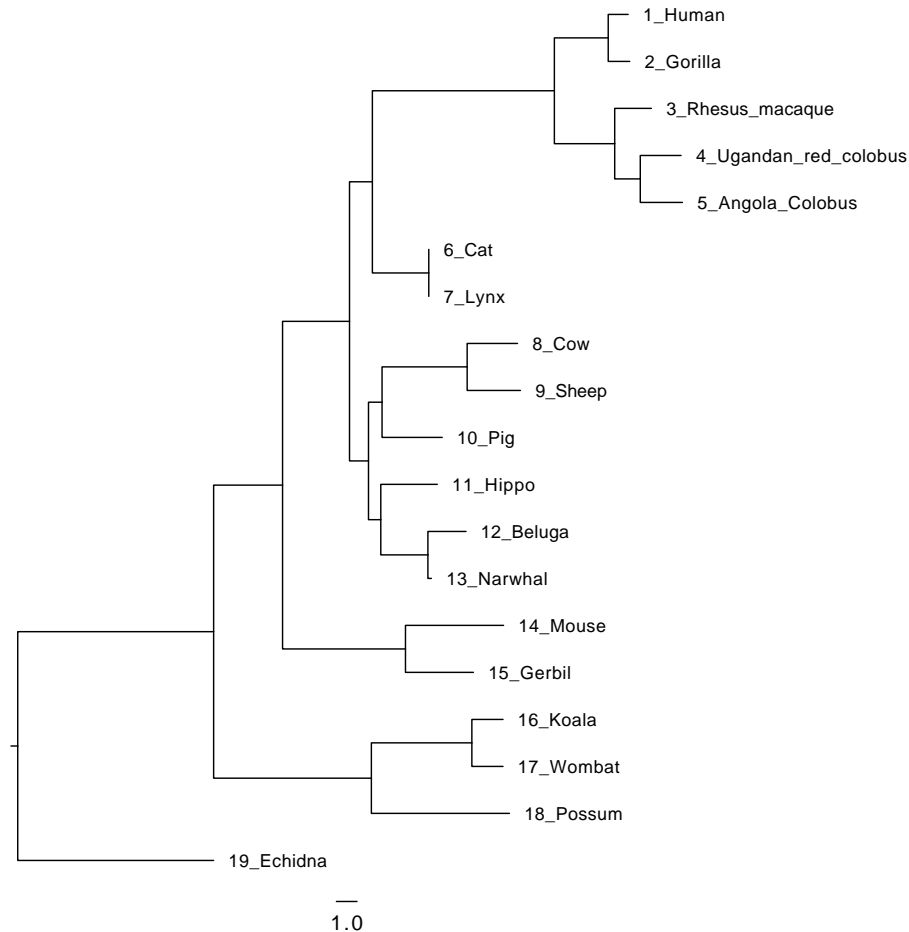


Figure 7. Phylogenetic tree estimated from the genetic differences matrix calculated on the core alignment. This was done on a computer using the Neighbor-joining (NJ) algorithm.

Questions for students:

- The scale bar indicates 1 DNA difference. Only horizontal distances matter in this phylogeny plot. How long are the 2 horizontal branches that connect to “Human” and “Gorilla”? How much would they be if you added them together? Does this match your distance matrix?
 - Answer: each branch is ~1 unit long, total difference of 2, which matches the distance matrix.
- Why do Cat and Lynx appear to not have horizontal branches connecting to them?
 - The sequences are identical (except for a few “-” gaps, which we ignore for counting differences). So the distance between them is 0.
- Are you surprised by the phylogenetic groupings of cow, sheep, pig, hippo, beluga, and narwhal? Should you be? (google it if necessary)
 - This was originally surprising to scientists, but it is now accepted that whales (cetaceans) are a subgroup of artiodactyls (even-toed hoofed ungulates). It turns out that the earliest whale fossils have legs and “hooves” (thick toenails). Strange but true! Famous fossils include *Pakicetus* (Pakistan whale, but it had 4 legs and walked!) and *Ambulocetus* (ambulatory whale, had 4 legs but was much more aquatic). The cetaceans + artiodactyls form a clade (phylogenetic group) now called Cetartiodactyls.

Nick Matzke & Wilda Laux

- All living mammals are divided into 3 groups: monotremes, marsupials, and placentals. Can you see these 3 groups in the *GULO* phylogeny? Do you think other genes/pseudogenes would give a similar phylogenetic tree? Why or why not?
 - Yes, we see those clades. And yes, other genes should give *similar* trees because all of these genes have been evolving in the same species with the same pattern of speciations (gene pool separations). (Note: trees from other genes will not necessarily be *identical*, due to the randomness of mutation, loss of similarity after many mutations, and other processes like hybridization).
- Why do you think the branches leading to the primate group species are longer than for the other mammals?
 - In these primates, the *GULO* gene is a pseudogene, so it has lost functional constraints on its sequence. Thus, there is no natural selection removing harmful mutations, and the sequence is “drifting” rather than being under stabilising selection.

References

De Tullio, Mario C. (2010). “The mystery of Vitamin C.” *Nature Education* 3(9):48.

<https://www.nature.com/scitable/topicpage/the-mystery-of-vitamin-c-14167861/>

Lents, Nathan H.; Cifuentes, Oscar E.; Carpi, Anthony (2010). “Teaching the process of molecular phylogeny and systematics: a multi-part inquiry-based exercise.” *CBE - Life Sciences Education*, 9, 513-523.

<http://dx.doi.org/10.1187/cbe.09-10-0076>

Mansueto, Alexander; Good, Deborah J. (2024). “Conservation of a chromosome 8 inversion and exon mutations confirm common gulonolactone oxidase gene evolution among primates, including *H. Neanderthalensis*.” *Journal of Molecular Evolution*, 92, 266-277. <https://doi.org/10.1007/s00239-024-10165-0>

<https://doi.org/10.1007/s00239-024-10165-0>

TOTA (2024). Long ocean voyages and the problem of scurvy. TOTA / Traditions of the Ancestors. Accessed 2024-07-31. <https://www.tota.world/article/113/>

Acknowledgements

We are indebted to Alexander Mansueto <alexander.j.mansueto@vanderbilt.edu>

Deborah Good <goodd@vt.edu> for providing the sequence data used in Figure 2 of Mansueto & Good (2024). Additional GULO sequence data used in this activity (for hippo, koala, wombat, echidna) was downloaded from NCBI GenBank. NJM is supported by the University of Auckland, NZ RSTA grants 18-UOA-034 and 21-UOA-040, and HFSP RGP0060/2021.