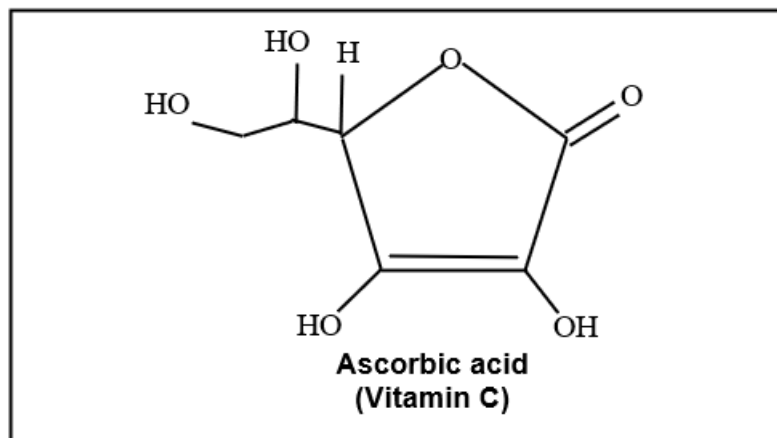


This resource is for educational purposes. For any other use, please consult the writers for permission

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

Student instructions



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

¹School of Biological Sciences, University of Auckland

²Department of Molecular Medicine and Pathology, University of Auckland

Aim: To use genetic information about the *GULO* gene/pseudogene to create a phylogenetic tree showing the evolutionary relationships between a selection of mammal species.

Background information:

The dataset in this activity uses a famous case in genetics, the *GULO* gene/pseudogene. *GULO* stands for L-gulonolactone oxidase, an enzyme involved in the biosynthesis of Vitamin C (ascorbic acid), a vital nutrient. Lack of Vitamin C in the diet 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, and “lassitude,” a lack of mental and physical energy.

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?

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.

We are using part of the *GULO* gene/pseudogene that remains in the genome, called Exon 12.

Methods:

1. Materials.

- An A3 sheet of paper copy with the unaligned DNA sequences of exon 12 of the *GULO* gene
- Scissors
- A piece of clear A3 paper to tape the sequences onto
- Tape

2. Align the DNA strips.

- Cut the strips of DNA sequences from the A3 sheet by cutting along the dashed lines. Each strip is a DNA sequence of exon 12 of *GULO/GULOP* from a different mammal species. You will have 19 strips.
- By hand, slide the sequences back and forth against each other until you are able to match the sequences.
- This soon gets hard to do without disturbing other paper strips, so tape your strips onto the clear A3 sheet to keep the strips in place.
- Question: Is there a pattern in the DNA data? Describe what you see.

3. Highlight the core alignment. Once aligned, these sequences have some leading and trailing sequence that we will ignore – the phylogenetic information is in the shared sequence.

- The “core alignment” corresponds to positions 75 through 197 on the Human sequence. This starts with ACTGTACCTCAAAGAA..., and ends with the end of the Human sequence (...CTACTGA).
- Mark the core alignment (e.g. indicate where it starts and stops by drawing arrows or a box)

4. Analyse the core alignment. Don’t worry about the extra sequences at the beginning/end that do not exactly match between the strips. Focus on the “core alignment” of exon 12.

- Which sequence seems most different from the others?

-
- Which pair of sequences seem most similar? Do the differences in each column appear randomly, or do they seem to have some structure?

5. In depth analysis: Counting differences between pairs of sequences. While scientists would use a computer to do this, it is not hard to count the number of differences between sequences by hand. This is done and recorded in what we call a *genetic distances matrix*.

- Fill in the blanks in the matrix below by counting the number of differences between pairs of sequences. (The teacher will decide which species you will look at. The teacher will also demonstrate how you will do this.)

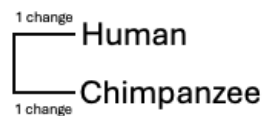
	Human	Gorilla	Rhesus_macaque	Ugandan_red_colobus	Angola_colobus	Mouse	Mongolian_gerbil	Cat	Lynx	Pig	Cow	Sheep	Hippo	Beluga	Narwhal	Possum	Koala	Wombat	Echidna
Human	0					27	25	16	16	18	23	22	19	17	17	32	35	35	38
Gorilla		0				25	25	16	16	18	23	22	19	19	17	32	35	35	37
Rhesus_macaque			0			28	27	17	17	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 1. Genetic distance matrix for 19 mammal species using information obtained for the *GULO* gene/pseudogene.

6. Hypothesise a phylogeny based on the *GULO* core alignment distances. Now, you will attempt to *draw* a phylogeny. This is exploratory, and does not have to be perfect.

The idea is to start by connecting the closest relatives first. The closest relatives will share the lowest number of differences as seen in a genetic distances matrix.

Let's say for example, if humans and chimpanzees have the most similar DNA sequences, and differ at 2 DNA positions over the 122 DNA bases in the core alignment. This would be seen by a score of '2' in a matrix between human and chimpanzee. We could draw their phylogeny like this:



If the next closest DNA sequence from human and chimpanzee is that of orangutans, and orangutans have '4' differences from both humans and chimps, we could add to the tree like this:



You can see that, by “walking along the branches”, the total number of differences between humans-orangutans is 4, and between chimps-orangutans is also 4.

Now, it’s your turn to use the results from the 19 species you have gathered to try to draw a phylogeny to show the evolutionary relationships between them.

As the number of DNA changes increases beyond a few, it will be hard to impossible to figure out exactly what length of branch is appropriate (a computer would use a complex averaging algorithm). We are just going to consider the “big picture” for the purpose of this activity. Note that more DNA difference equals longer branches.

More importantly, we should be able to figure out the grouping information from the DNA – which species are most similar to which other species?

Questions to guide you:

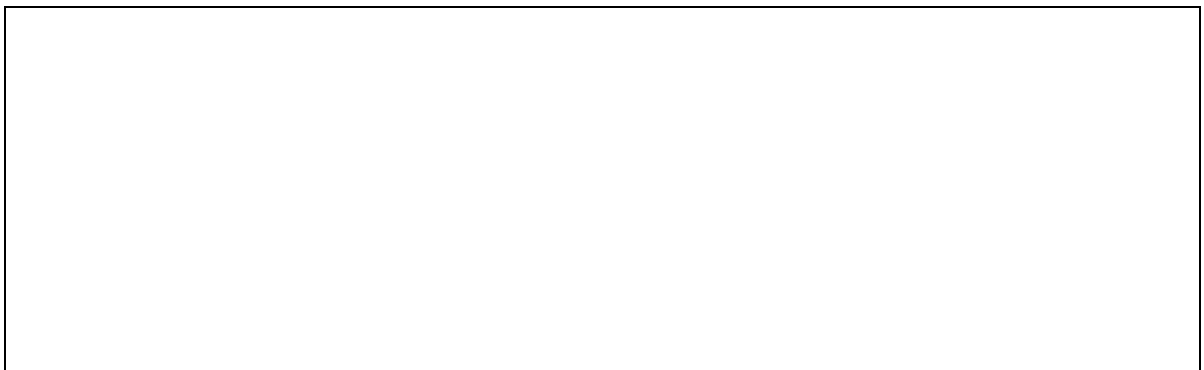
- Which two species are most similar genetically?
- Which two species are next most similar genetically?
- Which two species are next most similar?
- Which two species are next most similar?
- What species is most similar to the colobus monkey group?
- Which group/species seems most genetically similar 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?

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

The teacher will give you a copy of the phylogenetic tree produced by running a “neighbor-joining” algorithm on the DNA genetic distances from the core alignment. The scale bar indicates “number of DNA changes”. Compare this tree to your by-hand efforts.

Questions:

- How does your tree compare to the computer-generated tree? What are the major similarities and differences?

A large, empty rectangular box with a black border, intended for the student to write their answer to the question about comparing their tree to the computer-generated tree.

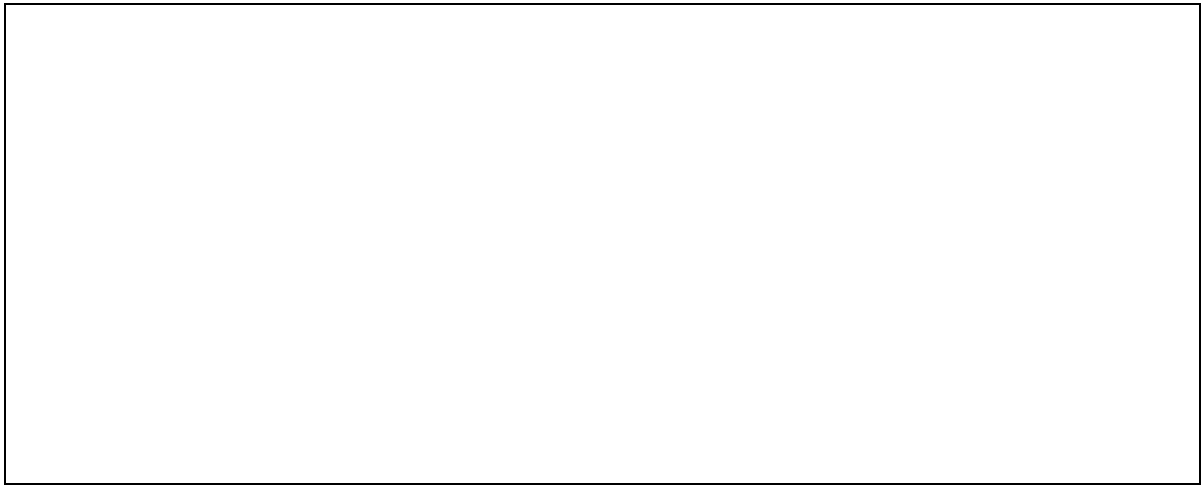
- The scale bar indicates 1 DNA difference. Only horizontal distances matter in this phylogeny plot. How long are the 2 horizontal branches that connect to “Human” and “Gorilla”? How much would they be if you added them together? Does this match your distance matrix?

- Why do Cat and Lynx appear to not have horizontal branches connecting to them?

- Are you surprised by the phylogenetic groupings of cow, sheep, pig, hippo, beluga, and narwhal? Should you be? (google it if necessary)

- 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?

- Why do you think the branches leading to the primate group species are longer than for the other mammals?



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>

Neanderthalensis." *Journal of Molecular Evolution*, 92, 266-277. <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.