Tracing the L-gulonolactone oxidase gene

BB226659

Introduction

Vitamin C, also known as ascorbic acid (AsA), is a necessary supplement that is involved in protecting cells, therefore responsible for keeping the skin, blood vessels, and cartilage healthy (NHS.uk, 2020) it also involved collagen biosynthesis in most mammals. While most animals are able to synthesise their own AsA from d-glucose with help of a particular enzyme (L-gulonolactone oxidase (GULO)) (Yang, 2013). Vertebrate GULO genes are highly conserved and share similar intron-exon patterns, but previous research found that sequencing analysis of the non-functional GULO gene from humans indicates that the gene has accumulated a substantial amount of insertion and deletion mutations and lost its function during the evolution (Nishikimi, 1994). A lot of primates and guinea pigs also lack the functioning GULO antioxidant and therefore must rely on Vitamin C amassed through their diet from exogenous sources (Nishikimi, 1992).

Basic Local Alignment Search (BLAST) is a program used to check for similarities between a supplied query and a nucleotide or protein sequence that is stored in the database (subject sequence) (Wheeler, 2007). The tool then calculates the statistical significance for each match, so the user can determine whether the results might have appeared by coincidence and choose only the ones that are of interest.

In this study, we will use a mouse GULO DNA coding sequence to perform blast searches for various species and attempt to figure out if a species lack the functioning gene and attempt to support our findings with scientific literature.

Materials & Methods

Running BLAST searches

For this study, we used the Mouse DNA coding sequence for the L-gulonolactone oxidase gene as a FASTA query, which was provided by the University of Edinburgh (table 1)

Three different tools: BLASTP, BLASTN and TBLASTN were used from the NCBI BLAST website on 5 following vertebrate species to determine whether these species have a functioning GULO gene. Details of the default (starting parameters) blast searches are described in table 2.

- *Homo sapiens* (taxid: 9606),
- *Xenopus tropicalis* (taxid: 8364),
- *Elephas maximus* (taxid: 9783),
- *Cavia porcellus* (taxid: 10141),
- *Branchiostoma lanceolatum* (taxid: 7740)

Mouse DNA coding sequence for the GULO gene
>Sequence_R ATGGTCCATGGGTACAAAGGGGTCCAGTTCCAAAACTGGGCGAAGACCTATGGCTGCAGTCCAGAGATGT ACTACCAGCCCACATCAGTGGGGGAGGTCAGAGAGGTGCTGGCCCTGGCCCGGCAGCAGAACAAGAAAGT GAAGGTGGTGGGTGGCGGCCACTCGCCTTCAGACATCGCCTGCACCGATGGCTTCATGATTCACATGGGC AAGATGAACCGGGTTCTCCAGGTGGACAAGGAGAAGAAGCAGGTCACAGTGGAAGCCGGTATCCTCCTGA CTGACCTGCACCCACAGCTGGACAAGCATGGCCTGGCCCTGTCTAATCTGGGAGCCGTGTCTGATGTGAC GGTTGGTGGCGTCATTGGGTCTGGAACACATAACACCGGGATCAAGCACGGTATCCTGGCCACCCAGGTG GTGGCCCTGACCCTGATGAAGGCTGATGGAACAGTTCTGGAATGTTCTGAGTCAAGTAATGCAGATGTGT TCCAGGCTGCAAGGGTGCACCTGGGCTGCCTGGGTGTTATCCTCACTGTCACCCTGCAGTGTGTGCCACA GTTCCACCTTCTGGAGACATCCTTTCCTTCGACCCTCAAGGAGGTCCTTGACAACCTGGACAGCCACCTG AAGAAGTCTGAGTACTTCCGCTTCCTCTGGTTTCCTCACAGTGAGAACGTCAGCATCATCTACCAAGATC ACACCAACAAGGAGCCCTCCTCTGCATCTAACTGGTTTTGGGACTATGCCATTGGGTTCTACCTCCTGGA ATTCTTGCTCTGGACCAGCACCTACCTGCCACGCCTCGTGGGCTGGATCAACCGCTTCTTCTTCTGGCTG CTGTTCAACTGCAAGAAGGAGAGCAGCAACCTCAGCCACAAGATCTTCTCCTACGAGTGTCGCTTCAAGC AGCATGTCCAAGACTGGGCCATCCCCAGGGAGAAGACCAAGGAGGCCCTGCTGGAGCTAAAGGCCATGCT GGAGGCCCACCCCAAGGTGGTAGCCCACTACCCCGTGGAGGTGCGCTTCACCCGAGGTGATGACATCCTG CTGAGCCCGTGCTTCCAGAGGGACAGCTGCTACATGAACATCATTATGTACAGGCCCTATGGGAAGGATG TGCCTCGGTTGGATTACTGGCTGGCCTATGAGACCATCATGAAGAAGTTTGGAGGCAGGCCCCACTGGGC AAAGGCCCACAATTGCACCAGGAAGGACTTTGAGAAAATGTACCCCGCCTTTCACAAGTTCTGTGACATC CGCGAGAAGCTGGACCCCACTGGAATGTTCTTGAATTCGTACCTGGAAAAGGTTTTCTACTAA

Table 1.

The table presents the FASTA formatted Mouse DNA coding sequence for the GULO gene, which was used for all the BLAST alignments in this study.

The algorithm:	**Version:**	**Molecule Type:**	**Database**	**Word size:**	**Gap costs:** (Existence, Extension)
BLASTP (protein-protein)	(2.13.0+)	Protein	All non-redundant GenBank CDS translations (PDB, SwissProt, PIR, PRF)	6	11,1
BLASTN (nucleotide-nucleotide)	(2.13.0+)	Mixed DNA	Nucleotide collection (nr/nt) (GenBank, EMBL, DDBJ, PDB, RefSeq)	11	5,2
TBLASTN (protein-nucleotide)	(2.13.0+)	Mixed DNA	Nucleotide collection (nr/nt) (GenBank, EMBL, DDBJ, PDB, RefSeq)	6	11,1

Table 2. Short description of each of the initial/default blast search programs used in this study and their key parameters. The parameters were later adjusted for optimization, described further in the results section.

In order to run the Blastp query on the coding sequence, we first needed to translate the nucleic acid sequence into its corresponding peptide sequence, this was done with the EMBL's software EMBOSS Transeq (Madeira, 2022). Blastp searches were carried out using the BLOSSUM62 matrix and conditional compositional score matrix adjustment was applied to compensate for the sequence composition of amino acids (Yi-Ku, 2005), thus providing us with more accurate E-values. All the blast sequences were run with the low-complexity filter 'on' in order to avoid 'bogus' hits and the 'expect value' cut-off limit was set to 0.05. As we can see from Table 2, most of the searches were run on a collection of databases, this was done to speed up the process and the very definitive and strong matches would be returned even with a large test size (no. of sequences tested). Only in the cases where the results were inconclusive or did not appear to be detected, we tried specific databases, to increase the E-value of the hits (by lowering the number of sequences tried, leading to fewer false positives) we were looking to find the most informative results. This is well explained by the following formula ($E = Kmne^{-\lambda S}$, where K is the database size)

While we are aware of the updated new blastn program called MegaBLAST, which is "about 10x faster than blastn" (Wheeler, 2007). However, it is only useful for nearly-identical sequences and considering

our work only demanded a short number of searches and time was not a limiting factor we decided to continue with a variation of blastn and discontiguous megablast searches in order to have more results to analyse and consider.

Analysing blast results

To make our results more reproducible we have recorded maximum and total bit scores, E-values, and accession numbers for all the hits in order to make our results reproducible. As expected not all blast searches gave us matches, which we interpreted as sufficient statistical evidence to suggest that there is no homology between the query and the species genome. We then focused on the highest scoring hits which produced the lowest E-values, but also analyzed the query cover, and percentage identity as well as visually assessed the coding sequence (CDS) to determine if the gene is functioning or not.

Results

Homo sapiens

BLAST	Accession:	E value:	Maximum Score:	Total Score:	Query coverage (%)	Percentage identity (%)
Blastp	NP_055577.1	9e-12	68.2	68.2	28	26.69
Blastn	NG_001136.2	2e-44	189	425	32	85.98
Tblastn	NG_001136.2	2e-10	70.5	272	39	78.05

When the blastp search was run on the default settings described in the methods using BLOSUM62 matrix, no hits were detected. The search parameters were then relaxed/optimised and BLOSUM45 matrix was used (gapcosts 15,2) and this search returned a lot of hits, the best match (Accession (Acc.) NP_055577.1) had an E-value of 3e-11 and total score of 68.2, but a very low Query Coverage. To remove some of the less likely hits and further optimise the search, a Model Organism database was used. This returned the same match but further decreased the E-value to 9e-12. The Query coverage of that hit was very low (28%), with as low percentage identity (%id) of 29.69%. It was therefore decided to ignore this and assume that there were no statistically significant matches.

The blastn search was much more promising. While using the default parameters a good match with GULOP pseudo gene was detected (Acc. NG_001136.2). The search was optimised with a discontiguous megablast and returned the same match with a much higher confidence (E-val of 2e-44), but the query coverage was lower (32% vs 40% with traditional blastn).

For tblastn we used the default parameters and got a match with the same GULO pseudogene (as in blastn), but with a relatively low query coverage (25%), but a relatively high E-value. Just to see the alignment better, the BLOSUM45 matrix was used to lower the threshold and sacrifice the statistical significance in order to increase the coverage to 39%, while keeping the %id the same. Because the same hit matched with a query as in blastn, there was no need to further increase our E-value by choosing a specific database etc.

During blastn and tblastn searches, there were other significant matches with the GULO genes on various chromosomes but all of these hits had low percentage coverage. When looking at the alignment between the human GULO gene and the sequence for a functioning GULO gene, we can see how exons have been mutated and this suggested that the gene has lost its function. The verdict for *homo sapiens* is that there is a GULO pseudogene and therefore it is not functioning in humans. This is supported by the findings in the paper by Hongwen Yang (2013), which states that humans have lost over half of the original 11 exons and the 6 ones that are still present are very poorly conserved with multiple substitutions, indel mutations and premature stop-codons.

Xenopus tropicalis

BLAST	Accession:	E value:	Maximum Score:	Total Score:	Query coverage (%)	Percentage identity (%)
Blastp	XP_031758963	0	659	659	99	71.82
Blastn	NC_030681.2	6e-22	112	427	42	74.56
Tblastn	XM_031903103.1	0	659	659	100	69.61

Blastp default search parameters worked optimally and returned one strong match (Acc. XP_031758963) with an E-value of 0 and no gaps.

Tblastn on default parameters returned one very strong match with 100% query cover (QC) and 70%id (Acc. XM_031903103.1). However, it was tagged as 'predicted' and although the genomic correction is not something to be afraid of and most likely it has been only done where appropriate, we still wanted to try out the gold-standard sequences stored in the RefSeq databases. This returned a statistically significant match with 92% QC and %id of 62.5% and after some extra research it was established that the predicted sequence found earlier was derived from exactly this Nigerian strain of Xenopus tropicalis chromosome 5 (Acc. NC_030681.2).

As expected, the blastn search returned very similar hits as tblastn, but to make our conclusion more supportive, we chose the top-level sequence from the RefSeq genomes database, Acc. NC_030681.2. When looking at the alignment we saw that there were 5, well-conserved coding regions, with high identity varying from 73-80% with no gaps in them

Previous research states that homologous sequences that match with more than 40% identity are widely assumed to have maintained the functional resemblance, which would suggest that our findings suggest that the GULO gene is present and functional in the *Xenopus tropicalis* (Pearson, 2013). Other research also confirms that a close polyploid homologue – African clawed frog (*Xenopus laevis*) has a functioning GULO gene and it allows the species to synthesise AsA in their kidneys. (Yajun Xie, 2016). In conclusion, it is difficult to say with certainty if the gene is functioning or not, but with the support of our results and literature we can assume that *Xenopus tropicalis* possesses a putative gene of GULO and the sequence similarities can be a sign of a similar function gene.

Elephas maximus

BLAST	Accession:	E value:	Maximum Score:	Total Score:	Query coverage (%)	Percentage identity (%)
Blastp	XP_049721450.1	0	827	827	99	91.44
Blastn	NC_064840.1	1e-67	265	1597	94	94.67
Tblastn	XM_049865493.1	0	827	827	100	87.76

The blastp searches on the PDB and SwissProt databases returned no hits with either BLOSUM62 or 45 matrices. As it turns out the default setting was the most optimal choice and returned a very strong match which was 90% identical over 99% query coverage, which is a strong sign of a fully functional gene. When combined with the results from blastn and tblastn we were already certain that the Asian elephant is certainly capable of carrying out the terminal step in Vitamin C production. Blastn program was optimised to a stricter search using the RefSeq genome database and returned a strong hit. It made 10 matches with identities varying from 82-95%. Because of the wobble in the last amino acid of the triplet, this makes the match seem even more significant and we can take this blastn result as another strong

suggestion that the species has a functioning GULO gene. Tblastn search led us to the same conclusion by giving us a hit with a “predicted” mRNA covering 100% of the query with 88% identity. That sequence was derived from the older version of the match recorded in blastn (Acc. NC_064840).

The research suggests that scurvy, a disease caused by Vitamin C deficiency, is very rare and Elephant milk contains 4 times the AsA level when compared with cow’s milk (Dhairykar, 2020), despite the fact that cows can synthesise vitamin C themselves, both in the liver and kidney (Weiss, 2019). Another argument that could aid us in determining whether the gene is active in the Asian elephant, is their diet, which mainly consists of tree bark, root, and leaves (Koirala, 2016). Unfortunately, this does not help us in reaching the conclusion as tree bark and leaves are a great source of vitamin C.

Because of all the factors above, the near-perfect match of *Elephas maximus* sequence to the mouse GULO coding sequence alignment is a significant enough sign that the Asian elephant species contains a functional GULO gene and therefore is able to synthesise its own Vitamin C, which could be an explanation of why the milk content is so rich with AsA.

Cavia porcellus

BLAST	**Accession:**	**E value:**	**Maximum Score:**	**Total Score:**	**Query coverage (%)**	**Percentage identity (%)**
Blastp	XP_012998768.1	0	581	581	99	65.65
Blastn	XM_013143314.2	0	821	1231	86	85.28
Tblastn	NT_176418.1	7e-19	95.9	610	75	84.00

BLOSUM62 matrix only returned one species when the search was run either against the RefSeq database or the default mixture of databases. BLOSUM45 matrix was also tried but returned a similar result with extra new hits which were not statistically significant and therefore discarded. The most prominent hit (Acc. XP_012998768.1) was chosen to study further. It is a corrected sequence tagged with a “low-quality protein” tag, therefore despite the great qc of 99% and relatively high %id of 66%, we had to investigate the alignment of blastn to check if there was a frame shift caused by a deletion of insertion, which could result in a loss of function. The results so far only show that the homology between the sequences is significant.

Blastn search was initially run on the RefSeq database and returned a strong hit (Acc. NT_176418.1), but the search was then tried on the default setting (as described in the methods section) and gave us a predicted sequence, which we will use to determine the conclusion. We chose to consider this hit because of the high score, qc and %id statistics. However, by studying the alignment we could note a vast number of errors and frame shifts between the matches, which would suggest that a larger part of the subject sequence, will not be translated correctly in the Guinee pig.

Tblastn search when run on a RefSeq database found the same initial hit as blastn search, which matched 9 different short sequences of a GULO-like protein, the matches had been found with varying frame shifts and low bits scores, which are a better way to standardise the score when sequences of varying lengths are being analysed. The statistical significance was lower, because of the higher number of tests run. All of that further suggests that a gene has undergone a significant number of mutations and could infer that the gene function has been lost.

The scientific literature has a vast number of studies that confirm our hypothesis. The absence of the activity is caused by a big, intragenic deletion or parts of exons 8 and introns 7 and 8 causing a frame shift (Lara Hasan, 2004). This supports our hypothesis and conclusions reached while studying the alignments. Other studies have shown that diet-induced illnesses were common in *Cavia porcellus* due to

a deficiency of AsA (Pernille Tveden-Nyborg, 2016). We, therefore, have concluded that despite close homology with the functioning GULO gene, guinea pigs have lost that functionality during their evolution.

Branchiostoma lanceolatum

BLAST	Accession:	E value:	Maximum Score:	Total Score:	Query coverage (%)	Percentage identity (%)
Blastp	CAH1255212.1	3e-158	488	488	99	49.77
Blastn	OV696687.1	4e-06	58.1	223	26	72.73
Tblastn	OV696687.1	2e-09	63.9	427	77	66.67

None of the top-quality protein databases (RefSeq, PDB) were able to provide us with any significant hits for the blast searches even with very relaxed parameters. When running the default database set with BLOSUM80 matrix on the blastp search, we were able get a few good hits, but all with relatively low %id. The blastn search returned a very poor hit (Acc. OV696687.1) on the default setting, therefore the parameters were relaxed and the word size of 7 was chosen to see if we can get any other hits, which unfortunately was not the case, and we were left with a hit which only matcher 26% of our query length (on very poor parameters). Tblastn search on the default parameters returned three hits, of which one was the same one we managed to match in blastn and had the best QC (46%) and %id (67%). They were sequences from 3 different chromosomes of *Branchiostoma lanceolatum*, but the interesting thing was that the match sequences were overlapping on different chromosomes, this is common in viruses and while some cases can be found in bacteria or eukaryotes, it is very rare (Tomohiro Nakayama, 2007).

This invertebrate amphioxus is one of the closest relatives to vertebrates still found in shallow seas near Norway (MICHAEL FUENTES, 2007). Because it is a relatively small animal its metabolic rate can be assumed as relatively high, which would suggest that they would require proportionally more vitamins than larger animals. There is a lack of new research concerning this species, so not much is known about their diet, besides the fact that they are classified as 'filter-feeders' (Svane, 1999). Research from the late 20th century suggests that phytoplankton and zooplankton are the main sources of nutrition, which are known to be relatively low in Vitamin C (A. M. Hapette, 1990).

Because of the poor alignment against a mouse GULO gene it is very difficult to conclude that *B. lanceolatum* can synthesise AsA itself. Even though a close ortholog (*Branchiostoma belcheri*) has a GULO-like gene, we cannot assume that it is functional. Therefore, the results are inconclusive.

Discussion

The purpose of this study was to use various blast searches and alignment information to determine whether 5 different have a functioning GULO gene necessary for the synthesis of Vitamin C. We have concluded with certainty that *Elephas maximus* has a functioning gene, as nearly identical sequences were found in their genome. *Xenopus tropicalis* has a putative GULO gene and this was determined by the fact that a sequence with a full query coverage and %id of over 70% were found in its genome, which infers that this species should be able to produce an enzyme with similar functionality. *Homo sapiens* and *Cavia porcellus* species had even worse alignments with the query, which suggested that the gene has amassed a substantial number of mutations which eliminated its functionality. Because the blast results could not give us definitive and binary answers in the last 3 cases described, we had to rely heavily on the literature. It is important to consider that humans and guinea pigs are the model organisms in studying the GULO gene. Therefore, their sequences and evolution have been studied to a greater depth than for example frogs. This might have had a biased impact on our conclusion. The least studied species that we

were working with was the European lancelet, it is impossible to conclude with 100% certainty, but it is very unlikely that a function of the GULO gene could have survived this number of mutations.

The cladogram presented in Figure 1 is a good depiction of the assumed phylogeny between the species. It supports most of our conclusions and is a nice way to summarise our results. It highlights how *H. sapiens* and *B. lanceolatum* have evolved furthest from the functioning gene, therefore suggesting that the pseudogenes present in these species are no longer functional. It also shows that the mutations acquired by *C. porcellus* have been the most dominant as it put that descendent on a separate branch and we know that it has lost its function, because of well-respected and peer-reviewed literature.

Overall, this study highlighted the strengths and weaknesses of the BLAST tool for determining the loss of function of a specific gene and showed the thought process behind how we can choose the parameters and use the data to trace the evolution of a specific gene.

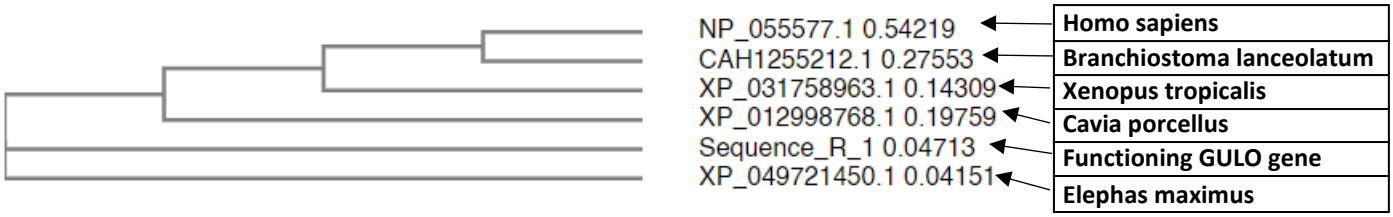

Figure 1. Phylogenetic Tree (cladogram) created with a Clustal Omega alignment tool (EMBL-EBI). It depicts the evolutionary descent of different species from common ancestor based on the top scoring hits of the blastp searches

References

A. M. Hapette, S. A. (1990). Variation of vitamin C in some common species of. *MARINE ECOLOGY PROGRESS SERIES*, 69-79.

Dhairykar, M. (2020). Management of nutrition in captive Asian elephants. *International Journal of Veterinary Sciences and Animal Husbandry*, 160-163.

Koirala R.K., R. D. (2016). Feeding preferences of the Asian elephant (Elephas maximus) in Nepal. *BMC Ecol*, 16, 54. doi:https://doi.org/10.1186/s12898-016-0105-9

Lara Hasan, P. V. (2004). Intragenic deletion in the gene encoding L-gulonolactone oxidase causes vitamin C deficiency in pigs. *Mamm Genome*, 323-333.

M Nishikimi, R. F. (1994). Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *J Biol Chem.*

M Nishikimi, T. K. (1992). Guinea pigs possess a highly mutated gene for L-gulono-gamma-lactone oxidase, the key enzyme for L-ascorbic acid biosynthesis missing in this species. *J Biol Chem*.

Madeira F, P. M. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*.

MICHAEL FUENTES, E. B. (2007). Insights Into Spawning Behaviorand Development of the European Amphioxus(Branchiostoma lanceolatum). *JOURNAL OF EXPERIMENTAL ZOOLOGY (MOL DEV EVOL)* , 484-493.

NHS.uk. (2020, 8 3). *Vitamin C*. Retrieved from https://www.nhs.uk/: https://www.nhs.uk/conditions/vitamins-and-minerals/vitamin-c/

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Curr Protoc Bioinformatics*.

Pernille Tveden-Nyborg, M. M. (2016). Diet-induced dyslipidemia leads to nonalcoholic fatty liver disease and oxidative stress in guinea pigs. *Translational Research*, 146-160.

Svane, H. U. (1999). Filter Feeding in Lancelets (Amphioxus), Branchiostoma lanceolatum. *Invertebrate Biology*, 423-432.

Tomohiro Nakayama, S. A. (2007). Overlapping of Genes in the Human Genome. *Int J Biomed Sci*, 14-19.

Weiss, B. (2019). Update on Vitamin Nutrition of Dairy Cows. *The Ohio State University*, https://dairy-cattle.extension.org/update-on-vitamin-nutrition-of-dairy-cows/#:~:text=The%20liver%20and%20kidney%20of,vitamins%20to%20prevent%20clinical%20deficiency.

Wheeler D, B. M. (2007). BLAST QuickStart: Example-Driven Web-Based BLAST Tutorial. *Comparative Genomics: Volumes 1 and 2.*, Chapter 9.

Yajun Xie, Y. L. (2016). Gulo Acts as a de novo Marker for Pronephric Tubules in Xenopus laevis. *Kidney Blood Press Res*, 794-801.

Yang, H. (2013). Conserved or Lost: Molecular Evolution of the Key Gene GULO in Vertebrate Vitamin C Biosynthesis. *Biochemical Genetics*, 413-425.

Yi-Kuo Yu, S. F. (2005). The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, 21(7):902-11.