**L-gulonolactone Oxidase (GULO) Expression: A Bioinformatic Exploration of L-gulonolactone Oxidase (GULO) protein in mammals**

Hameed Sanusi

Department of Biology, Georgia State University

Advisor: Dr. Eric Gaucher

Reader: Dr. Eric Gaucher, Dr. Jessica Carter

April 19th, 2024

**Table of Contents**

List of Figures and Tables 4

Acknowledgments 5

**Introduction** 6

     1.1Importance of Vitamin C 6

     1.2 **Vitamin C Synthesis and L-gulonolactone Oxidase (GULO)** 7

     1.3 Evolutionary Loss of GULO in Mammals 9

     1.4 Bioinformatics analysis of GULO mutation 11

1.5 Research Objectives 13

1.6 Hypothesis 13

**Materials and Methods** 14

* 1. GULO Protein Sequence Collection 14
  2. Multiple Sequence Alignment 14
  3. Phylogenetic Analysis of GULO sequences 15
  4. Reconstruction of Ancestral Sequences 15

**Results and Discussions** 14

* 1. Phylogenetic Analysis 1
  2. Reconstruction of Ancestral Sequences with Amino Acid Replacement 19

**Discussion and Conclusion** 21

**References**

**List of Figures and Tables**

Figure 1. A pathway showing the biosynthesis of ascorbic acid (Vitamin C) …...……………...8

Table 1:Parameters used for the phylogenetic analysis using PAML………………………….12

Figure 2: A collapsed phylogram showing the evolutionary relationship of several mammalian species GULO protein sequences.....………………………………………………………….....14

**Acknowledgments**

بِسْمِ ٱللَّٰهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

In the name of God, the most gracious, the most merciful.

First and foremost, I express my deepest gratitude to Dr. Eric Gaucher, who has been a terrific mentor throughout the course of this project. I walked into your doors ignorant of how to approach scientific problems and I can say that I am more confident about finding a place to advance the frontiers of science. His guidance and the materials he provided ensured the smooth completion of this work.

To Dr. Jesscia Carter, you have been an amazing mentor, and I am truly grateful for your help in helping me find a place in science.

At the later stages of the project, I am grateful for the input of Nuria Ramirez and Allison for their help in trying to provide context to some of the results I was presenting at our meetings. To our wonderful lab manager Lily, I am grateful for your help in helping get situated into the lab.

To my parents, Mr and Mrs. Sanusi, you have done the best in raising me, I cannot thank you enough for your constant prayers and ensuring that I get the best education.

To my Uncle, Mukaram Enifeni, thank you for being a shoulder of support and an enabler of dreams.

To my dear siblings, especially, my sister Azeezat, thank you for listening to my rants and the continuous motivation to never give up.

To my beloved, Zainab, love gives purpose, a reason to fight, to stand tall, and to never give up. I cannot appreciate enough, your presence in my life.

I am truly grateful to my friends, Moshood, Funmi B, Oreoluwa and Abdul Hameed, your emotional support throughout this period of carrying out this project is invaluable.

To everyone mentioned above, and to all the others who have supported me along the way.

**1.0 Introduction**

**1.1 Importance of Vitamin C**

Vitamins are defined as organic compounds that are essential for growth and development and are usually required in trace quantities in diets.  As an organic compound, vitamins are distinct from the major macromolecules – fats, carbohydrates, and proteins (Combs and McClung, 2019). It occurs naturally in components of food, typically present in minute quantities, yet crucial for normal physiological function, including maintenance, growth, and development. Notably, its deficiency can lead to a specific syndrome when the organism cannot synthesize it in sufficient amounts to meet its needs or uptake it from diets (Combs and McClung, 2019).

Vitamins can be sourced from plants, dietary supplements and animal food products.  This inability to synthesize in sufficient quantities, or at all, necessitates a dietary intake (Drouin *et al.,* 2011). This highlights the organism's dependence on specific enzymes for vitamin biosynthesis, and the potential consequences of their absence or limited activity. Vitamin C (ascorbic acid) is one example.

Ascorbic acid possesses various biological functions, many of which involve its redox properties (Combs and McClung, 2019), positioning it as a key player in cellular antioxidant defense (Combs and McClung, 2019). Additionally, maintaining adequate ascorbic acid levels in the body is essential for collagen, catecholamine, and carnitine biosynthesis (Michels, 2012). Ascorbic acid is the primary water-soluble antioxidant in plasma and tissues which allows to participate in the contribution of redox recycling of α-tocopherol, enhance the bioavailability of non-heme iron, and maintain optimal oxidation states of enzyme-bound metals for proper enzymatic function (Combs and McClung, 2019). As a result, compromises in these processes are known contributors to the onset of vitamin c deficiency.

Higher primates such as tarsier, orangutans, apes, humans belonging to the *Haplorhini* order and humans as well as guinea pigs are among the mammalian species that are unable to synthesize Vitamin C and consequently rely on their diets to obtain this vitamin to prevent scurvy – a pathophysiological onset because of Vitamin C deficiency (Nishikimi *et al*., 1994).

**1.2 Vitamin C Synthesis and L-gulonolactone Oxidase (GULO)**

The synthesis of vitamin C in mammals from glucose occurs via the glucuronic acid pathway (Combs and McClung, 2019). In mammals, the liver is responsible for the synthesis of vitamin C while the kidney is responsible for its synthesis in fishes, amphibians, older bird orders and reptiles. Interestingly, there has been change in the organ responsible for the synthesis of vitamin C twice during evolution: one from the kidney to the liver in mammals and in a similar manner in birds (Drouin *et al.,* 2011). For egg-laying mammals, ascorbic acid synthesis of vitamin C occurs only in their kidneys, and many marsupials can use both their liver and kidneys for vitamin C synthesis (Combs and McClung, 2019). This switch across organs has been explained to be because of selective pressures to balance biochemical homeostasis under stressful conditions (Drouin *et al.,* 2011).

In the pathway shown in figure 1 below, GULO is the protein that catalyzes the last step in the biosynthesis of Vitamin C. A mutation or disruption of the expression of GULO results in a deficiency of Vitamin C production meaning that, the conversion of L-gulonolactone to L-ascorbic acid will not happen.

A diagram of a chemical reaction

Description automatically generated

**Figure 1**: A pathway showing the biosynthesis of ascorbic acid (Vitamin C).

**1.3 Evolutionary Loss of GULO in Mammals**

Despite the biological significance of Vitamin C, mammalian species such as primates, some bat species and guinea pig are unable to produce it and this is due to a loss of a functional GULO gene that has become pseudogene (Henriques *et al*., 2019). Pseudogenes are described as DNA segments that contain a high degree of sequence similarity to functional genes but are non-functional (Zhang *et al*., 2019). Through sequence comparison of lineages that have function and non-functional GULO gene, the inactivation dates of GULO have been placed at 14 and 61 MYA respectively for guinea pig and primate genes (Lachapelle & Drouin, 2011).  Pseudogenes can contain disruptive effects, and the pseudogenization of a gene that was once functional is likely caused by a single mutational event. Such events of loss and regaining of function in GULO have been reported in some species of bats and fishes where GULO gene that were once thought to be nonfunctional have regained function (Cui *et al.,* 2011; Cho *et al.,* 2007). These events can introduce premature stop codons, disrupt splice junctions, cause frameshifts in the coding sequence, or impair the functionality of transcriptional regulatory elements (Zhang *et al*., 2019).

Pseudogenes particularly valuable for functional studies is unitary pseudogenes of which GULO is an example. These are unprocessed pseudogenes lacking functional counterparts within the same genome that arise from disruptive mutations in functional genes, rendering them incapable of being transcribed or translated successfully (Zhang *et al*., 2019).

**1.4 Bioinformatics analysis of GULO mutation**

Bioinformatics analysis describes the process of using computational technology to retrieve, process, analyze and predict biological information. Usually, the first step involved is collecting sequences of interest from biological data banks depending on the macromolecule of interest.

Once sequences are collected, comparative analysis of sequences from species of interest are done through multiple sequence alignment (Ramsden, 2023).  Basic Local Alignment Search Tool (BLAST) is one of such tools that checks for similar regions between biological sequences by comparing nucleotide or protein sequences to sequences in the database and calculates the statistical significance (Altschul et al., 1997). The comparison of sequences can allow for the inference of biologically significant relationships as well as identifying gene and protein families (Altschul et al., 1990).

Through phylogenetic analysis, computational and mathematical approaches are used to analyze biological datasets using maximum likelihood. This analysis generates trees based on likelihood scores that can give evolutionary insight into the divergence and relatedness of species. Additionally, through the reconstruction of ancestral sequences, amino acid replacements that are highly probable can give insight into the loss of function or otherwise in a protein of interest. By subjecting the GULO protein sequences across several mammalian species to these bioinformatics processes, gaining insight into the possible loss of function of GULO which results in the inability of Vitamin C synthesis in some higher mammals can be identified.

**1.5 Research Objectives and Hypothesis**

While a good number of mammalian species are currently able to produce Vitamin C due to a functional GULO gene, some mammalian species possess GULOP – a pseudogene.  Reconstruction of ancestral sequences can provide evolutionary insights into the mutations and gene conservations that are present in these genes across the several species analyzed. This study utilized ancestral sequence reconstruction and phylogenetic analysis to investigate the evolutionary divergence of the Gulonolactone Oxidase (GULO) gene into pseudogenes (GULOP) in specific mammalian lineages. In particular,this research focused strongly on identifying predicted key mutations of GULO in the *Strepsirrhini* sub-order of primates following ancestral sequence reconstruction. This is because the primates are divided into two suborders: Strepsirrhini and Haplorrhine and they are known to have evolved from a common ancestor. The haplorrhine orders containing humans and apes are known to possess pseudogenes and are unable to synthesize Vitamin C.

We believe that the amino acid replacements that happened in the internal branch prior to when the pseudogene evolved can provide a starting point to unravelling the loss of GULO activity in the *Haplorrhine* order.

**2.0 Materials and Methods**

The bioinformatic analysis of GULO protein sequences was carried out by first collecting sequences from the National Center for Biotechnology Information (NCBI) and querying for the sequences using Basic Local Alignment search Tool (BLAST). Following the sequences collection, multiple sequence alignment was carried out on the European Bioinformatics Institute job dispatcher tool using the Clustal Omega algorithm and the result of the alignment was visualized using Jalview.

After the sequence alignment, the sequences were then passed to Randomized Axelerated Maximum Likelihood (RAxML) to generate a phylogenetic tree. Mesquite was used to readjust the tree branches to conform to a Zoonomia project tree that used whole genome alignment of 240 phylogenetically diverse species of eutherian mammals. Finally, ancestral sequences were then reconstructed using Phylogenetic Analysis by Maximum Likelihood (PAML) to provide evolutionary insight into the divergence of the current mammalian sequences to early mammalian sequences.

**2.1 Sequence Collection**

Initially, GULO protein sequences were searched for by entering the word “L-gulonolactone Oxidase” on NCBI protein database. The results were then filtered to include only sequences for mammals. However, this search was not comprehensive enough as the number of mammalian species with the GULO protein sequence was limited.

A [BLAST](https://blast.ncbi.nlm.nih.gov/Blast.cgi) (Altschul *et al.,* 1997) protein search was then conducted using *Mus musculus* protein sequence with accession number NP\_848862 as the query search. The non-redundant protein sequences (nr) were selected as the choice database on BLAST, blastp (protein-protein BLAST) as the algorithm and the maximum target sequences was set to 5000.

**2.2 Multiple Sequence Alignment**

         Over 190 mammalian GULO sequences were collected and 3 non mammalian species belonging to two birds and one fish (table 2, supplemental figures).  All 197 mammalian sequences were entered into the EBI Clustal Omega job dispatcher (<https://www.ebi.ac.uk/jdispatcher/msa/clustalo>) tool (Madeira, 2022) for multiple sequence alignment (MSA) with the default settings and the fasta sequence generated were visualized using the Jalview version 2 software (Waterhouse, 2009).

Gaps were removed from the alignment on Jalview using the edit tool. Finally, trimming was done manually to get rid of poorly aligned regions from the multiple sequence alignment and even some mammalian species sequences were completely removed. Finally, an MSA fasta file containing 123 mammalian sequences and 3 non-mammalian sequences (available in table 3 in supplemental page) was then used for phylogenetic analysis.

**2.3 Phylogenetic Analyses**

         All 126 sequences were entered into RAxML GUI 2.0 (Edler *et al.,* 2021) for mac which uses the RAxML tool version 8.0 (Stamatakis, 2014) and RAxML-NG v1.2.1 (Kozlov *et al.,* 2019) for phylogenetic analysis. A model test based on ModelTest-NG (Darriba *et al.,* 2020) was initially run and the parameters (JTT+I+G) for the substitution models with the highest score and weight was selected. Analysis was carried out with the ML + Transfer + Boostrap Expectation + Consensus based on RaxML-NG tool (Kozlov *et al.,* 2019). *Acipenser trasnmontanus* (White sturgeon fish) with accession number ABO15549 was selected as the outgroup. Gamma (mean) was selected for rate heterogeneity, 100 for replicates and the +I (ML estimate) for the proportion of invariant sites.

         Following the phylogenetic analysis with RAxML, the best tree generated (figure 2) visualized with FigTree v1.4.4 was then redrawn to model the Zoonomia project phylogenetic tree of mammalian species (Zoonomia consortium, 2020) using Mesquite 3.81.

**2.4** **Reconstruction of Ancestral Sequences**

The tree output newick format from mesquite and all 126 MSA sequences were entered into PAML version 4.9j (Yang, 1997) as tree file and sequence file respectively following the CODEML program specification and parameters in table 1 below.

**Table 1:** Parameters used for the phylogenetic analysis using PAML.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Outfile | mlc |
| aaRatefile | wag.dat |
| Noisy | 9 |
| Model | 3 |
| fix\_alpha, clock, RateAncestor | 1 |
| verbose, seqtype, | 2 |
| runmode, getSE, aaDist, NSsites, Mgene, alpha, Malpha, getSE, cleandata, fix\_blength, method | 0 |
| Small\_Diff | .5e-6 |

**3.0** **Result**

**3.1 Phylogenetic analysis**

Following the analysis by RAxML, the tree in figure 2 was redrawn using mesquite to follow the Zoonomia project phylogenetic tree of mammalian species. A black background with different colored arrows

Description automatically generated

**Figure 2**: **A collapsed phylogram showing the evolutionary relationship of several mammalian species GULO protein sequences**. The tree was prepared using Figtree v1.4.4

**3.2 Reconstruction of Ancestral Sequences**

After the sequences were ran through PAML, node 195 shown in figure three below (figure not drawn to scale) containing mammalian species belonging to the order *Strepsirrhini* is of interest.

A chart with numbers and a number

Description automatically generated with medium confidence

**Figure 3**: **A section of the phylogram showing the node label for the mammalian species belonging to the *Strepsirrhini* order**. The tree was prepared using Figtree v1.4.4

Reconstruction of ancestral sequences using PAML showed that moving from node 195 to 196, the ancestral amino acid K (Lysine) at position 102 has a high probability (0.998) of being replaced by Q (Glutamine) with a probability of 1.000. Similarly, at position 137, the ancestral amino acid A (Alanine) is predicted to be replaced by S (Serine) with a high probability of 0.999. In total, moving from node 195 to node 196, a total of 15 amino acid replacement is predicted and their probabilities are shown in table 2 below.

**Table 2:** Summary of amino acid changes along node 195 to 196.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Amino Acid Position** | **Amino Acid at node 195** | **Probability of being replaced** | **Amino Acid at node 196** | **Probability of being replaced** |
| 102 | Lysine (K) | 0.998 | Glutamine (Q) | 1.000 |
| 137 | Alanine (A) | 0.988 | Serine (S) | 0.999 |
| 146 | Leucine (L) | 0.966 | Methionine (M) | 1.000 |
| 200 | Lysine(K) | 0.988 | Asparagine (N) | 1.000 |
| 204 | Aspartic acid (D) | 0.998 | Asparagine (N) | 0.999 |
| 211 | Lysine (K) | 0.997 | Arginine (R) | 0.999 |
| 268 | Cysteine (C) | 1.000 | Serine (S) | 1.000 |
| 283 | Asparagine (N) | 0.354 | Asparagine (N) | 1.000 |
| 284 | Glycine (G) | 0.778 | Histidine (H) | 1.000 |
| 297 | Threonine (T) | 0.995 | Serine (S) | 0.999 |
| 325 | Threonine (T) | 0.980 | Serine (S) | 0.623 |
| 330 | Histidine (H) | 0.999 | Glutamine (Q) | 1.000 |
| 348 | Aspartic acid (D) | 0.998 | Asparagine (N) | 0.999 |
| 366 | Methionine (M) | 0.999 | Isoleucine (I) | 0.999 |
| 385 | Threonine (T) | 0.916 | Asparagine (N) | 1.000 |

Lastly, the ancestral sequence in supplemental data along node 195 was also predicted using PAML.

**4.0** **Discussion and Conclusion**

In figure 2, the phylogenetic tree reveals the relationship between several groups of mammalian species. The tree starts out with placing *Acipenser transmontanus,* a species of fish as the outgroup supporting the theory that life evolved first on water. Also, the relationship between species belonging to the order lagomorphs (*Ochotona pirnceps)* and rodents (*Mus musculus* and *Rattus Novergicus*) is well established showing a descent from a common ancestor. Importantly, the population of rodent species is worthy of mention. Perhaps, the representative population in rodents will indicate the success of GULO expression as studies (Ohta & Nishikimi, 1999; Nishikimi *et al.,* 1994)   have described the presence of their twelve exons that result in a functional GULO as a basis for comparing the GULO in species such as human that have only six out of the eleven conserved exons responsible for GULO expression.

In table 2, the amino acid replacements that have occurred can serve a starting point to questioning the effect of those changes on the GULO expression since the node is the site of divergence between the haplorrhine and steprhinni species. Amino acid change can have effects on the structure and function of the protein. This is because, change in amino acid properties such as the hydrophobicity and charges can affect protein function, interactivity with other molecules or enzymatic activity.

While this study does not proceed to provide a structure for the protein that can be predicted from the generated ancestral sequence, it serves as a basis to further question whether the pseudogenization of the GULO gene are result of these amino acids replacement.

**References**

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.

Andrew Rambaut. FigTree Tee Figure Drawing Tool. Institute of Evolutionary Biology, University of Edinburgh. Available from:<http://tree.bio.ed.ac.uk/software/figtree/>

Cho, Y. S., Douglas, S. E., Gallant, J. W., Kim, K. Y., Kim, D. S., & Nam, Y. K. (2007). Isolation and characterization of cDNA sequences of L-gulono-gamma-lactone oxidase, a key enzyme for biosynthesis of ascorbic acid, from extant primitive fish groups. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *147*(2), 178-190.

Combs Jr, G. F., & McClung, J. P. (2016). *The vitamins: fundamental aspects in nutrition and health*. Academic press.

Cui, J., Pan, Y. H., Zhang, Y., Jones, G., & Zhang, S. (2011). Progressive pseudogenization: vitamin C synthesis and its loss in bats. *Molecular biology and evolution*, *28*(2), 1025-1031.

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular biology and evolution*, *37*(1), 291-294.

Drouin, G., Godin, J. R., & Pagé, B. (2011). The genetics of vitamin C loss in vertebrates. *Current genomics*, *12*(5), 371-378.

Edler, D., Klein, J., Antonelli, A., & Silvestro, D. (2021). raxmlGUI 2.0: a graphical interface and toolkit for phylogenetic analyses using RAxML. *Methods in Ecology and Evolution*, *12*(2), 373-377.

Henriques, S. F., Duque, P., López-Fernández, H., Vázquez, N., Fdez-Riverola, F., Reboiro-Jato, M., ... & Vieira, J. (2019). Multiple independent L-gulonolactone oxidase (GULO) gene losses and vitamin C synthesis reacquisition events in non-Deuterostomian animal species. *BMC Evolutionary Biology*, *19*, 1-12.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453-4455.

Lachapelle, M. Y., & Drouin, G. (2011). Inactivation dates of the human and guinea pig vitamin C genes. *Genetica*, *139*, 199-207.

Maddison, W. P., & Maddison. D.R. (2023. Mesquite): a modular system for evolutionary analysis. Version 3.81. http://www.mesquiteproject.org.

Madeira, F., Pearce, M., Tivey, A. R., Basutkar, P., Lee, J., Edbali, O., ... & Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic acids research*, *50*(W1), W276-W279.

Nishikimi, M., Fukuyama, R., Minoshima, S., Shimizu, N., & Yagi, K. (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. *Journal of biological chemistry*, *269*(18), 13685-13688.

Ohta, Y., & Nishikimi, M. (1999). Random nucleotide substitutions in primate nonfunctional gene for L-gulono-γ-lactone oxidase, the missing enzyme in L-ascorbic acid biosynthesis. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1472*(1-2), 408-411.

Ramsden, J. (2023). *Bioinformatics: an introduction*. Springer Nature.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Stipanuk, M. H., & Caudill, M. A. (2012). *Biochemical, Physiological, and Molecular Aspects of Human Nutrition-E-Book: Biochemical, Physiological, and Molecular Aspects of Human Nutrition-E-Book*. Elsevier health sciences.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189-1191.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences*, *13*(5), 555-556.

Zoonomia Consortium (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245. https://doi.org/10.1038/s41586-020-2876-6