The state of the enzyme-coding gene GULO recorded biological databases

Introduction

Vitamin C (ascorbic acid) biosynthesis is possible in some animals through the oxidation of L-Gulono-gamma-lactone by L-gulonolactone oxidase, encoded by the GULO gene[1]. The current research model for GULO action is Mus musculus, in which the GULO gene has 11 exons [2].

In this investigation, the animals used include the European lancelet (Branchiostoma lanceolatum), Tropical clawed frog (Xenopus tropicalis), Asian elephant (Elephas maximus), Guinea pig (Cavia porcellus) and Human (Homo sapiens). These species grant some insight into differences in the GULO gene over evolutionary stages: lancelets are regarded as a good model for vertebrate ancestors [3] because they are aquatic, and a theory to the evolutionary requirement to biosynthesise vitamin c was the change drastic increase in oxygen exposure when species evolved into terrestrial beings from aquatic environments [4], where vitamin c was a strong antioxidant the frog species could then show how successively more terrestrial species developed and required GULO to survive. After this, guinea pig acts as an outgroup to the functioning GULO in mice within the rodent classification, and African elephant could be an outgroup for humans as the only other large placental mammal in the list.

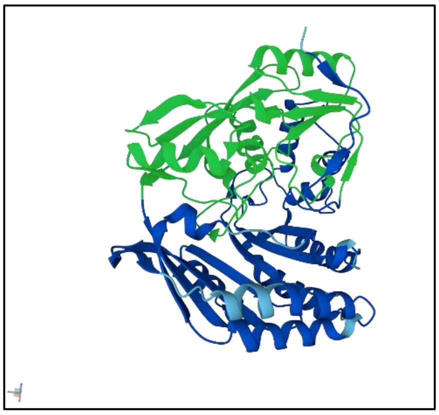

Figure 1 The FAD binding region found by Uniprot Prosite annotation in mouse GULO, positions 17-187 (highlighted in image from PDBe, which is a model of the Mus musculus GULO predicted with high confidence by Alphafold [1]). This residue gives the protein the ability to oxidise.

By using a species with a known functioning GULO gene as a query sequence, we can use BLAST[5] to assess the status of this gene in these Classes of animals and speculate why certain species are not able to biosynthesis ascorbic acid.

Materials and methods

The mRNA sequence for Mouse GULO was the starting point for investigation, with a BLASTx [6] search of default parameters providing its protein translation with 100% similarity ('Mus musculus L-gulonolactone oxidase, NCBI Reference Sequence: NP_848862.1'). In subsequent investigations, BLAST results filtered out low complexity regions and most hit lists were filtered by the names of the subject species.

With the Mouse GULO Protein sequence NP_848862.1 as reference, **BLASTP** was run to ascertain the existence of functioning L-gulonolactone oxidase proteins in our subject species and the hits list was filtered using the species names. This was done speculatively with the Non-redundant protein (nr) database, and repeated with default parameters using RefSeq Select proteins to search for higher quality, curated entries and then the UniProtKBBBB/Swiss-Prot database to establish if any functional GULO genes existed in our species. These searches returned no hits for the subject species.

To search from less similar sequences, the search was repeated in the UniProt database but with a BLOSUM45 matrix rather than the default BLOSUM62. Here, three hits were given for human (Q15392.2, Q8N465.3, O00116.1) including the protein Delta(24)-sterol reductase as it possesses the same FAD binding domain as the query [7]. Similarly, this search gave a single hit for Guinea pig containing the same domain (Alkyl-DHAP synthase, P97275.1[8]). At this stage, the small E-values for these results were misleadingly showing these 'hits' for the GULO gene when the protein only shared a functional domain.

The state of the enzyme-coding gene GULO recorded biological databases

By reverting to the lower-quality database of nr proteins, with the BLOSUM45 matrix and relaxed gap costs of 15,2, sequences were obtained for elephant, tropical clawed frog and guinea pig. However, there were still no hits for human or European lancelet GULO genes.

BLASTN searches were performed with the original mouse mRNA sequence as query to identify any homologues to GULO orthologs. The first search used the default nr/nt database but with a discontiguous megablast since previous results had indicated that there was little homology in our species at the protein level. This returned exon entries for the human GULO gene and mRNA sequences corresponding to the BLASTP successes. Checking for curated entries in the RNASeq database only gave nucleotide results for Mus musculus and using the EST database gave a hit for an experimental Xenopus cDNA read (GenBank: DN054249.1).

TBLASTN was performed with default parameters to assess evidence for an ancestral or pseudogene (non-functional) of GULO in the genome of our species, specifically Human and lancelet, which had not produced many results in other BLAST flavours. The XM entries for elephant, tropical clawed frog and guinea pig found in the normal BLASTn were confirmed, but not the other species. Since there were still no hits for human or European lancelet, the results implied that if these species possessed any trace of a GULO gene, that the sequences were too divergent for the scoring matrix. Consequently, the scoring matrix was loosened to PAM250, providing hits across the lancelet genome and repeated the human exons.

Results 1500-2500 words

Xenopus tropicalis

Xenopus tropicalis BLAST results					
BLAST flavour	Accession number	Maximum bit score	Total bit score	E value	Percentage Identity
BLASTP	XP_031758963.1	685	685	0	71.82%
BLASTN	XM_031903103.1	737	737	0	72.86%
TBLASTN	XM_031903103.1	658	658	0	69.55%

Obtaining results for the African Tree Frog (Atf) required a relaxed BLASTP search and remained somewhat elusive throughout investigation. The genome of the African Tree Frog is well sequenced due to its long term use as a genetic research model [9], however finding a hit in BLASTP required the use of the nr protein database. The extracted BLAST output hit score of 685 for the protein sequence with a high percentage identity suggested a high likelihood that the protein could be an ortholog, and the E-value of 0 made this a more confident assessment. Subsequently, the BLASTN mRNA result showed that the sequence was indeed a predicted L-gulonolactone oxidase protein with high coverage by RNASeq alignments. The MSA in the Appendix show that there is similar coverage across the entire length of the query Mouse peptide, however it should be noted that many residues are substituted throughout this chain and this could indicate adverse effects on protein shape. Without further avenues to explore the ensemble entry for the mouse GULO gene [10] was consulted and the clawed frog GULO gene was confirmed directly. There is a functioning GULO gene here.

Elephas maximus

Elephas maximus BLAST results					
BLAST flavour	Accession number	Maximum bit score	Total bit score	E value	Percentage Identity

BLASTP	XP_049721450.1	862	862	0	91.14%
BLASTN	XM_049865493.1	1585	1585	0	86.6%
TBLASTN	XM_049865493.1	827	827	0	87.73%

As is seen in the MSA of the Appendix and the percentage Identity of the above BLASTP result, the similarity between the Asian elephant sequence for the GULO gene and the query Mouse sequence is very close. High bit scores across the hits gathered throughout the investigation and strong E-values all act as evidence suggesting that a coherent L-gulonolactone oxidase protein might be present in this species. In the MSA, the major difference from the Mouse's functioning GULO gene is a 29 leading chain of transcribed peptides in the elephant's protein, however the functional parts of the gene seem unchanged. This would suggest that there would be a functioning GULO gene here, but consulting the ensemble orthologue page indicates that Loxodonta Africana is the elephant species which expresses a GULO gene that shares 91% identity and 99% coverage with the Mouse GULO gene. This near-identical sequence may be of some use to our investigation of the Asian elephant, as a putative gene in the Asian elephant could be inferred given its existence in a near genetic neighbour. On the other hand, the mutations may have relatively recent and cause the deactivation of the gene in the time since speciation between elephant types. Overall, the evidence here is inconclusive on the status of a functioning GULO gene in Elephas maximus.

Cavia porcellus

Cavia porcellus BLAST results					
BLAST flavour	Accession number	Maximum bit score	Total bit score	E value	Percentage Identity
BLASTP	XP_012998768.1	592	592	0	91.14%
BLASTN	XM_013143314.2	821	1231	0	85.28%
TBLASTN	XM_013143314.2	578	578	0	62.6%

Although the first Guinea pig sequence was discoverable with the same BLAST search parameters as the elephant and Atf sequences in the investigation, the evidence for a coherent GULO gene in the Guinea pig was weaker. Most evidently, the first result is deemed a 'LOW QUALITY PROTEIN', likely due to the detection of indels by the Gnomon gene prediction algorithm [11]. This lower quality is also reflected in the lower total bit score of 592 for this species compared to the first two species at this stage of the investigation. This annotation was also beneficial to the investigation, as it creates a conflicting image with the other BLAST outputs for the sequence which list an identity of 91.14%, a percentage cover of 99% and E-value of 0, which all suggest that a GULO gene is present. This conflicting image warranted further investigation, but similarly to the Atf entry, the mRNA of the Guinea pig is predicted to be a L-gulonolactone oxidase protein, with a high bit score of 1231 and high percentage identity (albeit with far weaker evidence for the prediction: fewer similar proteins and coverage of only 37% of annotated features seen in RefSeq entries). As will be seen in the Discussion, this evidence for a functioning GULO protein was not supported in the literature, which states that the Cavia porcellus species does not express the functioning protein.

As is visible in the MSA in the Appendix, there is a 29-residue insertion at position 125 of the Guinea pig sequence which is not present in the Mouse sequence, which occurs in middle the FAD-binding domain of the protein. This might disrupt the protein's folding and additionally its ability to perform the catalytic function in the biosynthesis of vitamin C. Therefore, whilst BLAST has rightfully identified that the GULO sequence is present in some form, it is possible that the protein is still produced but has been altered beyond function. This leads to the conclusion that a non-functioning GULO gene is present in Guinea pig, possibly to the point of pseudogenism.

Homo Sapiens

Homo Sapiens BLAST results						
BLAST flavour	Accession number	Maximum bit score	Total bit score	Query cover	E value	Percentage Identity
BLASTP	(NP_055577.1)	68.2	68.2	28%	2.90e-11	29.69%
BLASTN	D17461.1	189	331	24%	2E-42	85.98%
	NG_001136.2	189	425	32%	2E-42	85.98%
	D17460.1	103	103	8%	2e-16	81.08%
TBLASTN	D17461.1	67.1	144	21%	2e-09	72.09%
	NG_001136.2	67.1	190	29%	3e-09	72.09%
	D17460.1	44	44	7%	0.011	71.88%

As explained in the methods section, the investigation to find hits for human GULO required scoring matrices geared towards sequences that diverge substantially from the query sequence. The lack of entries in the first stage (BLASTP searches) when using the either the curated UniProt database or the lower quality Nr sequences indicated that there were no coding sequences for a coded GULO protein in humans.

After relaxing the search criteria and undertaking a nucleotide analysis in BLASTN, the results began showing the fragmented and few exons of the GULO gene. This was the first sign that the gene had undergone was unidentifiable or lost pseudogenisation as it was separated over a large region and was interspersed with numerous indels. The statistics of the results above show that the few matching segments in the human genome had low query cover but high percentage identity. This is to be expected as pseudogenes with substantial insertions and deletions will spread the original gene sequence over a large space, possibly beyond the query sequence's frame. In this case, the subjective decision to largely ignore the query cover can enable further investigation. TBLASTN searches were done to establish possible pseudogenisation, and although the same results were produced in TBLASTN as in BLASTN, the results have been included here to demonstrate that the database search parameters chosen during BLAST investigations can have a drastic effect on the statistics of hits. For example, the change of 425-190 total bit score for the NG_001136.2 entry, and extreme jump of the E-value to 2e-09 are evident in this scenario where alternative BLAST flavours can be compared against.

To determine how much of the coding region had persisted, the amino acid sequences from BLASTN were joined into a single synthetic FASTA sequence and put in the MSA alongside the other subject genes. The human construct was missing most of the FAD binding domain which is crucial for the catalytic role in ascorbic acid biosynthesis, as well as other regions which likely serve to preserve structural integrity.

Branchiostoma lanceolatum

Branchiostoma lanceolatum BLAST results						
BLAST flavour	Accession number	Maximum bit score	Total bit score	Query cover	E value	Percentage Identity

The state of the enzyme-coding gene GULO recorded biological databases

BLASTP	-	-	-	-	-	-
BLASTN	-	-	-	-	-	-
TBLASTN	OV696687.1	59.4	366	73%	9e-07	66.67%
	OV696695.1	72.4	126	42%	6e-11	28.28%

Gathering evidence for the status of the GULO gene in the lancelet was the most challenging. As can be seen in the table above, even the most lax BLASTP and BLASTN searches were unable to find any entries for the gene in this species' recorded genes. Eventually, using a PAM250 scoring matrix, there were two hits in lancelet from different parts of the genome. The entry from chromosome 2 seems to have more identity with the query sequence, so this was used in the MSA. As can be seen there, only half of the FAD region is seen in the lancelet sequence and the majority of the species' entry is non-identical to the Mouse query. Using BLAST's graphic summary of the alignment between the query and the two hits, the level of matching becomes apparent, and raises the question of an ancestor's GULO gene being duplicated in a similar event to that found in Tetradon lineage [12] , with a subsequent loss of the pseudogene explaining why there is no functional gene now.

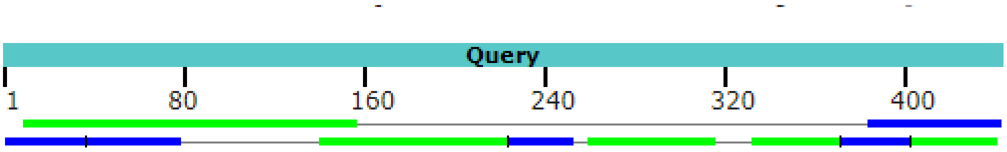

Figure 2 Both Branchiostoma lanceolatum results aligned with the mouse query sequence in the BLAST graphical summary viewer. The substantial islands of identity is apparent, despite the fact that the uppermost result in this alignment is on chromosome 10 and the lower on chromosome 2.

Based on the evidence available, this investigation concludes that Branchiostoma lanceolatum's does not have a functioning GULO gene, and that, if there once was such a functioning gene, it may have been duplicated into a paralog and lost as a pseudogene.

Discussion 400 words

Similar to the present study, lancelet analysis has been historically difficult due to poor elucidation of the genome [2] , however it could be theorised that no functioning GULO was ever selectively required in this species because lancets were not exposed to the atmospheric oxygen levels in the same way that ancestors to the tree Atf and other terrestrial species beyond were. Further investigation of close members of the Branchiostoma group may grant an inferred image of the status of GULO in B. lanceoatum

In this investigation, the conclusion regarding the Guinea pig GULO gene was that there might be a pseudogene or extremely mutated protein which did not serve its function but was still part of a coding sequence. Literature suggests that an expressed protein would likely still be visible through western blot techniques, and no such result has been seen in previous work [13]. It can be concluded confidently with reference to the literature that the Guinea pig species does not produce GULO [14].

The reason for the loss of GULO in terrestrial mammals is not known [15] , however, the literature supports our data here that humans and guinea pigs have suffered GULO inactivation over evolutionary timescales [16, 17]. Humans GULO pseudonisation and the loss of function to biosynthesis vitamin C may have been advantageous as the synthesis process also results in the production of reactive oxygen species as a by-product [18] . Humans can bypass the lack of biosynthesis by consuming exogenous ascorbic acid nutritionally [19], and it has been proposed that the requirement to acquire ascorbic acid purely through ingestion may have served as way to more tangibly control the levels of the vitamin, allowing better control of the transcription factor HIF1α [20] . Ultimately, the current understanding of the lack of GULO across the animal kingdom is incomplete, and is regarded to be a neutral trait given the ability of many species to survive without endogenous vitamin C biosynthesis [21] .

References

1. UNIPROT. GGLO_MOUSE. Available at: https://www.uniprot.org/uniprotkb/P58710/entry#function. Accessed Nov 4, 2022.

2. Yang H. Conserved or Lost: Molecular Evolution of the Key Gene GULO in Vertebrate Vitamin C Biosynthesis. *Biochem Genet*. 2013;51:413-425.

3. Holland LZ, Laudet V, Schubert M. The chordate amphioxus: an emerging model organism for developmental biology. *Cell Mol Life Sci*. 2004;61:2290-2308.

4. Nandi A, Mukhopadhyay CK, Ghosh MK, Chattopadhyay DJ, Chatterjee IB. Evolutionary Significance of Vitamin C Biosynthesis in Terrestrial Vertebrates. *Free Radical Biology and Medicine*. 1997;22:1047-1054.

5. Altschul SF, Wootton JC, Gertz EM, et al. Protein database searches using compositionally adjusted substitution matrices. *The FEBS journal*. 2005;272:5101-5109.

6. Nishikimi M, Yagi K. Molecular basis for the deficiency in humans of gulonolactone oxidase, a key enzyme for ascorbic acid biosynthesis. *The American journal of clinical nutrition*. 1991;54:1203S-1208S.

7. UNIPROT. DHC24_HUMAN. Available at: https://www.uniprot.org/uniprotkb/Q15392/entry. Accessed Nov 4, 2022.

8. UNPROT. ADAS_CAVPO. Available at: https://www.uniprot.org/uniprotkb/P97275/entry#names_and_taxonomy. Accessed Nov 4, 2022.

9. Grainger RM. Xenopus tropicalis as a model organism for genetics and genomics: past, present, and future. *Methods Mol Biol*. 2012;917:3-15.

10. ENSEMBL. Gulo Orthologues. Available at: http://www.ensembl.org/Mus_musculus/Gene/Compara_Ortholog?db=core;g=ENSMUSG00000034450;r=14:66224235-66246656;t=ENSMUST00000059970. Accessed Nov 4, 2022.

11. NCBI. Gnomon. Available at: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/. Accessed Nov 4, 2022.

12. Jaillon O, Aury J, Brunet F, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*. 2004;431:946-957.

13. Nishikimi M, Kawai T, Yagi K. 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*. 1992;267:21967-21972.

14. Nishikimi M, Koshizaka T, Ozawa T, Yagi K. Occurrence in humans and guinea pigs of the gene related to their missing enzyme L-gulono-gamma-lactone oxidase. *Arch Biochem Biophys*. 1988;267:842-846.

15. Nandi A, Mukhopadhyay CK, Ghosh MK, Chattopadhyay DJ, Chatterjee IB. Evolutionary Significance of Vitamin C Biosynthesis in Terrestrial Vertebrates. *Free Radical Biology and Medicine*. 1997;22:1047-1054.

16. Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K. 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*. 1994;269:13685-13688.

17. Lachapelle MY, Drouin G. Inactivation dates of the human and guinea pig vitamin C genes. *Genetica*. 2011;139:199-207.

18. Halliwell B. Vitamin C and genomic stability. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis.* Netherlands: Elsevier B.V; 2001;475:29-35. Available from: https://dx.doi.org/10.1016/S0027-5107(01)00072-0.

19. Ha MN, Graham FL, D'Souza CK, Muller WJ, Igdoura SA, Schellhorn HE. Functional rescue of vitamin C synthesis deficiency in human cells using adenoviral-based expression of murine l-gulono-γ-lactone oxidase. *Genomics*. 2004;83:482-492.

20. Grano A, De Tullio MC. Ascorbic acid as a sensor of oxidative stress and a regulator of gene expression: The Yin and Yang of vitamin C. *Medical hypotheses*. 2007;69:953-954.

21. Drouin G, Godin J, Pagé B. The Genetics of Vitamin C Loss in Vertebrates. *Current Genomics*. 2011;12:371-378.

22. Sievers F, Higgins DG. Clustal omega. *Curr Protoc Bioinformatics*. 2014;48:3.13.1-16.

23. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25:1189-1191.

<u>The state of the enzyme-coding gene GULO recorded biological databases</u>

Appendix

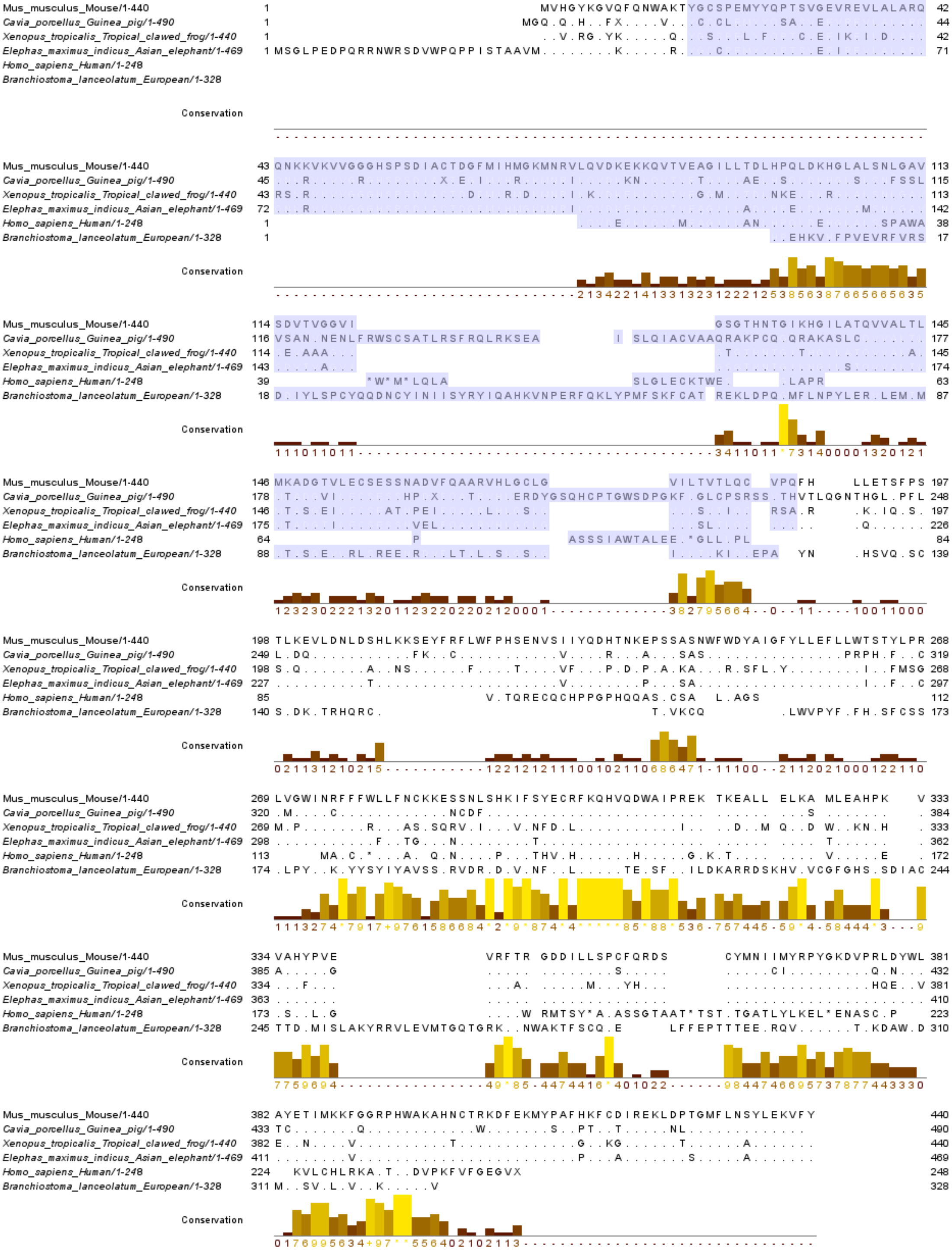
Mus_musculus_Mouse/1-440
Cavia_porcellus_Guinea_pig/1-490
Xenopus_tropicalis_Tropical_clawed_frog/1-440
Elephas_maximus_indicus_Asian_elephant/1-469
Homo_sapiens_Human/1-248
Branchiostoma_lanceolatum_European/1-328
Conservation

Figure 3 Figure 4 MSA of sequences, generated using Clustal Omega [22] and Jalview [23] . Conserved residues are represented by full stops, whilst any visible amino acid abbreviation letters indicate a lack of conservation in that position and gaps are whitespace The alignment uses the Mus musculus GULO peptide sequence as a reference sequence, and the lilac coloured block is the 'FAD-binding PCMH-type domain' identified by UniProt's automatic annotation [1].