Mutation ratios of proteoglycan synthesis genes in organisms with different bone classifications

Flaviu Vadan
Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada

Ian McQuillan
Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada

Brian F. Eames
Department of Anatomy & Cell Biology
University of Saskatchewan
Saskatoon, SK, Canada

April 7, 2020

Abstract

Proteoglycans are a class of macromolecules comprised of proteins to which carbohydrate groups are attached. They serve important functions in the synthesis of bone and cartilage. In addition, proteoglycans are known to function as extracellular signal molecules and have been identified intracellularly as well. The loss of proteoglycans is associated with debilitating diseases such as osteoarthritis. Therefore, an enriched knowledge of proteoglycans may aid osteoarthritis research by suggesting significant genes of the synthesis pathway, which can become subjects for more detailed studies. This research focused on collecting *Homo sapiens* orthologs of the proteoglycan synthesis pathway genes. Data for this study were collected from Ensembl. Multiple sequence alignments were performed to investigate gene sequence conservation across organisms. Phylogenetic analyses and an investigation of non-synonymous vs. synonymous mutation ratios of proteoglycan synthesis genes were conducted with the intent of verifying whether organisms that synthesize bone and cartilage exhibit higher mutation ratios compared to organisms that do not synthesize bone and cartilage. Organisms that synthesize bone and cartilage were hypothesized to showcase higher non-synonymous vs. synonymous mutation ratios compared to organisms that do not as a consequence of evolutionary pressures caused by environmental factors, such as the transition from water to land. Contrary to expectation, the research identified that organisms that synthesize bone and cartilage exhibit lower non-synonymous vs. synonymous mutation ratios compared to organism that do not synthesize bone and cartilage. [1]

[1]The code for data collection, parsing, and visualization, is available at: https://github.com/flaviuvadan/proteoglycan-pathway-evolution Study data are available in the same repository.

1 Introduction

Proteoglycans are a class of macromolecules comprised of core proteins to which glycosaminoglycans are attached [1]. The glycosaminoglycans are carbohydrate chains that are constructed from different types of sugars. Their configuration dictates the structure and function of the proteoglycans [2]. Proteoglycans can be found in the extracellular matrix [3], where they are involved in synthesizing articular cartilage and have been identified to function in processes such as growth factor signalling. Proteoglycans have been identified to be present intracellularly as well [4]. Proteoglycans are composed of a core protein onto which different carbohydrates are attached. The function of the proteoglycan and its tissue presence is dictated by the types of surface-attached carbohydrates. For example, chondroitin sulfate has repeating disaccharides of glucoronic acid and N-acetylgalactosamine whereas heparan sulfate has glucoronic acid and N-acetylglucosamine repeats [2].

The loss of extracellular proteoglycans is associated with debilitating diseases such as osteoarthritis. In addition to being known to be present in the extracellular matrix, recent findings that suggest proteoglycan to be located intracellularly have helped gain a better understanding of the consequences of proteoglycan loss [4, 5, 6]. In the extracellular matrix, proteoglycans are known to bind collagen to facilitate the formation of cartilage [7, 8]. For example, *Acan* encodes a highly sulfated protein that is known to facilitate bone shock absorption because of its capacity to bind water effectively [6, 8, 9].

One of the current problems in the field of skeletal evolution is the lack of understanding of the evolutionary trends of the genes present in the proteoglycan synthesis pathway. While efforts to understand genes independently have been previously pursued ([10]), an understanding of the evolutionary path of the genes of the proteoglycan synthesis pathway may guide future efforts towards studying specific genes in relation to their potential contribution towards the onset of osteoarthritis and guide decisions of organism model choices. In addition, the practical utilities of a richer understanding of evolutionary biology have been explored previously [11].

One of the standard approaches for performing comparative genomics analyses involves the use of multiple sequence alignments and phylogenetic profile similarities. Multiple sequence alignments are an essential tool in computational biology and bioinformatics for analyzing the homology of multiples sequences. Multiple sequence alignments involve aligning nucleotide, or protein, sequences to identify regions of similarity between sequences [12]. Phylogenetic profiling and phylogenetic trees are techniques used for identifying genes or proteins that coevolve and aid with identifying evolutionary relationships between proteins as interacting proteins tend to have similar tree topologies [13, 14]. Lastly, an approach for studying molecular evolution is the use of statistical methods such as comparisons between non-synonymous and synonymous substitution rates in protein-coding genes. Briefly, non-synonymous mutations are mutations that change the amino acid sequence of protein-coding genes while synonymous mutations are silent - they do not alter the amino acid composition but change codon composition. The non-synonymous to synonymous mutations ratio measures the evolutionary pressure that hypothetically acted on the selection of genes - a ratio less than 1 represents purifying/negative selection, a ratio of approximately 1 suggests neutrality, and a ratio greater than one represents positive selection [15, 14].

In this study, we investigated the homology of the genes of the proteoglycan synthesis pathway by employing comparative genomics techniques such as multiple sequence, phylogenetic, clustering, and dN/dS analyses. The group hypothesized that genes of the proteoglycan synthesis pathway exhibit higher non-synonymous vs. synonymous mutation ratios in organisms that synthesize bone/cartilage as a consequence of evolutionary pressures imposed by environmental factors such as the migration from water to land. The study performed a multiple sequence analysis of each

gene of the proteoglycan synthesis pathway using genomes from 187 organisms from Ensembl. Phylogenetic analyses have been performed to indicate gene presence and gene clustering for all organisms. Non-synonymous vs. synonymous mutation ratios have been evaluated using genomes from several representative organisms. Out of 187 organisms, 21 were selected as representatives in an effort to reduce the dimensionality of the dataset and facilitate the visualization of results. Finally, the results suggest that, contrary to the initial hypothesis, organisms that synthesize bone and cartilage tend to have lower rates of non-synonymous vs. synonymous mutation ratios compared to organisms that do not. In addition, groups of genes that perform similar functions tend to contain a subset of genes that exhibit sparse sequence conservation by comparison to other intergroup genes that have regions of high conservation.

2 Objectives

- Understand how genes of the proteoglycan synthesis pathway are distributed across multiple organisms;
- Investigate whether phylogenetic relationships between a subset of representative organisms from different taxonomic ranks is indicative of conservation of gene function
- Quantify the mutation ratios between genes of organisms from different clades

3 Materials and Methods

The following section will describe how data used for this research was collected and parsed. Specifically, the section touches on how *Homo sapiens* orthologs were collected from Ensembl and how multiple sequence alignments were generated for the respective orthologs' genes. In addition, the section covers how orthologs' taxonomic information was collected and how genes were mapped to their respective functions. Lastly, a discussion of how the phylogenetic trees were produced is provided and directions on how non-synonymous vs. synonymous mutation ratios were computed for the genes of a subset of the organisms collected from Ensembl.

3.1 Collection of orthologs

The 51 genes of the proteoglycan synthesis pathway were taken from [2] and their respective Ensembl IDs were manually curated by searching the Ensembl *Homo sapiens* build. The *Homo sapiens* orthologs of the 51 genes were collected using the REST homology information API hosted by Ensembl. The following parameters were passed to the API:

- $type = orthologues$;
- $sequence = dna$;
- $cigar\_line = 0$.

The returned homology information was parsed to extract the returned sequences and the corresponding species. The API requests were executed using the Python3 package "requests", version 2.18.4. Organisms with automatically annotated genomes were removed. Specifically, *Vicugna pacos*, *Tursiops truncatus*, *Erinaceus europaeus*, *Procavia capensis*, *Echinops telfairi*, *Pteropus vampyrus*, *Pongo abelii*, *Ochotona princeps*, *Sorex araneus*, *Choloepus hoffmanni*, *Tupaia burgeri*, and *Notamacropus eugenii* were excluded from the identified organisms. The gene sequences of the collected organisms were grouped based on the 51 genes used in the study.

3.2 Multiple sequence alignments (MSAs)

An MSA was generated for each of the 51 genes using EBI's *kalign* alignment tool [16]. The alignments were performed with default parameters and saved in FASTA format.

An MSA column similarity was defined as the ratio between the number of non-gap characters in an alignment's column and the total number of characters in the column. Column similarity ratios were computed for each MSA and the results were visualized using [17] (version 3.1.0).

3.3 Collection of taxonomic information

Taxonomic information was obtained for all the organisms identified in the orthologs collection step by accessing the REST API hosted by EBI's Taxonomy Service. All organisms' taxonomic information was collected by fetching data using their respective scientific names (e.g *Homo sapiens*). The data were parsed to obtain the following taxonomic ranks: kingdom, subkingdom, phylum, clade, subphylum, clade, class, subclass, superorder, order, suborder, subsuborder, family, genus, and species. The API requests were executed using the Python3 package "requests", version 2.18.4.

3.4 Taxonomic and gene function mapping

Taxonomic frequencies were computed for each organism class and visualized using [17] (version 3.1.0). Similarly, genes were binned according to the function classification outlined in [2] and the classification was visualized using the same version of [17].

3.5 Phylogenetic trees

The phylogenetic relationship between the organisms used in this study was modelled using NCBI's Common Tree [18, 19]. The data were saved in phylip format and a central species phylogenetic tree was created. As a visualization technique, fifty one species phylogenetic trees were generated based on the topology of central tree. For every gene, the organisms that do not have the respective gene in their genome were marked in the phylogenetic tree. The generated species phylogenetic trees were grouped based on the organisms that are marked to have the gene that is represented by the tree. An NCBI nucleotide BLAST gene search was performed against the set of organisms of each tree whose genomes were reported to not contain the gene of interest.

Algorithm 1 was used to group trees based on their organism gene presence similarities. The algorithm initializes a queue that initially contains all the 51 phylogenetic trees. As long as the queue is not empty, the algorithm removes a tree from the queue and finds another tree in the queue. Similarity is assessed based on an index obtained by taking the ratio between the intersection of the number of elements in the list of organisms represented by each tree and the total number of organisms in the study [20]. If a similarity of 0.8 (on a scale from 0 to 1) is identified, the algorithm joins the two trees by creating a new phylogenetic tree that represents the organisms from the original intersection of the trees, discards the trees that were initially extracted from the queue, and adds the newly created tree back to the queue. If the similarity threshold of 0.8 is not satisfied, the removed tree cannot be joined with another tree so it is added to the final list of trees, which is returned when the algorithm halts. Algorithm 1 represents the pseudocode for grouping trees based on similarity.

```
Result: A collection of grouped phylogenetic trees
trees = an empty list;
Q = an empty queue;
Add all the trees to Q;
while Q is not empty do
    T = Q.pop();
    similarT, score = findSimilarTree(T, Q);
    if score ≥ 0.8 then
        newT = joinTrees(T, similarT);
        Q.remove(similarT);
        Q.push(newT);
    else
        trees.append(T)
    end
end
return trees;
```

Algorithm 1: The algorithm that was used for grouping similar trees

3.6 *Reductio ad significans*

Clade representatives from each taxonomic group were selected and used as model organisms. The organisms were chosen based on their frequent use as model organisms and their record of good quality genomes.

3.7 Computing the ratio of non-synonymous to synonymous mutation ratios (dN/dS)

The ape package in R was used to compute mutation ratios [21]. ape was accessed using the Python3 rpy2 package (version 3.2.4). For every gene, ape's "dnds" function was used to compute the mutation ratios between every pair of organisms whose genomes include the respective gene. The results were visualized in 2D grids using the Python3 seaborn package (version 0.9.0). In addition, distributions of dN/dS ratios were constructed for each significant organism. Genes that were only present in a single organism out of a pair of organisms that were used for computing dN/dS were given scores of 1 for neutrality.

4 Results

This section offers a description of how many organisms were collected, along with the reasoning behind removing some of them. The MSAs of all the genes will be described by offering a intra-group comparison between genes that have similar function e.g sulfatases. An outline of an observation of the types of organisms the scientific community has access to through Ensembl will be constructed. The phylogenetic relationships between clades of organisms with different genetic traits will be described. All the generated trees are available as supplementary information. Lastly, the differences in dN/dS ratio distributions of different organism groups will be described.

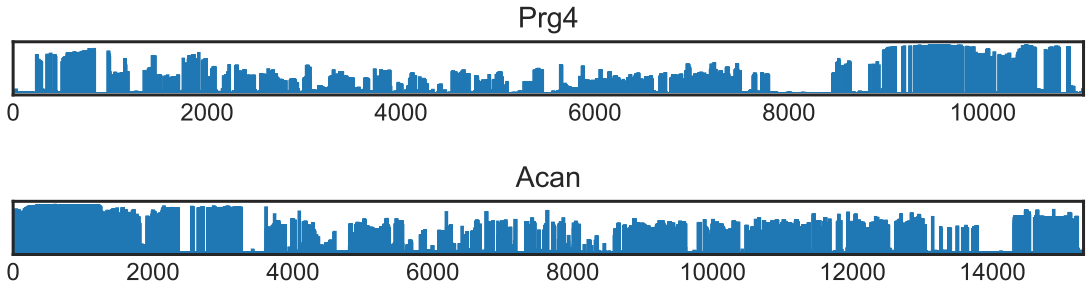

Figure 1: Multiple sequence alignments of the core protein genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

4.1 Collection of orthologs

The collection of orthologs from Ensembl resulted in 199 unique organisms. Of 199 collected organisms, 187 were used. The 12 organisms that were not used represented Ensembl genome records that are projection builds, which are low coverage genomes that have genes annotated by mapping to the human genome.

4.2 Multiple sequence alignments (MSAs)

The genetic sequences were grouped based on the 51 genes of the proteoglycan synthesis pathway. An MSA of the ortholog sequences for each gene was constructed, which resulted in 51 MSA plots.

4.2.1 Core proteins

The two core genes that were evaluated are *Acan* and *Prg4*. The MSA of *Acan* contains regions that were successfully aligned across the entire length of the gene. Multiple column similarities values that are not higher than 50% are present throughout the alignment (Figure 1). The MSA of *Prg4* showcases similarities with the MSA of *Acan* but contains a higher proportion of regions of column similarities that are less than 50%.

4.2.2 Xylosyltransferases

The two xylosyltransferase genes that were evaluated are *Xylt1* and *Xylt2*. The MSA of *Xylt1* and *Xylt2* showcase similar patterns (Figure 2). Both MSAs showcase regions that exhibit high conservation based on column similarity indices but have flanking regions with low similarity/a large number of alignment gaps.

4.2.3 Galactosyltransferases

The evaluated galactosyltransferase genes are *B3galt6, B4galt7, Chsy1, Chsy3, Csgalnact1,* and *Csgalnact2*. The MSA of *B3galt6* showcases two regions of high conservation between nucleotide 1500 and the end of the sequence (Figure 3). The MSA of *B4galt7* exhibits four regions of high conservation, which are separated by shorter regions of several hundreds of nucleotides. The MSA of *Chsy1* showcases high conservation based on the number of regions that are conserved and

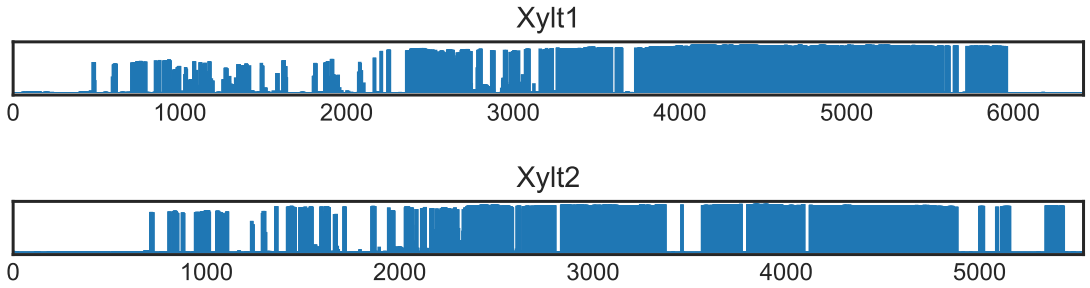

Figure 2: Multiple sequence alignments of the xylosyltransferase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

the consistently high ratio of column similarity. In addition, the region spanning the first 600 nucleotides showcases low column similarity. The MSA of *Chsy3* showcases similar conservation patterns to *Chsy1* as both genes have a starting region spanning several hundreds of nucleotides with low column similarity but exhibit wide regions of conservation. The alignments of *Csgalnact1* and *Csgalnact2* exhibit wide regions of similarity. However, *Csgalnact2* contains a region of low similarity from position 0 to position 100 of the alignment. In addition, *Csgalnact2* contains 3 regions of low similarity within the alignment.

4.2.4 Glucuronyltransferases

The evaluated glucuronyltransferase genes are *B3gat3, Chpf,* and *Chst1*. The alignment of *B3gat3* contains two regions of approximately 400 nucleotides that are conserved across all organisms' genes (Figure 4). In addition, the alignment showcases interspersed short regions that are conserved - approximately 50-100 nucleotides - with varying column similarities. The MSA of *Chpf* showcases conserved regions across the lengths of the aligned genes. However, the first 250 nucleotides were either not aligned and resulted in multiple gaps or were aligned but resulted in low similarity scores. Lastly, the alignment of *Chst1* showcases a conserved region of approximately 750 nucleotides with the rest of the alignment containing short regions of approximately 50 nucleotides that were aligned.

4.2.5 Glycosyltransferases

The evaluated glycosyltransferase genes are *Ext1* and *Ext2*. The MSA of *Ext1* showcases conservation across the gene with the exception of a short region of approximately 400 nucleotides at the beginning of the alignment (Figure 5). The alignment of *Ext2* exhibits conservation across the second half of the gene with the first half of approximately 2000 nucleotides showcasing low or not conservation - low column similarity score or gaps across the entire column.

4.2.6 Sulfotransferases

The evaluated sulfotransferase genes are *Chst1, Chst11, Chst12, Chst13, Chst14, Chst15, Chst2, Chst3, Chst4, Chst7, Chst9,* and *Ust*. The alignment of *Chst1* presents a conserved region from nucleotide 2200 to 3100 with the rest of the alignment showcasing either low column similarity or no conservation (Figure 6). *Chst11* exhibits a conserved region from nucleotide 1000 to

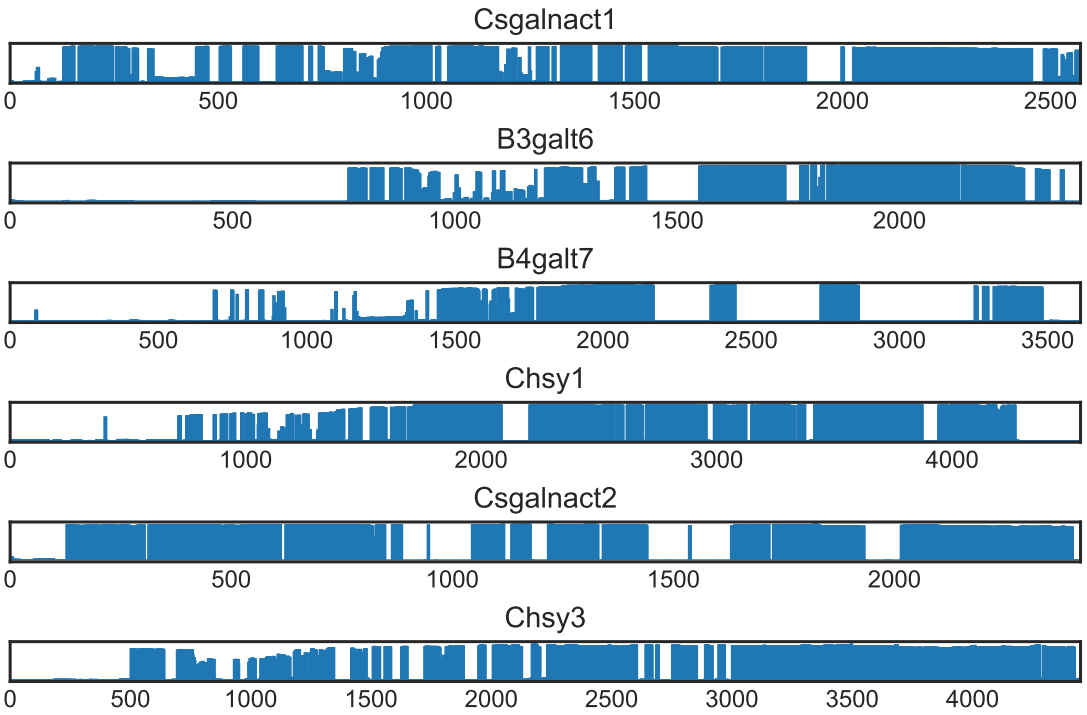

Figure 3: Multiple sequence alignments of the galactosyltransferase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

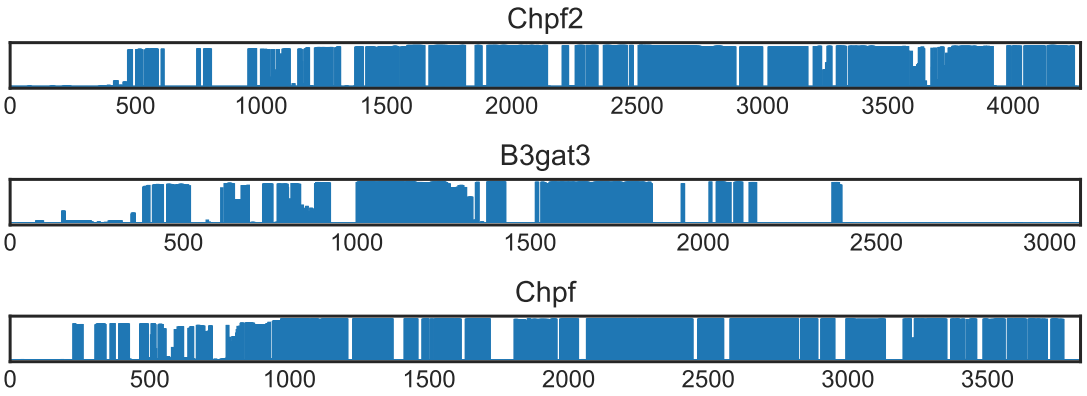

Figure 4: Multiple sequence alignments of the glucuronyltransferase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

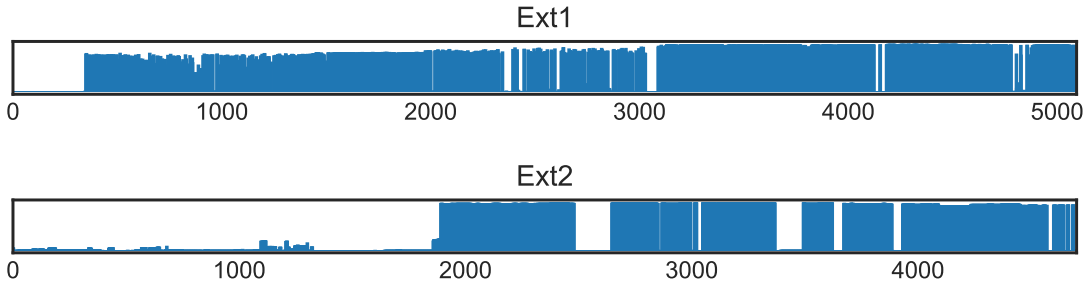

Figure 5: Multiple sequence alignments of the glycosyltransferase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

approximately 1800 with the rest of the alignment containing short regions of high similarity interspersed between regions of low or non-existing column similarity. The alignment of *Chst12* showcases conservation from location 1600 to the end of the gene sequence with multiple short regions of several hundreds of nucleotides that are conserved. *Chst13*'s alignment contains a conserved region from nucleotide 1600 towards the end of the gene alignment. The first half of the alignment showcases 6 short regions that exhibit conservation with the rest being either low conservation or no conservation. The alignment of *Chst14* showcases a conserved region of approximately 750 nucleotides towards the end of the alignment with the first half containing either non-existing or low conservation patterns. The MSA of *Chst15* showcases long conserved regions of approximately 1500 and 500 nucleotides, respectively, separated by a short region of low column similarity of approximately 100 nucleotides. In addition, the beginning of the alignment contains a region of approximately 300 nucleotides that showcase low column similarity. The alignment of *Chst2* showcases a continuous region of high similarity between nucleotides 1000 and approximately 2600. The alignment region before nucleotide 1000 showcases column similarity values between 50 and approximately 80%. The alignment of *Chst3* showcases a conserved region from nucleotide 1000 to approximately 2400 with the region between nucleotide 500 and 1000 showcasing high and low column similarity ranging from approximately 10 to 80%. The alignment of *Chst4* showcases 3 highly conserved regions between nucleotide 750 and 1000, 1100 and 1250, and 1300 and 2000, respectively. The first 500 nucleotides in the alignment showcase low to no column similarity. The alignment of *Chst7* showcases regions of similarity clustered towards the middle of the alignment from nucleotide 400 to 2200. The conserved regions contain column similarity values ranging from approximately 10 to 100%. Lastly, the alignment of *Chst9* showcases a conserved region from nucleotide 800 to 1750 and from nucleotide 1800 to 2200. The first 800 nucleotides contain short conserved regions with length ranging from 20 to 100 nucleotides.

4.2.7 Sulfatases

The evaluated sulfatase genes are *Arsa, Arsb, Arsc, Arsd, Arse, Arsf, Arsg, Arsi, Arsj, Arsk, Galns, Gns, Ids, Sulf1, Sumf1,* and *Sumf2*. The alignment of *Arsa* showcases multiple short conserved regions ranging in length from approximately 50 nucleotides to 400 nucleotides (Figure 7). The beginning of the alignment contains a region of 500 nucleotides that showcase low conservation while the end of the alignment showcases a region of approximately 300 nucleotides

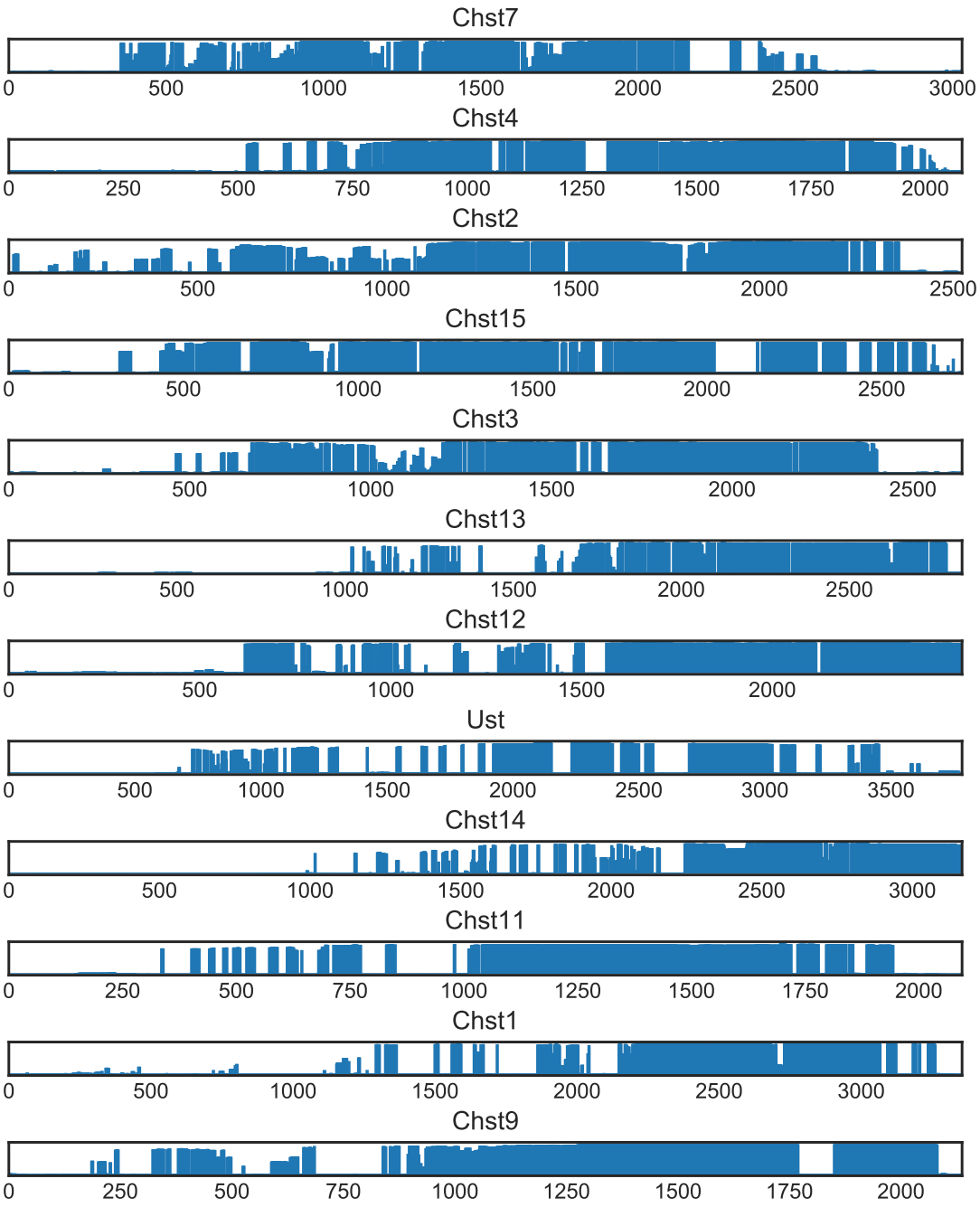

Figure 6: Multiple sequence alignments of the sulfotransferase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

with low conservation. The MSA of *Arsb* showcases long conserved regions from nucleotide 500 to approximately 2700 with the first 400 and the last 300 nucleotide regions showcasing low similarity. The MSA of *Arsc* showcases interspersed regions of high similarity between nucleotide 1100 and 4000 with the start and end regions of the alignment showcasing low similarity ranging from approximately 10 to 40%. The alignment of *Arsd* exhibits a conserved region from nucleotide 1500 towards the end of the sequence. The first 1500 nucleotides showcase low similarity ranging from approximately 10 to 80%. The alignment of *Arse* showcases conserved regions ranging from approximately 10 to 100% across a region from nucleotide 500 to 5200 with the flanking regions of the alignment showcasing no similarity. The alignment of *Arsf* showcases similar patterns to the alignment of *Arse* with conserved regions from nucleotide 500 to 4500 and flanking regions of low similarity. The alignment of *Arsg* showcases multiple conserved regions. The conserved regions range from nucleotide 500 to 1500 and 2200 to 4500 with interspersed regions of low similarity. The alignment of *Arsi* showcases 9 regions of high similarity separated by a region of no similarity from nucleotide 2500 to approximately 3800. The MSA of *Arsj* showcases high similarity spread throughout the alignment with short regions of low similarity of lengths ranging from 10 to 50 nucleotides. The alignment of *Arsk* showcases multiple regions of high similarity ranging from nucleotide 1500 to 3300 with multiple short regions of lengths ranging from 20 to 50 nucleotides surrounding the regions of high similarity. The flanking regions of the alignment represent conservation regions of 500 nucleotides. The MSA of *Galns* showcases a conserved region from nucleotide 1000 to approximately 3000, which are separated by approximately 8 regions of low similarity of length 10 to 50 nucleotides. The rest of the alignment showcases multiple short regions of 10 to 40 nucleotides that showcase high similarity across the alignment. The MSA of *Gns* showcases a conserved gene region from nucleotide 1700 to 3500. The first half of the alignment contains a highly conserved region from nucleotide 1000 to 1400 with the rest of the alignment showcasing low to no similarity. The MSA of *Ids* showcases multiple regions that are highly conserved across the gene with lengths ranging from 50 to 200 nucleotides. The first 1800 nucleotides of the alignment showcase column similarity values ranging from approximately 10 to 50%. The alignment of *Sulf1* showcases 3 highly conserved regions of approximately 500 nucleotides with multiple, short, highly conserved regions spread throughout the alignment. The region from nucleotide 5000 to 7000 contains similarity values ranging from 40 to approximately 80%. The MSA of *Sumf1* contains two regions of high similarity from nucleotide 1500 to 2000 and 2200 to 2400, respectively. The region from nucleotide 500 to 1500 showcases similarity values ranging from 20 to 80% with the flanking regions of the alignment showcasing low column similarity values of approximately 10%. Lastly, the MSA of *Sumf2* showcases conserved regions ranging from the start of the gene alignment to nucleotide 2000. The second half of the alignment contains either low similarity or no similarities.

4.2.8 Kinases

The evaluated kinase genes are *Fam20a, Fam20b,* and *Fam20c.* The MSA of *Fam20a* contains a highly conserved region from nucleotide 4000 to the end of the alignment (Figure 8). The first 4000 nucleotides contain interspersed regions of short length - approximately 50 nucleotides - but high column similarity values - approximately 80%. The alignment of *Fam20b* showcases 9 distinct nucleotide regions that are highly conserved with regions of low conservation scattered between. The longest region of low similarity ranges from nucleotide 400 to approximately 800. The alignment of *Fam20c* showcases three regions of high similarity with length ranging from 200 to 400 nucleotides. The rest of the alignment contains short regions of 20 to 100 nucleotides that are highly conserved.

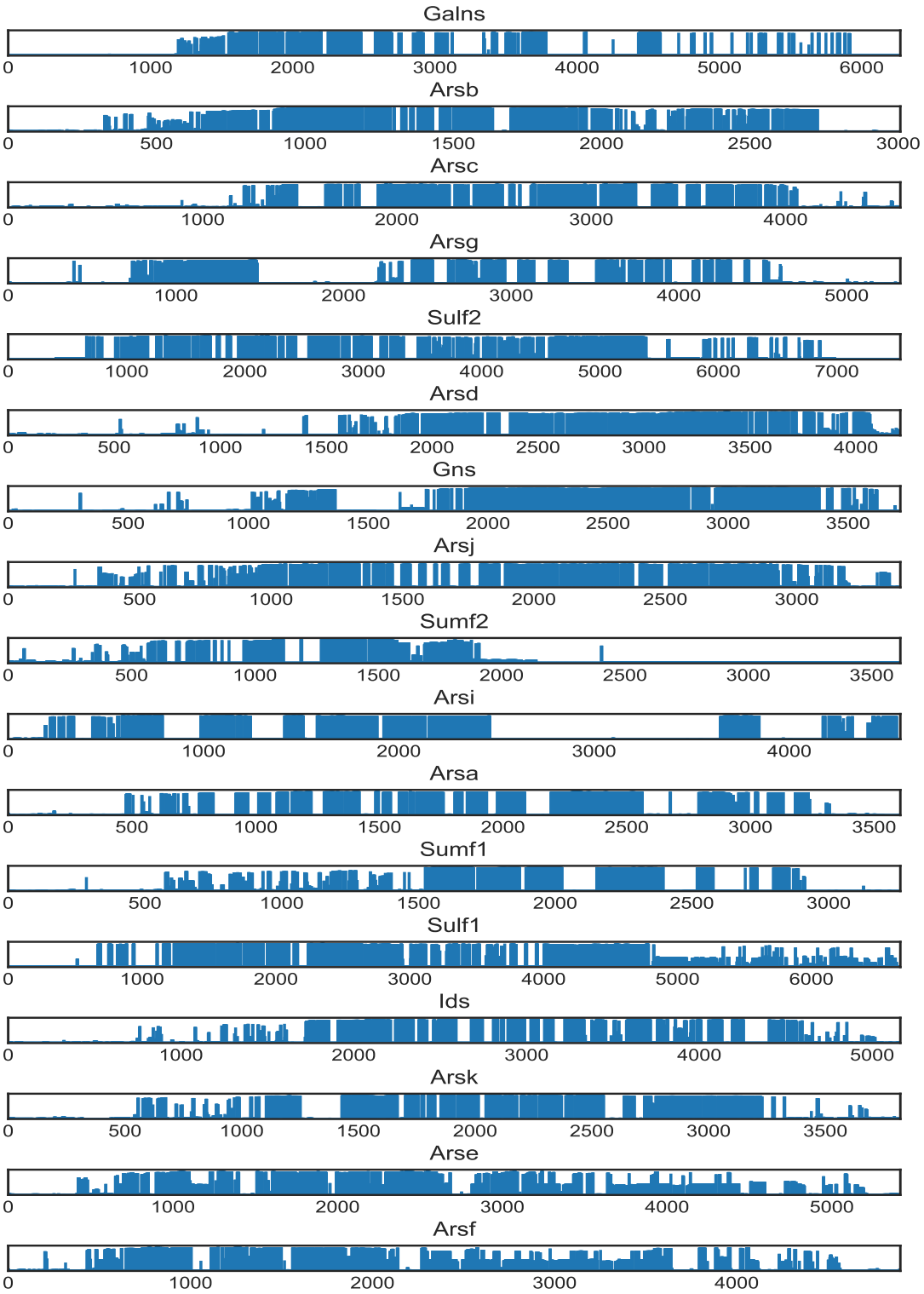

Figure 7: Multiple sequence alignments of the sulfatase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

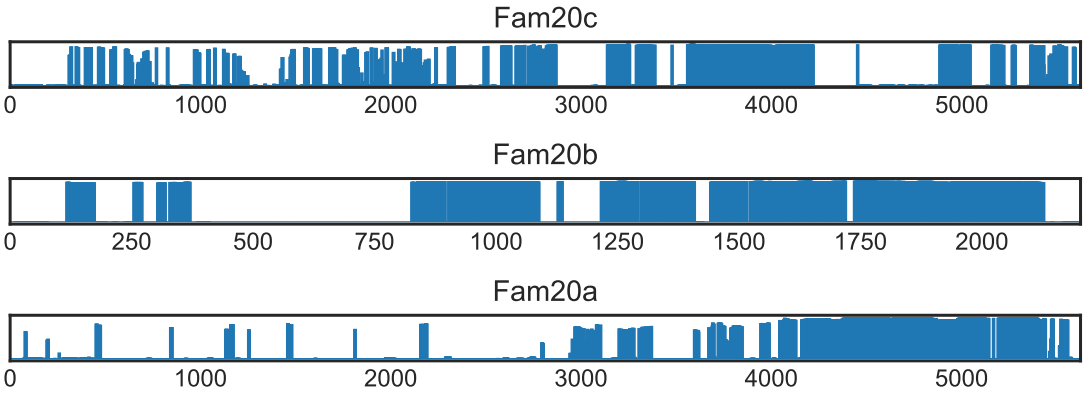

Figure 8: Multiple sequence alignments of the kinase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

Figure 9: Multiple sequence alignments of the epimerase gene. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

4.2.9 Epimerases

The evaluated epimerase gene is *Dse*. The MSA of *Dse* showcases a highly conserved region of approximately 3000 nucleotides. In addition, the alignment showcases multiple short regions ranging from 50 to 200 nucleotides that are highly conserved (Figure 9). The first 600 nucleotides showcase low similarity across the alignment.

4.2.10 Glycosidases

The evaluated glycosides gene is *Gusb*. The alignment of *Gusb* contains several highly conserved regions. Specifically, the alignment showcases 4 regions of lengths ranging from 200 to 600 nucleotides that are highly conserved (Figure 10). In addition, the alignment showcases multiple short regions of length ranging from 20 to 100 nucleotides that are highly conserved.

4.2.11 Sulfohydrolases

The evaluated sulfohydrolase gene is *Sgsh*. The MSA of *Sgsh* contains a region that is highly conserved between nucleotide 2800 to 3700 (Figure 11). In addition, the alignment showcases 4

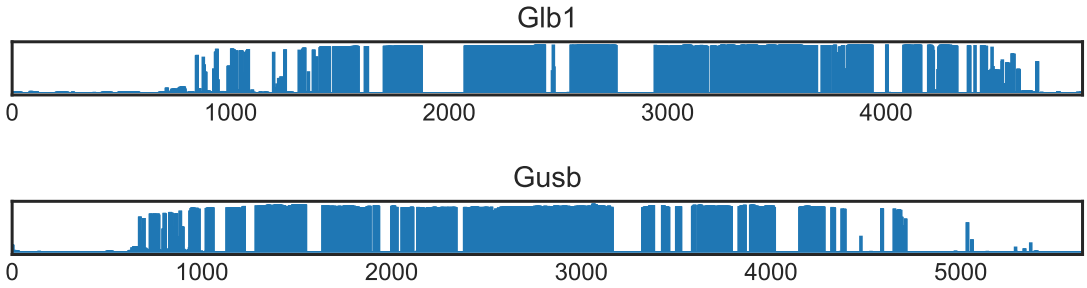

Figure 10: Multiple sequence alignments of the glycosidase genes. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

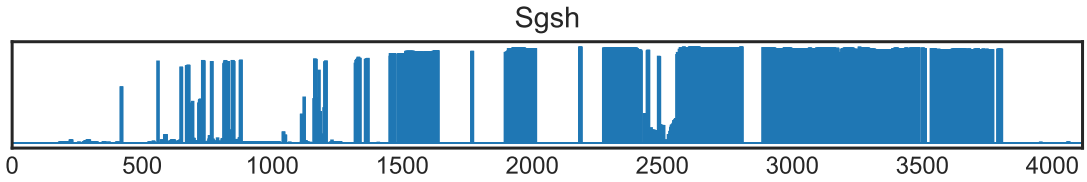

Figure 11: Multiple sequence alignments of the sulfohydrolase gene. The x axis represents the length of the alignment (number of nucleotides) whereas the y axis represents the column similarity - the ratio between the number of non-gap characters in a sequence alignment column and the total number of characters in the column.

regions that are approximately 200 nucleotides in length and showcase high similarity. The rest of the alignment contains scattered regions of low column similarity.

4.3 Taxonomic information

The collection of taxonomic information resulted in a record being produced for every organism used in this study. Organisms were grouped by taxonomic class. Figure 13 provides a summary of the number of organisms in each taxonomic category that was used for categorizing organisms. Significant organisms were selected from each taxonomic group outlined in Figure 13, which resulted in the 21 organisms outlined in Table 1.

4.4 Phylogenetic relationships

The phylogenetic relationships between all the genes used in this study were captured and visualized, which resulted in 51 "gene trees". Figure 12 represents a sample "gene tree" that was constructed from a phylogenetic tree that represents the evolutionary relationship between all the organisms used in this study. Trees were grouped based on organisms' genetic similarity using Algorithm 1, which resulted in a reduction from 51 trees to 16. Table 2 outlines the "gene trees" that were combined into a single tree.

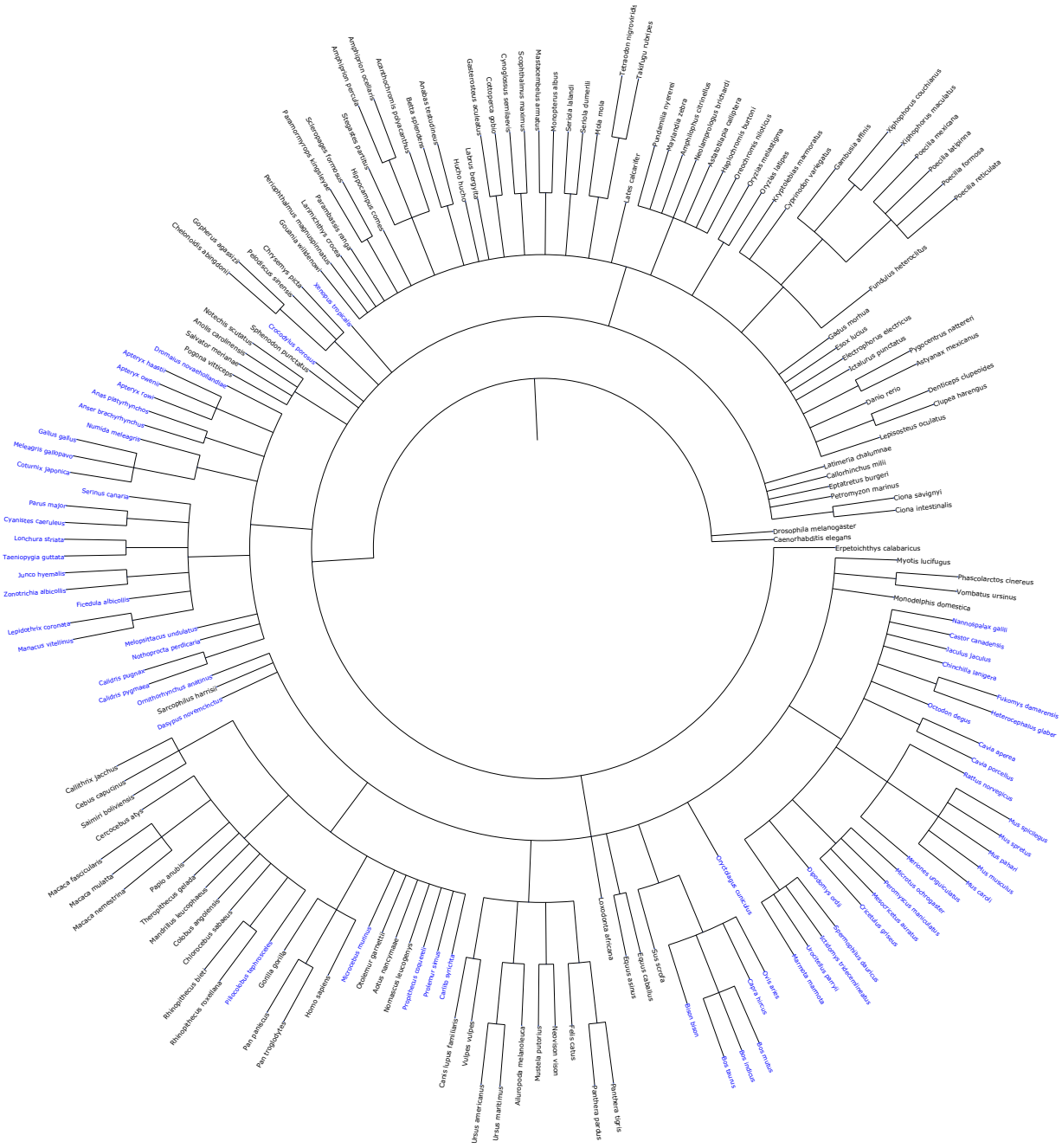

Figure 12: The gene tree for *Arsf*. The blue leaves represent organisms that do not have *Arsf* in their genomes whereas black leaves represent organisms that have *Arsf* in their genomes. The tree was constructed from a global phylogenetic tree representing the relationships between the organisms included in this study.

Class	Organism	Terrestrial	Aquatic	Cartilage	Bone
Mammalia	*Homo sapiens*	✓		✓	✓
	Mus musculus	✓		✓	✓
	Ornithorhynchus anatinus	✓		✓	✓
Actinopterygii	*Danio rerio*		✓	✓	✓
	Oryzias latipes		✓	✓	✓
	Takifugu rubripes		✓	✓	✓
	Lepisosteus oculatus		✓	✓	✓
Archelosauria	*Anas platyrhynchos*	✓	✓	✓	✓
	Gallus gallus	✓		✓	✓
	Chrysemys picta bellii		✓	✓	✓
	Crocodylus Porosus	✓	✓	✓	✓
Lepidosauria	*Notechis scutatus*	✓		✓	✓
	Anolis carolinensis	✓		✓	✓
Amphibia	*Xenopus tropicalis*	✓	✓	✓	✓
Holocephali	*Callorhinchus milii*		✓	✓	
Coelacanthiformes	*Latimeria chalumnae*		✓	✓	
Hyperoartia	*Petromyzon marinus*		✓	✓	
Hyperotreti	*Eptatretus burgeri*		✓	✓	
Phlebobranchia	*Ciona intestinalis*		✓		
Pterygota	*Drosophila melanogaster*	✓			
Rhabditina	*Caenorhabditis elegans*	✓	✓		

Table 1: An outline of the subset of organisms chosen as representatives.

Tree	**Genes**
1	*Arsc, Arsi, Dse, Ext1*
2	*Arsd, Arsj*
3	*Arse*
4	*Arsf*
5	*B3galt6, Chst13*
6	*Chst3, Xylt2, Gns, Acan, Chst1*
7	*Chst4*
8	*Chst7, B4galt7, Fam20b*
9	*Chst9*
10	*Chst12, Chst15, Glb1, Arsg, Fam20a, Chst2, Chpf, Gusb, Chsy1, Sumf1*
11	*Chst14*
12	*Chsy3*
13	*Sulf2, Chst11, Csgalnact2, Arsb, Galns, Fam20c, Prg4*
14	*Sumf2*
15	*Ust, Chpf2, Ext2, Arsa, Sulf1, Arsk, Xylt1*

Table 2: An outline of the gene trees that were combined based on organisms' similarity.

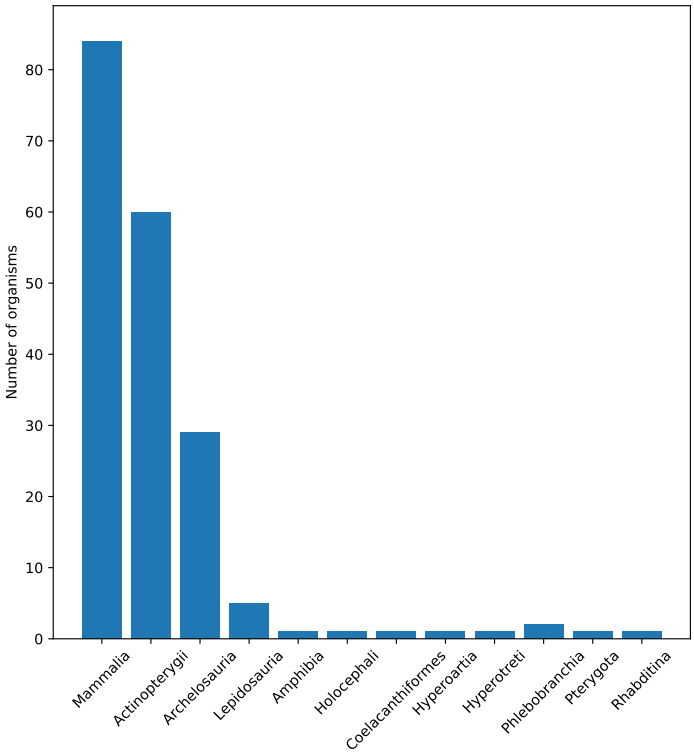

Figure 13: Taxonomic frequency of the analyzed organisms. The presented classes are: Mammalia - mammals; Actinopterygii - ray-finned fish; Archelosauria - turtles and archosaurs; Lepidosauria - scaly reptiles; Amphibia - amphibians; Holocephali - cartilaginous fish; Coelacanthiformes - bony fish (ancient); Hyperoartia - jawless bony fish; Hyperotreti - hagfish; Phlebobranchia - sea squirts; Pterygota - winged insects; Rhabditina - nematodes.

4.5 dN/dS ratios

The dN/dS ratios were computed for the chosen significant organisms, which lead to the creation of 21 dN/dS plots. A sample plot is outlined in Figure 17 with the rest of the dN/dS plots available as supplementary information. The results of all 21 dN/dS plots were summarized in Figure 14.

The dN/dS summaries of *Homo sapiens, Mus musculus, Ornithorhynchus anatinus, Danio rerio, Oryzias latipes, Takifugu rubripes, Lepisosteus oculatus, Anas platyrhynchos, Gallus gallus, Chrysemys picta, Crocodylus porosus, Notechis scutatus, Anolis carolinensis, Xenopus tropicalis, Callorhinchus millii*, and *Latimeria chalumnae* showcase bimodal distributions with an average of 48% of the dN/dS values being less than 1 while 52% of the values are greater than or equal to 1. The dN/dS summaries of *Petromyzon marinus, Eptatretus burgeri, Ciona intestinalis, Drosophila melanogaster*, and *Caernohabditis elegans* showcase the values are approximately normally distributed with a high proportion of the values clustered at the mean and a small variance. Approximately 23% of the dN/dS values are less than 1.

The genes dN/dS distributions illustrated by the classification of organisms based on bone, cartilage, or absence of bone and cartilage do not showcase clear statistically significant general differences between the distributions of the genes. The median of the dN/dS distribution for *Fam20b* in organisms without bone or cartilage is approximately 1, compared to the distributions of organisms with bone and organisms with cartilage only that showcase a median of approximately 0.1. By comparison to organisms with bone and organisms without bone and cartilage, the distribution of *Chst4* in cartilaginous organisms is clustered at 1. The dN/dS distribution of *Acan, Chst15, Ext1, Chst13, Chst12, Chst11, Chst1*, and *Chst9* are all clustered at 1 compared to the corresponding distributions in bony organisms and cartilaginous organisms.

The genes dN/dS distributions illustrated by the classification of organisms based on terrestrial, aquatic, or both terrestrial and aquatic habitats do not showcase clear statistically significant inter-group differences. The distribution of *Sumf2* in organisms that live in terrestrial and aquatic environments is clustered at 1 compared to the distribution of the same gene in organisms that are terrestrial and organisms that are aquatic.

5 Discussion

In this research, the sequence conservation, phylogenetic relationships, and mutation rates of genes of the proteoglycan synthesis pathway were inspected. The study was initiated by first collecting all the proteogylcan synthesis pathway gene orthologs of *Homo sapiens* from Ensembl. *Homo sapiens* was chosen as the model organism as it contains all the 51 genes that are involved in the synthesis of proteoglycan. Ensembl was chosen as the source database as the genomes of the organisms present in Ensembl are annotated and curated. Of the 199 collected organisms, only 187 were used. The 12 organisms that were excluded were projection builds, which were removed due to the uncertainty associated with the correctness of gene presence. For example, *Vicugna pacos*' genome is reported to not contain *Acan*, which is a core protein-coding gene involved in synthesizing proteoglycan. Errors such as the previously mentioned example may serve to motivate additional effort expenditure in the area of genome annotation, which is critical for performing inferences and testing hypotheses.

Collected gene sequences were aligned, which resulted in 51 MSAs. The alignment results were enumerated by grouping genes according to their function as it was expected for genes with similar function to showcase intergroup alignment similarities. The group of core proteins showcase alignments that contain conserved regions across their entirety. Conservation in the alignments of core proteins suggests that the sequences of these core proteins is generally

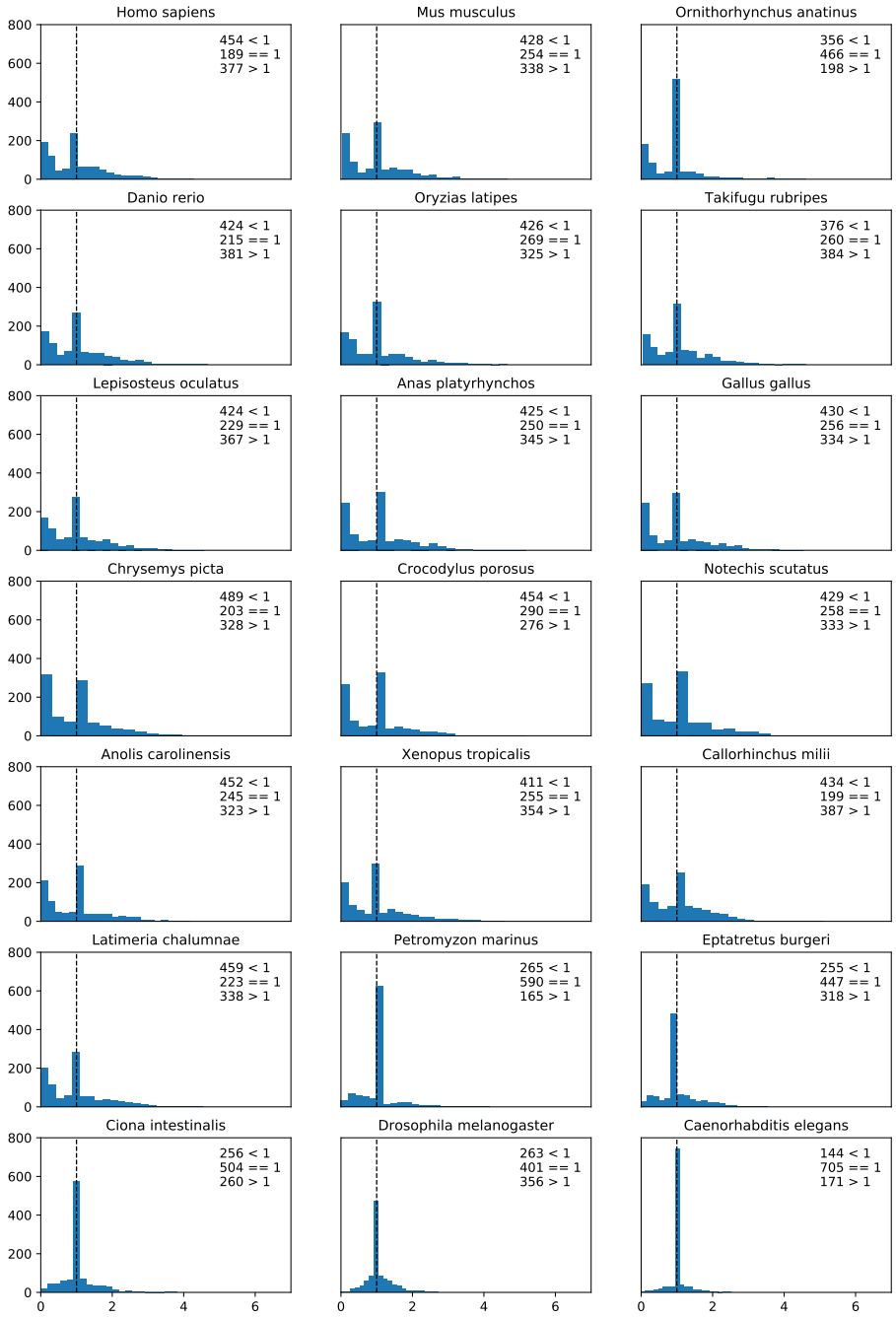

Figure 14: Histograms representing the distribution of dN/dS values for each significant organism. The histograms include a marker at the value of 1 to represent neutrality. The number of dN/dS values less than and greater than or equal to 1 are indicated.

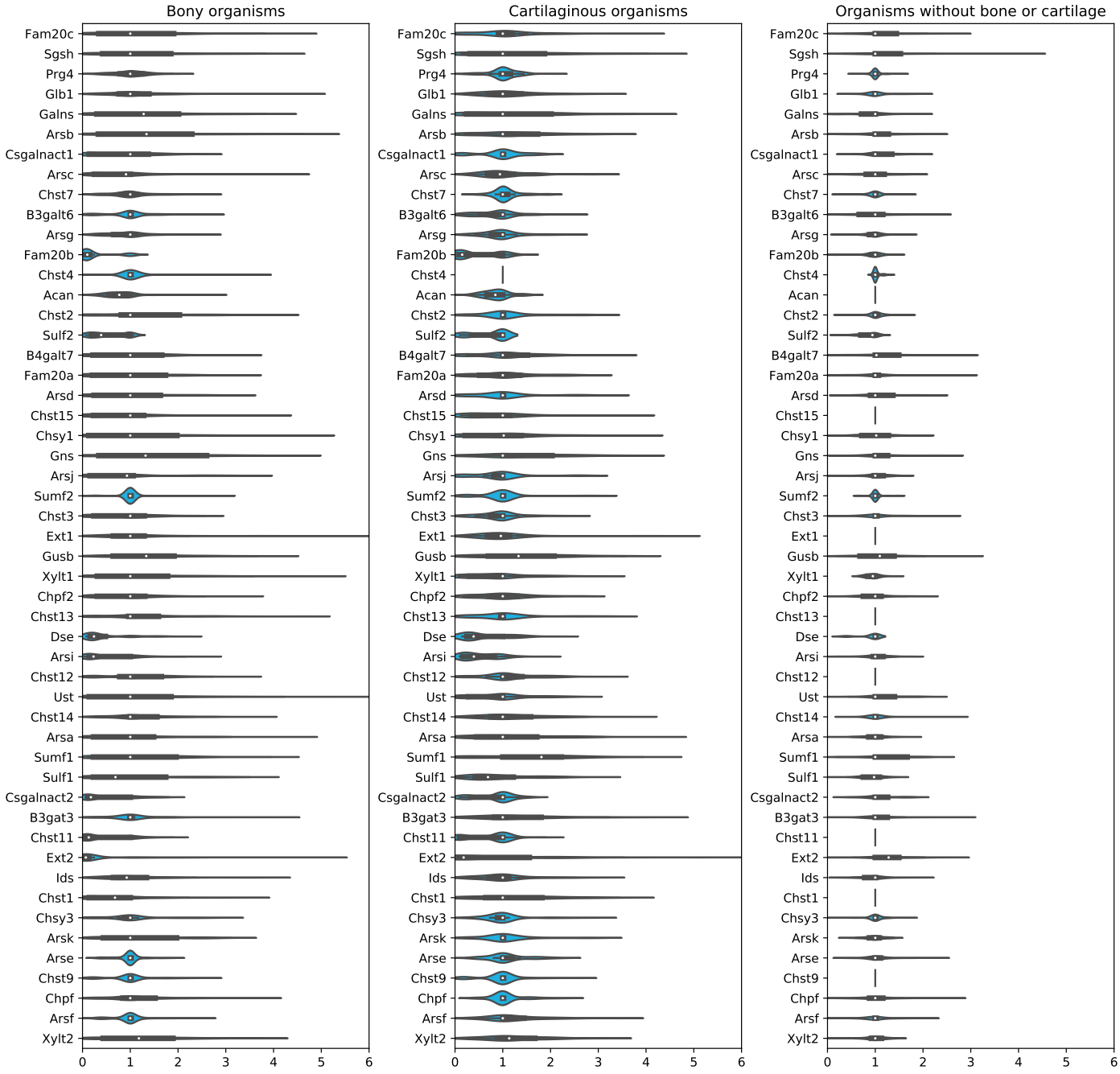

Figure 15: Violin plots for the dN/dS values of each independent gene binned by organism bone or cartilage presence presence.

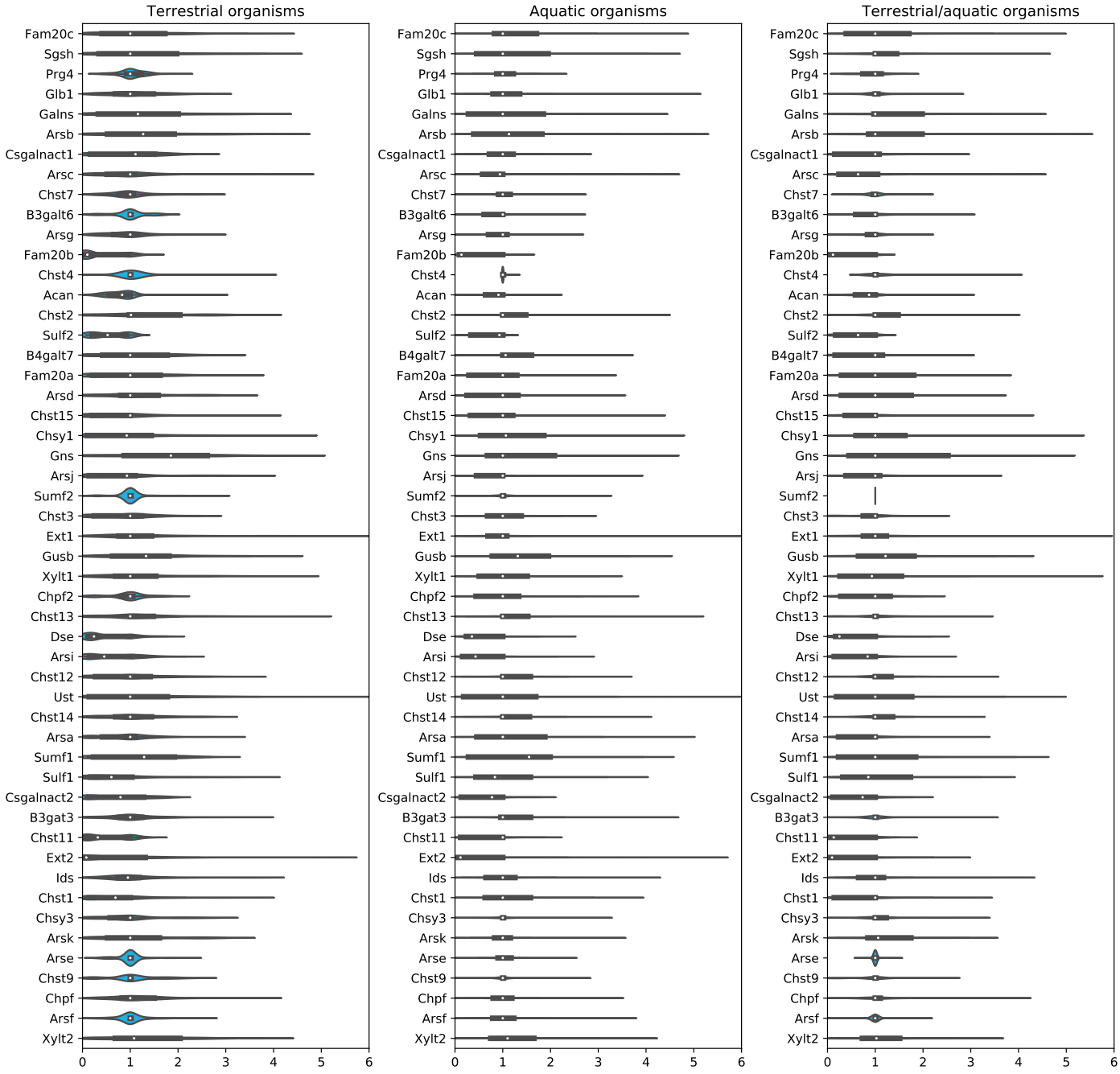

Figure 16: Violin plots for the dN/dS values of each independent gene binned by organism habitat.

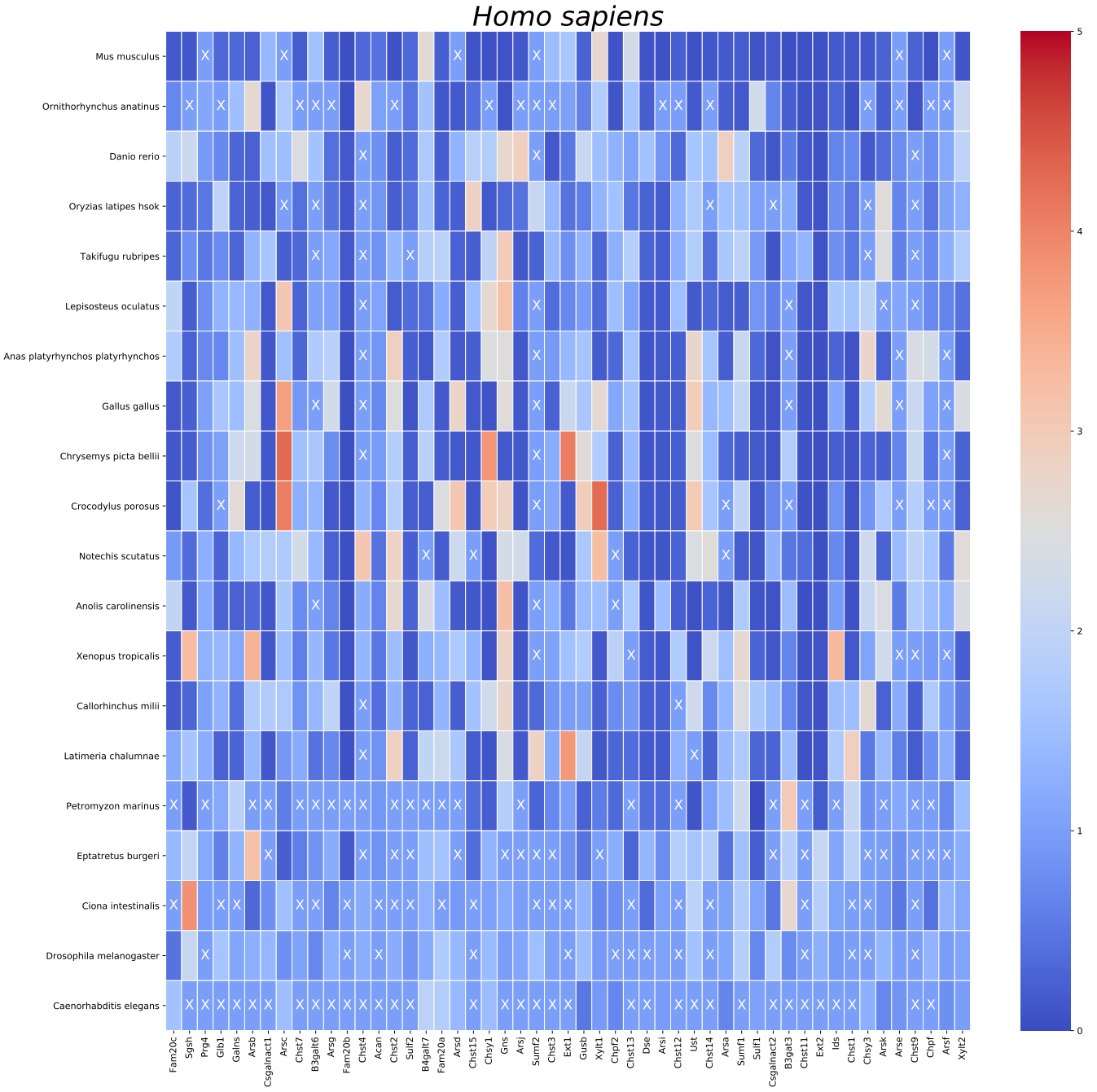

Figure 17: A sample dN/dS 2D grid for *Homo sapiens* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

undergoing negative selection (conservation). This is an expected outcome since it is accepted that *Acan* is an important protein-coding gene involved in binding water molecules within cartilage [6], suggesting it's sequence should be highly conserved across organisms. Xylosyltransferase showcased similar conservation patterns within the alignment but contained flanking regions that were less conserved. It is possible that sections of the sequence of *Xylt1* and *Xylt2* are critical for normal proteoglycan function, since xylosyltransferases catalyze the carbohydrate chain addition to core proteoglycan proteins [22]. Galactosyltransferase genes showcased several differences between sequence conservation patterns, which is inconsistent with previous findings that illustrate the importance of their conservation. Significant parts of alignments have not been identified to have high column similarity, as exemplified by the alignment of *B3galt6*. While it has been previously suggested that sequence conservation in galactosyltransferase genes is important, it may be manifested independently within genes rather than between them. Since the group of galactosyltransferases contains 6 genes (19), it is possible for gene redundancy to manifest as some genes may have higher transcription rates compared to others [23]. Glucuronyltransferases showcased similar sequence alignment patterns to galactosyltransferases but exhibited more sparsity, which suggests this group of genes may undergo higher mutation rates compared to galactosyltransferases. *Ext1* and *Ext2*, which are glycosyltransferase protein-coding genes, showcased unexpected alignments as *Ext1* exhibited highly conserved nucleotide regions across the entire alignment by comparison to *Ext2*, which showcased almost no conservation across the first 2000 nucleotides. Since *Ext1* binds *Ext2* to form a complex that modifies heparan sulfate, it is possible that *Ext1*'s amino acid sequence to contain multiple binding regions for *Ext2*, while *Ext2* contains one highly conserved region that *Ext1* attaches to with the first 2000 nucleotides performing a different function across organisms [24]. The 12 sulfotransferase genes that were included in the analysis showcased no general regularity in alignment patterns. However, *Chst7, Chst4, Chst12, Chst15*, and *Chst3* showcased a tendency for conserved region presences in the middle of the alignment while *Chst1, Chst11, Chst14, Chst12*, and *Chst13* showcased nucleotide region conservation in the second half of the alignment. It is possible that the conserved regions that are showcased by the two gene subgroups of sulfotransferases to contain critical amino acids that influence binding behaviour. It has been previously suggested that loss of *Chst3* function is associated with a mutation in R304, which corresponds to the conserved region showcased by the MSA of *Chst3* in the nucleotide region 500-1100 [25]. There are 16 sulfatases present in the proteoglycan synthesis pathway. The alignments of the 16 genes are similar to the ones observed in sulfotransferases - no general patterns within the group or between genes in the group. It is worth noting that all the alignments of the sulfotransferases group showcase an element of sparsity, which is supported by the presence of short and numerous nucleotide regions that showcase low column similarity suggesting how conservation. The *Fam* family of kinases have been extensively studied. The sequences of the *Fam* kinases showcase a higher number of conservation regions for *Fam20c*, which is consistent with previous research that provided evidence for the importance of conserved regions of *Fam20c* that serve to form dimers with *Fam20a*[10]. It is possible that the conserved nucleotide region observed towards the end of the alignment of *Fam20a* represents the region that is bound by *Fam20c*. By comparison to *Fam20c*, *Fam20b* showcases multiple regions of high conservation with a region of low conservation between nucleotides 300 and 800, which may be explained by the precedence in development of *Fam20b* ahead of *Fam20a*, as previously suggested [10]. The only epimerase that is involved in proteoglycan synthesis is *Dse*. The alignment of *Dse* showcases a large region that is conserved across all organisms' genetic sequences, with the exception of a region of 500 nucleotides at the beginning of the gene. In *Homo sapiens*, *Dse* is known to be present in multiple major tissues such as spinal cord, heart, intestines, kidney, etc. Given the highly conserved region after the first 500 nucleotides, it is possible for that region to code for important amino acids that are critical for the binding patterns

of the protein encoded by *Dse*. The evaluated glycosides are *Galns* and *Gusb*. By comparison to *Gusb*, *Galns* shows more sparsity in its MSA, suggesting this gene has a higher mutation frequency. Since both *Galns* and *Gusb* are responsible for the breakdown of glycosaminoglycans in the lysosome [26, 27], it is possible that this is another example of gene redundancy due to function and localization. The only evaluated sulfohydrolase is *Sgsh*. The alignment of *Sgsh* is comparable with that of *Galns* as it contains several regions that are conserved but also exhibits high sparsity. *Sgsh* performs a similar function to *Galns* and *Gusb* and is present in the same organelle - the lysosome [28]. It is also worth noting that *Homo sapiens* mutations in all three genes - *Galns, Gusb*, and *Sgsh* - lead to mucopolysaccharidosis disorders, supporting the notion of co-localization of the three genes and similar function.

The collection of taxonomic information resulted in 187 records representing individual taxonomic information for each organism in this study. The collection resulted in an unequal distribution of taxonomic groups, as outlined in Figure 13. Due to the high dimensionality of the data that would result from analyzing all of the organisms, this research used a subset of the organisms as a representative sample (Table 1). Organisms were chosen based on to their importance as models in biological research, accessibility to higher quality genome data for the models, and their representation of specific types of organisms. The evident high number of organisms from the clades Mammalia, Actinopterygii, and relatively high number of organisms in the clade Archelosauria, by comparison to the other clades, may motivate additional efforts and resources to be allocated towards sequencing organisms from clades with a low organisms representation. It is worth mentioning that the study searched and collected sequences that were orthologs of *Homo sapiens*. It is possible that other organisms from clades with a low representation in this study to be present in public databases but contain no orthologs of proteoglycan synthesis pathway genes of *Homo sapiens*.

The phylogenetic analysis and grouping resulted in a reduction from 51 phylogenetic trees containing gene presence markers to 15 (Figure 12). Phylogenetic analysis is typically used to infer and propose hypotheses regarding the co-evolution of collections of genes [13]. In addition, gene clustering based on vectors that indicate gene presence, as illustrated in Appendix Figure 20, is an additional approach for inferring gene co-evolution. Both phylogenetic analysis and gene clustering were pursued for identifying genes of the proteoglycan synthesis pathway that may have co-evolved. However, the evidence that has been identified is inconclusive. No clear consistencies have been observed between the clustering that resulted from grouping trees based on 75, 80, and 85% similarity (Algorithm 1) and the clustering presented in Appendix Figure 20.

The research started with the initial hypothesis that dN/dS values of organisms that synthesize bone and cartilage are higher by comparison to dN/dS values of organisms that do not synthesize bone and cartilage as a consequence of evolutionary pressures caused by environmental factors, such as the transition from water to land. The transition from water to land exposes organisms to the force of gravity by comparison to an aquatic environment, in which the primary environmental factor that may influence bone presence is the drag associated with movement through water. The summary of dN/dS distributions outlined in Figure 14 suggests that organisms that synthesize bone, such as *Homo sapiens, Mus musculus, Danio rerio*, etc, exhibit lower dN/dS values compared to the expected trend. The study hypothesized that this might be caused by higher conservation rates of important genes in the proteoglycan synthesis pathway for organisms that synthesize bone and cartilage. In addition, representative organisms that synthesize bone are higher-order organisms that are more complex by comparison to, for instance, *Ciona intestinalis*. Organisms that do not synthesize bone or cartilage, such as *Drosophila melanogaster*, may have higher dN/dS ratios because their rates of reproduction are much higher compared to more complex organisms, resulting in higher mutation rates over time. A comparison of dN/dS values from Figure 17 between *Homo sapiens* and *Drosophila melanogaster* and *Homo sapiens* and

Mus musculus is evidence of the difference between the conservation of proteoglycan synthesis pathway genes in *Homo sapiens* compared to the other two organisms - *Drosophila melanogaster* showcases higher dN/dS values compared to *Mus musculus*. It is also worth mentioning organisms that synthesize bone and/or cartilage, such as *Homo sapiens*, exhibit a bimodal distribution, as outlined by Figure 14, which may be caused by a collection of low dN/dS values caused by comparing complex organisms with other complex organisms, such as *Homo sapiens* and *Mus musculus*, which result in lower values due to conservation of genes, and *Homo sapiens* and *Drosophila melanogaster*, which differ significantly and have higher dN/dS values. The group found no significantly different gene distributions while inspecting the dN/dS distributions created based on grouping significant organisms based on bone and cartilage presence, or absence thereof, and habitat such as terrestrial, aquatic, or both.

6 Conclusions

This research started with the hypothesis that organisms that synthesize bone and cartilage have higher mutation rates, as exhibited by dN/dS ratios, compared to organisms that do not. We have performed a multiple sequence analysis of all the genes across 187 of 199 organisms whose gene sequences were collected from Ensembl. A phylogenetic analysis was performed with the intent of finding genes that have potentially co-evolved. Of all the organisms that were collected, a subset was selected to for analysis. Surprisingly, the opposite trend to the expected one was identified - organisms that synthesize bone and cartilage have lower dN/dS ratios compared to organisms that do not. Lastly, the evidence for significantly different dN/dS distributions of genes of the proteoglycan synthesis pathway is inconclusive and cannot be used to suggest specific genes that may be investigated further as an attempt to identify genes that may be important to research for enriching the field's knowledge of osteoarthritis.

7 Future Work

In light of current results, it may be beneficial for future work to investigate genes independently, similar to how previous work focused on the *Fam20* family of genes [10]. It is possible that studying an ensemble of genes concurrently adds too much noise to data, resulting in most distributions to tend to normality, which can be observed in the trends presented in Figure 14 - there are numerous values centred at 1. In addition, it may be beneficial for future work to focus on small subsets of genes, such as suflotransferases, to investigate how they have co-evolved and whether evolutionary patterns are indicative of gene redundancy, which may lead to function loss. Effort should be directed at investigating groups of genes similarly to how this research has conducted an analysis of the whole proteoglycan synthesis pathway. Furthermore, different approaches of grouping genes, such as grouping based on organelle presence, may result in different perspectives, which may offer further insight into the evolution of the pathway.

References

[1] A. R. Poole. Proteoglycans in Health and Disease: Structures and Functions. *Biochemistry*, 236:1–14, 1986.

[2] D. S. Brown and B. F. Eames. Emerging Tools to Study Proteoglycan Function During Skeletal Development. In *Methods in Cell Biology*, chapter 15, pages 485–530. Elsevier Press, Amsterdam, The Netherlands, 2016.

[3] A. Aspberg. The Different Roles of Aggrecan Interaction Domains. *Journal of Hystochemistry & Cytochemistry*, 12(60):987–996, 2012.

[4] C. A. Pataki J. R. Couchman. An Introduction to Proteoglycans and Their Localization. *Journal of Hystochemistry & Cytochemistry*, 12(60):885–897, 2012.

[5] D. Chen et al. Osteoarthritis: Toward a Comprehensive Understanding of Pathological Mechanism. *Bone Research*, 5, 2017.

[6] T. Hardingham and M. Bayliss. Proteoglycans of Articular Cartilage: Changes in Aging and in Joint Disease. *Seminars in Arthritis and Rheumatism*, 20(3):12–33, 1990.

[7] J. Nam M. Maldonado. The Role of Changes in Extracellular Matrix of Cartilage in the Presence of Inflammation on the Pathology of Osteoarthritis. *BioMed Research International*, 2013.

[8] et al A. A. Young. Regional Assessment of Articular Cartilage Gene Expression and Small Proteoglycan Metabolism in an Animal Model of Osteoarthritis. *Arthritis Research & Therapy*, 7(4), 2005.

[9] L. Schaefer R. V. Iozzo. Proteoglycan Form and Function: A Comprehensive Nomenclature of Proteoglycans. *Matrix Biology*, 42:11–55, 2015.

[10] H. Zhang et al. Structure and Evolution of the Fam20 Kinases. *Nature Communications*, 9, 2018.

[11] M. Heino et al. A. P. Hendry, M. T. Kinnison. Evolutionary Principles and Their Practical Application. *Evolutionary Applications*, 4:159–183, 2011.

[12] R. D. Sleator J. Daugelaite, A. O'Driscoll. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. *International Scholarly Research Notices*, 13, 2013.

[13] O. Keskin et al. Predicting Protein-Protein Interactions From the Molecular to the Proteome Level. *Chemical Reviews*, 116:4884–4909, 2016.

[14] Jean-Francois Gout. Molecular and Genome Evolution by Dan Graur. *The Quarterly Review of Biology*, 92(4):476–477, 2017.

[15] Z. Yang and J. P. Bielawski. Statistical Methods for Detecting Molecular Adaptation. *Trends in Ecology & Evolution*, 15(12):496–503, 2000.

[16] J. Lee et al. F. Madeira, Y. M. Park. The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Research*, 47:636–641, 2019.

[17] J. D. Hunter. Matplotlib: A 2d Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[18] D. A. Benson et al. E. W. Sayers, T. Barrett. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37:5–15, 2009.

[19] D. J. Lipman et al. D.A. Benson, I. Karsch-Mizrachi. GenBank. *Nucleic Acids Research*, 37:26–31, 2009.

[20] P. Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11:37–50, 1912.

[21] J. Claude et al. E. Paradis, B. Bolker. Package 'ape'. *Online*, 2011.

[22] et al R. Hao, Z. Zheng. Cloning and characterization of O-xylosyltransferase gene from *Pinctada fucata martensii*. *Journal of Applied Animal Research*, 47(1):408–416, 2019.

[23] M. Nakajima et al. Mutations in B3GALT6, which Encodes a Glycosaminoglycan Linker Region Enzyme, Cause a Spectrum of Skeletal and Connective Tissue Disorders. *American Journal of Human Genetics*, 92(6):927–934, 2013.

[24] C. McCormick et al. The putative tumor suppressors EXT1 and EXT2 form a stable complex that accumulates in the Golgi apparatus and catalyzes the synthesis of heparan sulfate. *Proceedings of the National Academy of Sciences*, 97(2):668–673, 2000.

[25] H. Thiele et al. Loss of chondroitin 6-O-sulfotransferase-1 function results in severe human chondrodysplasia with progressive spinal involvement. *Proceedings of the National Academy of Sciences*, 101(27):10155–10160, 2004.

[26] S. Tomatsu et al. Mutation and polymorphism spectrum of the GALNS gene in mucopolysaccharidosis IVA (Morquio A). *Human Mutation*, 26(6):500–512, 2005.

[27] R. Vervoot et al. Molecular analysis of patients with beta-glucuronidase deficiency presenting as hydros fettles or as early mucopolysaccharidosis VII. *American Journal of Human Genetics*, 58(3):457–471, 1996.

[28] A. Meyer et al. The mutation p.Ser298Pro in the sulphamidase gene (SGSH) is associated with a slowly progressive clinical phenotype in mucopolysaccharidosis type IIIA (Sanfilippo A syndrome). *Human Mutations*, 29(5), 2008.

A Genes

This section includes supplementary material generated for visualizing organism gene frequencies, gene functions, and genes' clustering.

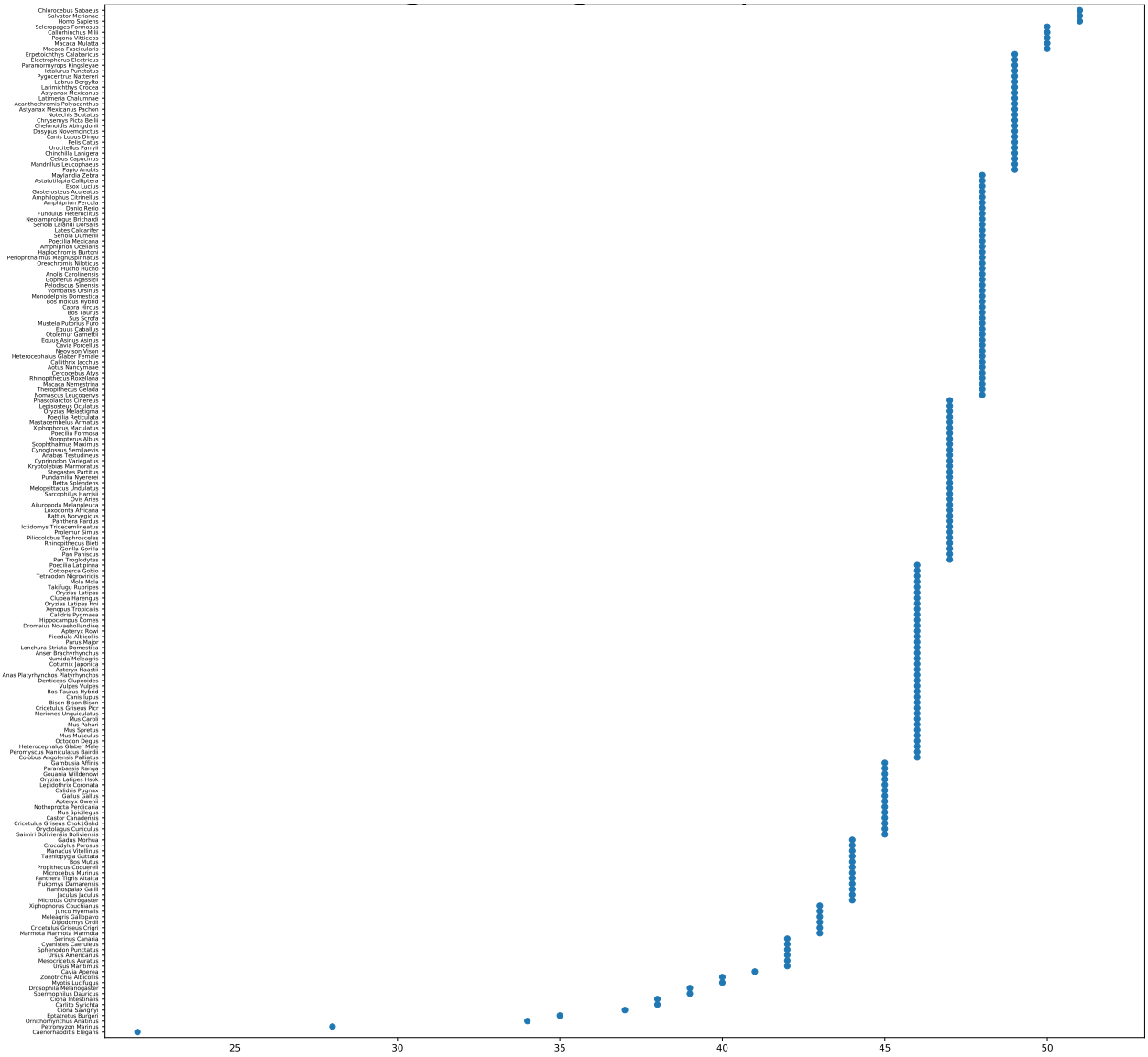

Figure 18: A plot of gene organisms vs. gene frequencies. The organisms were sorted in ascending order based on the number of genes their genomes include out of the 51 gene present in the proteoglycan synthesis pathway.

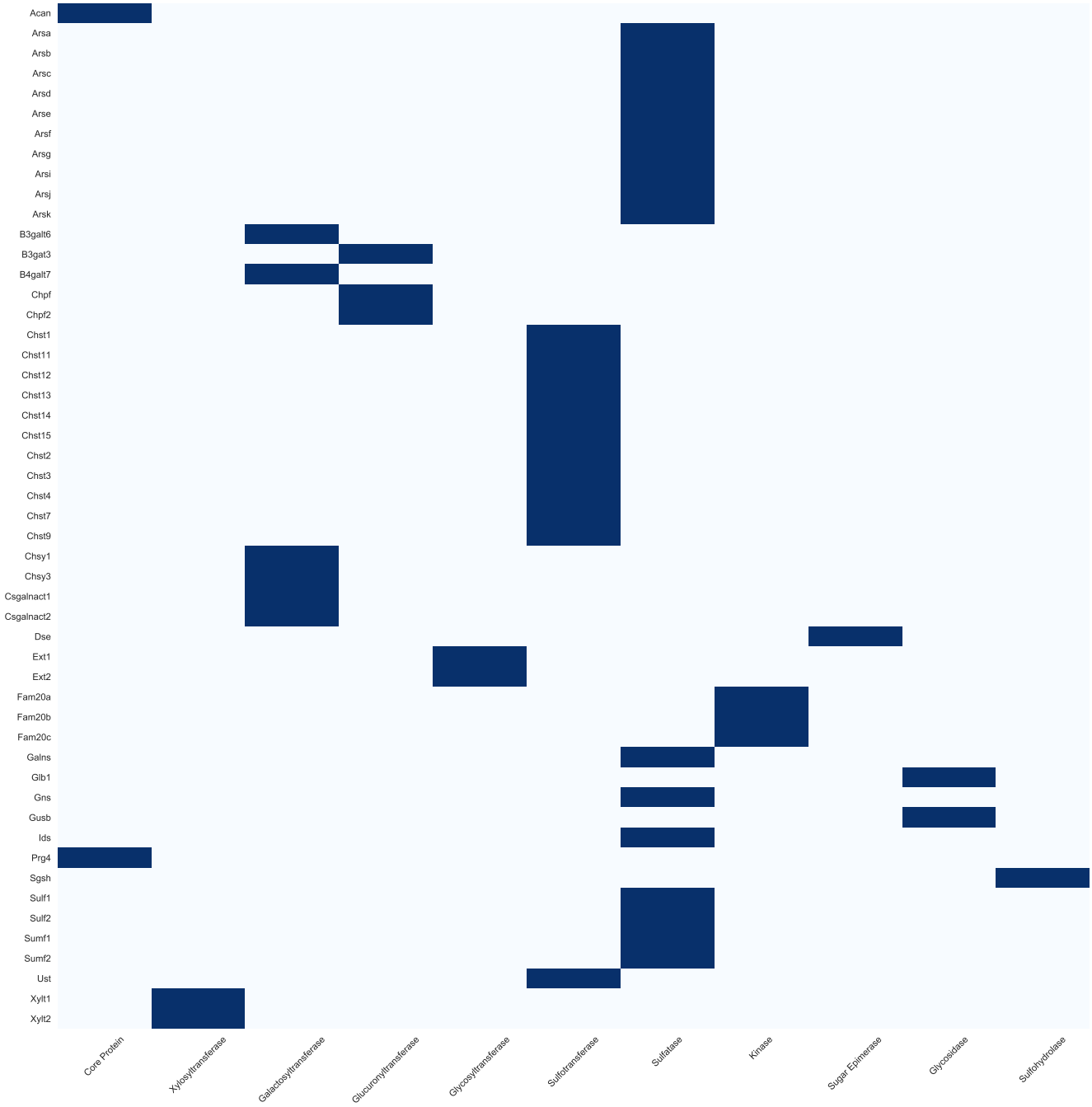

Figure 19: A matrix representation of gene function. Gene are plotted against known functions. Dark blue coloured cells indicate that gene on axis y performs the corresponding x axis function.

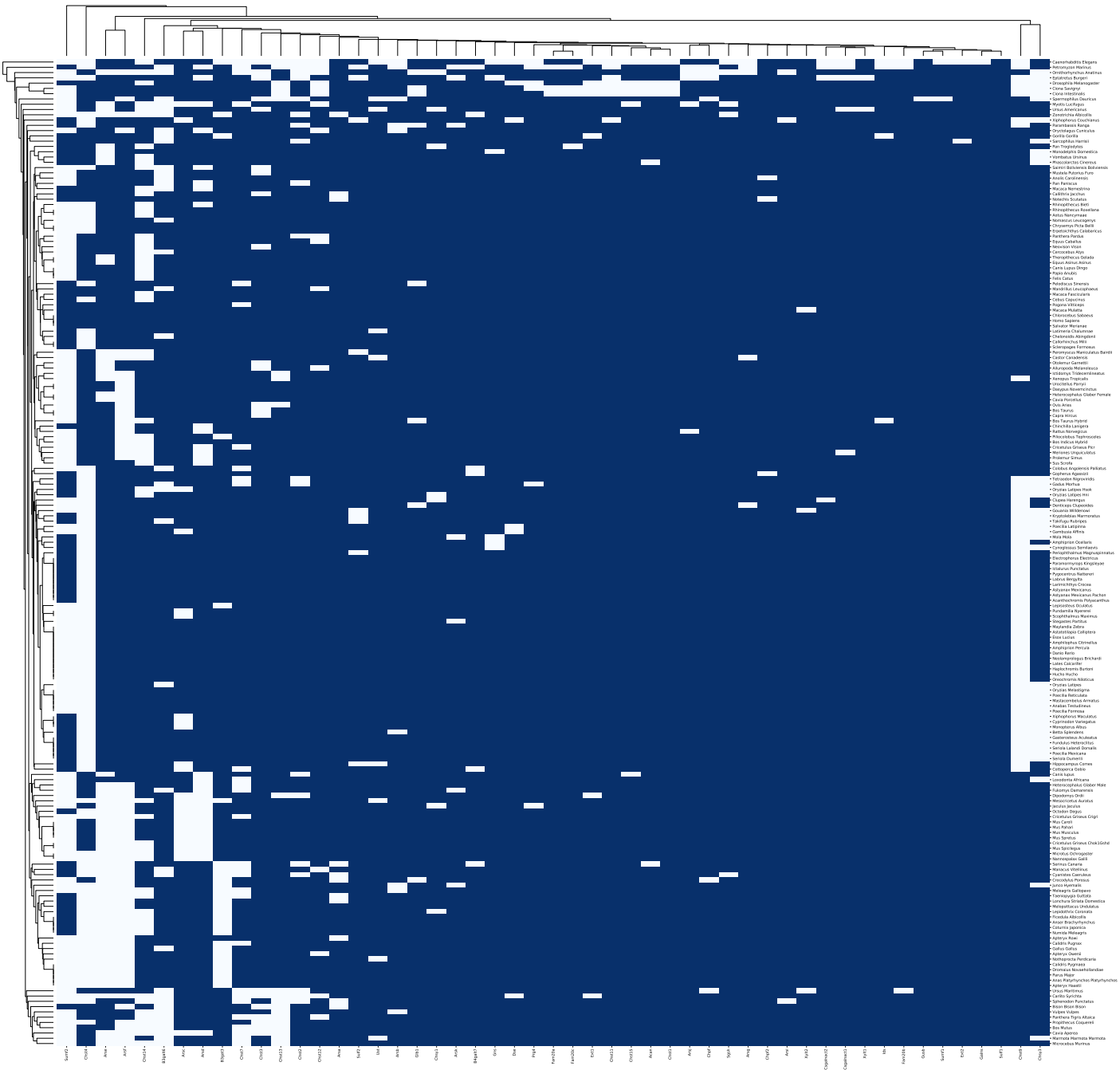

Figure 20: A clustermap of organisms vs. genes. Dark blue coloured cells indicate that the genome of the organism listed on the y axis contains the corresponding gene on the x axis. Organisms were clustered based on gene presence using the Euclidean distance between binary vectors indicating what genes are present in a specific genome.

B Phylogeny

This section includes supplementary material generated for visualizing paired phylogenetic trees.

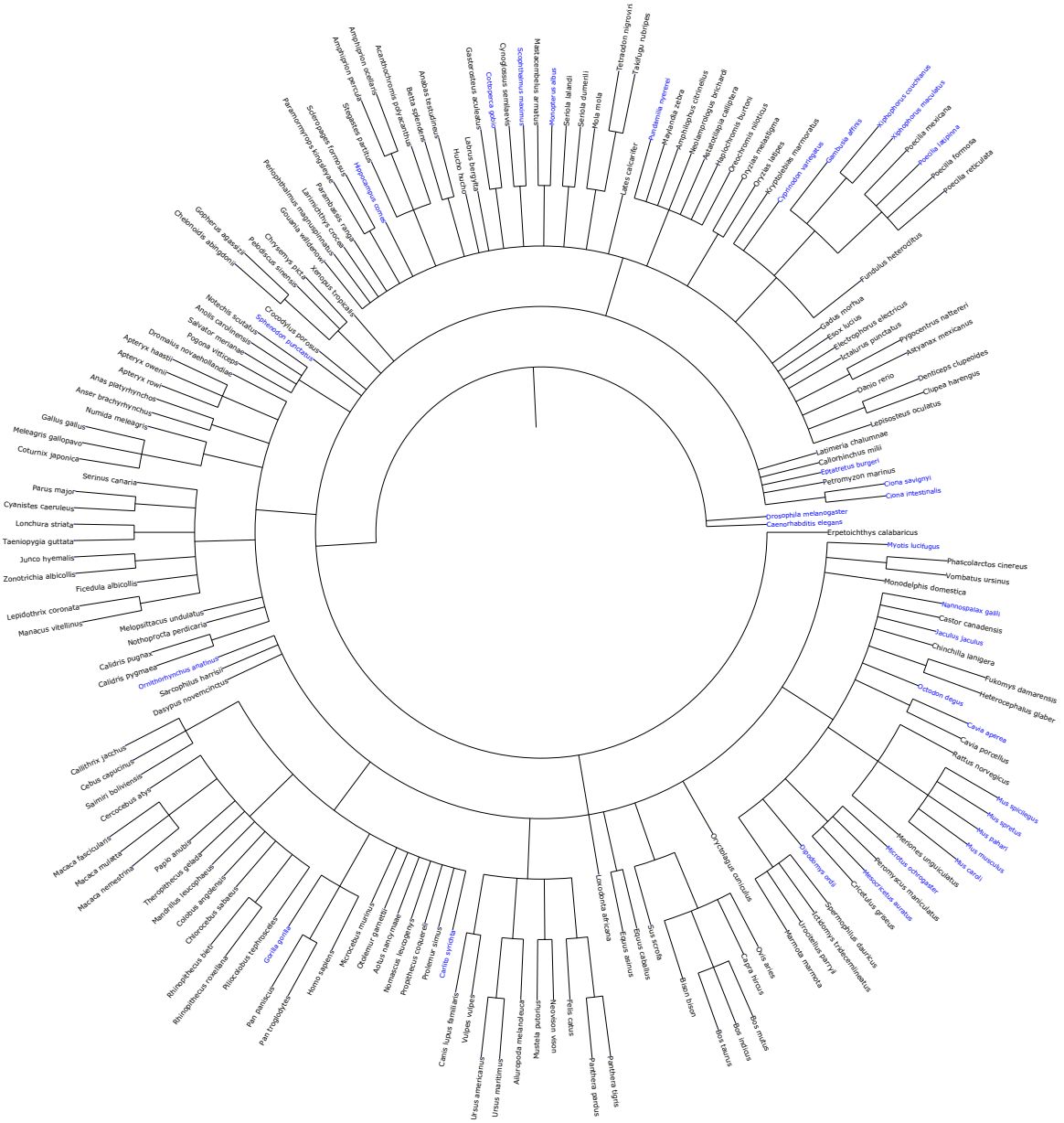

Figure 21: The paired gene tree for *Arsc*, *Arsi*, *Dse*, and *Ext1*.

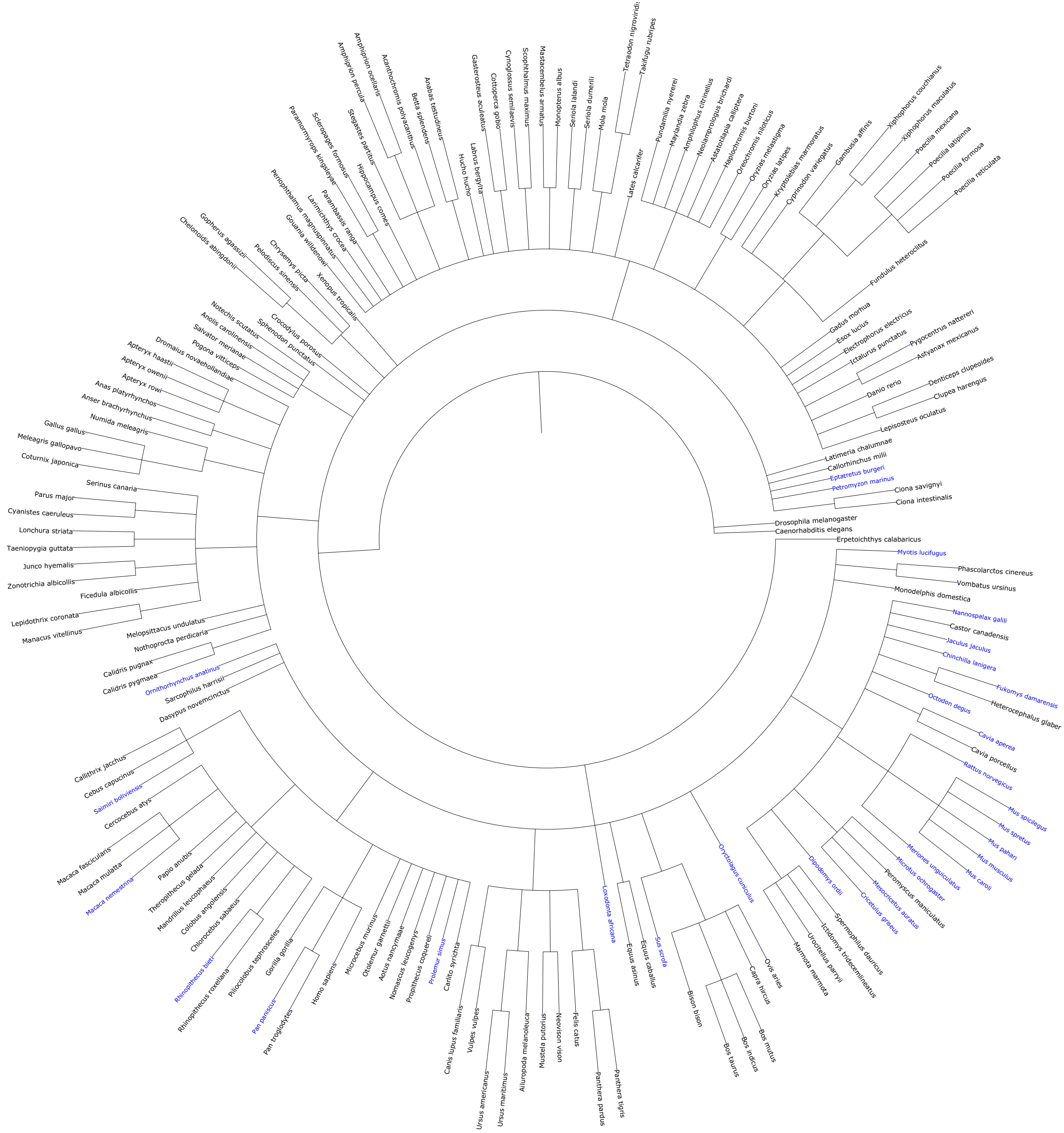

Figure 22: The paired gene tree for *Arsd*, and *Arsj*.

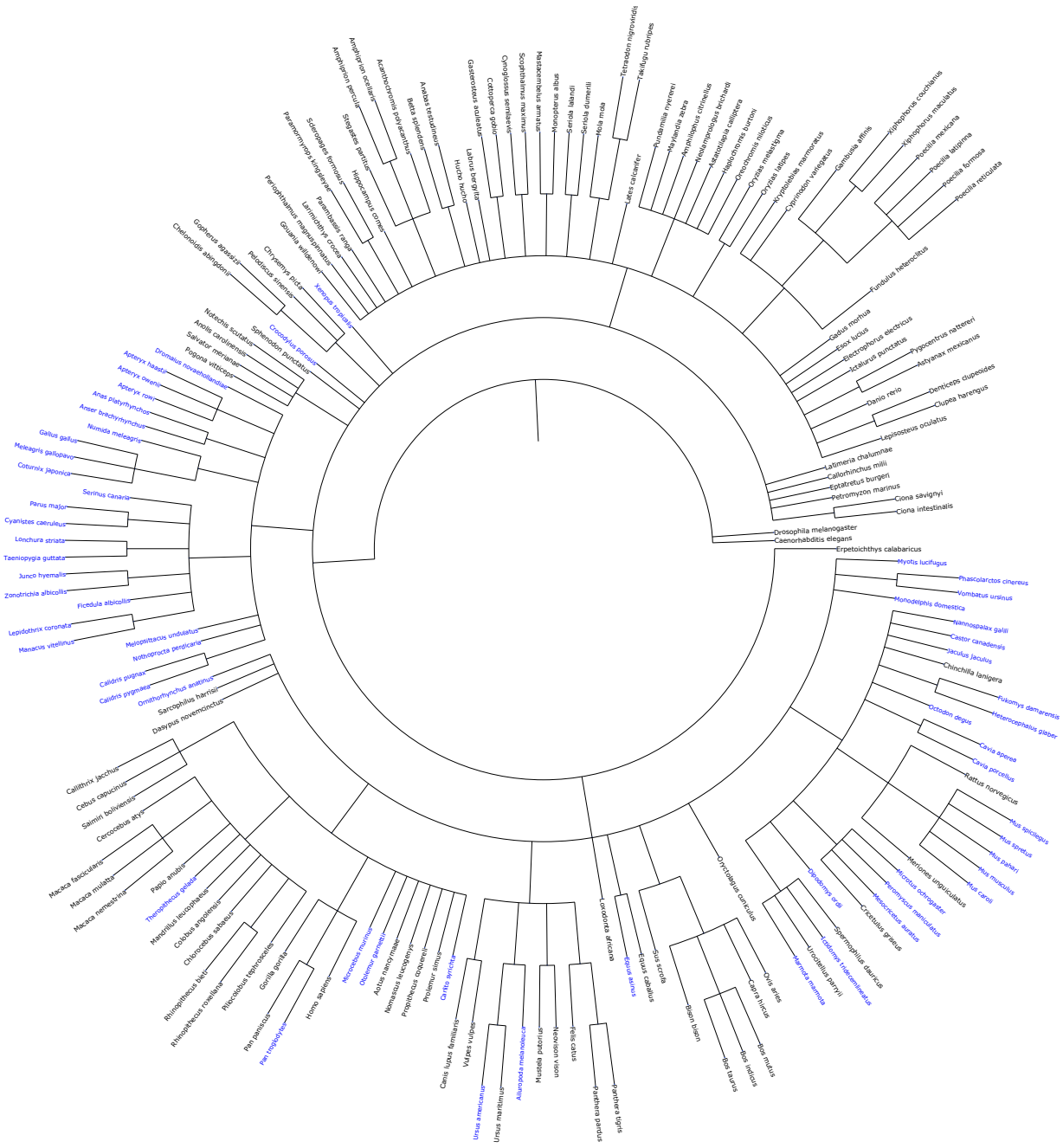

Figure 23: The gene tree for *Arse*.

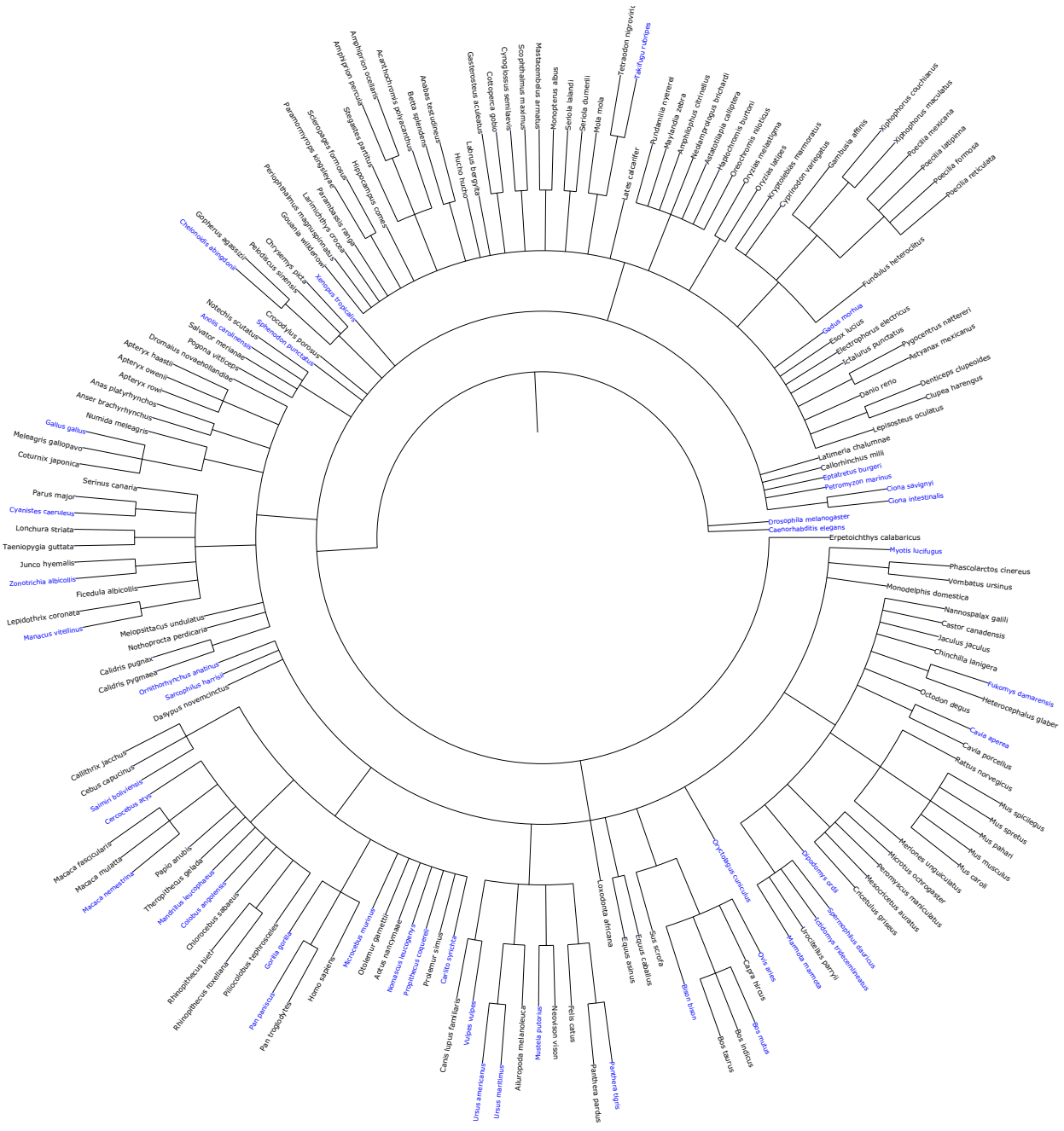

Figure 24: The gene tree for *B3galt6* and *Chst13*.

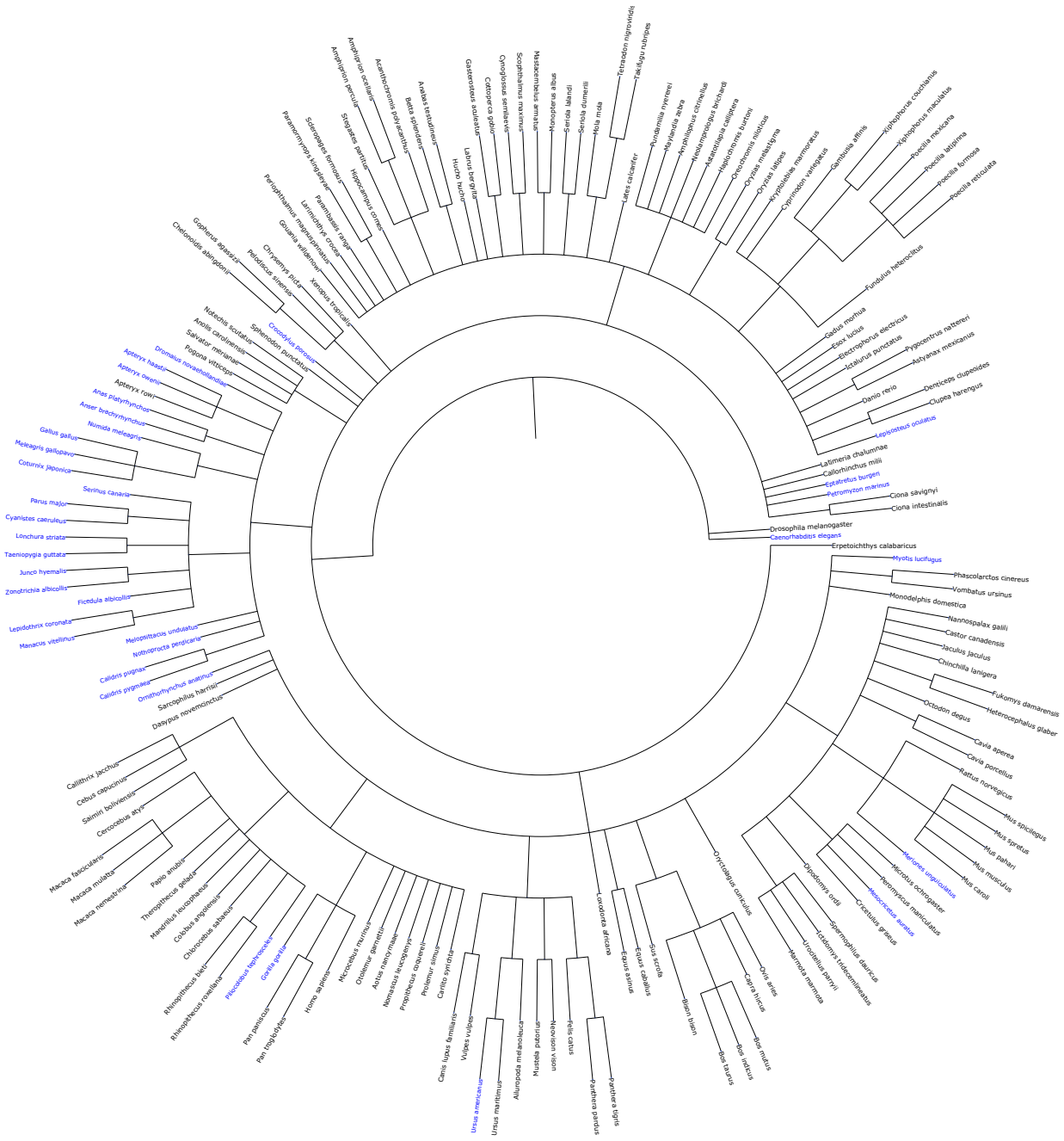

Figure 25: The gene tree for *B3gat6*, *Sgsh*, *Csgalnact1* and *Ids*.

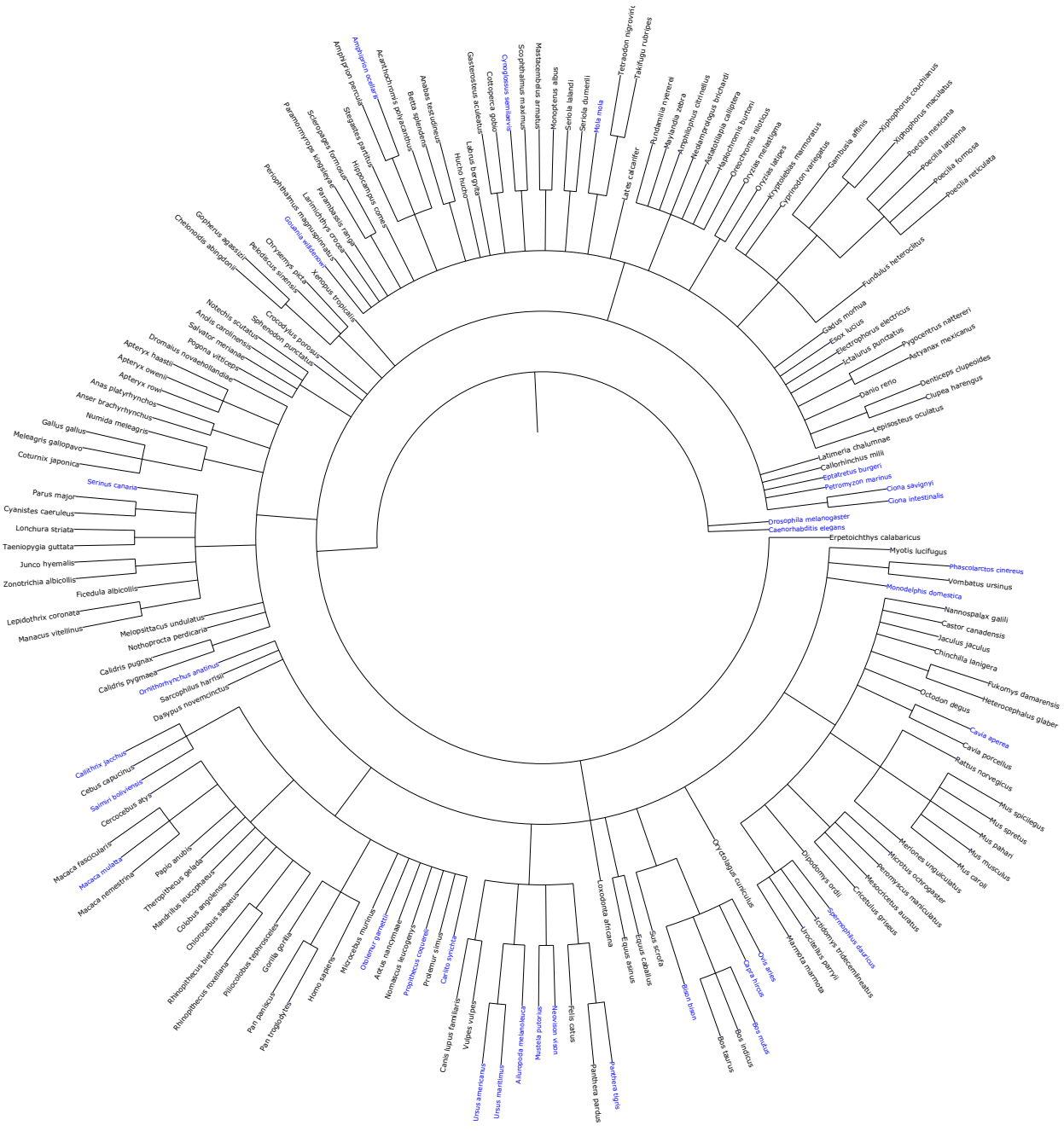

Figure 26: The gene tree for *Chst3*, *Xylt2*, *Gns*, *Acan* and *Chst1*.

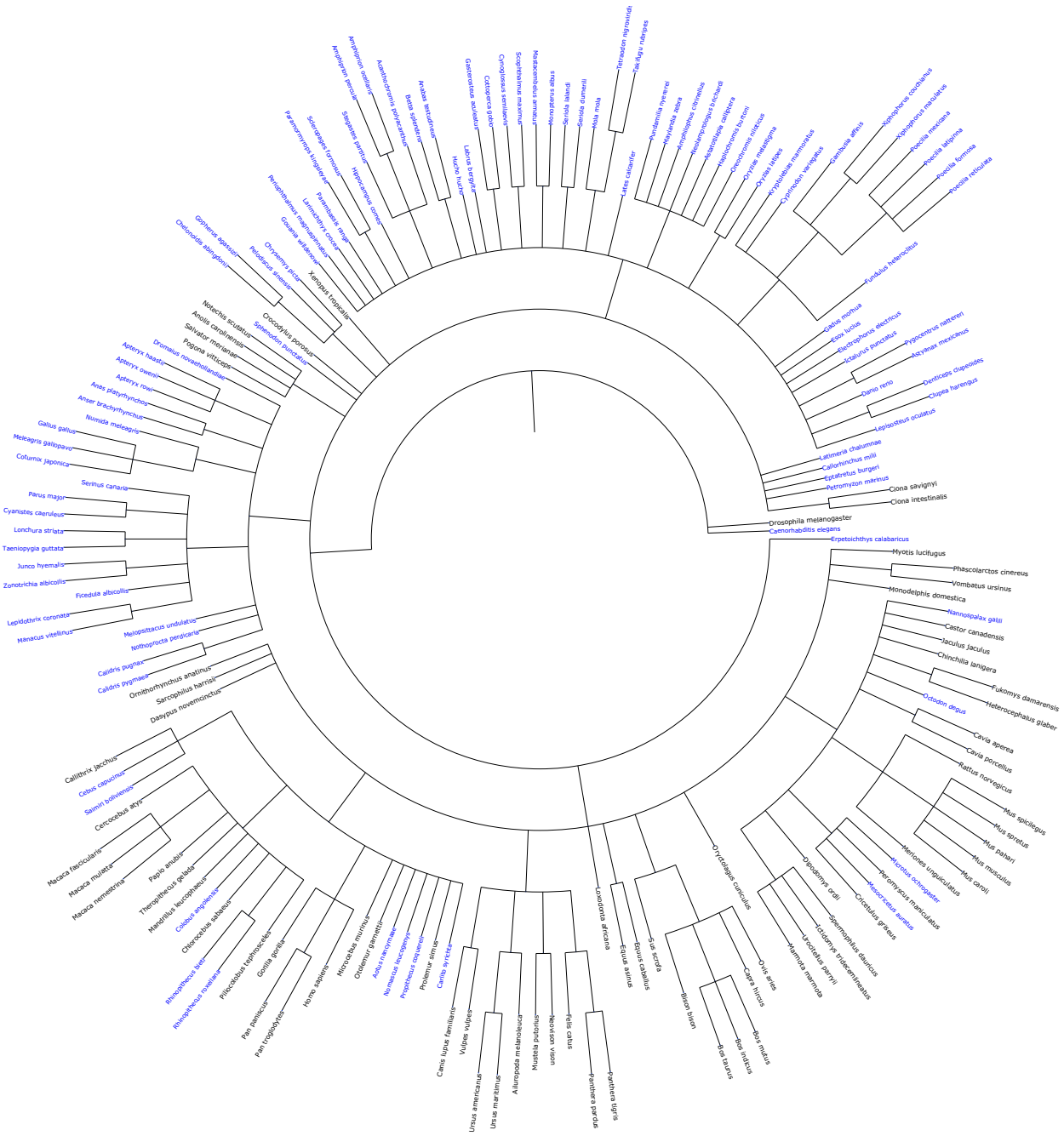

Figure 27: The gene tree for *Chst4*.

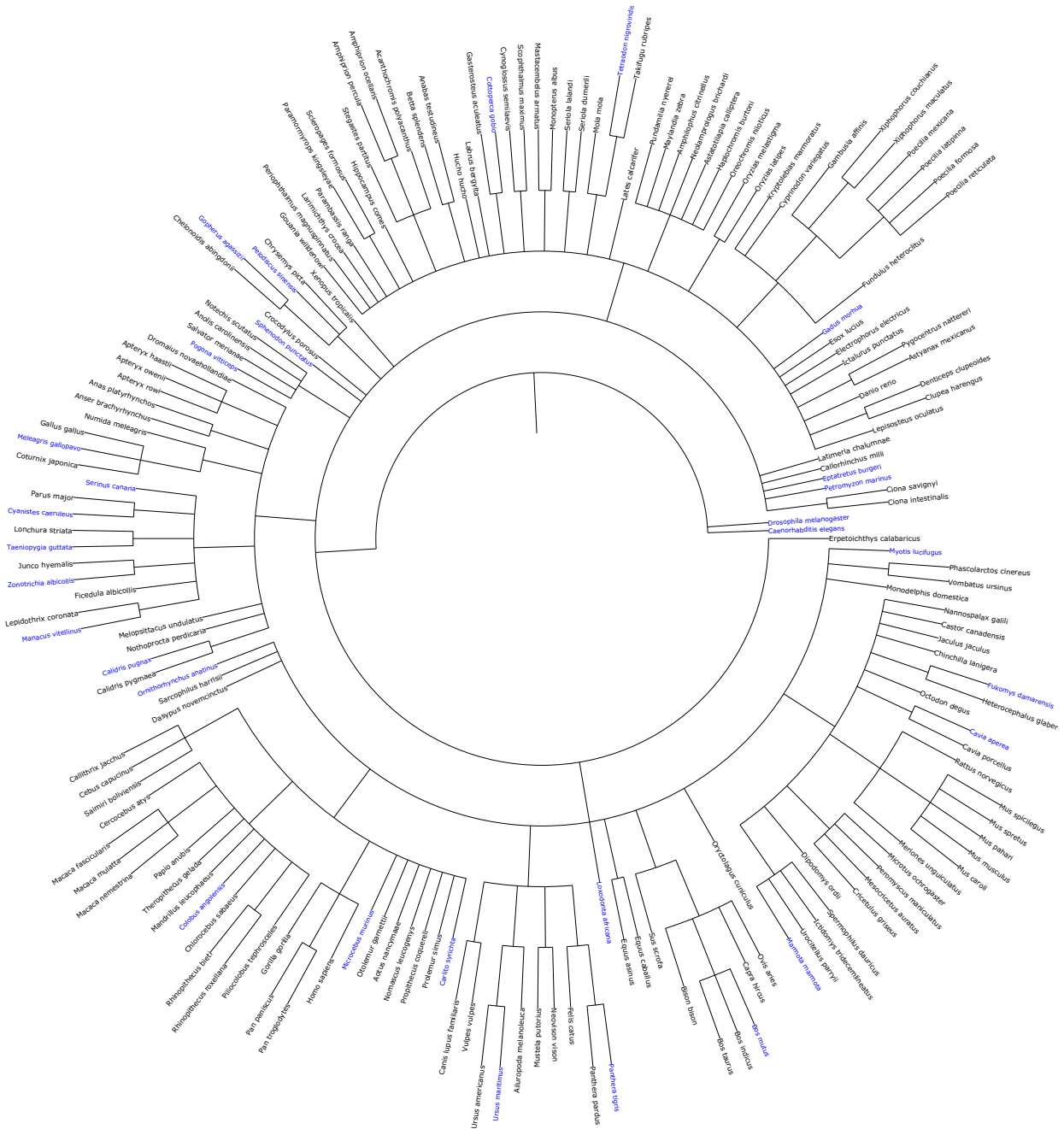

Figure 28: The gene tree for *Chst7*, *B4galt7*, and *Fam20b*.

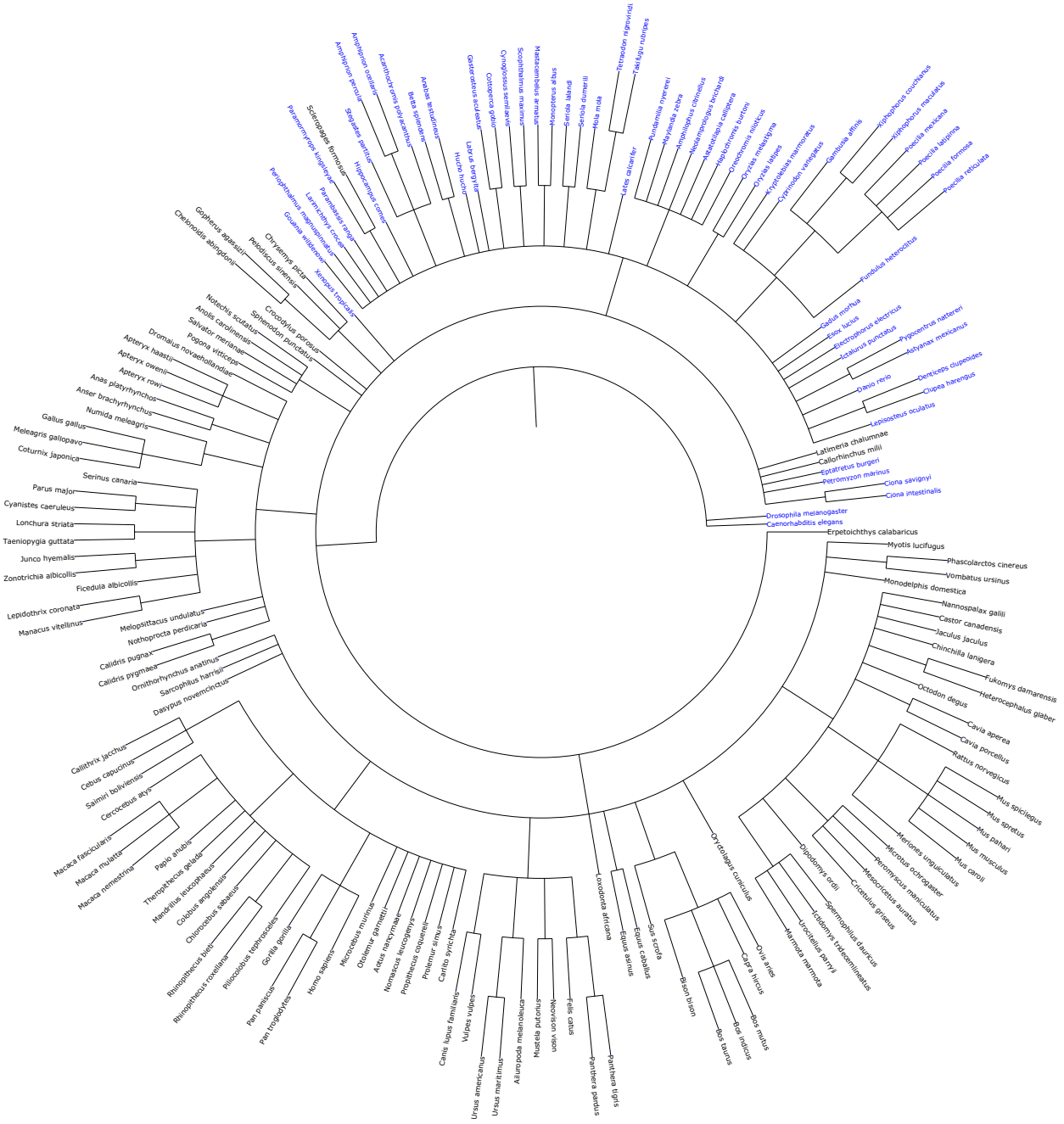

Figure 29: The gene tree for *Chst9*.

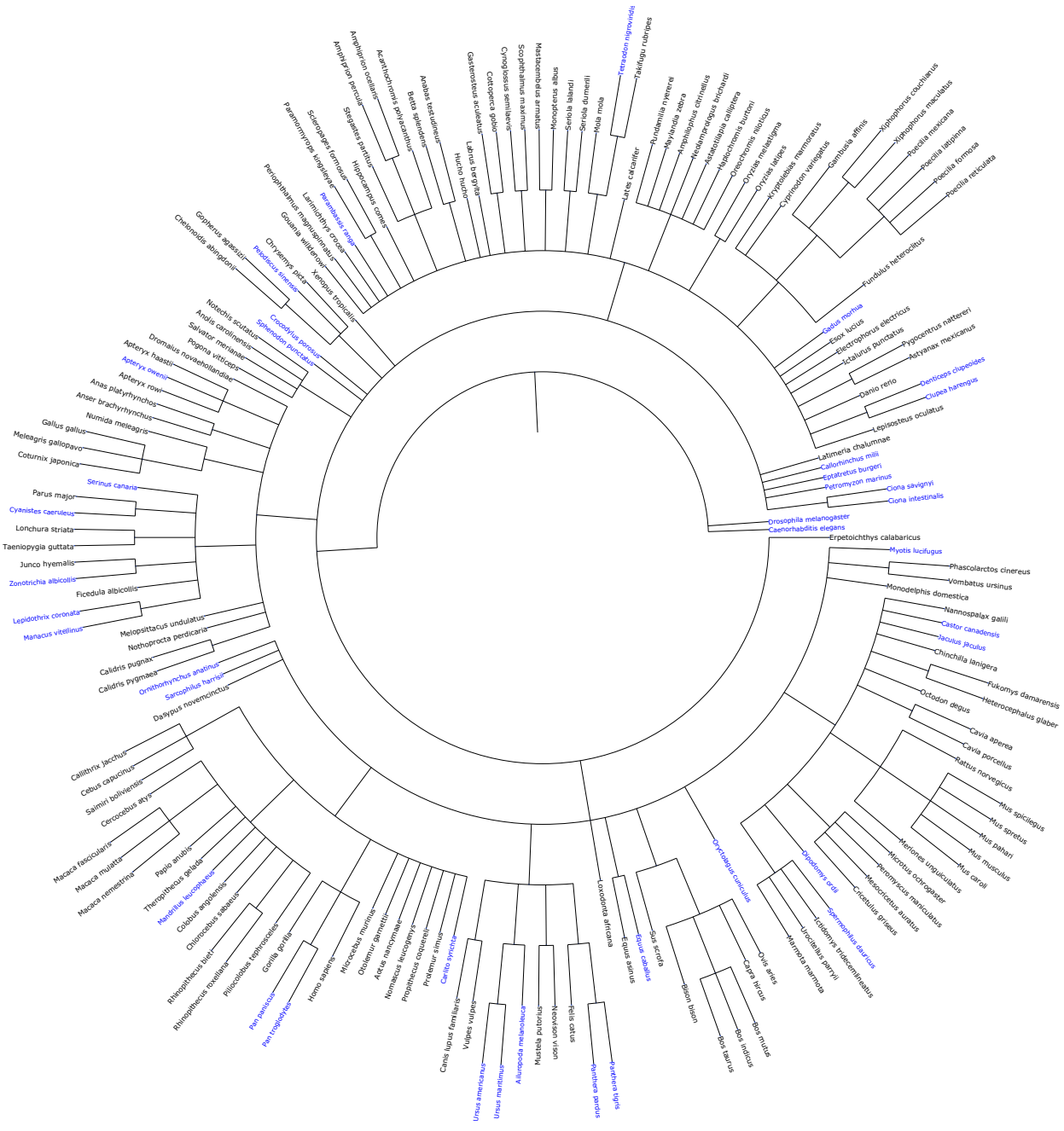

Figure 30: The gene tree for *Chst12*, *Chst15*, *Gib1*, *Arsg*, *Fam20a*, *Chst2*, *Chpf*, *Gusb*, *Chsy1* and *Sumf1*

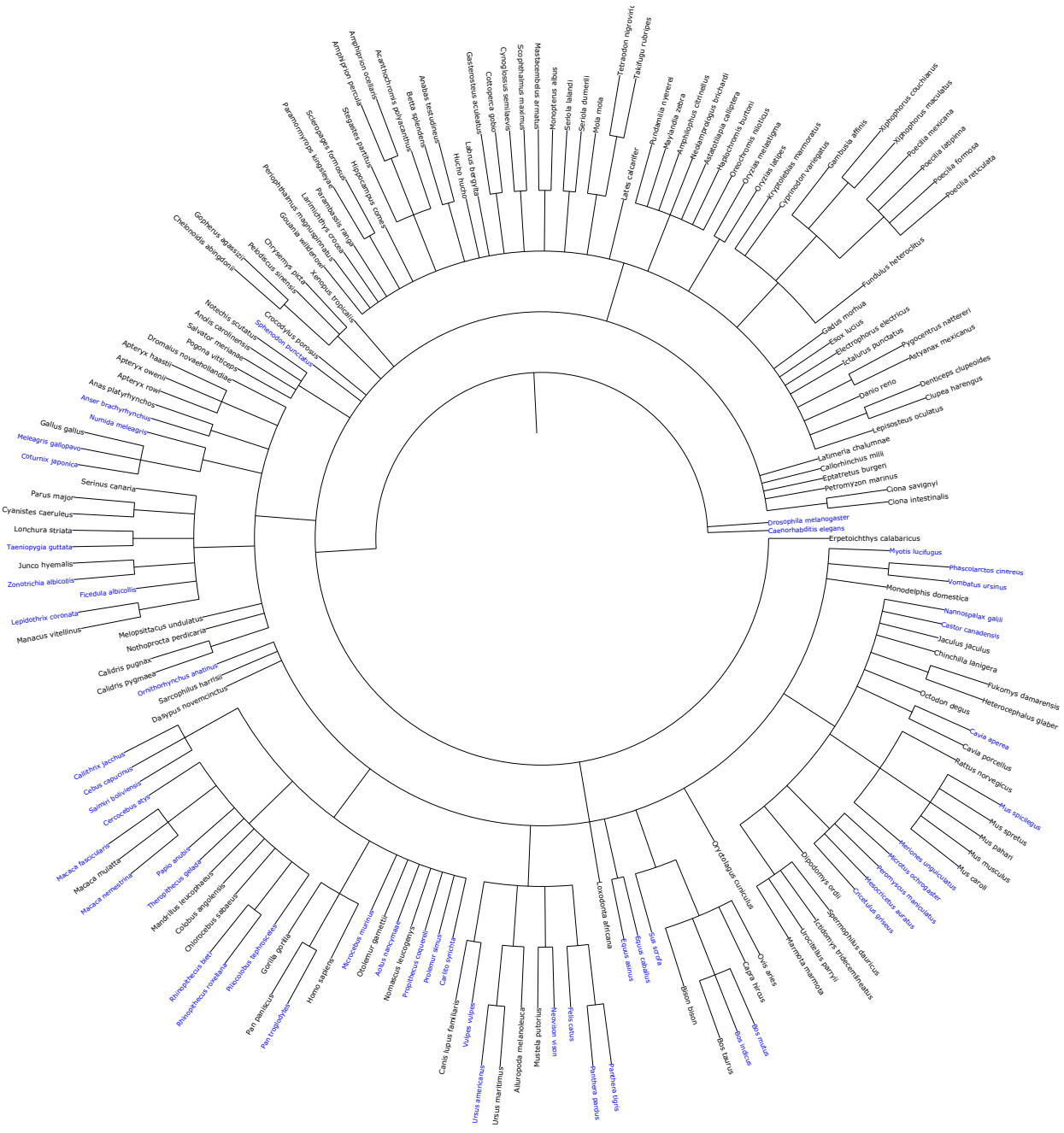

Figure 31: The gene tree for *Chst14*.

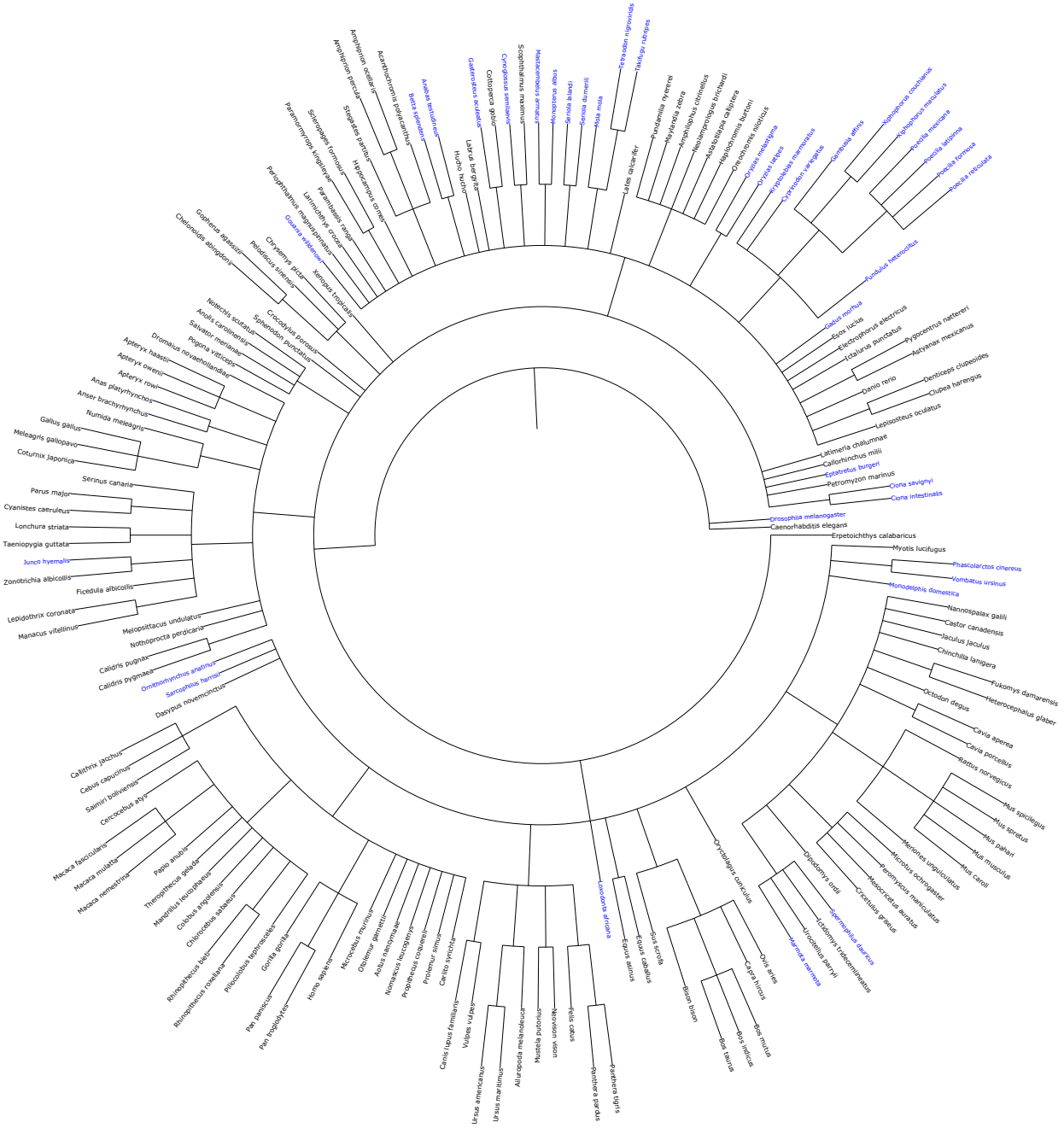

Figure 32: The gene tree for *Chsy3*.

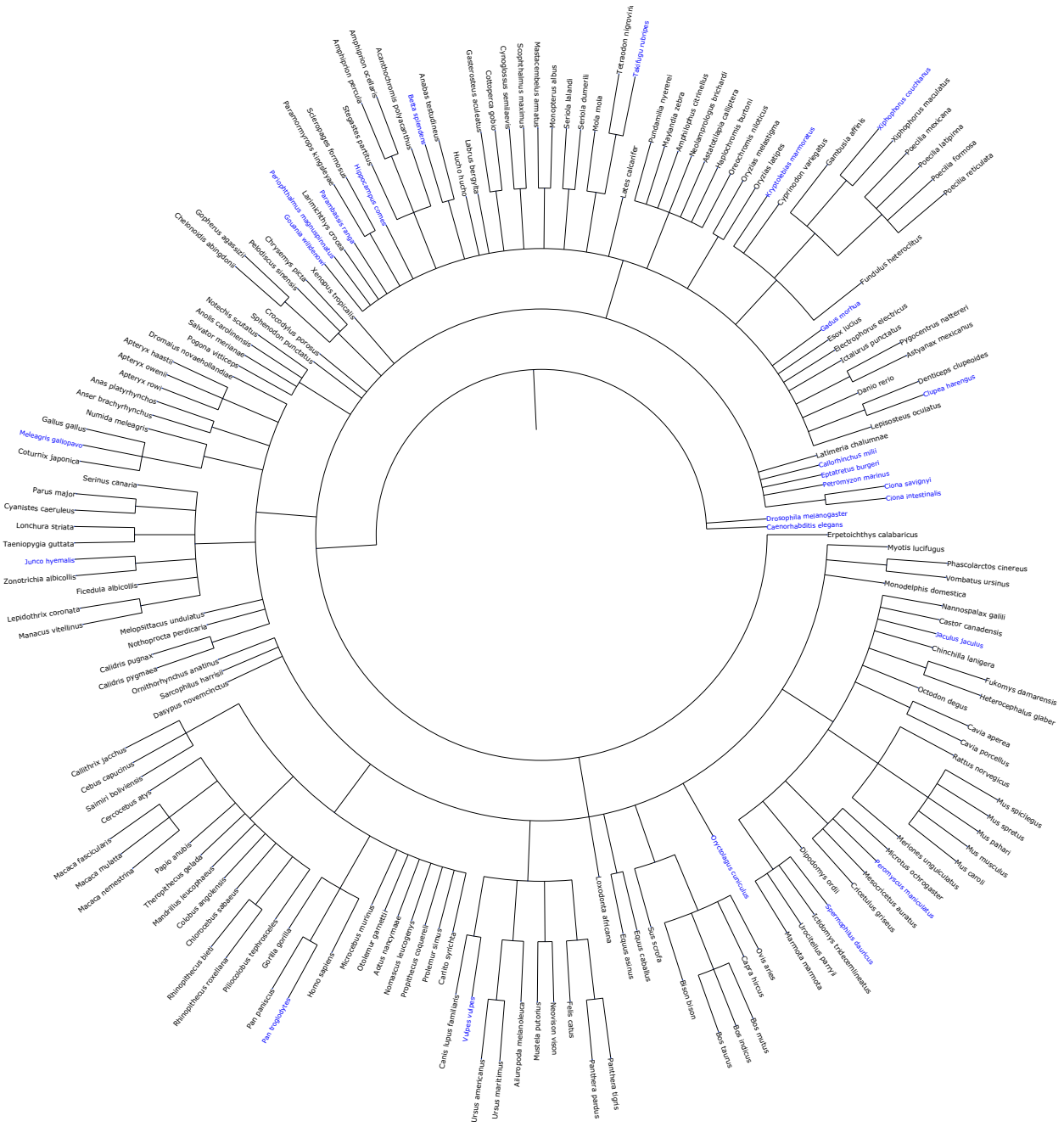

Figure 33: The gene tree for *Sulf2*, *Chst11*, *Csgalnact2*, *Arsb*, *Galns*, *Fam20c*, and *Prg4*.

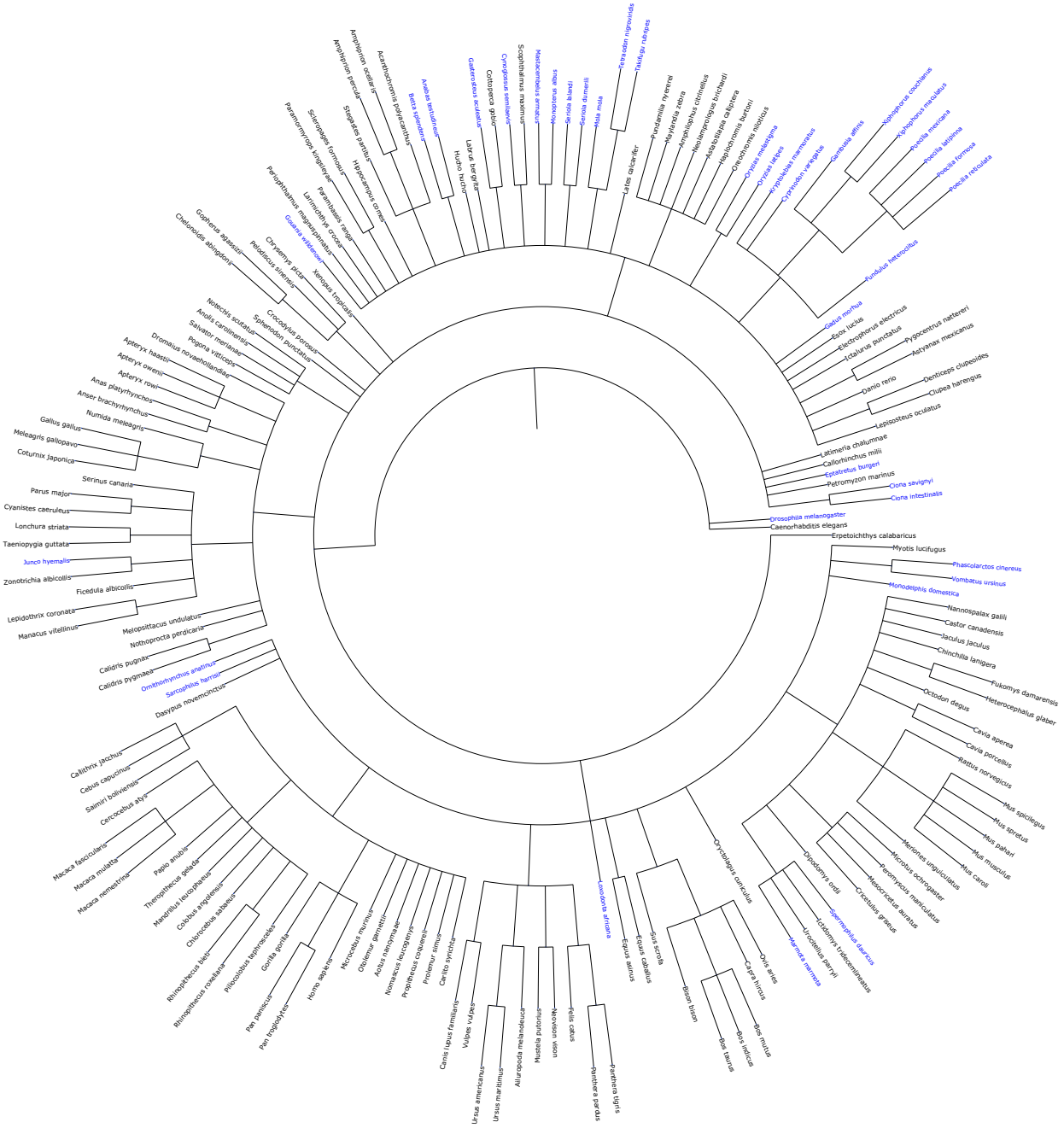

Figure 34: The gene tree for *Sumf2*.

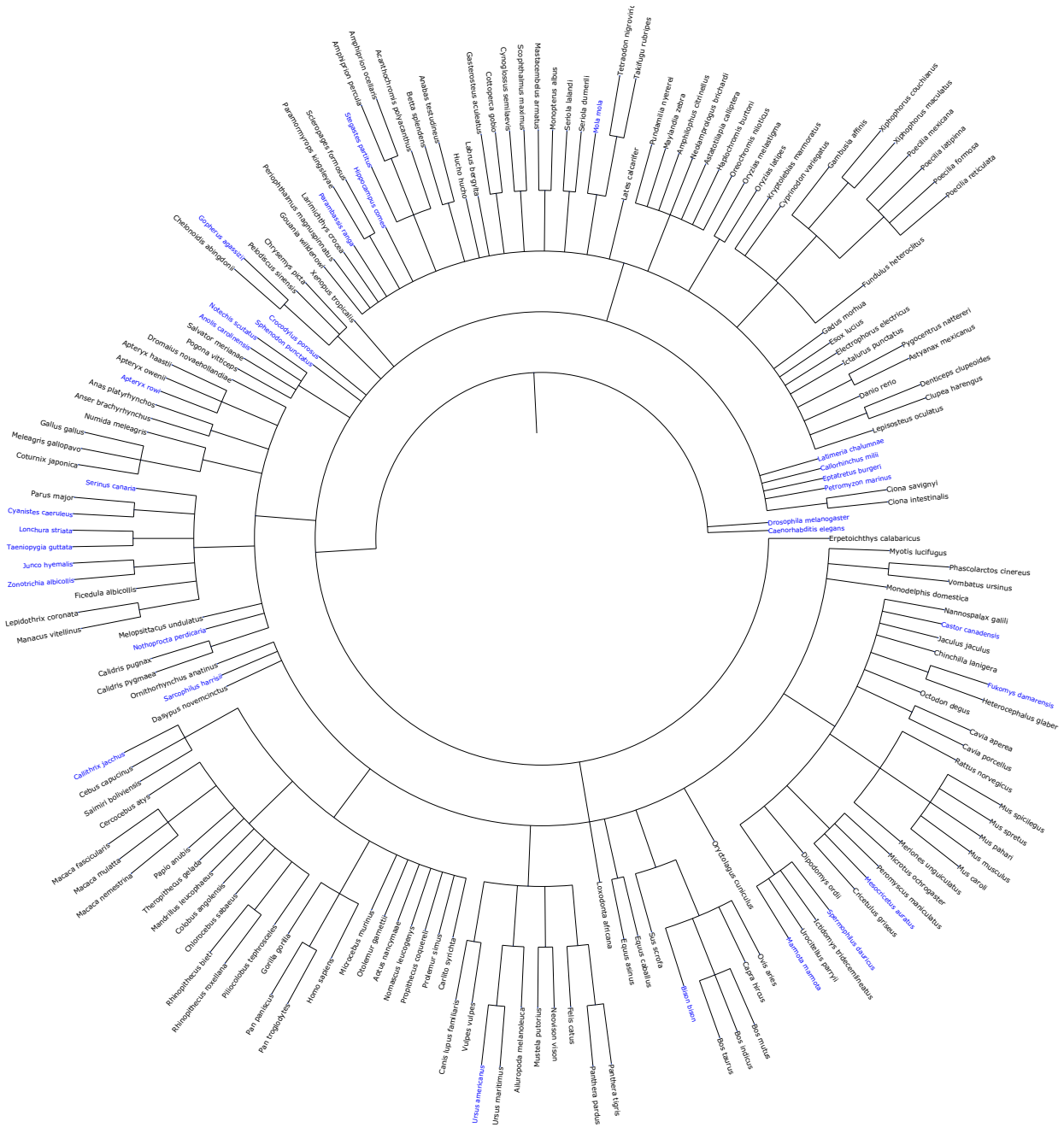

Figure 35: The gene tree for *Ust, Chpf2, Ext2, Arsa, Sulf1, Arsk* and *Xylt1*.

C dN/dS

This section includes supplementary material generated for visualizing the dN/dS ratios for all the significant organisms chosen to analyze in this study.

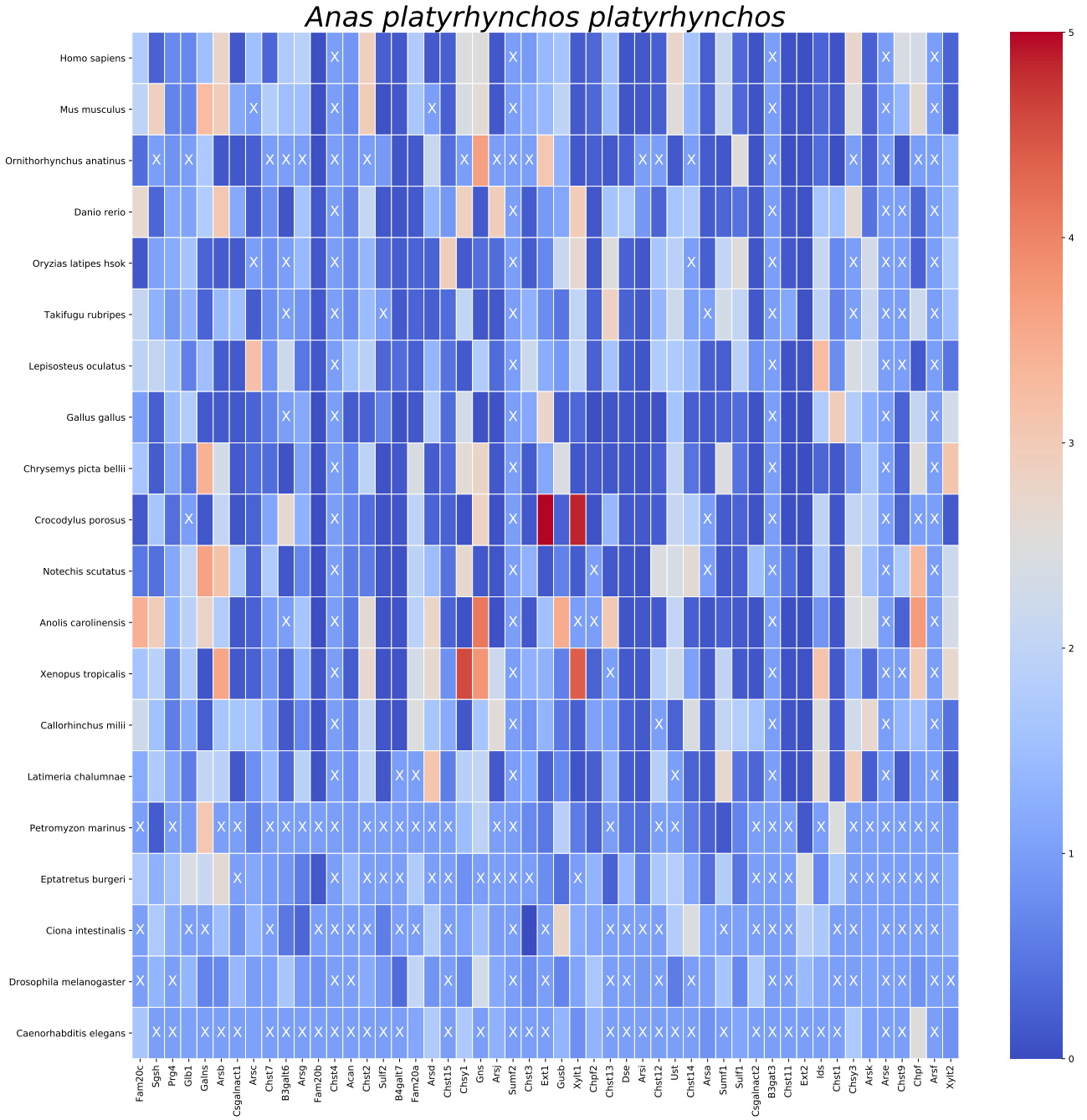

Figure 36: dN/dS 2D grid for *Anas platyrhynchos* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

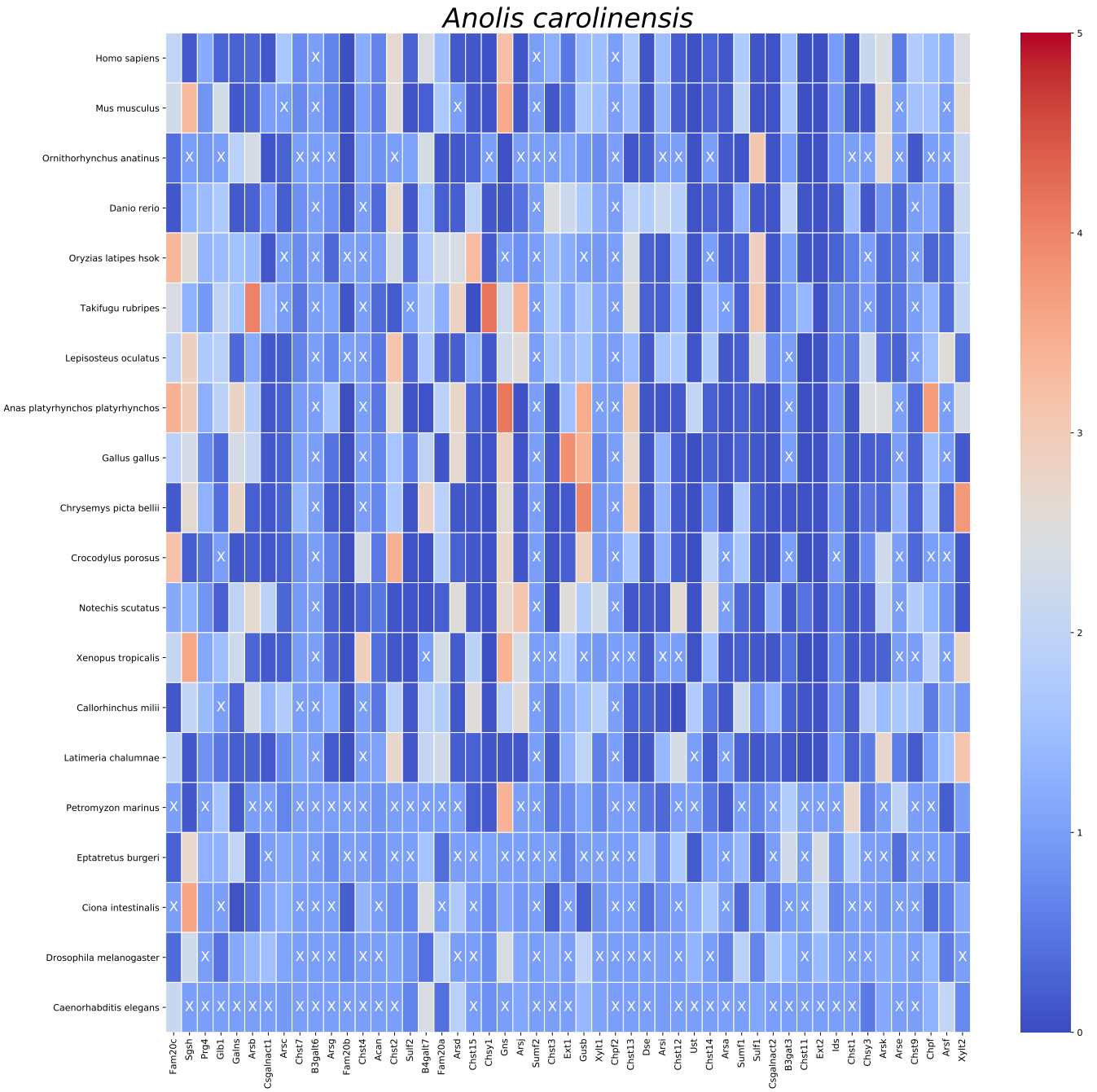

Figure 37: dN/dS 2D grid for *Anolis Carolinensis* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

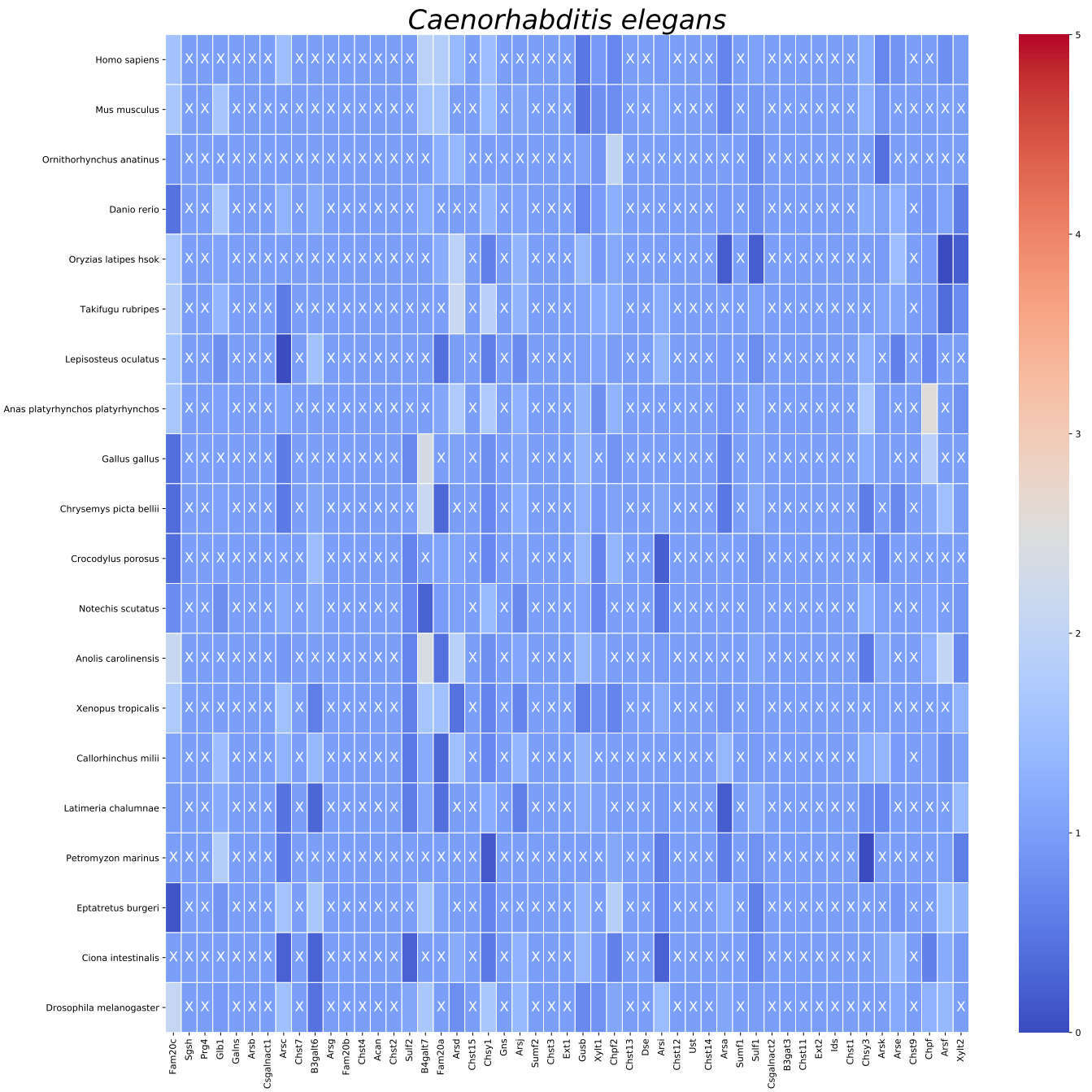

Figure 38: dN/dS 2D grid for *Caenorhabditis elegans* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection.

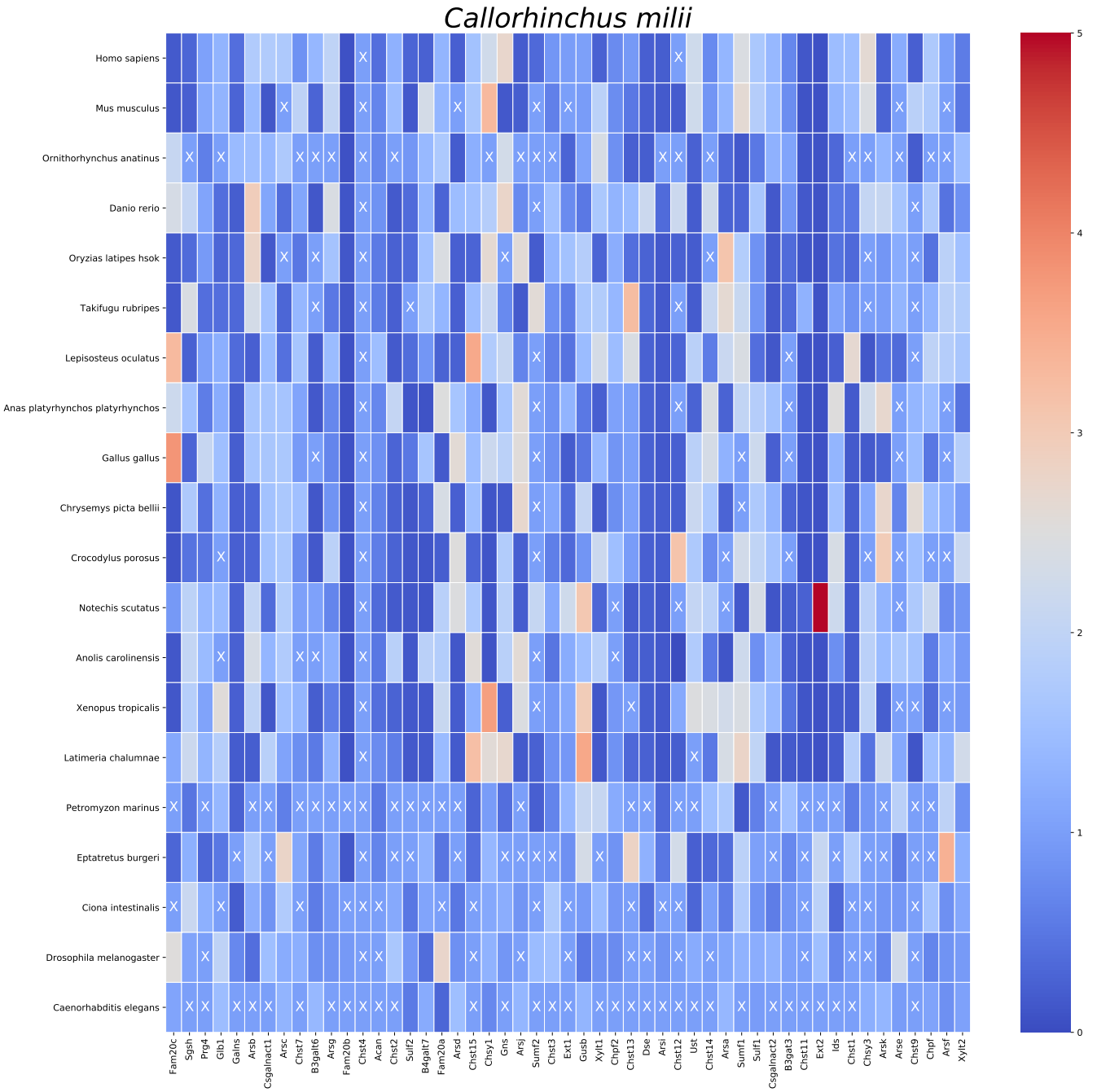

Figure 39: dN/dS 2D grid for *Callorhinchus millii* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

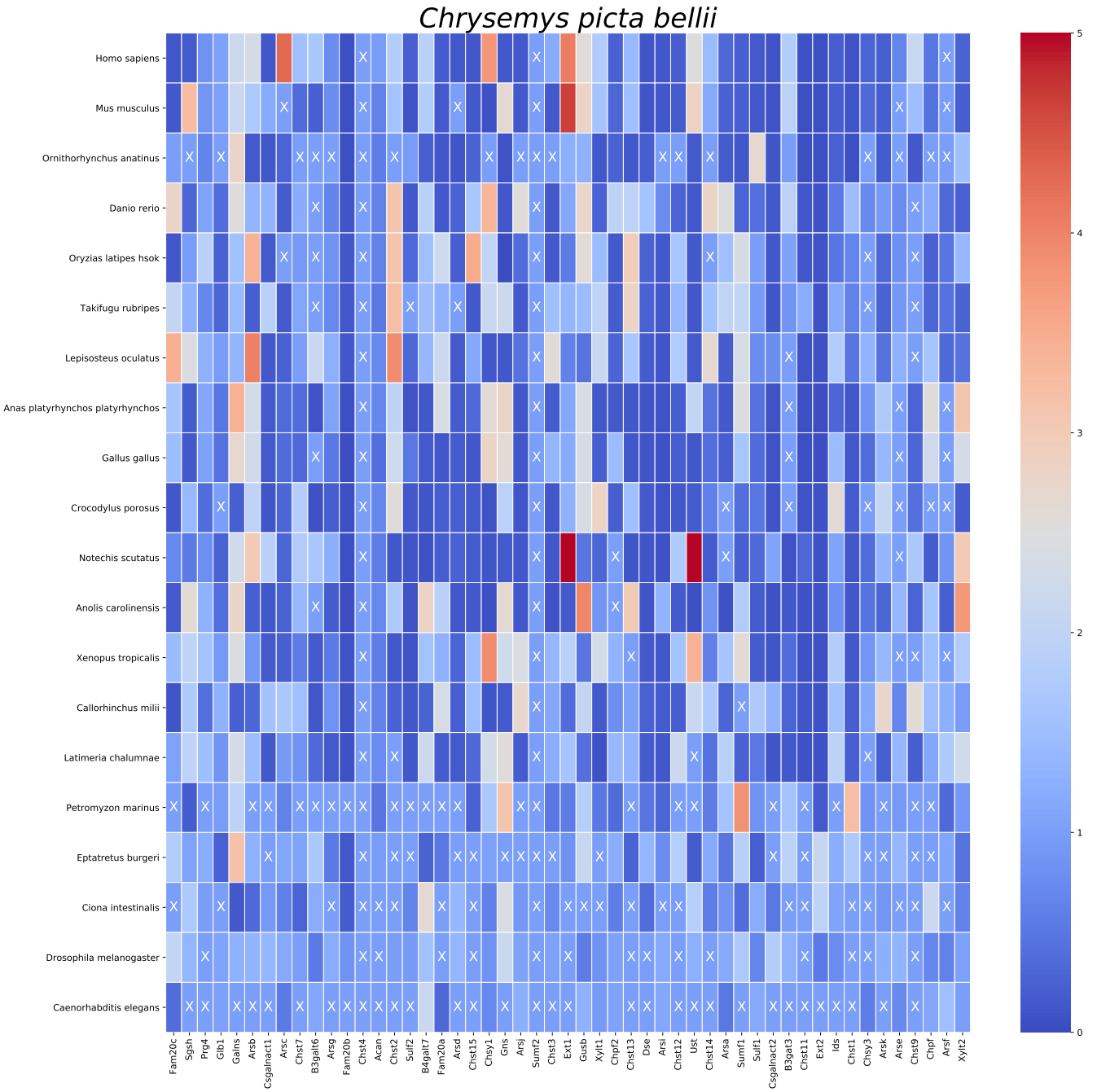

Figure 40: dN/dS 2D grid for *Chrysemys picta bellii* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

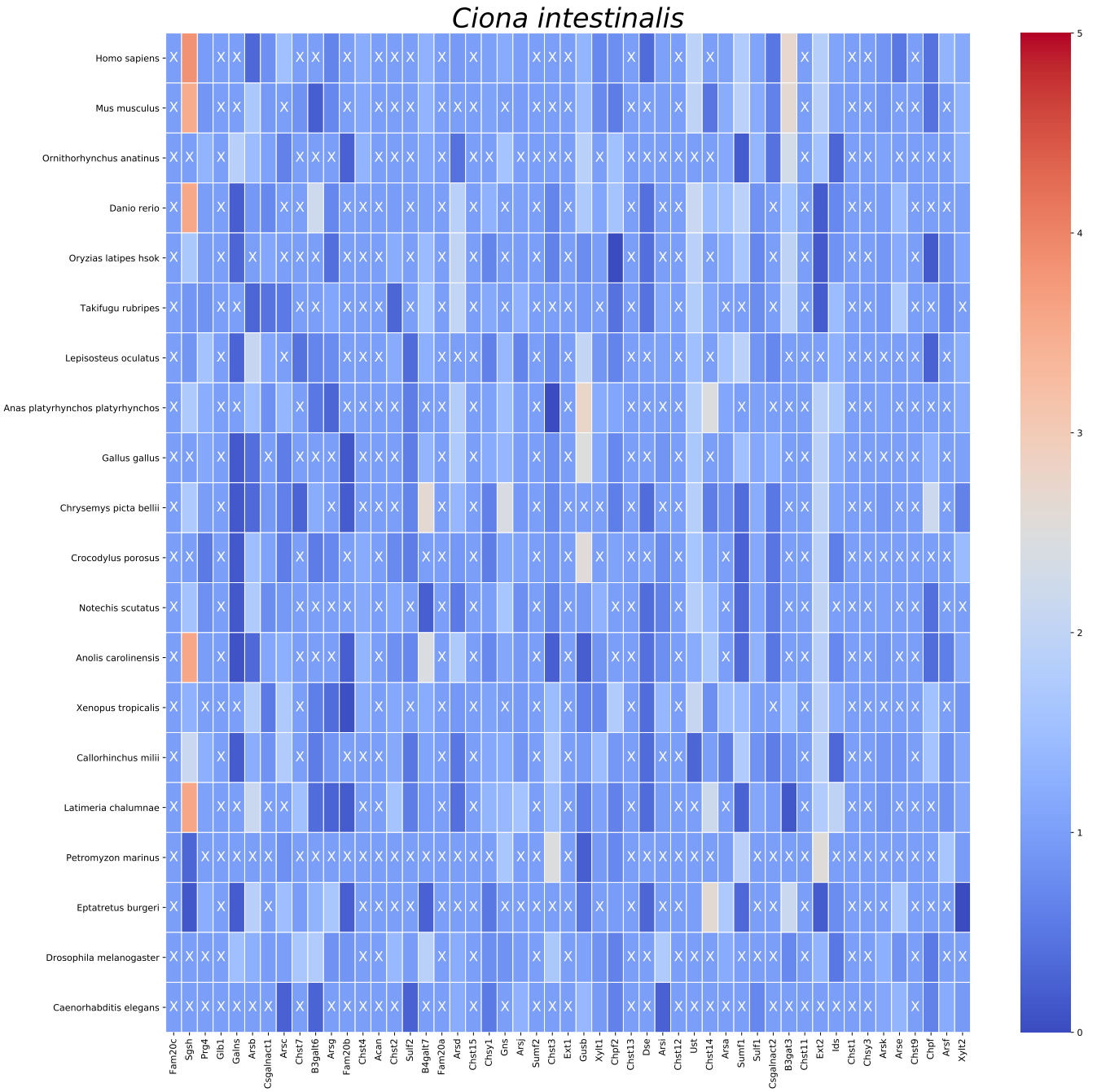

Figure 41: dN/dS 2D grid for *Ciona intestinalis* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

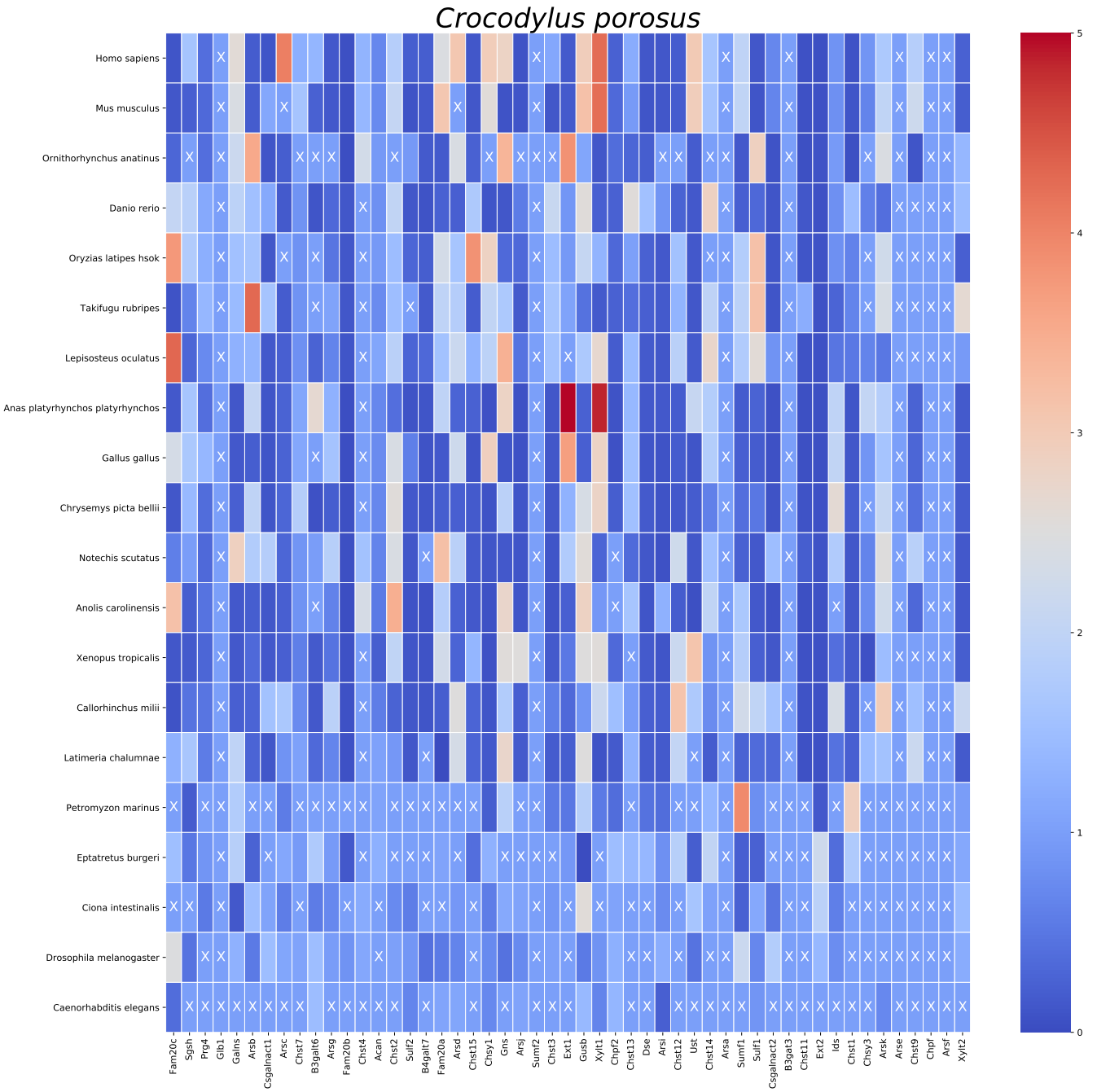

Figure 42: dN/dS 2D grid for *Crocodylus porosus* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

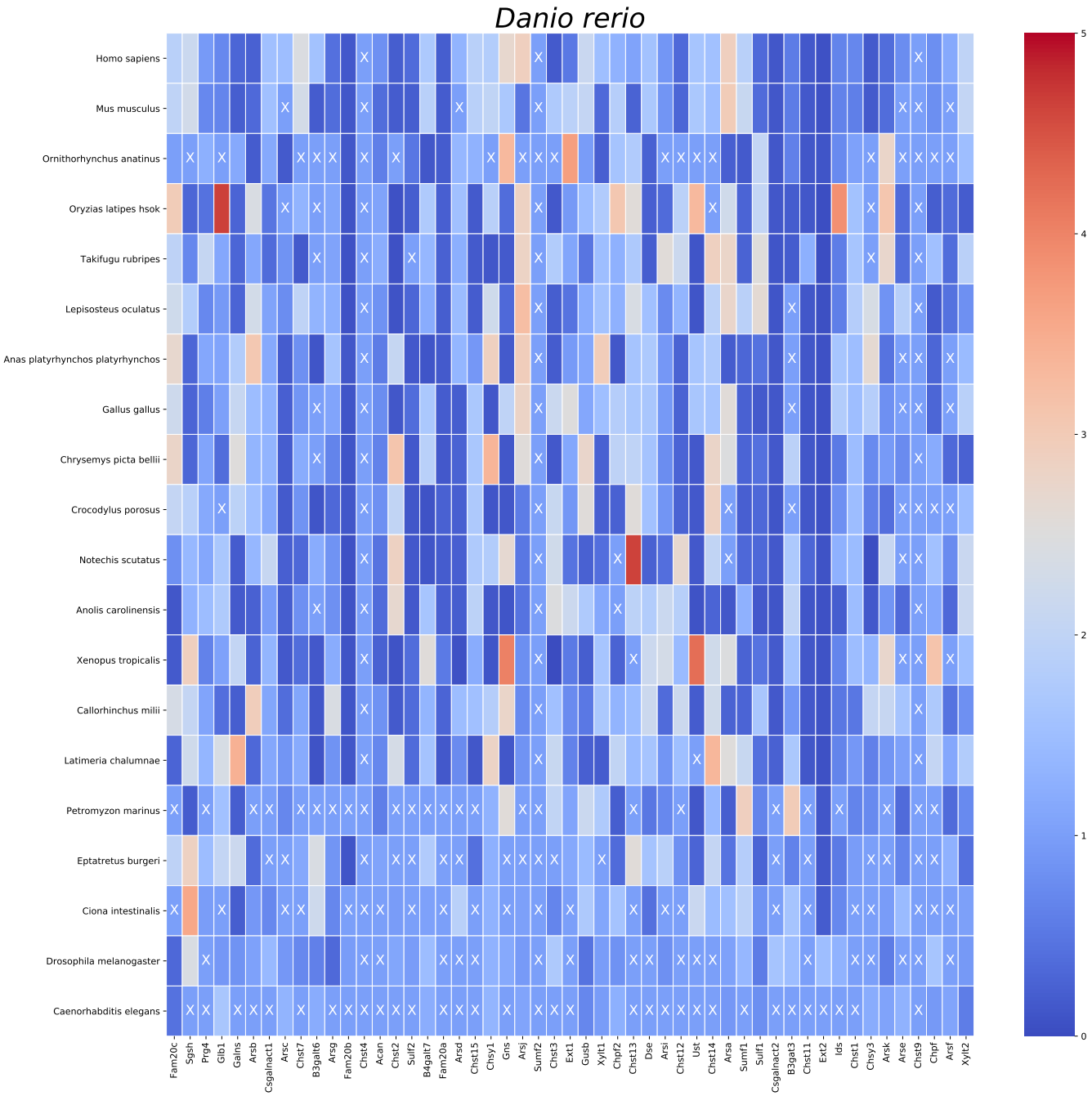

Figure 43: dN/dS 2D grid for *Danio rerio* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

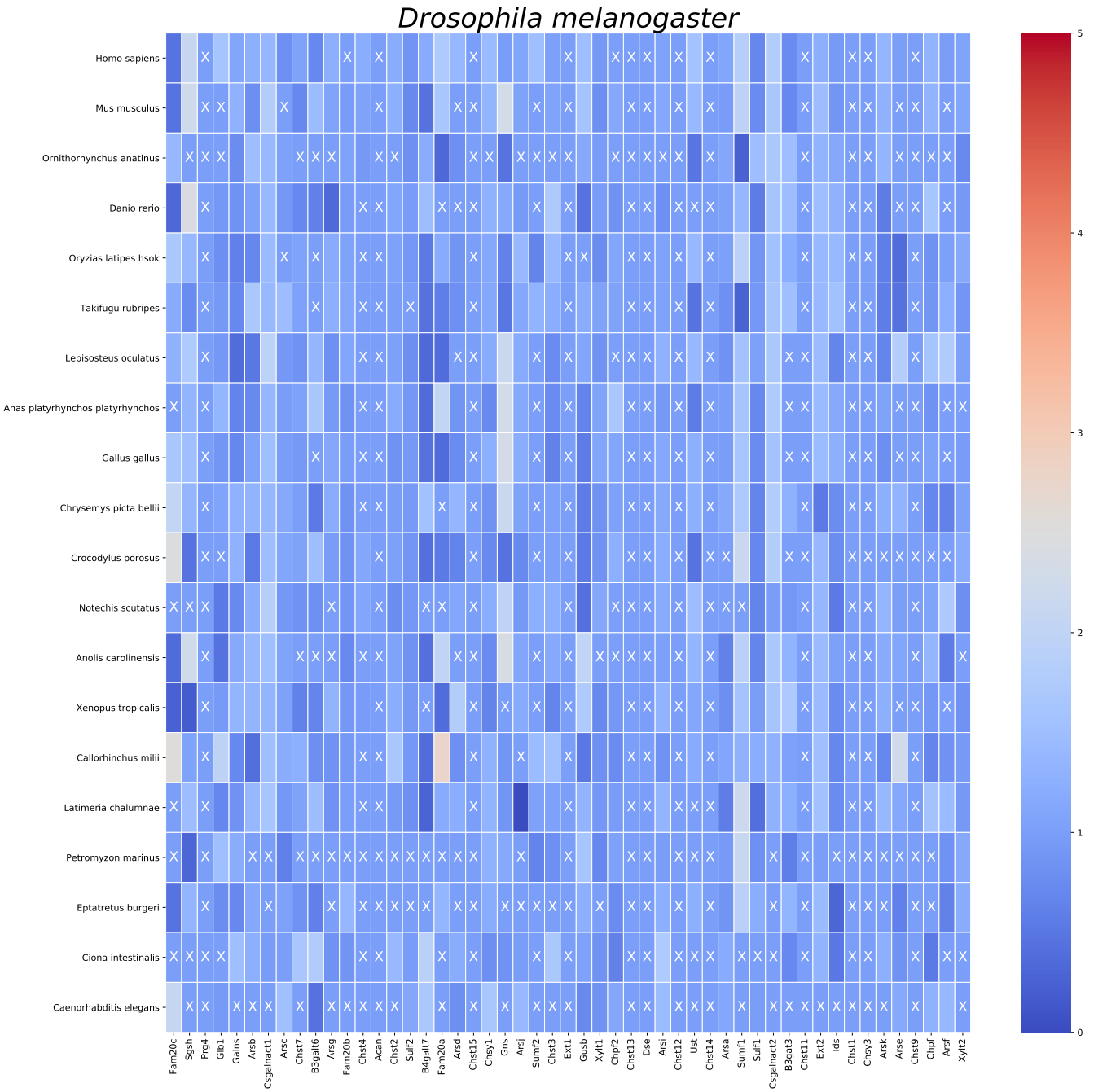

Figure 44: dN/dS 2D grid for *Drosophila melanogaster* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

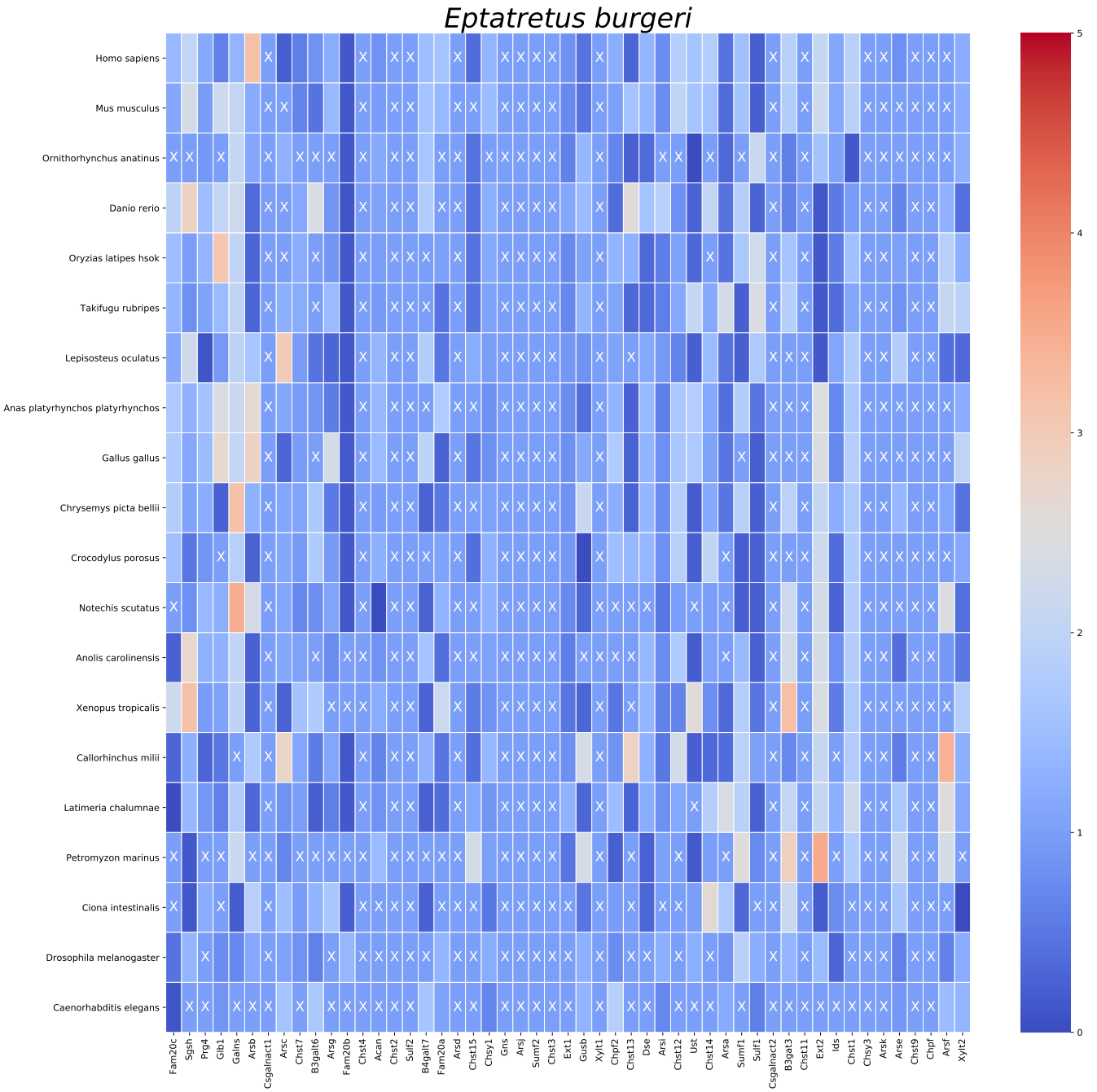

Figure 45: dN/dS 2D grid for *Eptatretus burgeri* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

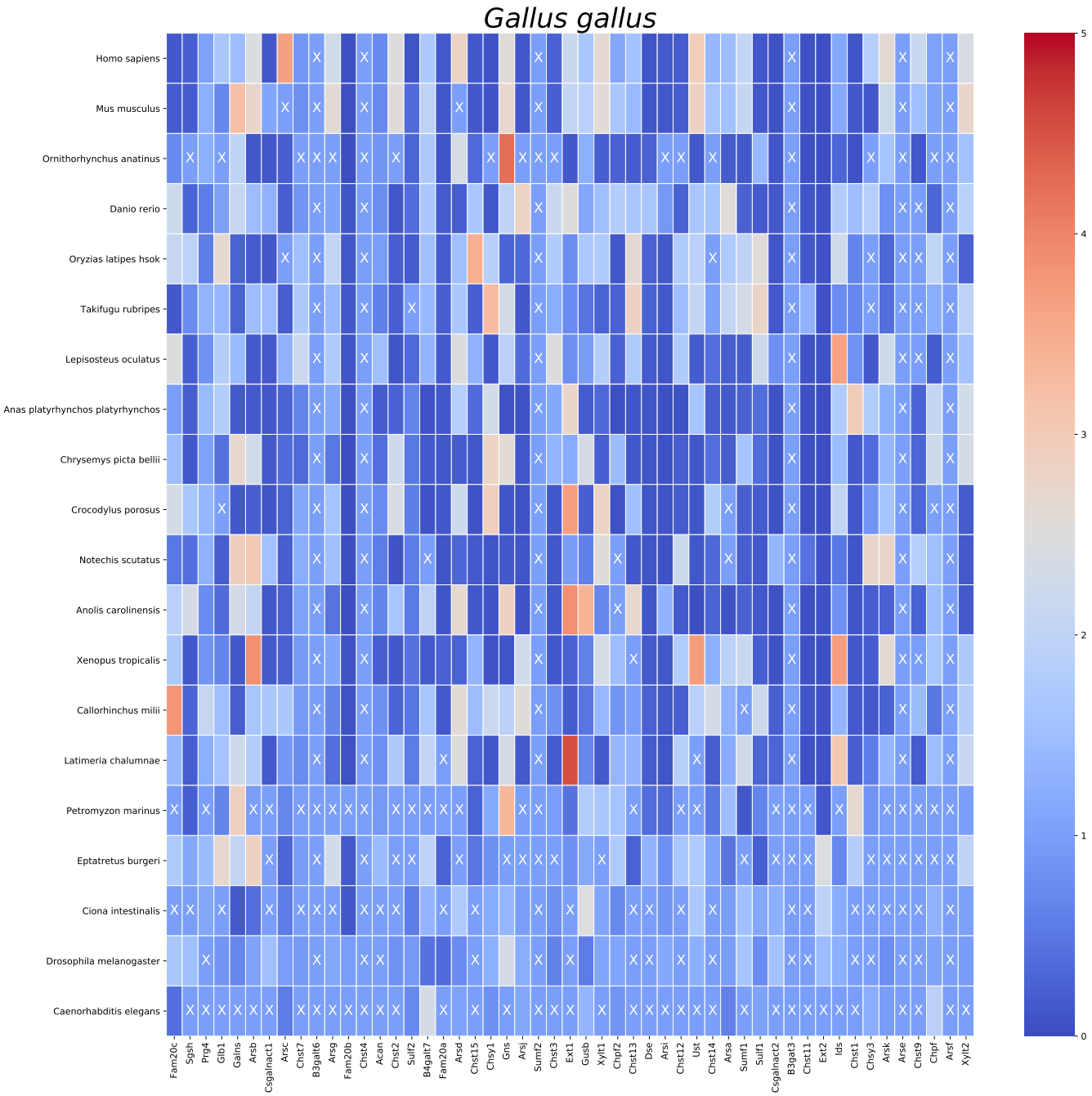

Figure 46: dN/dS 2D grid for *Gallus gallus* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

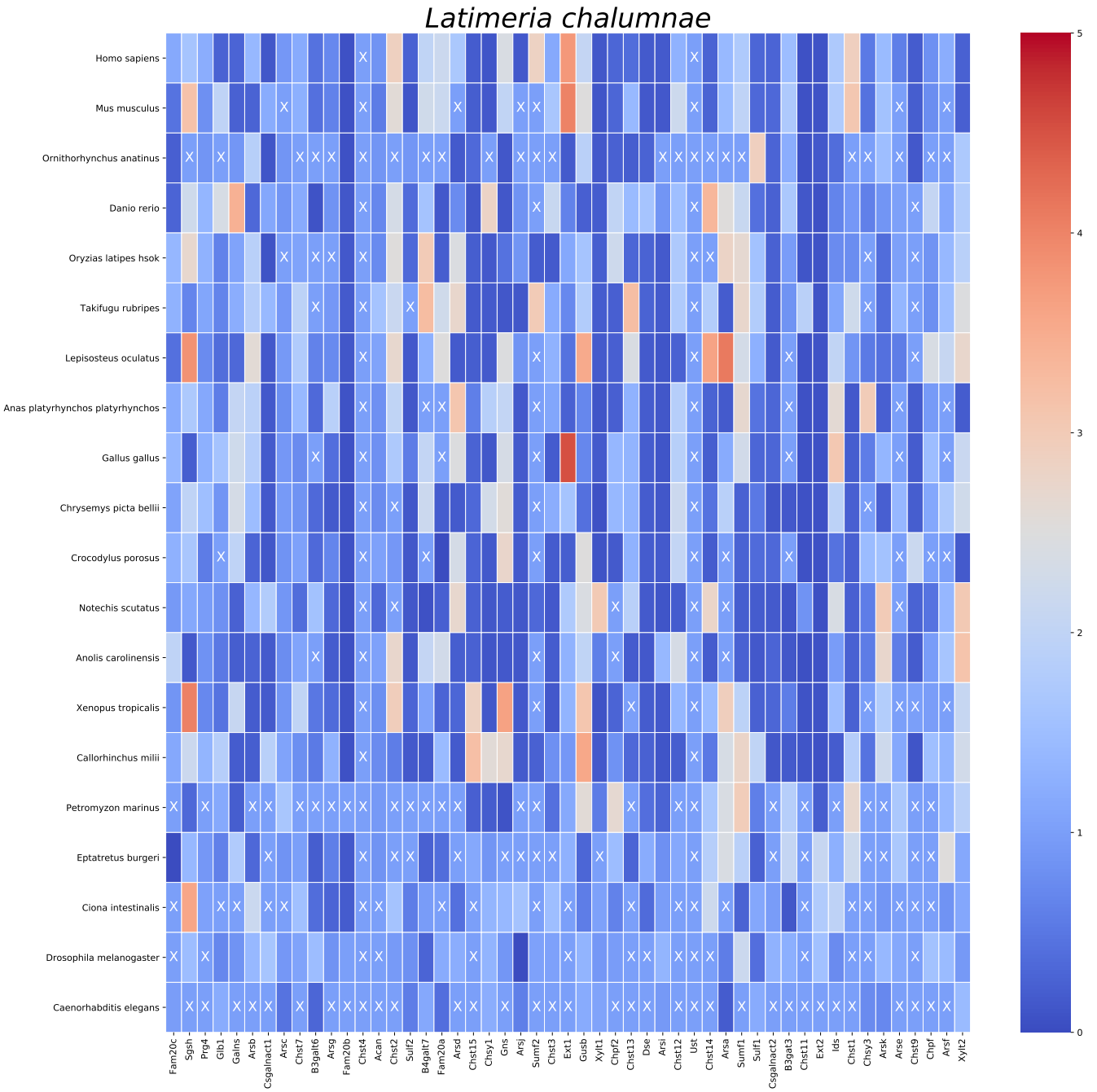

Figure 47: dN/dS 2D grid for *Latimeria chalumnae* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

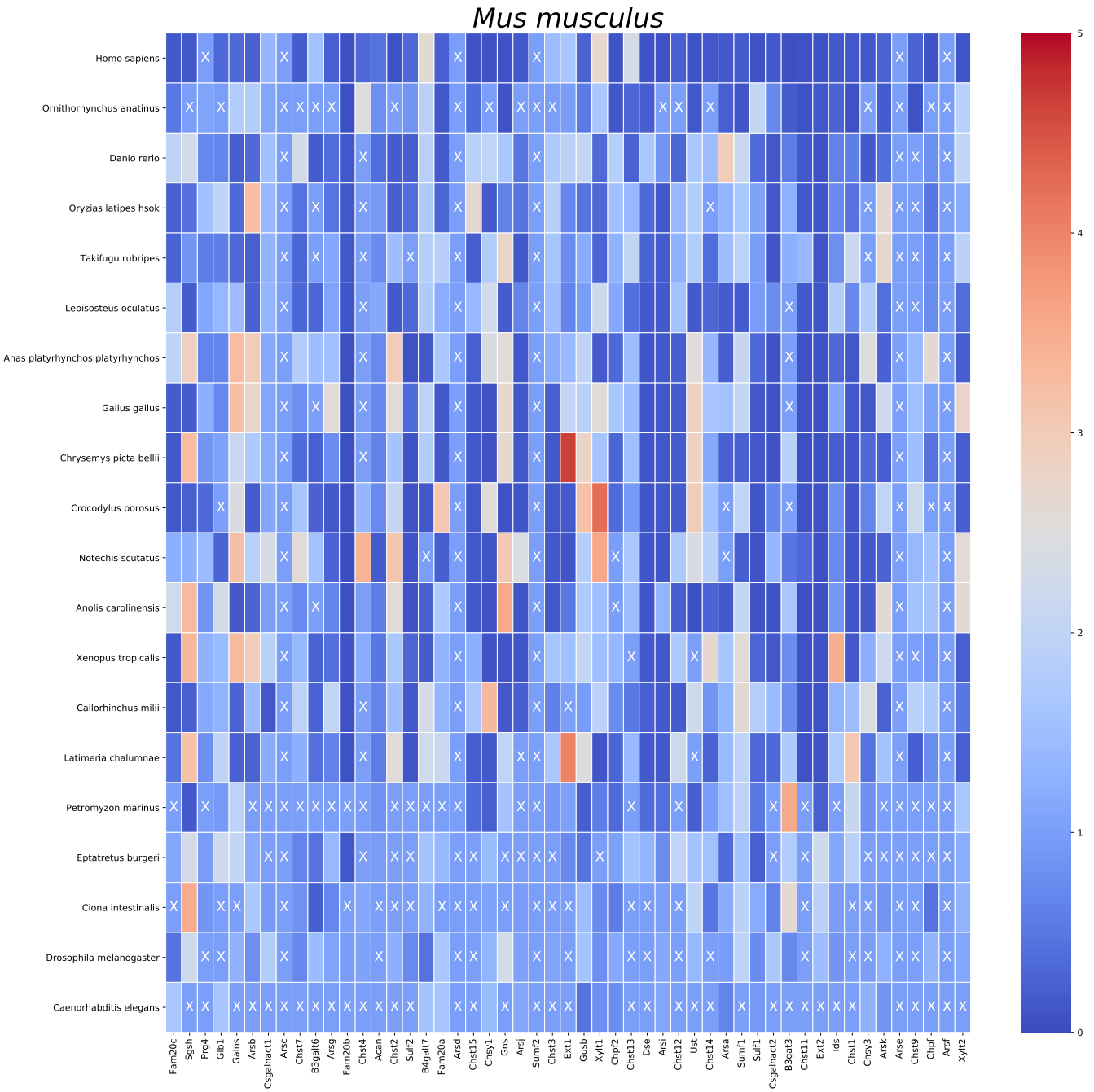

Figure 48: dN/dS 2D grid for *Mus musculus* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

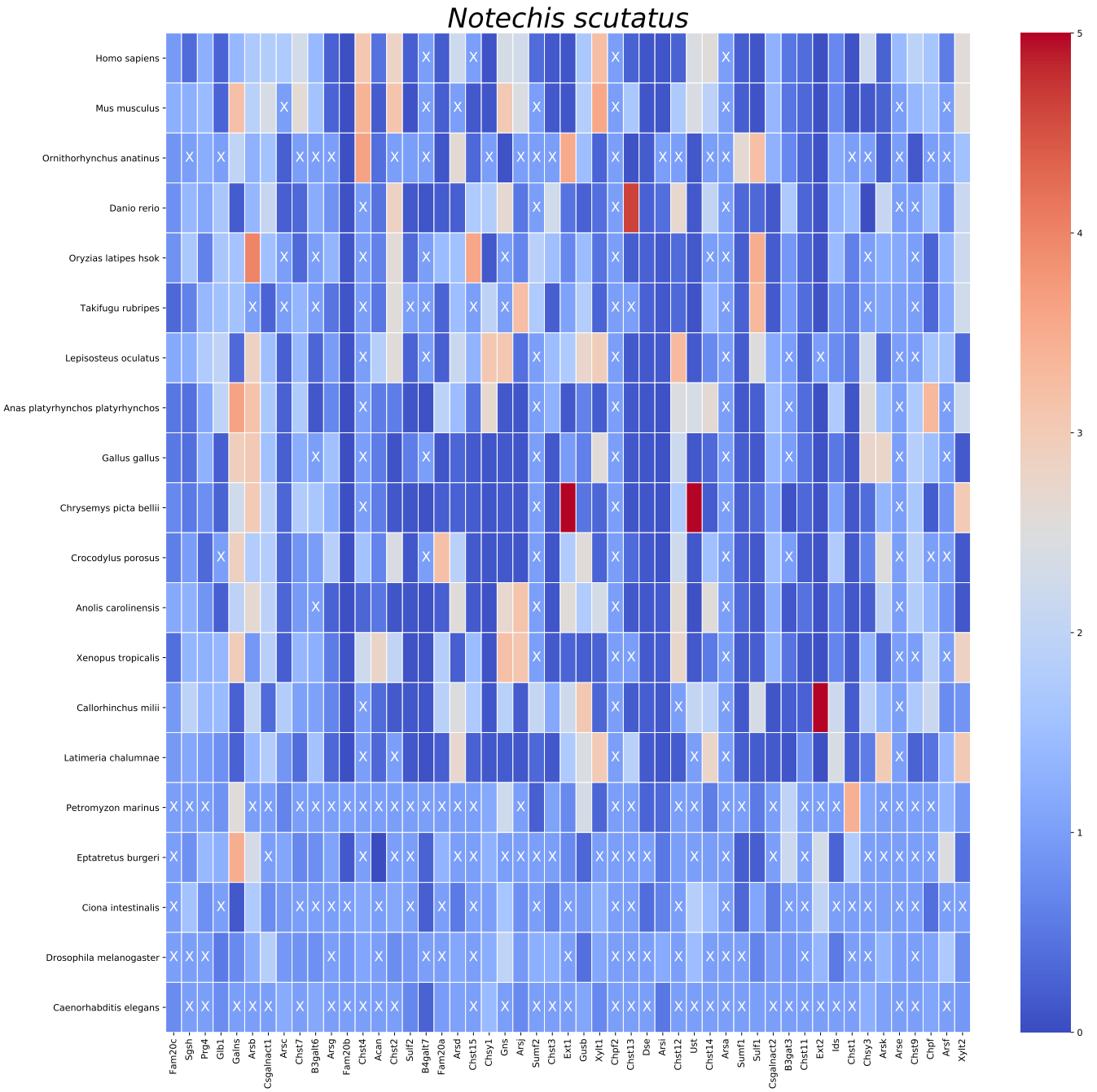

Figure 49: dN/dS 2D grid for *Notetis scutatus* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

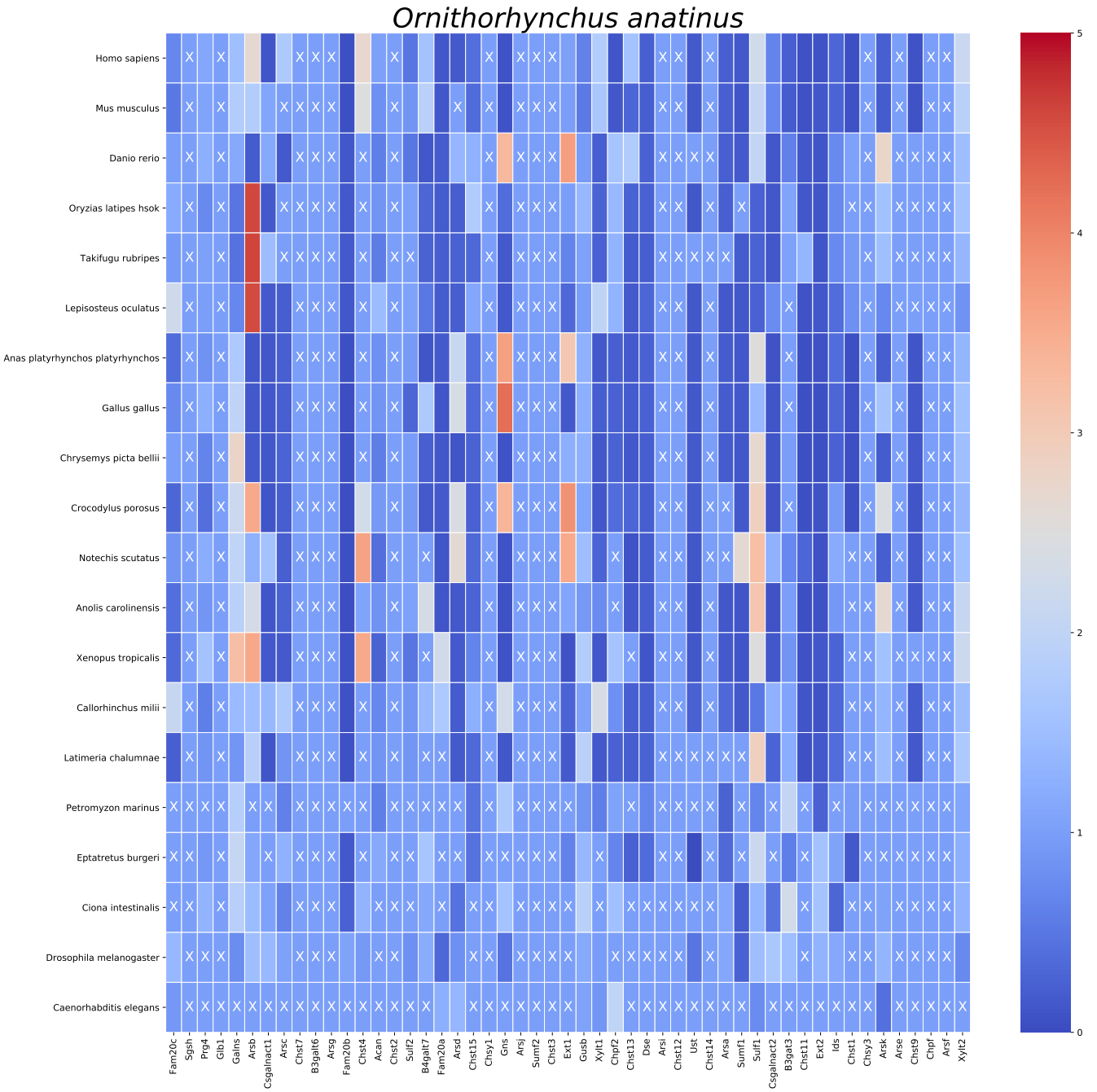

Figure 50: dN/dS 2D grid for *Ornithorhynchus anatinus* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

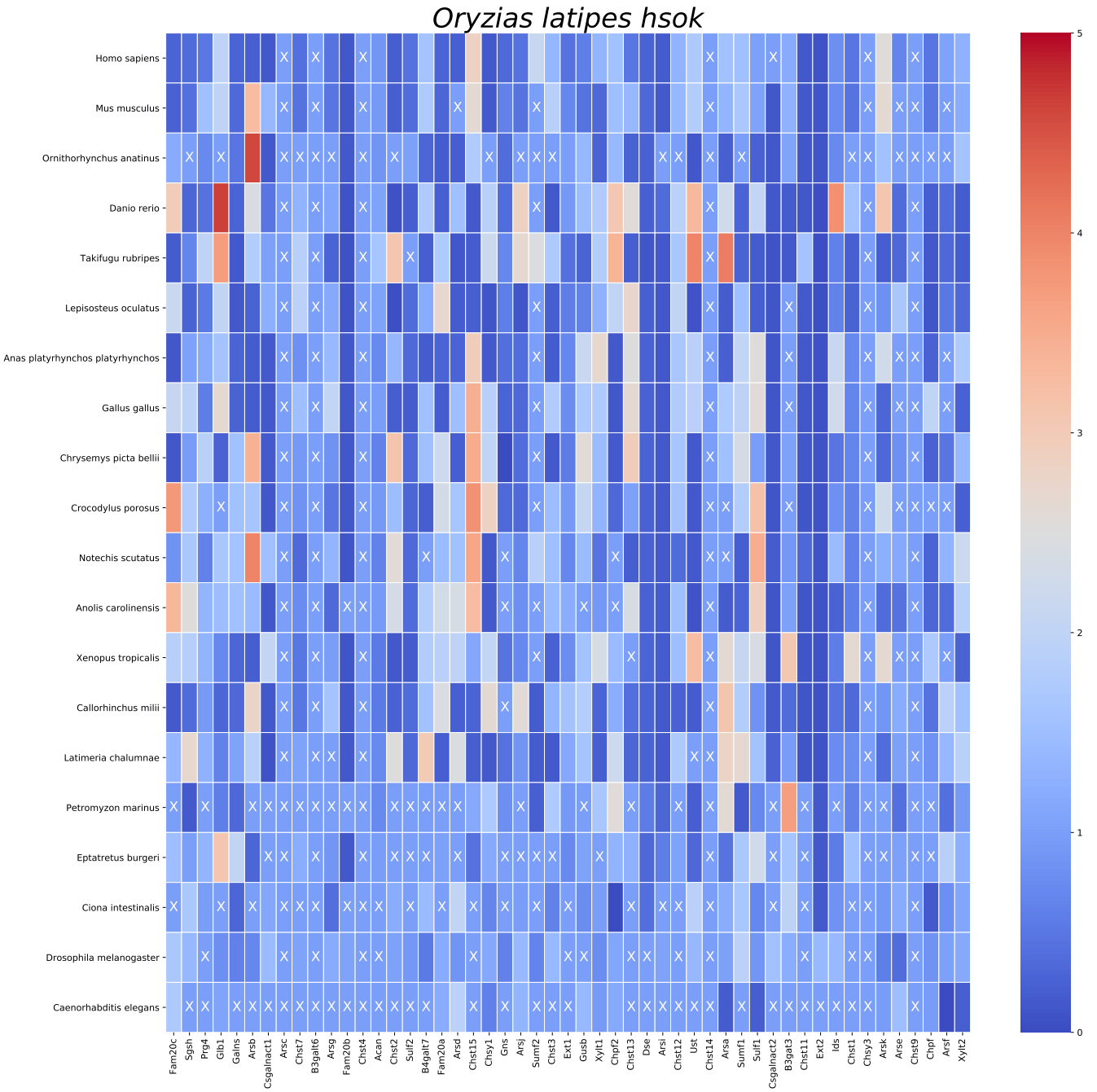

Figure 51: dN/dS 2D grid for *Oryzias latipes* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

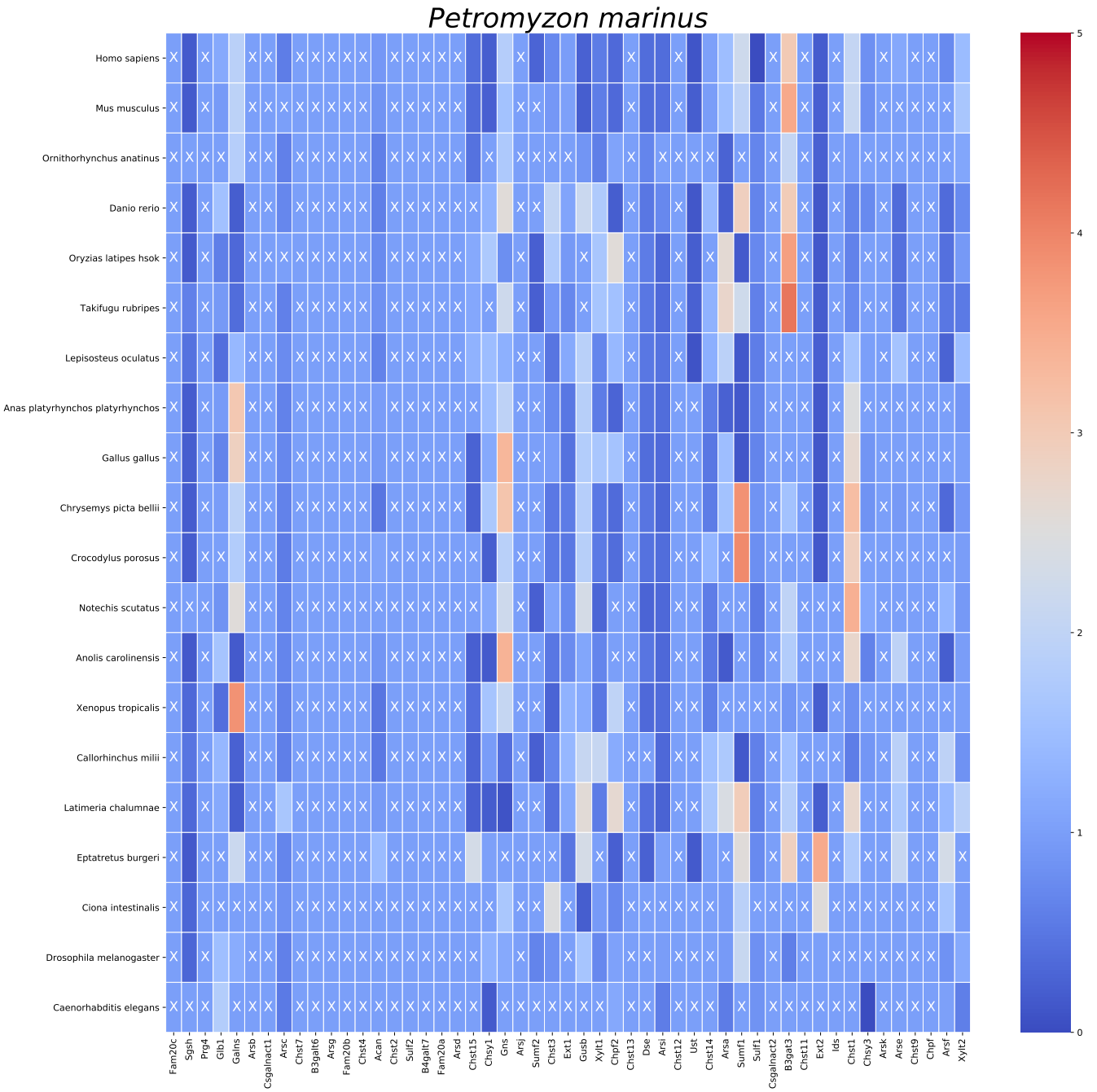

Figure 52: dN/dS 2D grid for *Petromyzon marinus* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

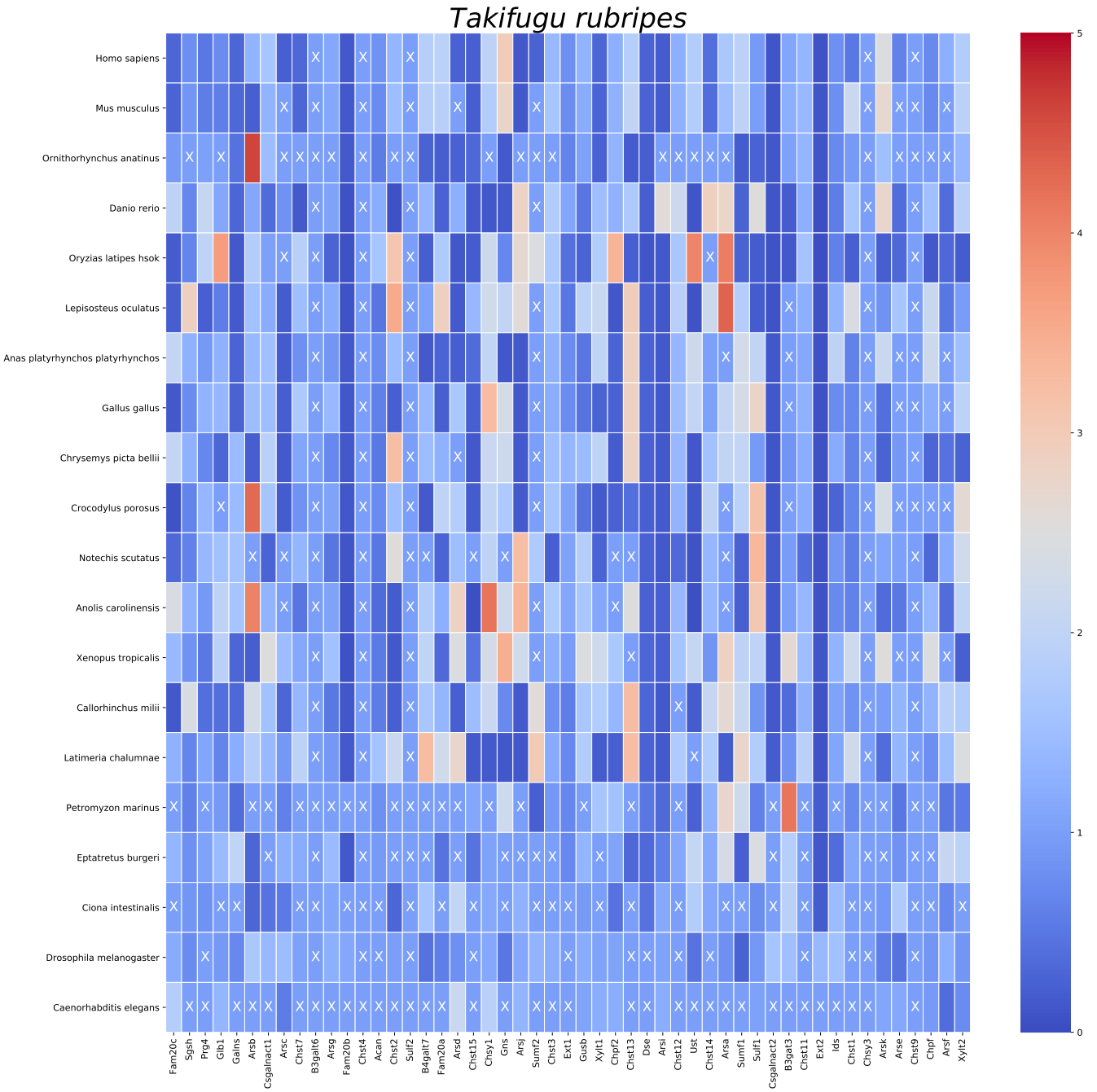

Figure 53: dN/dS 2D grid for *Takifugu rubripes* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.

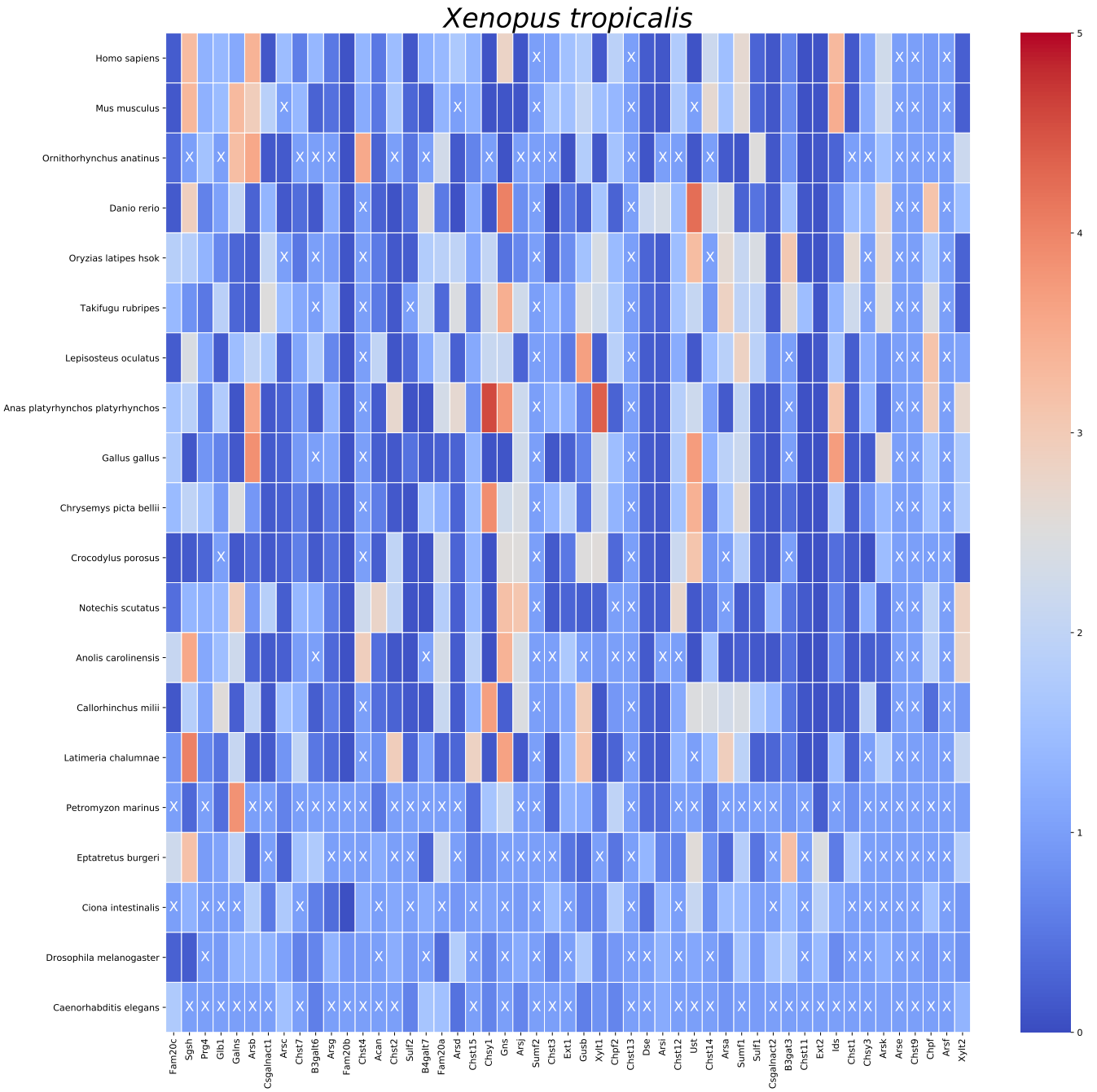

Figure 54: dN/dS 2D grid for *Xenopus tropicalis* computed against the other significant organisms. The intensity values represent the dN/dS ratio - values lighter in colour are less than 1 and purifying selection while values darker in colour are greater than 1 and represent positive selection. Grid boxes that are marked with "X" indicate that either the organism the grid represents, or the organism that dN/dS is computed against, does not have the gene associated with the box.