Protein repeats show clade-specific volatility in Aves

Sandhya Sharma[1], Lokdeep Teekas[1], Nagarjun Vijay[1]

[1]Computational Evolutionary Genomics Lab, Department of Biological Sciences, IISER Bhopal, Bhauri, Madhya Pradesh, India

*Correspondence: nagarjun@iiserb.ac.in

Sandhya Sharma: 0000-0002-0987-2521

Lokdeep Teekas: 0000-0003-2761-3977

Nagarjun Vijay: 0000-0002-8703-6991

Supplementary text and Supplementary figures

Dataset preparation

From Ensembl Biomart release 105, we downloaded a list of all human protein-coding genes and excluded all mitochondrial, read-through, and LINC genes [1]. Using NCBI datasets, we retrieved the remaining Avian clade protein-coding genes [2]. The downloaded gene sequences are subsetted into 29 avian orders (**Table S1** and **S2**: 56 species in total). The species tree and order classification of the Aves clade is downloaded from BirdTree [3] with 1000 bootstraps. The phylogeny presented in birdtree.org is inferred using a combination of molecular and morphological data. Specifically, the tree is based on a large-scale analysis of DNA sequences from different bird species and a comprehensive review of morphological features, such as skeletal structures, plumage, and behavior. Overall, the phylogeny presented in birdtree.org is based on a rigorous and comprehensive analysis of both molecular and morphological data and represents the most widely accepted evolutionary relationships among bird species.

Based on the following method, we selected the amino acid sequences for each gene that was the most similar: (i) We selected and aligned all of the “NP” designated sequences for a specific gene. The consensus sequence is created using the “NP” aligned sequences, (ii) The sequence is regarded as the final sequence of that species if it is the only “NP” sequence for that gene, (iii) If a species has more than one “NP” sequence for a gene, each sequence is aligned with the consensus sequence, and the sequence with the best per-base alignment score is preferred, (iv) If a species lacks an “NP” sequence for that gene, all of its “XP” sequences are aligned with the consensus sequence. The candidate with the highest score for per-base alignment is chosen.

Filtering and alignment

All of the chosen amino acid sequences are mapped to the corresponding CDS sequence. We remove all the sequences with incomplete ORF (absent START codon, absent STOP codon, contains non-nucleotide characters, contains in-frame STOP codon, or sequence not a multiple of three). We discard a multifasta file for a gene if it only contains three or fewer species. Codon alignment is performed on the remaining files using the GUIDANCE program [4] with the MUSCLE aligner [5] using 100 bootstraps.

Repeats identification and related analyses

We use fLPS2.0 [6] to find low-probability segments in the amino acid sequences. Repeats are characterized as stretches that are longer than four amino acids, have more than 70% composition, and contain fewer than five unique amino acids. A customized script maps the repeat stretch's coordinates onto the corresponding aligned CDS sequence. The overlapped repeat regions are classified as orthologous repeats.

Each order can have a taxon-specific abundance of particular repeat types. We determine the proportion of the ten most frequent repeats in each order. Using the proportion of these repeats, we create a stacked barplot in R by coloring each repeat distinctly (**Fig. 1**). The remaining repeats' proportions are added to make the proportion 1 in each clade.

The number of distinct repeat types and their relative abundance are the two criteria that Simpson's diversity index (D) requires. We calculated the number and relative abundance of different amino acid repeats in R. The Simpson's diversity index (D) is calculated as $\mathbf{1 - (\sum(n*(n-1))/(N*(N-1)))}$. Here, n is the count of each amino acid repeat type, and N is the summation of all the counts of all the amino acid repeats. The D ranges $0 \leq D \geq 1$, with 0 implying no diversity and 1 representing maximum diversity. For each order, we utilized this to calculate the D and then used the results for the barplot (**Fig. S1** and **Table S3**). The barplot shows the distribution of D of all the orders.

Clade-specific quantification of repeat-containing genes

We subsetted the order-wise list of protein-coding genes and repeat-containing genes. For the input of the UpSet plot using the UpSetR package [7] in R, we created a presence-absence matrix in an order-wise manner for the repeat-containing genes. Along with the number of unique genes with repeats present in order(s), their order-wise proportion (the genes containing repeats divided by the total number of genes in that order) is also calculated and presented in the UpSet plot (**Fig. S2**).

We counted the number of different amino acid repeat types in all the orders and represented the top thirty in the barplot in R (**Fig. S3**). A total of 45 repeat-containing genes are common across all the orders (**Table S4**).

GO enrichment analyses

Using ShinyGO 0.76.1 [8], the Gene Ontology (GO) enrichment analysis is carried out on the list of genes containing repeats with the background set to all protein-coding genes. We performed GO analysis for cellular components, molecular functions, and biological processes (**Fig. S4** and **Table S5**).

Detecting orthologous repeat length contrast between species groups

A custom R script is used to identify orthologous repeats and detect considerable differences in repeat length between species groups (repeatplotter_everything.r). The workflow of the script is as follows:

1. **Input required:**
    - File with clade name, species name, gene name, unaligned and aligned coordinates of repeats, and aligned gene length
    - Species tree
    - Clade name
    - Gene name
    - Proximity window (in nucleotides; max 20)
    - Compiled file with site-specific positive selection details
2. **Main workflow:**
    i) *Subsetting the dataset:*
        (1) We subset the repeat coordinate file according to clade and gene name. This gives us species-wise coordinates of the repeats in that specific gene in the clade.
        (2) We also subset the compiled file with site-specific positive selection details according to gene and clade. This will provide the details of positively-selected sites in the gene in the clade.
        (3) We list unique species mentioned in the subsetted repeat coordinate file.
        (4) We prune the species tree according to the species mentioned in the above-created list.

    ii) *Identification of orthologous repeats:*
    **Important note: The length of the gene can vary between species, potentially changing the orthologous repeat coordinates between species in unaligned coordinates. But in the aligned file, the orthologous repeats will be aligned. We

primarily use the aligned coordinates to identify orthologous repeats. Moreover, the length of the orthologous repeats can also vary between species.

(1) We uniquely select the aligned end coordinates of the repeats in the file.
    (a) If all repeats have the same aligned end coordinate, there is only one orthologous repeat.
    (b) If there is more than one unique aligned end coordinate of the repeat in the aligned file:
        (i) Either there exists more than one repeat in the gene for species, or the end coordinates of the orthologous repeat do not align perfectly.
        (ii) We will make a histogram of the distribution of aligned end coordinates by keeping the bin size of "proximity value."
        (iii)The end coordinates of the repeats within the same bin are considered orthologous.

*The list of detected orthologous repeats is provided in **Table S6**.

iii) Identifying the repeat length contrast between species groups:

(1) We calculate the repeat length at the internodes using phylogenetically independent contrast (PIC) for each set of orthologous repeats.
(2) We select the internodes with repeat lengths lesser than the 1st percentile and greater than the 99th percentile from the distribution of orthologous repeat length.
(3) We select internodes having two child internodes with at least three species in both child internodes.
(4) We calculate the orthologous mean repeat length for both sets of child internodes.
(5) The set of species with a larger mean repeat length is considered expanded, while the set of species with a smaller mean repeat length is considered contracted.

Species-wise quantification of repeat length contrast

We quantified the proportion of repeats expanded or contracted for each species from the considerable repeat length variation results. The proportion is calculated based on the number of results with significant length variation to the number of results either in expansion or contraction in a species-wise manner. The obtained quantitative result is visualized as a pie chart on the avian phylogeny (**Fig. 2**). In the pie chart, the numbers in brackets indicate the number of times the species has an expansion or contraction of repeats. The sum of the numbers in each bracket will indicate the number of times each species has a considerable difference in repeat length.

Furthermore, we filtered out the results of considerable length variation with less than three species in either of the compared species groups. The genes with considerable repeat length variation between two adjacent species groups are listed in **Table S7**.

Secondary structure prediction

The amino acid sequence of MAN1C1 of *Anas platyrhynchos* and *Taeniopygia guttata*, as a representative species from the contracted and expanded repeat groups, respectively, is provided to PROTEUS2 [9] for secondary structure prediction. We predicted the coordinates of the membrane helix using the PROTEUS 2 web server. The structures are visualized in the Protter (version 1.0) [10] web server by providing the coordinates of the repeat and membrane helix (**Fig. S5**).

Molecular evolutionary analyses

On aligned gene sequences, we compare the site models M7 and M8 of PAML [11] to find positively selected sites. Sites with a posterior probability greater than 0.95 are considered significant positive sites.

Data and script availability

All the data and scripts necessary to generate the results are provided on the GitHub link:

https://github.com/Sandhyash/Aves_repeats

References

1. Howe K.L. et al. 2021. Ensembl 2021. Nucleic Acids Res. **49**, D884–91.
2. Sayers E.W. et al. 2020. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res.
3. Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature.
4. Sela I., Ashkenazy H., Katoh K., Pupko T. 2015. GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. **43**, W7–14.
5. Edgar R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**, 1792–7.
6. Harrison P.M. 2021. fLPS 2.0: rapid annotation of compositionally-biased regions in biological sequences. PeerJ. **9**, e12363–e12363.
7. Conway J.R., Lex A., Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. **33**, 2938–40.
8. Ge S.X., Jung D., Jung D., Yao R. 2020. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. **36**, 2628–9.
9. Montgomerie S., Cruz J.A., Shrivastava S., Arndt D., Berjanskii M., Wishart D.S. 2008. PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res. **36**, W202–9.

10. Omasits U., Ahrens C.H., Müller S., Wollscheid B. 2014. Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics. **30**, 884–6.

11. Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. **24**, 1586–91.

12. Kosakovsky Pond S.L., Frost S.D.W., Muse S. V. 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics. **21**, 676–9.

Supplementary Figures

Figure S1

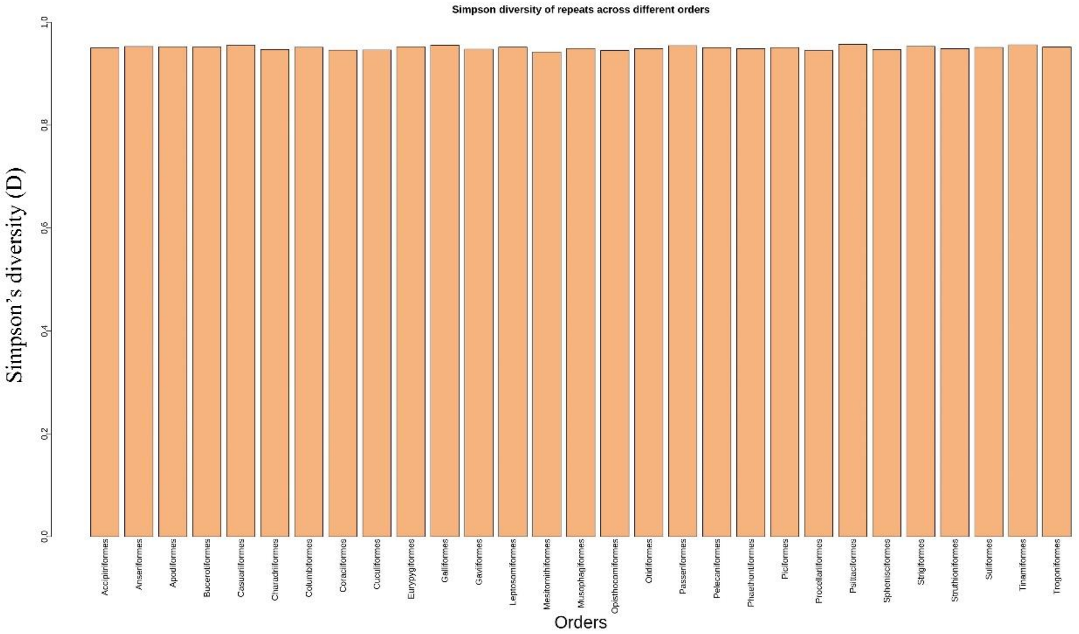

Figure S1: Simpson's diversity (D) of amino acid repeats in different Avian orders. The Simpson's diversity varies between 0 to 1, where 1 represents maximum diversity, and 0 represents no diversity. The D values for the above-mentioned avian orders range between 0.94 to 0.96, implying a high diversity of repeats.

Figure S2

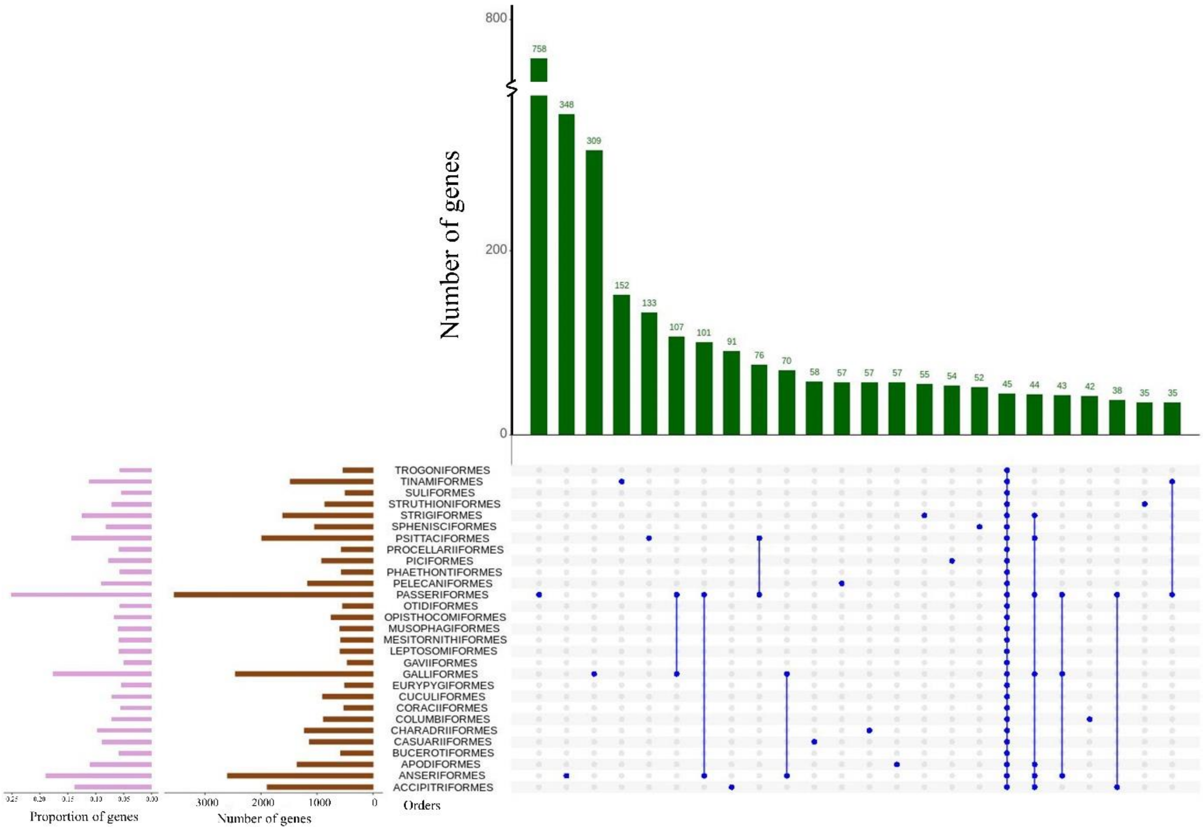

Figure S2: The UpSet plot of repeat-containing genes and their proportion in each order. The green colored bars represent the number of genes with repeats with numbers provided at the top of the bars, and the blue colored dots indicate the intersection between orders. The brown bars (bottom left side) represent the total number of genes with repeats in each order. The pink bars represent the proportion of repeat-containing genes (compared to the total genes in that order) for each order.

Figure S3

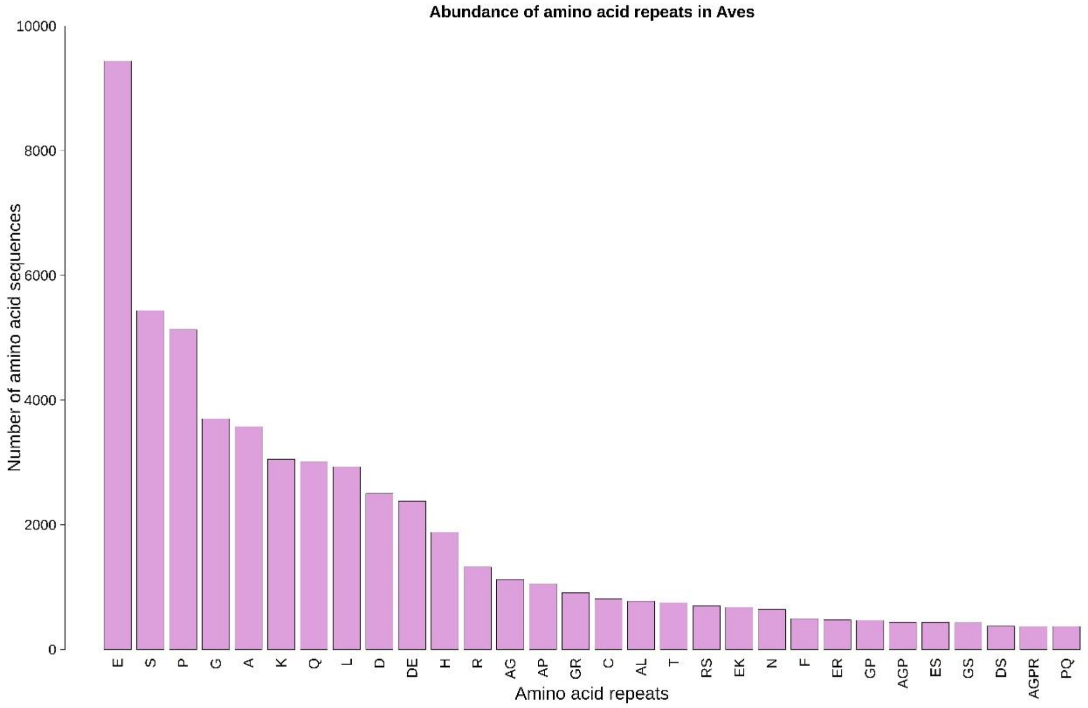

Figure S3: Abundance of top 30 amino acid repeat types in the Avian clade. The glutamic acid (E) repeat occurs most abundantly. The abundance of occurrence of serine (S) and proline (P) amino acid repeats ranges from 4000 to 6000, after which there is a gradual decrease in other amino acid repeat types.

Figure S4

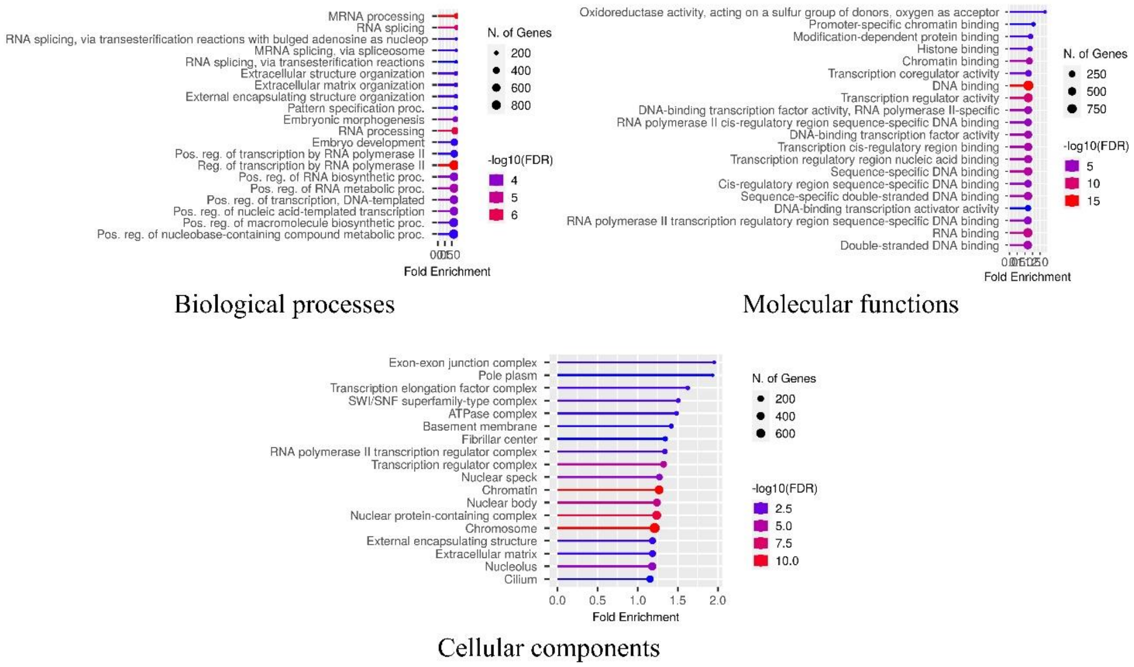

Figure S4: GO enrichments of genes with repeats in Biological processes, Molecular functions, and Cellular components. The GO enrichment is performed on the ShinyGO web server.

Figure S5

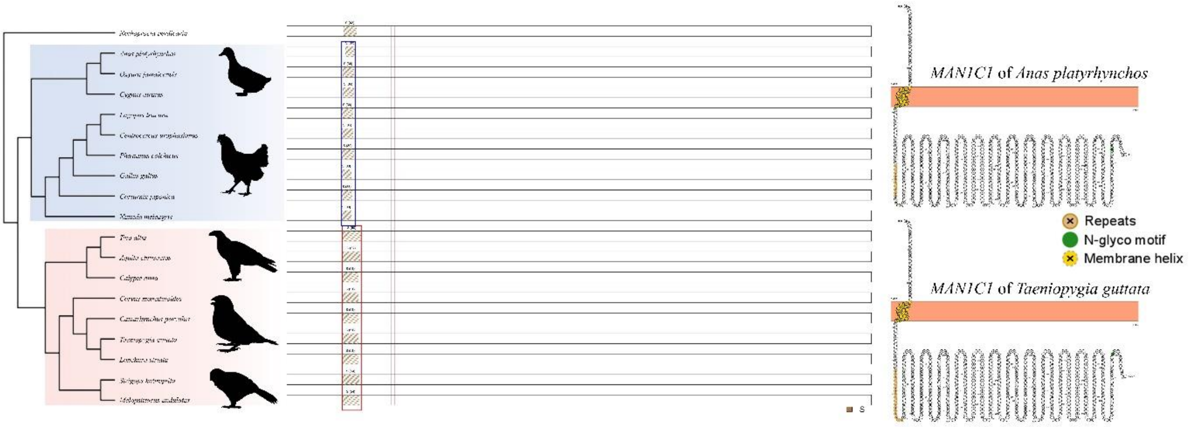

Figure S5: Length volatility of polyS repeat in MAN1C1 across Avian orders. The relative position of the orthologous repeat in the gene (depicted by burlywood3 colored dashed lines) shows the contracted length (species inside the blue box) in Galloanseriformes compared to Neoaves (species inside the red box). The numbers in brackets indicate the number of nucleotides in that repeat stretch. The species tree is downloaded from the BirdTree website. The vertical brown lines represent positively selected sites. The right side images are the schematic representation of the secondary structure of *MAN1C1* in *Anas platyrhynchos* and *Taeniopygia guttata*. The *MAN1C1* gene primarily localizes in the Golgi apparatus. The

peach color bar represents the Golgi membrane, and the orange and yellow colors represent the repeat region and membrane helix.

Figure S6

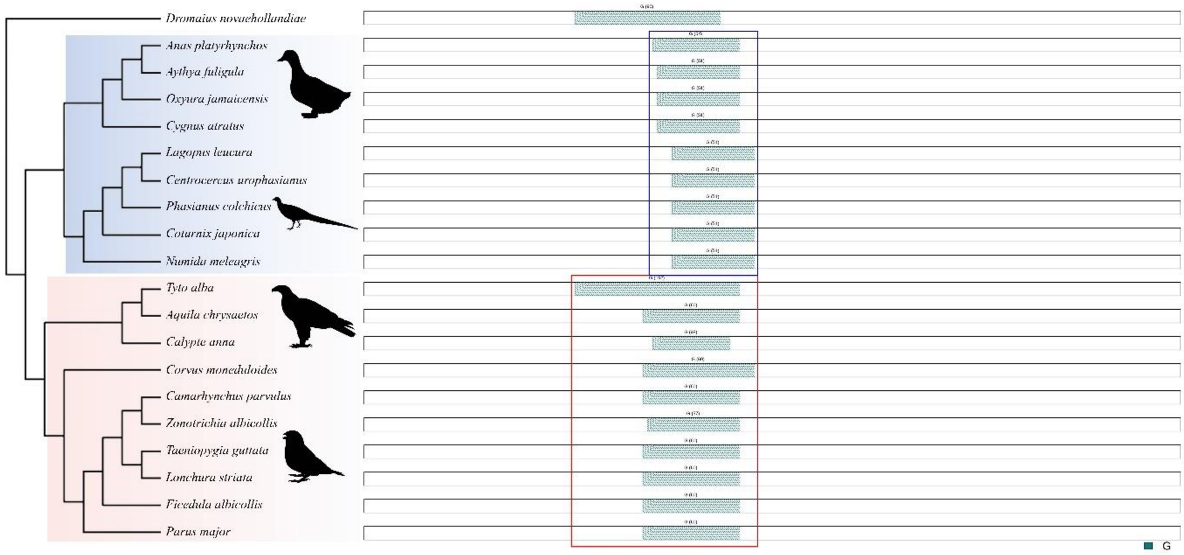

Figure S6: Overall contracted polyG repeat of NHLH2 in Galloanseriformes than Neoaves. Colored boxes indicate the relative position of the repeat in the gene, and the amino acid and its nucleotide length in brackets are provided above the box. The red and blue colored boxes around the species represent the group of species with substantially greater repeat lengths compared to the other species having shorter repeat lengths, as detected by PIC. The phylogenetic tree is downloaded from the BirdTree website.