Terminal regions of a protein are a hotspot for repeats and selection

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

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

*Correspondence: nagarjun@iiserb.ac.in

Supplementary text

Dataset preparation

We downloaded a list of all human protein-coding genes from Ensembl Biomart release 105 (Howe et al. 2021) and filtered out all the mitochondrial, read-through, and LINCs from the list. We downloaded the remaining protein-coding genes using NCBI datasets (Sayers et al. 2020). The downloaded gene sequences are subsetted into 13 clades (Afrotheria, Rodentia, Chiroptera, Carnivora, Perissodactyla, Primates, Artiodactyla, Squamata, Testudines, Aves, Lagomorpha, Marsupials, and Amphibia: 308 species in total). The species tree of the Aves clade is downloaded from BirdTree (Jetz et al. 2012), while the remaining clades' species tree is downloaded from TimeTree (Kumar et al. 2017). Within each clade, we selected the most similar amino acid sequences for each gene based on the following procedure: (i) We choose all the "NP" denoted sequences for a gene in a clade and align the sequences. Using the aligned sequences, we generate a consensus sequence, (ii) If a species has only one "NP" sequence for a gene, the sequence is considered the final sequence of that species, (iii) If a species has more than one "NP" sequences, each sequence is aligned with the consensus sequence, and the one with best per base alignment score is selected, and (iv) if no "NP" sequence exists for a species, all the "XP" sequences of that species for that gene are aligned with the consensus sequence. The one with the best per base alignment score is selected.

Filtering and alignment

We map all the selected amino-acid sequences to their respective CDS sequence. We remove all the sequences with incomplete ORF (absent START codon, absent STOP codon, contains non-nucleotide characters, or sequence not a multiple of three). If a multifasta file for a gene in a clade contains less than four species, we drop it. Surprisingly, after applying the filtering criteria, the Lagomorpha clade gets completely filtered out and thus is not included in molecular evolutionary analyses. The remaining files are codon-aligned using the GUIDANCE (Sela et al. 2015) program with MUSCLE (Edgar 2004) aligner with 100 bootstraps.

Repeats identification and related analyses

We use fLPS2.0 (Harrison 2021) to identify stretches of low probability in the amino acid sequences. Stretches that are longer than four amino acids, have more than 70 % composition, and are composed of less than five different amino acids are considered repeats. The coordinates of repeat stretch are mapped on their respective aligned CDS sequence using a custom script. The stretches with overlap are considered orthologous repeats.

GO enrichment analyses

The Gene Ontology (GO) enrichment analysis is performed on the list of genes containing repeats with all protein-coding genes as background using ShinyGO 0.76.1 (Ge et al. 2020). We performed GO analysis for biological processes, molecular functions, and cellular components (**see Figure S1, S2, and S3**). Similarly, we performed GO enrichment analyses for biological processes, molecular functions, and cellular components on genes with positively selected sites with all protein-coding genes as background. We extended our GO enrichment analyses by classifying repeats and positively selected sites (PSS) according to their position in a gene. The repeats or PSS occurring in the first 20 % or last 20 % of the gene are termed as terminal (terminal-repeat or terminal-PSS), while the rest are termed as mid-repeats or mid-PSS, respectively. Below we have summarized the gene sets and GO analyses:

Foreground gene set	Background gene set	GO Analyses
Repeat-containing genes	All protein-coding genes	BP, MF, CC
Terminal-repeat genes	Repeat-containing genes	BP (NS), MF (NS), CC

Mid-repeat genes	Repeat-containing genes	BP, MF, CC
PSS-containing genes	All protein-coding genes	BP, MF, CC
Terminal-PSS genes	PSS-containing genes	BP (NS), MF, CC (NS)
Mid-PSS genes	PSS-containing genes	BP, MF, CC

The table summarizes all the GO enrichment analyses with foreground and background gene sets. The list of genes used for GO enrichment analyses is mentioned in Supplementary Table S2.

BP: Biological process

MF: Molecular function

CC: Cellular component

PSS: Positively selected site

NS: Non-significant result

For the violin plot of length distribution of repeats, we selected all the repeats for each species and calculated their length in nucleotides. The violin plot is generated in R (R Core Team 2021) (**see Figure 1**).

Simpson's diversity index requires two parameters: the number of different types of repeats and their relative abundance. We calculated the number and relative abundance of different amino acid repeats in R. Using this, we calculated the Simpson's diversity index for each species and used the values for the boxplot (**see Figure 2**). The boxplot shows the distribution of Simpson's diversity index of all the species in that clade. Similarly, we computed the boxplot using Shannon's diversity index (**see Figure S4**).

Species available on public datasets have varying levels of assembly and annotations and affect the number of complete ORFs we can recover in a species. The species with fewer complete ORFs or annotated amino acid sequences will show less number of overall repeats detected. The number of complete ORF and the number of repeats detected shows a linear correlation on the log scale for each clade (**see Figure S5**). We calculated the number of amino acid sequences present for each species and the number of unique repeats detected. We use this information to correlate the log(number of amino acid sequences) with the log(number of unique repeats) for each clade.

Each clade can have a clade-specific abundance of particular repeat types. We list each clade's twenty most abundant repeats and calculate their proportion. We plot the stacked barplot using the proportion values of these repeats by coloring each repeat distinctly (**see Figure 3**). All the remaining repeats' proportion is added to make the proportion 1 in each clade. The stacked barplot is generated in R.

For each homopolymer repeat (i.e., a repeat stretch consisting of only one type of amino acid), we wanted to calculate their abundance of occurrence on a normalized gene length. We selected all the repeats of a particular amino acid from all the clades and converted the mid-position of the repeat to a normalized gene position. For example, if an L repeat stretches from 30 to 60 base-pair positions on a gene length of 1000. The mid-position of the repeat is 45 (mid of 30 and 60) on the gene of length 1000. On normalized gene length (i.e., converting gene length to 100), the repeat's mid-position becomes 4.5. We bin the normalized repeats' position in an interval of five, i.e., all the repeats position between 0 to 5 will be in the same bin. Similarly, all the normalized repeat positions between 20 to 25 will be in the same bin. This gives us a total of twenty bins. Later we calculate the proportion of the repeat in each bin. If the repeat is distributed completely randomly, the proportion of repeats in each bin should be close to 0.05 (or 5%). We consider the repeat over-abundant in a bin if the occurrence is more than 8% and the under-abundant for less than 2%. **Figure S6** shows the L amino acid repeat distribution along the normalized gene position. The grey box represents the frequency of random occurrences.

Molecular evolutionary analyses

We compare site models M7 and M8 of PAML on aligned gene sequences to detect positively selected sites. The sites with posterior probability > 0.95 are considered significant positive sites. The positive sites are binned on normalized gene position in bins of 10. The ω (dN/dS: non-synonymous substitution rate/synonymous substitution rate) for each lineage is calculated using the branch-free model of PAML.

We binned repeats mid-position in bins of 10 on normalized gene position and visualized the frequency of occurrence in each bin using a histogram. Similarly, we visualized the frequency of positive sites on the same

histogram (**see Figure 4**). We plotted the alignment coverage of all the aligned files on a normalized gene along with repeat and positive site abundance. We divided the alignment coverage by 20 to visualize on the same y-axis margin as positive sites and repeat proportion (**see Figure S7**).

Previous studies suggest a higher GC% of gene sequences with repeats than genes without repeats. The studies were conducted on a limited number of genes and species. Here, we calculated the GC% for each gene sequence and sorted the sequences according to "repeat status." We plotted the distribution of GC% of genes with repeats and without repeats for each clade using a boxplot (**see Figure S8**). The distribution of GC% is compared using Wilcoxon's test in R.

Similarly, we subsetted all the lineages' ω values less than two according to their "repeat status," compared using boxplot, and used Wilcoxon's test for the significant difference in the distributions (**see Figure 5**). We also compared the distributions using all the ω values (**see Figure S9**). The barplots are shown without outliers for better visualization. The value above the bars shows the p-value of Wilcoxon's test. The values below the bar represent the number of sequences in that distribution (represented by "n") and the mean of ω (represented by x-bar).

References

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

Ge SX, Jung D, Jung D, Yao R (2020) ShinyGO: A graphical gene-set enrichment tool for animals and plants. Bioinformatics 36:2628–2629. https://doi.org/10.1093/bioinformatics/btz931

Harrison PM (2021) fLPS 2.0: Rapid annotation of compositionally-biased regions in biological sequences. PeerJ 9:. https://doi.org/10.7717/peerj.12363

Howe KL, Achuthan P, Allen J, et al (2021) Ensembl 2021. Nucleic Acids Res 49:D884–D891. https://doi.org/10.1093/nar/gkaa942

Jetz W, Thomas GH, Joy JB, et al (2012) The global diversity of birds in space and time. Nature. https://doi.org/10.1038/nature11631

Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34:1812–1819. https://doi.org/10.1093/MOLBEV/MSX116

R Core Team (2021) R: A Language and Environment for Statistical Computing

Sayers EW, Beck J, Brister JR, et al (2020) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz899

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–W14. https://doi.org/10.1093/nar/gkv318

Supplementary Figures

Figure S1

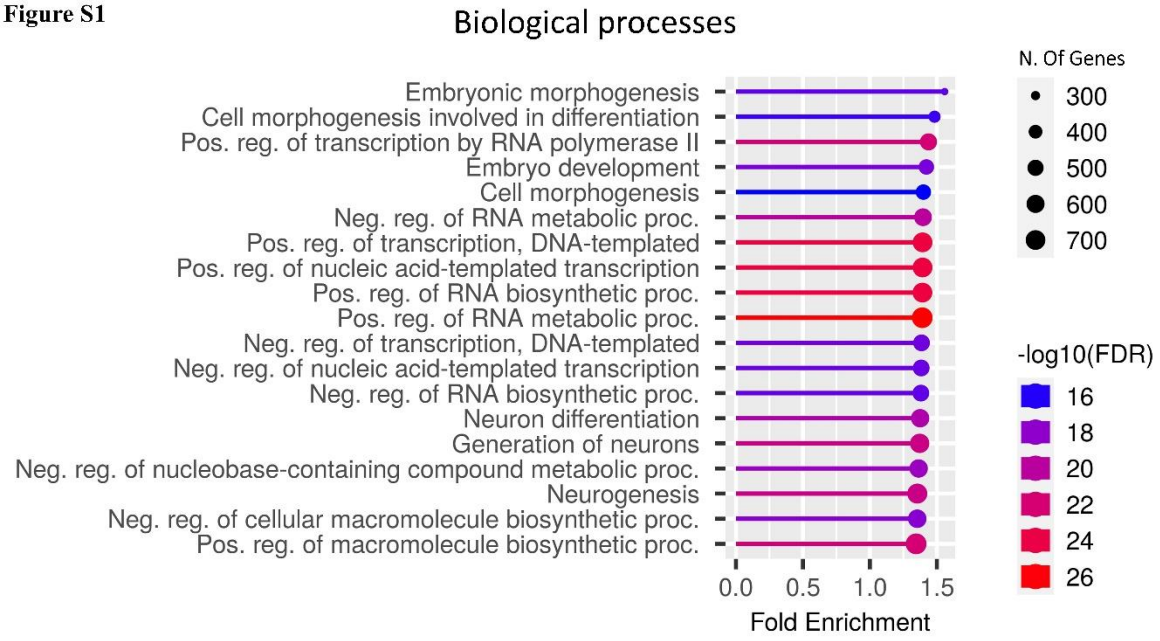

Figure S1: Gene ontology (GO) enrichment analysis of genes with repeats for biological processes.

Figure S2

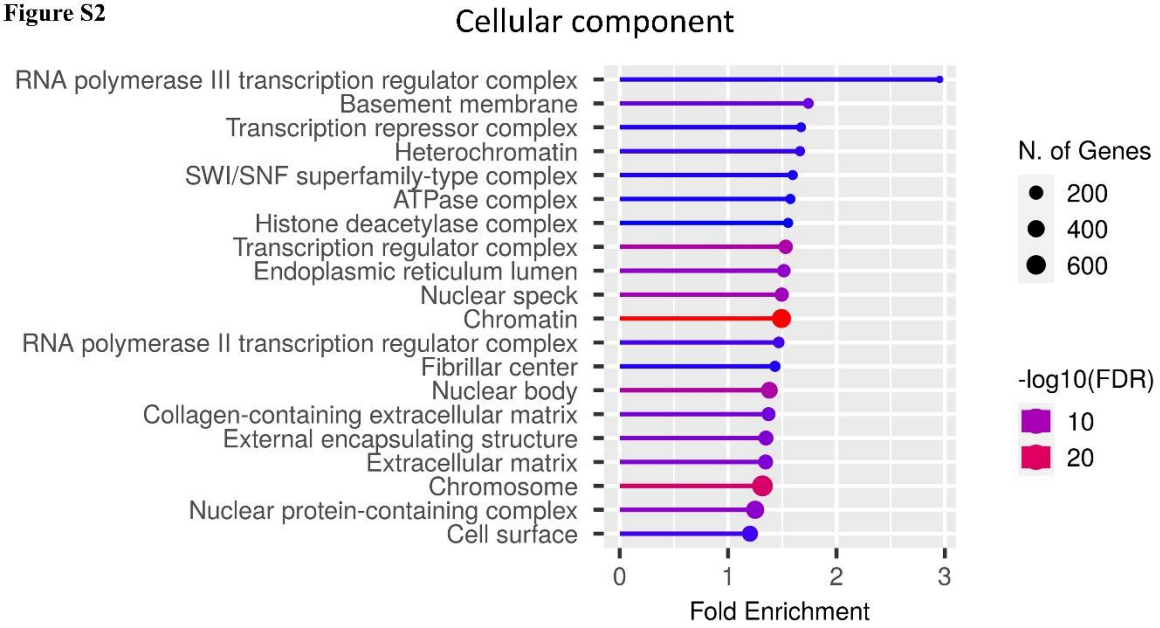

Figure S2: The figure shows enrichment for cellular components using GO enrichment analysis in genes with repeats.

Figure S3

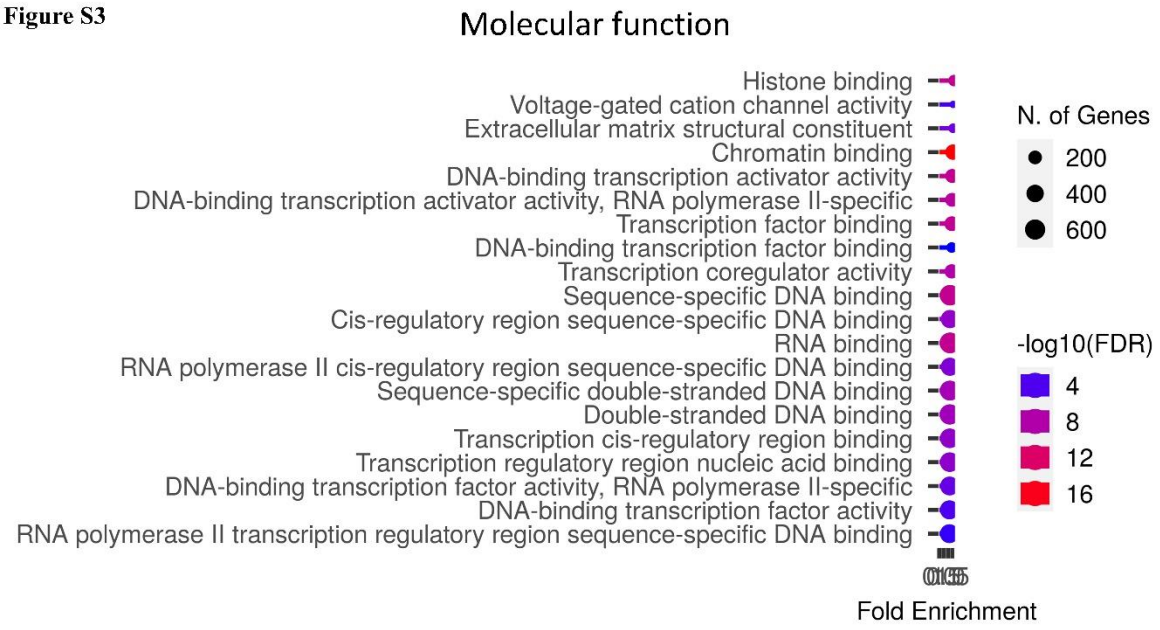

Figure S3: Enrichment in molecular function in genes containing repeats predicted using GO enrichment analysis.

Figure S4

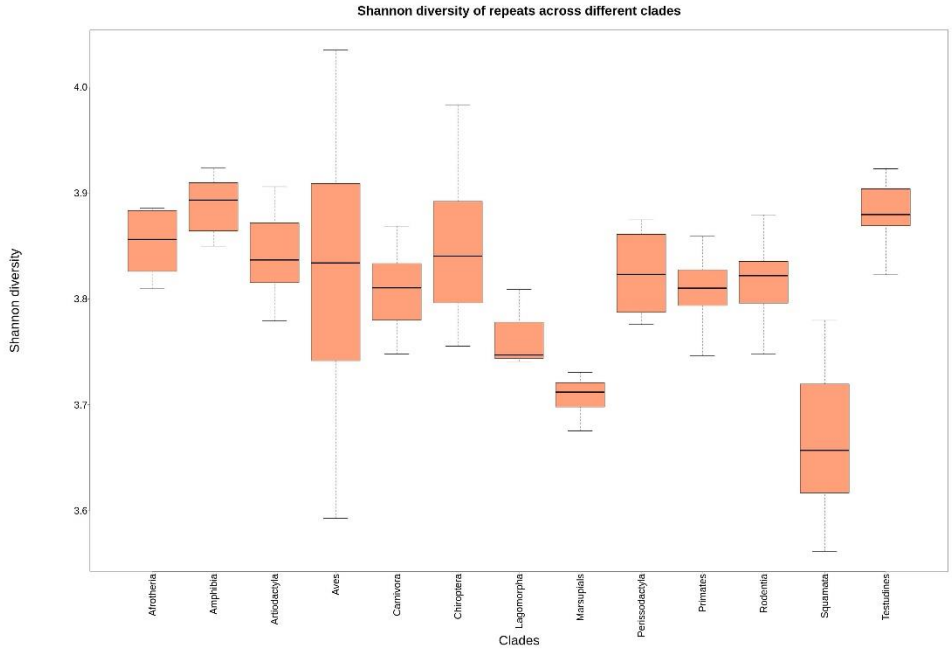

Figure S4: Distribution of Shannon's diversity index of repeats in each clade.

Figure S5

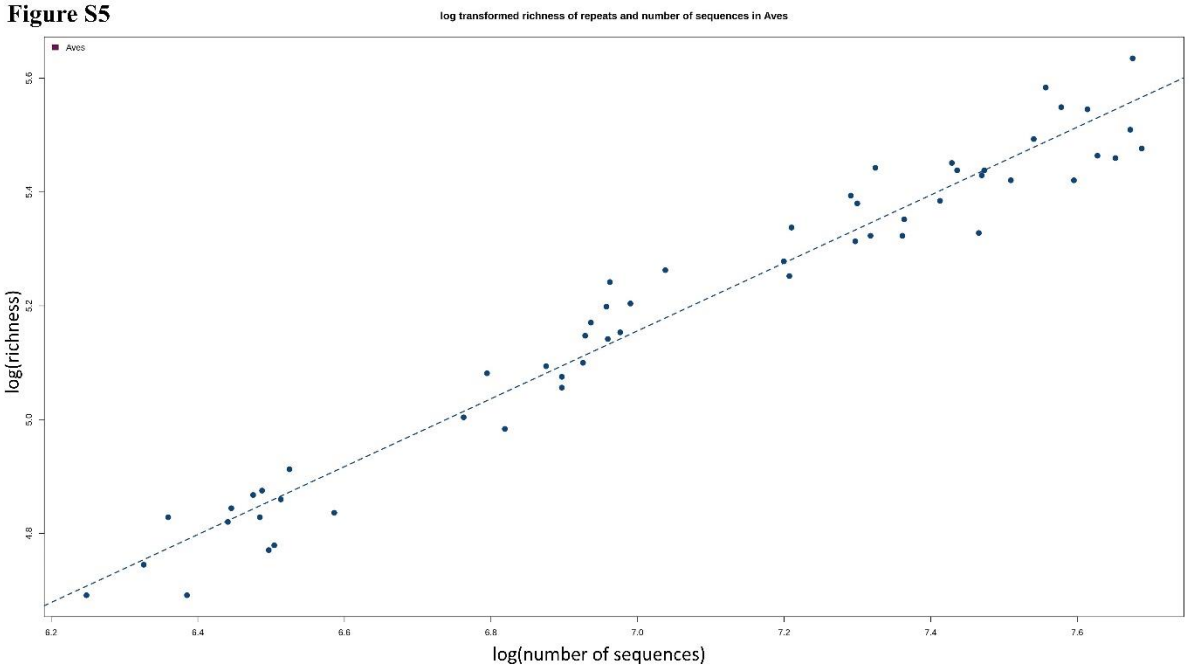

Figure S5: Correlation of log(types of repeats) or log(richness of repeats) with log(number of gene sequences).

Figure S6

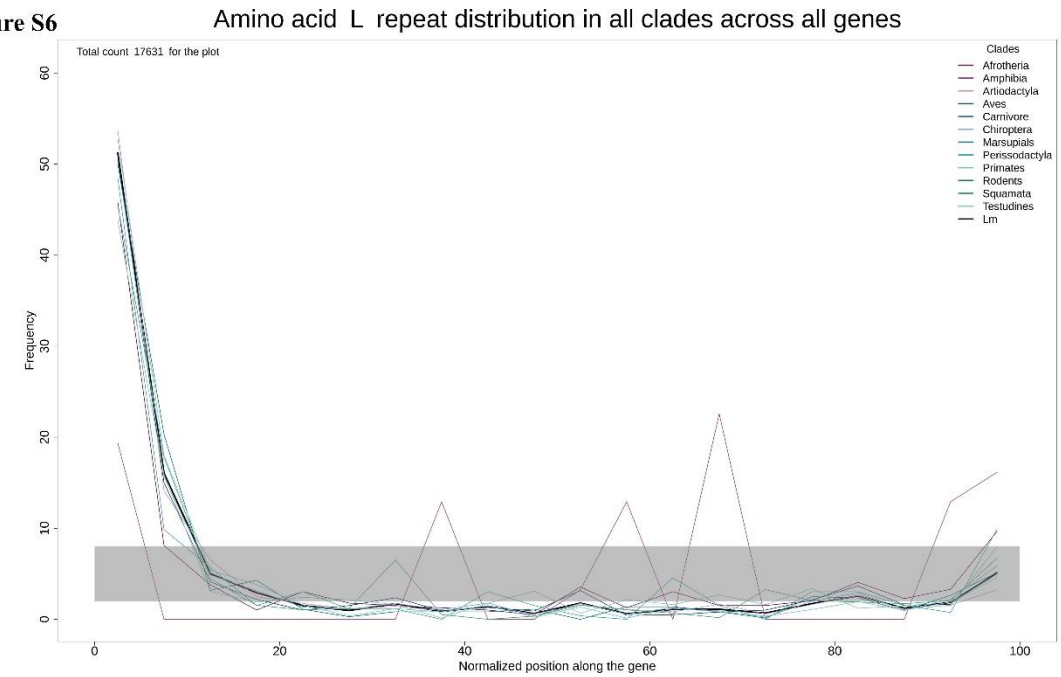

Figure S6: Frequency of Leucine (L) amino acid occurrence repeat along a normalized gene length.

Figure S7

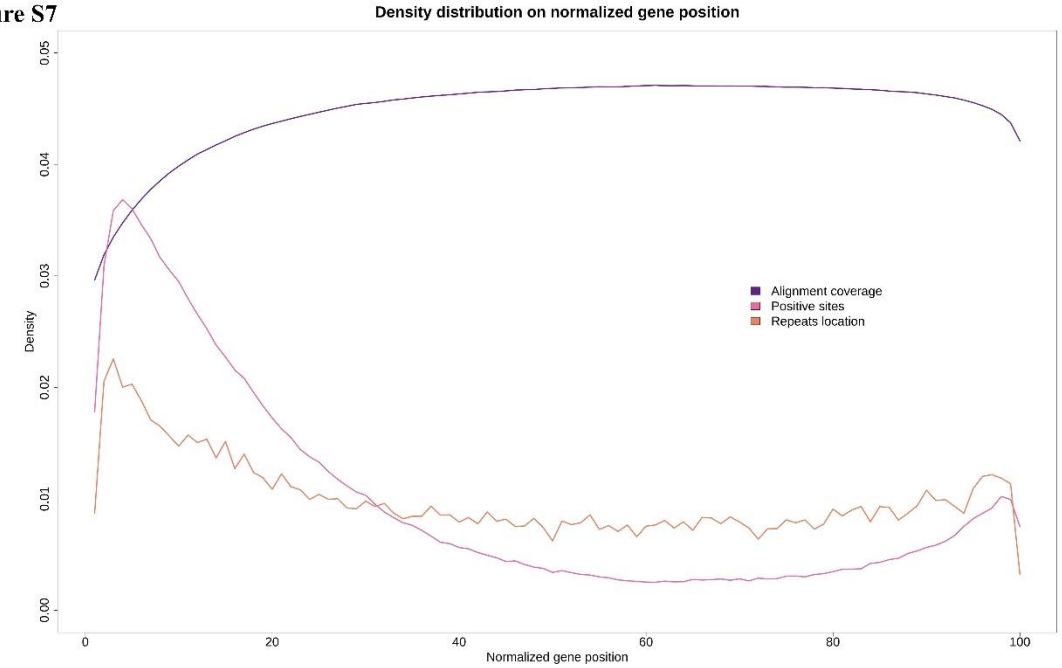

Figure S7: Proportion of positively selected sites, repeats, and alignment coverage (original alignment coverage is divided by 20 for better visualization with positive sites and repeats) along the normalized gene length.

Figure S8

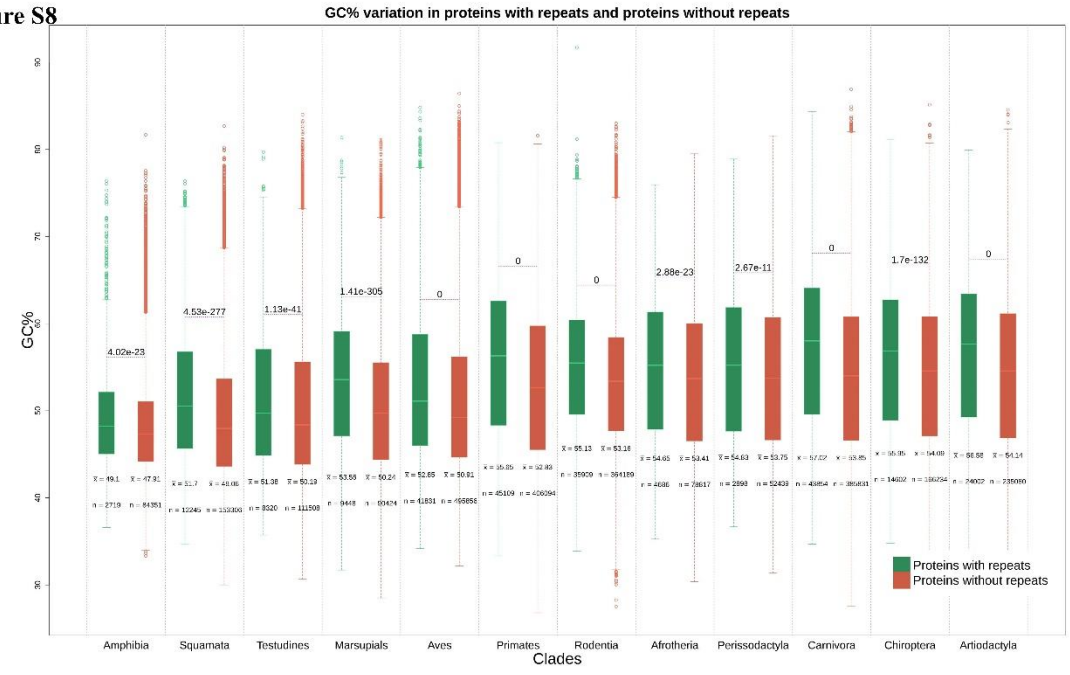

Figure S8: Distribution of GC% of genes with repeats and without repeats in a clade-wise manner.

Figure S9

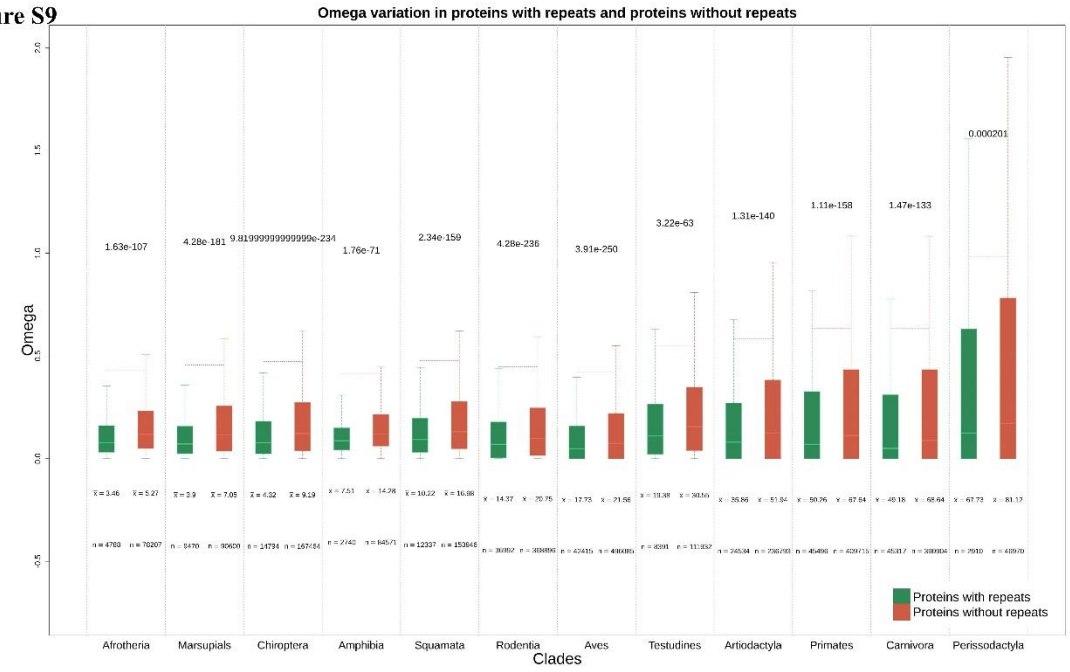

Figure S9: Distribution of ω (dN/dS: non-synonymous substitution rate/synonymous substitution rate) of genes with and without repeats.