Supplementary Material

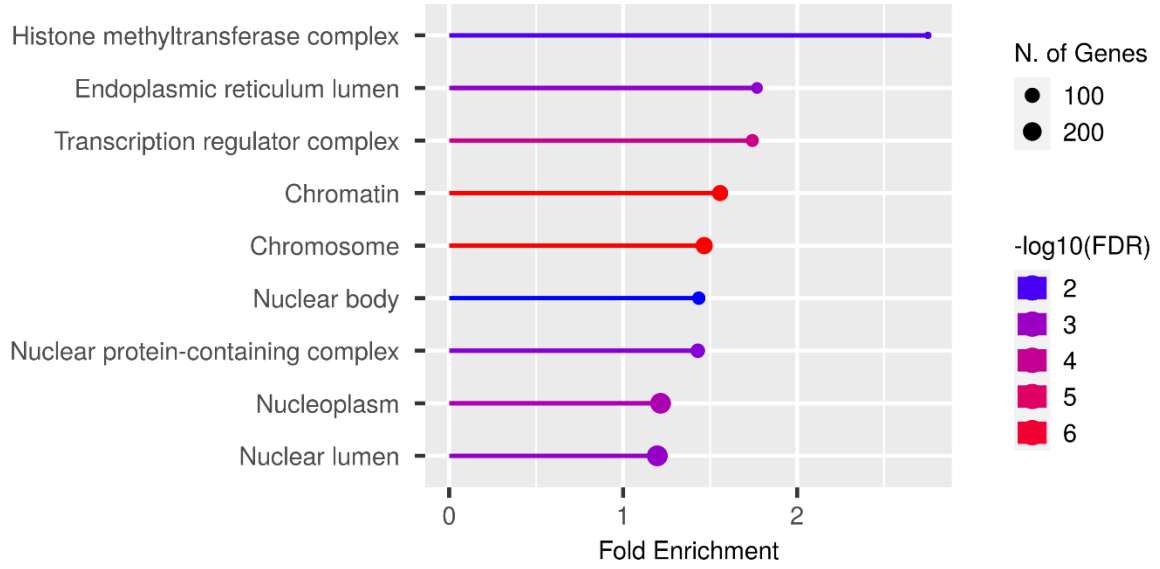
Lineage-specific protein repeat expansions and contractions reveal malleable regions of immune genes

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

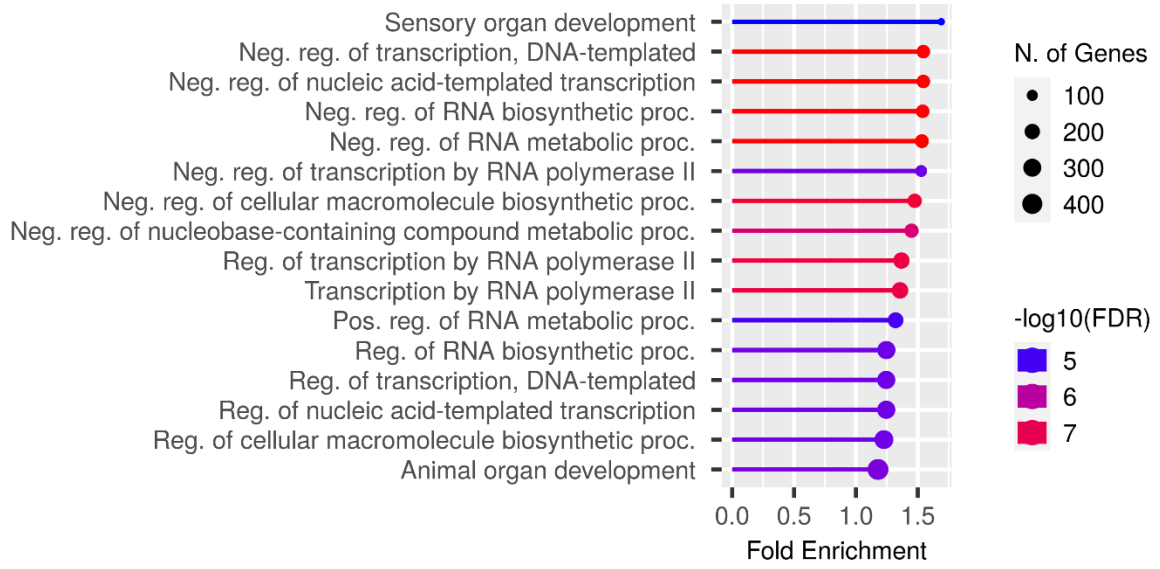
[1]Computational Evolutionary Genomics Lab, Department of Biological Sciences, IISER Bhopal, Bhauri, Madhya Pradesh, India

*Correspondence: nagarjun@iiserb.ac.in

Suppl. Fig. 1



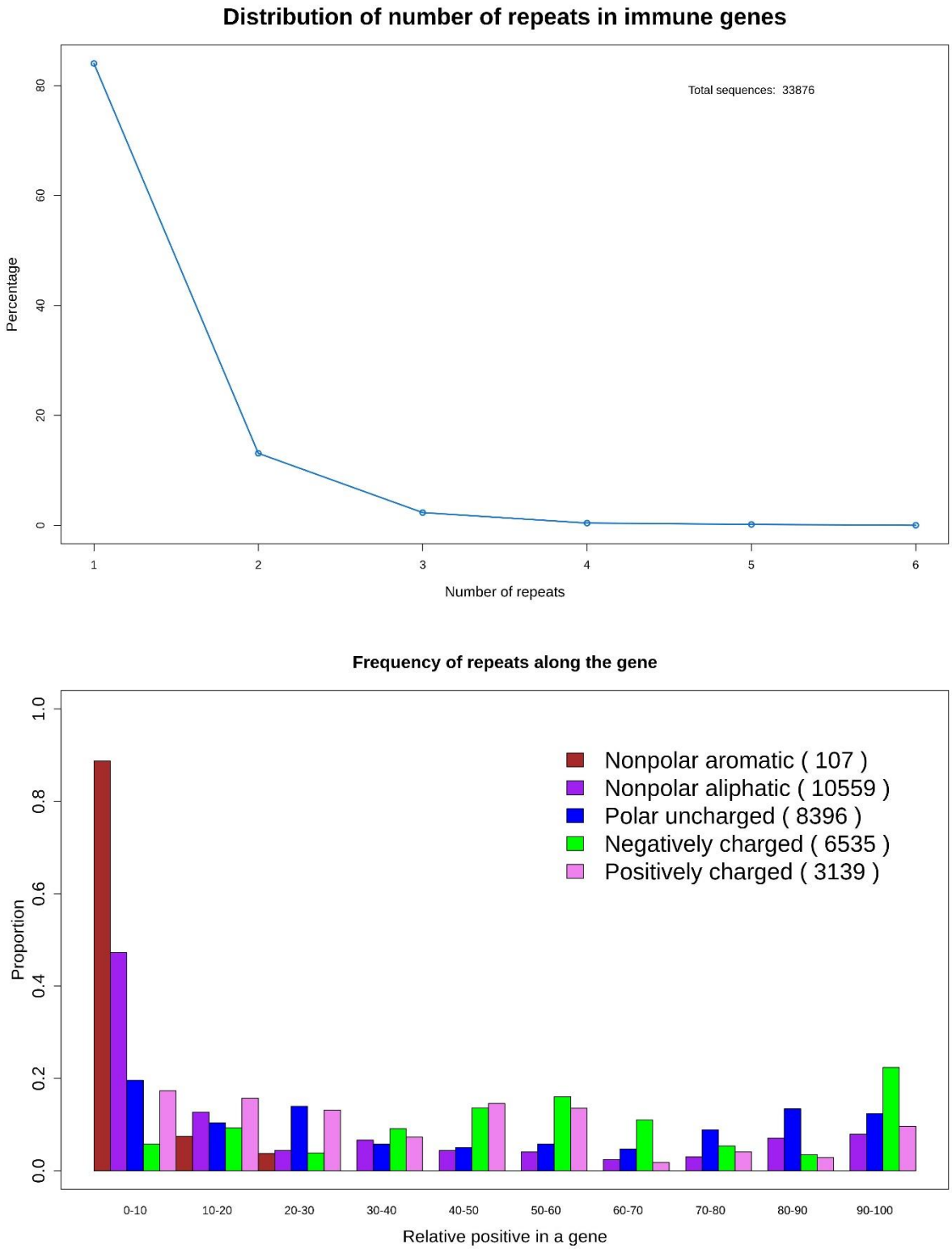
Cellular components



Biological Process

Suppl. Fig. 1: The GO enrichment of immune system process genes for cellular components and for biological processes.

Suppl. Fig. 2



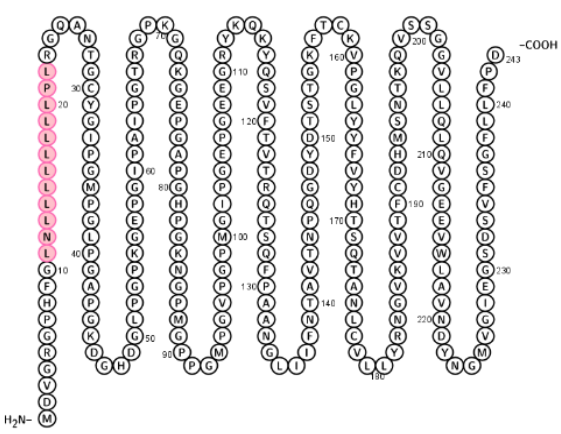
Suppl. Fig. 2: Distribution of repeats in immune genes. **A** Percentage of immune genes containing different number of repeats. Most of the immune genes contain only one repeat. **B** Distribution of repeats by properties in normalized gene position. Most of the repeats appear in the beginning of

the gene. The numbers in the bracket represent the number of homopolymer repeats detected in immune genes.

Suppl. Fig. 3

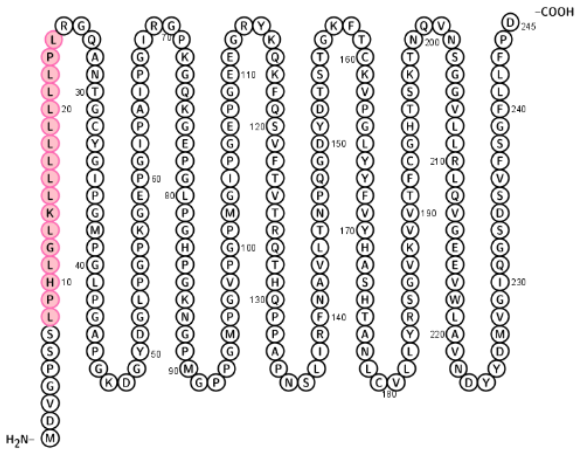
Suppl. Fig. 3: The pie chart shows the frequency of expanded and contracted genes in the Amphibia clade. Red represents the genes with significantly expanded repeat lengths, and blue are the genes that have contracted significantly repeat lengths. The number of genes under significant expansion and contraction in each species is given in brackets.

Suppl. Fig. 4

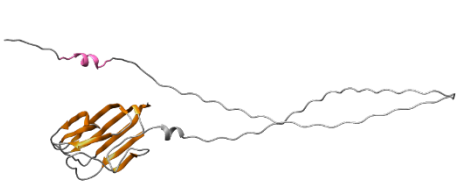


Repeats

C1QC of *Cebus imitator*

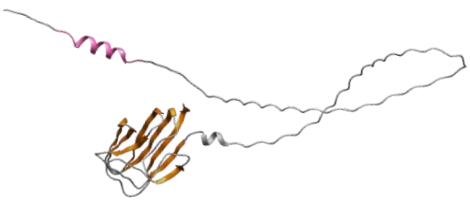


C1QC of *Homo sapiens*



Strands

Repeats



C1QC of *Cebus imitator*

C1QC of *Homo sapiens*

Suppl. Fig. 4: Repeat length distribution of *C1QC* gene across different species in Primates clade with phylogenetic tree downloaded from the TimeTree website. The repeats are plotted at their respective relative position along the gene. The number in brackets represents the number of nucleotides making the repeat stretch. The red box represents the group of species with longer repeat lengths, while the blue box depicts significantly shorter repeat lengths. The purple box represents the positively selected species detected by aBSREL, and the brown vertical lines are the positively selected sites detected by codeML. We downloaded the secondary structures of one species with significantly expanded and one species with significantly contracted from the Protter website, and the repeat regions are highlighted in pink. We visualized the protein structure of the *C1QC* gene in ChimeraX and highlighted the repeat region. SWISS-MODEL is used to model the protein structure of the C1QC gene of the *Cebus imitator,* while the human protein structure is downloaded from the AlphaFold Protein Structure Database.