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

Lokdeep Teekas
https://orcid.org/0000-0003-2761-3977

Sandhya Shama
https://orcid.org/0000-0002-0987-2521

Nagarjun Vijay (nagarjun@iiserb.ac.in)
IISER Bhopal https://orcid.org/0000-0002-8703-6991

Keywords: immune genes, positive selection, protein repeats, repeat length variation, low complexity sequences

Posted Date: July 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1761690/v1

License: ☒ ☛ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at Genes & Immunity on October 6th, 2022. See the published version at https://doi.org/10.1038/s41435-022-00186-4.
Abstract

Functional diversification, a higher evolutionary rate, and intense positive selection help a limited number of immune genes interact with many continuously evolving pathogens. Repeats in protein-coding regions are a well-known source of rapid functional diversification, adaptive variation, and evolutionary novelty. Repeats play a crucial role in biochemical functions like functional diversification of transcription regulation, protein kinases, cell adhesion, signaling pathways, morphogenesis, DNA repair, recombination, and RNA processing. Repeat length variation can change the associated protein's interaction, efficacy, and overall protein network. Repeats have an intrinsic unstable nature and can potentially evolve rapidly and expedite the acquisition of complex phenotypic traits and functions. Because of their ability to generate rapid, adaptive variations over short evolutionary distances, repeats are considered "tuning knobs." Repeat length variation in specific genes, like RUNX2 and ALX4, is associated with morphological and physiological changes across vertebrates. Here we study repeat length variation as a potent source of species-specific immune diversification across several clades of tetrapods. Moreover, we provide a clade-wise comprehensive list of immune genes with repeat types for future studies of morphological/evolutionary changes within species groups. We observe significant repeat length variation of FASLG in Rodentia and C1QC in Primates' contrasting species groups.

Introduction

Immune genes are well-known hotspots of intense positive selection in several clades of vertebrates. The abundance of selection signatures in immune genes results from an ongoing arms race between the host and the pathogen. Adaptive evolution occurs in most interferon-stimulated genes. In addition to the adaptive evolution of genes, changes in the gene content have helped deal with the changing pathogen repertoire. Host genomes have also responded to pathogens through other strategies such as splice isoforms and several forms of non-coding RNA that counter pathogens. Hence, understanding the source of genomic novelty which can take on immune roles is an important question. Repeats in protein-coding genes are a source of evolutionary novelty but have received less attention.

Proteins of diverse life forms contain recurring motifs like tandem repeats, repetitive domains, and periodic structures. In the human genome, around 20–28% of proteins have at least one stretch of homo-peptide residues. We refer to such amino acid repeat sequences of low complexity as repeats. Repeats play a crucial role in biochemical functions like functional diversification of transcription regulation, protein kinases and cell signaling pathways, cell adhesion, morphogenesis, DNA repair, recombination, development, and RNA processing. Repeats have an intrinsic unstable nature and can potentially evolve rapidly and expedite the acquisition of complex phenotypic traits and functions. Because of their ability to generate rapid, adaptive variations over short evolutionary distances, repeats are considered "tuning knobs." A combination of relaxed purifying selection and positive selection drives fast divergence of the repeat sequences, presumably yielding new functions. These novel functions may be of adaptive significance in different ecological niches.
Repeats are highly unstable and prone to contraction or expansion by recombinational mechanisms (like gene conversion or unequal crossing-over) or replication-slippage\textsuperscript{22}. This volatile nature of repeats contributes to functional and morphological diversity, such as the repeat length variation in the case of the \textit{RUNX2} gene\textsuperscript{28,30}. \textit{RUNX2} is a DNA-binding transcription factor that plays a crucial role in skeletal development in vertebrates\textsuperscript{31}. \textit{RUNX2} contains two repeats: poly-Q and Poly-A adjacent to each other\textsuperscript{32}. Studies have found a positive correlation between Q/A ratio with facial shape and angle of facial bones compared to skull length between different breeds of dogs\textsuperscript{28}. The QA repeat length variation in \textit{RUNX2} contributes to the evolution and diversity of skeletal structure across vertebrates\textsuperscript{30}.

The importance of repeats in parasites for cell adhesion, invasion, and immune evasion of host cells is overwhelming\textsuperscript{33–36}. Several repeat-containing proteins are involved in host-parasite interactions, such as host cell entry, survival, adaptation to the intracellular environment, and escape from the host immune system\textsuperscript{33–36}. Furthermore, specific repeats like Leucine-rich repeats (LRRs) are known to be exploited by both hosts and pathogens in host-pathogen interactions. The pathogens use LRRs to attach to the host cell and enter it. In contrast, the host cells utilize LRRs to recognize pathogen-specific molecular patterns\textsuperscript{34,37}, where they sense specific pathogen-associated molecules and activate the innate immune system\textsuperscript{38,39}. Due to their variable nature, high abundance in immune-related genes, and consequences on function and morphology, immune genes' repeats can potentially contribute to the clade-specific diverse immune repertoire.

In contrast to previous studies, which focus on the functional implication of specific repeats, we characterize and quantify lineage-specific changes in several Tetrapoda clades by investigating the role of repeats in all immune genes. Repeats occur in a third of the immune genes and have experienced lineage-specific changes in length and composition. We reasoned that protein repeats occur in malleable regions of immune genes and are a potent source of evolutionary novelty. Our analysis focused on identifying the global patterns of a repeat occurrence and expansion/contraction. In our multi-clade dataset, most repeats occur in the initial 20\% or the final 10\% of the protein and genes with a specific range of GC content. The proportion of genes with repeat expansion or contraction is species-specific. Our in-silico study identifies candidate genes with species-specific changes in repeat length, which may have functional implications.

**Materials And Methods**

**Dataset preparation**

To use a well-curated and stable set of genes with immune functions, we rely upon gene ontology (GO) annotation. We compiled a list of 2945 immune genes by downloading human genes annotated with the term "immune system process" (Accession # GO:0002376) from Ensembl release 106\textsuperscript{40} (see Supplementary Table S1). We obtained the full-length coding sequences and corresponding protein sequences of human immune genes and their one-to-one vertebrate orthologs for 216 species spanning
12 clades from NCBI\textsuperscript{41} (see Supplementary Table S2). We filtered out read-through annotations and those genes which lacked ortholog annotation across at least four species for each clade resulting in a final set of 2488 genes. In the case of multiple splice isoforms, we chose the sequence with the most similar length and sequence identity to the other orthologs. We downloaded the species tree for non-avian clades from the TimeTree\textsuperscript{42} and the BirdTree\textsuperscript{43} websites for birds.

Identification of protein repeats

We identified the repeats in protein-coding regions in protein sequences using fLPS2.0\textsuperscript{44} with default settings. For each gene, we provided the clade-specific composition matrix as a background. We filtered out repeats identified by fLPS2.0 if (a) they were shorter than a four amino acid stretch or (b) the repeat purity of the stretch was less than 70%, or (c) they consisted of more than four different amino acids. This criterion ensured that we only focused on a conservative, reliable set of repeats. We removed repeats containing "X," corresponding to gaps in the sequence or annotation errors in the protein. The complete list of repeat locations is available in Supplementary Table S3.

Comparison of orthologous sequences

We first excluded coding sequences with premature stop codons and subsequently removed genes with sequences in less than three species within each clade. We performed multiple sequence alignment of the coding sequences (codons) using the MUSCLE\textsuperscript{45} aligner with 100 iterations. We mapped the location of protein repeats onto the aligned nucleotide coding sequences. Subsequently, we identified orthologous repeats based on positional overlap in the alignment. We performed a phylogenetic correction of the length for each orthologous repeat using the Phylogenetically Independent Contrasts (PIC) method implemented in the phytools\textsuperscript{46} R\textsuperscript{47} package. The most extreme values of the phylogenetically corrected repeat lengths for the internodes were considered significantly contrasting between the corresponding species groups. The group with a higher mean repeat length was expanded, while the other group was contracted between the contrasting species groups (see Supplementary Table S4). We filtered out those cases in which either of the contrasting species groups had less than three species to ensure reliability of the results (see Table 1).

Molecular evolutionary analysis

We identified lineage-specific positive selection using the aBSREL\textsuperscript{48} program from the HyPhy\textsuperscript{49} package. We also identified positively selected sites for each gene using codeml from the PAML\textsuperscript{50} package by comparing models M7 and M8.

Characterization, visualization, and analyses

We did gene enrichment analysis using ShinyGO 0.76 server\textsuperscript{51} (see Supplementary Table S5). The protein structures of genes of interest are downloaded from Alphafold Protein Structure Database\textsuperscript{52}. If the protein structure of a specific gene for a species is unavailable on Alphafold, we predicted the structure
on the SWISS-MODEL server. All the protein structures are visualized in ChimeraX v1.16. To test whether overlap of repeat regions with positively selected amino acids occurs more than chance, we used Bedtools fisher (v2.26.0). All other necessary analyses are done in R.

**Results**

**Protein repeats occur in a third of the immune genes**

Of the 2488 immune genes with well-annotated sequences across multiple species (Supplementary Table S1 and S2), we found repeats in 823 genes. The Gene Ontology (GO) analysis of molecular functions of repeat-containing genes is enriched for molecular binding (Integrin binding, transcription factor binding, DNA binding, receptor binding, and kinase binding) (Fig. 1A and Supplementary Table S5). Moreover, repeats are enriched in genes involved in immune cells’ differentiation, development, and activation (Supplementary Fig. 1A). Interestingly, we also found that repeat-containing genes show significant enrichment in the nBAF complex (Supplementary Fig. 1B). We found that of the 722 genes in adaptive immunity (GO: 0002250), 141 genes contain repeats. Of the 922 genes involved in innate immunity (GO:0045087), 198 have repeats (Fig. 1B).

**Repeats are abundant in specific GC content genes**

We observed that most repeats occur at the beginning of the genes (initial 20% of the gene) or the terminal (final 10% of the gene). We also observed GC content of the genes with repeats ranging between 42-65% (Fig. 2A and Supplementary Table S6). To test whether there is a difference in GC content of repeat-containing genes and genes without repeats, we compared the GC% of both classes for each clade. Our multi-clade dataset shows that the repeat-containing genes have significantly higher GC content than genes without repeats in each clade (Fig. 2B, Wilcoxon test for significance).

**Immune genes may contain multiple repeats**

Several immune genes consistently have multiple orthologous repeats across clades (Fig. 3A). Some genes have up to five repeats, for example, in the case of ARID1B. We also quantified the overlap of positively selected sites with repeat regions using Fisher’s exact test and found that 59 immune genes favor positively selected sites in repeat regions (Supplementary Table S7). Of these genes, SETD1A significantly preferred amino acid sites in repeats in three clades, of which Primates is shown in Fig. 3B. We also found that 2223 lineages across the clades contain repeats and are under positive selection (Fig. 3C). We shortlisted the genes for further analyses with the same orthologous repeat in at least six species in at least three clades. Of the shortlisted genes, we highlighted the repeats in the 3D structure of proteins for DHX9 and SETD1A genes in Homo sapiens (Fig. 3D and 3E).

**Repeats are highly variable in length**
Orthologous repeats can have species-specific or clade-specific changes in length. We hypothesized that longer genes might tolerate more variations in repeats than shorter ones. To test this, we correlated the gene and orthologous repeat lengths for all the species (Fig. 4A). We observed that repeat length correlates significantly, but weakly, with gene length (Kendall's rank correlation: $\tau = 0.0308$, $p < 2.2e-16$, $n=40532$). We also observed that most genes with more than ten thousand nucleotide gene sequences have smaller repeats of similar length (e.g., KMT2A, LRP1, LYST, PKHD1L1, and RNF213). Only two genes, PRG4 and HRG, show large length variation for orthologous repeats between different clades (380-2876 nucleotides for PRG4 and 123-867 nucleotides for HRG). We also observed that most genes with more than ten thousand nucleotide gene sequences have smaller repeats of similar length (e.g., KMT2A, LRP1, LYST, PKHD1L1, and RNF213). Only two genes, PRG4 and HRG, show large length variation for orthologous repeats between different clades (380-2876 nucleotides for PRG4 and 123-867 nucleotides for HRG). We also visualized the repeats of HRG and PRG4 across all the clades (Fig. 4B and 4C). We noticed that orthologous repeats show length and composition variation between different species. HRG gene contains only one repeat across all the species considered. In Aves, the gene comprises a repeat of amino acid composition HP (histidine and proline). In contrast, the repeat primarily consists of amino acids GHP (glycine, histidine, and proline) in the rest of the clades. Moreover, the repeat is longer in Aves (237-867 nucleotides long) than in the rest of the clades (123-507 nucleotides long). Similarly, for PRG4, the repeat varies in length and composition across different clades (EPT and PEAT amino acid composition).

Repeats show lineage-specific length variation

Since hypervariability of repeats is a source of rapid, adaptive variations in a short evolutionary time\textsuperscript{27,28}, we tested for significant length variation of orthologous repeats for each clade between species groups (Table 1). Furthermore, we quantified the number of immune genes for expansion and contraction of repeats for each species within their clade (Fig. 5A). In the Primates clade, out of 7 Hominoidea species considered in the study, six species show more genes under contraction of repeats. In the Amphibia clade, Gymnophiona clade (three species considered: Microcaecilia unicolor, Geotrypetes seraphini, and Rhinatrema bivittatum) showed a higher number of genes with contraction of repeats compared to Anura clade (Xenopus tropicalis, Bufo bufo, Nanorana parkeri, and Rana temporaria) (Supplementary Fig. 2). Repeats have unstable nature and can vary in length within species of the same clade, potentially giving rise to lineage-specific evolutionary novelty. To test this, we shortlisted the genes which contain orthologous repeats across the clades and calculated the standard deviation of the length (Fig. 5B). CREB1 and RAB35 show no length variation between species of the same clade, while RUNX2, KAT6A, PRRC2C, and HAND2 show length variation between species. Maintenance of the same repeat length across different species can indicate selective constraints for the same\textsuperscript{22}.

**FASLG**, a transmembrane protein-coding gene, shows lineage-specific length variation of repeat

**FASLG** gene, a member of the tumor necrosis factor superfamily, contains a proline repeat. The proline-rich repeat shows a significant change in length in the Rodentia clade with expansion in Cricetidae (represented by Onychomys torridus, Peromyscus leucopus, Peromyscus maniculatus, Arvicolam amphibius, Microtus ochrogaster, Cricetulus griseus, and Mesocricetus auratus) against Muridae (represented with Meriones unguiculatus, Mus caroli, Arvicanthis niloticus, Rattus norvegicus, and Rattus rattus) (Fig. 6A). The structure of the transmembrane gene, **FASLG**, contains the repeat in the cytoplasmic
side of the protein (Fig. 6B). The protein structure visualization showed the repeat near an α-helix (Fig. 6C). Similarly, within Primates, the C1QC gene has a significant repeat length contraction of leucine repeat in new-world monkeys (12 amino acid long stretch) than the rest of the species (17 amino acids) (Supplementary Fig. 3).

Discussion

The rapid evolution of protein repeats is essential in immune genes like leucine-rich repeats in innate and adaptive immunity. This study focuses on the potential significance of repeat abundance, characterization, and length evolution in tetrapod immune genes. Our study found that around one-third of the immune genes contain repeats which are on par with previous studies of complete gene sets\textsuperscript{18,56,57}. We also found that repeats are enriched in functions related to molecular binding, immune cells' differentiation, development, and function. Innate immunity primarily works based on molecular structure recognition unique to microorganisms\textsuperscript{58} using pattern recognition receptors (PRRs). Each PRR of the host can bind to many molecules of a common structural pattern\textsuperscript{59}. PRRs have several distinct classes, of which toll-like receptors (TLRs) are best characterized. TLRs are transmembrane receptors that recognize bacterial and viral products and activates downstream immune response against it\textsuperscript{60}. Like TLRs, node-like receptors (NLRs) are a well-characterized class of cytosolic receptors in immunity. At the C-terminus, all NLRs contain a leucine-rich-repeat (LRR) and a nucleotide-binding oligomerization domain (NOD)\textsuperscript{61,62}. The NLRs recognize pathogenic particles and activate downstream pathways of the immune response. T-cell and B-cell receptors play a central role in antigen recognition in adaptive immunity. Processes like somatic hypermutation and recombination, non-templated nucleotide addition, and gene conversion generate a diverse receptor repertoire to identify almost any antigen\textsuperscript{63}. The presence of repeats in the genes involved in necessary immunity mechanisms emphasizes the role of repeats in immune pathways.

Among repeat-containing immune genes, we observed GO enrichment for genes of the nBAF complex. The BAF complex is an ATP-dependent complex for chromatin remodeling in mammals consisting of at least 15 subunits\textsuperscript{64,65}. BAF complex regulates DNA accessibility by evicting or sliding nucleosomes. Furthermore, the complex also induces chromatin structure's conformational changes by recruiting histone variants\textsuperscript{64}. The nBAF complex is a specialized neural BAF complex. Mutational studies of the nBAF complex in mice suggest that the complex plays an essential role in the adult brain, and its dysfunction is associated with behavioral abnormalities\textsuperscript{66}. During viral infection, the BAF complex remodels the nucleosome at IFN-β and IFN-α gene promoters and helps in the induction of IFN-α-responsive genes and IFN-β expression\textsuperscript{67}. BAF complexes are required to suppress CD4\textsuperscript{68} and the expression of CIITA to activate MHC class II\textsuperscript{69}. It is worth mentioning that repeat expansion and contraction are associated with numerous neurological disorders\textsuperscript{70}. Enrichment of genes in the nBAF complex further enhances the importance of repeats in neuronal development and immunity.
A relationship between GC content and the occurrence of repeats in mammalian genes is well known. Nucleotide compositional constraints would expedite the formation of repeats in the GC-rich region of the gene. Previous studies suggest positional preference of repeats in protein sequence, with most repeats occurring at the beginning or end of gene sequence and avoiding the middle region. For example, leucine (PolyL) repeats occur either at the beginning or at the end of the gene sequence and have a very low frequency in the middle of the gene sequence. In contrast, PolyT frequently occurs in the middle region of the sequence. We also noticed that most repeats in immune genes occur in the 40–65% GC content range with high frequency in the beginning and end of the gene sequence and very low frequency in the middle region. The positional preference could result from avoiding disorder in domain folds of the protein structure. Genes with repeats have significantly higher GC% than those without repeats across all the tetrapod clades studied. This result suggests that GC content-rich genes have more possibility of acquiring a repeat. As GC-biased codons mainly encode repeats, they increase the overall GC content of the gene even further.

Many proteins lack a well-defined 3D structure and possess their functional unit in the intrinsically disordered or unstructured regions. The unstructured state facilitates increased interaction capacity, protein interaction networks, enhanced association rates, and adaptability to diverse partners. These disordered regions or proteins are often generated by the variable length of internal repeat regions and can potentially give rise to novel functions or protein variants. Repeat regions play a crucial role in protein-protein interaction networks, especially in hub proteins (highly connected proteins), by providing flexibility, adaptability, and an extended area for interaction. More repeats in an amino acid sequence facilitate more scope for functional diversification and the generation of new interaction networks in a short evolutionary time. Variable repeat regions provide genetic variation and an opportunity for selection to work on it. The occurrence of positively selected sites in repeat regions suggests that repeats are malleable regions that provide functional diversity and possess functionally important sites for interaction with other proteins. Overlap of only around 10 percent of lineages under positive selection and having repeat regions indicates that repeat regions help provide only length variability and may not lead to positive selection.

Longer genes may tend to have longer repeats and, hence, have more nucleotide length variation in orthologous repeats. This factor can influence the result of significant repeat length variation and possibly bias the overall results towards longer genes with repeats. We noticed that longer repeats and longer genes have a positive but weak correlation and that most longer orthologous repeats have similar lengths (e.g., RNF213, LRP1, KMT2A, LYST, and PKHD1L1). We observed the most length variation of orthologous repeats in PRG4 and HRG genes, possibly resulting from annotation/assembly artifacts.

Histidine-rich glycoprotein, HRG, is a plasma glycoprotein that can regulate immune complex clearance, complement activation, phagocytosis of apoptotic cells, cell adhesion and migration, and modulate angiogenesis. Human HRG contains two cystatin-like regions at the N-terminus, a histidine-rich region (HRR), a proline-rich region (PRR), and a C-terminal domain. In conditions similar to tissue injury or...
tumor growth, like low pH or interaction with Zn$^{2+}$, the histidine-rich region of the molecule enhances ligand binding$^{83,85}$. We noticed variation in the composition of the repeat region of HRG among different clades (HP and GHP amino acid composition) with varying lengths (with the most extended repeat length observed in Aves). Variation in size and composition of HRG repeat across species reflects different evolutionary trajectories under varying selective constraints on the biological process of the gene.

As discussed earlier, variation in repeat length can give rise to functional and evolutionary novelty in a short time. Repeats undergo accelerated expansion once length variation is above a critical length, while minor repeat length variation is proposed to be functionally neutral$^{22}$. Purifying selection constraints the repeat length in a specific range above which the length variation is deleterious$^{12,22}$. We wanted to quantify the abundance of contraction or expansion of the number of repeat-containing immune genes in a species-specific manner within each clade. A bias towards either expansion or contraction will indicate a change in an overall gene-interaction network due to repeat length variation. We observed Hominoidea to have more genes under the contraction of repeats. Apes (Hominoidea), compared to Asian and African monkeys, exhibit increased sensitivity to immune stimulation by viruses and bacteria$^{86}$. Apes also activate a broader array of defense molecules potentially beneficial for early pathogen killing$^{86}$. Widespread repeat contraction of immune genes in apes might have contributed to the different immune responses between apes and monkeys. Similarly, in the Amphibia clade, Gymnophiona shows more genes under significant contraction than Anura clade. The number of species included in these two clades of Amphibia is limited and needs more data to test the consistency of the result.

In the case of strong purifying selection, the orthologous repeats will tend to maintain the same composition and length across distinct species$^{22,87}$. Of the genes with the same orthologous repeats across clades, we randomly selected 11 genes to look for the standard deviation of repeat length. RUNX2, KAT6A, PRRC2C, and HAND2 show length variation between species, of which RUNX2 is a well-known example of morphological variation in vertebrates due to repeat length variation$^{28,30}$. Notably, KAT6A also showed repeat length variation between species across clades. KAT6A helps regulate transcription and gene expression by acetylating lysine-9 residue of histone 3. KAT6A is also known to participate with RUNX2 in osteogenesis$^{88,89}$. Correlation and association of repeat length variation of RUNX2 and KAT6A can be a potential target for future studies of morphological variations.

Clade-specific expansion and contraction of a repeat in protein sequence can reflect the overall functional diversification of that protein's interaction network or structure. In the case of Rodentia clade, we observed expansion of proline-rich repeat of FASLG in Cricetidae compared to Muridae. FASLG is a transmembrane protein whose binding with its receptor induces apoptosis$^{90,91}$. FASLG also participates in neural degeneration/regeneration$^{92}$ and T cell activation$^{93}$. The cytoplasmic side of FASLG contains a distinct proline-rich region consisting of several SH3 binding domains flanked by arginine and lysine residues and two conserved casein kinase I (CKI) binding sites$^{93,94}$. The proline-rich repeat is responsible for sorting Fas ligand (FasL) to secretory lysosomes$^{94}$, influencing T cell activation by reverse signaling$^{93}$. Deleting the proline-rich domain in cells with secretory lysosomes leads to surface expression of FasL$^{94}$. 
Costimulation and signaling through a T-cell receptor are necessary for optimal T-cell activation. Deleting 45–54 amino acids of the repeat leads to failure of costimulation\textsuperscript{93}. The contraction/expansion of proline-rich repeat in two clades of Rodentia can potentially indicate an alteration in the interacting network associated with the repeat functioning.

Our in-silico analysis identifies several promising candidate genes whose repeat length has changed drastically between species or clades, rapidly generating evolutionary novelty. However, the practical implications of these changes are unclear. Hence, the immediate next step would be to investigate the consequences of these changes on the structure and function of these proteins. Changes in the repeat regions of genes can change interaction partners, structural conformation, the prevalence of post-translational modifications, protein stability, etc. Moreover, genes that contain more than one repeat have multiple substrates for rapid change in the protein sequence. Protein repeats in living organisms have survived selective constraints and persist due to beneficial or neutral functional consequences. Based on these surviving repeats, we identify malleable regions of immune genes. The availability of high-throughput genome editing technologies provides a new avenue to synthesize genes with protein repeats of various lengths and compositions\textsuperscript{95,96}. Such synthetic biology approaches will be able to provide a comprehensive picture or fitness landscape of which regions of what genes can tolerate repeats and how such repeats affect their functions. Such synthetic biology approaches will be able to provide a comprehensive picture or fitness landscape of which regions of what genes can tolerate repeats and how such repeats affect their functions.

This study focused on clade-specific orthologous repeat length variation in a phylogenetically corrected framework in immune genes. We characterized and quantified the different types of repeats and their overlap with positively selected lineages and sites. Moreover, we enhance previously reported morphological/physiological consequences of repeat length variation across vertebrates and provide new candidates for future studies. Our study of the repeat region in immune genes highlights the importance of the repeat-rich domain in immunological processes. It suggests repeats as a potent source of the immense diversity of immune functions with a limited number of genes.

**Limitations**

A conservative list of ~1500 genes are involved in immune response\textsuperscript{97}. After including genes with other immune-related functions, a more comprehensive gene set can be considered immune genes. However, our current understanding of immune-related genes is far from complete and continues to improve with the availability of high-throughput assays\textsuperscript{98,99}. Moreover, our analysis strategy requires removing species-specific genes and those with poor annotation and genome assembly challenges. Hence, our list of immune genes is not exhaustive and may have missed several genes with protein repeats that are difficult to assemble.

Similarly, the repeat sequence may be incorrect due to assembly errors and can confound the identification of repeat expansion and contraction. We consider 12 vertebrate clades for our analysis (see
Supplementary Table S2) to have an extensive phylogenetic sampling. However, five of these clades (Afrotheria, Amphibia, Marsupials, Perissodactyla, and Testudines) have less than ten species with genome assemblies. Although the Aves clade has 56 species with genomes, the quality of these genomes is highly heterogeneous, and many genes are missing from the genome assembly. Hence, sparse sampling and missing data will likely limit our understanding of protein repeat evolution. Accurate, high-quality genome assemblies and availability of population-level genomic samples will provide a more detailed picture of protein repeat evolution. We use a phylogenetic correction approach that cannot distinguish between empirically significant and biologically relevant repeat changes. For instance, we do not consider within-gene position-specific heterogeneity, amino-acid type, sequence saturation, repeat purity, etc., while shortlisting candidate genes. Overall, our criteria are stringent and may result in the exclusion of essential genes.

Declarations

Acknowledgment

We thank the Ministry of Human Resource Development for fellowship to LT and SS. Computational analyses were done on the Har Gobind Khorana Computational Biology cluster established and maintained by combining funds from IISER Bhopal under Grant # INST/BIO/2017/019, IYBA 2018 from the Department of Biotechnology (Grant no. BT/11/IYBA/2018/03), and ECRA from Science and Engineering Research Board (Grant no. ECR/2017/001430).

Competing interest statement

None to declare

Availability of data

All data associated with this study are available in the Supplementary Materials.

Contributions

LT and SS wrote the manuscript with NV. LT, and SS analyzed the data and generated the results. All authors reviewed the manuscript.

References


Figures

Figure 1
A GO enrichment of the immune system process genes for molecular functions. B Venn diagram showing the overlap of the number of genes with repeats to immune genes. Adaptive immune response genes (GO:0002250, 722 genes), innate immune response genes (GO:0045087, 922 genes), and immune system process genes with repeats (823 genes) are represented with dodgerblue, green4, and orchid-colored circles, respectively. There are 47 genes common in the immune system process, adaptive immune response, and innate immune response genes.
Figure 2

A The 2D histogram of GC% and the frequency of occurrence of the genes with repeats at their relative positions along the gene. The figure shows more abundance of repeats within the GC% range at the beginning and end of the gene. B The boxplot shows significantly higher GC% in genes having repeats than genes not having repeats across all the clades. The $\bar{x}$ represents the mean, and the $n$ is the total number of sequences used for comparison in that clade.
**Figure 3**

A. The amino acid sequence of species containing multiple repeats. The colored boxes represent the amino acid repeat at their position in the gene sequence with their length (in nucleotides) at the top. 

B. The repeat length distribution for *SETD1A* gene along the phylogeny of Primates clade. The red box
represents the group of species with significantly longer repeat length, and the blue box is around the group of species showing significantly contracted repeat length. C Venn diagram represents the number of positively selected lineages and the lineages having repeats. D The structure of the DHX9 protein of \textit{Homo sapiens} contains two repeats at two distinct positions. E The SETD1A protein sequence contains five different amino acid repeats in \textit{Homo sapiens}. The structures are downloaded from AlphaFold Protein Structure Database and visualized in ChimeraX. The repeat stretches are colored and enclosed with blue lines for better visualization.
Figure 4

A The scatter-plot of repeat length with gene length. Some of the genes show a high degree of variable orthologous repeat length. The dark blue dots are for outliers. PRG4 gene shows a varying length of the orthologous repeat (shown in sky blue color) across the species. B The figure shows repeat length distribution for the HRG gene along the phylogeny across different species for all clades. C The figure
shows varying lengths of the repeat for the \textit{PRG4} gene across different species for all clades. The phylogenetic trees are downloaded from TimeTree website.

\textbf{Figure 5}
A The pie chart represents the frequency of expanded and contracted genes in the Primates clade. Red color represents the proportion of genes with expanded repeat lengths, and blue is the proportion of genes with contraction of repeat length. The brackets represent the number of significantly expanded and contracted genes for each species. B The heatmap of the standard deviation of the orthologous repeat lengths in genes across different clades. White-colored boxes represent the absence of the gene for that clade, the green color represents that all repeat lengths are the same in that clade, i.e., the standard deviation is zero, and the red represents a high standard deviation.

Figure 6

A Repeat length distribution of the FASLG gene across different species in the Rodentia clade. The phylogenetic tree is 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 significantly longer repeat length, and the blue box represents significantly shorter repeat length. B We downloaded the secondary structure of the FASLG gene from the Protter website for Cricetulus griseus and Rattus norvegicus. C The protein structure of the FASLG gene is visualized in ChimeraX. The top protein structure is of the FASLG gene of Cricetulus griseus modeled by SWISS-MODEL. The bottom protein structure is the FASLG gene in Rattus norvegicus downloaded from AlphaFold Protein Structure Database.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf
- SupplementaryTables.xlsx
- Table1.xlsx