Yusra Nazneen
Department of Biology,
McGill University, Montreal

Finding evolutionary innovations of Protein Domains across Animals: focus on intrinsically disordered, prion-like and other compositionally unusual Domains

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

Table of Contents

0.1 Structured Abstract

Objective

The objective of this study is to observe novel innovations of intrinsically disordered regions (IDR), prion-like domains (PLD), and compositionally biased domains (CBD) in animals, and to analyse specific case studies of circadian proteins and disease-linked intrinsically disordered proteins (IDP) protein subsets.

Methods

A systematic analysis of proteins was conducted using 58 proteomes dispersed within different clades in the animal kingdom to observe IDRs, PLDs, and CBD annotations in proteins within their orthologue group at each clade level, also known as an orthogroup, going up the kingdom's tree. The presence and significant alteration of these domains in the hierarchal path of their orthogroup were noted and a novel annotation was observed if the annotations in the different orthogroup clades within the hierarchal path were sufficiently different. These differences were categorized by the novel presence of the annotation, an increased observance of the number of tracts of the domain, a drastic change in length of the domain, or a gain in a CBD signature within the orthogroup clade. These novel annotations were then summed by clade or species to get a total count of novel annotations at all levels. Further, they were analysed by their gene-ontology (GO) terms for possible functional inferences and specific case-studies were analysed in more depth. These tasks were mostly accomplished using a combination of bash/awk, R, and python scripts.

Results

A total of 1 828 orthogroups with novel annotations for IDRs or PLDs were determined with *Mollusca* harboring the greatest proportion of the novel IDR-long annotations, most novel annotations were due to a a novel appearance or significant alteration of an IDR. Combining the novel annotations observed for both IDR-long and IDR-short, 571 orthogroups with novel annotation by novel presence were identified, 572 orthogroups

containing novel annotations by increased count of domain tracts, 435 by novel length, and 250 orthogroups with a novel annotation for a novel appearance of a PLD. For novel CBD single-amino acid (AA) signatures accompanied by low-complexity domain (LCR) annotation, a staggering 35 569 orthogroups with novel annotations were identified, the majority within *Mammalia*. Of these 187 overlapped with protein orthogroups containing either a novel IDR or a novel PLD annotation as well. Summarizing all novel annotations, the most frequent GO biological terms were related to regulation of gene transcription often through positively or negatively regulating transcription by RNA polymerase II; the highest GO molecular terms were for identical protein binding, followed by metal ion, ATP, and then RNA and DNA binding. From analysing the circadian protein case-study, Cry1 and HOX9 were found to have gained a PLD domain from mammals onwards, additionally, HOX9 gained a short IDR and Cry1 a long IDR also within mammals onwards. Within the disease-linked protein subset, Gsk3a was observed to gain a novel PLD annotation whereas p53 lost a PLD within mammals onwards. Also, of note within hominids BCL2 gained a short IDR. In both data subsets numerous single and multiple-AA novel CBDs and LCRs were identified.

Conclusion

The study indicates novel innovations of IDR, PLD, and CBDs in the hierarchal path of animals. For novel IDR and PLD annotations *Mollusca* was the most common clade and for novel CBD *Mammalia* was the highest-ranking clade. The most common GO terms related with the novel annotations were linked to transcription or translation initiation and predominantly molecularly involved with DNA, RNA, ATP, or protein binding. These findings have important implications for understanding the evolution of these protein domains and may inform future research on the role of these domains in disease pathophysiology and other biological and cellular processes.

0.2 Résumé de la Thèse

Objectif

L'objectif de cette étude est d'observer de nouvelles innovations de régions intrinsèquement désordonnées (RDI), de domaines de type prions (PLD) et de domaines de composition biaisée (CBD) chez les animaux, et d'analyser des études de cas spécifiques: protéines circadiennes et protéines liées à la maladie.

Méthodes

Une analyse systématique a été menée en utilisant 58 protéomes dispersés dans différents clades du règne animal pour observer les IDR, les PLD et les annotations CBD dans les protéines de leur groupe orthologue à chaque niveau de clade, connu sous le nom d'orthogroup, remontant l'arbre kingdom's. L'altération significative de ces domaines dans le chemin hiérarchique de leur orthogroupe a été notée et une nouvelle annotation a été observée si les annotations dans différents clades d'orthogroupe dans le chemin hiérarchique étaient suffisamment différentes. Ces différences ont été classées selon la présence nouvelle, le nombre de secteurs de domaine, le changement de longueur du domaine ou le gain dans une signature CBD. Ces nouvelles annotations ont ensuite été additionnées par clade ou espèce pour obtenir un nombre total de nouvelles annotations à tous les niveaux. De plus, ils ont été analysés par leurs termes d'ontologie génique (GO) pour d'éventuelles inférences fonctionnelles et des études de cas spécifiques ont été analysées plus en profondeur. Ces tâches ont été accomplies en utilisant une combinaison de scripts bash/awk, R et python.

Résultats

Un total de 1 828 orthogroupes avec de nouvelles annotations pour les IDR ou les PLD ont été déterminés avec Mollusca hébergeant les plus grandes annotations IDR-long; la plupart des annotations nouvelles étaient dues à un nouvel IDR. En combinant les nouvelles annotations observées pour IDR-long et IDR-court, 571 orthogroupes par nouvelle présence ont été identifiés, 572 par comptage accru, 435 par nouvelle longueur et 250 à partir d'un nouveau PLD. Pour de nouvelles signatures de CBD mono-

aminoacides (AA) accompagnées d'annotations de domaine de faible complexité (LCR), 35 569 orthogroupes avec de nouvelles annotations ont été identifiés, la majorité chez Mammalia. De ces 187 chevauchements avec des orthogroupes contenant soit un nouvel IDR ou une nouvelle annotation PLD. Résumant toutes les annotations nouvelles, les termes biologiques GO les plus fréquents étaient pour la transcription des gènes souvent par la régulation de la transcription par l'ARN polymérase II; les termes moléculaires GO les plus élevés étaient pour la protéine identique, l'ion métallique, ATP, puis ARN et ADN se liant. De l'étude de cas de protéine circadienne, Cry1 et HOX9 se sont avérés pour avoir gagné un domaine de PLD des mammifères en avant, en plus, HOX9 a obtenu un IDR court et Cry1 un IDR long également chez les mammifères et au-delà. Dans le sous-ensemble de protéines liées à la maladie, Gsk3a a obtenu une nouvelle annotation PLD alors que p53 a perdu un PLD chez les mammifères; aussi, chez les hominidés, BCL2 a obtenu un IDR court. Dans les deux sous-ensembles, de nombreux CBD et LCR uniques et multiples ont été identifiés.

Conclusion

L'étude indique de nouvelles innovations d'IDR, de PLD et de CBD dans le chemin hiérarchique des animaux. Pour les nouvelles annotations IDR et PLD, Mollusca était le clade le plus courant et pour le nouveau CBD Mammalia. Les termes GO les plus courants liés aux nouvelles annotations étaient pour l'initiation de la transcription ou de la traduction et moléculairement impliqués dans l'ADN, l'ARN, l'ATP ou la liaison aux protéines. Ces résultats ont des implications importantes pour comprendre l'évolution de ces domaines protéiques et pourraient éclairer les recherches futures sur le rôle de ces domaines dans la physiopathologie des maladies et les processus cellulaires.

0.3 Acknowledgments

This work would not have been possible without several people supporting me, reassuring me, and guiding me through the learning processes required to perform this research. Most notably, I'd like to thank my kind supervisor Dr. Paul Harrison to whom I am enormously grateful for being very supportive and patient and constantly providing constructive criticism and introducing me to many computer tools to assist in coding. I am grateful also to my ever-supportive family, my lovely parents and cheerfully humorous brothers and my dear husband who are the reason I am very blessed to research with minimal worry. I am also very thankful to my wonderful supervisory committee comprising of Dr. Brandon Xia and Dr. Melania Cristescu whose insight has directed me to clearer thought and direction and whose encouragement has been very lifting at each meeting. I would also like to thank Dr. Traian Sulea for giving me the opportunity to intern at the NRC Molecular Modeling Team and Dr. Christopher Corbeil and Dr. Francis Gaudreault who supervised and mentored me patiently to code in python and familiarize with SQL in the early stages, which was invaluable. And lastly, I am very thankful to the McGill Biology Department for making all bureaucratic necessities effortless, particularly I am thankful to Ancil Gittins.

0.4 Contribution of Authors

Technical advice was given throughout this undertaking from Dr. Paul Harrison. The supervisory committee, Dr. Brandon Xia and Dr. Melania Cristescu also provided valuable feedback at every major step along the way. All chapters were written and edited by me with editorial help from Dr. Paul Harrison.

0.5 List of tables and figures

List of Tables

List of Figures

0.6 List of abbreviations

AA: amino acid

BT: biological terms

DOT: disorder to order transition

CBD: compositionally biased domain

FB: folding-upon binding

GO: gene ontology

HOG: hierarchal orthogroup

IDR: intrinsically disordered regions

IDP: intrinsically disordered proteins

LCR: low complexity region

LLPS: liquid-liquid phase separation

MAP: microtubule associated protein

MSA: multiple sequence alignement

MT : molecular terms

NF : novel factor

OG : orthogroup

PLDs: prion-like domains

PPI : protein-protein interation

PTM: post-transnational modification

RBP : RNA-binding protein

RNAP : RNA polymerase

0.7 Introduction

With the exciting advent of fast and cheap genome sequencing an unprecedented storage of sequence information is now readily available online, most notably in the UniProt and NCBI databases. This great triumph is limited however by the growing demand for bioinformaticians to analyze and annotate the colossal library of sequence data and organize it to discern meaningful patterns and connections to forge new insights. Over the last few decades several domains and annotations have been discovered in proteins, often widely conserved, with conservation conventionally directly relating with function. However, observations have been made of some domains that emerge independently from their ancestral origin or disappear seemingly randomly within certain clades or species. This study proposes to identify these domains; specifically, targeting intrinsically disordered regions (IDR), prion-like domains (PLD), and compositionally biased domains (CBD). These domains are often directly responsible for specific changes in protein conformation and thereby protein function and their importance is underscored in appreciating their function in regulating nearly all cellular processes. More than 70% of known signaling proteins possess disordered regions, if altered these domains are often linked to a wide breadth of pathologies (Uversky et al., 2015). Since different protein families utilize these domains for different functions, identification of significant alterations in these domains over evolutionary history contributes significantly in better understanding and predicting differences in protein behavior and function between protein orthologs.

Eukaryotes harbor the largest proportion of intrinsically disordered domains; interestingly, increased number of disordered domains has been linked with increasing organism complexity (Gao et al., 2021). Intrinsically disordered proteins (IDP) lack a stable conformation, they are primarily composed of IDRs; herein the protein structure can readily transition between a myriad of possible transient conformations depending on the environment and binding partners. This conformational plasticity and increased flexibility allow IDPs to be promiscuous having multiple binding partners and function in almost any regulatory capacity: as transcription factors, initiation factors and coactivators, conductors for post-translational

modifications (PTM), hubs for protein-protein interaction (PPI) networks, scaffolds hosting giant protein complexes, in addition to other roles involving multiple binding partners (Wright and Dyson, 2015).

Notably, IDRs are often characterized as compositionally biased for single or multiple amino-acid (AA) residue biases having CBDs or low-complexity regions (LCR). These LCR tandem repeats have been linked to specific functions and are common in PLDs (Alberti et al., 2009). Asn-Gln LCRs are characteristic of PLDs, coined 'prion-like' for their similar composition to yeast prion proteins. They are only 'prion-like' however since albeit exhibiting the prion protein property of forming aggregates they lack their ability to be 'infectious', incapable of changing the conformation of other proteins to mimic their aggregation-prone conformation (Cascarina et al., 2014). Interestingly, aggregation of proteins with these domains is a common hallmark of neurodegenerative disease including Amyotrophic lateral sclerosis (ALS), Alzheimer's, Parkinson's, and frontotemporal dementia (King et al., 2012). Similarly, LCRs often are encompassed within IDRs or PLDs and certain LCRs are linked with stress granule formation and

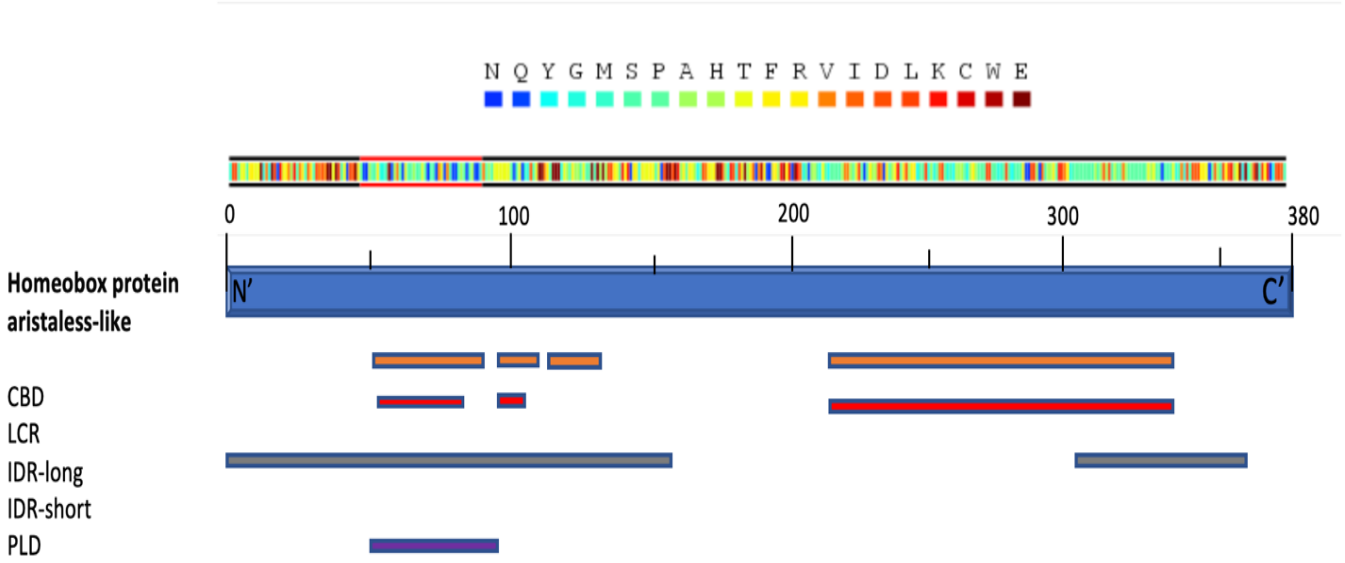

Figure 1.0: Example protein for showcasing overlap of annotations of interest. Homeobox protein aristaless-like protein was selected randomly from proteins harboring PLD and less than 500 amino acids. UniProt accession id: A0A158NE44. The annotations shown are derived from the selected annotation softwares used in this study. Color code visualization of sequence provided by PLAAC webserver. CBD: provided by fLPS; LCR: provided by fLPS; IDR-long (30 or more amino acid in length): provided by confirmation by IUPred2a and DisoPred; IDR-short (20-29 amino acid in length): provided by confirmation by IUPred2a and DisoPred; PLD: provided by PLAAC.

liquid-liquid phase separation (LLPS), a process by which LCRs can separate out of the cytoplasmic solution into liquid droplets (Banani et al., 2017). However, aberrant phase separation and subsequent aggregation of these granules under chronic stress is directly involved in the pathophysiology of a wide horde of diseases including neurodegenerative as well as cancer and cardiovascular diseases (Mo et al, 2022; Wang et al, 2021). Figure 1.0 shows an example of the relation and overlap between the mentioned domains of interest.

For this study, protein sequences from selected proteomes were grouped by multiple sequence alignments (MSA) to identify and group together orthologs, proteins in different species derived from the same protein sequence and thereby sharing a template with a common ancestor. For instance, the hemoglobin protein in mice and the hemoglobin protein in humans are orthologs of one another. By tracking the descent of orthologues down the animal tree, novel innovations can be observed by relation to the hierarchal orthogroups (HOG) of the annotation in different clades, with the term orthogroups defined as a group of proteins within one clade which are all orthologues to each other at that clade-level. By noting the disappearance and appearance of these annotated domains across the phylogenetic tree patterns of emergence or disappearance of these domains can be tracked, please refer to Figure 1.1 for an example. Since CBD, PLD, and IDR have significant functional relevance, the tracking of these domains is valuable for gaining insight into the changes in protein function and behavior across different clades and species.

For this study 58 proteomes from the animal kingdom are analyzed first globally and then more case-specifically from smaller data subsets of circadian proteins and common disease linked proteins. Circadian proteins are involved in multitudinous regulatory processes and harbor extensive IDRs harboring several PTM sites wherein phosphorylation is a common means for circadian rhythm regulation. As for disease-linked proteins, this larger protein subset is largely comprised of aggregation-prone proteins with LCRs found to be essential for LLPS in physiological conditions requiring stringent controls for regulatory processes within the cell, a process if perturbed leads to various pathologies. A closer inspection on the novel annotations

for these subsets can bring interesting observations that may reflect on the different means of circadian regulation, mechanisms for maintaining protein stability under stress, and different propensities for LLPS across different animals.

Herein novel annotations have been described in four different ways: presence or absence, count, length, and CBD signature. For the first, presence or absence, novel annotations are marked for the sudden appearance of a domain, either IDR, PLD, or CBD, from orthologs of one clade in relation to the orthologs of the same protein in a hierarchal clade or hierarchal orthogroup (HOG). Regarding novel CBDs, each different type of bias is denoted as a 'signature', importantly, this tracking is significant since different signatures have been observed to be associated with different protein functions (Wright and Dyson, 2015). Novel annotations by count or length are considered for IDRs and are observed respectively if the average number of counts or the average length of the annotation in the orthogroup is markedly different, more than three tracks or 100 AA residues in length or greater respectively, from the average number of counts or the average length of the annotation in its HOG.

In this research endeavor I identify novel IDR, PLD, and CBDs annotations within Animalia and subsequently analyze these novel emerging domains for functional relevance either globally or with a focus on certain protein subgroups: circadian and disease-linked proteins.

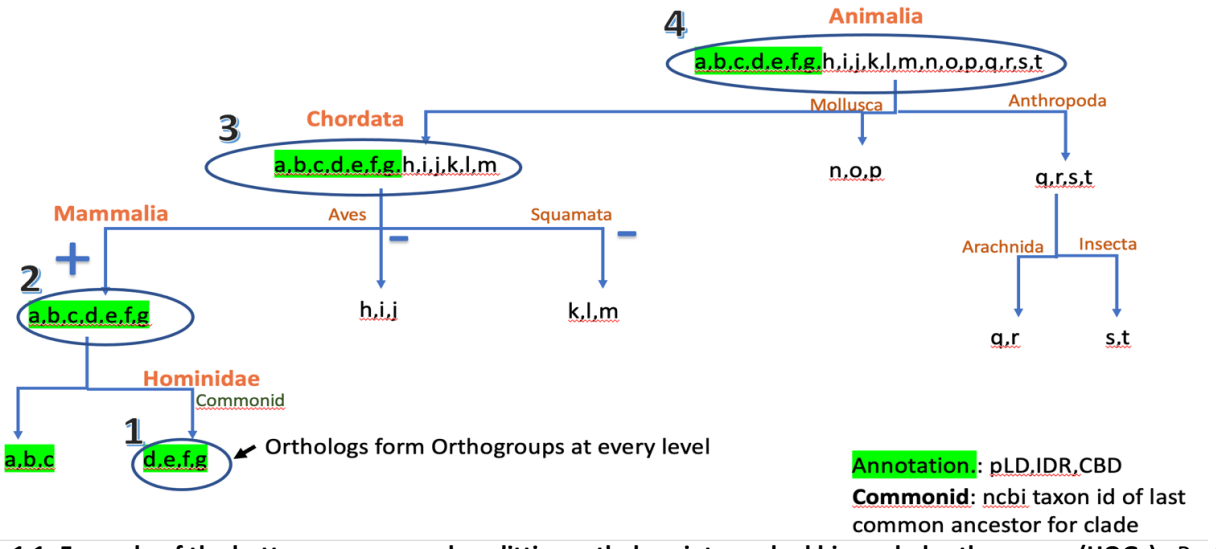

Figure 1.1: Example of the bottom-up approach: splitting orthologs into ranked hierarchal orthogroups (HOGs). Protein orthologues at shown clades are symbolized by letter characters and ranked from bottom to top from 1-4 as the four HOGs shown in the example, with a novel gain in annotation within the 2nd HOG Mammalia in this example.

0.8 Literature Review

0.8.1 Introduction

Proteins are functional biomolecules essential for life, and their structural and functional complexity is due in part to the presence of discrete domain blocks. These domain blocks contain specific regional structures and functions and when combined in different combinations they create unique proteins with specialized functions. Understanding the evolution of these domains is critical for gaining insights into the diversity of proteins and the intricate mechanisms of regulation in different species, in addition to potentially gaining insight into the onset and exacerbation of various pathologies. In this thesis, I explore the novel occurrences of three protein domains - IDR, PLD, and CBD - within the animal kingdom and how they may emerge or have been altered between clades. The animal kingdom was chosen for its direct relevance to the evolution of complex regulatory mechanisms and due to its complex social behavior observances, which may be linked to novel domains. Moreover, the prevalence of IDRs in eukaryotes and the large extent to which environmental and lifestyle inputs affect regulatory processes, especially in mammals, make this topic particularly interesting.

The findings of this study hold considerable value and contribute to our understanding of the diversity of life on Earth. The bioinformatics research conducted on protein domains involved in regulatory processes has potential applications in various fields, including medicine, biotechnology, and conservation efforts. The large number of uncharacterized proteins in the online growing databases highlights the significance of gaining insights into the types of protein domains they encompass, since it provides clues towards their functions. Certain protein domains may play a critical role in regulating an animal's physiology and homeostasis, and mutations within these domains may be important for specific species. Knowledge of these domains may aid in developing therapeutics and identifying new therapeutic targets, genetic screening risk flags, and understanding differences in species physiology. However, it is essential to note that this study has certain limitations. 58 selected proteomes were selected screened for high quality by meeting specific requirements for selection, but due to the unequal representation of proteomes across certain clades available in the UniProt database,

some clades were slightly overrepresented while others underrepresented. While efforts were made to select a diverse and inclusive repertoire of proteomes by selecting at least two to three species with high-quality proteomes for all clades, it is possible there is some bias due to this uneven distribution. Additionally, limitations may exist from false positives or negatives from annotation software, which will be discussed in later in this review. Despite these limitations, the findings of this study provide intriguing insights into the evolution of protein domains involved in regulatory processes and their potential applications in various fields.

The objective of this literature review is to provide a comprehensive and cohesive overview of the topics relevant to the research question and their interrelatedness. The review begins by examining the plausible mechanisms underlying protein domain evolution within the animal kingdom. Subsequent sections focus on each domain of interest individually, elucidating their unique features, functions, involvement in various cellular processes and mechanisms, their distribution and evolutionary history within animals, and their roles in diverse pathologies. Due to the high degree of overlap between these domains, IDRs are given the most attention as they encapsulate features of LCRs and PLDs to efficiently execute their central role in regulation. Thereafter, the review discusses the rationale for selecting specific annotation software and potential limitations caused therefrom, and following, the computational methods used for the analysis and then concluding with a succinct review.

0.8.2 Evolutionary mechanisms that drive the emergence of new protein domains

Protein domains can appear, disappear, or be altered through various mechanisms that drive protein evolution in animals; foremost, these mechanisms entail gene duplication, exon shuffling, horizontal gene transfer (HGT), and alternative splicing. The extent to which these mechanisms are adopted in different clades within the animal kingdom vary, herein the mentioned mechanisms will be discussed and the differential tendencies for the different branches of life in the animal kingdom to employ them. The combination and frequency of employment of these mechanisms can drive the evolution of multi-domain proteins with more intricate functions or even the generation of new proteins.

Gene duplication is a common occurrence within eukaryotes resulting in duplicated genes referred to as paralogs. Akin to orthologs, paralogs share a common template, however, these arise from duplication events within the same species. There are various types of duplications that can occur resulting in chromosomal rearrangements or exon shuffling. Common mechanisms include whole genome duplication which is most prevalent in plant species, while tandem and segmental duplications are common in animals. Tandem duplications can occur from errors in DNA replication or non-allelic homologous gene recombination during crossing-over in meiosis or from aberrant DNA repair processes. Additionally, duplication of segments of the genome through transposable elements is a common occurrence in animals via retroposition. In this mechanism, a copy of the original gene is created by duplication, coined a "retrocopy," which is fully equipped with the necessary tools to insert back into the genome referred to as "jumping" to a new region not neighboring the parent where it is free to accrue mutations while not compromising the fitness of the original gene. Interestingly, thousands of retrocopies have been identified in the human genome, moreover, segmental duplications encompassing LCRs account for roughly 13.7% of the human genome (Lallemand et al., 2020).

As mentioned, different domains of life seem to have differential tendencies for the proportion to which these mechanisms are employed for protein domain evolution, animals seemingly favor gene duplication and domain accretion mechanisms, with exon shuffling being less common (Zhang et al., 2020). Interestingly, a higher proportion of domain accretion events in deutereosomes including Chordates relative to protostomes including mollusks and arthropods was observed in a study by Zhang et al, possibly suggesting a greater gain in functional capacity within deutereosomes (Zhang et al., 2020). Additionally, splicing together introns containing host exon fragments is another mechanism other than the employment of transposable elements to achieve a combination of exons derived from different genes referred to as ‘exon shuffling’. Notably, exon shuffling has been thought to contribute to the emergence of various multi-domain proteins, including new immunoglobulin domains and zinc-finger domains (Inan et al., 2010).

Lesser common in animals is HGT, more commonly employed in bacteria and unicellular organisms. HGT is the transfer of genetic material from symbiotic or parasitic organisms like bacteria, viruses, fungi, or mobile genetic elements, thus resulting in the seemingly random emergence of genes which may not be in other genomes of the same clade or hierarchal clades; overtime these domains may be able to provide advantages in adaptive evolution and fitness of the organism by providing new functionable capabilities. The gain of these domains can offer several fitness advantages to the organism by conferring them with new adaptive capabilities: metabolism of new nutrients, improved response to stress, detoxification of environmental toxins, or better adaptation to new or changing ecological niches. For instance, a gain of a PLD was found in a rotifer species, bdelloid, linked with increased tolerance of desiccation (Boschettit, 2012).

It is worth noting protein domain evolution can also be driven by alternative splicing, one of several mechanisms utilized in fine-tuning genome transcription regulation. Alternative splicing leads to the production of multiple mRNA isoforms from a single gene, which subsequently gives rise to various protein isoforms, increasing the functional diversity of the protein. Incorporation of intronic or non-coding DNA, such as from transposable elements during splicing can lead to the exonization of these elements and the creation of new protein domains. This gain of new functional domains can significantly impact an animal's evolution by providing an adaptive advantage. Furthermore, alternative splicing may contribute to the evolution of new regulatory mechanisms since it is observed to be tissue and developmental stage dependent. Proteins involved in alternative splicing and their regulators may be inferred through IDRs and CBDs since proteins involved in alternative splicing are often regulated epigenetically through PTMs often located within IDRs (Singh et al., 2020).

0.8.3 Intrinsically Disordered Regions

0.8.3.1 Relevance of molecular and structural properties of IDRs to function

IDRs, as previously briefly described in the introduction, are domains within proteins lacking a stable three-dimensional structure under physiological conditions, as such they flexibly fluctuate their conformation in response to environmental cues and binding partners. They generally constitute of polar and charged amino acid (AA) residues which afford them high solubility and flexibility, often incorporating these residues in LCRs with repeated motifs, and presenting them on the exposed surface of the protein thereby being available to respond to various stimuli via binding to multiple proteins, DNA, RNA, and small molecules partners through forming electrostatic interactions. For their central role as master orchestrators for protein regulation and signaling, IDRs are well-suited, with an impressive repertoire of functional capabilities employing various mechanisms to achieve regulation of these highly interactive processes. They regulate nearly all cellular processes including protein regulation through instructing on protein localization, modification, stability, and degradation, as well as myriads of cellular processes foremost of which is their involvement with signal transduction, cytoskeletal organization, DNA damage, stress response and transcriptional and translational regulation (Wright & Dyson, 2015). All the while they also form structures capable of acting as molecular scaffolds, flexible linkers, molecular switches, entropic bristles, and entropic springs detailed shortly in the following paragraphs in this section (van der Lee et al., 2014).

Importantly, the efficient coordination of these processes requires the modulation of the activities of myriads of proteins, thus regulatory proteins require IDRs granting them the functional capability of binding with a large assortment of protein binding partners. The high conformational flexibility offered to them by encompassing IDRs coupled with their low structural stability under physiological conditions with only transient secondary structures allows them the structural plasticity to readily perform dynamic actions with a greater propensity to undergo conformational changes and therefore can achieve greater functional versality. Moreover, within these regions they are often endowed with short linear motifs

(SLiMs), protein domain recognition sequences generally less than ten AAs in length, which mediate specific protein-protein interactions (PPI). It has been observed from NMR spectroscopy studies that IDRs can undergo local compaction and transient collapse allowing them to form little pockets which facilitate small molecule recognition and binding (Uversky, 2013). Another level of regulatory control IDRs offer is through their LCRs which have the propensity to undergo LLPS by reversibly and transiently clustering together to form membranelles organelles within cells, usually from the onset of stress stimuli, which allows for protective compartmentalization of mRNA and RNA-binding proteins (RBPs) in the cell within these discrete units due to which they can avoid being targeted for destruction (Banini et al., 2017).

Although master regulators of the cell, IDRs are also regulated themselves by being host sites for PTMs. PTMs in IDRs such as phosphorylation, acetylation, or ubiquitination, are commonly observed mechanisms to regulate their conformational state and thus binding affinity to other molecules. In addition to affecting PPIs, PTMs within IDRs also influences the protein's stability and other functions. Transcription factors contain IDRs which bind to multiple DNA targets regulating their transcription, additionally, transcriptional co-activators also contain large IDRs to bind and regulate various transcription factors and thereby regulate gene transcription in response to the cellular microenvironment (Wright et al., 2015). An extensively studied example of this is found in the p53 transcription factor that contains a large disordered domain hosting several PTM sites for phosphorylation, phosphorylation at this site dictates the proteins' ability to bind to DNA and other molecules (Bullock et al., 2001). Furthermore, IDRs regulate gene expression by their presence at the N-terminal domain of histones where they modulate the open access of the gene for transcription factors to initiate gene expression via PTMs sited hosted by the IDR. RNA polymerase II similarly encompasses several PTMs within a disordered domain that regulate its functional capacity for binding to other partners and activating transcription (Hsin et al., 2012).

IDRs can regulate using PTMs as result of their flexible nature which allows them to act as molecular 'switches' by transitioning between different conformational states in response to PTMs, binding to different ligands, PPIs, or in response to environmental changes such as pH, temperature, or ionic strength. This 'disorder-to-order' transition is characteristic of molecular recognition features (MoRF), short stretches of IDRs which recognize specific molecules and undergo disorder-to-order transition upon binding, a critical process for functional activation. Interestingly, their regulatory capacity also extends to the protein they inhabit, having the capacity to modulate other domains in the protein they reside in. Moreover, as master orchestrators of signal transduction, IDRs act as intermediates linking a wide network of upstream and downstream signaling molecules, these 'adaptor' proteins mediate the interaction between various protein partners, oftentimes receptors, and are involved in cell proliferation, differentiation, apoptosis, and cellular immune response among other essential cellular processes (Wright & Dyson, 2015).

Despite not having a stable conformation, IDRs can assist structured proteins in gaining stability and achieving their native folding by acting as chaperones or as scaffolds to facilitate PPIs through disorder-to-order transition of IDRs in the proteins and thereby allowing the formation of stable structures in complex. Moreover, the flexible nature of their polypeptides allows them to act as entropic bristles, tethering proteins and distancing them from each other by providing repulsive forces and consequently preventing proteins from coming close to one another and aggregating in events of misfolding (Uverky and Dunker, 2010). This "macromolecular crowding" allows for the reduction of protein surface area exposure to denaturing agents and proteases. They also prevent protein aggregation by acting as shields for hydrophobic patches of proteins preventing them from coming into contact and erroneously aggregating and via masking or exposing certain protein domains and thereby facilitating proper protein folding. IDRs further aid in protein folding by forming entropic springs, a mechanism by which they utilize their flexible backbones to provide an efficient means for storing and releasing energy for conformational sampling (Uverky and Dunker, 2010). Nevertheless, IDRs are also implicated in causing misfolding and protein aggregation, often when they are very long and amenable to

aggregation prone conformations, in such states they are implicated in various diseases as discussed later.

0.8.3.2 Prevalence, distribution, and evolution of IDRs across animals

Common to all domains of life, IDRs are widely prevalent and conserved across animals, with an estimated 30% of all eukaryotic proteins, and 40% of all animal proteins containing at least one IDR (van der Lee et al., 2014). Their wide prevalence and conservation suggest their essential role in animal biology, however, within the different phyla in the animal kingdom IDRs are observed at various frequencies, interestingly chordates were found to house a significantly higher number of IDRs relative to non-chordates; furthermore, vertebrates were observed to have higher frequency of IDRs compared to invertebrates (van der Lee et al., 2014). The correlational observance of their presence in larger and more complex animals may contribute in their more complex animal traits such as the development of a nervous system and may arise from an increased demand for more intricate degrees of regulation for the increased complexity of vertebrate genomes (van der Lee et al., 2014). However, regarding the proportion of the proteome comprised of IDRs they were found in higher proportion in intracellular parasites, insects, and nematodes, whereas they were lowest in chordates (Sickmeier et al., 2007). The distribution and prevalence may be influenced by protein size and function, large multi-domain proteins involved with regulatory functions require high flexibility and efficient conformational sampling capability, as such they have much greater prevalence of IDRs (Uversky, 2013). Another factor for variance is the length of the IDR, some phyla house IDRs which are markedly longer. Although, their presence varies between different animal clades, the greater variance is between different protein families. Naturally, their frequency is much higher in intrinsically disordered proteins (IDPs) involved in cell-cycle regulation, transcriptional regulation, intracellular signalling proteins, cell-adhesion molecules, cytokines, and those involved with cytoskeletal organization (Xie et al., 2007).

Worth noting is the rapid evolution of certain protein families enriched in IDRs such as transcription factors and immune system proteins which may be the result of selective

pressures requiring adaptation critical to surviving in changing environmental conditions or the need for novel immune defense mechanisms. For example, hypermutations and somatic recombinations in the IDR comprising the variable region of antibodies allows for their ability to generate a colossal library of antigen-binding sites providing defence against an inexhaustible number of threats in the universe. The evolution of multi-domain proteins through domain accretion also enables the existence of more complex and intricately regulated proteins that target different sets of genes in response to different signals, as observed in large transcription factors. Moreover, IDRs contribute to the specificity of kinases and phosphatases by acting as 'docking sites' and undergoing conformational changes upon binding allowing for more selective interactions and fine-tuning of cellular signaling pathways (Uversky, 2013).

0.8.3.3 Functional implications of IDRs in disease pathophysiology

Since IDRs are essential for regulating cellular processes, loss of function mutations causing loss or gain in a toxic function mutation opens the doors to a wide expanse of diseases in animals. Hence, they are involved in the pathophysiology of different cancers particularly breast, prostate and lunger cancers, multiple neurodegenerative diseases, cardiovascular diseases, and even infectious diseases. Mutations inside the IDRs of oncogenes can give rise to cancer, as observed in the p53 transcription factor (Iakoucheva et al., 2016). The pathogenesis of neurodegenerative diseases is implicated to drastically progress, if not be borne, from the aggregation of IDRs forming amyloids in devastating and debilitating diseases such as Alzheimer's, Parkinson's, Amyotrophic lateral sclerosis (ALS), and frontotemporal dementia (FTD) (Uversky, 2017; Harrison and Shortner, 2017). The link of IDRs to cardiovascular diseases is multi-faceted. An early step that can lead to developing atherosclerosis, the buildup of plaque inside the arteries, is the abnormal function of endothelial cells lining the blood vessel. Mutation in the IDR of several proteins can exacerbate heart disease, for instance endothelial cell proteins involved with regulating endothelial cell function in the blood vessels, activated inflammatory proteins, or proteins involved with lipid metabolism can disrupt endothelial cells function as well as cause chronic inflammation and dyslipidemia, all key events and risk factors leading to heart disease. IDRs are also involved with infectious diseases by aiding the replication

and assembly of viral proteins within hijacked cells (Cortese et al., 2020). Infectious diseases such as prion diseases, a classic example being Creutzfeldt-Jakob disease, directly relate their pathophysiology to IDRs (Uversky, 2017). By elucidating the various ways in which IDRs operate in disease pathophysiology novel therapeutic targets can be realized for drug development.

0.8.4 Prion-like Domains

0.8.4.1 Relevance of molecular and structural properties of PLDs to function

PLDs share structural and biochemical properties with prion proteins, most notably their ability to adopt a self-templating conformation from a soluble conformation to an insoluble β-sheet rich conformation that can self-propagate and thereby form aggregates (Toombs et al., 2010). PLDs contain both ordered and disordered regions often with LCR encompassing disordered regions in the middle and ordered flanking regions. Notably, intrinsic disorder is a key characteristic of PLDs and required for their essential role in regulation which they are equipped for due to their several advantageous properties: PLDs have high polar and charged AA residue sequence count, high conformational flexibility, disorder-to-order transition upon binding capability, capability for forming electrostatic interactions with multiple proteins and nucleic acids, and possession of MoRFs (Uversky, 2013). They play an essential role in the function of transcription initiation factors, transcription factors, RNA-binding proteins, membrane binding proteins, and chromatin-associated proteins (Wang et al., 2018). By adopting a range of conformations, the disordered region of PLDs can interact with various RNA sequences and recognize different RNA targets (Chen et al., 2019). For membrane proteins, PLDs are required for their localization and trafficking function by mediating different PPIs at the cell surface (Das et al., 2018). In addition, PLDs have been implicated in stress response, ribosome biogenesis, protein degradation, formation of protein complexes, RNA processing, RNA metabolism, and signal transduction (Uversky, 2013; Wang et al., 2018).

Furthermore, as previously mentioned PLDs overlap and share structural and molecular properties with LCRs, specifically the ability to undergo LLPS, allowing them to form dynamic and self-templating aggregates or granules. These LCRs within PLDs are often enriched in

glycine, asparagine, glutamine, and tyrosine AA residues (Kim et al., 2013). Phase behavior and aggregation propensity are both dependant on the length of PLDs. With the capabilities of self-assembly and binding with multiple protein partners they are critical for the normal function of RBPs, transcription factors, and scaffolding proteins which are equipped with PLDs to perform their functions (Harrison and Shortner, 2017). Within regulatory proteins such as transcription factors PLDs modulate their interaction with other proteins and nucleic acids, thereby being essential for modulating the regulatory capacity of regulatory proteins. Akin to IDRs, the activity of PLDs, for instance aggregation behavior, are regulated by PTM sites, commonly through phosphorylation and ubiquitination (Lie et al., 2020). Interestingly, PLDs are distinguished not only from IDRs by their high propensity to aggregate and form amyloid-like fibrils but also from consisting of repeating units or LCRs, these repeating units allow them to interact with multiple binding partners (Harrison and Shortner, 2017).

0.8.4.2 Prevalence, distribution, and evolution of PLDs across animals

PLDs are widespread across the animal kingdom and highly conserved as expected since they play essential roles in fundamental biological processes. Naturally, they are enriched in certain protein families, such as RBPs, transcription factors, and signalling proteins. They have been observed in a wide range of animal taxa, including cnidarians, flatworms, arthropods, mollusks, and chordates, however, are relatively scarce in invertebrates, such as insects and nematodes (Uversky, 2018; Neme & Tautz, 2016). In a paper analyzing 300 animal proteomes high conservation of PLDs was observed across animal phyla with a positive correlation between increased organismal complexity and prevalence of PLDs (Zhang et al., 2021). The prevalence of PLDs may be a result of gene duplication, sequence divergence, and selection pressure (Neme & Tautz, 2016). Moreover, PLD length varies significantly across animal phyla and protein families with longer PLDs were found in higher eukaryotes suggesting an involvement in more complex cellular processes (An & Harrison, 2016). For example, the longest PLDs are found in the human FUS protein, a crucial protein for RNA processing and transport (Sun et al., 2018).

Interestingly, PLDs are particularly abundant and highly conserved in neuronal proteins in the human brain with varying prevalence and distributions across different neuronal cell types and brain regions suggesting a crucial role in neuron function. For example, the TDP-43 protein contains a PLD and is involved in RNA processing and transport, it is found in the nucleus and cytoplasm of neuronal cells suggesting a role in regulating gene expression and consequently protein production in neurons (Chen et al.,2019). Of note, higher brain regions such as the cortex, hippocampus, and cerebellum contain higher levels of PLDs relative to other regions such as the basal ganglia (Zhang et al., 2021). Further, neuronal cell types such as motor neurons and Purkinge cells in the cerebellum have greater prevalence of PLDs relative to other neuronal cell types potentially suggesting they may play a greater role in the regulation of these cell types (Iguchi et al., 2013).

0.8.4.3 Functional implications of PLDs in disease pathophysiology

Since PLDs are essential for the proper function of myriads of regulatory proteins, thus, any deleterious alteration to a PLD can have far-reaching devastating consequences on the function and stability of the protein and consequently cause the dysregulation of fundamental cellular processes. PLDs have been linked to the pathogenesis of a wide berth of neurodegenerative diseases as abovementioned including Alzheimer's, Huntingtons, Parkinson's and ALS. In Alzheimer's and Huntington's respectively, the PLD in the amyloid-β peptide and huntingtin, has been demonstrated to be essential for its aggregation and toxicity, a hallmark of the disease's pathology (Lashuel et al., 2002). In Parkinson's, the PLD of TDP-43 is linked to its aberrant self-assembly and accumulation into toxic cytoplasmic inclusions (Iguchi et al, 2013; Chen et al., 2019). As for ALS, mutations in the PLD of the FUS protein is linked significantly to both FTD and ALS development and progression, furthermore, the PLD in α-synuclein is implicated in the formation of Lewy bodies, toxic protein aggregates exacerbating ALS progression (Harrison and Shortner, 2017). Further, PLDs have been directly implicated in disrupting synaptic plasticity and long-term memory formation; overexpression of RBP TIA-1 forms cytoplasmic aggregates or RNA granules within murine neurons disrupting long-term potentiation required for memory formation (Li et al., 2013).

In addition to their direct involvement in the pathophysiology of several neurodegenerative diseases and memory formation, PLDs are heavily implicated in cancer playing a role in oncogenic signaling, transcriptional regulation, apoptosis regulation and protein degradation leading to tumor progression and metastasis. Since LCRs within PLDs undergo phase separation, mutations in the PLD can cause aberrant protein interactions in LLPS exacerbating the progression of some cancers. Additionally, since PLDs are often integral to regulating gene transcription, they can upregulate the transcription of oncogenes, for example the PLD containing PCNP nuclear protein is linked with increased cell proliferation, migration, and invasion in gastric and ovarian cancer cells (Zhang et al., 2019). Moreover, PLDs are even involved with the pathophysiology of viral infections aiding in viral replication, assembly, and immune system evasion within the hijacked cell (Anastassopoulou et al., 2020).

0.8.5 Compositionally Biased Domains and Low Complexity Regions

0.8.5.1 Relevance of molecular and structural properties of CBDs to function

CBD domains are characterized by a high frequency of specific AA residues or physiochemical properties such as hydrophobicity or charge size. A specific subset of these domains, known as LCRs, are distinguished by a high frequency of repetitive AA residues like glycine, serine, and alanine, often with tandem repeat regions, and linked to LLPS capability. CBDs, particularly those containing LCRs, provide regulatory proteins with the ability to form protein-protein and protein-nucleic acid interactions. Moreover, they are crucial for proper protein folding and stability and determinant of aggregation propensity. An example of their involvement in regulation is though alternative splicing; the ser-arg rich domain is found in various proteins involved with RNA processing and serves as a binding site for splicing factors playing a critical role in alternative splicing (Long and Caceres, 2009). Notably, C-terminal ser-arg residue pairs of some protein families cause changes in regional conformation upon phosphorylation; these domains are often heavily phosphorylated with phosphorylation impacting the regional conformation and disrupting the interaction between the CBD and RNA (Blaustein et al., 2005). Furthermore, repeats containing high combinations of lys-ala-pro are

enriched in the linker DNA region of histone 1 flanking the nucleosome core where they may play a role in regulating gene expression or other chromatin-related regulatory processes since mutations in this domain have been shown to alter chromatin structure and gene expression (Fan et al., 2005). Moreover, sequences enriched in proline, threonine, or serine are observed to have a phase-transition tendency and form stress granules. These tendencies increase upon mutation in the LCRs of RBPs often leading to neuronal impairment (Zbindin et al., 2020). Notably, aliphatic residues comprise the highest AA residue frequencies in some disease-linked amyloid fibrils, including Aβ-42 (Ponte et al., 2004; van der Lee et al., 2014; MacLea et al., 2015).

0.8.5.2 Prevalence, distribution, and evolution of CBDs across animals

The prevalence and distribution of CBDs is widespread and expansive across the animal kingdom, like PLDs and IDRs, they are more highly prominent within protein families involved directly with regulation of cellular processes such as cell signalling, DNA replication, transcription, and autophagy. They are particularly highly conserved in transcription factors, DNA-binding proteins, and signalling proteins (Gallagher et al., 2022). Protein families involved in these processes have multiple CBDs which are usually highly conserved, meanwhile protein families not involved with regulation can have no CBDs. Not only are CBDs not equally distributed across proteins, within proteins their distribution is also favored at certain regional conformations of the protein such as loops or the region between α-helices and β-strands (Mizanty et al., 2020). Although the evolution of CBDs is still poorly understood some studies have suggested their emergence through gene duplication events following a divergence in sequence and protein function in response to specific environmental pressures such as evading host immune responses or changing environmental conditions (Belshaw and Jiggins, 2009). Moreover, proteins involved with stress response and DNA repair within tardigrades are particularly highly enriched with CBDs, an interesting observation given the highly adaptative capabilities of tardigrades to survive in extreme environmental conditions (Boothby at al., 2020).

0.8.5.3 Functional implications of CBDs in disease pathophysiology

Since CBDs in regulatory proteins are often functionally linked to protein interaction, a loss of function mutation or a mutation resulting in decreased optimization of a function within these domains, akin to IDRs and PLDs, is implicated in the pathophysiology of neurodegenerative diseases such as Alzheimer's and Parkinson's, cancer, and autoimmune diseases. For example, in ALS the CBD of FUS is essential for binding RNA and localizing it to stress granules, a normally reversible and tightly regulated protective process by which mRNA and RBPs are sequestered upon the onslaught of stress signals and thereby prevented from degradation or unfavorable translation, however, under chronic stress these granules form pathological aggregates exacerbating the development of ALS and FTD (Harrison and Shortner, 2017). Moreover, disease linked mutations in LCRs are observed to perturb the domains biophysical properties, specifically the competitive binding of RNA to RNA-recognition motifs. If unable to bind to these motifs RNA is free to bind to other LCRs for stress granule self-assembly consequently leading to an increased propensity towards LLPS and stress granule formation; an example of this is the glycine rich LCR of TARDBP leading to neuronal impairment (Lin et al., 2015; Zhang et al., 2015; Mackenzie et al., 2017; Gitler et al., 2017).

0.8.6 Selection of annotation software

As the number of sequenced proteomes increases annually, in tandem rises an increasing demand for accurate annotations as they become increasingly paramount in identifying new compositionally biased and disorder prone regions which may be functionally relevant. Inaccurate annotations from high false positives makes it difficult to find annotations with functional relevance, particularly for discovering meaningful novel mini motifs. Sifting through the annotations for discovering a motif with functional relevance requires additional information to increase prediction accuracy. Notably, current annotations for PLDs are often from biased predictors (i.e., N/Q bias) therefore several potential PLD sequence regions are overlooked, and consequently false negatives are high for several PLD predictors, particularly since other repeat motifs can also exhibit prion-like aggregation propensity such as the Y/G motif (Nielson and Mulder, 2019). To mitigate this, this study used LCRs determined by the fLPS

software and PLDs determined by PLAAC. Moreover, several predictors are not trained with fully ordered proteins and tend to predict them as disordered, thus, substantial number of false positive rates are observed as well on more balanced datasets (Nielson and Mulder, 2019; Liu et al., 2020). To reduce the false positives for IDR annotations, this study used both IUPred2a and DisoPred software where an IDR was identified with high confidence if both annotation software were able to identify the region as disordered.

0.8.7 Combinatorial analysis for novel annotations and meaningful insights

Information about IDR sequence length, position, motifs/repeats, conservation type, disorder classification, bias signature, overlap of domains of interest and degree of disorder all are relevant in determining possible function; considering these factors in juxtaposition gives multiple lines of support for increasing the likelihood of possible functions. Novel annotations were analysed in various ways: overlap of domains, novel presence, novel length, novel counts, and novel character. They were then further inspected by their GO terms; this was especially useful in determining PTM sites as well as RBPs or transcription factors. Screening for PTM sites is a useful exercise as the modulatory effect on this site has been repeatedly observed to be relevant to altering protein function and linked to disease progression. For example, phosphorylation of the PTM site in the microtubule-binding domain of tau was found to facilitate LLPS downstream leading to the aggregation of amyloid tangles (Wegmann et al., 2018). Determining novel annotations for RBPs is also useful because of their involvement in alternative splicing, a process that occurs exceptionally high rates in the central nervous system, whereby the larger number of isoforms may require more intricate and refined means of regulation making neurons highly sensitive and susceptible to mutations altering RBP behaviors (Heravi et al., 2022).

Moreover, it is useful to observe novel dramatic changes in length as length of the IDR or PLD is extremely relevant in disease pathophysiology. For instance, reducing the length of the LCR within the C-terminal IDR of RNA Pol II (RNAPII) led to decreased RNAPII clustering and recruitment of the transcription apparatus, meanwhile extending the region had the opposite

effect (Boehning et al., 2018). Length of LCRs within PLDs are important in determining the stability and rate of amyloid-fibril formation of amyloid forming proteins, moreover RBPs with longer LCRs more efficiently self-assemble and gravitate towards LLPS or stress granule assembly upon changes in the domain (Heravi et al., 2022). As for short IDRs, they are often observed to function as linkers containing specific MoRFs or linear motifs/repeats, whereas longer IDRs act as entropic chains possibly containing multiple motifs and domains and are commonly involved with protein recognition and tend to be in phosphatases and kinases; in both cases these are observed within flexible disorder regions (Lobley et al., 2007, van der Lee et al., 2014). Very long IDRs are involved with transcription related functions (Lobley et al., 2007). As for repeats, often disordered tandem repeats of one to two AA residues have functional roles.

An interesting example of the functional relevance of disorder conservation, AA composition, position, and length taken together, is the strictly conserved (for aromatic and charged residues) C-terminal loop in the non-NQ rich Het-s fungal homologs, this domain was determined essential, via mutagenesis screening, for prion propagation (Daskalov et al., 2014). Position is doubly important since the site of the IDR and its neighboring region/context provides functionally relevant information. The N'terminal tail of IDRs is observed to be involved with DNA-binding proteins, whereas C-terminal IDRs are observed to be more frequently involved with transcription factor regulation compared to IDRs at other sites (van der Lee et al., 2014). Analysing the residues within context of neighboring residues can be critical for determining the domain's role. IDRs may expose highly conserved short 3-8 AA peptide motifs which allow association with other proteins (Weathers et al., 2007). The identification of novel and possibly uncharacterized IDRs, PLDs, and CBDs will give greater insight into the biological role of these compositionally biased and/or disorder prone regions. Please refer to Figure 1.2 for an overview of the multiple considerations which give rise to confidence in predicting domain function.

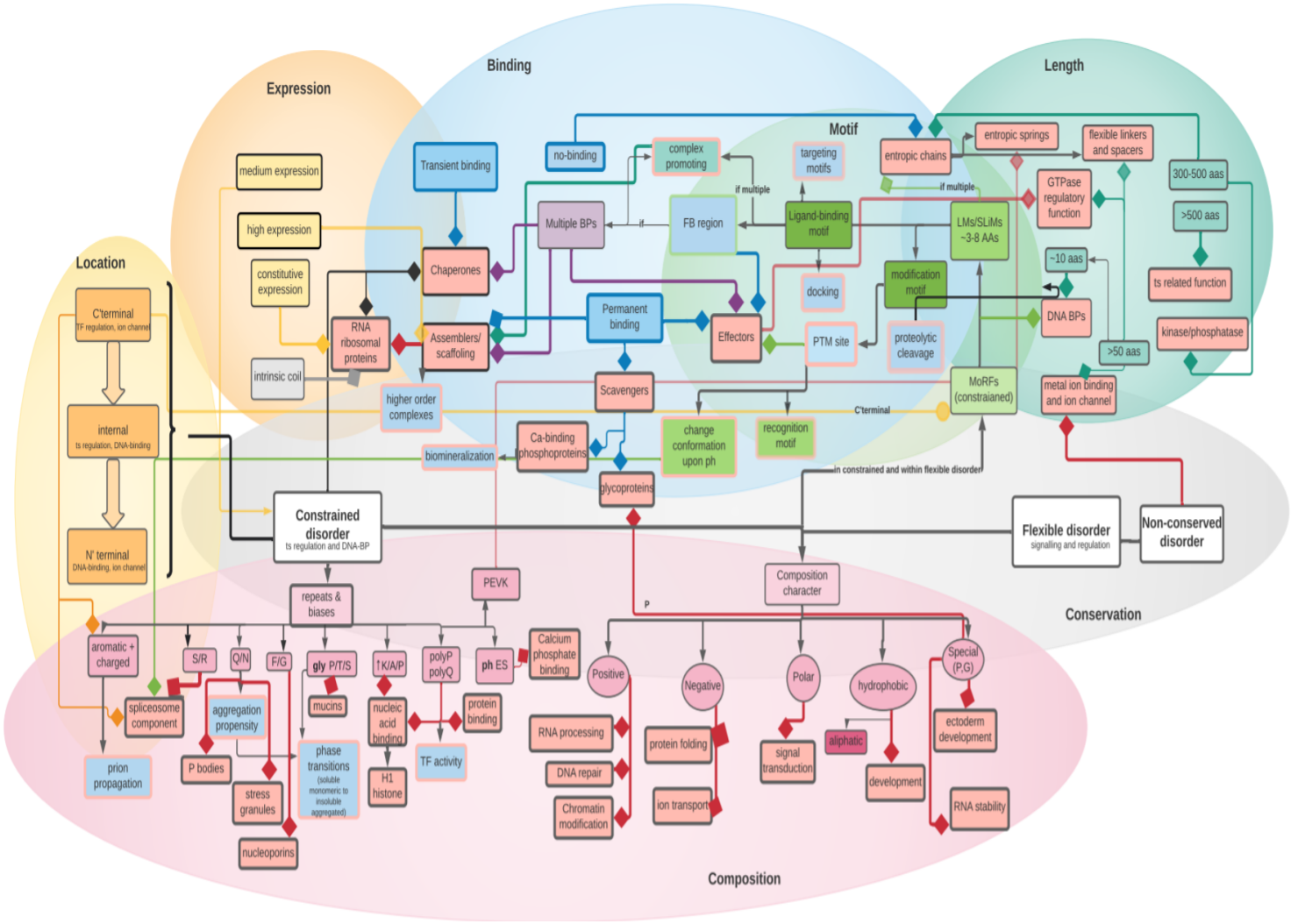

Figure 1.2 Flowchart depicting the order within disorder. A general framework for building multiple supports for domain function from sequence-based multi-dimensional analysis of different features as discussed in the review. Ts= transcription. Multiple sources were compiled for the creation of this figure (van der Lee et al., 2014; Daskalov et al., 2014; Bellay et al., 2011; Lobley et al., 2007; Moesa et al., 2012)

0.9 Methods

0.9.1 Data collection, processing, and dataset generation

58 animal proteomes, including common animal model organisms, were downloaded from UniProt based on an initial screening criterion for quality, a Benchmarking Universal Single-Copy Orthologs (BUSCO) complete score of at least 85 was required to meet high quality standards (UniProtKB, 2022) (Release 2021_05). The proteomes selected represent a diverse range of different clades within the kingdom with a maximum of four representatives within a clade and an average representation of three members per clade. Due to gross over-representation and under-representation within the database of high-quality proteomes in certain clades a few clades only have two member representatives. The dataset in total comprises 2.6 million proteins.

To create different orthogroups at different clade levels OrthoFinder was run, selected for its higher accuracy and ease of use for larger datasets (Emms et al., 2015). After the initial sorting of all proteins into orthogroups placed within major clade levels of the kingdom and excluding species-specific orthogroups as they provided no comparison to hierarchal orthogroups (HOG) for determining novel domains, the selected annotation predictors were run for all orthogroups within every clade on different computer clusters. These predictors were selected because of their higher performance accuracy relative to other available annotation software coupled with their availability as standalone programs able to be run on terminal and their ease of use for larger datasets (Erdos, 2020; Lancaster, 2014; Harrison, 2021).

For the screening of IDRs IUPred2a was used, these annotations were then cross validated using DisoPred to achieve higher annotation confidence. IUPred2a and DisoPred output were parsed using an awk script to find disordered regions, sequence tracts with every residue having a disorder score of 0.5 or above. These were then classified as ‘long’ is they were 30 or more residues long, or ‘short’ if between 20-29 residues. Furthermore, since position is relevant for inferencing function the position of the annotation was determined and denominated within five groups using a python script: end-N (eN)terminal, N-terminal, middle, C-terminal, and end-

C (eC) terminal. The annotation belonged to the end terminals if it began within the flanking 10% of the protein, if it was located within 11-20% of the protein from each end it was classified as N or C-terminal depending on the end, otherwise it was classified as middle. For the determination of PLDs, primarily PLAAC was used for screening using a core length of 40 residues and requiring a log likelihood ratio score above 0 with the weighting of background probabilities obtained from the input sequences. Additionally, LCRs found from fLPS coinciding with PLDs found from PLAAC gave higher confidence annotations (Lancaster et al., 2014). fLPS was run on the data twice to find CBDs and LCRs, once with the default settings screening for bias domains with a sequence tract length between 10-1000 residues and adjusted for a p-value cut-off of 10e-7, the second run was to find smaller LCRs with a tract length between 5-25 residues again with a p-value cut-off at 10e-7.

All annotations were then gradually compiled into a large overarching dataset using different computer clusters on Compute Canada. Subsequently the data was cleaned and parsed using R, a hierarchical id (HID) was created to track orthogroup identity comprising the tree node, number of the clade, and the last common ancestor of the proteins within the clade. Using these id-tags the proteins belonging to the same orthogroup were aggregated together reducing the 2.6 million rows with every protein as one row to 257 111 rows with each row containing all the proteins and respective annotations within orthogroups belonging in different clades. Subsequently, to track each distinct orthogroup path up the tree, or to identify all hierarchal orthogroups (HOG) within one path, a unique id was created for each orthogroup within the same path using a python script. Further, using python scripts novel annotations were identified through comparing relative differences which were sufficiently novel according to the selection criteria for each type of novel annotation between hierarchal orthogroups. Subsequently, a global analysis was conducted totalling the novel annotation counts at each clade level and by species using R.

0.9.2 Determination of novel annotations

For determining novel annotations four criteria were considered: presence, counts, length, and signature. A novel annotation by presence within an orthogroup was observed

when the percentage of the orthogroup having the annotation was 50% dissimilar to a respective higher or lower clade in its path and 50% or more of the difference was contributed by the annotation from that orthogroup, for an example please reference Figure 1.3. Accordingly, if the required threshold for either the change in percentage of orthogroup with the annotation or the contribution to that different coming from that orthogroup was more selective than 50%, for instance 60%, then the novel annotations will also be reduced in number but become more significant. The selection criterion thresholds were thus all selected to be moderate. Briefly, the percentage was determined by the proteins within the orthogroup with the annotation divided by the total number of proteins within the orthogroup, if the annotation percentage difference between the orthogroup and another hierarchal orthogroup (HOG) was 50% or higher and 50% or more of the resulting difference was attributed a result of the orthogroup, then a novel presence was marked for the orthogroup contributing the 50% or higher difference. Contribution for the orthogroup was determined by the number of proteins having the annotation within the orthogroup divided by the number of proteins having the annotation within the HOG. By adding a weighted component for contribution only the most novel annotations were mimed out of the data. Novel annotations for presence with these thresholds were done for both long and short IDRs and PLDs.

The second novel annotation criterion was based on the average number of counts of the annotation within the orthogroup. An average annotation count per protein was calculated for the orthogroup and compared to HOGs, if the difference in the calculated average annotation count per protein representing the orthogroup was three or greater between HOGs, then a novel annotation by count was observed for the orthogroup where the higher counts were observed. The difference in average counts between the HOGs were noted in a column, in addition, the average counts observed for all HOGs were kept in another column as a list for reference. Novel annotations by count were considered for IDR-long and IDR-short annotations.

Regarding novel annotations by length, if the average length of the annotation in the orthogroup, calculated by the summation of the length of all IDR-long annotations within the

orthogroup divided by the number of IDR-long annotations in the orthogroup, observed a difference of 100 AA residues or greater to the average length in a respective HOG then a novel annotation by length was observed for the orthogroup where the length was greater. Akin to the other novel annotations, a numerical value for the average difference was noted in addition to the average lengths of all HOGs listed for reference. Novel annotations for length were considered only for IDR-long annotations.

Lastly, novel annotations were considered for CBDs by novel signature: a signature was deemed 'novel' when a unique single-AA signature, categorized as a CBD with a LCR bias, was present within a lower-ranking HOG but absent in its immediate higher-ranking counterpart (besides within the orthologues in the higher-ranking HOG coming from the lower-ranking HOG). These comparisons were exclusively made between adjacent HOG clades to permit precise and reliable identification of novel annotations by signature. LCR signatures and CBD signatures were compared separately, for novelty the combination of both the same novel signature found for the LCR and for the CBD was required unless mentioned otherwise. Additionally, novel signatures were also grouped by different physiochemical characters: polar, aliphatic, aromatic, acidic, basic, or unique. This methodology was also extended to identify novel annotations in multiple-AA residue biases in a similar fashion.

0.9.3 Creation of data-subsets for case studies

For a more in-depth look at novel annotation occurrences subsets of the data were created for circadian genes and disease-linked proteins mining the global dataset for all orthogroups with the *homo sapian* protein for the selected proteins. The circadian proteins data subset was comprised of proteins well established for their central roles in circadian regulation including CLOCK, casein kinase I, Cry1, Cry2, HOX9, and BMAL1 as well as all their respective orthologues and paralogs within the orthogroups containing the hominid orthologues (Saini et al., 2015). For disease-linked proteins, APO-E, APP, BCL2, β-secretase, BRCA1, CHCHD10, FUS, MAP tau, NUPR1, p53, PAX5, PrP, PSEN1, PSEN2, α-synuclein, TDP-43, and Gsk3a were included with all their respective associated orthologues and paralogs in a similar manner (Chang et al., 2019; Lopez-Quilez et al., 2020; Nikolic et al., 2017). These

proteins were selected as they are highly cited and extensively studied in literature for their involvement in neurodegenerative disease or cancer.

0.9.4 Gene Ontology (GO) analysis

GO terms were accessed for each novel IDR and PLD annotation using the Gene Ontology Resource (accessed January 14th, 2023), GO terms for biological process, molecular function, cellular component, and any corresponding notes for PTMs were retrieved from the database for the orthogroup of proteins with the novel annotation. To preserve the integrity of clade-specific features and adaptations, a GO enrichment analysis was deliberately omitted from this study. Pooling data across diverse clades for such an analysis could risk diluting or obscuring the unique functional attributes inherent to individual evolutionary lineages.

0.9.5 Figures

iTOL was used to generate the phylogenetic trees, and a webserver tool was used to create venn diagrams (Letunic et al., 2007; Oliveros, 2015). Most other figures were made using R with the help of various packages including ggtree, tidyverse, and phylotool, while a few figures were made using Microsoft excel.

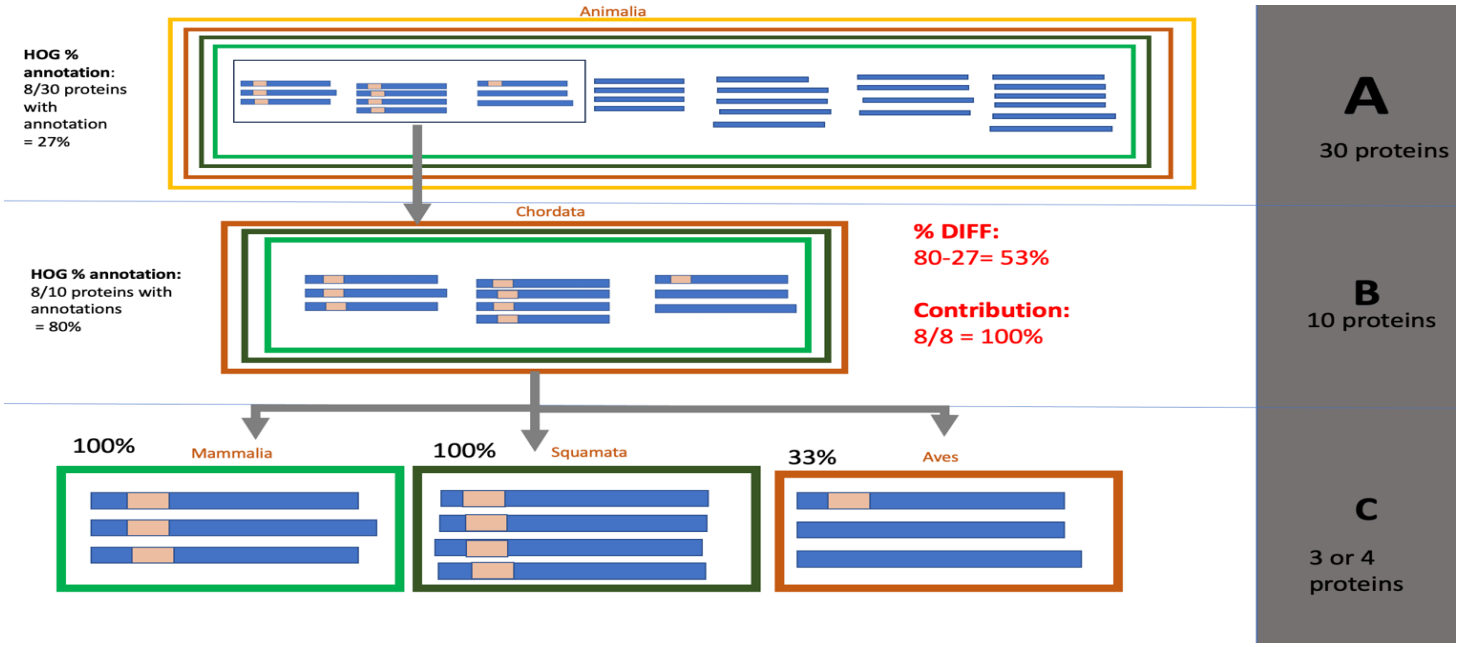

Figure 1.3: Example representation for determining % difference of annotation and contribution to difference between HOGs. A, B, and C are represented as different HOGs, each HOG represents all the orthologues of the protein within that clade. Protein sequences for orthologues are shown in blue bars with the annotation in cream. In this example *Chordata* has a novel annotation as the determined % difference and % contribution between it and its respective HOG in *Animalia* are both greater than the set 50% threshold.

1.0 Global Results

1.0.1 Novel IDR annotations

Three ways for determining novel IDRs were classified: by novel presence, by increase in the count of tracts, and by increased IDR length. To reiterate, a novel presence was considered if the annotation appeared within a clade but was not apparent in hierarchal orthogroups (HOGs). For this two-criterion needed to be fulfilled: first a percent difference of at least 50% between the orthogroups for the percent of the orthogroup containing the domain had to be observed and then secondly at least 50% of the observed difference when comparing the relative HOGs had to be contributed by the observed clade with the novel domain presence. A novel annotation by count was observed when there was a difference of three or more IDR tracts between the average count of tracts per protein within the orthogroup relative to the average number of tracts per protein in a HOG. Lastly, a novel annotation by length was considered when the difference of the average length of the annotation per protein between relative HOGs was 100 AA residues or greater. These thresholds were considered moderate as they outputted a moderate number of novel observations; with a more broader threshold range, for instance 30%, the number of novel observations jumped to thousands whereas with a more stringent threshold such as 80% there were only a few select observations. Also, note only annotations overlapping in both IUPred2a and DisoPred annotations were selected for IDR novel annotations for all categories to ensure greater confidence in results. The majority of novel IDR-long, long for disordered regions 30 AA or above, and IDR-short annotations, short for disordered regions between 20-29 AAs, for all novel categories were observed to be located at clades above the species-level. The only prominent species level annotations which is comparable to counts in the higher-ranked clades was for IDR-long presence for the crustacean species *Amphibalanus Amphitrite*. Refer to the phylogenetic tree in Figures 1.4-6 for more detail. One important consideration is the number of HOGs above the clade with the novel annotation, as more HOG steps is relevant for considering the significance of the novelty; more HOG steps allow for a more nuanced comparison for lower HOGs, however, it is also important to note clade representation in higher-ranked HOG clades with less steps is much better.

1.0.1.1 IDR innovations determined by 'novel presence'

A total of 345 orthogroups with high-confidence novel annotations for novel IDR-long appearance were found and 226 orthogroups containing a novel appearance for an IDR-short annotation. The largest gene ontology (GO) biological terms (BTs) represented for novel IDR-long annotations by presence were for protein phosphorylation, followed by transcription regulation by RNA polymerase II (RNAPII), and regulation of DNA-templated transcription. For IDR-short the top GO BTs were similar except proteolysis replaced protein phosphorylation in the top three terms. For GO molecular terms (MTs), the largest represented category for IDR-long was metal-ion binding, then ATP-binding, and DNA/RNA binding tied for third. IDR-short fared similarly, with the exception that RNAPII-specific was placed instead of ATP-binding within the top three GO MTs.

Of the 345 novel orthogroups identified by novel presence, predominantly the orthogroups were in *Mollusca* and *Acari* clades. From within the novel IDR -long annotations found within molluscs, the top GO MTs were for metal, ATP, and GTP binding. The mapped GO BTs for the novel annotation were very widespread within the clade with 57 mapped terms with protein phosphorylation, albeit the highest count only occurring thrice. The top GO MTs were similarly widespread within the novel annotations in *Acari*, with 29 mapped terms and the top count for regulation of DNA-templated transcription occurring but thrice. For the next highest-ranking clade *Hexanauplia*, ATP-binding and metal-ion binding were the top GO MTs of 35 terms, and protein phosphorylation for the top GO BT of 38. Hexanauplia was followed by *Hominidae* as the next highest-ranking clade for most novel IDR-long counts. Within *Hominidae* the top GO MTs were for DNA-binding transcription factor activity, followed by RNAPII specific and RNAPII cis-regulatory region sequence-specific DNA binding of 79 mapped terms, and regulation of transcription by RNAPII as the top mapped GO BT of 215 terms. *Hominidae* observed significantly more mapped GO terms likely due to more intensive research and annotation efforts within this clade.

For novel presence of IDR-short annotations, the top-ranking clades were similar to the top clades for novel IDR-long presence. Molluscs again were the top clade and with similar top GO MTs: ATP, nucleic-acid, and metal-ion binding of the 36 mapped terms. However, the top GO BTs within the clade were somewhat different, with intracellular protein transport ranking first from 29 terms. *Acari*, again, the second highest clade had similar top GO MTs: ATP, actin, and metal-ion binding holding the top spots of 36 mapped terms; moreover, here again intercellular protein-transport was the top GO BT. *Hominidae*, also ranking within the top five clades, with identical protein, RNA, and metal-ion binding as its top GO BTs, and proteolysis as the top term within an expansive range of GO MTs. Kindly refer to table 1.2 for further breakdown details for GO terms for top-ranking clades.

1.0.1.2 IDR innovations determined by 'novel count'

As for novel IDR annotations by count, 433 orthogroups with novel annotations were determined for IDR-long and 139 for IDR-short. *Mollusca* accounted for almost 20% of the novel IDR-long annotations by count, followed by *Hexacorallia*, and *Endopterygota*. *Hominida*e represented 10.5% of the novel annotations in this criterion. Perhaps not surprisingly more novel annotations were found for novel counts of IDR compared to novel presence as the IDR annotation software can breakup long IDR domains into smaller chunks at times. For novel IDR-long annotations by count the top three GO BTs were all related to transcriptional regulation with regulation or positive regulation of RNAPII as first- and second-ranked terms, followed by regulation of DNA-templated transcription. The top GO BTs for IDR-short were similar, however, including protein phosphorylation, and microtubule-based movement in addition to positive regulation by DNA-templated transcription.

The top mapped GO MTs within the top two ranking clades, *Mollusca* and *Hexacorallia*, were for metal, ATP, and zinc-ion binding, the top GO BTs for molluscs were for phosphorylation, whereas within *Hexacorallia* the highest GO BT counts were for DNA repair followed by terms specifying different mechanisms for DNA repair. Similar results were observed for the third highest ranking clade for novel IDR-long annotations by count, *Hominidae*, with metal-ion and

DNA binding as the top GO MTs of 135 mapped terms, and the top five GO BTs either related to positive or negative regulation of transcription by RNAPII or of DNA-templated transcription. For novel counts for IDR-short, the top clade *Hexacorallia* had ATP and microtubule binding as the top recurring GO MTs, and microtubule-based movement as the top recurring GO BT. Similarly, *Mollusca*, the next clade with the highest novel IDR-short by count annotations, had microtubule, zinc-ion, and ATP binding as the top recurring GO MT and microtubule-based movement as the top GO BT. For *Hominidae* the top recurring GO MTs were in relation to microtubule binding and positive regulation of microtubule polymerization for the top biological terms; whereas for *Endopterygota*, tied with *Hominidae* as the third top clade for novel IDR-short by count, metal and chromatic binding were the top recurring GO MTs, and protein phosphorylation as the top GO BT.

1.0.1.3 IDR innovations determined by 'novel length'

For novel IDR-long annotations by length, 433 orthogroups with novel annotations were identified. This threshold was set to extract more exaggerated cases to compensate for the annotation software's tendency to breakup IDR domains. As expected, the more extreme cases of these largely different lengthy annotations held proteins with much greater lengths compared to their fellow orthologues and co-orthologues. Molluscs were also the top clade within this category, herein top GO MTs were for RNA/DNA, metal-ion binding, ATP, and phosphatidylinositol binding and protein serine/threonine kinase activity of 59 total mapped terms; the top GO BTs were much broader with the top term only having three counts for protein phosphorylation. *Hexacorallia*, akin to novel IDR-long by count, was the second highest scoring clade, with top GO MTs for ATP, RNA, and nucleic acid binding of 63 mapped terms, and top GO BTs tied for protein phosphorylation and regulation of transcription by RNAPII, followed by regulation of DNA-templated transcription and signal transduction of 125 total mapped terms. Again, as with novel IDR-long by count, here too *Hominidae* ranked third, the top associated GO MTs were for RNA, ATP, and identical protein binding of 106 mapped terms; the top GO BTs included extracellular matrix organization, followed by brain development and regulation of transcription by RNAPII from a total of 360 mapped terms.

1.0.1.4 Novel IDRs in Hominidae and importance of HOG steps

One important consideration when inspecting the novel annotations by different clades is the number of HOG steps present to reach the clade, for this HOG steps are considered and convey how many HOG clades were above the clade with the novel annotation present in the context of this research. For instance, in the case of molluscs only five HOG steps are present, whereas for *Hominidae* ten steps are present as shown in Figure 2.6. This is relevant because, the increase in number of steps is indicative of a more significant novelty as there is more specificity in its divergent path and can indicate better how the HOGs are diverging for the annotation overtime down the tree. Another aspect to consider is clade-representation, *Hominidae* also has better representation with three species representing it whereas *Mollusca* only has two; naturally, higher-ranked clades with less HOG steps will have much better representation since they coalesce species from multiple lower-ranked clades. Best represented are lower-ranked clades with the most HOG-steps and most species-representation, for instance *Neognathae*. Interestingly, novel IDR by length was the most marked, or highest count, novelty to occur in *Hominidae* from the inspected novel annotation categories. For more details, please see Table 1.0.

1.0.2 Novel PLD annotations

Only novel annotations for novel presence within a clade with respect to relative HOGs was considered for PLDs, here 250 orthogroups encompassing novel PLDs by novel presence were identified spread across various clades, with the largest represented clade, *Mollusca*, only accounting for 16.4% of these novel annotations, followed by *Endopterygota* comprising 12%, and then *Hominidae* at 9.2%. Interestingly *Chordata*, *Mammalia*, and *Hominidae* were amongst the top five represented clades for novel PLD presence annotations. Interestingly, most novel annotations in the most highest-ranking clades were for novel PLD presence, please refer to the phylogenetic trees in Figure 1.4-6 provides for a more comprehensive overview. Among the 250 novel PLD annotations, 167 contained novel LCRs as well and of this set the associated top GO MTs were tied between DNA-binding transcription factor activity, RNAPII specific, and

metal-ion activity each with 21 counts, followed by RNA and DNA binding. The top four GO BTs from this set were for regulation of transcription by RNAPII, either positive or negative or DNA-templated, with the next highest counts between seven-eight for protein transport, cell differentiation, and protein phosphorylation.

From the novel PLD annotations residing within molluscs, the top associated GO MTs relate to metal-ion, DNA, and actin binding, with top associated GO BTs tied for actin cytoskeleton organization, protein transport, and ubiquitin dependant protein catabolic process from 35 mapped terms. *Endopterygota*, the second highest ranked in this category, had top GO MTs for DNA-binding transcription factor activity, followed by RNAPII specific, and metal-ion binding; associated top GO BTs include regulation of transcription by RNAPII, chromatin organization, and intracellular signal transduction of 106 mapped terms. *Hominidae* ranked third and observed a very widespread range and distribution of GO MTs and GO BTs, of them ATP and cysteine-type deubiquitinase activity for GO MTs, and immune response and spermatogenesis for GO BTs with only a maximum count of three amidst 100 mapped terms. The top GO MT for *Mammalia* was tied between DNA-binding transcription factor activity and RNAPII specific with ten counts each, considering there were only 20 novel annotations within this clade, 50% were associated with transcriptional regulation, moreover, the next leading terms were for DNA-binding and RNA-binding from a total of 68 mapped terms. In accordance the top GO BTs for novel PLD within mammals relate to positive or negative regulation of transcription by RNAPII, and negative regulation of DNA-templated transcription of 271 total mapped GO BTs. For *Chordata*, the top term only had 4 counts for ATP-binding, followed by DNA-binding transcription factor activity, RNAPII cis-regulatory region sequence-specific, identical protein binding and phosphatidylinositol binding of 72 total mapped terms. Associated top GO BTs were for transcription by RNAPII, cell differentiation, and innate immune response of 272 mapped terms.

1.0.3 Novel CBD annotations

A total of 35 569 orthogroups with novel single-AA signature CBD annotations with LCRs were identified where there was a novel signature gain observed in both the default CBD and the LCR run on fLPS, moreover, to avoid species-specific orthogroups there was an additional requirement for the HOG steps to be above one. Further, to refine the selection only novel presence of the CBD, rather than absence, was considered. Here *Mammalia* was significantly the most highest-ranking clade, followed by *Chordata*, and then *Bilateria.* Please refer to Figure 1.7 to view the novel annotations in a phylogenetic tree and Table 1.0 for more details. Upon further breakdown of the different signatures, 4 364 novel aromatic signatures and 14 618 novel aliphatic signatures were identified, with Chordates ranking as the top clade in both physiochemical groups, for reference please see Figure 1.8. As for novel acidic signatures, 10 369 were observed with *Bilateria* ranking as the top clade followed closely by *Mammalia* and *Chordata* and a total of 16 188 novel basic signatures were unveiled with *Chordata* and *Mammalia* nearly tied for top place. As for novel polar CBD signatures, 19 560 were identified, with novel serine signature presence accounting for most of the novel polar CBD signatures. Overall, the most novel signatures were from the unique AA classification (either glycine or proline) with 19 824 novel CBDs. Individually, the top signature with the most novel annotations was decidedly proline (15 826 novel counts), followed by serine (11 121 novel counts), glycine (10 995 novel counts), arginine (10 825 novel counts), and alanine (9077 novel counts). The least novel CBD annotations were observed mostly within the aromatic signatures with tryptophan (892 novel counts) having the least followed by tyrosine (997 novel counts) and then the aliphatic methionine (1 282 novel counts). Refer to Figure 1.9 for a phylogenetic tree reporting the novel polar, acidic, basic, and unique single-AA CBDs with LCRs and Table 1.0 for more details.

Additionally, some multiple-AA residue biases were inspected, considered as one in different possible combinations, mentioned earlier for functional significance in the literature review. For novel serine-arg (SR) biases, known to be involved as a binding site for splicing factors during alternative splicing of RNA processing, there were 2 854 observations with top GO MTs

associated with identical protein, metal-ion, ATP, and RNA/DNA binding and top GO BTs for signal transduction, positive and negative regulation of transcription by RNAPII, protein phosphorylation, and positive and negative regulation of DNA-templated transcription; herein the top ranking clades were *Metazoa, Bilateria, Chordata, Gnathostomata,* and *then Hominidae* after which there was a significant drop-off with clades only comprising 1.5% at most of the novel SR counts. Note novel counts in *Metazoa* were from those orthogroups which have a novel signature at the highest orthogroup comprising all metazoans, or orthologues from all animals considered within the study but not present in the lower OG; its presence from the top HOG disappeared in the lower clade. Another multiple bias signature considered was lys-ala-pro (KAP). As mentioned in the literature review, KAP bias is linked to regulating gene expression via regulating chromatin packaging and gene access, here 182 novel KAP signatures were identified with *Chordata* placing as the highest-ranking clade, followed by *Metazoa, Bilateria, Gnathostomata* and *Mammalia*. As expected, frequently occurring GO MTs in this category include DNA, chromatin, and histone binding, followed by ATP, protein kinase, and RNA binding. The most frequently appearing GO BTs were for positive regulation of DNA-templated transcription, chromatin remodelling, cell division, positive and negative regulation of transcription by RNAPII, and protein phosphorylation of 456 mapped BTs. Another multiple-AA residue bias considered was gln-asn (QN) with 1 220 novel signatures found primarily within *Metazoa, Bilateria*, and *Chordata*, clades were similarly ranked for ser-pro-thr (SPT) bias with 1 990 novel annotations. A note, for triple-AA biases like SPT, LCRs were not required to accompany the novel CBDs although were often present in double or single-AA residues of the bias (ie. SP or PT). This was acceptable since serine, proline, and threonine are all also individually known to increases proclivity towards phase transition and formation of stress granules from their single-AA residue biases as well as in combination. Most frequently occurring GO MTs for SPT bias were for identical protein-binding, DNA-binding transcription factor activity and RNAPII cis-regulatory region sequence specific DNA-binding. The top three GO BTs were all for transcriptional regulation: regulation of transcription by RNAPII, followed by positive, and negative regulation of transcription by RNAPII. Signal transduction and cell differentiation followed these terms.

1.0.4 Overlap of Novel Annotations

Although the novel CBD analysis only mainly concerned single-AA residue gains, nevertheless, thousands of single-AA CBDs with LCRs were identified within various physiochemical groups. The complexity and breadth of these findings necessitated a more meticulous inspection, focusing on novel CBDs with LCRs which overlap with novel IDR and PLDs to reduce noise in the data and gain more insight from the results. Within these constraints, 187 unique protein orthogroups were identified, although, if the single-AA signature only had to appear in either the CBD or LCR this number increased to 305. Additionally, if multiple-AA residue novel CBD signatures were considered that also overlapped with novel IDR or PLDs this number further increased to 850 novel protein orthogroups with CBD signatures. Considering the most restricted case of 187 novel overlapping single-AA CBDs with LCRs, the top emerging clades hosting these novel annotations were *Mammalia* comprising almost 30% of the novel overlapping signatures, followed by *Amphibalanus Amphitrite* at 10%, and *Acari* at 9.6%. Kindly refer to Figure 2.3 for the top clades from different physiochemical groups of these novel overlapping CBDs. From these overlapping novel single-AA signatures there were hardly any aromatic or aliphatic novel signatures besides alanine with a novel gain observed in 44 different protein OGs. The most novel signature gains were for proline with a count of 71, followed by serine at 50, and glutamine and glutamic acid tied at 45 counts. Tyrosine, isoleucine, and tryptophan had the least novel counts with only two to three novel cases. From all novel single-AA CBDs, the top GO BTs were for transcription regulation, either positive or negative, by RNAPII or DNA-templated transcription, other top terms included signal transduction, protein phosphorylation, cell migration, and regulation of cell cycle from 609 total mapped terms. Top GO MTs were for metal-ion, DNA, ATP, RNA, and chromatin binding. Interestingly from the 187 novel overlapping CBDs, 25 had documented PTM observances as provided by the GO database. Of these 25, 18 were OGs within Mammals with the top associated GO BTs for protein phosphorylation, signal transduction, and positive regulation by RNAPII. Of these 25 known to undergo PTMs five had GO BTs pretaining to brain development, and five for nerve

development or maintenance, with only one overlapping instance between the groups and seven of the nine total annotations from the two groups within mammals.

Also interesting was the overlap of PLD annotations and LCR with novel QN bias. 65 of the 250 novel PLD annotations have a CBD for either Q or N or in combination, accounting for 26% of the novel PLD annotations. The top three clades for this overlap are *Bilateria, Mammalia*, and *Chordata* with most frequently occurring GO MTs for DNA-binding transcription factor activity, RNAPII specific, and metal-ion binding with top GO BTs for either positive or negative regulation of transcription by RNAPII.

Moreover, as expected, there was considerable overlap between clades/species harboring novel annotations, particularly with novel IDR annotations and novel PLD annotations overlapping in all clades/species in which they occur and novel CBD present in almost all cases of these. More than 50% of novel IDR-long annotations (including IDR-long presence, IDR-long counts, and IDR-long length) overlap within the same clade/species. Overlap of novel annotations within the same species was much less observed, at the species-level generally only one or two novel annotations categories were found to overlap. Refer to Figure 2.1 and Figure 2.4 for more detail, and Table 1.1 for a complete breakdown of overlapping clades/species for novel IDR and PLD annotations.

Lastly the overlap between the different annotations considered is best highlighted in the Pearson correlation matrix in Figure 2.5, the most significant clustering in the top left corner is indicative of the overlap of the annotations identified reflecting the percent of each OG comprising that annotation, clustering shows a significant overlapping of long IDRs and short IDRs identified from IUPred2A and DisoPred. There is also moderate clustering observed for the variables encapsulating the novelty of long IDR by presence and short IDR by presence within the OG, and between novel long IDR by counts and short IDR by counts.

1.0.5 Overview and recap of findings

The results of this study are based on the bioinformatical analysis of proteome-wide annotations to identify novel patterns of protein domain evolution for IDRs, PLDs, and CBDs in the animal kingdom. Novel annotations were discovered at both internal node clades and at the species level excluding species-specific proteins. Highlights include the *Mollusca* clade hosting the most novel annotations for IDR and PLD for all novel categories, however, *Mammalia* held the most novel CBD signatures for novel single-AA gains. At the species-level *Amphibalanus Amphitrite* from crustaceans had the most novel IDR, PLD, and CBD annotations overall. Figures 2.0 and 2.2 sum the top clades with the most novel IDR, PLD, and CBD annotations separately for each category. Not surprisingly, there were thousands more orthogroups with novel CBDs identified, a total of 33 569, compared to the total high-confidence IDR-long and IDR-short novel annotations from all categories, totaling 1 598, and novel PLD annotations totalling 250. Of the 1 598 orthogroups with novel annotations for IDR-long and IDR-short, 1233 were from IDR-long: 345 from novel presence, 433 from novel number of counts, and 455 from novel length. Since CBDs encapsulate IDRs which often encapsulate PLDs, it was unsurprising the most novel annotations were found for novel CBDs, followed by IDRs, and then PLDs. As for novel CBDs, the greatest single-AA signature gain was found for prolines, and the least for tryptophan. Novel functional multiple-residue biases for SR, KAP, PTS, and QN were also analyzed which contained GO terms reinforcing their discussed putative functional roles.

1.1 Figures representing global results.

1.1.1 Phylogenetic Trees

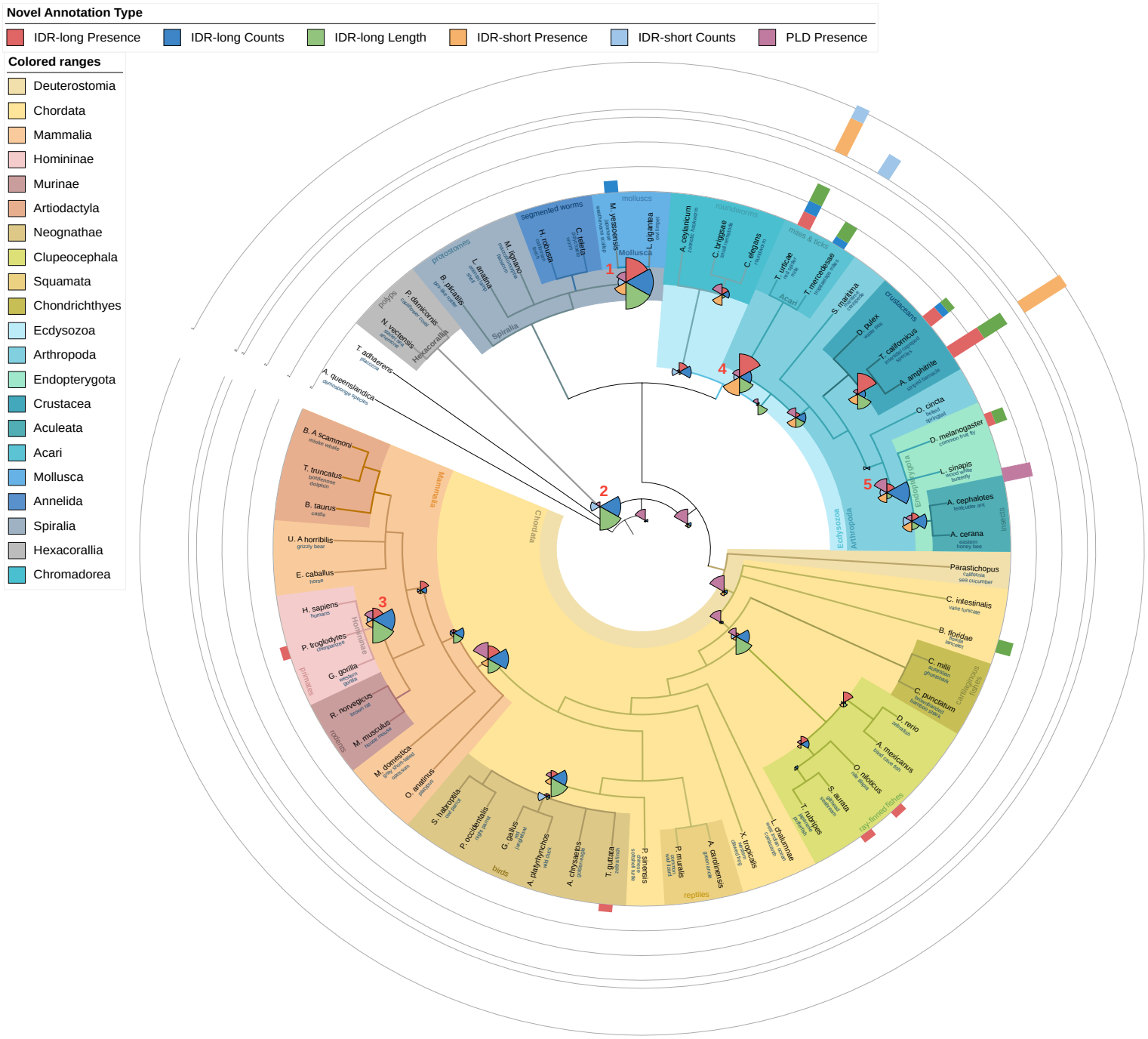

Figure 1.4: Animalia phylogenetic tree with IDR-long, IDR-short, and PLD novel annotations. Top four clades with greatest total number of novel annotations are labeled on the tree (1. *Mollusca, 2. Hexacorallia, 3. Hominidae, 4. Acari, 5. Endopterygota*). Polar pie charts are shown for the internal novel annotations present throughout the tree with the pie-chart radii determined by log 10 of the total novel annotations at the clade level. At the species – level, two levels of stacked bar charts are shown with all annotations on a log 10 scale.

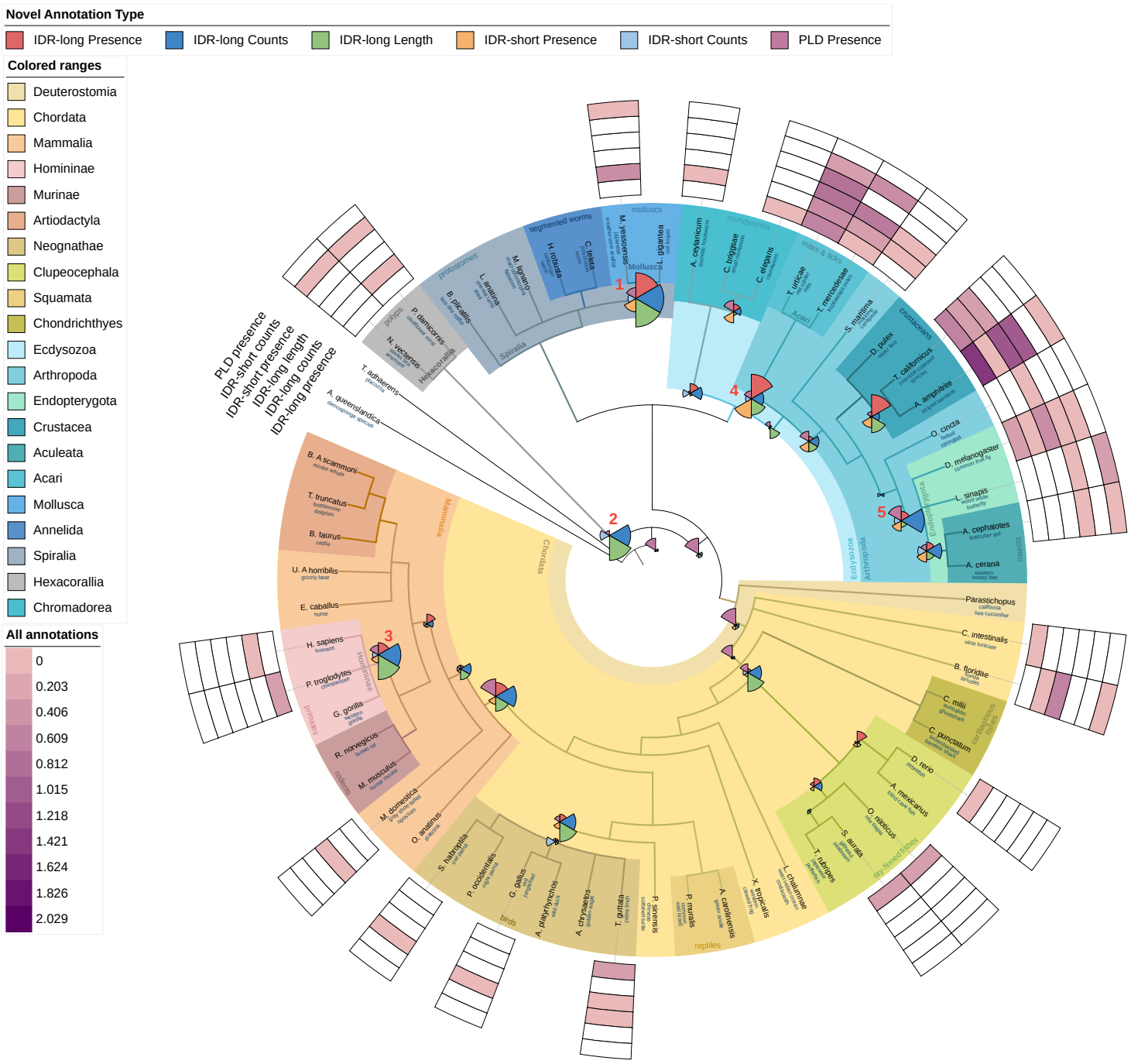

Figure 1.5: Animalia phylogenetic tree with heat map for log 10 of IDR and PLD novel annotations at the species level and pie-charts for novel annotations at internal nodes. Pie-chart radii are logarithmically represented by the total number of novel annotations at the clade level of the chart.

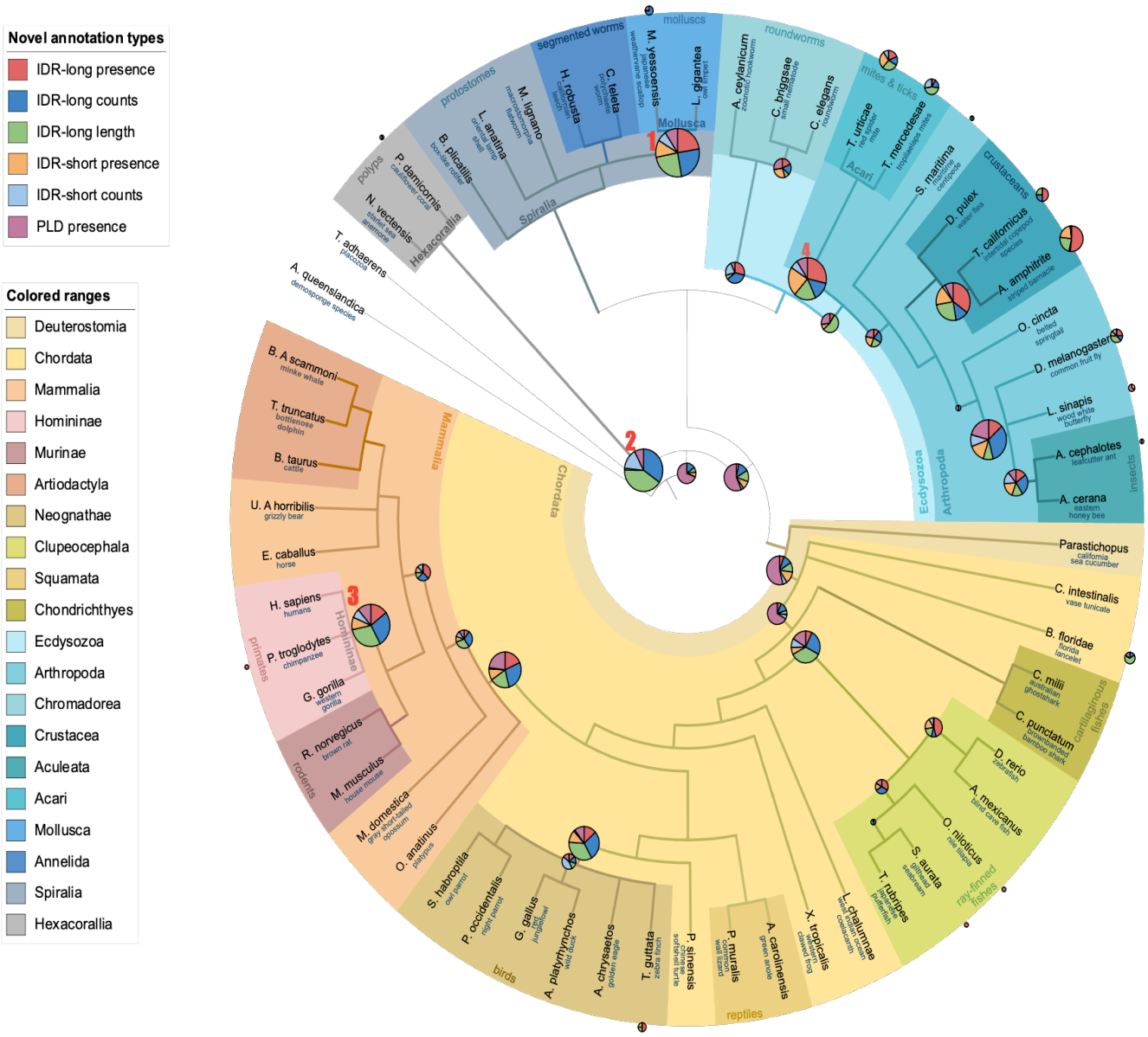

Figure 1.6: Animalia phylogenetic tree with pie-chart depiction of all internal and species-level IDR and PLD novel annotations. All novel annotations are shown in pie-charts including at the species-level for greater clarity on the relative total number of annotations identified at each clade level. Pie-charts radii are determined by log 10 of the total number of novel IDR and PLD annotations at the clade or species level.

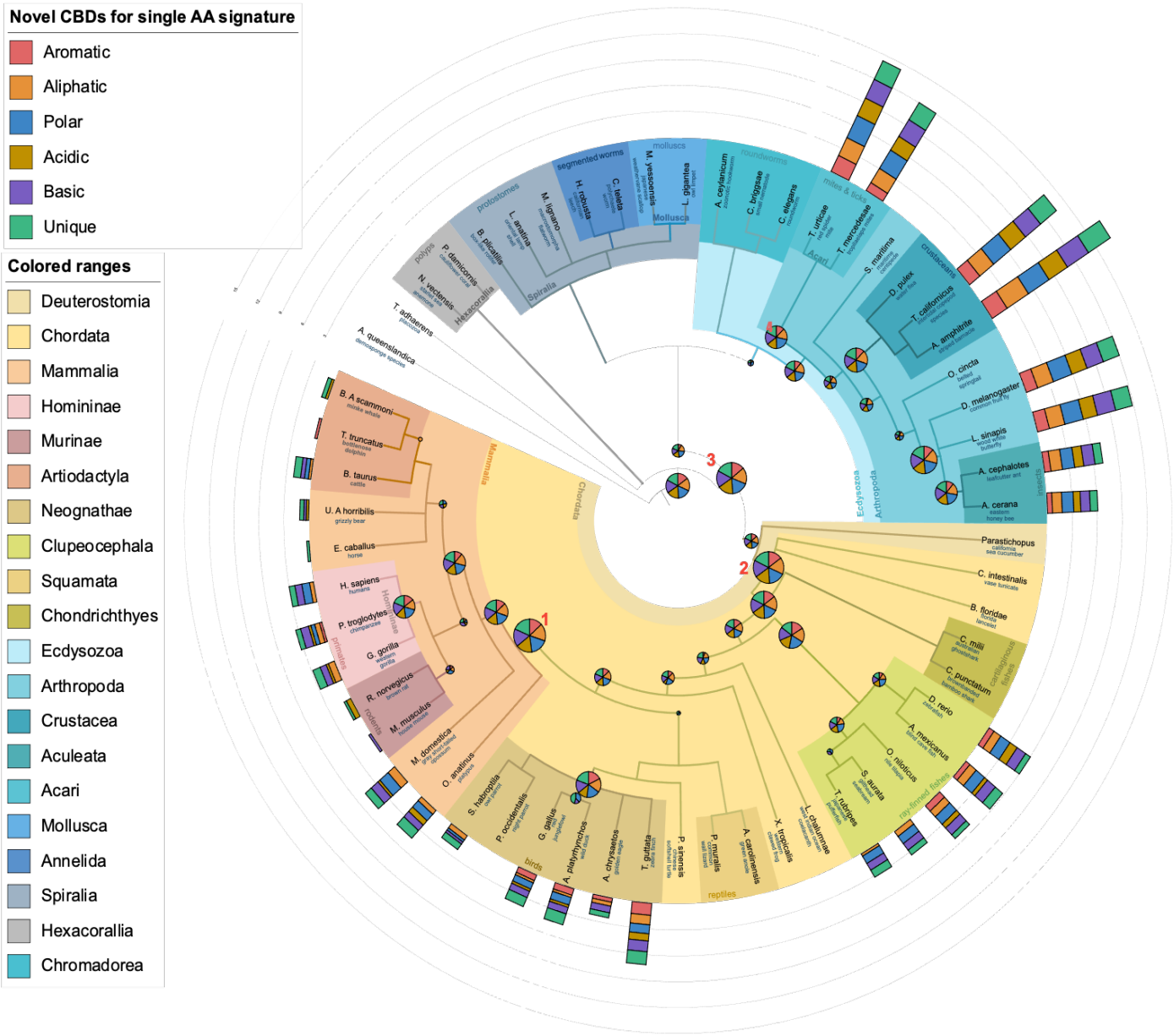

Figure 1.7: Animalia phylogenetic tree with novel single-AA CBDs with LCRs sorted into physiochemical groups. Internal pie-chart radii are determined by log 10 of the total number of novel single-AA CBDs with LCRs in the clade and stacked bar plots are shown on a log 10 scale highlighting novel single-AA CBD annotations sorted into physiochemical groups at the species-level. Top 3 clades are noted (1. *Mammalia*, 2. *Chordata*, 3. *Bilateria*).

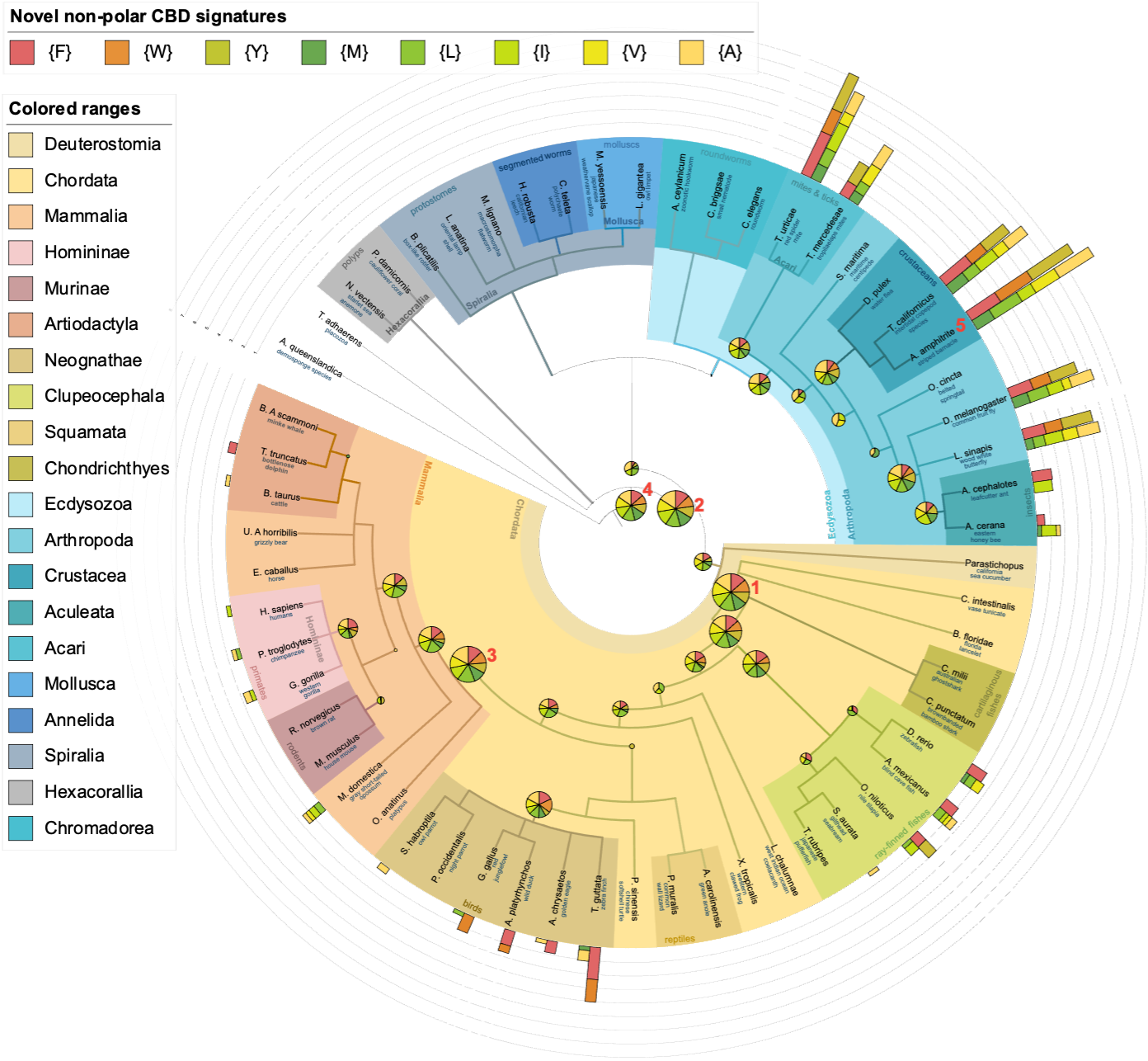

Figure 1.8: Animalia phylogenetic tree with novel single-AA non-polar CBDs with LCRs. Novel annotations in internal clades are represented in pie-charts with radii equivalent to log 10 of the total novel single-AA non-polar CBDs with LCRs, and similarly at the species-level but represented in stacked bar charts. Top 5 clades are denoted (*1. Chordata, 2. Bilateria, 3. Mammalia, 4. Metazoa, 5. Amphibalnus Amphitrite).*

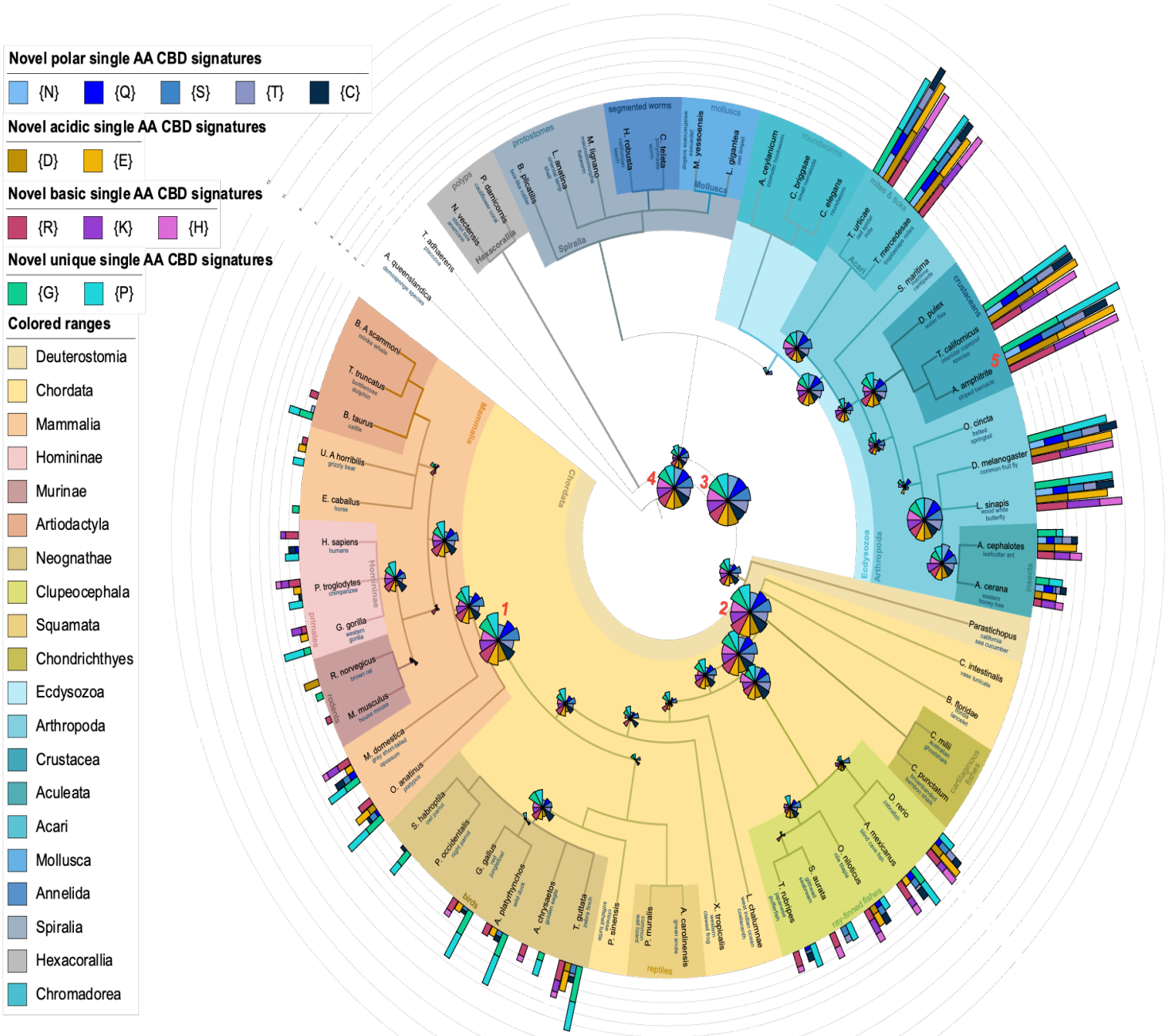

Figure 1.9: Animalia phylogenetic tree with novel single-AA polar, acidic, basic, and unique CBDs with LCRs. Novel annotations in internal clades are represented in pie-charts with radii equivalent to log 10 of the total novel single-AA polar CBDs with LCRs in the mentioned groups, similarly, performed for the species-level but represented in stacked bar charts. Top 5 clades are denoted (*1. Mammalia, 2. Chordata, 3. Bilateria, 4. Metazoa, 5. Amphibalnus Amphitrite).*

1.1.2 Other figures

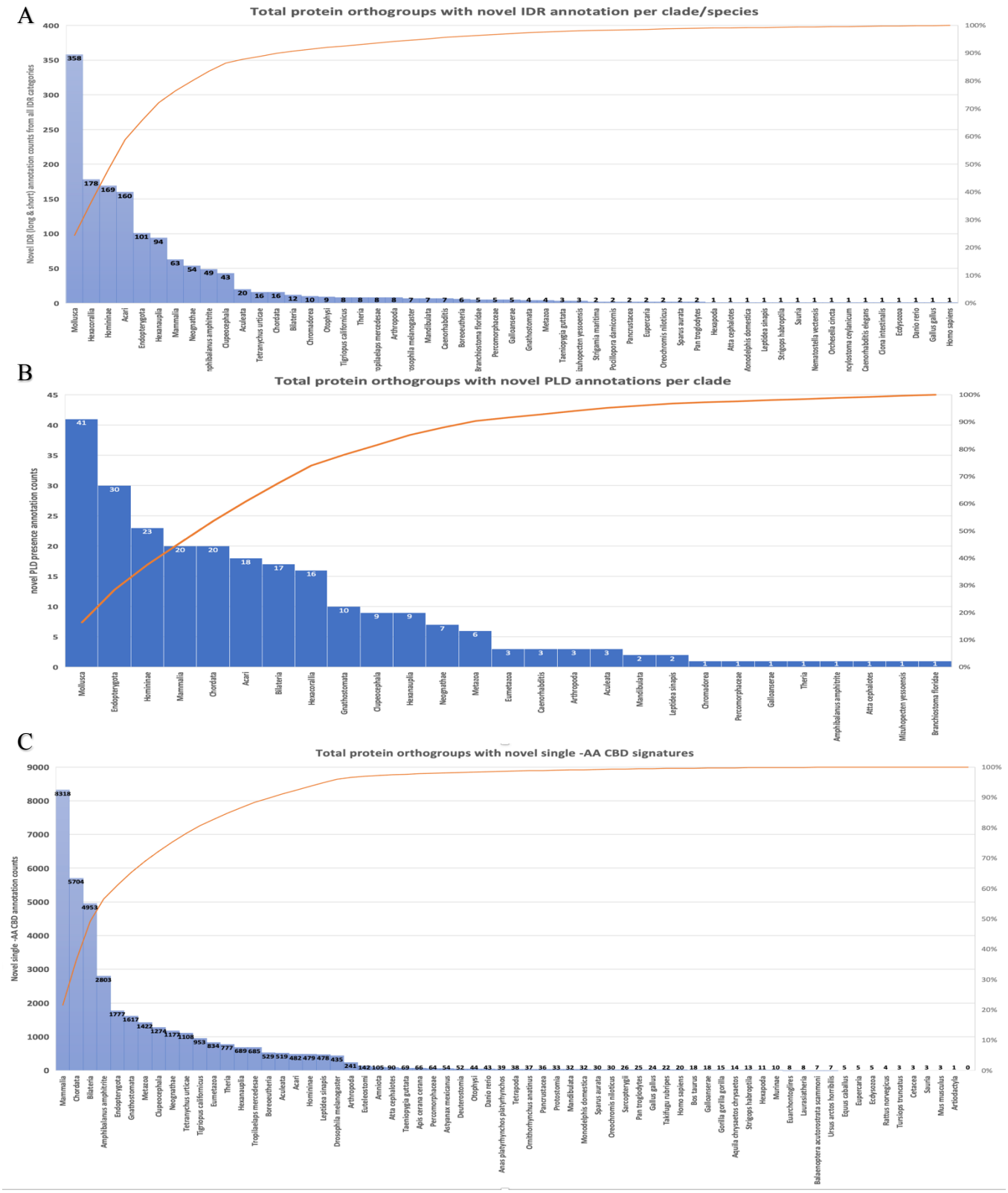

Figure 2.0: Bar plots depicting top clades or species with novel IDR (A), novel PLD (B), and novel CBD with LCR annotations (C). Trendline covers percentage of all novel annotations covered.

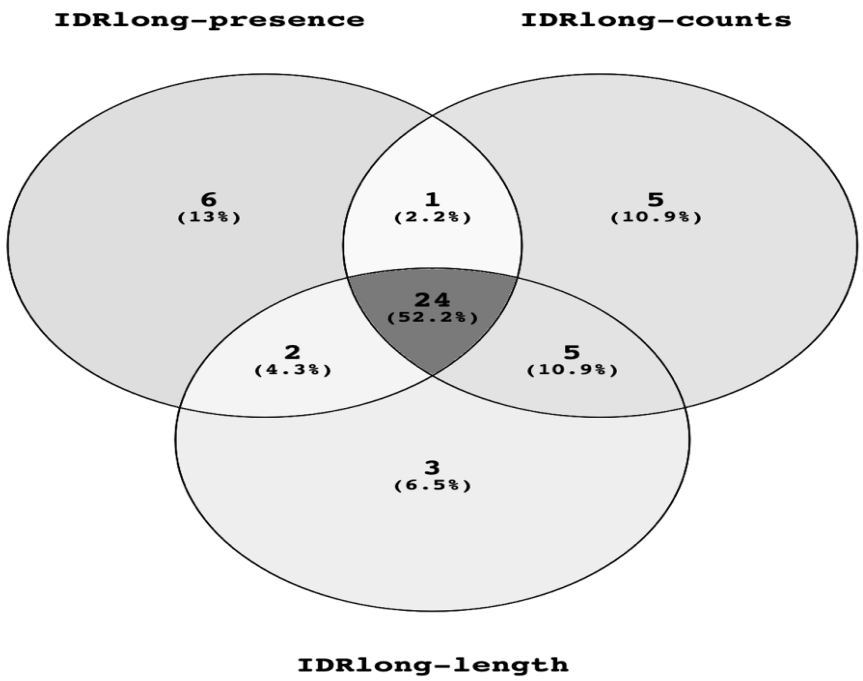

Figure 2.1: Venn diagram of all novel IDR-long annotation counts of different species/clades and their overlap with each other. 24 clades/species or 52% of all clades/species have overlapping novel annotations in all IDR-long categories.

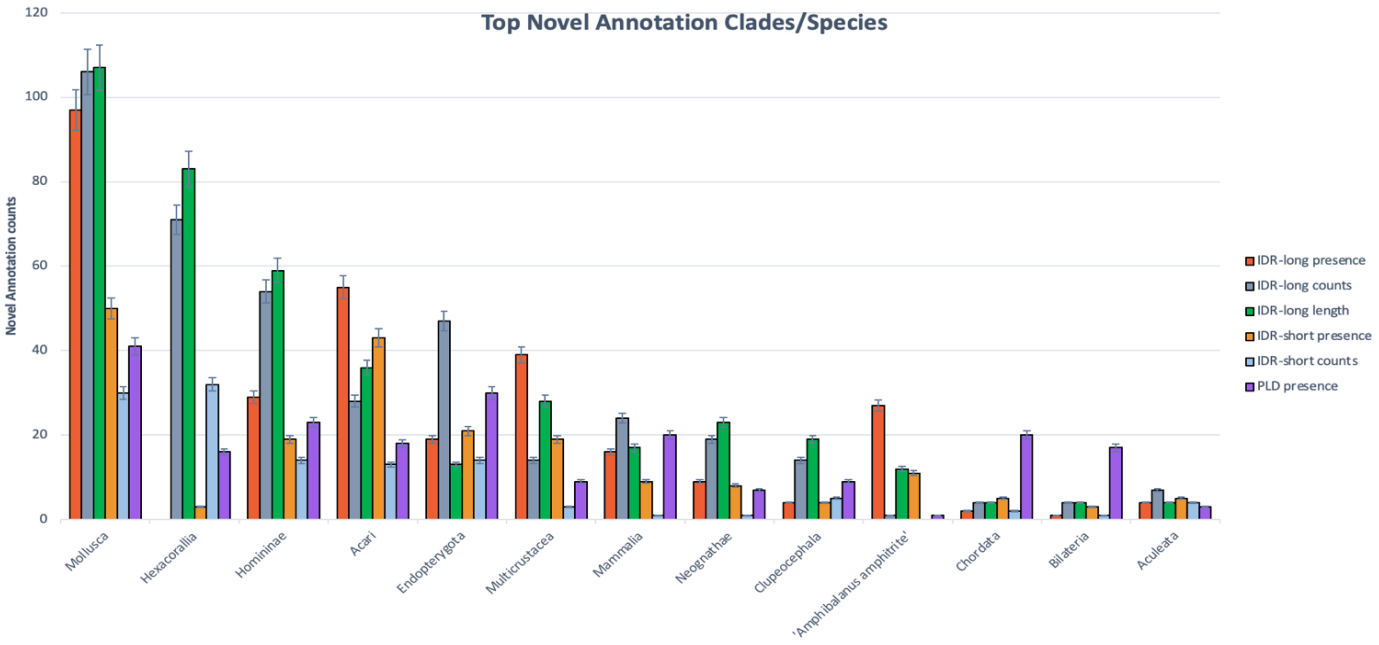

Figure 2.2: Stacked bar plot for the breakdown of different novel annotation types for the top 12 clades/species identified for having the most total number of novel IDR and PLD annotations. Percentage error bars are shown.

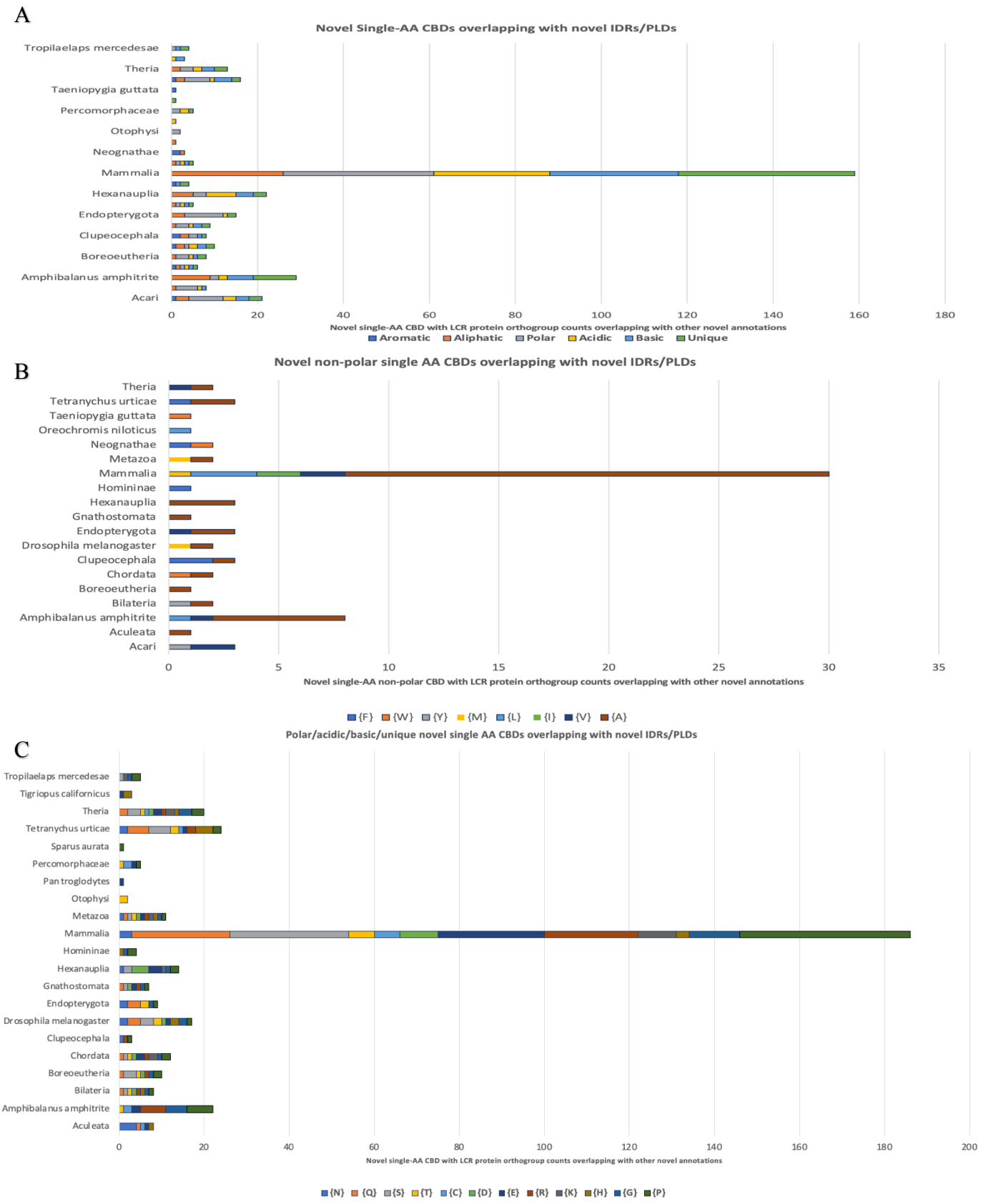

Figure 2.3: Stacked bar plots showcasing clades/species with novel single -AA CBDs with LCRs overlapping with other novel annotations. A. novel signatures sorted into physiochemical groups; B. non-polar novel signatures; C. polar,acidic,basic, or unique novel signatures.

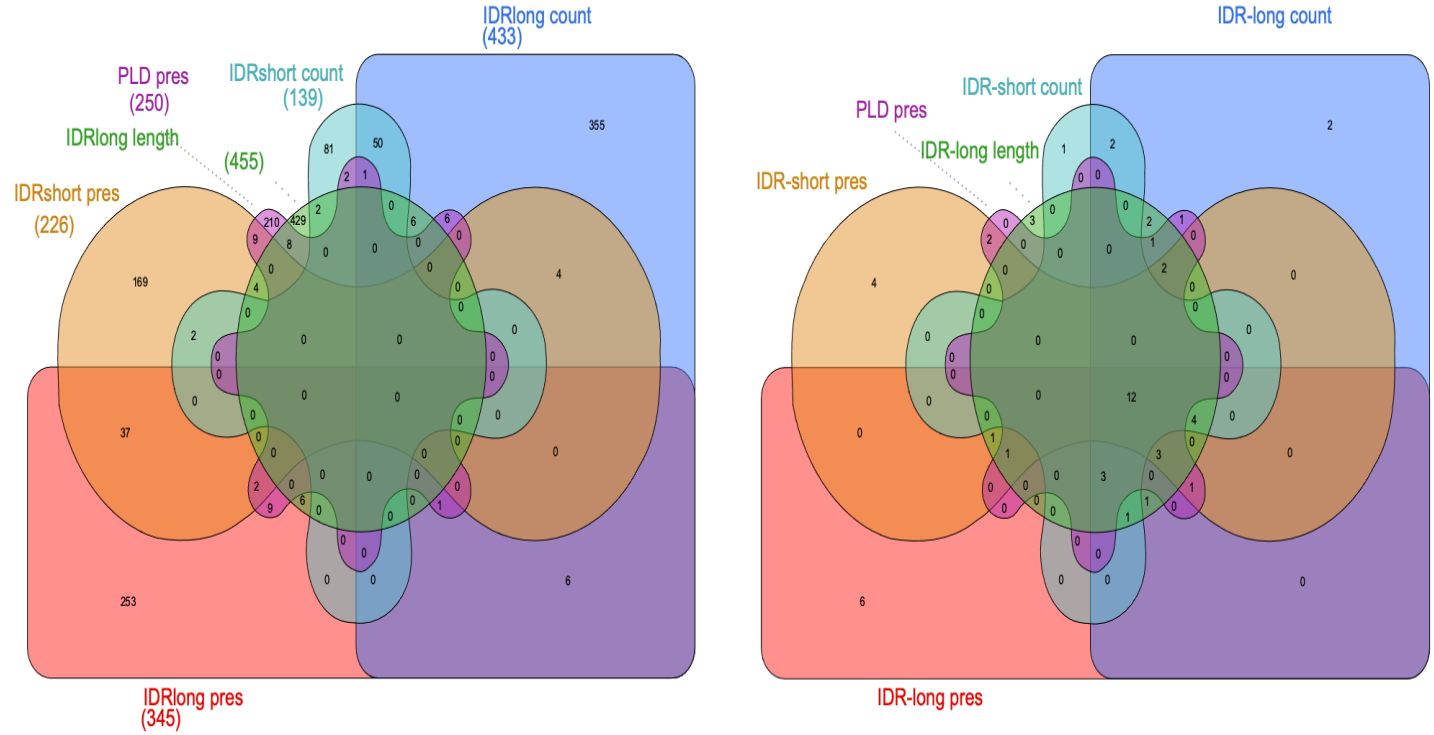

Figure 2.4: Venn Diagram highlighting the overlap of different novel IDR and PLD annotations. A. Between different protein orthogroups. B. Between different clades/species. Total protein orthogroups unique to the novel annotation category shown within parenthesis in A under the novel category.

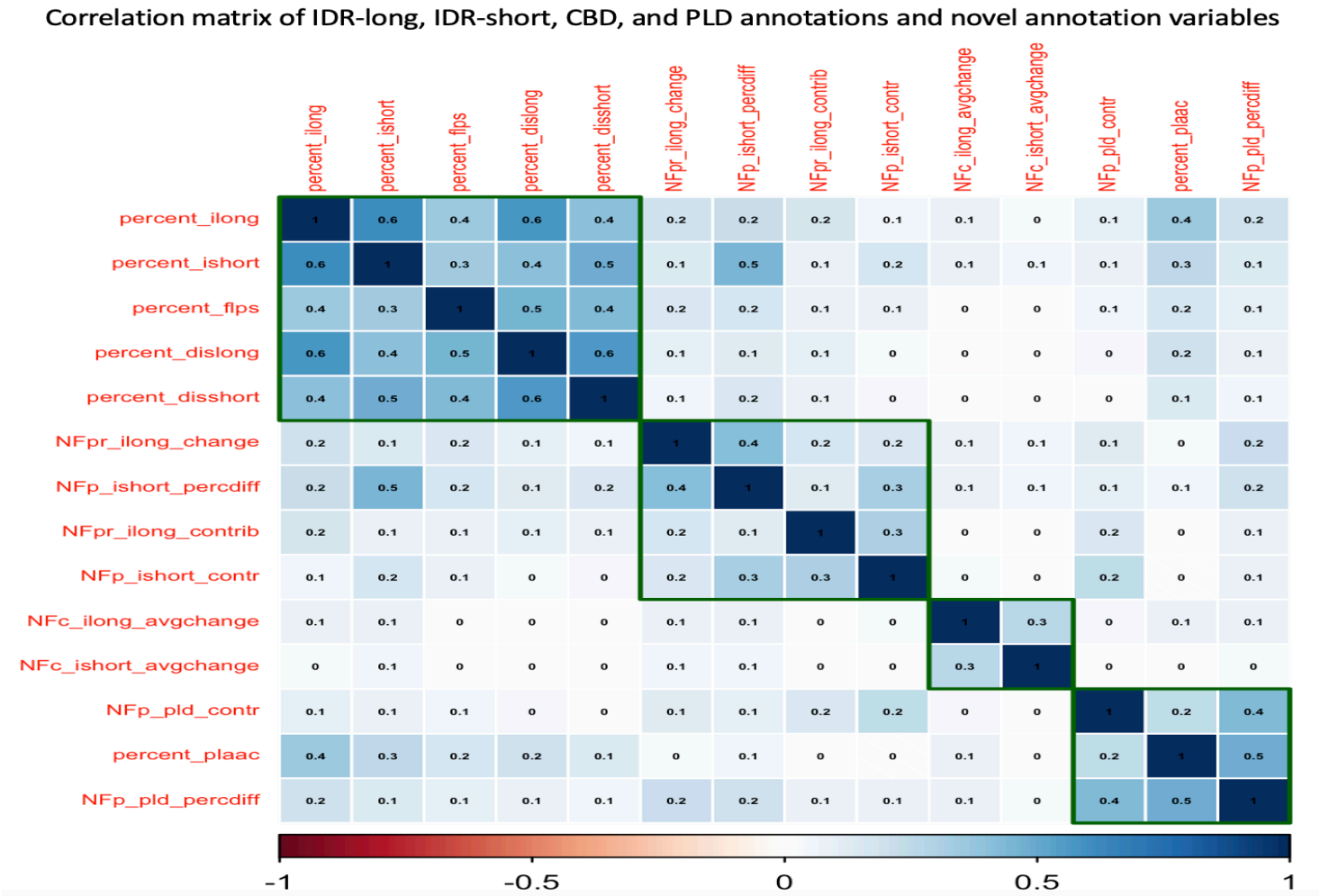

Figure 2.5: Correlation matrix for annotation statistics and novel annotation statistics. Pearson correlation coefficient used (linear dependence), ordered by Hclust clustering, green rectangles to better show clustering patterns.
From order shown in matrix: **percent_ilong:** percentage of the orthogroup with IDR-long annotation by IUPred2a; **percent_ishort**: percentage of the orthogroup with IDR-short annotation by IUPred2a**; percent_flps**: percentage of the orthogroup with CBD annotation by flps; **percent_dislong**: percentage of the orthogroup with IDR-long annotation by DisoPred; **percent_disshort**: percentage of the orthogroup with IDR-short annotation; **NFpr_ilong_change:** change between HOGs for IDR-long presence annotation; **NFp_ishort_percdiff**: percentage change between HOGs for IDR-short annotation presence; **NFpr_ilong_contrib**: contribution responsible for difference in IDR-long annotation presence change between HOGs; **NFp_ishort_contr:** contribution responsible for difference in IDR-short presence annotation change between HOGs; **NFc_ilong_avgchange**: change in the average count number of IDR-long annotation between HOGs; **NFc_ishort_avgchange**: change in the average count number of IDR-short annotation between HOGs**; NFp_pld_contr:** contribution responsible for difference in PLD presence annotation change between HOGs; **percent_plaac**: percentage of the orthogroup with PLD annotation by PLAAC; **NFp_pld_percdiff:** percentage change between HOGs for PLD annotation presence.

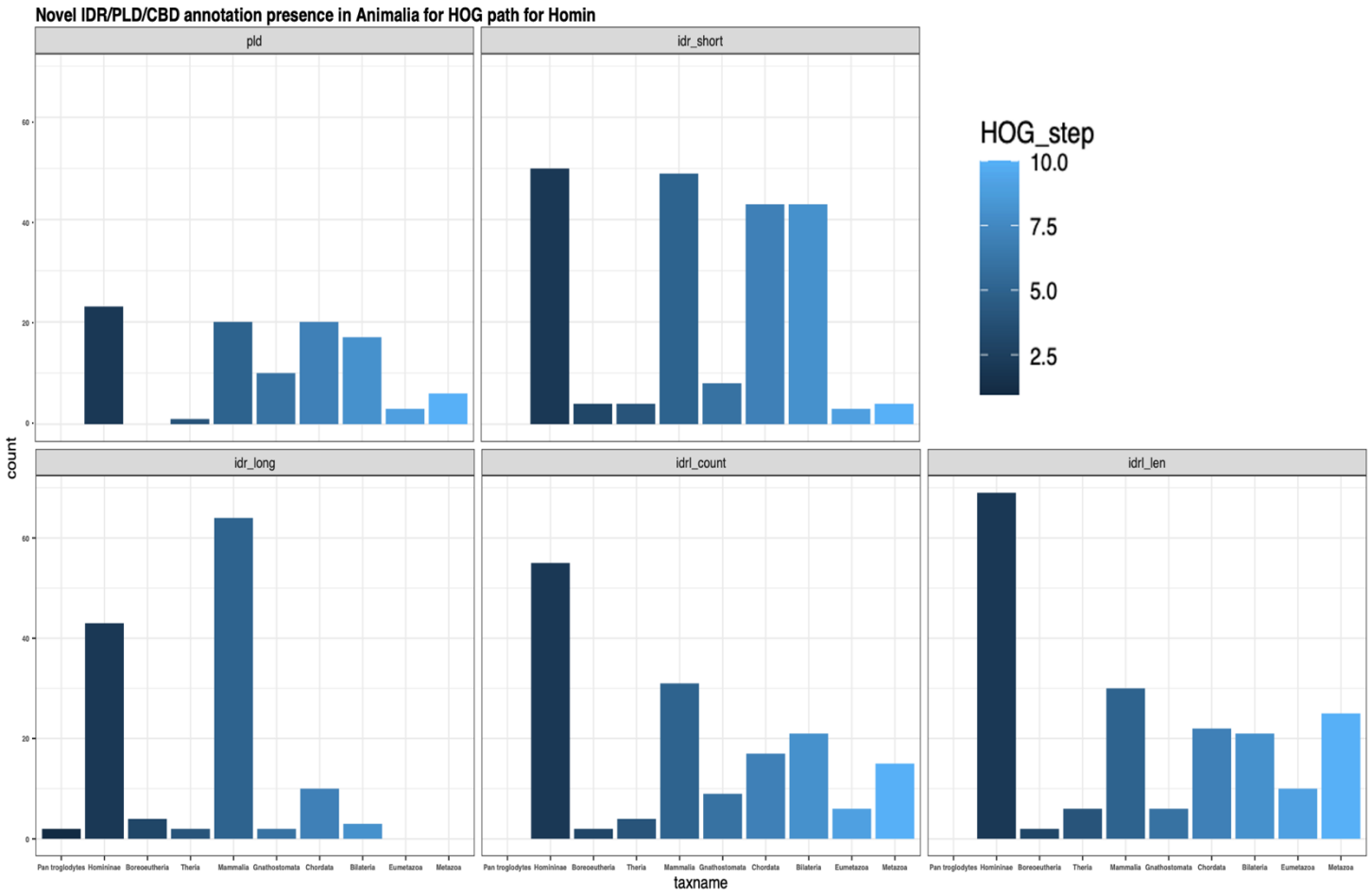

Figure 2.6: Novel annotations for all HOGs going down to the *Homindae* clade. From right to left, HOG steps are demarcated by blue gradient going up the tree where the gradient is lightest with the greatest number of steps (furthest travel up the tree). Pld: novel PLD; IDR_short: novel IDR-short by presence; idr_long: novel IDR-long by presence; idrl_count: novel IDR-long by count of tracts; idrl-len: novel IDR-long by length.

1.1.3 Tables summarizing global results.

Novel annotation	Threshold for extraction	Counts (unique)	Top Clades: % Total novel annotation of category
IDR-long presence	NF presence difference ≥ 0.5 & NF presence contribution ≥ 0.5	449 HC: 345	*1) Mollusca: 22%* *2) Acari: 15.9%* *3) Hexanauplia: 11%* *4) Hominidae: 8.4%* *5) Amphibalanus Amphitrite: 7.8%*
IDR-long counts	NF avg count difference ≥ 3	529 HC: 433	*1) Mollusca: 24.5%* *2) Hexacorallia: 16.4%* *3) Hominidae: 12.5%* *4) Endopterygota: 10.8 %* *5) Acari: 6.5%*
IDR-long length	NF avg length difference ≥ 100	603 HC: 455	*1) Mollusca: 23.5%* *2) Hexacorallia: 18.2%* *3) Hominidae: 13%* *4) Acari: 8%* *5) Hexanauplia: 6.2%*
IDR-short presence	NF presence difference ≥ 0.5 & NF presence contribution ≥ 0.5	484 HC: 226	*1) Mollusca: 22%* *2) Acari: 19%* *3) Endopterygota: 9.3 %* *4) Hominidae & Hexanauplia: 8.4%* *5) Amphibalanus amphitrite: 4.9%*
IDR-short counts	NF avg count difference ≥ 3	180 HC: 139	*1) Hexacorallia: 23%* *2) Mollusca: 22%* *3) Hominidae & Endopterygota: 10%* *4) Acari: 9.4%* *5) Clupeocephala: 3.6%*
PLD presence	NF presence difference ≥ 0.5 & NF presence contribution ≥ 0.5	250	*1) Mollusca: 16.4%* *2) Endopterygota: 12%* *3) Hominidae: 9.2%* *4) Mammalia & Chordata: 8%* *5) Acari: 7.2%*
CBD signature total 1AA	Hog steps > 1 (remove species-specific cases) Both Flps and LCR are present for signature in clade Gain in signature	35 569	*1) Mammalia: 21.6%* *2) Chordata: 14.8%* *3) Bilateria: 12.8%* *4) Amphibalanus Amphitrite: 7.3%* *5) Endopterygota: 4.6%*
CBD signature 1AA - aromatic	Novel signature is from {F,W,Y}	4364 {F} = 2783 {W} = 892 {Y} = 997	*1) Chordata: 23.6%* *2) Bilateria: 14.5%* *3) Metazoa: 10.9%* *4) Mammalia: 9.8%* *5) Gnathostomata: 8.7%*
CBD signature 1AA- aliphatic	Novel signature is from {M,L,I,V,A}	14 618 {M} = 1282 {L] = 4163 {I} = 2151 {V} = 1624 {A} = 9077	*1) Chordata: 19.1%* *2) Mammalia: 18.4%* *3) Bilateria: 14.5%* *4) Amphibalanus Amphitrite: 8.2%* *5) Metazoa: 7.2%*

CBD signature 1AA - polar	Novel signature is from {N,Q,S,T,C}	19 560 {N} = 4118 {Q} = 6807 {S} = 11 121 {T} = 4138 {C} = 5861	1) *Bilateria: 18.2%* 2) *Mammalia: 17.5%* 3) *Chordata: 15.9%* 4) *Metazoa: 6.2%* 5) *Gnathostomata: 5.5%*
CBD signature 1AA - acidic	Novel signature is from {D,E}	10 369 {D} = 4883 {E] = 8305	1) *Bilateria: 20.4%* 2) *Mammalia: 18.8%* 3) *Chordata: 17.9%* 4) *Metazoa: 8.8%* 5) *Gnathostomata: 5.8%*
CBD signature 1AA - basic	Novel signature is from {R,K,H}	16 188 {R} = 10 825 {K} = 5451 {H} = 4686	1) *Chordata: 19.4%* 2) *Mammalia: 19%* 3) *Bilateria: 17.4%* 4) *Metazoa: 7.1%* 5) *Amphibalanus Amphitrite: 6.7%*
CBD signature 1AA – unique	Novel signature is from {G,P)	19 824 {G} = 10 995 {P} = 15 826	1) *Mammalia: 24.8%* 2) *Chordata: 19.4%* 3) *Bilateria: 14.8%* 4) *Amphibalanus Amphitrite: 7.1%* 5) *Metazoa: 6.2%*
CBD signature – RS	Novel signature gain includes RS or SR signatures from both novel LCR and CBD fLPS runs	2 854	1) *Metazoa: 21.2%* 2) *Bilateria: 19.9%* 3) *Chordata: 17%* 4) *Gnathostomata: 7.1%* 5) *Hominidae: 6.3%*
CBD signature - KAP	Novel signature gain includes either KAP/KPAKP/APK/PKA/PAK from either LCR or CBD signatures	359	1) *Chordata: 21.2%* 2) *Metazoa: 17.8%* 3) *Bilateria: 15.3%* 4) *Gnathostomata: 8.6%* 5) *Mammalia: 8.4%*
CBD signature- QN	Novel signature gain of Q,N,NQ,QN in LCR or NQ,QN in CBD	1 220	1) *Metazoa: 27.5%* 2) *Bilateria: 22%* 3) *Chordata: 13.4%*
CBD signature- SPT	Novel signature gain includes any triple-AA bias combination of SPT, or two-letter or single amino-acid bias from SPT in LCR	1 990	1) *Metazoa: 18.8%* 2) *Bilateria: 18.6%* 3) *Chordata: 16.8%*
Novel single AA CBD overlapping with novel IDR/PLD	Novel single AA CBD and single AA LCR + novel IDR/PLD + HOG steps ≥ 2	187	1) *Mammalia: 29%* 2) *Amphibalanus Amphitrite: 10%* 3) *Acari: 9.6%*
Novel single AA CBD - aromatic Overlapping with novel IDR/PLD	Novel matching single AA CBD and LCR signature within {F,W,Y} + novel IDR/PLD + HOG steps ≥ 2	10 {F} = 5 {W} = 3 {Y} = 2	1) *Clupeocephala & Neognathae* 2) *Chordata* 3) *Acari*

Novel single AA CBD - aliphatic Overlapping with novel IDR/PLD	Novel matching single AA CBD and LCR signature within {M,L,I,V,A} + novel IDR/PLD + HOG steps ≥ 2	62 {M} = 3 {L} = 5 {I} = 2 {V} = 7 {A} = 44	*1) Mammalia* *2) Amphibalanus Amphitrite* *3) Hexanauplia*
Novel single AA CBD - polar Overlapping with novel IDR/PLD	Novel matching single AA CBD and LCR signature within {N,Q,S,T,C} + novel IDR/PLD + HOG steps ≥ 2	88 {N} = 18 {Q} = 45 {S} = 50 {T} = 24 {C} = 15	*1) Mammalia* *2) Endopterygota* *3) Acari*
Novel single AA CBD - acidic Overlapping with novel IDR/PLD	Novel matching single AA CBD and LCR signature within {D,E}+ novel IDR/PLD + HOG steps ≥ 2	55 {D} = 21 {E} = 45	*1) Mammalia* *2) Hexanauplia* *3) Acari*
Novel single AA CBD - basic Overlapping with novel IDR/PLD	Novel matching single AA CBD and LCR signature within {R,K,H} + novel IDR/PLD + HOG steps ≥ 2	64 {R} = 39 {K} = 17 {H} = 17	*1) Mammalia* *2) Amphibalanus Amphitrite* *3) Tetranychus urticae & Hexanauplia*
Novel single AA CBD - unique Overlapping with novel IDR/PLD	Novel matching single AA CBD and LCR signature within {G,P} + novel IDR/PLD + HOG steps ≥ 2	80 {G} = 32 {P} = 71	*1) Mammalia* *2) Amphibalanus Amphitrite* *3) Theria, Acari, & Hexanauplia*

Table 1.0: Top clades for every novel annotation category with thresholds for requirement for each category. NF: novel factor term for measurement for determining weight (contribution) and percent difference threshold for novelty; HC: high confidence where both DisoPred and IUPred2a gave an annotation, otherwise only IUPred2a (shown above). Top clade percentages are determined for the total number of novel annotations of that category, HC counts are used for all analysis.

Novel annotation categories	Total clades/species	Clades/Species
IDR-long counts, IDR-long length, IDR-long presence, IDR-short counts, IDR-short presence, PLD presence	12	Theria, Homininae, Bilateria, Endopterygota, Multicrustacea, Aculeata, Clupeocephala, Chordata, Acari, Mollusca, Mammalia, Neognathae
IDR-long counts, IDR-long length, IDR-long presence, IDR-short counts, IDR-short presence	4	Otophysi, 'Tetranychus urticae', 'Drosophila melanogaster', Boreoeutheria
IDR-long counts, IDR-long length, IDR-long presence, IDR-short presence, PLD presence	3	Caenorhabditis, Mandibulata, 'Amphibalanus amphitrite'
IDR-long counts, IDR-long length, IDR-long presence, IDR-short counts, PLD presence	3	Gnathostomata, Chromadorea, Galloanserae
IDR-long counts, IDR-long length, IDR-long presence, IDR-short counts	1	'Tropilaelaps mercedesae'
IDR-long counts, IDR-long presence, IDR-short presence, PLD presence	1	Percomorphaceae
IDR-long length, IDR-long presence, IDR-short presence, PLD presence	1	Arthropoda
IDR-long counts, IDR-long length, IDR-short presence, PLD presence	2	Eumetazoa, Hexacorallia
IDR-long counts, IDR-long length, IDR-long presence	1	'Tigriopus californicus'
IDR-long length, IDR-long presence, IDR-short presence	1	'Taeniopygia guttata'
IDR-long counts, IDR-long length, PLD presence	1	'Branchiostoma floridae'
IDR-long counts, IDR-long length	2	'Strigamia maritima', Eupercaria
IDR-long counts, IDR-short counts	2	'Pocillopora damicornis', Hexapoda
IDR-long counts, PLD presence	1	'Mizuhopecten yessoensis'
IDR-short presence, PLD presence	2	'Leptidea sinapis' ,'Atta cephalotes'
IDR-long presence	6	'Pan troglodytes', 'Danio rerio', 'Oreochromis niloticus', 'Ciona intestinalis' ,'Sparus aurata' ,'Caenorhabditis elegans'
IDR-long counts	2	'Homo sapiens', 'Ancylostoma ceylanicum'
IDR-long length	3	'Monodelphis domestica', Pancrustacea, 'Orchesella cincta'
IDR-short presence	4	'Gallus gallus', Ecdysozoa, 'Strigops habroptila', Sauria
IDR-short counts	1	'Nematostella vectensis'

Table 1.1: Overlap of clades/species for different IDR and PLD annotations. Reference table for Figure 2.4 diagram B.

Annotation Type		terms	terms
IDR-long presence	1) GO:0006468: protein phosphorylation, 18 2) GO:0006357: regulation of transcription by RNA polymerase II, 16 3) GO:0006355: regulation of DNA-templated transcription, 8	1) GO:0046872: metal ion binding, 33 2) GO:0005524: ATP-binding, 31 3) GO:0003677: DNA-binding, 16;	1) GO:0005634: nucleus, 64 2) GO:0016020: membrane, 49 3) GO:0005737: cytoplasm, 26
IDR-long counts	1) GO:0006357: regulation of transcription by RNA polymerase II, 32 2) GO:0045944: positive regulation of transcription by RNA polymerase II, 28 3) GO:0006355: regulation of DNA-templated transcription, 21	1) GO:0046872: metal ion binding, 73 2) GO:0005524: ATP-binding, 46 3) GO:0003677: DNA-binding, 45	1) GO:0005634: nucleus, 139 **2)** GO:0005737: cytoplasm, 78 3) GO:0005654: nucleoplasm, 51
IDR-long length	1) GO:0006357: regulation of transcription by RNA polymerase II & GO:0006468: protein phosphorylation, 24 **2)** GO:0006355: regulation of DNA-templated transcription, 20 **3)** GO:0045944: positive regulation of transcription by RNA polymerase II, 19	1) GO:0046872: metal ion binding, 46 **2)** GO:0003723: RNA-binding, 45 **3)** GO:0005524: ATP-binding, 37	1) GO:0005634: nucleus, 118 2) GO:0016020: membrane, 69 **3)** GO:0005737: cytoplasm, 66
IDR-short presence	1) GO:0006357: regulation of transcription by RNA polymerase II, 20 2) GO:0006508: proteolysis, 10 **3)** GO:0006355: regulation of DNA-templated transcription, 7	**1)** GO:0003677: DNA-binding, 26 2) GO:0046872: metal ion binding, 22 3) GO:0000981: RNA-polymerase II specific, 18	1) GO:0005634: nucleus, 52 2) GO:0016020: membrane, 40 3) GO:0005737: cytoplasm, 17
IDR-short counts	1) GO:0045893: positive regulation of DNA-templated transcription, 9 2) GO:0006468: protein phosphorylation, 8 3) GO:0007018: microtubule-based movement, 7	1) GO:0005524: ATP-binding, 17 2) GO:0008270: zinc ion binding, 16 3) GO:0008017: microtubule binding, 15	1) GO:0005737: cytoplasm, 33 2) GO:0005634: nucleus, 29 3) GO:0016020: membrane, 23
PLD presence	1) GO:0006357: regulation of transcription by RNA polymerase II, 35 2) GO:0045944: positive regulation of transcription by RNA polymerase II, 17 3) GO:0030154: cell differentiation, 15; GO:0045892: negative regulation of DNA-templated transcription, 15	**1)** GO:0000981: RNA-polymerase II specific, 35 2) GO:0046872: metal ion binding, 29 **3)** GO:0003677: DNA-binding, 27	1) GO:0005634: nucleus, 87 2) GO:0016020: membrane, 62 3) GO:0005737: cytoplasm, 48
All novel IDR and PLD annotations	1) GO:0006357: regulation of transcription by RNA polymerase II, 1180 2) GO:0045944: positive regulation of transcription by RNA polymerase II, 990 3) GO:0000122: negative regulation of transcription by RNA polymerase II, 757 4) GO:0007165: signal transduction, 674 5) GO:0045893: positive regulation of DNA-templated transcription, 628	1) GO:0042802: identical protein binding, 1521 2) GO:0046872: metal ion binding, 1420 3) GO:0005524: ATP-binding, 1057 4) GO:0003723: RNA-binding, 958 5) GO:0003677: DNA-binding, 934	1) GO:0005634: nucleus, 4166 2) GO:0005829: cytosol, 3292 3) GO:0005737: cytoplasm, 3183 4) GO:0005654: nucleoplasm, 2569 5) GO:0005886: plasma membrane, 2441

Table 1.2: Top GO biological, GO molecular, and GO cellular terms for novel IDR and PLD annotations. GO terms given followed by: description, followed by comma and number of counts.

1.2 Discussion of global results

The bioinformatical analysis from this study unveiled a broad distribution of novel IDR, PLD, and CBDs across several taxonomic levels from high-ranking clades to specific species, with most novel annotations falling into internal clades, particularly within chordates and mammals, sub-clades within *Anthropoda,* and most particularly within *Mollusca*. The different kinds of novel annotations determined can provide different insights into the role and evolution of the different regions and will be discussed individually below.

1.2.1 Significance of novel IDR

1.2.1.1 Significance of novel IDRs by presence

The emergence of a novel IDR within a clade may be indicative of a unique adaptation or functional requirement within the clade driving its novel presence. In the context of this study the high number of novel IDR-long presence annotations identified within *Mollusca* coupled with the most frequently appearing gene ontology (GO) terms within the novel annotations in this clade including metal, ATP, and GTP-binding related GO molecular terms (MTs) and protein-phosphorylated related biological terms (BTs) may be indicative of metabolic adaptations which have been essential for survival in diverse marine environments. The two molluscs considered in this study *Mizuhopecten yessoensis* and *Lotta gigantea* dwell in differing niches in the Pacific with contrasting environmental pressures, the former inhabits subtidal cold waters of the North whereas the later resides within an intertidal and constantly variable conditions with significant environmental fluctuations in temperature, salinity, moisture, and exposure to air (Sagarin et al, 2007; Silina AV, 2023). The observed novel emergence of IDR within these species may arise from differing adaptive responses to these contrasting environmental stressors which reflect the distinct molecular footprints of their ecological pressures considering most GO BT terms associated with the proteins with novel IDR annotations by presence in *Mollusca* are in relation to enzyme activity, intracellular protein transport, or initiation of transcription or translation with associated GO cellular terms located mostly in the membrane or nucleus. As for *Hominidae*, the IDR-long novel annotations seem to largely belong within proteins involved with regulation of gene expression, translation, and protein modification suggesting

opportunities to adaptions for more complex and intricate gene regulatory networks, potentially contributing to the higher cognitive function development seen in hominids. For instance, the POU domain, class 3, transcription factor 1, and transcription factor 3, UniProt id Q03052 and P20264, in *Hominidae* contained novel IDR by presence and have associated GO BTs involved with forebrain development. A further example is the *Hominidae* orthogroup containing Tyrosine-3-monooxygenase, UniProt id P07101, containing a novel IDR annotation by presence and having associated GO terms with cerebral cortex development and encompassing several serine sites for phosphorylation from different kinases.

Moreover, as discussed since the length of the IDR influences its role and potential for functional promiscuity, nonetheless, there was significant overlap between both novel IDR-long and IDR-short novel annotations by presence for top ranking clades and GO terms within those clades, notably for activities involving ATP, nucleic-acid, and metal-binding. These findings suggest roles in modulating metabolic processes possibly in response to environmental stressors. Intriguingly, intracellular protein transport emerged repeatedly as a more significantly emerging GO BT within IDR-short novel annotations. The prevalence of novel IDR-short orthogroups with protein intracellular transport GO BTs was particularly dominant within molluscs, interestingly, these terms were often accompanied by GO terms for metal-ion binding, such as calcium-ion binding, or binding of ATP, RNA, lipid, or small GTPase and involved with nucleocytoplasmic transport such as snRNA export from nucleus or endocytosis and late endosome to vacuole transport. This observation might point to a unique role of short IDRs in protein trafficking, short-IDRs may function as linkers and contain individual linear motifs or MoRFs, moreover, the specificity and simplicity of shorter IDRs might make them particularly suitable for guiding proteins to their correct locations within the cell (van der Lee et al, 2014). Acting as flexible linkers short IDRs can adopt multiple conformations required for the complex interactions required during the transportation process, often involving disorder-to-order transition upon binding, for instance, with the kinesin protein (Hyeon et al., 2007).

1.2.1.2 Significance of novel IDRs by count

While the emergence of novel IDRs within a clade often signals the advent of new functions, quantitative prevalence of IDR tracts may suggest enhanced specialization within existing functional pathways. Moreover, the acquisition of multiple short IDRs confers several advantages ranging from modular flexibility, allowing individual protein regions to adopt a variety of distinct local conformations, to augmented binding capacity for simultaneous interactions with diverse proteins and molecules. These advantages can facilitate rapid regulatory responses to evolving microenvironmental cues and signals which is essential for the intricately fine-tuned regulation observed in signaling pathways. Moreover, IDR-short regions benefit from a reduction in folding barriers that long IDRs can pose.

In the context of this study, *Mollusca* and *Hexacorallia* had the greatest number of novel IDR annotations by count both for long and short IDRs. The associated top GO BTs were related to metal, ATP, and zinc-ion binding possibly hinting towards playing a role in the sophisticated regulation of metal homeostasis within molluscan proteins. Given the broad spectrum of GO BTs present, the novel IDRs by count participate in a diverse array of cellular functions ranging from DNA repair to intracellular transport and protein homeostasis. This extensive participation, particularly for DNA repair functions, may underpin complex adaptations to counter varying environmental stressors. Similarly, the novel annotations by count within *Hominidae* may signify the evolution of specialized regulatory networks considering most of the novel IDRs by count within *Hominidae* have GO terms involved with gene expression regulation, with most of the top GO BTs pertaining to transcriptional regulation, chromatin remodelling, and DNA repair mechanisms.

1.2.1.3 Significance of novel IDRs by length

Akin to novel annotation by count, a novel annotation by length suggests the development of a more complex and sophisticated process, as opposed to a new functional adaptability. However, where novelty by tract count may indicate towards an increase in sophistication in regulation by novel participation of the protein in multiple transient interactions, novelty by

length may hint towards a focus on a single more complex multifaceted role in a single pathway or process.

Here too molluscs emerged as the top clade with top GO terms relating to DNA/RNA binding, signal transduction, and enzymatic activities. The longer IDR tract may influence the evolution of diverse and complex molecular functions, for instance through the addition of additional PTM sites more precise physiological adaptations may result, these may be required within the clade for the range of environmental conditions it faces. For novel annotation by increased IDR-long length the *Hexacorallia* clade followed *Mollusca* as the second highest-scoring clade with ATP, RNA, and nucleic acid binding related molecular functions. The functional versatility provided by the longer IDR may be significant to contributing to the mechanisms observed within this clade allowing it to withstand fluctuating ocean temperatures and other stressors (Liew et al., 2020). Intriguingly, *Hominidae* ranked third with top GO BTs relating to extracellular matrix organization and brain development. Here longer IDRs can benefit in the fine-tuning of several cellular processes employing a variety of mechanisms including more complex regulation by hosting more PTM sites and an expanded interface for protein-protein interactions allowing for enhanced interaction versatility and scaffolding functions. Since IDRs can modulate critical protein functions required for brain development and function, and alterations within IDR-containing proteins are implicated in neurodevelopmental and neuropsychiatric disorders, the novel expansion of IDRs may be essential for the precise orchestration of brain development paving the way for complex cognitive functions.

1.2.2 Significance of novel PLDs

The novel emergence of PLDs within a clade may be indicative of an evolutionary step towards more refined regulatory control, attributed to their distinctive capacity for conformational flexibility and enhanced capacity to mediate protein-protein interactions (PPI). In this study, *Mollusca, Endopterygota*, and *Hominidae* were found to have the highest incidence of novel PLDs, each associated with unique biological and molecular roles. In general,

the top GO terms for novel PLD presence pertained to various aspects of maintaining and regulating cellular function and homeostasis.

Within *Mollusca*, the top GO MTs were for metal-ion, DNA, and actin binding, accordingly, the top GO BTs involved fundamental cellular processes like protein transport and cytoskeletal organization. This corroborates with previous demonstration of the role of PLDs in enhancing cellular fitness under diverse conditions through their domain flexibility and thereby significant influence on PPIs (Chakrabortee et al., 2016). The *Endopterygota* clade ranked second for most novel PLDs with top associated GO BT functions for regulating DNA transcription either by RNAPII or regulating transcription factor binding to DNA. PLDs have been observed in several transcription factors which display prion-like behavior, aberrations to which are consequential in augmenting neurodegenerative disease (Harrison & Shortner, 2017). *Hominidae* was the third highest ranking clade with the broadest range of mapped GO molecular and biological terms spanning from ATP and cysteine type deubiquitinase activity to immune response and spermatogenesis. Interestingly, biological processes with high cell-to-cell variability were often associated with novel PLDs, particularly within *Hominidae*. The conformational flexibility of PLDs allows them to rapidly change shapes affecting their interactions with other cellular components; this dynamic behavior enables PLDs to perform various functions, contributing to cellular heterogeneity—the phenomenon where genetically identical cells can behave differently based on their molecular interactions. Consequently, PLDs play crucial roles in complex cellular processes such as development, differentiation, and the cellular response to environmental stresses.

This is particularly relevant in the context of immune system responses where immediate adaptation to a plethora of threats is imperative. Since responding rapidly to a broad range of potential threats is essential for the efficient job of the immune system, the employment of PLDs for such tasks is likely a significant contributor in facilitating the versatility required to achieve these tasks. Moreover, PLDs can contribute to the fast response required from the immune system by their essential participation in the formation of functional amyloids, these

can form components of inflammasome complexes involved in innate immunity (Lu et al., 2014).

An intriguing finding from analyzing novel annotations was that the novel annotations present in most higher-ranking clades were largely comprised of novel PLDs, consider Figures 1.4-1.6. For instance, going down the tree within *Metazoa, Bilateria, Chordata,* and *Gnathostomata* most of the observed novel annotations were novel PLDs. This prevalence suggests that these domains may serve pivotal roles in the regulation and conservation of fundamental cellular and biological processes, especially in the context of evolutionary progression within these clades. Moreover, the predominance of novel PLDs in these clades might hint at a nuanced interplay between protein structure and function which influences various physiological processes. Accordingly, PLDs in these higher clades may be involved in more diversified roles, potentially participating in, or regulating a myriad of cellular events, from the regulation of signal transduction to morphogenesis. Indeed, six of the top 10 GO BTs pooled from these four higher-ranked clades were for regulation gene transcription either positive or negative by RNAPII or DNA-templated transcription, the other four terms were cell differentiation, protein phosphorylation, innate immune response, and a tie at tenth for axon guidance, cartilage development, cell cycle, central nervous system development, and neuron differentiation of 550 mapped terms.

1.2.3 Significance of novel CBDs

The discovery of novel CBDs sheds significant light on the multifaceted aspects of protein structure and function. This study unveiled 35 569 novel LCR single-AA CBDs found across the animal tree in different clades. *Mammalia* dominated as the highest-ranking clade with the most novel signature counts, followed by *Chordata* and then *Bilateria*; this may be attributed to the relative complexity and diversity of these clades which encompass a plethora of protein structures and functions required for successful adaptation to their complex environments.

Interestingly from the identified single-AA residue signatures, novel proline signatures were most abundant. The unique cyclic side-chain structure of prolines which harbors a secondary amino group influences protein folding by constraining local flexibility, it provides structural hinges within the protein landscape which enhances protein stability due to its rigidity. It can also influence protein interactions as proline-rich motifs serve as recognition sites for multiple protein domains including the SH3, WW, EVH1, and proline-rich domains (PRDs) (Ball et al., 2005). The association of PRDs with phase transitions and stress granule formation has been established and is another interesting aspect as the high occurrence of novel proline signatures may hint towards the necessity for stress granule formation under cellular stress (Ball et al., 2005; Kim et al, 2008). The abundance of serine and glycine novel signatures also warrants a mention as in the former case serine's ability to undergo phosphorylation, critical for signal transduction, coupled with its frequently observed involvement in catalytic functions reflect its significant role in cell signalling, whereas glycine's small size allows it to facilitate small and tight turns within the protein contributing to its proper folding.

From the multiple-residue biases inspected, the novel emergence of SR bias, relevant to alternative RNA splicing, was identified in 3000 cases; GO terms for RNA processing and signal transduction often accompanied this bias implying novel roles within the clades where they appear for transcriptional regulation via RNA splicing. Novel gain of KAP bias, linked with chromatin related regulatory processes, was found in 182 cases, potentially indicating a novel role gain in regulation of gene expression via chromatin remodeling. Also worth noting are cases of observed novel Q-N and S-P-T biases, either individually or in combination, due to their association with stress granule formation via LLPS, the GO terms for these biases were linked to transcription regulation, signal transduction, and cell differentiation, thus further corroborating their involvement in a regulatory capacity.

1.2.4 Integrated discussion of correlated findings

Significant overlap was found between novel PLD annotations and novel gain in LCR CBDs, revealing 26% of novel PLDs overlapped with either novel glutamine (Q) or asparagine (N)

biased regions, or both. This overlap is noteworthy as it suggests an expanded functional versatility of these proteins possibly contributing to their wide-ranging GO associated terms. For instance, LCRs, particularly those with Q or N bias, are known to participate in phase separation of membrane-less organelles via LLPS contributing to the regulation of various cellular processes as mentioned in the literature review including RNA metabolism, signal transduction, and stress response (Boeynaems et al., 2018). A particular example of this is immune system responses, associated with three novel cases of PLD with QN bias, liquid condensates formed by phase-separated proteins can compartmentalize key signaling molecules, thereby concentrating, and accelerating the immune response (Banani et al, 2017). The overlap between novel PLDs and LCRs could potentially enhance the capacity of these proteins to undergo LLPS and participate in the regulation of various biological processes, of the 386 GO BTs mapped for novel PLDs 65 also encompassed a novel QN bias.

In the analysis that identified 187 unique protein orthogroups featuring novel single-AA CBDs with LCRs which also incorporated or overlapped with a novel IDR or novel PLD, *Mammalia* emerged as the most dominant clade constituting 30% of these novel overlapping annotations. The data spotlighted a pronounced presence of signatures for proline, serine, glutamine, and glutamic acid, in marked contrast to the paucity of novel aromatic and aliphatic signatures apart from alanine. As previously described, the recurrence of novel proline signatures may be related with the benefits offered by its inherent ring-like structure, suggesting potential implications for enhanced structural robustness, protein stability, and serving as potential recognition motif for certain protein domains. Furthermore, novel serine signatures indicate putative roles as phosphorylation sites pivotal in signal transduction pathways, this was in alignment with the top-associated GO BTs for signal transduction observed for novel serine signatures. Novelty of glutamine and glutamic acid signatures are also interesting, glutamine has been previously recognized for its pivotal role in stress-response in cellular functions, while the acidic side chain of glutamic acid is implicated in calcium binding via its carboxylate group, and participation in various catalytic functions and PPIs (Haber-Pohlmeier et al., 2007; Kan et al., 2015).

Amidst the 187 overlapping novel CBD annotations, 25 contained PTM annotations retrieved from the GO database, of these the majority belonged to *Mammalia*. Within these 25 orthogroups, the most common LCR signature was for proline with 14 novel counts, followed by a tie between alanine, glutamine, glutamic acid, and serine each with 10 novel counts. Beyond the general GO BTs linked to protein phosphorylation, signal transduction, and RNAPII regulation, a significant proportion of these annotations known to undergo PTMs were affiliated with GO BTs related to mammalian brain or nerve development. Such a relationship accentuates the importance of CBDs with affiliated PTMs in orchestrating intricate neural processes, setting the stage for a more focused investigation on these novel domains. In this study, 10 of the 25 PTMs had GO terms related to the brain, including some oncogenes such as proto-oncogene c-Crk and among these the top novel LCR signature was for glutamic acid and serine each with seven counts, followed by alanine, proline, and glutamine with six counts.

The discernible scarcity of aromatic novel signatures is likely indicative of their emphasis for conservation. Aromatic residues are often strategically positioned to execute specific roles within protein structures such as in π-stacking configurations at nucleotide binding domains or interfaces facilitating PPIs (Santos et al., 2013). These aromatic residues, given their bulky nature, could impose spatial constraints or have heightened energetic costs for sidechain rotamer transitions. Accordingly, alanine may be a more prevalent novel signature among the aliphatic residues because of its small size and simplicity, carrying only a methyl group it is ideal for allowing structural flexibility while forming hydrophobic bonds, for instance for helix formation.

1.2.5 Limitations and potential sources of errors

One of the key limitations of this study was the under-representation of certain clades within the animal tree due to the under-representation of high-quality proteomes available in UniProt which met the selection criteria for certain clades. Due to this there may be a slight bias for underrepresented clades for novel annotations as they have less species within the clade for

comparison and most likely less HOG steps from the top to observe the divergence among. Additionally, the potential inclusion of false positives or false negatives when identifying annotations is another limitation, although the likelihood of having false positives was mitigated by requiring an overlap of two annotations software for IDR annotations, and the default CBD run and LCR run using fLPS for the novel CBD signatures. Furthermore, PLD annotations were identified primarily using PLAAC, although LCRs were also run separately which often encompass PLDs; nevertheless, any false negatives inherent from how PLAAC determines PLDs will then affect the rest of the subsequent analysis. Some LCRs which are prion-like may be represented as false-negatives not picked up from PLAAC and identified only as LCR in the dataset.

Another limitation for IDR annotations stems from their binary classification system entailing either disordered or ordered states. This is unsatisfactory in describing physiological states, *in vivo* molecular functions involving IDRs often have a range of intermediate dynamically shifting protein assemblies ranging between completely ordered and completely disordered and this wide spectrum accounts for different degrees of disorder. Hence, disorder is more appropriately described as a continuous spectrum and the binary classification is limiting. Moreover, an only sequence-based analysis approach is limiting for thoroughly identifying evolutionary relationships in proteins as many evolutionary distant relationships between proteins are masked and not detectable at only a sequence-level analysis but can be observed by structural similarities since structural similarity can be incredibly high albeit very low sequence similarity. However, since this study targets disordered regions, this relation is less significant than for ordered protein relations; nevertheless, since structural schemes for folding upon binding and proteins complexed with ligands are now becoming more available, threading the disordered sequence through known bound protein models, or recognized folding-upon binding structural motifs may identify structural similarities between orthologues which can help shed more light on how these protein domains are evolving when coupled with the sequence analysis approach.

1.2.6 Relevant therapeutics and future directions

The identification of novel IDRs, PLDs, and CBDs unearth previously uncharted territories in protein domain biology bringing forth various exciting opportunities. The association of novel IDRs with intricate gene regulatory networks, particularly within the *Hominidae* clade, offers a compelling research direction. The expansive interface of particularly long novel IDRs allows for multiple PPIs and has greater likelihood of hosting more PTMs sites, thus making an excellent target for further investigation. Investigating novel IDRs related to extracellular matrix organization and brain development in *Hominidae* is another compelling research direction. Moreover, the association of certain novel IDRs with metal-binding functions is a fascinating area to explore as metal homeostasis has been increasingly recognised as vital in several biological processes with dysregulation having implications in amyloid formation and thereby disease implications (Sadakane et al., 2018). In an evolutionary context, novel IDRs in different clades from *Mollusca* to *Hexacorallia*, present an interesting avenue for deciphering evolutionary responses to environmental stimuli, specifically how organisms like molluscs can harness these novel domains as adaptive mechanisms to cope with environmental stressors; this may lead towards providing solutions in conservation Biology efforts.

Additionally, considering the potential role of certain PLDs in orchestrating immune responses, there is a pressing need to delve deeper into their functions; advanced immunotherapies could be designed, targeting proteins with a novel QN bias, especially for conditions rooted in dysfunctional immune responses. Further, the intricate roles of serine and glutamic acid in cell signaling pathways warrants deeper exploration into specific regulation contexts where they are involved. Addressing disturbances in these pathways may be helpful, particularly for disorders like cancers. Another interesting lead is considering novel SR biases and their implications for RNA splicing, this may help in furthering our understanding of diseases like spinal muscular atrophy and certain neurodegenerative conditions since improper RNA splicing can lead to the accumulation of toxic proteins in neurons (Li et al., 2021). Moreover, targeting novel proline-rich motifs which are key influencers of protein folding and signal transduction through participating in PPIs may be another promising area to explore. The ability to regulate

these PPIs, especially in diseases where these interactions have become aberrant, could be valuable in unlocking new therapeutics (Ball et al, 2005). Moving forward, integrating these novel sequence-based findings with structural analyses is imperative. For instance, cryo-EM can serve to validate in-silico hypothesis and thereby help in understanding the multifaceted nature of protein domains. Better insights into IDR, PLD, and CBD evolution provides a framework for manipulating protein properties to achieve specific functions and design novel therapeutic biomolecules, with greater understanding of the emergence and specific features of these essential regulatory domains more efficient applications of biotechnology can be made.

1.3 Case Studies

For the analysis of the specific case-study data subsets hierarchal paths to *Mammalia* and *Hominidae* orthogroups were the focus; the subsets were created using the *homo sapian* orthologue and all orthogroups containing the hominid orthologue were gathered from the global dataset. Novel domain disappearances were also considered in addition to novel appearances within clades and for novel annotation presence all HOGs were compared to one another. Further, for novel annotation by presence the contribution weight was not necessary to note the annotation in the findings, as it very much limited options, having a 50% or higher percentage difference of the annotation between HOGs sufficed. To compensate however, sensitivity was raised for detecting IDR annotations with either IUPred2a or DisoPred annotations being sufficient as opposed to the more specific scrutiny within the global dataset requiring the presence of both annotations; increased sensitivity resulted in smaller percent differences between HOGs. Only novel presence and novel signatures annotations were noted as the two inspected subsets did not unveil any novel annotations from other IDR categories. Further, due to the high volume of novel CBD and LCRs within these subset, novel gains within *Hominidae* and *Mammalia* HOG clades were focused and reported; CBDs without identical LCRs were also reported for these clades. For multiple-AA novel residue biases mentioned specifically, combinations of the bias were considered as a single multiple-AA bias as performed for the global dataset, with smaller combinations of the bias were absorbed into the larger bias (ie. PSQ, SGP, GSPQ, PQGS, or SPQG as one multiple-AA bias for GSPQ).

1.3.1 Circadian protein subset

1.3.1.1 Circadian protein subset results

The circadian protein subset scrutinized six circadian proteins integral to the mammalian circadian clock and their respective orthologues within different HOGs: CLOCK, Casein kinase I, Cry1, Cry2, HOX9, and BMAL1. A novel gain in both IDR-short and PLD annotations was observed for HOX9 starting from *Mammalia* compared to the upper *Bilateria* and *Gnathostomata* HOG clades respectively; of the two novel annotations, the novel PLD annotation was more significant with an 81% difference in presence between the HOGs. Cry1 observed a novel IDR-long and PLD annotation at *Hominidae* relative to *Chordata* and the rest of *Metazoa* and a novel IDR-short disappearance relative to these clades as well, see Figure 2.7-2.8 for reference. Cry2 observed similarly between *Hominidae* and *Chordata*, however, instead of gaining lost a PLD domain with a difference close to 50% between these clades. Notably, Cry1 and Cry2 share the same *Chordata* and *Metazoa* HOGs, their orthologues being absorbed together from *Chordata* upwards. Another observation of note close to the threshold cut-off is the disappearance of a PLD domain observed for BMAL1 starting from *Gnathostomata* relative to the top *Animalia* HOG at a 43% HOG annotation difference, from 48% of all (197) *Animalia* clade orthologues having the PLD annotation to only 5% of the total orthologues (60) in *Gnathostomata* possessing it and 0% in *Hominidae* (three orthologues) thus having a 48% difference in annotation between *Animalia* and *Hominidae*, see Figure 2.8 for reference. Moreover, BMAL1 experienced a loss of a PLD starting from within *Gnathostomata* compared to the top *Animalia* HOG, and a gain in an IDR-short annotation difference close to the threshold at 42% in *Hominidae* relative to *Mammalia*, please see B in Figure 2.7.

A novel gain in CBD or LCR was only observed within the *Hominidae* clade for Cry1 and Cry2 proteins responsible for repressing CLOCK-BMAL1 mediated transcription of circadian genes (Saini et al., 2015). Cry1 had a novel valine and val-arg (VR) CBD and novel ala-gly (AG) novel CBD with LCR within *Hominidae*, it also observed a novel pro-ser-glu (PSE) and novel trp-arg-phe-cys (WRFC) CBD signature. Within *Mammalia* several novel gains in signatures were found for CBDs with accompanying LCRs: CLOCK observed a gain in proline, glutamine, and serine,

BMAL1 observed a novel gly-ile-pro (GIP) CBD with LCR, Cry1 a novel glycine, arginine, serine, cys-gly-ser (CGS) and gly-arg-pro (GAP) CBD with LCR signature, and HOX9 with a novel ala-pro-ser-gly (APSG) CBD with LCR in several combinations.

1.3.1.2 Circadian protein subset discussion

The novel changes observed within the six circadian proteins may provide valuable insights into how adaptations might have occurred for PPIs, cellular heterogeneity, and even disease states. Generally, the domains within these proteins have been well-conserved as expected due to their paramount importance in regulation, nevertheless, some interesting observations were found. BMAL1, another critical circadian protein, displayed a substantial reduction in PLD annotation in the orthogroup from *Gnathostomata* and further still for mammals and lower clades compared to the top *Animalia* HOG. This change may signify a shift in the mechanism BMAL1 uses within *Gnathostomata* and mammals for localization or interaction with other proteins compared to its orthologues pooled in *Animalia* from other clades. Interestingly, the loss in a PLD was accompanied by a nearly reaching threshold gain in a short IDR within *Hominidae*, perhaps as a compensatory mechanism for the potential loss of LLPS properties. As for HOX9, the novel IDR short and PLD acquired within mammals and hominids may imply an increased role in LLPS, cellular differentiation, and development.

Furthermore, the emergence of new CBDs with LCRs in multiple proteins within the mammalian clade could serve numerous purposes from evolutionary adaptations to environmental stresses to enhancing protein stability. Specifically, in mammals for CLOCK a high number of novel LCRs and CBDs were observed, see Figures 2.9-3.0. Moreover, if these novel biases correspond with IDRs, they may also encompass PTM sites, an adaptability particularly pertinent for circadian protein regulation under varying environmental conditions. The high number of novel emergences of CBDs and LCRs across various clades could indicate evolutionary pressures on the clade to facilitate complex or more specific protein interactions, LLPS, or cellular stress responses and perhaps thereby contribute towards clade-specific refinements for regulatory mechanisms involved with cellular heterogeneity.

Moreover, it is of interest to note the background context within which these novel annotations were identified, here considering relative differences between the number of orthologues or the average protein length between HOGs can provide some context. With a greater difference in HOG size, the percent annotation differences between HOGs can easily become more prominent resulting therefore in the identification of more novel annotations; hence, novel annotations found between HOGs with smaller size differences become more relevant. Moreover, it is also interesting if the average length of the protein sequence between HOGs is more than modestly different; a greater difference in length allows for more probability of novel annotations, equally important a significant difference in average protein sequence length is an interesting evolutionary departure and observance of novelty by itself. From the novel annotations observed in the circadian protein subset, the average protein size for the *Mammalia* HOG for Cry1, where a novel IDR-long and novel PLD annotation was identified relative to chordates, may thus suggest a gain in these novel domains within its increased protein sequence length, please refer to Figure 3.2 for reference. The subtle increase in the average protein length between *Gnathostomata* and *Mammalia* is also concurrent with the novel gain in a PLD and short IDR within mammals for HOX9. Moreover, HOG size is a telling indication of the number of observed novel CBD and or LCR signatures within the HOG, please refer to Figures 2.9-3.2.

Although the study unveiled several intriguing findings for the circadian protein case-study, before embarking in future directions that explore these potential avenues that warrant further investigation a confirmation should be conducted with all orthologues of the protein being investigated since this study is limited by its reliance on the annotations provided by the selected annotation software for analysis which, although kept minimal due to increased sensitivity requiring either IUPred2a or DisoPred annotations for IDR annotation, is open to false positives and false negatives. Nevertheless, these observations open doors to understanding the molecular mechanisms underlying the circadian rhythm's evolutionary dynamics and its potential vulnerabilities to diseases and alterations to environmental changes.

1.3.1.3 Circadian protein subset figures

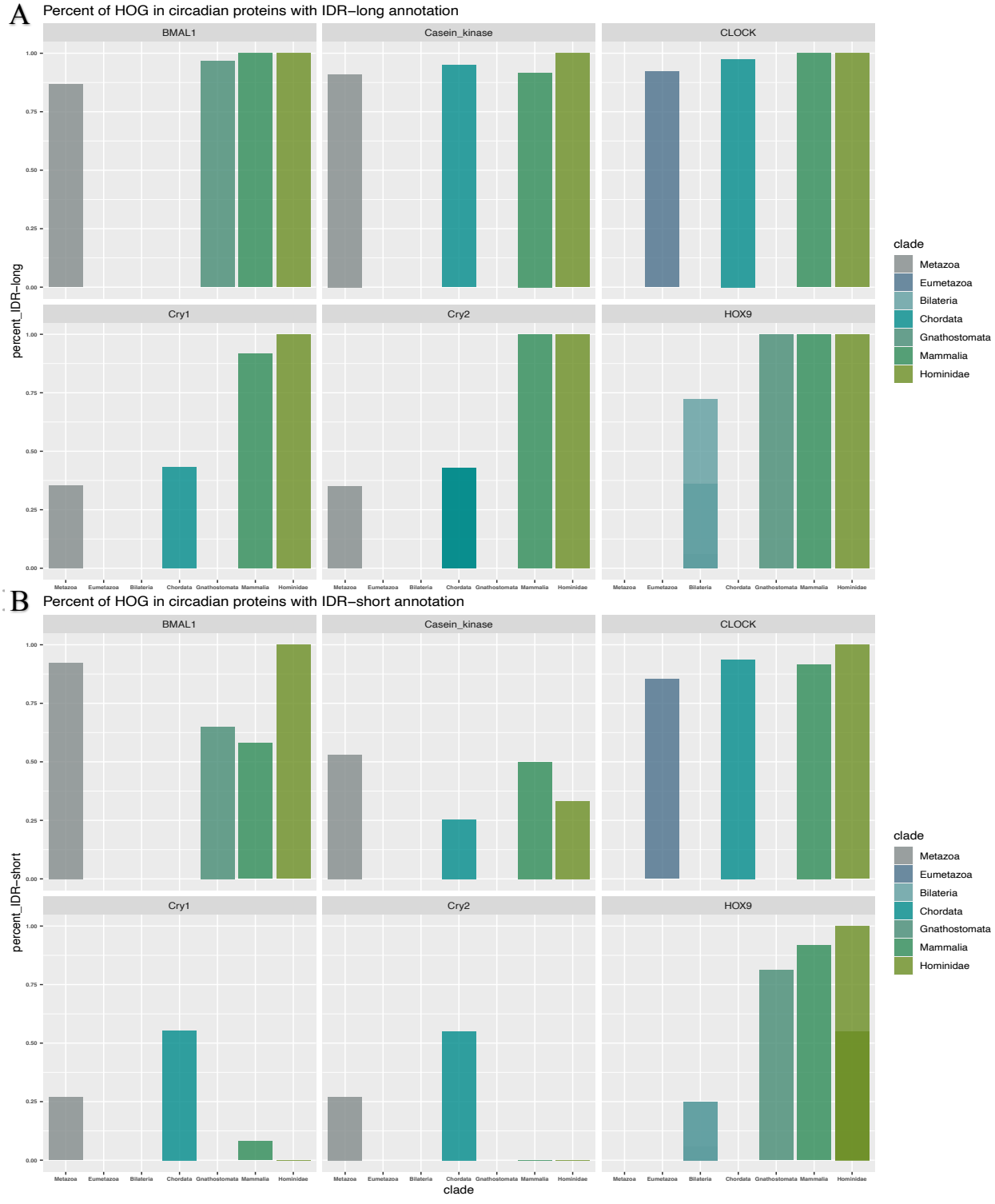

Figure 2.7: IDR-long and IDR-short annotations for circadian protein subset for different HOGs. The UniProt accession ids for the *homo sapien* orthologues of the proteins are: O15516, Q9HCP0, Q49AN0, Q16526, B5DFK3, O00327. Orthogroups are ordered left to right along the x-axis from the highest-ranked HOG to the lowest.

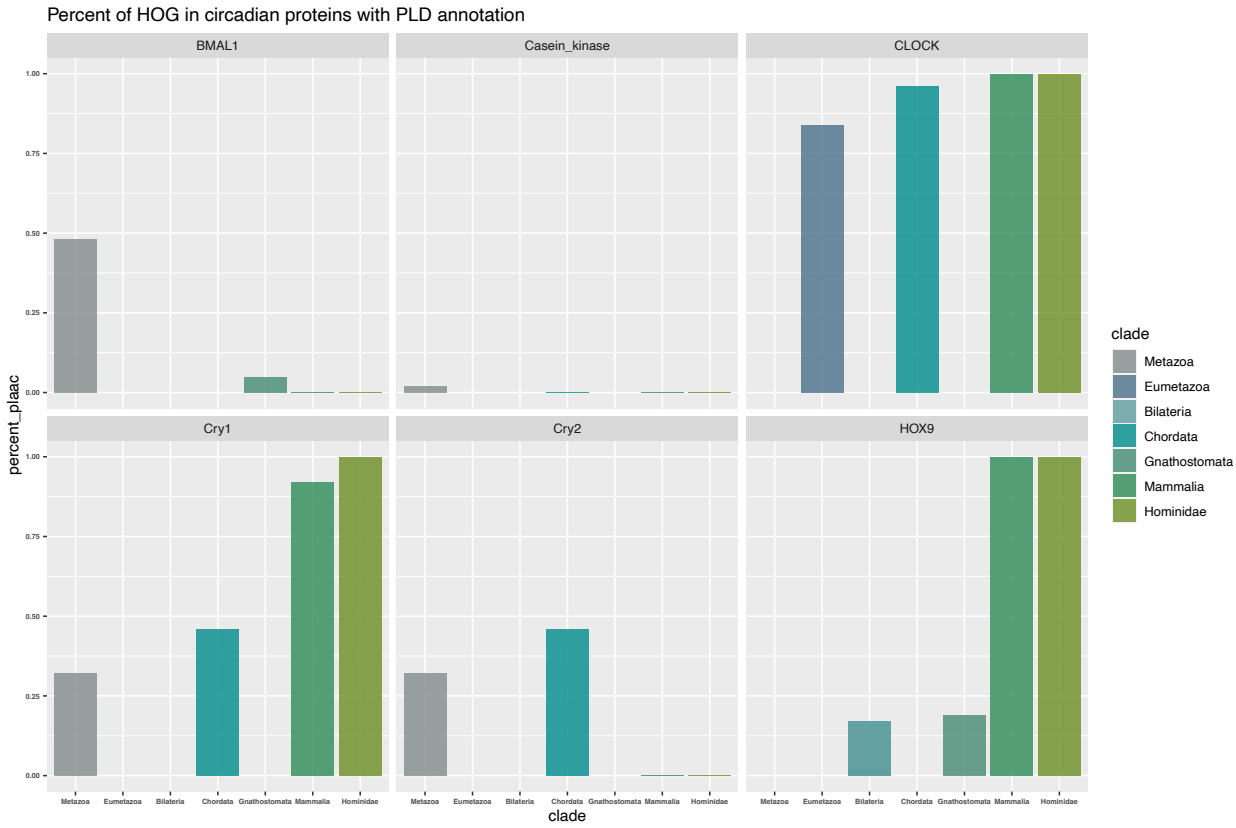

Figure 2.8: PLD annotations for circadian protein subset for different HOGs.

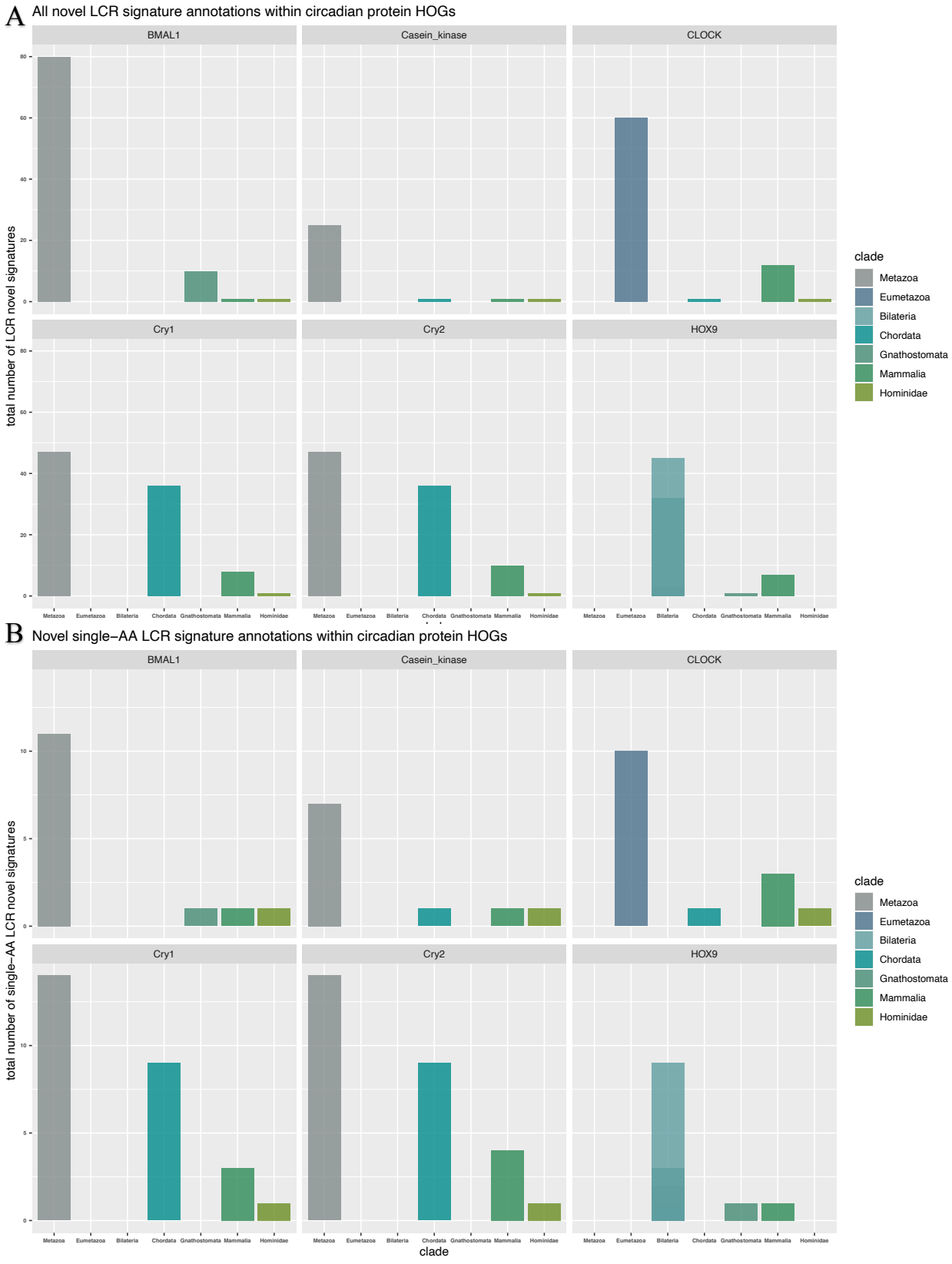

Figure 2.9: Novel LCR annotations for circadian protein subset for different HOGs.

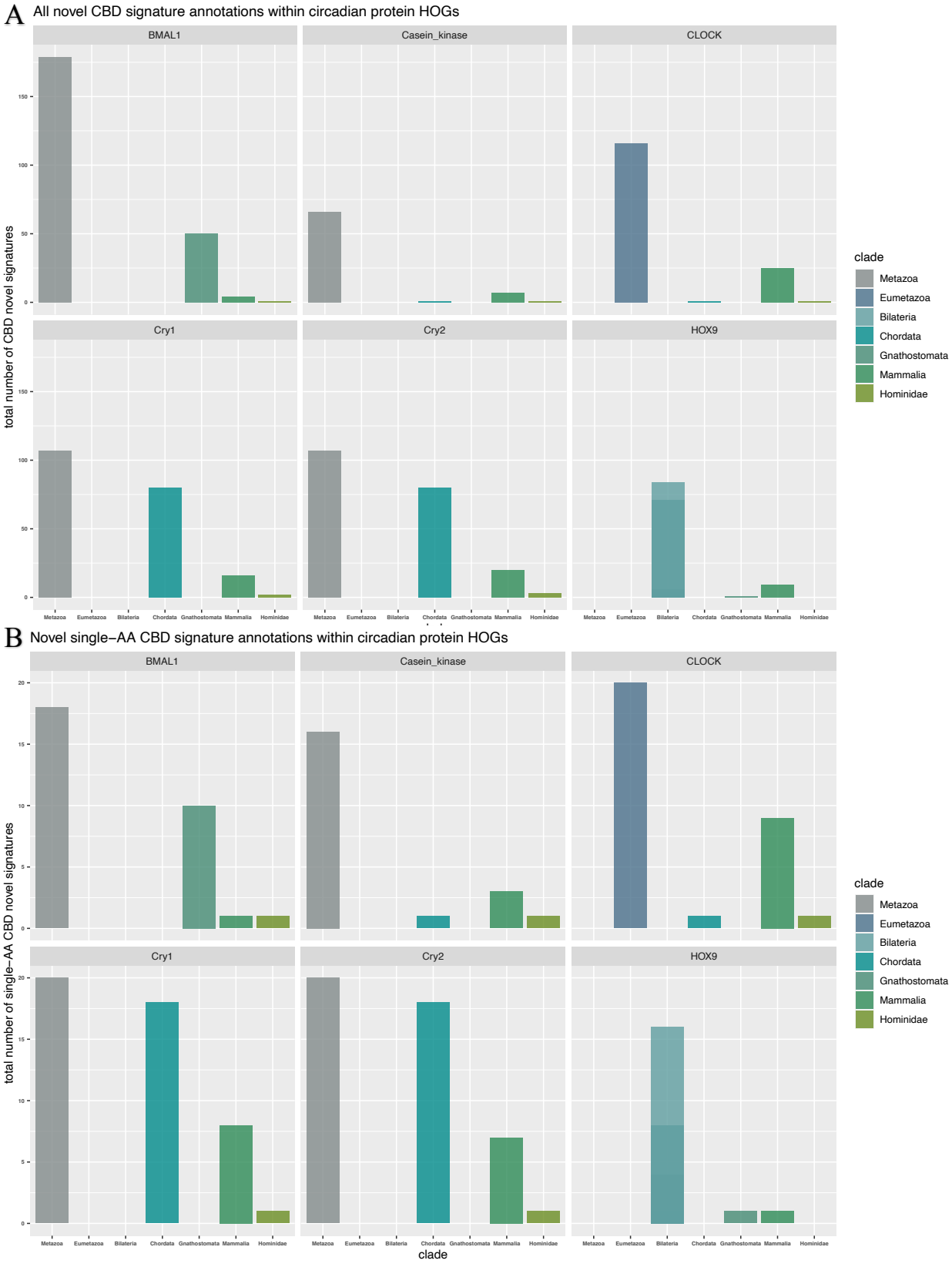

Figure 3.0: Novel CBD annotations for circadian protein subset for different HOGs.

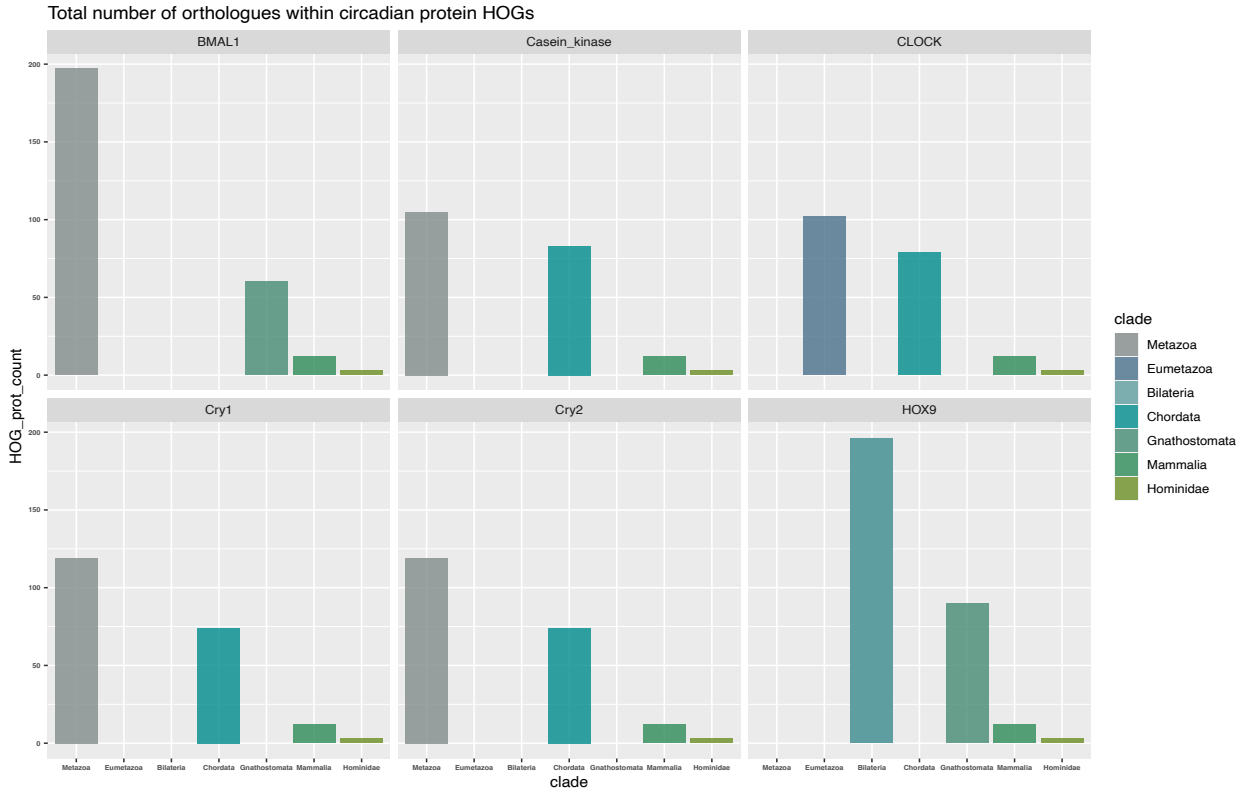

Figure 3.1: Number of orthologues within circadian protein HOGs.

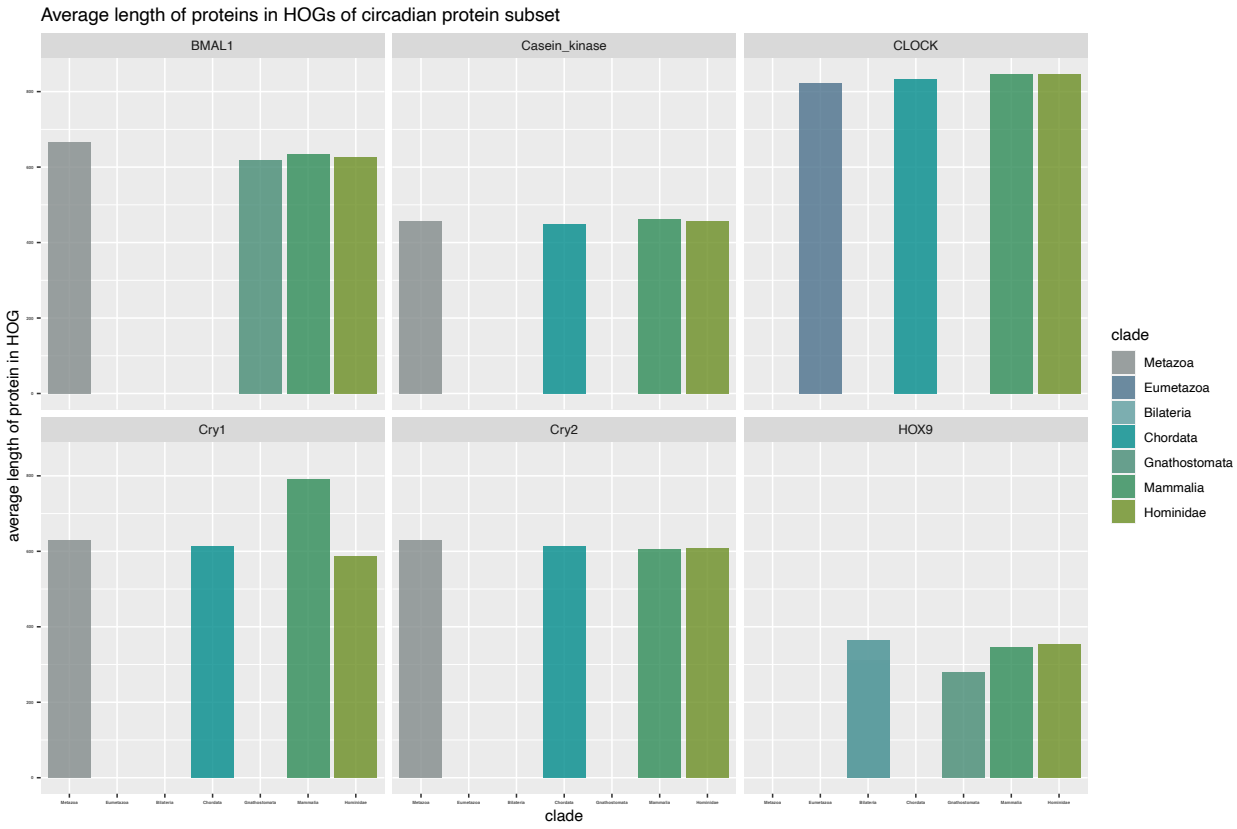

Figure 3.2: Average protein length of different HOGs in circadian protein subset.

1.3.2 Disease-linked IDP subset

1.3.2.1 Disease-linked protein subset results

The disease linked subset of IDPs consists of 17 proteins and their orthologues: APO-E, APP, BCL2, β-secretase, BRCA1, CHCHD10, FUS, MAP tau, NUPR1, p53, PAX5, PrP, PSEN1, PSEN2, SYNCAIP (α-synuclein), TDP-43, and Gsk3a. There were only a few novel IDR and PLD annotation appearances and disappearances observed within this subset, however, they were coupled with the observance of several novel signature annotations. In PSEN1 and PSEN2 within *Chordata* relative to *Metazoa*, with the proteins sharing both HOGs, there was a novel IDR-long gain, please see Figure 3.2. Additionally, Gsk3a experienced a novel PLD gain from *Mammlia* onwards whereas p53 observed a novel disappearance of a PLD in *Mammalia* onwards relative to its *Chordata* and *Animalia* HOGs, please see Figure 3.4. Moreover, within *Hominidae,* relative to mammals, BCL2 experienced a short IDR gain, whereas CHCHD10 experiences a near threshold short IDR loss.

As for novel LCRs, within *Hominidae* there is a gain in a novel phenylalanine LCR in p53, pro-ala (PA) in BCL2, threonine in Gsk3a, ser-gly-asn-gln (SGNQ) for TDP-43, and arg-gln (RQ) for BRCA1. There were also several novel multiple-AA residue gains in CBDs not accompanied with LCRs within *Hominidae*, including a novel gain in glu-pro-gly-cys (EPGC) and val-leu-tyr-met (VLYM) in PSEN2, pro-ser-trp-gly (PSWG) in BCL2, arg-gly (RG) in NUPR1, and variations of pro-lys-gly-ser (PKGS) for tau. For novel CBD gains within *Mammalia,* there were several novel single-AA CBD with LCRs including, alanine for PAX5, glutamic acid and phenylalanine for PSEN1, alanine and proline for CHCHD10, proline for NUPR1, asparagine and serine for α-synuclein (SNCAIP), aspartic acid, glutamic acid, arginine and threonine for APP, threonine for PrP, glycine, glutamine, arginine, and serine for RBP FUS, and proline for BCL2. In addition, there are several cases of either novel LCR or novel CBD gains separately, for instance, for microtubule associated protein (MAP) tau there is a gain in an aspartic acid novel CBD signature and a gain in glycine, histidine, and leucine LCR signatures. Moreover, there were several occurrences of novel multiple-AA biases encompassing both LCR and CBD signatures including ala-pro-tyr (APY) for PAX5, glu-gln (EQ) and gln-ser (QS) for PSEN1, ala-pro-gly (APG) and pro-gln (PQ) for CHCHD10,

cys-arg (CR), gly-pro (GP) ,and arg-pro (RP) for β-secretase, pro-ser (PS) for α-synuclein, asp-thr-glu-val (DTEV) for APP, gly-trp (GW) for PrP, pro-ser-gly (PSG) for MAP tau, arg-glu-gln-trp (REQW) for Apo-E, gly-ser-gln (GSQ) for RBP FUS, pro-cys (PC) for BCL2, asn-gln (NQ) for BRCA1, and gly-ser-asn-gln (GSNQ) for TDP-43.

1.3.2.2 Disease-linked protein subset discussion

Several curious alterations within the IDR, PLD, and CBD protein landscape of the 17 disease-linked proteins analysed were found. Unsurprisingly, there was a great deal of conservation, particularly for IDRs and PLDs, and predominantly the novel observances came from novel gains in CBD or LCR signatures. Within this subset there was one IDR-long gain and one IDR-short gain, in the former case within *Chordata* observed for PSEN1 and PSEN2 which begin sharing higher-ranked HOGs starting from chordates and in the latter case for BCL2 within *Hominidae*. As discussed extensively, due to the central participation of IDRs in signalling transduction and PPIs, these annotations may imply evolving functionalities that warrant further investigation. Moreover, increase in disorder within chordates for PSEN1 and PSEN2 relative to *Metazoa* opens doors to investigating possible altered protein behavior and functionalities from the increase in disorder and possible implications for neurodegenerative pathophysiology. Even more so, the absence of a PLD in p53 within the mammalian clade and further down within hominids was unexpected as the loss of a PLD may result in losing functional contribution towards phase separation, aggregation, or signaling cascades (King et al., 2012). It is tempting however to speculate plausible benefits to the clade as the absence may be resultant of evolutionary pressures to suppress certain interactions or refine the protein for more species-specific functions.

As for novel signatures found within the disease-linked proteins, they are likely consequential of lineage-specific adaptations in the proteins for evolving protein functionalities. Novel LCR and CBD signatures were especially abundantly noted within mammals, particularly novel single-AA CBD and LCRs relative to their novel observance in other respective HOGs, specifically for α-syncuclein (SYNCAIP), Gsk3a, and PSEN1 and PSEN2, see Figure 3.5-3.6. Single-AA LCRs

are especially compelling areas to explore for future directions as their novel observance points towards an AA signature novelty usually not observed within even multiple-AA LCR biases of the upper HOG, moreover, they potentially point towards clade-specific regulatory roles as LCRs are well established for playing essential roles in LLPS and the formation of membrane-less organelles as previously discussed (Boeynaems et al., 2018).

Regarding specific novel signatures, in *Hominidae*, the appearance of a novel phenylalanine LCR in p53 could allow for new PPIs or allow cell membrane attachment to facilitate clade-specific adaptations to some of the numerous biological functions p53 is involved with ranging from cell cycle control to apoptosis. Similarly, a novel gain in the PA signature in BCL2 within hominids may affect protein folding or stability allowing interactions with other proteins due to the regionally stabilizing effect of proline as earlier discussed. Furthermore, the novel gain in a glutamic acid signature for PSEN1 within *Mammalia* may indicate a novel functional capacity; incorporating the negatively charged AA signature may well change the electrostatic properties of PSEN1 and considering acidic regions often engage in ionic interactions with basic regions on other proteins the bias may regionally affect protein interactions and functions. Additionally, negatively charged LCRs could influence the subcellular localization of the protein by interacting with positively charged ions or other positively charged molecules or proteins within the cell. Importantly, novel acidic LCR gains may also have functional relevance in regulating propensity for protein aggregation; their negative charge may mitigate or exacerbate the rate of aggregation (Mavadat, 2023). Beyond this, NQ LCRs are also particularly interesting for several aforementioned reasons: they are associated with prion-like behavior, potentially contribute towards phase transition, involved with the formation of either functional or pathological aggregates (Boeynaems et al., 2018; Malinovska et al., 2013).

Moreover, considering the background context of the HOGs within which these novel annotations are identified is important to realize the extent of their significance. In this case-study there seemed to be much less overlap between the difference in the increase of average protein sequence length within the HOGs and observance of novel annotations, see Figure 3.8 for the average protein lengths of different HOGs. Although, more modest differences in

average protein length were absorbed as almost non-existing due to the large averages of other HOGs within the figure, nevertheless, some interesting and relevant observations can be made. There is a decrease in average HOG protein sequence length for p53 in mammals and onwards, this observance is interesting coupled with the novel loss of a PLD within mammals as well for p53. Moreover, in cases where a novel annotation resides within an orthogroup and is novel relative to a HOG which is near it in the number of orthologues it contains, for instance *Mammalia* usually with twelve orthologues and *Hominidae* with three orthologues resulting in only a difference of nine orthologues, it makes the novel annotation more significant as was the case for the short IDR gain in BCL2 between these clades, see Figure 3.7 for reference. Moreover, unexpectedly the novel CBDs or LCRs observed within the disease-linked subset were much less correlated with relative HOG size or relative average HOG protein sequence length in contrast to the circadian protein case-study; there were a significant number of novel CBD and LCR annotations found within *Chordata* and *Mammalia* clades of several of the proteins relative to other HOGs despite the clades smaller size relative to higher-ranked OGs.

Despite various interesting curiosities emerging from this case-study, it is crucial to underscore the broader implications of these observations requires further experimental validation and biomolecular studies to test putative functional insights. While the analysis provides a rich framework for understanding the evolutionary landscape of these protein domains in the context of IDRs, PLDs, and CBDs, it is limited by its dependency on accurate annotations, nevertheless it serves as an invitation for future studies to delve deeper into more uncharted territories of protein domain evolution, particularly with a focus on proteins acting in a regulatory capacity.

1.3.2.3 Disease-linked protein subset figures

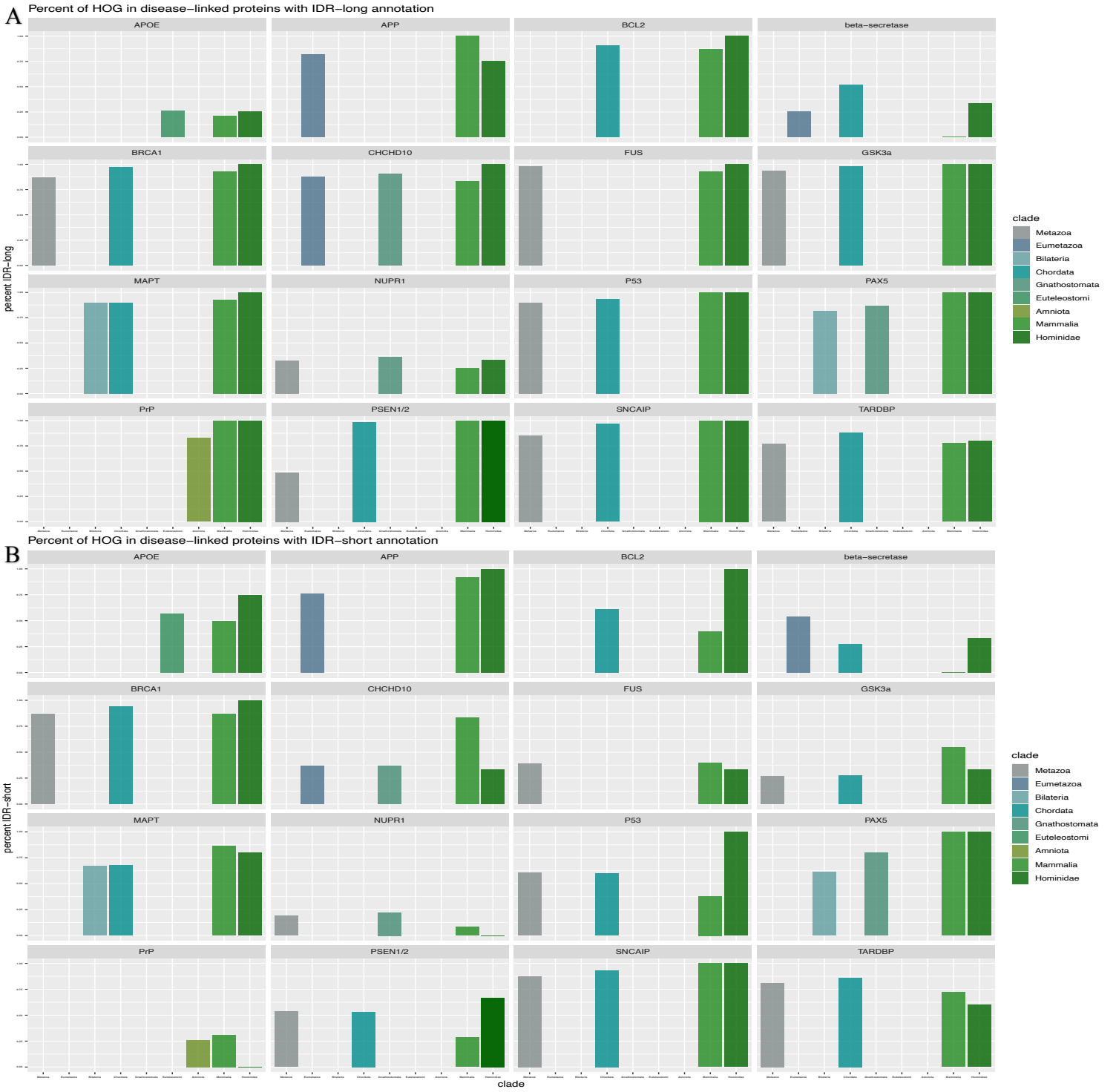

Figure 3.3: IDR-long and IDR-short annotations for disease-linked protein subset for different HOGs. Orthogroups are ordered left to right along the x-axis from the highest-ranked HOG to the lowest-ranked. The UniProt accession ids for the *homo sapian* orthologue of the proteins are **APP**: P05067; **MAPT**: P10636, **PSEN1**: P49768; **PSEN2**: P49810l; **APOE**: P02649; **GSK3A**: P49840; **beta-secretase 1**: P56817; **NUPR1**: O60356**; PAX5**: Q02548; **BCL2**: Q12983; **SNCAIP**: Q9Y6H5; **p53**: P04637; **BRCA1**: P38398; **PrP:** P04156; **TARDBP**: Q13148; **CHCHD10**: Q8WYQ3; **FUS**: P35637. Note PSEN1/2 are combined for shared HOGs of higher-ranked clades, *Chordata* and *Metazoa,* and no novel observances within their lower ranked individual orthogroups.

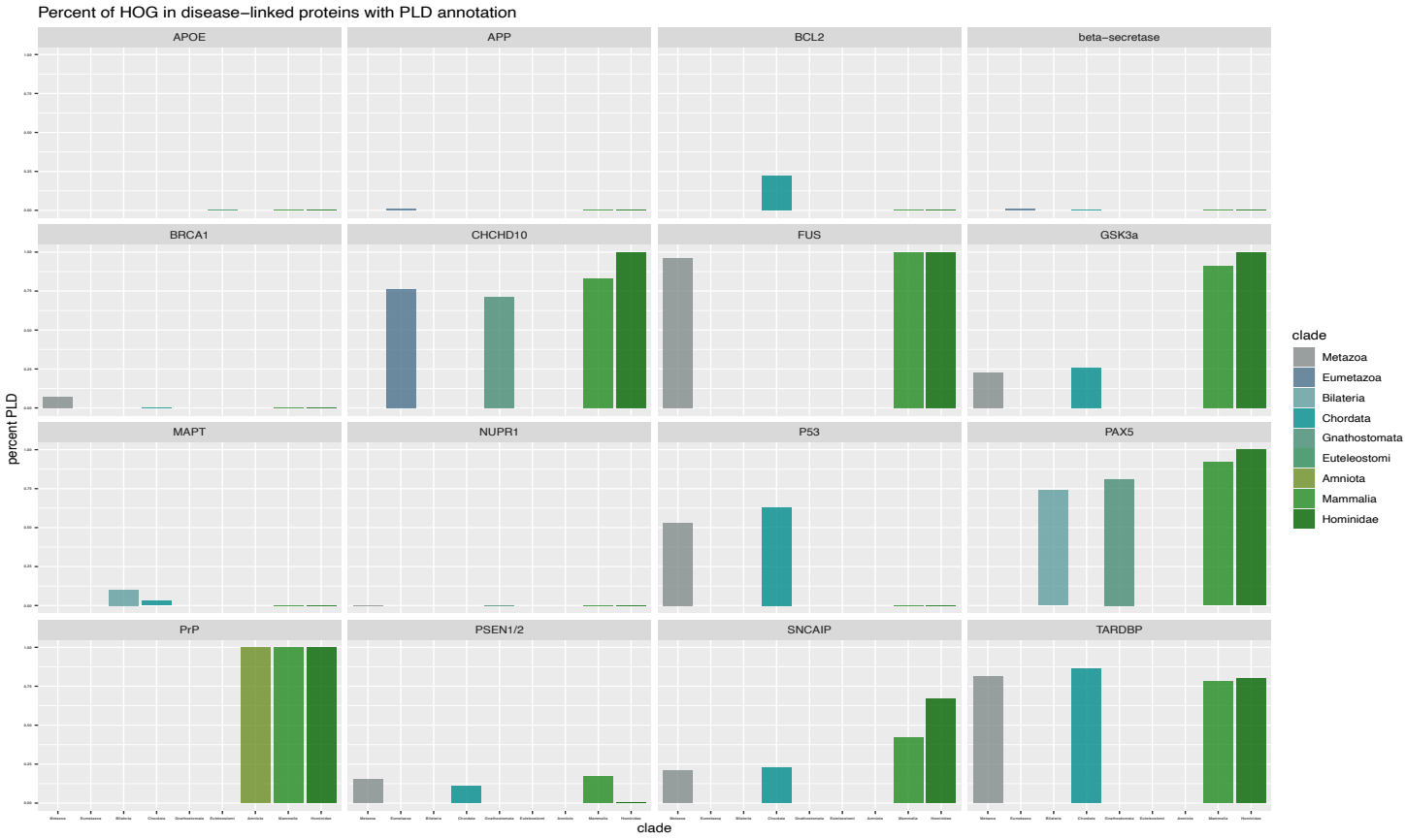

Figure 3.4: PLD annotations for disease-linked proteins for different HOGs.

Figure 3.5: Novel LCR annotations for disease-linked protein subset for different HOGs.

Figure 3.6: Novel CBD annotations for disease-linked protein subset for different HOGs.

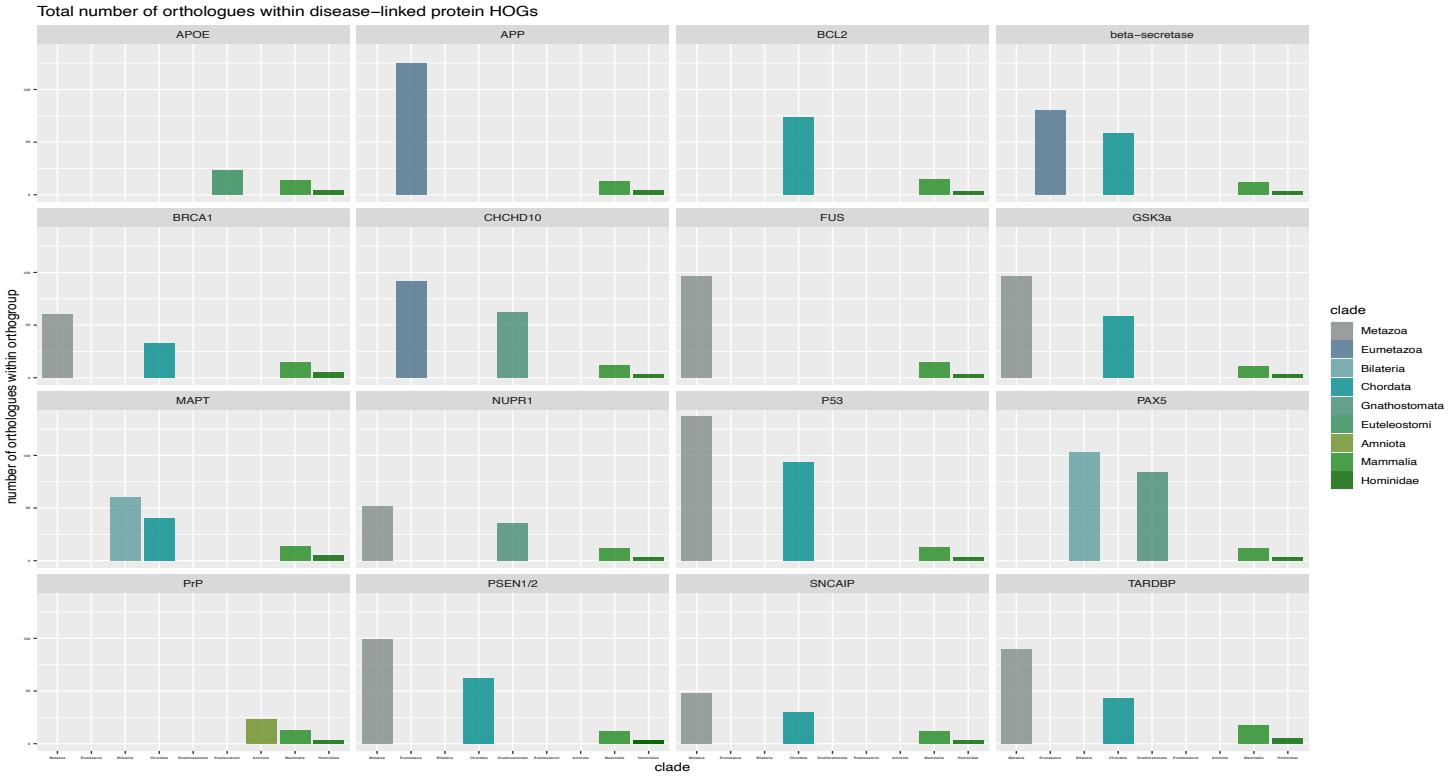

Figure 3.7: Number of orthologues within disease-linked protein HOGs.

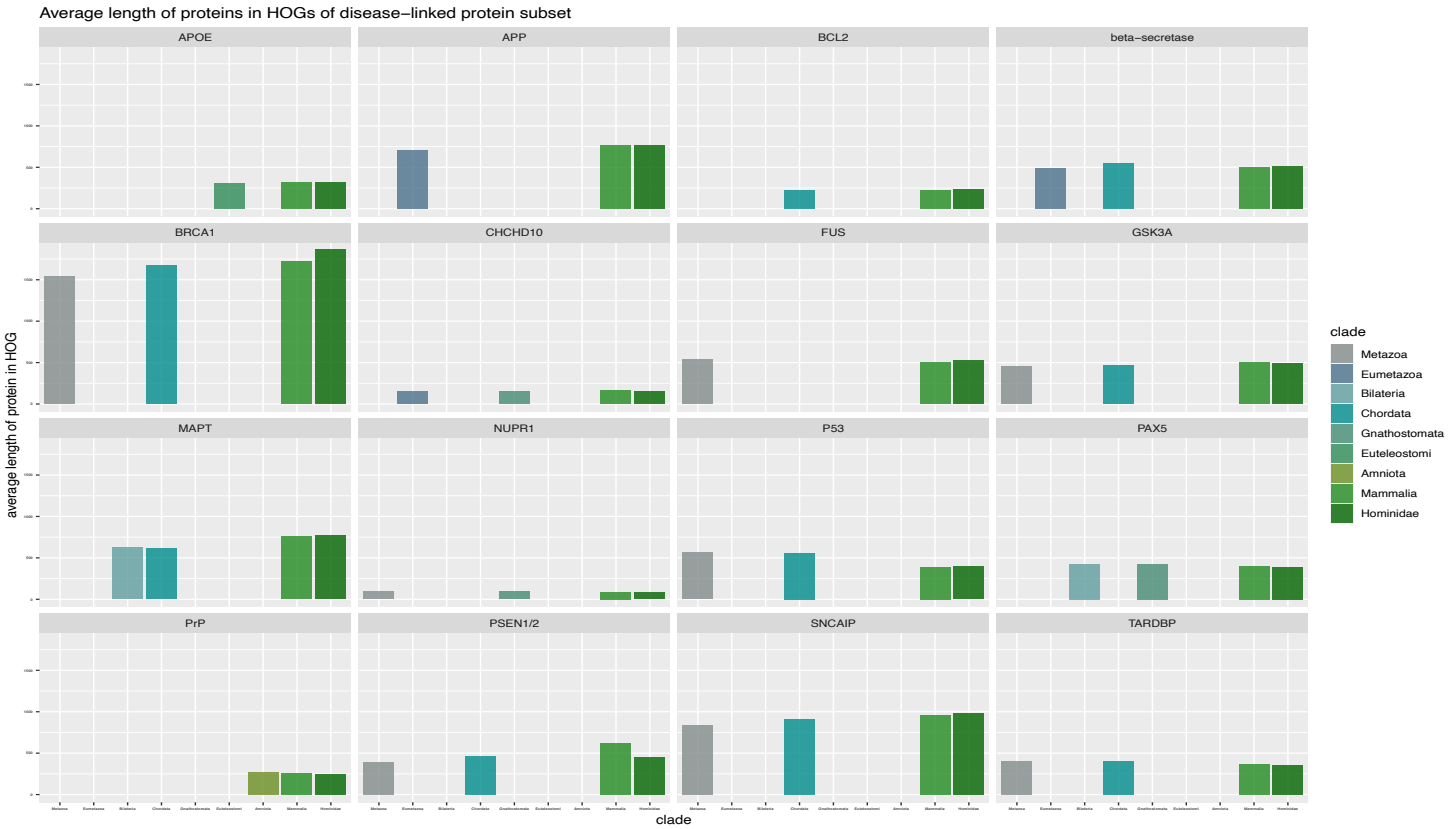

Figure 3.8: Average protein length for different HOGs of disease-linked protein subset.

1.4 Conclusion

The exhaustive investigation into the dramatic alterations within IDRs and PLDs and both the dramatic and subtle alterations within CBD annotations has identified various orthogroups containing novel domains and providing a basis for intriguing insights into the multifaceted role and adaptability of these domains for specialized protein functions across different clades in the animal kingdom. The study unveiled 365 orthogroups with novel short IDRs and 1 233 orthogroups containing novel long IDRs. Of the novel IDR annotations, the breakdown was as follows for the different novel classifications: 571 orthogroups were identified harboring a novel annotation for novel presence, 572 for novel count, and 435 for novel length which only considered long IDRs. The differential roles of these novel IDRs stand as a testament to their functional diversity in cellular biology. Long IDRs, given their extended length, frequently serve as hotspots for post-translational modifications (PTMs), thus playing a central role in orchestrating intricate cellular signaling pathways and mechanisms, whereas shorter IDRs, are frequently employed in protein-protein interactions (PPIs), facilitating interactions for cellular processes that require rapid and specific association and dissociation of protein complexes and binding partners. The exploration into PLD annotations in the different orthogroups of the animal kingdom highlighted the identification of 250 orthogroups with novel PLD annotations. This is particularly interesting given the inherent capability of PLDs to undergo phase separation and form membrane-less organelles underscoring their critical role in regulating cellular functions; within the larger biological context they are especially important for regulating development, differentiation, homeostasis, and ensuring efficient immune responses among various other processes (Das et al., 2018).

As for single-AA CBDs with LCRs, 35 569 orthogroups with novel signatures were identified. emphasizing the vast adaptive potential and evolutionary trajectory across different clades, with *Mammalia* being notably dominant for this category of novel annotations. Novel signatures of key residues, for instance, proline with its inherent structural rigidity or serine due to its pivotal role in phosphorylation and signal transduction, showcase the variety and depth of the potential functional implications harbored within these novel CBDs. When these novel

signatures are considered for the specialized potential functional roles of their residues, particularly in multiple-AA biases or within LCR signatures, a more comprehensive understanding of the evolutionary forces sculpting protein folding and protein structures also emerges. Moreover, a clear overlap was observed between novel LCRs and PLDs, the complementation of these domains provides stronger support for their putative molecular role in regulating critical cellular functions.

From the case study data subsets for circadian proteins and disease-linked proteins some intriguing novel annotations were also observed. Notably, within the circadian protein subset a PLD was found within *Mammalia* for Cry1 and HOX9; also, within mammals Cry1 gained a long IDR and HOX9 a short IDR. From the disease-linked dataset Gsk3a gained a novel PLD annotation within *Mammalia*, and p53 lost a PLD from *Mammalia* onwards. BCL2 also observed a short IDR gain within *Hominidae.* Accompanying these novel annotations was a gain in several single-AA and multiple-AA CBDs with LCRs observed for almost all proteins in both case studies. The possible implications of these novel annotations likely point towards lineage-specific adaptations of protein functions to accommodate more specific PPIs, protein-molecule interactions, increased protein stability, or otherwise alterations resulting in the improved regulatory capacity of the protein to respond more efficiently to the environmental stresses pertinent to the clade.

Beyond this, the identification of novel IDRs, PLDs, and CBDs provides promising directions for future research paving the way for scientific inquiries in several avenues of the life sciences, particularly within physiology and the intricate regulation of signalling pathways. For instance, investigating if novel long IDR annotations act as additional host sites for PTMs thereby providing added layers of regulatory control, particularly in the case of novel IDRs by length, is an interesting avenue for further study. Moreover, future research should involve experimental validation of the putative functions of these novel annotations, particularly the influence on PPI networks; it would be interesting to test if these novel annotations are facilitating new PPIs or disrupting the affinity of existing interactions. This may be achieved with the use of specific

binding assays to assess how these novel domains influence binding affinities with different protein or molecular partners or through targeted mutagenesis analysis using gene-editing techniques like CRISPR/Cas9 to investigate how these novel domains modulate protein function and cellular processes *in vivo*. Additionally, *in vivo* studies allow for investigating the putative stress response roles of the proteins containing the novel annotation, this is relevant since IDRs and PLDs often are heavily involved with stress response.

Furthermore, domain disappearances are also of interest, for instance, from the disease linked IDP subset investigating if the loss of the PLD in p53 within mammals has effect on cancer or neurodegenerative disease pathophysiology in mammals relative to other chordates where it is present. And perhaps the novel gain in PLDs warrants even more attention due to the potential role of the novel domain in propensity for phase separation which may be tested through phase separation assays. For instance, from the circadian protein case-study, investigating the role of the novel PLD in Cry1 for mammals and its affect in regulating circadian rhythms relative to its orthologues within chordates may expand understanding of the possible functional role of the novel PLD in Cry1 emergent within mammals.

The emergence of novel IDRs, PLDs, and CBDs in general potentially point towards specific evolutionary pressures favoring the development of these proteins for specialized specific functions and adaptations within and unique to the clade. This is particularly noteworthy in the abundance of novel CBDs with LCRs observed within mammals which may reflect an adaptability that facilitates molecular interactions that contribute to the range of features unique to mammals including but not exclusive to metabolic optimization, advanced cognitive function, complex social behavior, extended lifespan, or adaptation to specific environmental niches that mammals occupy. For example, a novel gain in a CBD with a LCR within circadian proteins might indicate an evolutionary adaptive strategy for survival by optimizing timekeeping mechanisms within the clade facing specific environmental pressures, this may potentially be achieved via increasing the affinity to binding partners or gaining affinity to new ligands, or even increased precision in regulation by the addition of more PTM sites.

Lastly, it is important to note while this study has provided various insights and opened the door to several avenues of intriguing further research, the study harbors limitations primarily due to its heavy reliance on the accurate annotation of the domains of interest. False positives and false negatives from the inherent shortcomings of the used annotation software, although mitigated as much as possible, and the potential caveats associated with a purely sequence-based evaluation is a drawback and limitation to the study, further research towards experimental validation will support the sequence-dependant analysis. Coupling sequence-based analysis with experimental validation in the future for specific proteins of interest should therefore be prioritized. Nevertheless, this comprehensive exploration into the world of IDR, PLD, and CBD evolution of proteins within animals and its potential functional implications has indicated several fascinating evolutionary trajectories and adaptive strategies through the development of specialized functions within proteins in different clades and set the stage to gain a more profound appreciation and understanding of the intricacies of regulation. These insights present a sturdy foundation for future directions in the world of Bioinfomatics and molecular and cellular Biology for expanding knowledge on the intricate mechanisms involved with the precise regulation of a plethora of biological processes, potential therapeutic innovations, and expanding upon this to investigate the evolutionary trajectories of protein domains of other kingdoms.

References

1. Akopian, D., Shen, K., Zhang, X., & Shan, S. O. (2013). Signal recognition particle: an essential protein-targeting machine. Annual review of biochemistry, 82, 693-721. DOI: 10.1146/annurev-biochem-072711-164732

2. Alberti, S., Gladfelter, A., & Mittag, T. (2019). Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. Cell, 176(3), 419-434.

3. Alberti, S., Halfmann, R., King, O., Kapila, A., & Lindquist, S. (2009). A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. Cell, 137(1), 146-158. doi:10.1016/j.cell.2009.02.044

4. An, L., & Harrison, P. M. (2016). The evolutionary scope and neurological disease linkage of yeast-prion-like proteins in humans. *Biology Direct*, *11*(1), 32. https://doi.org/10.1186/s13062-016-0134-5/

5. An, L., & Harrison, S. C. (2016). The evolutionary scope of prion phenomena. PLoS One, 11(12), e0166751.

6. Anastassopoulou, C., Fuchs, J., & Kortemme, T. (2020). Protein intrinsic disorder in viral replication: Emerging antiviral drug targets. Current Opinion in Virology, 44, 107-119.

7. Ball LJ, Kühne R, Schneider-Mergener J, Oschkinat H. (2005). Recognition of proline-rich motifs by protein-protein-interaction domains. Angew Chem Int Ed Engl. 44(19), 2852-69. DOI: 10.1002/anie.200462121

8. Banani, S. F., Lee, H. O., Hyman, A. A., & Rosen, M. K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nature Reviews Molecular Cell Biology, 18(5), 285-298. https://doi.org/10.1038/nrm.2017.7

9. Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B. J., Boone, C., Bader, G. D., Myers, C. L., & Kim, P. M. (2011). Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biology*, *12*(2), R14. https://doi.org/10.1186/gb-2011-12-2-r14

10. Belshaw, R., & Jiggins, F. M. (2009). Comprehensive phylogenetic analysis reveals conserved evolutionary trajectories in the eukaryotic armorome. Proceedings of the National Academy of Sciences, 106(43), 17610-17615. doi: 10.1073/pnas.0907540106

11. Blaustein, J. B., Oballe, A. J., & Robinson, A. S. (2005). Phosphorylation-mediated regulation of alternative splicing in cancer. International journal of cell biology, 2005(4), 135-143.

12. Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., ... & Darzacq, X. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. Nature structural & molecular biology, 25(9), 833-840.

13. Boeynaems, S., Alberti, S., Fawzi, N. L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L., Tompa, P., & Fuxreiter, M. (2018). Protein phase separation: A new phase in cell biology. Trends in cell biology, 28(6), 420-435.

14. Boothby, T. C., Tapia, H., Brozena, A. H., Piszkiewicz, S., Smith, A. E., Giovannini, I., ... & Goldstein, B. (2020). Tardigrades Use Intrinsically Disordered Proteins to Survive Desiccation. Molecular cell, 80(5), 876-887.

15. Boschetti, C., Carr, A., Crisp, A., Eyres, I., Wang-Koh, Y., Lubzens, E., . . . Tunnacliffe, A. (2012). Biochemical diversification through foreign gene expression in bdelloid rotifers. PLoS Genetics, 8(11), e1003035. doi: 10.1371/journal.pgen.1003035

16. Bullock, A. N., & Fersht, A. R. (2001). Rescuing the function of mutant p53. Nature Reviews Cancer, 1(1), 68-76. https://doi.org/10.1038/35094077

17. Buratti, E., & Baralle, F. E. (2008). Multiple roles of TDP-43 in gene expression, splicing regulation, and human disease. Frontiers in bioscience, 13, 867-878.

18. Cascarina, S. M., & Ross, E. D. (2014). Yeast Prions and Human Prion-like Proteins: Sequence Features and Prediction Methods. Cellular and Molecular Life Sciences, 71(9), 1737-1755. doi:10.1007/s00018-013-1515-9

19. Chakrabortee, S., Kayatekin, C., Newby, G. A., Mendillo, M. L., Lancaster, A., & Lindquist, S. (2016). Luminidependens (LD) is an Arabidopsis protein with prion behavior. Proceedings of the National Academy of Sciences, 113(21), 6065-6070

20. Chang, C. C., & Chen, Y. J. (2019). Tau is essential in cancer proliferation and tumor suppression: Implication on its dual role in cancer and Alzheimer's disease. Biomolecules, 9(12), 779. https://doi.org/10.3390/biom9120779

21. Chen, H., Xue, L., Huang, H., Yao, X., & Guo, L. (2019). Structural insights into RNA recognition by the prion-like domain of TDP-43. Science China Life Sciences, 62(3), 347-356.

22. Cortese, M. S., Uversky, V. N., & Dunker, A. K. (2020). Intrinsic disorder in scaffold proteins: getting more from less. Progress in biophysics and molecular biology, 150, 45-56. doi: 10.1016/j.pbiomolbio.2019.08.006

23. Csizmok V, Felli IC, Tompa P, Banci L. (2008). Intrinsically disordered proteins at the interface of structural and regulatory biology. Chem Rev. 118(3):1169-1206. doi: 10.1021/cr500241e.

24. Daskalov, A., Gantner, M., Wälti, M. A., Schmidlin, T., Chi, C. N., Wasmer, C., Schütz, A., Ceschin, J., Clavé, C., Cescau, S., Meier, B., Riek, R., & Saupe, S. J. (2014). Contribution of Specific Residues of the β-Solenoid Fold to HET-s Prion Function, Amyloid Structure and Stability. *PLoS Pathogens*, *10*(6). https://doi.org/10.1371/journal.ppat.1004158

25. Das, R. K., Huang, Y. H., Basu, S., Verdine, G. L., & Shivashankar, G. V. (2018). Formation of protein phase transitions and membrane-less organelles by prion-like proteins in cells. Trends in biochemical sciences, 43(10), 818-830.

26. Dass, R., Mulder, F. A. A., & Nielsen, J. T. (2020). ODiNPred: comprehensive prediction of protein order and disorder. *Scientific Reports*, *10*(1), 14780. https://doi.org/10.1038/s41598-020-71716-1

27. David M Emms, Steven Kelly, Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench, *Genome Biology and Evolution*, Volume 12, Issue 12, December 2020, Pages 2258–2266, https://doi.org/10.1093/gbe/evaa211

28. Emms, D.M., Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16, 157 (2015). https://doi.org/10.1186/s13059-015-0721-2

29. Erdős, G., & Dosztányi, Z. (2020). Analyzing Protein Disorder with IUPred2A. *Current Protocols in Bioinformatics*, *70*(1), e99. https://doi.org/10.1002/cpbi.99

30. Fan, Y., Nikitina, T., Zhao, J., Fleury, T.J., Bhattacharyya, R., Bouhassira, E.E., Stein, A., Woodcock, C.L., and Skoultchi, A.I. (2005). Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. Cell 123, 1199–1212.

31. Gao, C., Ma, C., Wang, H., Zhong, H., Zang, J., Zhong, R., He, F., & Yang, D. (2021). Intrinsic disorder in protein domains contributes to both organism complexity and clade-specific functions. *Scientific Reports*, *11*, 2985. https://doi.org/10.1038/s41598-021-82656-9

32. Haber-Pohlmeier, S., Abarca-Heidemann, K., Körschen, H. G., Dhiman, H. K., Heberle, J., Schwalbe, H., Klein-Seetharaman, J., Kaupp, U. B., & Pohlmeier, A. (2007). Binding of Ca2+ to Glutamic Acid-Rich Polypeptides from the Rod Outer Segment. *Biophysical Journal*, *92*(9), 3207–3214. https://doi.org/10.1529/biophysj.106.094847

33. Hanson, J., Yang, Y., Paliwal, K., & Zhou, Y. (2017). Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, *33*(5), 685–692. https://doi.org/10.1093/bioinformatics/btw678

34. Hanson, J., Paliwal, K. K., Litfin, T., & Zhou, Y. (2019). SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics, Proteomics & Bioinformatics*, *17*(6), 645–656. https://doi.org/10.1016/j.gpb.2019.01.004

35. Harbi, D., Parthiban, M., Gendoo, D., Ehsani, S., Kumar, M., Schmitt-Ulms, G., Sowdhamini, R., & Harrison, P. (2011). PrionHome: A Database of Prions and Other Sequences Relevant to Prion Phenomena. *PloS One*, *7*. https://doi.org/10.1371/journal.pone.0031785

36. Harrison, A. F., & Shorter, J. (2017). RNA-binding proteins with prion-like domains in ALS and FTLD-U. Prion, 11(4), 251-263.

37. Harrison, P. M. (2021). fLPS 2.0: rapid annotation of compositionally biased regions in biological sequences. *PeerJ*, *9*, e12363. https://doi.org/10.7717/peerj.12363

38. Hsin, J. P., & Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes & Development, 26(19), 2119-2137. https://doi.org/10.1101/gad.200303.112

39. Hu, G., Katuwawala, A., Wang, K. *et al.* flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat Commun* 12, 4438 (2021). https://doi.org/10.1038/s41467-021-24773-7

40. Huang, L., Wang, J., Ma, Z., Zhang, Y., & Zhang, X. (2019). Convergent evolution of prion-like domains across eukaryotic proteomes. Molecular biology and evolution, 36(9), 2041-2051.

41. Hyeon, C., & Onuchic, J. N. (2007). Mechanical control of the directional stepping dynamics of the kinesin motor. Proceedings of the National Academy of Sciences, 104(42), 17382-17387. DOI: 10.1073/pnas.0707902104

42. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., & Obradovic, Z. (2016). The importance of intrinsic disorder for protein phosphorylation. Nucleic acids research, 44(14), 1-12. doi: 10.1093/nar/gkw401.

43. Iguchi, Y., Katsuno, M., Niwa, J., Takagi, S., Ishigaki, S., Ikenaka, K., Kawai, K., Watanabe, H., Yamanaka, K., & Sobue, G. (2013). Loss of TDP-43 causes age-dependent progressive motor neuron degeneration. Brain: A Journal of Neurology, 136(Pt 5), 1371–1382. https://doi.org/10.1093/brain/awt029

44. Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics, 11(2), 97-108. doi: 10.1038/nrg2689

45. Kan, C.-C., Chung, T.-Y., Juo, Y.-A., & Hsieh, M.-H. (2015). Glutamine rapidly induces the expression of key transcription factor genes involved in nitrogen and stress responses in rice roots. *BMC Genomics*, *16*(1), 731. https://doi.org/10.1186/s12864-015-1892-7

46. Kim, H. J., Kim, N. C., Wang, Y. D., Scarborough, E. A., Moore, J., Diaz, Z., ... & Lee, S. J. (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. Nature, 495(7442), 467-473.

47. Kim, J.-E., Ryu, I., Kim, W. J., Song, O.-K., Ryu, J., Kwon, M. Y., Kim, J. H., & Jang, S. K. (2008). Proline-Rich Transcript in Brain Protein Induces Stress Granule Formation. *Molecular and Cellular Biology*, *28*(2), 803–813. https://doi.org/10.1128/MCB.01226-07

48. King, O. D., Gitler, A. D., & Shorter, J. (2012). The Tip of the Iceberg: RNA-Binding Proteins with Prion-like Domains in Neurodegenerative Disease. Brain Research, 1462, 61-80. doi:10.1016/j.brainres.2012.01.016

49. Lancaster, A. K., Nutter-Upham, A., Lindquist, S., & King, O. D. (2014). PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics*, *30*(17), 2501–2502. https://doi.org/10.1093/bioinformatics/btu310

50. Lallemand, Y., & Leducq, J. B. (2020). Chromosomal rearrangements and transposable elements. In Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts (pp. 207-225). Springer.

51. Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23(1), 127-128. https://doi.org/10.1093/bioinformatics/btl529

52. Li, D., McIntosh, C. S., Mastaglia, F. L., Wilton, S. D., & Aung-Htut, M. T. (2021). Neurodegenerative diseases: a hotbed for splicing defects and the potential therapies. *Translational Neurodegeneration*, *10*, 16. https://doi.org/10.1186/s40035-021-00240-7

53. Li, Y. R., King, O. D., Shorter, J., & Gitler, A. D. (2013). Stress granules as crucibles of ALS pathogenesis. Journal of cell biology, 201(3), 361-372. https://doi.org/10.1083/jcb.201302044

54. Lie, Y. S., Sun, L., Ji, Y. H., Wei, Y. P., Gao, Y., Li, S. S., & Li, X. M. (2020). PTM-sites: A database of protein post-translational modifications sites. Database, 2020.

55. Liu, Y., Wang, X., & Liu, B. (2020). RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins. *Briefings in Bioinformatics*, *bbaa018*. https://doi.org/10.1093/bib/bbaa018

56. Liew, Y. J., Aranda, M., & Voolstra, C. R. (2020). Reefgenomics.Org- a repository for marine genomics fata. Database (Oxford), 2016.

57. Lobley, A., Swindells, M. B., Orengo, C. A., & Jones, D. T. (2007). Inferring Function Using Patterns of Native Disorder in Proteins. *PLoS Computational Biology*, *3*(8). https://doi.org/10.1371/journal.pcbi.0030162

58. Long, J. C., & Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. The Biochemical Journal, 417(1), 15–27. https://doi.org/10.1042/BJ20081501

59. Lopez-Quilez, A., Mytko, K., Jain, S., Zalucki, O., & Witting, P. K. (2020). Emerging role of RNA-binding proteins in the control of early stage breast cancer metastasis. Journal of Oncology, 2020, 1-11. https://doi.org/10.1155/2020/5365471

60. Lu, A., Magupalli, V. G., Ruan, J., Yin, Q., Atianand, M. K., Vos, M. R., ... & Egelman, E. H. (2014). Unified polymerization mechanism for the assembly of ASC-dependent inflammasomes. Cell, 156(6), 1193-1206.

61. MacLea, K. S., Paul, K. R., Ben-Musa, Z., Waechter, A., Shattuck, J. E., Gruca, M., & Ross, E. D. (2015). Distinct Amino Acid Compositional Requirements for Formation and Maintenance of the [PSI+] Prion in Yeast. *Molecular and Cellular Biology*, *35*(5), 899–911. https://doi.org/10.1128/MCB.01020-14

62. Malinovska, L., Kroschwald, S., Alberti, S. (2013). Protein disorder, prion propensities, and self-organizing macromolecular collectives. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1834(5), 918–931.

63. Mavadat, E., Seyedalipour, B., Hosseinkhani, S., & Colagar, A. (2023). Role of charged residues of the “electrostatic loop” of hSOD1 in promotion of aggregation: Implications for the mechanism of ALS-associated mutations under amyloidogenic conditions. *International Journal of Biological Macromolecules*, *244*, 125289. https://doi.org/10.1016/j.ijbiomac.2023.125289

64. Mizianty, M. J., Stach, W., Chen, K., & Kedarisetti, K. D. (2020). Compositional biases and weak disordered regions drive domain formation and phase separation in intrinsically disordered proteins. Current Opinion in Structural Biology, 63, 1-8. https://doi.org/10.1016/j.sbi.2019.12.002

65. Mo, Y., Feng, Y., Huang, W., Tan, N., Li, X., Jie, M., Feng, T., Jiang, H., & Jiang, L. (2022). Liquid–Liquid Phase Separation in Cardiovascular Diseases. *Cells*, *11*(19), 3040. https://doi.org/10.3390/cells11193040

66. Moesa, H. A., Wakabayashi, S., Nakai, K., & Patil, A. (2012). Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Molecular BioSystems*, *8*(12), 3262–3273. https://doi.org/10.1039/C2MB25202C

67. Necci, M., Piovesan, D., Predictors, C., Curators, D., & Tosatto, S. C. E. (2020). Critical Assessment of Protein Intrinsic Disorder Prediction. *BioRxiv*, 2020.08.11.245852. https://doi.org/10.1101/2020.08.11.245852

68. Neme, R., & Tautz, D. (2016). Evolution: dynamics of de novo gene emergence. Current biology, 26(18), R737-R739.

69. Nielsen, J. T., & Mulder, F. A. A. (2019). Quality and bias of protein disorder predictors. *Scientific Reports*, *9*(1), 5137. https://doi.org/10.1038/s41598-019-41644-w

70. Nikolić, N., Gajić-Veljić, M., Radosavljević, D., & Janković, R. (2017). Implications of APOE, APP, BACE1, PSEN1, PSEN2, and MAPT in Alzheimer's disease pathogenesis and their genetic association with cancer development. Journal of Molecular Neuroscience, 63(3-4), 382-392. https://doi.org/10.1007/s12031-017-0901-2

71. Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. https://bioinfogp.cnb.csic.es/tools/venny/index.html

72. Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling--a review. Gene, 238(1), 103-114. doi: 10.1016/s0378-1119(99)00228-0

73. Ponte, I., Vila, R., & Suau, P. (2003). Sequence complexity of histone H1 subtypes. *Molecular Biology and Evolution*, *20*(3), 371–380. https://doi.org/10.1093/molbev/msg041

74. Ryan, Veronica H, and Nicolas L Fawzi. “Physiological, Pathological, and Targetable Membraneless Organelles in Neurons.” *Trends in neurosciences* vol. 42,10 (2019): 693-708. doi:10.1016/j.tins.2019.08.005

75. Sadakane, Y., & Kawahara, M. (2018). Implications of Metal Binding and Asparagine Deamidation for Amyloid Formation. *International Journal of Molecular Sciences*, *19*(8), 2449. https://doi.org/10.3390/ijms19082449

76. Sagarin, Raphael & Ambrose, Richard & Becker, Bonnie & Engle, John & Kido, Janine & Lee, Steven & Miner, C. & Murray, Steven & Raimondi, Peter & Richards, Dan & Bell,

Christy. (2007). Ecological impacts on the limpet Lottia gigantea populations: Human pressure over a broad scale on island and mainland intertidal zones. Marine Biology. 150. 399-413. 10.1007/s00227-006-0341-1.

77. Saini, C., Brown, S. A., & Dibner, C. (2015). Human peripheral clocks: applications for studying circadian phenotypes in physiology and pathophysiology. Frontiers in neurology, 6, 95. https://doi.org/10.3389/fneur.2015.00095

78. Santos, J., Iglesias, V., Santos-Suarez, J., & Gil-Longo, J. (2013). Importance of aromatic residues in peptides for DNA binding: Structure-affinity relationships. Journal of Peptide Science, 19(9), 585-593.

79. Shelkovnikova, T. A., Robinson, H. K., Southcombe, J. A., Ninkina, N., Buchman, V. L., & Buchman, A. (2014). Recruitment into stress granules prevents irreversible aggregation of FUS protein mislocalized to the cytoplasm. Cell cycle, 13(23), 3711-3722. https://doi.org/10.4161/15384101.2014.965065

80. Silina AV. Effects of temperature, salinity, and food availability on shell growth rates of the Yesso scallop. PeerJ. 2023 Feb 21;11:e14886. doi: 10.7717/peerj.14886. PMID: 36846447; PMCID: PMC9951806.

81. Singh, B., Kinnebrew, M., & Alshareedah, I. (2020). The Role of Intrinsically Disordered Regions and Chaperones in the Regulation of Alternative Splicing. International Journal of Molecular Sciences, 21(19), 7145. https://doi.org/10.3390/ijms21197145

82. Shen, B., Chen, Z., Yu, C., Chen, T., Shi, M., & Li, T. (2021). Computational Screening of Phase-separating Proteins. *Genomics, Proteomics & Bioinformatics*, *19*(1), 13–24. https://doi.org/10.1016/j.gpb.2020.11.003

83. Shin, Y., & Brangwynne, C. P. (2017). Liquid Phase Condensation in Cell Physiology and Disease. Science, 357(6357), eaaf4382. doi:10.1126/science.aaf4382

84. Sun, Y., Chakrabartty, A., & Li, Y. (2018). Prion-like domains in RNA binding proteins are essential for higher-order assembly of stress granules. Journal of Biological Chemistry, 293(27), 11030-11041. Doi: 10.1074/jbc.RA118.002881

85. Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A. K., Daughdrill, G. W., & Uversky, V. N. (2013). The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins*, *1*(1), e24360. https://doi.org/10.4161/idp.24360

86. Toombs, J. A., McCarty, B. R., & Ross, E. D. (2010). Compositional determinants of prion formation in yeast. Molecular and cellular biology, 30(1), 319-332.

87. UniProt Consortium. (n.d.). UniProtKB - Swiss-Prot and TrEMBL. Retrieved May 5, 2021, from https://www.uniprot.org/

88. Uversky, V. N. (2013). A decade and a half of protein intrinsic disorder: Biology still waits for physics. Protein Science, 22(6), 693-724. https://doi.org/10.1002/pro.2261

89. Uversky, V.N. (2015). Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. Frontiers in Aging Neuroscience, 7, 18. doi:10.3389/fnagi.2015.00018

90. Uversky, V. N. (2017). Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. Current Opinion in Structural Biology, 44, 18-30.

91. Uversky, V. N., & Dunker, A. K. (2010). Understanding protein non-folding. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1804(6), 1231-1264. doi: 10.1016/j.bbapap.2010.01.017

92. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., & Babu, M. M. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, *114*(13), 6589–6631. https://doi.org/10.1021/cr400525m

93. Wang, Y., Liu, J., Liu, X., & Lu, F. (2018). Regulatory roles of prion-like domains in RNA binding proteins. International journal of molecular sciences, 19(12), 3627.

94. Wang, B., Zhang, L., Dai, T., Qin, Z., Lu, H., Zhang, L., & Zhou, F. (2021). Liquid–liquid phase separation in human health and diseases. *Signal Transduction and Targeted Therapy*, *6*(1), 1–16. https://doi.org/10.1038/s41392-021-00678-1

95. Weathers, E. A., Paulaitis, M. E., Woolf, T. B., & Hoh, J. H. (2007). Insights into protein structure and function from disorder-complexity space. *Proteins*, *66*(1), 16–28. https://doi.org/10.1002/prot.21055

96. Wegmann, S., Eftekharzadeh, B., Tepper, K., Zoltowska, K. M., Bennett, R. E., Dujardin, S., Laskowski, P. R., MacKenzie, D., Kamath, T., Commins, C., Vanderburg, C., Roe, A. D., Fan, Z., Molliex, A. M., Hernandez-Vega, A., Muller, D., Hyman, A. A., Mandelkow, E., Taylor, J. P., & Hyman, B. T. (2018). Tau protein liquid–liquid phase separation can initiate tau aggregation. *The EMBO Journal*, *37*(7), e98049. https://doi.org/10.15252/embj.201798049

97. Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1), 18-29. https://doi.org/10.1038/nrm3920

98. Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., & Obradovic, Z. (2007). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. Journal of Proteome Research, 6(5), 1882-1898.

99. Zbinden, A., Pérez-Berlanga, M., De Rossi, P., & Polymenidou, M. (2020). Phase Separation and Neurodegenerative Diseases: A Disturbance in the Force. *Developmental Cell*, *55*(1), 45–68. https://doi.org/10.1016/j.devcel.2020.09.014

100. Zhang, G., Xie, Y., Jin, Y., Hu, Q., Chen, X., & Shen, X. (2019). PCNP is overexpressed in human gastric cancer and promotes proliferation and invasion of gastric cancer cells. Molecular medicine reports, 20(5), 4545-4552.

101. Zhang, Y., Lu, X., Zhao, Y., Ren, J., & Gong, X. (2021). Evolutionary and structural analysis of prion-like domains across animal proteomes. Frontiers in Molecular Biosciences, 8, 619243. doi: 10.3389/fmolb.2021.619243