Computational insights into alternative splicing driven proteome diversification: An evolutionary perspective.

Paras Verma

PH17060

A thesis submitted for the fulfillment of

the degree of Doctor of Philosophy

Department of Biological Sciences

Indian Institute of Science Education and Research Mohali

Knowledge city, Sector 81, SAS Nagar, Manauli PO, Mohali 140306, Punjab, India.

April 2024

Dedicated to Maa and Papa

Declaration

I have carried out the work presented in this thesis under the guidance of Dr. Shashi Bhushan Pandit at the Indian Institute of Science Education and Research Mohali. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgment of collaborative research and discussions. This thesis is a bona fide record of my original work, and all sources listed within have been detailed in the bibliography.

Paras Verma

Date:

Place:

As the supervisor of the candidate's thesis work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Shashi Bhushan Pandit

Date:

Place:

Acknowledgments

I extend my heartfelt gratitude to the many individuals who have supported me throughout my Ph.D. journey.
Firstly, I am deeply grateful to my Ph.D. advisor, Dr. Shashi B. Pandit, for his unwavering patience, guidance, and constant support. His mentorship has been instrumental in shaping my research path.
I sincerely appreciate my doctoral committee members, Dr. Kuljeet S. Sandhu, and Dr. Kamal P. Singh, for their insightful suggestions and ongoing interest in my work.
A special acknowledgment goes to the Computational Biology Group members, past and present, for their camaraderie, advice, and discussions. My gratitude extends to Deeksha, Preeti, Rivi, Dr. Nidhi, Arshdeep, and Meenakshi for their support.

My friends Deeksha (again), Archit, Yogesh, Arpita, and Swati have been my pillars of strength, and their companionship has made this journey memorable.

I appreciate the entire IISER community, administrative staff, and support services for their contributions to my academic journey. Special thanks to Mr. Prateek and Dr. Samrat for ensuring consistent supply of clean RO water.

Additionally, I would like to thank DST-SERB for funding my academic trip to Germany and providing me the opportunity to share my work at an international forum.

To my family, your love, prayers, and encouragement have been my foundation. I thank my parents, Mr. Naresh Kumar and Mrs. Shikha and my sister Chetna, for their unwavering support and positive mindset. I am deeply grateful to the Almighty for blessing every aspect of my life.

Lastly, I would like to acknowledge and appreciate this terrific journey, which moulded me into the person I am today. The lessons and skills learnt here will be contributing to the rest of my life. I would like to end this by acknowledging my decision to embark on this journey.

(Special thanks to Deeksha (again) for reminding me to acknowledge the contribution of the most important person, i.e., myself!)

Thank you all for your presence and support!

Paras Verma

Synopsis

Proteins are one of the most essential and versatile molecules in biological systems, playing crucial roles in organismal complexity and evolution. Protein sequence/structure encompasses a wide array of features, including binding motifs, short linear interaction motifs, evolutionarily conserved surface patches, and modular domains. These facilitate proteins to perform diverse functions in cells ranging from structural role, binding activity to catalytic function. The interaction of proteins with other cellular components is essential and crucial to perform their biological processes, where niche of housekeeping complexes form consistent interactions and evolutionary younger proteins form specific interactions in distinct tissue or cellular stages. Minute protein changes over the evolutionary course may lead to functional innovation while performing complex regulatory and signaling pathways. These feature-encompassing capabilities of proteins and their interplay with transcriptional and post-transcriptional processes have provided an edge in the adaptive evolution of eukaryotes while instilling phenotypic complexity.

One such process, alternative splicing (AS), generates transcriptome diversity in eukaryotes through variably spliced mRNA transcripts. Their translation contributes to the proteome expansion and increasing the functional repertoire of genes. The prevalence of AS throughout the eukaryotic kingdom and its contribution to expanding the proteome offers a plausible explanation for the observed perplexing disparity between the count of genes and required proteome diversity for organismal complexity. While significant experimental and computational efforts have enhanced our understanding of the AS impact on transcriptome diversity, there have been limited studies detailing its contribution to proteome expansion. In my thesis, I have devised an innovative framework to uniquely annotate exons that facilitates comparative analyses of proteome variation generated by various AS events (exon skipping, mutually exclusive exons, alternate splice sites, and intron retention). The significance of this innovative framework was emphasized in deciphering unique AS events, which were impossible without carefully integrating transcriptome and proteome counterparts of spliceoforms. These events were compared for five representative model organisms: *Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, Mus musculus,* and *Homo sapiens,* and are documented in a publicly accessible database. Through our analysis, we illustrate the complex interplay of AS and alternate transcription (AT) in diversifying human proteome. Subsequently, their impacts were assessed in the context of imparting functionality

and diversity within intra-gene isoforms. In addition to splicing analysis, I have performed modeling and simulation studies of β-sheet nanocrystal regions in spider silk protein (inspired from a computational materiomics design perspective) and separately investigated the role of Y321 in oligomerization of Vibrio cholerae Cytolysin toxin (Molecular Dynamics) and network analyses. Below is a brief overview of the work performed during my Ph.D. duration, that is arranged in six chapters.

Exon Nomenclature Annotation and Classification in Transcripts (ENACT): A framework to uniquely annotate exons and transcripts of genes

Regarding the prevalence of AS in eukaryotic genes and genomes, recent RNAseq-based studies have shown that ~95% of multi-exon human genes undergo alternative splicing, which has the potential to translate into protein products. These isoforms are differentially expressed in tissues/developmental stages and are essential in regulating cellular processes. AS-driven proteome abundance is more complex to quantify than its transcriptional counterpart due to constraints on analyzing protein expression, mass spectrometry-based detection, and tissue-wise expression. However, recent advancements in proteogenomics and ribosomal profiling techniques have started revisiting and unraveling the role of splicing-induced isoform variation in normal cellular and disease stages. Considering their importance, many primary databases, such as NCBI and Ensembl, and eukaryote-specific databases (UCSC genome browser) provide documentation of AS variants in eukaryotic genes with regular updates. In general, splicing events have four defined categories a) exon skipping events (ES), b) mutually exclusive events where two exons hardly came together in one isoform (MXE), c) splice site various (5’/3' or both), and d) Intron retention events. All the above-listed splicing events can occur exclusively in UTR regions or the middle region of the protein sequence or their interface at the translation start and termination sites. Definitions of such alternative events have solid foundations in the transcriptome. However, their occurrence may not necessarily impact the protein product directly as changes in the UTR region are also frequent that may alter translation rate or introduce upstream ORFs to halt the ribosome. Splicing and choice of alternative promoter sites (APS), alternative polyadenylation (AP), and of alternative translation initiation and termination sites (ATIT) further complicate the above event definitions and how they may affect the primary protein product of the gene. All above-listed events in combination and isolation have immense potential to modulate the termini regions of isoforms very often, and similar has been the observation of many studies. However, the inference of what exon regions are participating and affecting the protein domains becomes

difficult to comprehend solely from genomic coordinates and four categories of splicing events. In the current evolving era of proteogenomics, to purpose and gain insights from such splicing and related events about protein sequences being modified and tuned, there is a need to annotate and systematically characterize the AS events, which could be mapped easily to proteins.

To address the above limitations and to incorporate protein sequence information on exons, we have designed a standardized framework system called Exon Nomenclature and Classification of Transcripts (ENACT). ENACT annotates exonic entities with features encapsulating a) the role of exons in all transcripts (coding/UTR or both in different transcripts), b) coding status in the current transcript, c) constitutive, constitutive like, or alternate nature of exon d), linear sequence position of exon in gene, e) their splice site changes and f) variant count of splice site changes. These characteristics are assembled into six-letter frameworks called EUIDs, enabling computational analysis of annotated genomes and enhanced illustration of splicing-induced transcript changes for a gene. To make these annotations publicly accessible, we have documented them in a visually appealing manner in a database called ENACTdb, available at http://www.iscbglab.in/enactdb. ENACTdb also renders predicted secondary structure, disorder content, and Pfam domain annotations to individual transcripts, further extending visual aid and helping interpret inter-transcript functional changes.

Distribution of AS events in representative genomes: An evolutionary perspective

Having exons documented in ENACTdb, we compared their features among *Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, Mus musculus,* and *Homo sapiens.* Their inter-organism comparison demonstrates ENACT's enhanced utility in uncovering distinct organismal preferences of AS events and transitions. We only compared the exon entities for protein-coding genes with ≥2 distinct protein-coding isoforms and ≥2 coding exons (2X2 dataset). We observed that constitutive coding exons are the most prevalent form of exons in all organisms. In the context of organismal complexity, we observed an increasing fraction of UTR, specifically their alternate subtype of exons. From lower to higher organisms, there is increasing occurrences of 'dual' exons, which are non-coding in some transcripts and coding in others. These are only identifiable with careful integration of protein sequences in AS events and are a hallmark of ENACT. Inter-genome comparison indicated differences in the prevalence of exon skipping (increased with organismal complexity) in higher organisms for coding exons and decreased preference to alter splice sites. This marked difference was further strengthened when the length distribution of exons for those categories were compared.

Additionally, we also observed ratio of alternate to constitutive UTR exons is higher than coding exons, indicating an increased extent of Alternate Transcription (AT) in conjunction with AS.

Detailed investigation of Alternative Splicing and Alternative Transcription induced changes in the proteome of *Homo Sapiens.*

Delving into details from the previously observed simultaneous preference of AT and AS in higher organisms, we carefully elucidated their footprint on human genes using a detailed investigation of exon variations in the human genome. We observed a considerable impact of the AT in coding gene architecture, where they contributed, on average, $1/3^{rd}$ of the protein sequence. Regarding inter-transcript coding region variability, alterations were often found to be more prevalent in AT driven region than in AS driven, which was noteworthy as AS driven region encodes far greater protein region ($2/3^{rd}$) than AT. Subsequently, we assessed the possible differences in their variability and observed different inclusion rates of alternate exons, especially for sub-type undergoing splice site changes (A(ss)). Both alternate (A) and 'A(ss)' exons were analyzed in detail for their impact on diversifying intra-gene transcripts. A(ss) cases showed marked differences in insertion/deletion (indel) lengths on AT and AS-driven regions, often affecting coiled secondary structures. The AT-driven region harbors more frame-altering events, especially when A(ss) cases undergo splice site events. The association of these splice events with the C-terminal region was observed that may provide mechanistic insights into previously observed enrichment of substitutions in the C-Terminal region of the protein. The utility of ENACT helped uncover the context of alternate exons, where functional associations of AT/AS-driven regions were detailed. In addition to the previously established roles of AT-driven regions in introducing intrinsically disordered fragments and phosphorylation sites, we observed a considerable fraction of theirs also being assigned to domains. Detailed comparison with AS-driven region elucidated that AT region encodes more intact domains (contained in a single exon) than AS, which has not been highlighted in the literature. The noted prevalence of their contributed domains and more intact domains may have overlooked plausible mechanism in generating functional diversity within intra-gene isoforms, where in addition to AS-driven region AT may also be a substantial variable in the equation of organismal complexity. It is important to note that AT contributes high variability in proteins despite contributing only 1/3 to the isoform region.

Optimal protein sequence design mitigates mechanical failure in silk β-sheet nanocrystal

Spider silk (SS) is an intriguing material with several attractive properties and is one of the toughest biomaterials known to humans. The molecular origin of SS's strength is mostly attributed to the hierarchical arrangement of β-sheet nanocrystals (laminated antiparallel beta sheets) in an amorphous-rich matrix. This complex hierarchical arrangement gives the SS fiber strength and elasticity, making it amenable for many material property applications and a key focus of materiomics initiatives. The β-sheet nanocrystalline region's stability and resilience have foundations in the highly organized hydrogen bond patterns. The nanocrystal regions predominantly contain repeat motifs of "Ala" or "Ala/Gly". We hypothesized that changing the basic constituents of those repeats by β -sheet favoring residues may help further increase the strength attributes of the modified silk protein.

We modeled the repeat amino acid sequences in nanocrystals and performed simulations to investigate the possibility of increasing the mechanical features of silk. Through careful modeling, we replaced "Ala/Gly" repeats with representative amino acids having different physiochemical properties. The representative amino acids were selected from polar (Threonine/Asparagine) and hydrophobic (Isoleucine/Valine) physicochemical groups that preferred to occur in β-strand. Moreover, we hypothesized that polar amino acids would make extensive hydrogen bond interactions and hydrophobic sidechain would increase strength through side chain interactions. The selected representative amino acids were classified based on their size and physicochemical properties into three broad categories: (i) Small amino acids (SAA/small-AA): poly-Alanine (pAla), poly-Alanine-Glycine (pAlaGly), poly-Glycine (pGly); (ii) Polar amino acids (PAA/polar-AA): poly-Threonine (pThr) and poly-Asparagine (pAsn); and (iii) large hydrophobic amino acids (HAA/hydrophobic-AA): poly-Valine (pVal) and poly-Isoleucine (pIle). The resilience of nanocrystals and fracture were compared on the modeled nanocrystal by performing SMD pulling. The ultimate tensile strength and toughness of pAla is the maximum followed by pAlaGly repeat containing nanocrystals. This elucidated that the β-sheet sequence occurring in the nanocrystalline region of the natural silk crystal optimized for SS nanomechanical features. We have also realized that though representative groups PAA and HAA have contributed to interactions by their sidechains, their additive effect altered the defacto pulling by compensating the reformation of the backbone hydrogen-bonded motif during signature slip-stick motion. Performing these *in silico* experiments shows that nature chooses amino acids amenable for optimal packing and interface sheet stacking, which necessitates resilient response to external load. Further, these experiments showed that the

maintenance of spatial hydrogen bonds and their direction to external force must be maintained when the external force is applied to it. The work is performed in collaboration with Dr. K. P. Singh (DPS, IISER Mohali).

Unraveling the Functional Implications of Y321A mutation in the *Vibrio cholerae* Cytolysin (VCC) through MD Simulations and network analysis

'Vibrio cholerae cytolysin' (VCC) from *V. cholerae* is a prominent virulence factor of the cholera pathogen. It belongs to β-Pore-forming toxins that utilize β-strands for the pore-formation mechanism to lyse the cells. VCC is a multidomain dimorphic protein that remains in monomeric form in solution and on membrane interaction it assembles into heptamer while organizing its pre-stem region into the β-barrel pore. To gain insight into this mechanism of transition, four aromatic amino acids that change their orientation and are in close vicinity to the pre-stem loop, which will transit into barrel were mutated to alanine (experimental details from the lab of Dr. Kausik (DBS, IISER Mohali)). One mutant Y321A leads to functional arrest of the protein and blocks its cytotoxicity and pore-forming capability, indicating its role in this structural reorganization. We designed and performed MD simulations of WT and mutant form (Y321A) to provide a mechanistic basis of this defect. Employing community network analysis and cross-correlation differences in those two ensembles, we showed that residue interactions involving the cradle loop, part of the pre-stem, β-Trefoil, and β-Prism domain are affected in the Y321A mutant. The mutation of Y321A in the hinge region of the pore-forming pre-stem motif appears to impose long-range defects within the VCC structure. Such defects, in turn, possibly affect the communications between different structural motifs/modules/domains surrounding the pore-forming pre-stem motif, thereby compromising structural/conformational reorganization, crucial for the oligomeric pore-formation.

List of Publications

Mondal AK, **Verma P**, Sengupta N, Dutta S, Pandit SB, Chattopadhyay K. Tyrosine in the hinge region of the pore-forming motif regulates oligomeric β-barrel pore formation by *Vibrio cholerae* cytolysin. *Mol Microbiol*. 2021; 115: 508–525. https://doi.org/10.1111/mmi.14631

Verma P, Panda B, Singh KP, Pandit SB. Optimal Protein Sequence Design Mitigates Mechanical Failure in Silk β-Sheet Nanocrystals. *ACS Biomaterials Science & Engineering* **2021** *7* (7), 3156-3165. https://doi.org/10.1021/acsbiomaterials.1c00447

Contents

List of Figures

List of Tables

Chapter 1

Alternative splicing: Role in diversifying transcriptome and proteome with challenges.

1.1 INTRODUCTION

Eukaryotic gene architecture comprises distantly placed segments called exons and introns, where the former takes part in mRNA and introns are alternatively excised during pre-mRNA processing (Long and Deutsch 1999). Alternative splicing is a co and post-transcriptional mechanism before the translation that enables generation of those different mRNA products from individual genes by differentially incorporating exons (Marasco and Kornblihtt 2022) (Berget, Moore, and Sharp 1977; Breitbart, Andreadis, and Nadal-Ginard 1987). It should be noted that not all exons are part of coding region, and significance of those exons will be discussed in the later section (Aspden, Wallace, and Whiffin 2023). Differential incorporation or excision of exons/introns is primarily done by four major event types of AS, as shown in Figure 1.1, which include: a) exon skipping (ES), where one or more exons are skipped (not form part) in a transcript; b) mutually exclusive events in which two exons are mutually exclusive as these do not co-occur in any alternatively spliced transcripts of a gene; c) alternate 5’ (5SS) or/and 3’ (3SS) splice site of the exon, where it has more than one splice sites (5’/3’), and d) Intron retention events (IR), where intron region between two exons is retained in a transcript. (Marasco and Kornblihtt 2022). Amongst varied forms of AS (Modrek and Lee 2002), exon skipping (ES), and intron retention (IR) are quite prevalent in higher eukaryotes, where the frequency of the former increased vastly in bilaterian ancestors. The latter is more abundant in lower eukaryotes such as fungi and protozoa (Grau-Bové, Ruiz-Trillo, and Irimia 2018). The relationship of AS with organismal complexity and associated functions and phenotypic correlations of species are discussed in the next section.

1.2 SPLICING AND ORGANISMAL COMPLEXITY

After the first draft release of the human genome, the disparity between the number of genes and that of the proteome counterpart came as surprise and has been an active interest to scientific community (Lander et al. 2001; Venter et al. 2001). Further refining of assembly revealed a lower-than-expected number of protein-coding genes. Initially discovered more than four decades ago (Alt et al. 1980), alternative splicing (AS) has been suggested as a key piece in the puzzle to unravel mystery behind achieving transcriptome diversity which possibly can lead to proteomic complexity from a limited gene pool (Nilsen and Graveley 2010). Alternative splicing was proposed as a candidate to explain the diversification in the number of cell types observed in some eukaryotic lineages (a higher number of cell types in each species is assumed to reflect increased organism complexity (OC)). In addition to AS, gene duplication has long been associated with functional innovation and was correlated to OC (Talavera et al. 2007; Hahn and Wray 2002). However, rates of gene duplications failed to reflect the diversification of cell types observed in several eukaryotic lineages, and its correlation to OC is only reliable, whence analysis is restricted to metazoans (Schad, Tompa, and Hegyi 2011). As more studies keep extending deeper into the puzzle of OC, AS has been confidently viewed as the primary driving force behind higher eukaryotes' transcriptome and proteome diversity (Nilsen and Graveley 2010). However, its advent, prevalence, and footprint on evolutionary trees had to await comparative genomics and more resounding transcriptomic evidence (Trapnell et al. 2010; Pan et al. 2008). Recent high-throughput sequencing data has shown that nearly all multi-exonic genes in model vertebrates and up to 95% in humans undergo splicing (Li et al. 2016; Pan et al. 2008; Barbosa-Morais et al. 2012). Evidence of splicing does exist in yeast, but infrequent observations for alternative events (Howe, Kane, and Ares 2003). Among diverse eukaryotes, considerable prevalence of AS (Bush et al. 2017; Singh and Ahi 2022) suggests its relation to phenotypic complexity in numerous biological occasions in diverse taxa, including but not limited to coat color in deer mice (*Peromyscus*) (Mallarino et al. 2016), flowering time in *Arabidopsis* (Macknight et al. 2002) and barley (Grützmann et al. 2014), thermogenesis homeostasis (Vernia et al. 2016), virulence in pathogenic fungi (Grützmann et al. 2014), neuronal maturations in primates (Calarco et al. 2007; Lin et al. 2010), gender determination by sex-lethal gene in insects and ability to sense infrared rays by vampire bats to locate their prey effectively by lowering the temperature of heat sensitive cation channel TRPV1 by using alt5' splice site to truncate carboxy-terminal domain (Gracheva et al. 2011).

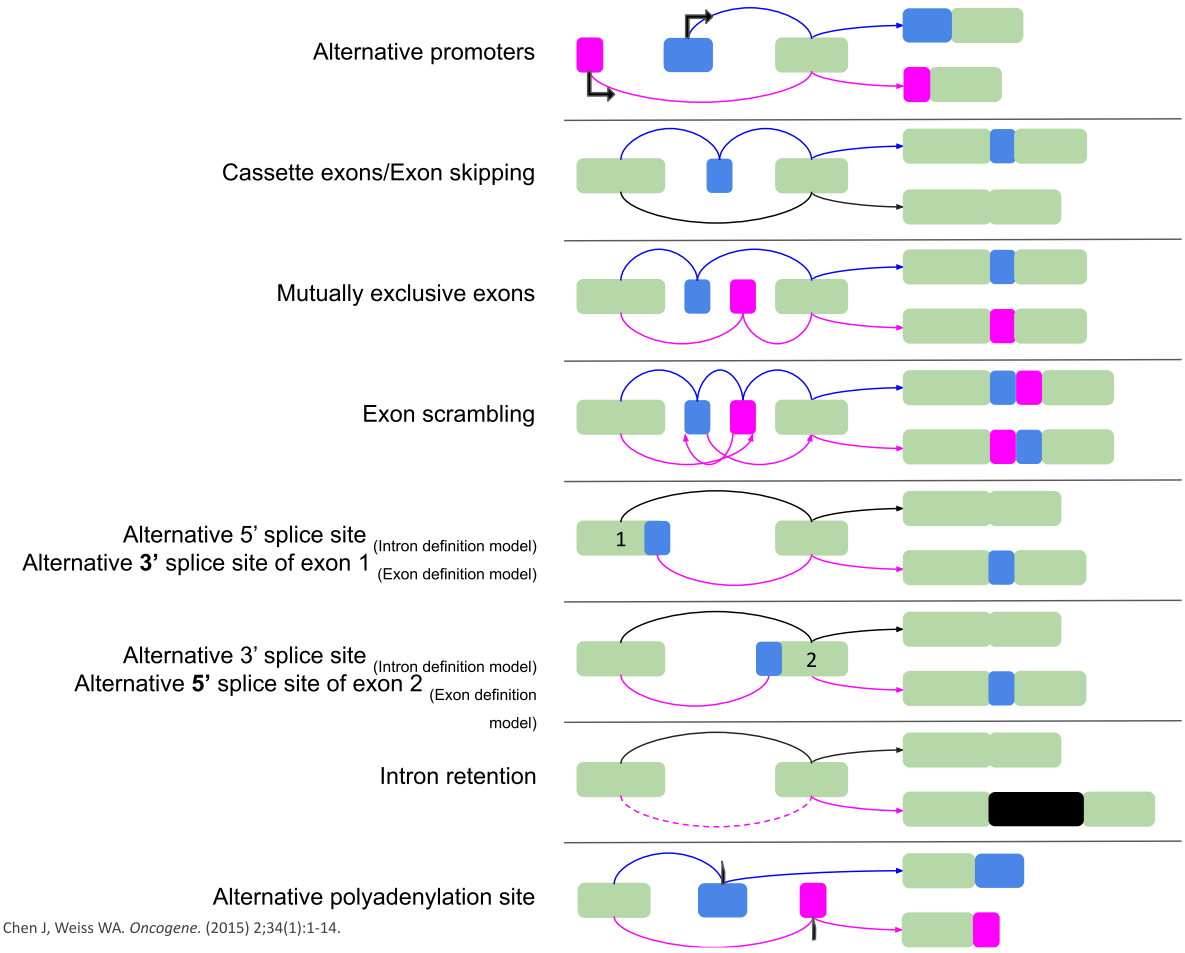


Figure 1.1: Different events of alternative splicing (AS) and alternative transcription (AT) processes. The combination of exons and their inclusion has been shown on the left, with arrow connectors indicating their combinations. The resulting mature transcripts are shown on the right. Constitutively spliced regions are colored green, whereas alternate exons are colored magenta and blue. Intron retention events (in absence of splicing) have been represented with the color black. The arrow symbol in Alternate promoters indicates corresponding transcription initiation sites and the vertical bar in Alternative polyadenylation sites indicates transcription termination sites.

In higher eukaryotes, assimilation of AS in gene architecture also plays several vital roles in cellular or molecular functions like enzymatic activities, transcription, apoptosis, autophagy, differentiation, cell/tissue fate determination (Baralle and Giudice 2017; Black 2003), and other developmental processes (Tang et al. 2013; Wang et al. 2008; Wang and Burge 2008). These processes are tightly regulated, and needless to mention, its dysregulation has also been implicated in several pathologies, including diseases like cancer and developmental and

neurological disorders (Zhang, Qian, et al. 2021), further implying its prioritizing preference to delineate associated molecular impact.

The extent of splicing and its presence among several lineages and assimilated roles has been emphasized, but how this complex mechanism evolved and matured among different organisms, especially with an increase in organismal complexity, is the focus of the next section.

1.3 EVOLUTIONARY ADAPTATION MECHANISMS IN CONTEXT OF AS

To gain insights into AS-driven evolutionary adaptations, the intricacies of splicing and associated demarcation of the introns and exons must be considered (Keren, Lev-Maor, and Ast 2010). Spliceosomal assembly carefully orchestrates this demarcation (Matera and Wang 2014), and it is of more than 300 distinct subunits of RNA, protein, and protein cofactors; and is considerably one of the largest macromolecular complexes (Fredericks et al. 2015; Nilsen 2003). The composition of spliceosomes defines *trans* factors, and they interact with *cis* sequence motifs in mRNA molecules that include splice donor/acceptor sites, branch sites, polypyrimidine tracts, and a range of other sequence motifs. Communication between *trans* and *cis-acting* factors is vital in determining whether a region of mRNA will be spliced in as exon or out as intron. Detailed comprehensive interactions are still being elucidated (Matera and Wang 2014), but existing studies detail that their communications are highly variable, transient, and of relatively low specificity (Fredericks et al. 2015). Nature of these interactions introduces immense plasticity in choosing regions to be included in mature mRNA molecules, as composition and biogenesis of spliceosome is variable in different cells and tissue types (Matera and Wang 2014). Considering the nature of exons, ‘constitutive’ are regions that are present throughout the collective mRNA gene pool of organism and are strongly promoted for their inclusion by ubiquitous splicing factors(Keren, Lev-Maor, and Ast 2010), and ‘alternative’ are regions that are variably present in different isoforms and are tightly regulated for their inclusion and involves specific cross talks of *cis* factors and *trans* factors (Castle et al. 2008; Das et al. 2007). Evolutionarily, exon lengths have undergone shortening, favoring the exon definition model and sensitized length factor for its influence on mRNA inclusion-exclusion criteria (De Conti, Baralle, and Buratti 2013a; Keren, Lev-Maor, and Ast 2010). Evolutionarily, constitutive exons are conserved, and alternative exons show a continuum of conservation,

wherein their subpopulation, which takes part in more isoforms (major form; more frequently present), are more conserved than other subpopulations with minor occurrences, indicating latter undergoing loss and gain in individual genomes (Modrek and Lee 2003; Keren, Lev-Maor, and Ast 2010). Differences in patterns and frequency of alternatively spliced exons are also a function of their lineage split in organisms, as detailed by human and mouse genome comparison. A study by Modrek et al. (Modrek and Lee 2003) established that a small fraction of pre-split exons (present in the common ancestor of mouse and human) undergo alternative participation (4%), and this fraction increases to 36% considering post-split exons. Among the latter, 92% (of 36%) were expressed as minor forms. Such differences between major and minor forms are not merely a consequence of selection pressure relaxation but are shaped by both regulatory processes and interplay of co-transcriptional, transcriptional, and post-transcriptional processes (Agirre et al. 2021). Specifically, many alternate exons are spliced post-transcriptionally in contrast to often co-transcriptional splicing of constitutive exons (Tilgner et al. 2012). The dynamics of these alternate exons to integrate with gene architecture and later to transcript assimilation is exhibited by their relaxed selection pressure, which may further open doors to incorporate cis-trans mutations and consequential impact in diversifying protein and RNA regulations. Incorporated alternate exons can undergo negative selection pressure if they lack frame-preserving attribute; however, they need not necessarily if the ancestral exon assimilated isoform keeps getting expressed at normal levels (Singh and Ahi 2022), as reported by (Xing and Lee 2005; Xing and Lee 2006) during mammalian evolution. Further, evidence of 3n (multiple of 3 nucleotide) exons amongst exon skipping events of vertebrates, which are not observed in plants and other eukaryotes, strengthens the fact that those correlated to phenotypic complexity as ORF integrity will usually be maintained (Grau-Bové, Ruiz-Trillo, and Irimia 2018). However, this is still understudied and needs further analysis to affirm uniform organismal prevalence.

1.4 SPLICING IN NON-CDS REGION AND THEIR CO-OCCURRENCE WITH OTHER PROCESSES

In the preceding section, we have reviewed differences between alternate and constitutive exons, their possible routes of evolutionary transitions, and their dynamic role in transcript structure makeup. Recent estimates showed that >95% of multi-exonic genes undergo splicing (Pan et al. 2008). However, their footprint is not exclusive to the CDS region and has been

observed to affect its upstream and downstream UTR regions(Aspden, Wallace, and Whiffin 2023). Comparison of fraction of alternate nucleotides has detailed that regions outside the CDS harbor strikingly >4 fold fraction (Shabalina et al. 2014) and indicate substantial possible impact in modulating gene expression and regulation, where 5'UTR regions are known to harbor promoter elements, influence RNA stability through their secondary structures, altering translation rate and translational efficiency. 3'UTR regions affect transcript expression and localization by altering the choice of polyadenylation sites and translation termination elements (Ji et al. 2011; Derti et al. 2012; Ni et al. 2013). Conceivably, such a high fraction of alternate nucleotides and exons results from weaker selection in UTRs than in CDS, which are constrained by protein structural features (Shabalina et al. 2010).

1.4.1 Co-occurrence of alternative splicing and alternate transcription initiation and termination

Such a higher fraction of alternative regions in UTRs is not an independent consequence of AS but a complex interplay with alternative transcription initiation (ATI) and termination regions (ATT) processes, as evidenced by numerous studies on mammalian gene expressions (Landry, Mager, and Wilhelm 2003; Shabalina et al. 2010). Delving deeper into their complicated interplay and their consequential impact on gene architecture, thorough comparative analysis of alternate nucleotides fraction, the mean number of isoforms, and intron loci features, the study of Shabalina et al. (Shabalina et al. 2010) demonstrated strong and distinct coupling between these processes within and between 5'UTR, CDS and 3'UTR regions (Figure 1.2) . In summary, they observed that

a) positive coupling between alternate exons of AT and AS in 5′ UTR,

b) tight positive correlation between AS in CDS and ATT in 3′ UTRs, and

c) anticorrelation between ATI and AS in 5′ UTRs with AS in the CDS (fig. 3).

Correlation between AT and AS in 5'UTR can be considered plausible, where AS is likely to splice long UTR exons chosen by distinct transcription starts sites during alternate transcription (Mignone et al. 2002; Shabalina et al. 2010; Lynch, Scofield, and Hong 2005) and is a necessity to confer efficient translation initiation later, which is dependent on optimal 5'UTR length. The positive correlation between AS in CDS region and Alternative transcription termination in 3'UTR must indicate the purposing of distinct proteins from AS to

different subcellular conditions by varying alternate polyadenylation sites (PAS) (Shyu, Wilkinson, and van Hoof 2008; Shabalina et al. 2010). Anticorrelation between AS in CDS and AS and ATI in 5'UTR is challenging to be reasoned from current literature. It may indicate exclusive modulation on only upstream UTR, controlling the translation rate of same resulting protein sequence variably in distinct tissues.

Nevertheless, their co-occurring roles in several genes may be involved in modulating transcription, translation, mRNA stability, and localization while fine-tuning gene expression and contributing to the complexity of gene regulation (Shabalina et al. 2010). Their corresponding impact and how they may be introducing functional variability are discussed for 5'UTR and 3'UTR regions in the following section:

1.4.2 5' UTR

The fraction of alternate nucleotides participating in 5′ untranslated regions (5′ UTRs) are far higher than that in coding sequences (CDSs) and 3' region (Shabalina et al. 2010; Shabalina et al. 2014). As previously described, 5'UTR regions harbor promoter elements, influence RNA stability, and affect translation rate and efficiency (Churbanov et al. 2005; Resch et al. 2009). The longer UTR region (or added exon count) and weaker selection pressure open the possibility of encountering more open reading frames (ORF) than the primary one. Similar has been observed in literature where initiation codons, upstream open reading frames (uORFs), in addition to primary ORF were noted. Their misregulation can lead to pathologic conditions and slow or even halt the ribosome, affecting translation rates and, in some scenarios, translational repression (Ji et al. 2004; Resch et al. 2009). Occasionally, they were also found to have a role in increased translation efficiency (Resch et al. 2009; Reynolds, Zimmer, and Zimmer 1996). Many of those uORFs are not merely a result of weak selection influenced recently evolved regions, as they were also found to be conserved, and their deeper functional significance has yet to be detailed (Resch et al. 2009; Churbanov et al. 2005). In addition to the previously observed higher coupling of AT and AS processes in 5'UTR, where former introduced extended UTR regions were supposed to splice out for optimal translation efficiency by AS, literature evidence of them acting in combination exists for a handful of genes, for instance, human gene *axin2* (a negative regulator of *Wnt/B*-catenin signaling) has three isoforms with different arrangements of upstream AUGs (uAUGs) and uORFs in 5'UTRs which confer different mRNA stabilities and translational efficiencies in different isoforms (Resch et al. 2009;

Churbanov et al. 2005; Hughes and Brady 2005). Similar instance has also been observed for *nNOS* (neuronal nitric-oxide synthase), where alternate exon in 5'UTR introduces a translational control element, later inhibiting mRNA translation (Resch et al. 2009; Churbanov et al. 2005; Newton et al. 2003) and in mu-opioid receptor gene, which exhibits leaky scanning of ribosome and corresponding translation repression in different isoforms by combining AS and alternative promoter usage (Resch et al. 2009; Song et al. 2007; Churbanov et al. 2005).

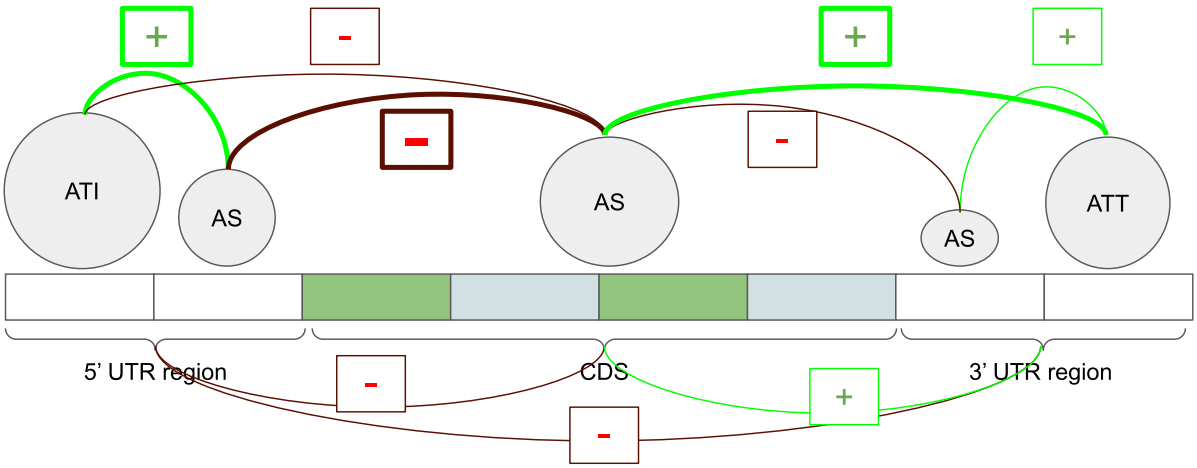


Figure 1.2: Coupling relationships between Alternative splicing (AS) and Alternative Transcription Initiation (ATI) and termination (ATT) processes. The rectangular block represents Gene Architecture, where colored segments are CDS and exons, and white empty rectangles upstream and downstream as UTR regions. ATI, AS, ATT circle size represents the approximate proportion of their prevalence in those regions in introducing alternate exons. Connected arrows are colored green and brownish red to depict positive and negative coupling, and their thickness indicates the strength of those couplings.

1.4.3 3' UTR

Alternate exon fractions as 3' UTR ends are less than that of the 5'UTR region and may be a consequence of high abundance of transcription termination signals and sparse intron density overall (Shabalina et al. 2014; Hong, Scofield, and Lynch 2006). Nevertheless, their low preponderance is not reflected in their functionality, and they play an essential role in modulating the choice of polyadenylation sites (PAS) (Proudfoot 2016). Revisiting previously observed positive coupling of AS in the CDS and ATT in the 3′ UTR, in the context of the above, indicates coordinated regulation of these processes and concomitant expression modulations as distinct transcript expression signatures of several tissues. Gene prevalence of these PAS has been observed to affect 70% of human genes (de Klerk and Hoen 2015) with an

average of two PAS per gene (Zhang et al. 2018) and has varied greatly among different studies (Derti et al. 2012; Ozsolak et al. 2010; Shepard et al. 2011). Further, the prevalence of PAS increases substantially if their signatures embeddings in introns are considered, but in normal cellular states, those are often suppressed and infrequently used (Yao et al. 2012). Inverse correlation between the increase in UTR length 3' and protein expression levels are reported in literature (de Klerk et al. 2012; Ji et al. 2011). Their positive coupling with AS in the CDS region may define the regulatory scope to introduce newly evolved CDS alternative exons and tune their expression (Ni et al. 2013) while possibly tuning its abundance compared to major isoform for evolutionary assimilation of alternate coding exons as discussed previously. In agreement with the above-suggested mechanism, detailed transcriptome-wide studies have illustrated the usage of different 3'UTR lengths in diverse tissues. Surprisingly, recently evolved neuronal genes can be observed to choose distal PAS as exemplified in the brain (Ji et al. 2009), where in addition, pancreatic islet, uterus, bone marrow, and ear also exercise expression of longer length mRNAs by preferring distal PAS. Conversely, blood, Retina, ovary, and placenta express shorter forms, likely reflecting abundant isoform expression. However, it should be noted that the correlation of distal to proximal choice of PAS to recently evolved alternate exon isoform expression is merely speculation based on data representation and needs to be investigated thoroughly to gain confidence.

1.5 AS EFFECT ON PROTEIN MOLECULES AND PROTEOME

1.5.1 AS helps rewire the proteome

Choice and dynamics of exon combinations for different transcripts of the same gene can significantly affect their properties and influence variable domain composition, distinct binding partners, intracellular localization changes, altered enzymatic activity, antagonist function, stability, and life span of protein (Resch et al. 2004; Xing, Xu, and Lee 2003; Stamm et al. 2005; Bush et al. 2017). The global impact of alternative splicing on the proteome has been extensively analyzed, where it plays a crucial role in regulating protein-protein interactions while introducing binding hotspots, specific functional motifs, and disordered regions within proteins (Hegyi et al. 2011; Buljan et al. 2012).

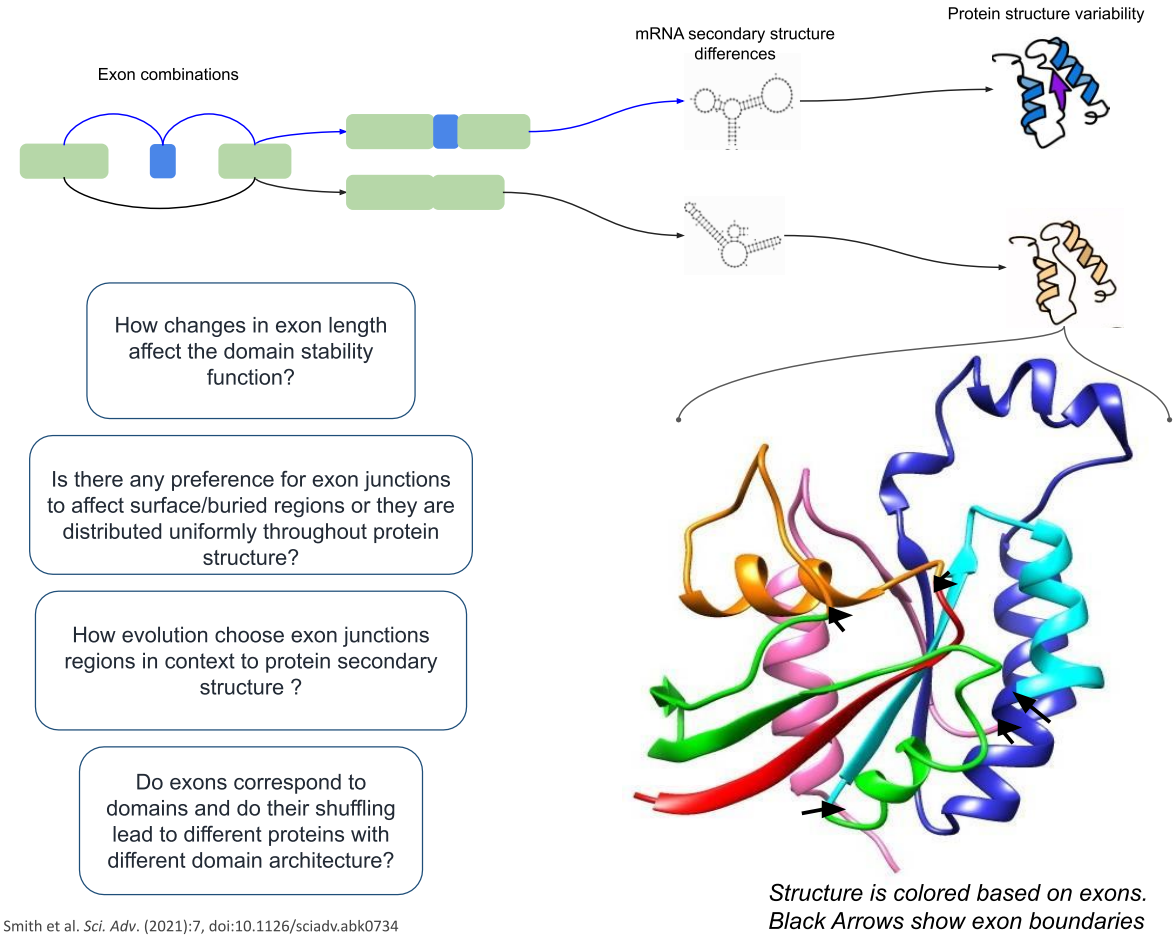


Figure 1.3: Protein accommodation scenario depiction based on dynamic exon choice. Protein accommodation scenario depiction based on dynamic exon choice. Schematic highlights alternate exon choice in transcripts and its impact on protein structure, emphasizing where exon junctions lie in the context of secondary and tertiary structure and protein packing and their possible relationships with protein domains. Exon regions are color-coded, indicating their positions within the protein structure.

AS regions preferentially encode residues located at the protein's surface (Bush et al. 2017; Wang et al. 2005), and maybe cross-indicative of them also play a role in modulating inter-protein interaction by altering isoforms. Comparisons of protein-protein interaction in AS context elucidate that intra-gene isoforms often lack half of the interactions among pairs. (Yang et al. 2016). This aspect gets more emphasized if their overall regional contribution of disordered fraction is considered and highlights alternative exon's (AE) role in adopting multiple stable configurations (Dunker et al. 2002), further strengthening the collective indications of them being associated with evolved regulatory networks in higher organisms while diversifying the proteome (Blencowe 2006; Jangi and Sharp 2014),

1.5.2 Impact of AS on individual protein molecules

In addition to global analysis, considering protein as the ultimate phenotypic product, a crucial question arises about the nature of exon-exon junctions in protein structures and how their variability is reflected when transcripts differ in exon combinations. A study by Wang et al. and others (Wang et al. 2005; Contreras-Moreira, Jonsson, and Bates 2003) revealed a preference to affect coiled regions more often than expected. This observation seems plausible, considering accommodating structured regions would be more costly and challenging than unstructured or coiled regions. This observation aligns well with a prior observation made in a study by Craik et al. (Craik et al. 1982), where exon junctions were found more frequently on the surface of proteins than in the buried regions and agrees well with our previous discussion on them likely to harbor disordered regions while helping rewiring protein interactome. Contrary to this, another study by Piwowar et al. (Piwowar et al. 2013) highlighted the crucial role of exons in providing structural stability and imparting hydrophobicity. However, the exact nature of alternate and constitutive junctions was not specified sufficiently, making direct comparisons challenging and demands revisiting in future studies. In addition to exon junctions affecting coiled residues, their preponderance towards termini was noted; and when they affect structured regions, the preference to remove the entire secondary structure (SS) unit was noted instead of altering SS unit in middle (Wang et al. 2005). Many previous studies favored data retrieval from UniProt to analyze splicing as reference genome databases were still in their initial stages, impeding limits to analyzing diverse splicing events from multiple sequence alignments of theirs. However, as data representation became more structured with exon boundaries and genome coordinates, details from complex AS events and their progression highlighted AS events and their effects on structured and conserved regions of proteins (Birzele, Csaba, and Zimmer 2008). A thorough structural analysis of AS impacts suggested an alternative mechanism to accommodate such drastic changes, displaying plasticity towards retaining similar or antagonistic functions while emphasizing protein structural robustness. In addition, the possibility of fold transitioning was also illustrated in the study by Birzele et al. in 2007 (Birzele, Csaba, and Zimmer 2008). In line with this perspective, the stability of such AS event was analyzed in MD simulation to conform stability of possible resulting modeled structure; for instance, Wang et al. in 2005 studied 16 amino acid substitutions and a 190 amino acid deletion in the *Cytochrome P450* (O64636-2) protein of Arabidopsis thaliana and highlighted the plastic character of protein while it exhibited uniform stability (Wang et al.

2005). The impact of alternative splicing (AS) on structured regions of proteins was not limited to only studying secondary structured regions or buried-exposed areas but was also extended to its relationship with protein domains, specifically their intersection with exon boundaries. A study by Hegyi et al. in 2010 (Hegyi et al. 2011) investigated this aspect and observed significant differences in domain truncation size during alternate selection of exons by AS. Interestingly, they found AS events showed a preference to avoiding globular domains altogether; however, when they affect, a tendency to preserve the hydrophobic surface area was noted, that may consequentially alters stability of resulting protein structures minimally.

These findings indicate that alternative splicing considerably impacts protein structure and domain organization while preferring to preserve stability and hydrophobic regions and exhibits control over protein structure-function relationships. Additionally, it should be noted, as many of these studies were conducted before the extent of AS and ATI/ATT regions were thoroughly analyzed on generating mammalian transcript diversity, it is unclear whether such regional preference was taken to be distinguished for those distinct processes, or in other words, whether many of observed events were introduced by AS or ATI/ATT.

1.5.3 Past disagreements over the extent of AS on the proteome counterpart

The considerable impact on protein structure from modeling analysis of spliceoforms from reference databases has been a concern in the field. Disagreement and contrasting views in favor and against the potential protein realization of spliceoforms are prevalent in literature, where a study by Tress et al.(Tress, Abascal, and Valencia 2017) termed splicing-impacted structural changes as more of a revolution than evolution while emphasizing potentially deleterious impact of AS events on protein structure and comparing those with subtle stepwise changes like evolutionary forces does. Follow-up studies discussing similar impacts have been repeated in the literature on diverse themes, where the study of (Birzele, Csaba, and Zimmer 2008) proposed alternative adjustment mechanisms to maintain folded states and relate them to diverse functional variations they may encompass, including activities of antagonism. In addition to concern about the structural reorganization of splice isoforms, disagreements between proteome realization of transcriptome splicing extent were also noted and largely stem from proteogenomic integration challenges and detection differences between protein and RNA. Proteins, unlike RNA, cannot be readily amplified, and their quantification differences have instilled a considerable lag in realizing the true extent of splicing. Since long, splicing

products were quantitated at the protein level using mass spectrometry, which though is a powerful tool to detect proteins, however not suitable to quantitate peptides corresponding to splice isoforms, as subset quantification of peptides mapping to unique exonic junctions of mRNA and their sufficient quantity to turnover in protein expression is a nontrivial task. This lag of appropriate, sensitive methods and its integration with transcriptome has also been discussed in the literature, where the study of Tress et al. (Tress, Abascal, and Valencia 2017) has scrutinized numerous LC/MS-MS proteomic experiments encompassing 100 human tissues in eight large-scale studies from several human cell lines and stages. Their study concluded to emphasize the predominant presence of a single main isoform per gene, where only 0.4% of all detected fraction correspond to alternate isoforms, underscoring limited concordance between the transcriptome and proteome. Nonetheless, a distinct letter by Blencowe BJ (Blencowe 2017), highlighted potential limitations associated with their study, where in addition to detection criteria, over-utilization of stringent filtering cutoffs was underscored, that may have potentially masked the expression of alternate isoforms. Despite being hardly 5 years old, their discussion re-emphasizes layer of challenges to integrate information from protein and mRNA abundance. Suggestions to improve protein quantification were also discussed, where Wang et al. (Wang et al. 2018) highlighted the usage of chymotrypsin over trypsin in the peptide digest stage for isoform quantification in mass spec, as the latter's target sites often match with exon junctions. Later with alternate technology development to quantify protein evidence of spliceoforms, ribosomal profiling studies confirmed evidence of mRNA spliceoforms diversity specific to cell and tissue-specific (Sterne-Weiler et al. 2013). Additionally, separate studies utilizing such ribosomal engaged fraction elucidate 75% of exon skipping events (Weatheritt, Sterne-Weiler, and Blencowe 2016). Although the precise profiling of exon skipping and other splicing events is yet to be elucidated entirely at the protein level, similar technology development and advancements in mass spec seem promising in addressing those.

1.6 CHALLENGES

In addition to discordance between high throughput RNA and protein level detection of spliceoforms, considerable other challenges can also be noted and are discussed below.

1.6.1 Lack of consistency in the databases:

Annotation models of genes in different reference databases show disagreement between their gene models, genomic coordinates, count of transcripts, protein-coding subsets, and even exon definitions, providing noteworthy challenges to the research community to choose one. It is worth mentioning that their discordance does not necessarily mean one is superior to the other, but differences in their algorithmic pipelines to annotate assembly. This also indicates the complexity and challenges in characterizing the genome. The breadth of this annotation problem can be further understood when two of those major databases (NCBI and Ensembl) had to come together almost 2 decades after the first Draft release of the human genome to partially address it by providing common standard annotations for a subset of transcripts in human genes under project MANE (Yates et al. 2020). Differences between those reference databases have been discussed in literature. Their comparison indicated that GENCODE has a high transcript average count per gene, and NCBI/UCSC and CCDS report a much lower average (Harrow et al. 2012; de Klerk and Hoen 2015). Their consequential impact on researchers using a preferred gene/transcript set to design experiments, report results, and communicate their findings is considerable. The impact of such differences on reference-based RNA seq assembly and differential gene expression is also dramatic, as discussed in (Zhao and Zhang 2015). This gets further exacerbated considering the sheer quantity of genomic and transcriptomic data being generated but lag in proteomic extent elucidation (Rodriguez et al. 2013) and represents a complicated scenario leveraged by enormous challenges in annotating the eukaryotic genome.

1.6.2 Co-occurrence with other processes

We highlighted several challenges and discussed disagreements regarding the splicing extent of proteome and individual protein molecules. However, those increase manifold when the specific impact of either AS or ATI/ATT has to be considered, as they are tightly knit in higher eukaryotes (Shabalina et al. 2010). A study in 2009 by Resch et al. (Resch et al. 2009) indicated that gene prevalence for 5'UTR varied from 12% (Nagasaki et al. 2006) to 22% (Modrek et al. 2001). For alternative promoters, it ranged from 10% (Zhang, Haws, and Wu 2004) to 18% (Trinklein et al. 2003). Their combined impact can be exemplified considering the case of gene *Dicer* (involved in RNA interference), which expresses several transcripts with a considerable number of upstream AUGs (uAUG) (Resch et al. 2009), and its transcripts show decreasing

translation efficiency with an increasing number of uAUG like we previously emphasized (Irvin-Wilson and Chaudhuri 2005). Similar phenomena were also noted for multidrug resistance-associated protein 2 (Mrp2) (Zhang, Li, and Vore 2007) and endothelium-specific receptor tyrosine kinase (*Tie 2*) required for blood vessel maturation (Park, Lee, and Pelletier 2006). Considering many of those uAUGs are conserved and show gradient continuum impact on translation efficiency with their conservation (Resch et al. 2009), individual impact elucidation of only AS may underappreciate complex regulation influencing that gene and its isoforms. For instance, it can be speculated for previously considered drastic modeling observation (Tress et al. 2007) to have splicing events in 5' and 3' direction, concomitantly impacting their translation efficiency and abundance while compensating its existence. Their impact elucidation in isolation may also be a consequence of Disagreements between fields for a considerable role in generating tissue-specific transcriptome signature, where Reyes et al. (Reyes and Huber 2018) favored ATIT and Buljan et al.,, (Buljan et al. 2012) favored tissue-specific cassette coding exons. Additionally, complex regulation of ATIT and alternate promoters were also found to be heavily influenced distal with enhancer regions (Consortium et al. 2014). In summary, functional elucidation and system-wide role of AS events and resulting isoforms are challenging to be realized from isolated impact elucidation studies and need more system biology inspired study designs involving influence of diverse processes influencing generation of different transcripts in gene, to synergistically govern impact of those on proteome and transcriptome.

1.7 Objectives of thesis

Broadly above listed challenges can be classified into two different themes:

a) Challenges pertaining to the fate of exons in different transcripts, their integration to transcript structure, and assimilated roles in CDS and UTR region.

b) Functional assessment of transcripts undergoing AS and the importance of considering the region impacted by ATIT while studying protein consequences from AS.

In this thesis (following 3 chapters), I have thoroughly studied theme a), which will help in future studies to understand more about theme b).

In Chapter 2, we have developed a framework to systematically characterize the exons and their protein-coding implications from NCBI's RefSeq resource. While this approach might not

encompass all variations and exhaustiveness of RNA seq data, it serves to prevent potential inaccuracies that may have been generated on its integration in a proteogenomic manner, which we previously discussed as the reason for disagreement of its extent. Later, in this chapter, leveraging the utility of our designed nomenclature, we have annotated splicing events in proteomes of 5 distinct model organisms and documented them in a visually appealing database (ENACTdb) available at URL: http://www.iscbglab.in/enactdb

In Chapter 3, we have compared the extent of splicing events and observed emerging trends of genes harboring a distinct population of 'Dual' and 'UTR' exons with subtype preference of alternative exons, indicating co-acting roles of AT and AS processes with observation of them also impacting CDS regions along with UTR regions.

In Chapter 4, the detailed extent of AT and AS processes on the human genome, as annotated by RefSeq, has been explored with emphasis on gene architecture and its coding region. Their contribution to imparting intra-gene transcript variability functional association by assessing Pfam domain contributions was also noted.

Apart from analyzing AS and AT extent, we have performed modeling and simulation studies in the last two chapters of the thesis. In Chapter 5, we modeled β-sheet nanocrystal regions in spider silk, which are responsible for the ultimate tensile strength of silk with beta-sheet favoring amino acids. Through multiple SMD pull studies on modeled nanocrystals (having beta sheet favoring residues), we found that naturally occurring sequence of silk achieves superior mechanical strength by optimizing side-chain interaction, packing, and main-chain hydrogen bond interactions. In another study (chapter 6), we investigated the role of Y321 in oligomerization of Vibrio cholerae Cytolysin toxin. MD and network analyses showed that Y321A mutation leads to a drastic change in network communities, suggesting a possible loss of coordinated motion required during oligomerization.

Chapter 2

Exon Nomenclature Annotation and Classification in Transcripts (ENACT): A framework to uniquely annotate exons and transcripts of genes

2.1 INTRODUCTION

Alternative splicing (AS) process generates transcript diversity and contributes to proteome diversity (Nilsen and Graveley 2010) in eukaryotes. Previously, the role of alternative splicing in transcriptome diversity has been extensively characterized using genome-wide microarray and RNAseq analyses. However, studies on the contribution of AS to the extent of proteome diversity have been limited due to technological challenges (Rodriguez et al. 2013; Sebestyén, Zawisza, and Eyras 2015), which have also possibly underestimated importance of alternative splicing in generating proteome diversity (Blencowe 2017; Tress, Abascal, and Valencia 2017). The recent technological advancements in proteogenomics and ribosomal profiling methods have unraveled the extent of proteome diversity. Moreover, comparing isoform expressions among disease states allows for detailed insights into their effect on biological processes (Zhang, Wang, et al. 2021; Ji 2018).

Alternative splicing is broadly classified into following four events (Marasco and Kornblihtt 2023): a) exon skipping (ES), where one or more exons are excluded from mature transcript; b) mutually exclusive events, in which two non-co-occurring exons in mature transcripts are referred to as mutually exclusive; c) alternate splice sites, wherein an exon harbors more than one splice site, which can be at 5’ (5SS) or/and 3’ (3SS); and d) Intron retention events (IR) are those where a region between two exons is retained in a mature transcript. Although AS events are described mainly in the context of coding exons, the untranslated regions (UTR) or partially

coding regions can also undergo similar events (Tapial et al. 2017; Leppek, Das, and Barna 2018). In the latter cases, these events usually do not result in changes in protein sequences. Nevertheless, can have other consequences arising from alterations in mRNA sequence such as sequence modification affects its susceptibility to non-sense mediated decay, has effect on translation rate due to changes in mRNA secondary structure, and can generate non-productive ORFs (Tamarkin-Ben-Harush et al. 2017). Furthermore, alternative splicing of UTRs may influence the choice of the promoter or polyadenylation site, which can potentially result in truncated ORF with consequences on its regulation and localization (Xin, Hu, and Kong 2008; Mayr 2017). The AS events occurring in the partial coding exons may also impact translation initiation or termination sites leading to protein sequence changes mostly at N or C termini (Ren et al. 2020; Tasic et al. 2002; James and Smyth 2018). More intricacies are observed when isoforms are generated from alternate ORFs or upstream ORFs (uORF), which differ from the main ORF of the gene (Kochetov 2008) and can potentially modulate termini region of isoforms or rate of translation (Wang et al. 2005; Shabalina et al. 2014). Among many verified spliced variants, ES is the most prevalent of the events and usually combines with other AS events giving rise to complex alternatively spliced variants (Climente-González et al. 2017; Reixachs-Solé and Eyras 2022). Many AS events can alter the reading frame, affecting the protein sequence of alternatively spliced transcripts.

Over the years, combined experimental and computational efforts have led to detailed documentation of AS events in databases such as NCBI (O'Leary et al. 2016), Ensembl (Cunningham et al. 2022), and eukaryote specific databases such as UCSC genome browser (Lee et al. 2022). These provide well-annotated representations of gene transcript(s). However, a systematic comparative analyses of various alternative splicing events such as prevalence of AS events across eukaryotes, tracing evolutionary conservation of exon in orthologous genes, and analyzing effects of AS types on isoforms features pose significant challenges. One of the approaches to tackle this challenge is by uniquely annotating exons based on their AS event(s). Previously, there have been limited attempts to identify exons with features linked to them. For instance, ASTRA and ALTAVISTA used concepts of naming a bit matrices to define AS and Alternate Translation Initiation (ATI) or Alternate Translation Termination (ATT) events followed by their conversion to a decimal system or symbolic event designation (Foissac and Sammeth 2007; Sammeth, Foissac, and Guigó 2008; Nagasaki et al. 2006). These databases majorly focused on characterizing the pairwise comparisons of transcripts or identifying local events. Moreover, description was complicated by incorporation of AS events from

genome/transcriptomic features. In an another approach, ASPicDB database developed exon repository, which documented multi-exon gene protein variants with their various predicted properties (Martelli et al. 2011). Unfortunately, some of these databases/tools are no longer maintained or updated with the latest data.

In the present chapter, we describe a standardized framework system developed for exon nomenclature and classification that are also mapped to amino acid sequence and associated features. Our innovative framework approach, Exon Nomenclature Annotation and Classification in Transcripts (ENACT), uniquely identifies and annotates exons based on its AS event(s) observed in alternatively spliced transcripts. Further, the amino acid sequence is mapped to exons simplifying the process for associating the predicted protein features such as secondary structure, domains, and disordered regions to exons. Such a description enables annotation and analyses of any incurred change(s) in exon features across various alternatively spliced transcripts. Each exon entity in ENACT is assigned a 6-character unique descriptor, wherein each character represents an attribute denoting AS event(s). The characters of exon descriptor in the order from left to right designates their following property: a) amino acid coding status of exons in transcripts such as coding/non-coding, b) amino acid sequence variation encoded in exons having the exact genomic coordinates, c) inclusion frequency of exons categorized as constitutive (present in all transcripts), or constitutive-like and alternate, d) the relative position of an exon in a gene, e) 5' and/or 3' exon splice site variations and the last character denotes f) counts of various observed splice site variations. The ENACT unique exon descriptor enables visualization or representation of each isoform as a combination of exon descriptors, alignment of isoforms using relative exon positions as equivalent aligned positions, and investigation into various types of AS events within or across genomes. Most importantly, exon descriptors can be documented in relational databases, allowing fast and easy computational analyses to investigate the abundance of various AS events in genomes. Moreover, in the era of proteogenomics, an approach to describe transcript(s) through annotated exon (s) can greatly assist in mapping the isoform diversity of a gene in an organism.

2.2 MATERIALS AND METHODS

2.2.1 Data source and association of sequence features

The exon coordinates were obtained from the NCBI RefSeq database in the gene table format and corresponding protein sequences of isoforms were retrieved using Biopython's (Cock et al. 2009) Entrez Efetch API's. These were assigned ENACT identities and a detailed description of the exon identity assignment Enact algorithm is described in section 2.2.3. The associations of protein features were performed after sub-setting their protein counterpart from listed protein identifiers in gene tables. We predicted following features of translated proteins (isoforms):

a) Pfam domains (Mistry, Chuguransky, Williams, Qureshi, Salazar, Sonnhammer, Tosatto, Paladin, Raj, and Richardson 2021) for each isoform were predicted using PfamScan (HMMERv3.2.1) (Madeira et al. 2022b).

b) Secondary structure prediction was performed using modified PSIPred (Jones 1999) from I-TASSER package (Roy, Kucukural, and Zhang 2010).

c) The disorder region was predicted using IUPred3 (Erdős, Pajkos, and Dosztányi 2021), where a score >0.5 was used to predict residue-level disorder.

We preformed annotations of exons in genes encoded in genomes of representative organisms *viz. Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, Mus musculus*, and *Homo sapiens.* The annotations were stored in a database for easy retrieval and presented with intuitive interactive interface in the ENACTdb. Detailed representations of transcripts are discussed in the results section.

2.2.2 Definitions of ENACT framework of exon

We have designed an intuitive exon nomenclature that enables an easy tracking of exon with their annotations and is human interpretable and suitable for automated computational analyses. Each exon is assigned a unique descriptor or Exon Unique IDentifier (EUID) comprising of six characters, each encompassing its one distinct attribute delimited by a dot ('.'). The detailed characteristics and information ingrained in these six characters with their notation are illustrated in Figure 2.1 and details are discussed below:

1. Amino acid coding attribute: Exon is assigned an attribute based on whether it harbors coding genomic coordinate *i.e.,* codes for amino acid sequence. The exon can be coding or non-coding in all or some transcripts. Based on these, amino acid coding feature is depicted by an alphabet character defined as follows:
    - U: shows an exon remains a part of the untranslated region (UTR) in transcripts whenever it is present in them.
    - M: is for a single nucleotide protein coding exon in Coding Genomic Coordinate (CGC) span.
    - T: depicts that an exon always contributes amino acids sequence or is constituent of Coding sequence (CDS) in whichever transcript it occurs.
    - D: shows that an exon consists of CDS in at least one transcript and is defined as UTR exon in at least one another transcript.
2. Coding status of an exon: It describes the amino acid sequence variations of an exon region having same genomic coordinates. The sequence variation could arise due to alternate promoter site, alternate translation initiation/termination site, and Frame Shift Events (FSE) in an exon region. Such changes are accommodated in nomenclature by a numeric character with following definitions:
    - -2: exon contributes no amino acid sequence
    - 0: as a placeholder to include M case (single nucleotide exons)
    - -1: depicts premature stop codon in upstream exon, hence, this exon does not contribute to the amino acids sequence, even though it has more than one nucleotide in CGC.
    - 1: exon contributes amino acids to the isoform (transcripts).
    - ≥2: it is a counter for the number of different amino acids observed for exons in comparison to reference isoform.

    It is pertinent to mention that we do consider variation in amino acid sequence for an exon while assignment as described later in the section 2.2.3 (B). But these are not considered in analysis as sequence variation.

3. Exon occurrence frequency: We computed weighted inclusion frequency (WIF) of exons as a ratio of exon occurrence in transcripts by the total number of transcripts. The WIF is discretized to groups exon into following categories:
    - G: exons having WIF of 1. These are constitutive exons

- A: exons having WIF < 1, are alternate exons
- F: exons having 5'/3' splice site variations but are present in all transcripts. These are also referred to as constitutive-like exons.

  Importantly, exons in single transcript genes are annotated with 'G'.

4. Exon relative position in a gene: We obtained non-overlapping exons or anchor exons as a Reference Set of Exons (RSOEx) in a gene by collating exonic region from all transcripts and processing by employing Enact algorithm as described in the section 2.2.3 Ai. The exons in RSOEx are sorted numerically based on their position of GC and are assigned relative positions starting from one to N (total number of exons) in a gene.
5. Alternate Splice Site (ASS): As described above, splice site variations (both at 5' and 3' with respect to exons) can be given with respect to exon GC of reference exon or parent exon listed in RSOEx. Importantly, the positional variation is described with respect to GC of exons in RSOEx. The splice site variations at 5', 3' or both in ASS is denoted with an alphabet described below:
   - n: denotes a 5' splice site change that can lead to either extension or shortening of ASS exon length keeping its 3' splice site unchanged.
   - c: shows a 3' splice site change leading to extension or shortening of exon length keeping its 5' splice site unchanged.
   - b: when both 3' and 5' splice sites are changed leading to extended or shortened exons with respect to exon in RSOEx.
   - 0: is used to describe the parent form of exon with its GC.
6. Occurrence of splice site changes: The last character of EUID is a numeric character, which is the count of alternate splice sites observed for an exon. We keep the default exon count value as 0 and it is incremented by 1 when ASS variation is observed for an exon. In essence, the number of ASS events either n/c/b can be obtained by looking at the number and total instances of exon ASS of an exon is summation of all variations observed for n/c/b cases.

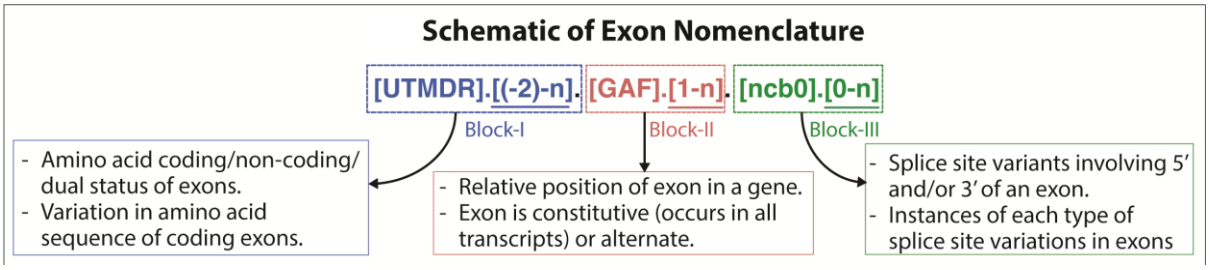


Figure 2.1: Exon nomenclature descriptor. Six characters encoding scheme for exon is shown with three blocks of description. Block-I shows pseudo-global feature, block-II embeds global information of relative exon position and its constitutive/alternate and block-III has information of alternative splice sites and their occurrences.

Intron retention (IR) events: These are considered as a special case for exon nomenclature as retention involves two exons and six character descriptor will be insufficient to capture its details. Therefore, we have incorporated additional characters to describe IR exons. To describe this, we use five identifiers combined with ':' colon symbol to discern from standard EUID. The first identifier is the alphabet 'R' to recognize that the exon is involved in IR, followed by a digit describing its amino acid coding attribute (the same notation is used as described before). The third and fifth identifiers are exon EUIDs, between which the intron/exon region is retained to form IR exon. The fourth identifier is a numeric character showing the number of retention events observed involving exons and their variants. We use 0 as the default value of this counter. An example of IR exon descriptor is shown below:

R:1:U.-2.A.**2**.n.1:0:T.1.A.**3**.0.0

The above IR exon depicts it as an amino acid coding exon, which consists of region retained between exons U.-2.A.2.n.1 to T.1.A.3.0.0. It is the first instance involving exons 2 and 3 (shown in bold as these are the relative position of exons in a gene).

Each EUID character uniquely identifies an exon's attribute either observed across transcripts (global) or seen in a transcript (local). For instance, the first character (amino acid coding) is a global feature, as the exon having 'T' at this position in EUID would show it is amino acid coding in transcripts wherever it occurs. On the contrary, the fifth position encodes a local property as it captures the alternative spliced site events occurring in a transcript. For each of interpreting exon features, we have grouped two consecutive characters of EUID into three sections (Figure 2.1). The first section (block-I) identifies exon's "coding status and amino acid variations" and has pseudo-global information. The second (block-II) contains information about the "relative position of an exon with its inclusion status". As it requires knowledge from

all transcripts, block-II encompasses global information. The last section (block–III) documents "all possible splice site variations with their counter" and it has local information content. Notably, the global information (block-II) of an exon will remain invariant across transcripts.

2.2.3 Description of ENACT algorithm

We implemented the exon extraction, decoding its feature in transcript followed by annotation using the nomenclature as described in previous section (2.2.2) in the ENACT algorithm. Briefly, we obtained the Genomic Coordinates (GC) and Coding Genomic Coordinates (CGC) of all protein coding transcripts in a gene from the NCBI database available in the gene table format. Since we are interested in analyzing the effect of AS events on protein sequence, we have mapped amino acid sequence to exons for each isoform. For mapping amino acid sequence of an isoform to its constituent exons, we consider non-overlapping triplets (three nucleotides or codon) from a transcript and assign amino acid to it. Since the length of coding genomic coordinate of Exon (L_{ex}) may not be multiple of 3 (codon length), we either borrow or donate a nucleotide to the successive exon based on the output by taking the modulus of L_{ex} by 3. If the remainder is 1, the right genomic coordinate of current coding exon is subtracted by 1 and added to following exon by shifting its left coding coordinate by 1; if remainder is 2, the nucleotide is borrowed from the successive exon, and coordinates are changed vice-versa to former case. Subsequent to assignment of amino acid sequence to exons for each isoform, we proceed to apply Enact algorithm to identify and assign nomenclature to exons. The outline of Enact algorithm is illustrated in Figure 2.2 and more details of nomenclature are given in Appendix A – Exon nomenclature description. The main algorithm steps are briefly discussed below:

a. <u>Identifying and assembly of anchor exons</u>

Our nomenclature requires a non-overlapping set of exons or anchor exons in a gene to be defined such that rest can always be identified either as an overlapping or retention instances with respect to them. Uniquely identifying anchor exons is a non-trivial task, as genomic coordinates of many of them have variable overlapping region in transcripts.

In the first step of Enact algorithm, we applied an empirical approach to define non-overlapping (anchor) exons. Since these may not be present in one single isoform, we begin with selecting an isoform having maximum number of exons using specified criteria. For a

given gene, we select an isoform having the maximum number of coding exons from a set of curated isoforms (RefSeq proteins having 'NP_' prefix) and define it as Reference ISOform (RISO). If the number of coding exons is same in ≥2 isoforms, then the one with the longest length is selected as RISO. In case, a gene has no 'NP_' prefixed isoforms, then reference isoform is chosen from all known isoforms using the criteria as described above. The RISO exons constitute the initial set of Reference Set Of Exons (RSOEx) (substep 1 Figure 2.2A). We define RSOEx exon data type as tuple consisting of GC, CGC and mapped amino acid sequence to it. The latter information is important for tracking changes in exon mapped protein sequence when both GC and CGC are same for two exons. RSOEx is populated with exons (from non-RISO transcripts) whose genomic coordinates do not overlap with anchor exons. To achieve this, we initially gathered exons from all non-RISO transcripts of the gene forming *NREx* set, represented as substep 2 in Figure 2.2A. This set is processed using *insertionFilter()* routine to identify those having GC overlap with anchor exon or select a representative among those overlapping with each other but not with anchor GC. The *insertionFilter()* routine sorts NRexon based on their length and iterates over them until the list is empty, following steps:

1. Exon having non-overlapping GC with members of NREx and RSOEx are moved from NREx to RSOEx.
2. Exons from *NREx* that do not overlap with RSOEx but do overlap among themselves are pooled together as *OlEx* and a representative member is selected as the one having a length of ≥ 10 amino acids (≥ 30 nucleotides). If no exon satisfies this condition, the smallest exon is chosen as the representative. The chosen representative member is then moved to RSOEx. The members that overlap with the representative exon are transferred to the Splice Site Relative (SSR) set, depicted in substep 3 of Figure 2.2A
3. Exons with GC overlap to RSOEx members are appended to SSR.
4. Exons having the same GC as RSOEx but differ in their amino acid sequence are moved to SSR.

Thus, the updated RSOEx consists of non-overlapping exons from all transcripts and SSR have their splice site variants. Notably, SSR may contain Intron Retention (IR) exon, which has GC overlap with at least two anchor exons (RSOEx). We used

retentionFilter() routine to identify IR exons and are moved to IR set (substep 4 in Figure 2.2A).

b. Detection of splice variants: In the step, every exon in SSR set is compared with those in RSOEx for their GC to define nature and type of splice variants or altered amino acid sequence. We consider splice site variants from exon definition model and not intron definition as former are preponderant with short exon with intervening long introns (De Conti, Baralle, and Buratti 2013b) therefore 5' or 3' in splice variants refers with respect to exon. Moreover, as ENACT framework entity definitions focus on encoded proteins, it is more appropriate to use exon definition model. In the *EventAnnotation()* routine, we used following criterion of define exon variants:

    a. If GCs of SSR and RSOEx exons are identical, they will vary in the mapped amino acid sequence because of different CGC, and frame shift. These are assigned aaChange events.

    b. If 5' GC of SSR and RSOEx exons are different and their 3' GC are identical, it is referred to as 'n' case.

    c. If 3' GC of SSR and RSOEx exons vary while having identical 3' GC, it is referred to as 'c' case.

    d. If both 5' and 3' vary in SSR with respect to RSOEx, it is referred to as 'b' splice site variation.

    The 'aaChange' and 'n/c/b' exon variation events are stored in "Event tracker", which maintains separate record of SSR exons and it relationship to RSOEx member. Apart from this, "Event Tracker" also records the number of these variant types for each of RSOEx members based on the GC. For instance, the frequency of 2 for an 'n' form of an exon means that this show has three distinct 'n' or 5' splice site variations.

c. Exon Unique Identifier assignment: In this step, each exon is assigned unique identifiers based on their features as has been described before in section 2.2.2. The "Event Tracker" data is used predominantly for the purpose of nomenclature, and it is cross-referenced to SSR/RSOEx for addressing various questions to decide identifier of an exon. These questions are: 1) Does the exon consistently remain non-coding/coding or switch between non-coding and coding states, based on its coding genomic coordinate, in transcripts

whenever it is present?; 2) Do coding exons occurring more than once have the same CGC across their occurrences, or do they exhibit change in amino acid sequence change?; 3) What is prevalence of an exon in transcripts of a gene?; 4) Does an exon show splice site variations (n/c/b) as defined before?, and 5) What is frequency of each splice site variation?

Based on exon nomenclature discussed in detailed in section 2.2.2, the above questions enable assignment of identified as follows:

a. Designating coding feature annotation (Block I): The questions 1 and 2 as mentioned before provide information to annotate the first two characters in EUID. If an exon whenever it occurs is non-coding (without CGC) in a transcript, it is assigned 'U' character. Similarly, coding exons are assigned 'T' and those which are coding in some transcript and non-coding in other is assigned 'D'. The 'M' tag is given to exons of single nucleotide. The second character is dedicated to track change in amino acid sequence change of an exon in a transcript relative to the one in RSOEx. Such changes are noted in step 2 of the Enact algorithm. The numeric code assigned as per nomenclature as discussed in section 2.2.2 (1 and 2).

b. Prevalence of an exon (Block II): As previously described (section 2.2.2 (3 and 4), block-II provides global information about an exon, *i.e.,* its relative position and inclusion in transcripts. The relative position of an exon in RSOEx is determined by numerically sorting anchor exons in an increasing order and numbering exons from 1 to N (total number of exons). Subsequently, splice variants of the RSOEx exons are assigned the same relative number. An exon that is present in all transcripts with same GC as in anchor exons is categorized as constitutive ('G'), while an exon not present in all transcripts is assigned alternate ('A') label. If an exon is present in all transcripts but has variable splice sites (listed in SSR as n/c/b), it is defined as 'majorly constitutive' and annotated as 'F' tag, indicating that all transcripts share the exonic region but not complete region.

c. Splice site variations and their frequency (Block III): The questions 4 and 5 above allow us to annotate block-III of EUID. Using 'Event Tracker', we identify splice site variants of RSOEx exon in SSR. For each exon, the variants are identified as n/c/b (discussed before), and this is the 5^{th} character in EUID. The number of n/c/b variants is assigned in the last and 6^{th} character. Importantly, the RSOEx (anchor) exons are assigned "0" for the last two characters indicating splice site variants are

derived from its GC. Exons that harbor splice site choice are only dependent or related to Block-II features and the block-I attributes could be completely independent for two splice site variants. For example, an exon EUID:U.-2.A.4.0.0 ('U' and -2 indicates it is UTR and no amino acid is associate with it (block-I); the exon is at 4th position in reference exons, and it is alternate exon (block-II) and the last 0.0 indicates it is anchor exon). The exon at same position can have EUID:T.1.A.4.c.1, which shows the 3' or 'c' splice variant of 4th exon is coding whenever it occurs in the transcript and is an alternate exon.

ENACT framework and modular construction

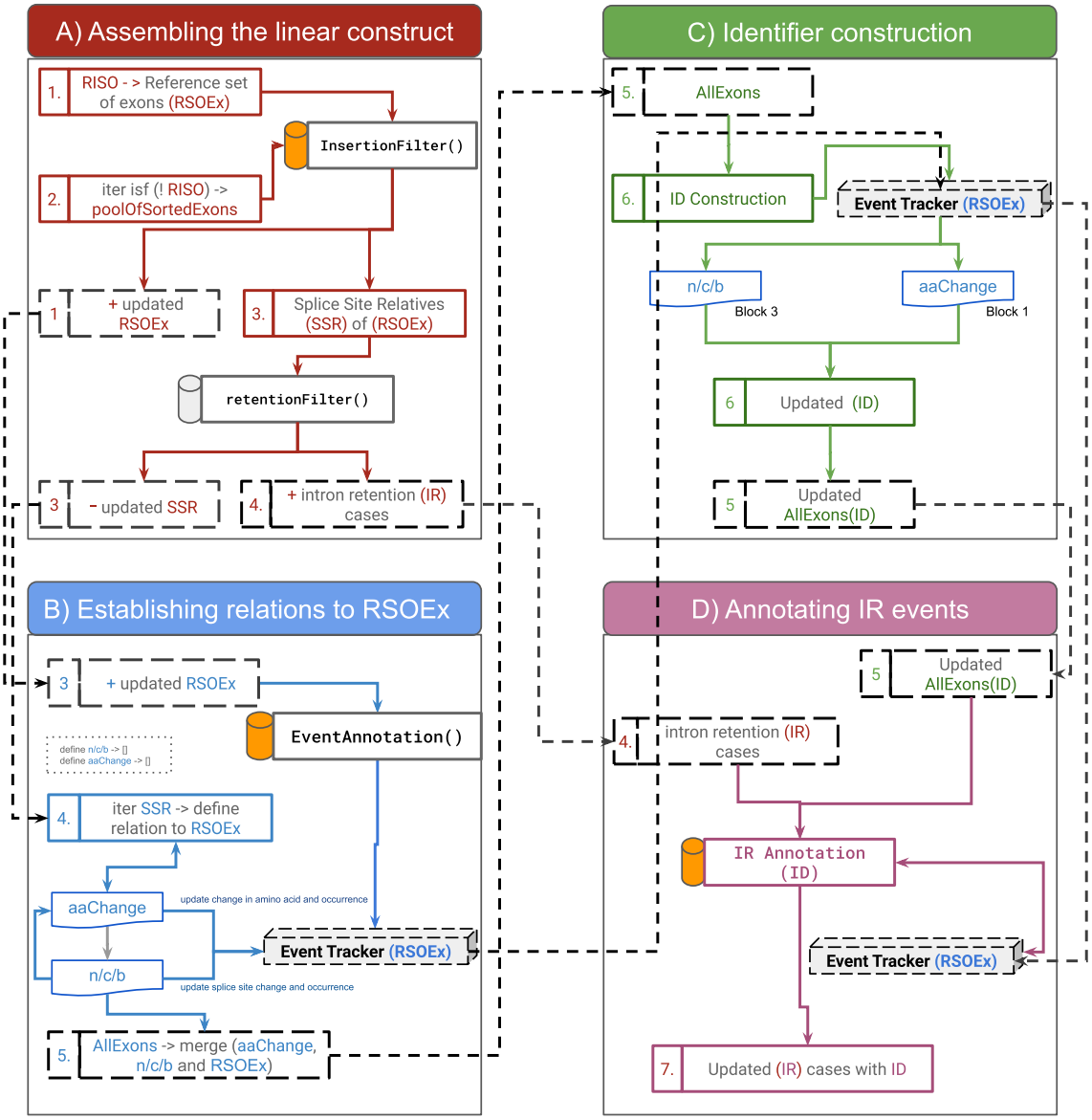


Figure 2.2: Overview of Enact algorithm. Four major steps of algorithm is shown in independent panels from A-D, listed integer (1-7) in them depicts substeps. A) The first step involves defining Reference ISOform (RISO) followed by constructing a set of non-overlapping exons (RSOEx) from RISO and other non-RISO isoforms. Importantly, exons overlapping to the set in RSOEx are defined into either Splice Site Relative (SSRs) or Intron Retention (IRs); B) In this step, SSRs are further classified into various types of splice site variants or amino acid change cases. These are carefully stored in EventTracker for further annotations; C) The descriptor of exons is constructed based on various on various information stored in EventTracker and same is updated in it; D) The final step deals with annotating IR exons. Black Dashed rectangles depict DataTypes and dashed lines donate its flow across four steps. The cylinder object refers to routine performed in the algorithm. Additional details of the ENACTdb Nomenclature are provided in the tutorial pages in http://www.iscbglab.in/enactdb and Appendix A – Exon nomenclature description.

d. Intron retention annotation: The routine *retentionFilter()* in the initial step of Enact algorithm detects exon, which show overlapping GC with more than one member in RSOEx. Such exons are listed as tentative IR exons. Since the RSOEx and SSR are already assigned exon nomenclature, in the last step EUID is assigned to IR exon. For this, we identify the first and last overlapping RSOEx exon with the IR exon based on GCs, subsequently, we check for congruence between 5' and 3' of IR with the 5' of first and 3' of the last RSOEx exons respectively. If these are congruent, then we assign EUID as discussed below, otherwise, we find appropriate splice site variant (from SSR) of the first/last exons, which matches to respective 5' or 3' GCs of IR exon. If no listed exon in (SSR/RSOEx) matches, then we define appropriate new splice variant solely for the purpose of annotating IR exon. Importantly, such new splice variant follows same protocol of annotation as in defined for SSR. Finally, EUID is assigned to IR exon based on exons with which it overlaps, coding nature and number of retention case occurrences from former exon (1 means first instance and 2 means second retention case originating and so on.

2.3 RESULTS AND DISCUSSIONS

2.3.1 ENACT representation of Alternative splicing events

Having developed nomenclature for exon, we examined the representation of various AS events using ENACT (Figure 2.3) for a hypothetical gene having all possible combination of events. We chose a hypothetical case as it is challenging to find an example showing all AS events. As shown in the example, the gene is composed of a total of 9 exons (coding/non-

coding) in it. Of these, two non-coding exons are assigned an Exon unique IDentifier (EUID) as U.-2.G.1.0.0 and U.-2.A.9.0.0. The former exon is observed in all transcripts (having same GCs) and is assigned constitutive (G) code, whereas we assign alternate (A) to this exon does not occur in all transcripts. The relative position can be inferred from the 4th character of EUID, which accordingly is 1st and 9th exons are non-coding. The coding attribute of the second exon 'D' or dual as it is part of CDS in IS-8 and non-coding in rest other isoforms. Accordingly, EUID of the second exon in IS-8 is D.1.A.2.0.0, however in rest others it is assigned with D.-2.A.2.0.0. As can be seen, the second character, which shows the local amino acid coding attribute, is changed from "-1" to "2" in IS-8. The coding exons 3 to 8 are assigned with global attribute of 'T' for coding status and G/F/A based on their occurrences in transcripts. For instance, exons 3 and 5 occur in all isoforms, but show splice site variations in some isoforms and these are assigned 'F' tag for inclusion frequency attribute. As has been discussed in methods, based on changes in exon features within an isoform its local attributes are changed accordingly. For example, exon 3 in IS-4 exhibits 3' splice site change therefore, the relevant local property in this transcript is assigned 3.c.1 showing that exon 3 undergoes 3SS and it is first instance of such a splice site variation. Similarly, exon 5 in IS-5 shows both 5SS and 3SS variations and the local attribute is modified from 5.0.0 to 5.b.1, wherein 'b' shows both 5' and 3' splice site variations with their first 'b' observation in it. We can see the splice site variations are observed for alternate exon 4 in isoforms 4 to 7, for exon 7 in IS-6 and IS-7. Based on the type of changes, the local attributes of EUID are changed accordingly. Importantly, we can keep a track of splice site changes and number of their variations such as exon 4 in IS-5, IS-6 and IS-6 undergoes 3 independent types of 5' splice site changes and is evident from the last character of EUID of exon-4 in these isoforms (4.n.1, 4.n.2, and 4.n.3).

Figure 2.3 shows AS events along with exons assigned with unique descriptor using ENACT nomenclature. As exons 6 and 7 do not co-occur, the IS-1 and IS-2 are examples showing mutually exclusive AS events for these exons. Interestingly, we can systematically identify all occurrences of mutually exclusive exons events in a gene. The exon skipping events can easily be identified as it finds when one or more exon is skipped in reference to a transcript. We have already discussed various splice site variations (5SS/3SS and 5SS with 3SS) representations in exons and the same is shown in the example. Alternate translation initiation sites can be identified either from dual exons or examining the change in the amino acid sequence of the first coding exons. Finally, the intron retention event is shown as merging of intron region between 3rd and 4th reference exons in IS-8. It is important to note that ENACT nomenclature

can even depict the IR event even between exons having other variations. Through our nomenclature, we can observe that multiple AS events can easily be shown in a single simplified representation of transcript

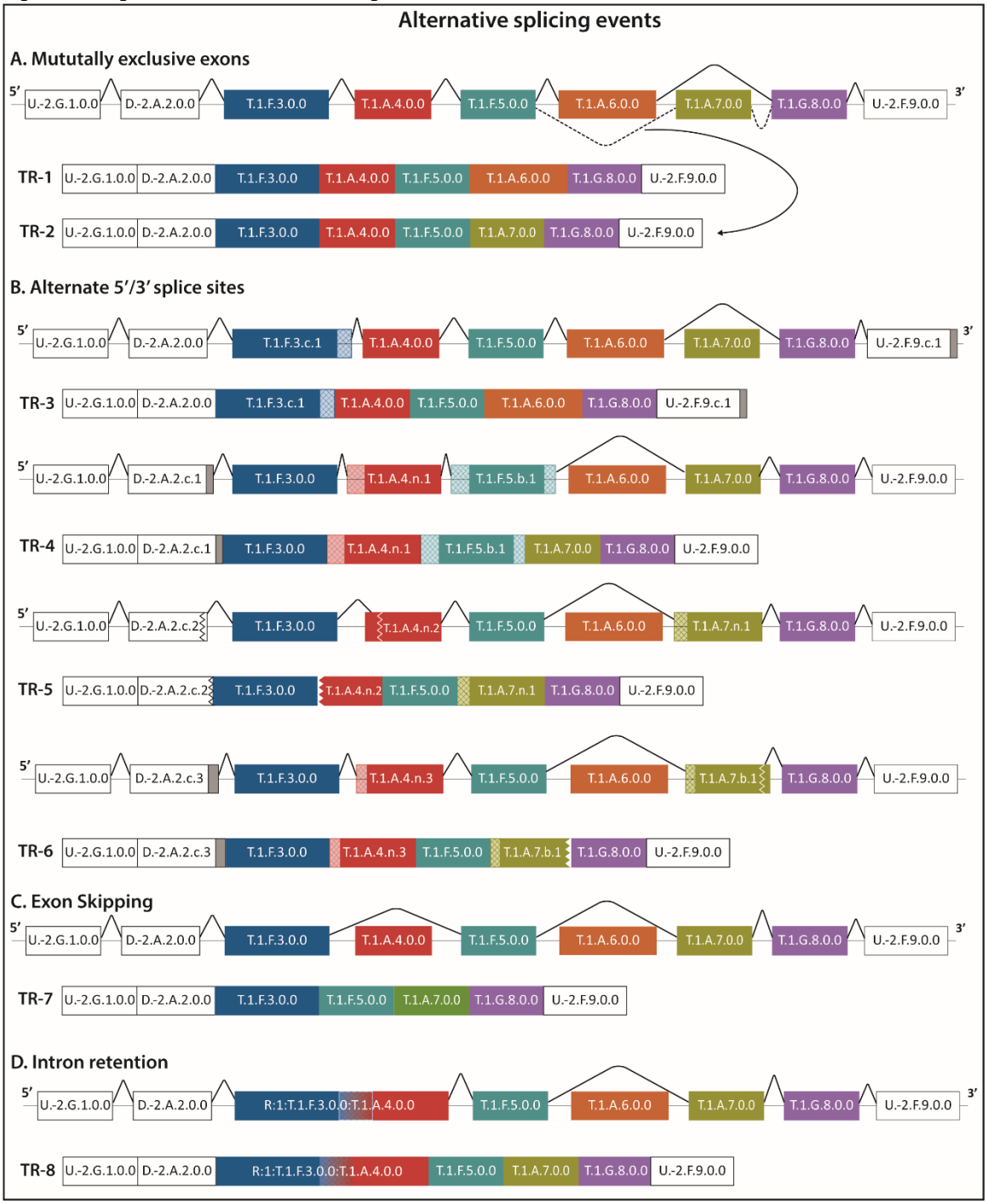


Figure 2.3: Depiction of alternative splicing events of ENACT annotated exons. A hypothetical example showing annotation of exons with ENACT in various alternative splicing events. Each coding, non-coding and partial coding exons are shown as colored, while and partial grey colored rectangular block. Different color schemes have been used to distinguish consecutive exonic entities and for easy visual interpretation.

2.3.2 ENACT exon annotation of genomes

We obtained genes from NCBI RefSeq of five representative genomes *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, and *H. sapiens* and processed genes using Enact algorithm to annotate exons in genes. The secondary structure, Pfam domain and disorder prediction of isoforms were performed as described in Section 2.2 and these are mapped on exons. The summary of genes, isoforms, and annotated exons are summarized in Table 2.1. The annotations are presented in ENACTdb database, which is available at. The ENACTdb is publicly accessible at URL: http://www.iscbglab.in/enactdb .

Table 2.1: Summary of gene, isoforms and exons annotated in representative organisms.

Organism	Number of protein-coding genes	Average number of Total (Coding) [UTR] exons/gene	Average number of total (Unique) isoforms/gene
C. elegans	19,972	6.3 (6.2) [0.1]	1.4 (1.4)
D. melanogaster	13,972	4.7 (4.1) [0.6]	2.2 (1.6)
D. rerio	26,374	10.2 (9.6) [0.6]	1.8 (1.7)
M. musculus	22,134	10.5 (9.4) [1.1]	4.2 (3.0)
H. sapiens	20,443	11.8 (10.4) [1.4]	6.4 (4.4)

2.3.3 Database design and Django manager

ENACT framework-based nomenclature entity was associated with protein features annotated per isoform (see section 2.1.1) and the corresponding data was processed and stored using in-house python script. We constructed object-oriented pickled serializers that we term parent object container (Figure 2.4A). These are transformed and designed information of exons/transcripts/genes is suitable imported in MySQL tables. The detailed database schema with relationship among entities is shown in Figure 2.4B. The database consists of 3 major tables called exonapp_genes, exonapp_transcripts and exonapp_exons. The exons are kept in separate tables because it reduces redundancy, as the same exon can be present in multiple transcripts. In general, the table '_genes', '_transcripts' and '_exons are collection of all genes in an organism, respective transcripts associated with a gene and all exons of a gene

respectively. Such a cross-reference of tables allows us to reconstruct transcripts with exon constituents easily for any given gene. The next set of tables lists exon entries (unique property rows) with a list of transcripts they are contributing. Although this increases manual mapping and transcript construction but reduces space requirements. Django server manages the backend of database such as MySQL, web server access, and fetching queries. The front-end GUI is rendered by ReactJS in a one-page application (https://reactjs.org) (Virtual DOM works by turning on/off different components and rendering content dynamically).

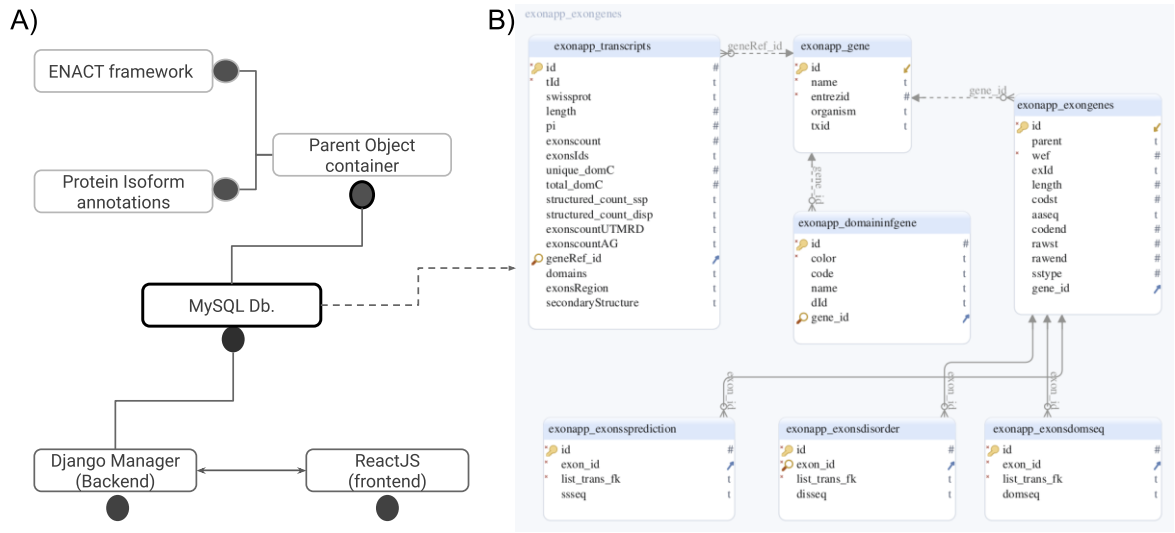


Figure 2.4: Framework to ENACTdb layout and database schema. Figure showing A) the layout, data types and flow of information maintained in ENACTdb and the panel B) has the ENACTdb relational database SQL tables and their relationship between entities of the SQL tables.

2.3.4 React JS rendered visual modules

The database can be queried with gene name or NCBI gene identifier, which results the occurrence of search term across genomes listed in the database. The user can subsequently select terms of choice and click to find a detailed representation of gene. The main gene page is rendered as a single page using native ReactJS with options to select and show content on demand from the user. The top section of the gene page has basic information about genes including the number of transcripts, coding, and non-coding exons. The bottom section of this page is divided into two subsections where the first has list of scrollable isoforms showing their amino acid length, exon count, fractions of predicted disordered and secondary structure regions (Figure 2.5). Apart from this, the names of isoforms are cross-referenced to the NCBI.

The detailed view of each isoform can be rendered in the second sub-section. To avoid cluttering this part of the page, we have enabled the 'show'/'hide' options link for isoform(s). Therefore, the views of transcripts can be shown on demand. We have developed two independent views for isoform showing as a combination of exons (mapped with amino acid sequence).

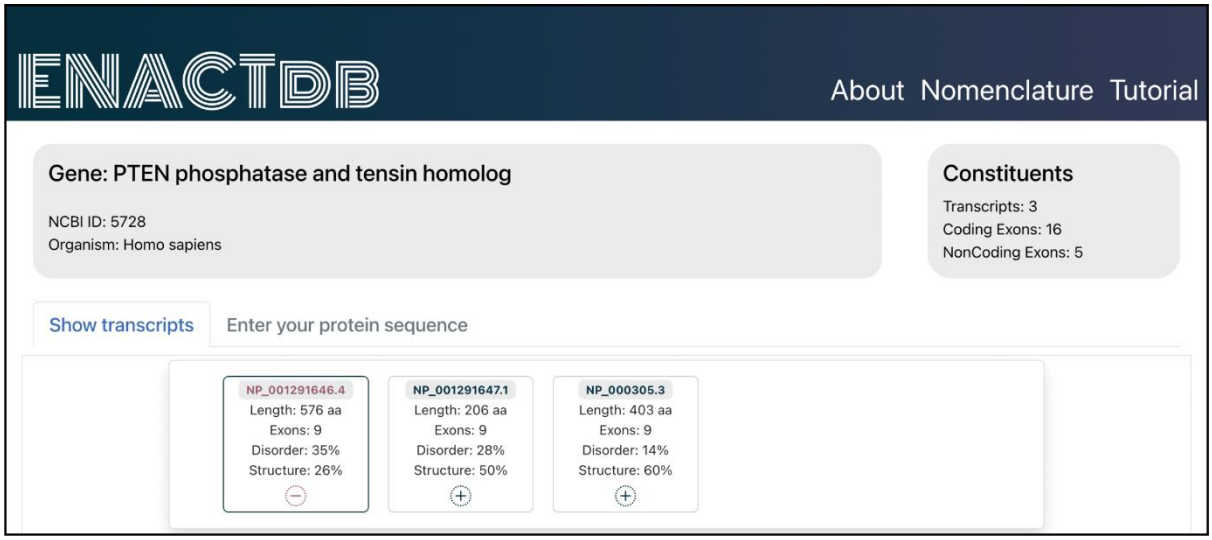


Figure 2.5: Gene page view of ENACTdb. A representative web page for a gene is shown for the human PTEN gene. It shows the number of exons (coding/non-coding) and isoforms with their length, fraction of secondary structure, and disordered region. The symbols (+) and (-) can be clicked to show and hide the transcript/isoform view and their annotated exons.

The transcript can be selected to render in block and nightingale views. The former needs selection of a feature for showing on an exon view whereas all features can be shown in one single view in the nightingale view. We have selected a gene eEF1 (eukaryotic translation factor) from *D. melanogaster* for showcasing various views obtained from ENACTdb.

2.3.4.1 Block view

It is a naive view of exons shown as exons as rectangle block with the mapped amino acid sequence on it (Figure 2.6a). The protein mapped on exons is the default view. The hovering of mouse of any exon will show EUID followed by detailed explanation of exon features on the right side of page (Figure 2.6a). We can display predicted features of proteins mapped on exon by choosing 'Show SS', which displays secondary structures, 'Show Dom' will show domains identified in the protein sequence with mapping on exon and 'Show Dis' will show

disordered regions of each exon (Figure 2.6b and 2.6c). To see any of the above features, it is important to select 'reset to aa', which will make it the default of block view.

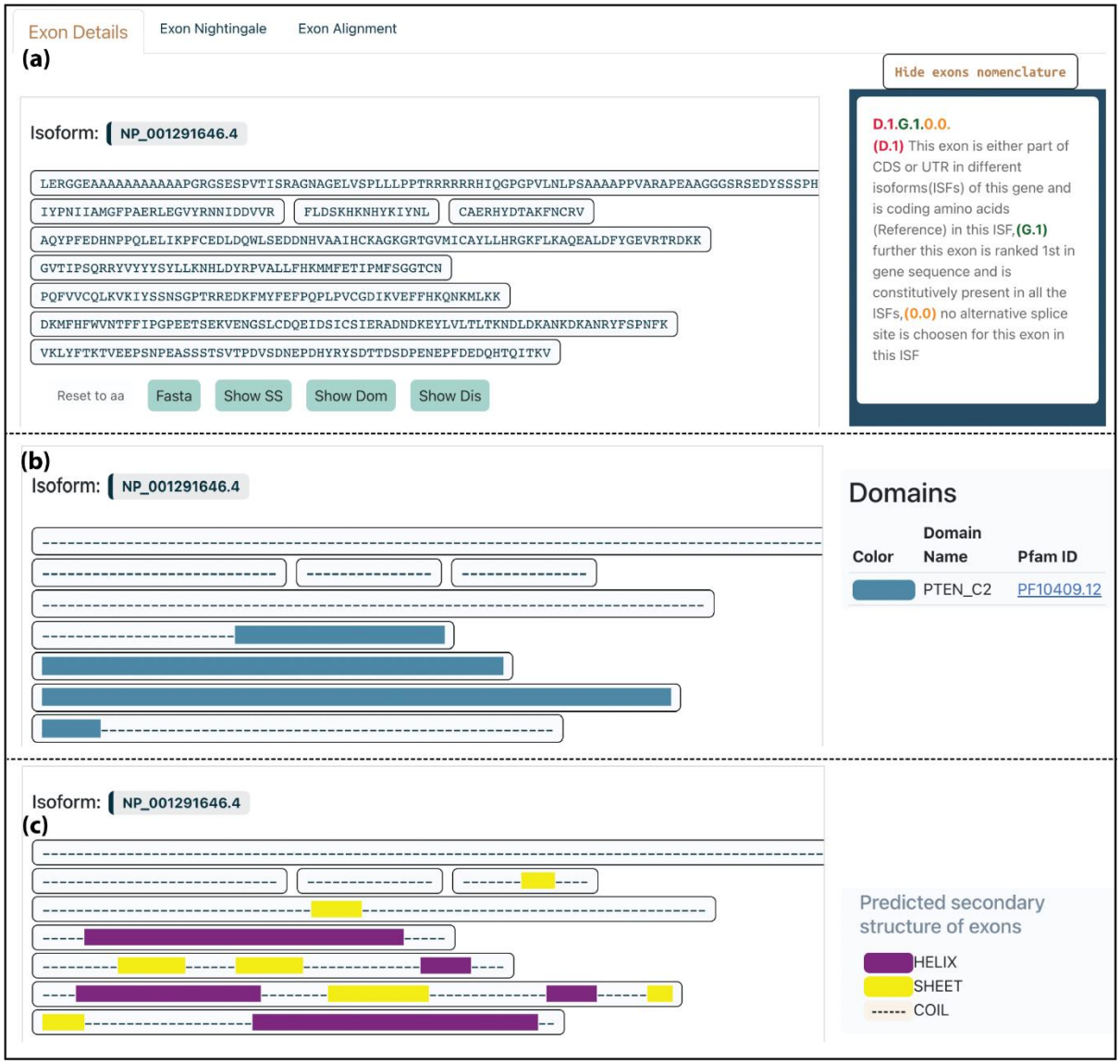


Figure 2.6: Block view of transcripts and exons. It is snapshot of ENACTdb showing the block view of transcript with exon EUIDs and other predicted features. The panel **(a)** is the default view with the exon region shown as rectangle and hovering over it displays the 6-character EUID with its detailed explanation. The view can be changed to show domain region covering the part of exon(s) as illustrated in **(b)** and it can also show secondary structure **(c)**. The view can be rotated from one to the other after clicking reset button 'reset to aa'.

2.3.4.2 ProtVista (Nightingale) view

The block has limitations as not all features are displayed in one single view. Therefore, we implemented Protvista module and implemented its nightingale view with all required

modifications in the script to make an All-In-One view representation of transcript. This nightingale view depicts predicted features on the protein sequence for each isoform. This allows all predicted protein properties to be displayed in one view allowing for better understanding as well as interpretation of exons. In this view, exons are shown as mapped regions on protein sequence, which can be zoom-in/zoom-out (Figure 2.7a). Below this, four rows display different features together. The mapped exons are shown on top with alternate colors to distinguish one exon from the next one. The next we show predicted secondary structures using arrow for strand and spring for helix views. The third is rectangle for disordered regions and the last is predicted Pfam domain(s) shown using different colored oval boxes. Importantly, we also provide the raw data in a table format where each of displayed content is marked under features. For instance, one can select 'exons' feature and it will show nomenclature, position in sequence and their lengths (Figure 2.7b). Interestingly, one can display all isoforms with all their features together on the same page.

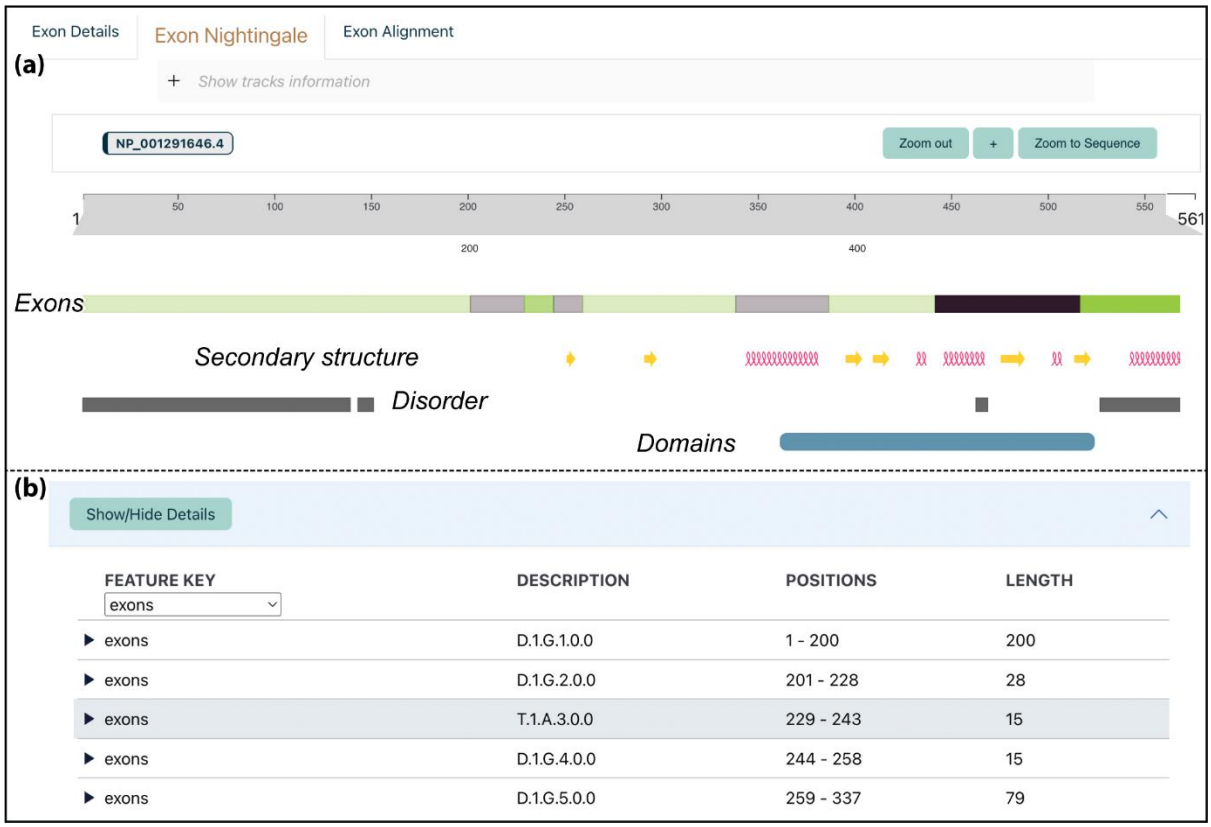


Figure 2.7: Nightingale view of transcripts and exons. A snapshot showing the nightingale view of transcript and annotated features of exon. **(a)** The amino acid sequence region can be zoom-in/zoom-out to show various sequence features mapped at the exon level concurrently. Every other exon is differently colored to distinguish from each other. The secondary structures are shown using symbols with helix represented by spring, and the strand is illustrated by arrow.

The gray rectangle region denotes the disorder region(s) and domain(s) is shown using a rounded rectangle (separate colors for different domains). **(b)** Descriptive features of the above representation can also be seen in the section below by clicking 'Show Details'. The image shows the annotation of exon with its nomenclature and position in the gene.

2.3.5 Additional features of ENACTdb

2.3.5.1 Exon Alignment View

One of the significant features of ENACTdb is the ability to display multiple alignment of exons anchored with their positions. Such a view can provide perspective of exon variations seen across transcripts and give insights into the effect of such exon inclusion/exclusion on protein predicted features (Figure 2.8a). As can be seen in the figure the second exon is missing in the second isoform, which leads to significant changes in the predicted secondary structure, which can also be appreciated in the nightingale view of the transcript.

2.3.5.2 Mapping exons on a new protein sequence

In many instances, a user may have a sequence but does not have knowledge of which exons are contributing to the isoform. Since we have mapped exon to amino acid sequence, we leveraged this information to make a mapping tool to search for an input protein sequence to exons provided a user also provides the gene from which the protein sequence is derived. In this tool, a user can provide a protein sequence, which is assumed to be due AS events, however, it is still not documented in the databases. The mapping search feature will output the exon(s) mapped on the submitted protein sequence. Here, we declare a mapping only when all amino acids of an exon are found in the input sequence. Therefore, the region of sequence, which does have an exact map to a known exon sequence, remains unmapped region. We submitted a sequence where we modified the region spanning the 3rd exon. The output showed match for exons 1 and 2, however, the last 3rd exon remains unmapped as the sequence was modified (Figure 2.8b). Our exon descriptor encapsulates its various attributes such as relative position, amino acid coding property, inclusion in transcripts, and splice site variations. Such descriptions facilitate enabling systematic computational analyses of occurrences of exons features in a genome or their comparison across genomes providing possible insights into evolution of alternative splicing and their contribution to transcriptome/proteome diversity.

2.4 CONCLUSIONS

In the present work, we have devised an innovative approach ENACT to annotate exon by assigning it a 6-chatacter descriptor from an observed pattern of alternative splicing events across transcripts. The ENACT nomenclature facilitates an easily illustratable representation of transcripts as a combination of exons EUIDs and allows for detailed interpretation of AS events. Significantly, it would simplify computational analyses for deciphering alternative splicing properties within or across various organisms. The integration of predicted features from isoforms to exons enhanced the information content, which could be used to gain insights into the effect on protein secondary structure or domains due to various AS events. Our nomenclature and Exon segmented protein properties could assist in exploring the exon-specific role in protein evolution, especially multidomain proteins.

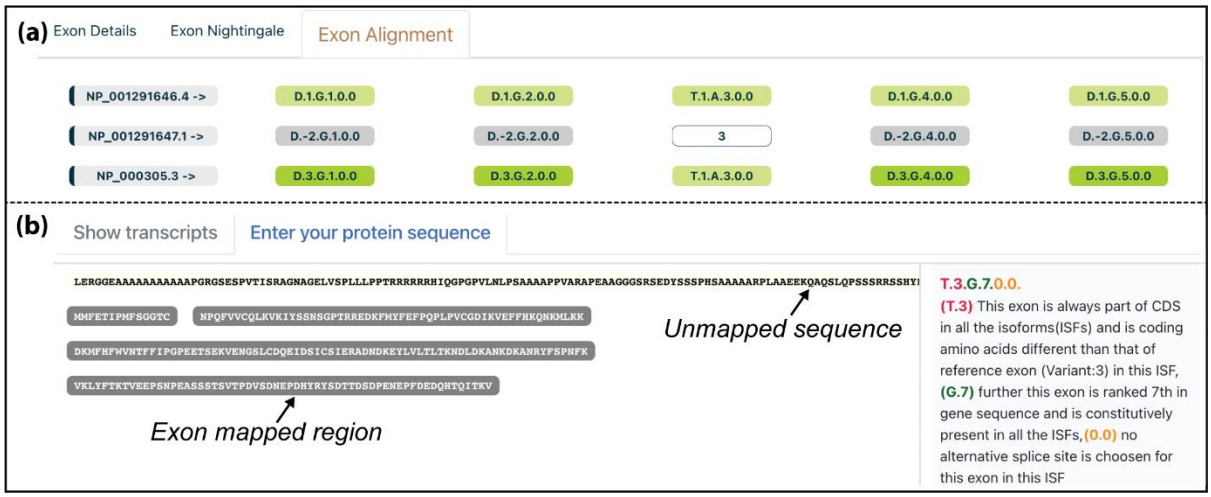


Figure 2.8: Exon alignment and protein search view of ENACTdb. (a) The figure is a snapshot of ENACTdb showing the exon alignment of 3 PTEN transcripts. It is important to note that alignment is anchored to the exon's relative position. **(b)** The output of exon map tool is as a result of searching a N-terminal edited PTEN sequence. The initial sequence could not be mapped because it was modified, however, the rest of other sequence is mapped to exons and shows facility to map a new protein sequence to exons of a gene.

We have provided exon annotations of genes encoded in the genomes of five model organisms in ENACTdb database. The exon attributes of a gene can be displayed in two different detailed views. The exon alignment can be performed to reveal the inclusion/exclusion of exons across transcripts providing a better representation and comparison of transcripts or isoform features. We provided a mapping tool to query a gene using a protein sequence to find known exons. If there are unmapped regions, it could be a

potentially modified region in protein due to AS event. In summary, our database allows enhanced visualization of isoforms having protein sequences and associates predicted features mapped to exons.

Chapter 3

Distribution of AS events in representative genomes:

An evolutionary perspective

3.1 INTRODUCTION

In the previous chapter, we established a standardized framework, ENACT, designed for uniquely annotating exons. This framework facilitates comparison of Alternative Splicing (AS) occurrences within individual organisms and across species. Using ENACT nomenclature system, we proceeded to annotate genes in five prominent model organisms: *C. elegans, D. melanogaster, D. rerio, M. musculus,* and *H. sapiens*. The comprehensive annotations of exons in these organisms are documented in the ENACTdb database.

Typically, alternative splicing events are predominantly assessed through pairwise comparison of transcripts, posing a difficulty in providing perspective of splice site or frameshift variations of exon(s). Moreover, it does not capture the association between exon preference of 5'/3' splice site variations to either the N or C terminal of proteins due to the absence of mapping between amino acids and exons. It is also not trivial to gain insights into exon inclusion/exclusion effects on secondary structures, sequence, or structural domains of proteins from pairwise comparison of transcripts. In such investigations, the significance of 'dual' exons is often underestimated and shifts in patterns along with the preponderance of various AS events across evolution tend to be unappreciated. The ENACT nomenclature makes some of these issues tractable and amenable for appropriate computational analyses to understand the distribution of various exon variations related to AS events either within a gene in the organism or across species. Within this chapter, we have utilized annotations from ENACTdb to conduct comparative analyses of occurrences and extent of noncoding and coding exonic variations, followed by investigating the effect of these on transcripts/isoforms sequence variation across five organisms. Moreover, we also examined the prevalence of these among constitutive and alternate exons.

3.2 MATERIALS AND METHODS

3.2.1 Construction of 2X2 dataset

The detailed description of identifying RISO transcript, RSOEx, and various steps of exon annotation is provided in Section 2.2.3. The present work explores proteome expansion/diversity arising from AS events and variations in constitutive and alternate exons. Therefore, we have restricted our study to genes having at least two different isoforms and two coding exons. For this, we filter an organism's genes that satisfy the abovementioned criteria to create a subset of genes referred to as 2X2 dataset. As constitutive/alternate exons are based on their occurrences in transcripts/isoforms, subsetting gene sets in this manner ensures appropriate categorization of exons in the study.

3.2.2 Pfam domain annotations

The detailed methodology of Pfam domains annotation is described in section 2.2.1. Briefly, we have used Pfam 35.0 domain definitions (Mistry, Chuguransky, Williams, Qureshi, Salazar, Sonnhammer, Tosatto, Paladin, Raj, Richardson, et al. 2021) for identifying domains for each isoform using PfamScan (HMMERv3.2.1) (Madeira et al. 2022a). An e-value cutoff of 0.01 and model length cutoff of ≥ 0.7 of a given Pfam domain is used for assigning a domain to protein sequence.

3.3 RESULTS AND DISCUSSIONS

3.3.1 Summary of exon annotations in five genomes

As described in methods, the latest genome sequence build (2022) from the RefSeq database (NCBI) for *C. elegans* (worm), *D. melanogaster* (fruit fly), *D. rerio* (zebrafish), *M. musculus* (mouse), and *H. sapiens* (human) were annotated and deposited in ENACTdb. The summary of exon annotation in these organisms is shown in Table 3.1. We initiated our analysis by examining the distribution of exon count in genes and analyzed the relative contributions of coding/noncoding exons. The average number of total, coding, and UTR exons per gene demonstrates an increasing pattern from lower to higher organisms. Remarkably, the zebrafish, mouse, and human genomes exhibit a large fraction of genes (~37%) having ≥ 10 exons, in contrast to ~12% in other organisms (Figure 3.1A). Next, we considered the contribution of

coding exons to each gene, and it showed a similar trend (Figure 3.1B), wherein a large fraction of genes contain more coding exons as organismal complexity increases (Table 3.1).

Table 3.1: Summary of exons/gene occurrence in genomes

Organism	Number of protein-coding genes	Average number of Total (Coding) [UTR] exons/gene	Average number of total (Unique) isoforms/gene	Number of genes in 2X2 dataset
C. elegans	19,972	6.3 (6.2) [0.1]	1.4 (1.4)	4413 (22%)
D. melanogaster	13,972	4.7 (4.1) [0.6]	2.2 (1.6)	3520 (25%)
D. rerio	26,374	10.2 (9.6) [0.6]	1.8 (1.7)	8111 (31%)
M. musculus	22,134	10.5 (9.4) [1.1]	4.2 (3.0)	11494 (52%)
H. sapiens	20,443	11.8 (10.4) [1.4]	6.4 (4.4)	13063 (64%)

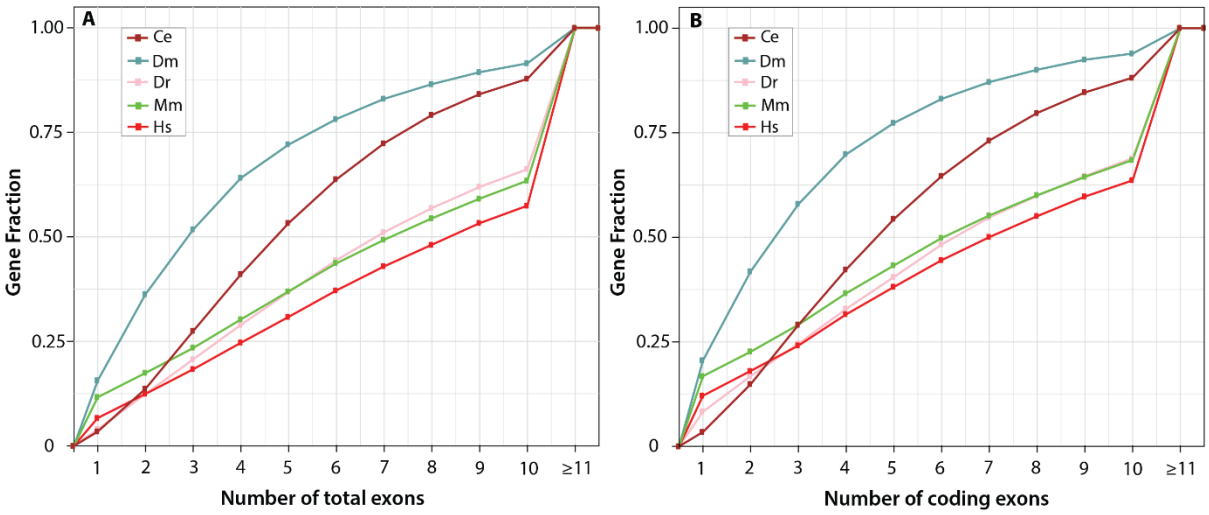


Figure 3.1: Cumulative distribution of total and coding exons in representative genomes. The gene fraction distribution consisting of N number of total exons is shown in (A), and the same for coding exons is shown in (B)

In the cases of fruit fly and worm, about 50% of genes possess ≥3 and ≥5 coding exons, respectively. Conversely, the same gene fraction of the zebrafish, mouse, and human genes consists of ≥6 exons. We compared the relative increase in average coding or noncoding exons per gene among genomes. Interestingly, the relative increase in human noncoding exons/gene is 1.3 times with respect to zebrafish genome, whereas the coding exons/gene increases only

by 0.08 times. This increase is considerable when we compare coding/noncoding exons of human with worm genome, indicating a relative expansion of noncoding exons in higher organisms.

Expanding on this observation, we compared total protein-coding variants (all listed transcripts) with unique isoform counts (distinct protein sequences) with an assumption that splicing involved UTR exons would result in no change in protein sequence but contribute to transcript diversity (Table 3.1). As can be seen, the difference between mean total and unique isoforms per gene increases from zero in the worm to 2.0 in humans, with an exception in zebrafish genome, which shows only a modest change (0.1). The same is evident from distribution of gene fractions having distinct protein-coding variants (Figure 3.2). Especially humans and mouse genomes show greater redundancy in isoform sequences than lower organisms. This observation is consistent with the previous findings of an elevated count of noncoding exons contributing to transcript diversity. In the case of the zebrafish genome, we speculated that modest change may be because it has many paralogous genes and probably generates isoform diversity through them. As seen in Figure 3.2B, ~1/3rd and ~1/2 of mouse and human genes encode ≥ 3 distinct isoforms, whereas in zebrafish, fruit fly, and worm, ≤ 1/10 of genes have ≥ 3 distinct isoforms suggesting that proteome diversity has significantly increased in organisms with greater complexity. Notably, the human genome consists of ~8% of genes having ≥11 protein-coding listed isoforms, and some examples in this category include MAP4 with 158 unique proteins, WNK1 and WNK2 genes (each having >50 isoforms). In the later sections, through ENACT elucidated annotation, we have divulged the occurrences and nature of UTR, coding, and 'dual' exons and their distributions in genomes.

3.3.2 Overview of 2X2 dataset and role of UTR exons in proteome expansion

The earlier analysis highlighted that numerous genes encode only a unique isoform rendering it inadequate for understanding genome-wide AS-derived proteome diversity. Additionally, we would also explore distributions of constitutive/alternate exons occurrences across genomes, and the reliability of such categorization improves by genes having more than one isoform. To facilitate this, we filtered each organism's genes to obtain those with at least two distinct isoforms and composed of at least two exons (see section 3.2.1), resulting in a 2X2 gene dataset.

After applying the above filtering criteria, over 50% of human and mouse genes are included in the dataset, whereas it ranges from ~1/3rd to ~1/4th genes from rest of other organisms (Table 3.1). The ensuing analysis is performed on the 2X2 dataset.

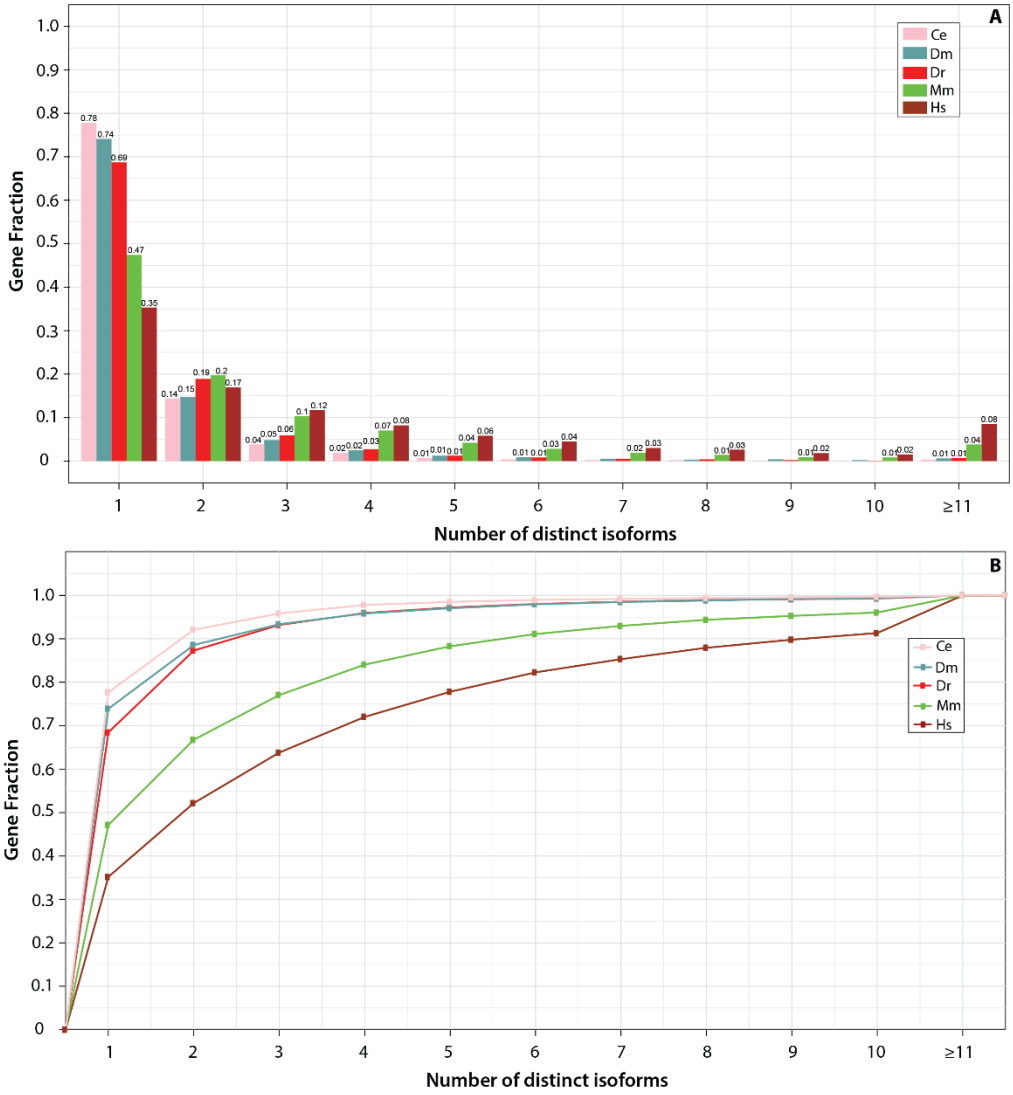


Figure 3.2: Distribution of distinct protein-coding isoforms. A) Bar chart showing gene fraction distribution for distinct (unique) isoforms and B) Cumulative distribution of the gene fraction with respect to unique protein isoforms.

To further investigate the extent splicing generated isoforms and their subsequent participation in proteome diversity, we initially examined the reduction in the total (listed) number of isoforms/gene when considering only distinct or unique isoforms. We found that the average count of listed transcripts (Total_ISF) and distinct (unique) isoforms (NR_ISF) per gene has an upward trend from worm to human genome (Figure 3.3a and 3.3b). The NR_ISF

per gene (Figure 3.3b) is greatly reduced in humans (~30%), followed by mouse (~28%), fruit fly (22%), and zebrafish (~8%), and no changes are observed in the worm genome. This suggests that numerous human, mouse, and fruit fly genes encode multiple isoforms (transcripts) having variation contributed only from noncoding exons and none from coding region. Moreover, indicating a possibility of multiple regulations in such genes at the level of mature mRNA. Subsequently, we probed the average number of distinct transcripts per isoform (NVar/NR_ISF) to explore the isolated extent of only UTR exon variations. So, a value of 1 would mean that there is only one listed isoform per distinct protein sequence. Value >1 indicates multiple transcripts listed per distinct protein sequence. As seen (Figure 3.3c), the average NVar/NR_ISF is higher than 1 in human, mouse, zebrafish, and fruit fly genomes, indicating they harbor multiple transcripts with identical protein sequences. Further, we delved into finding the maximal possible variations in the noncoding region of isoform observed per gene by considering the maximal count of transcripts listed per distinct protein in them, computed as $NVar_{max}$/NR_ISF. It is evident from Figure 3.3d that human and mouse genes have, on average, a variation of 2.6 and 2.2, respectively. This suggests extensive variation is embedded in the UTR region of genes in these organisms, and its consideration alongside a rising count of coding protein isoform indicates an additional role of AS in translational regulation. Detailed mechanistic insights into the role of noncoding exons and their impact on coding repertoire have been discussed elsewhere (Aspden, Wallace, and Whiffin 2023). The maximum UTR variation in humans is observed for gene SNRPN (small nuclear ribonucleoprotein polypeptide N), which has 26 noncoding and 8 coding exons. These combine through alternative splicing to generate 112 protein variants yet only result in 5 distinct protein isoforms. Thus, both coding and noncoding exons contribute to transcript or proteome diversity.

To comprehensively investigate various details of proteome expansion/diversity inferred by the 2X2 subset, we utilized the ENACT framework to map changes embedded at the level of exons or how these vary across genomes. In the subsequent discussions, we present exon/gene fractions variations across the genome using the three blocks of ENACT exon annotation.

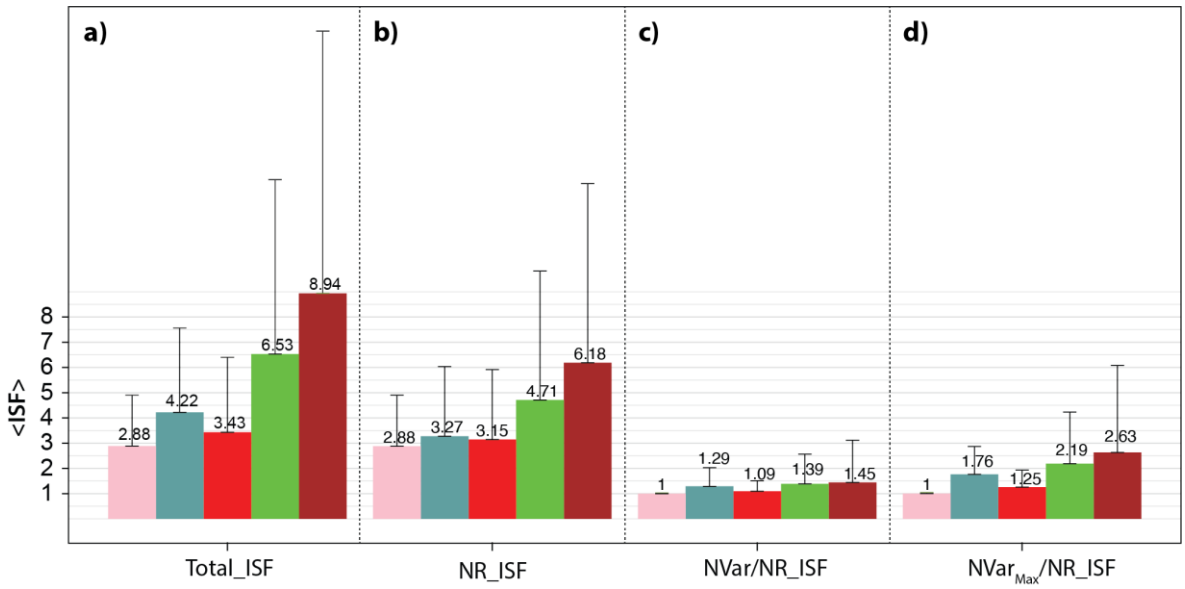


Figure 3.3: Mean number of isoforms in 2X2 dataset across five organisms. Panel (a) shows the bar chart of average listed isoforms per gene, and (b) shows the same for unique or distinct isoforms per gene. Panels (c) and (d) shows the mean number of listed transcripts per distinct isoform; and the maximum number of variant per distinct isoform per gene in the genome, respectively.

3.3.3 Analyses of coding/noncoding and dual exons across genomes

The coding status and amino acid sequence variations are encompassed in Block-I of exon nomenclature. We used Block-I annotations to investigate the preponderance of coding/noncoding/dual exons across representative organisms. Traditionally, exon is classified as coding or noncoding based on whether it contributes amino acid sequence to the isoform. However, this classification overlooks exons, which are coding in some transcripts or non-coding in others. ENACT categorizes these as 'dual' exons (see section 2.2.3, subsection C, point 'a'). As previously elaborated (see section 2.2.3, subsection C, point 'a'), an exon is tagged as 'T', representing the translated region, consisting of Coding Genomic Coordinate (CGC) and identical GCs in all listed transcripts. An exon is designated as 'U' for UTR exon having the same GCs but lacking CGC in all transcripts. The dual or 'D' exons have the same GCs but possess CGC in at least one transcript and are absent in others. Notably, ENACT nomenclature also introduces dual exons. The second character in exon nomenclature captures variation in the amino acid sequence of coding exons with respect to the same sequence in the RSOEx set. The amino acid sequence of an exon could differ from reference sequence because of following reasons:

a. Alternate promoters start or termination of translation.

b. Frameshift in the amino acid sequence due to inclusion/exclusion of alternate exons.

c. While considering triplets to associate amino acid sequences to exon, we notionally borrow or donate nucleotide to make triplets. In this way, the same exon can be associated with two amino sequences differing in length by one amino acid. This artifact of our association protocol does not mean the same exon encoding multiple amino acid sequences.

The first two points, (a) and (b), are valid cases of amino acid sequence change; however, c) is associated with 'aa' assignment to non-3n divisible exons and was ignored in our analyses. In brief, Block-I describes the coding nature and amino acid sequence variation of coding exons. The amino acid change categories are commonly referred to as Frameshift events (FSE).

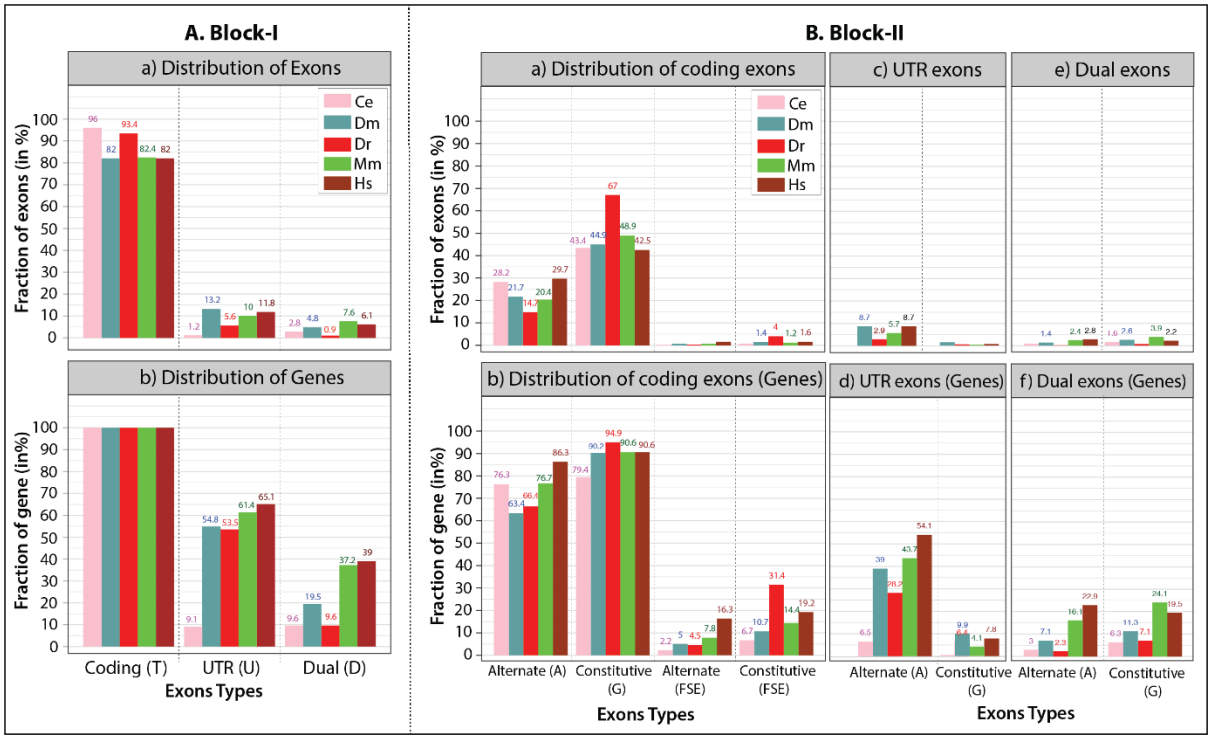


Figure 3.4: Comparative distribution of major exon types defined in ENACT and their occurrence in genes. A) The bar chart describing the distribution of coding/noncoding(UTR)/dual exons as per Block-I of ENACT and their abundance in genes are shown in sub-panels a and b, respectively. B) This panel shows the distribution of various exon types (U/T/D) defined as alternate/constitutive (Block-II). The sub-panels a) and b) shows bar chart showing the exon relative frequency of coding exons classified as alternate, constitutive, alternate (FSE), and constitutive (FSE) and their relative occurrences in genes across genomes, respectively. The sub-panels c) and d) show the relative occurrence of noncoding exons and their occurrences in genes, respectively. The same for dual exons are shown in panels e) and f) for exons and occurrences in genes, respectively.

In the present analyses, we evaluated major exon types (T/U/D) occurrences and their respective contributions to genes in an organism. The coding (T) exons are the most prevalent (>80%) across genomes and comprise the highest fraction of exons in *C. elegans* (Figure 3.4A). Next to most prevalent were noncoding exons (>10%) in the fruit fly, mouse, and human genes, with low occurrences in worm and zebrafish genomes. The 'dual' exons are relatively abundant in mice and humans (>6%), suggesting them as recently gained features of gene architecture in higher organisms. Despite their relatively lower occurrence, they assimilate in >35% of human and mouse genes. The noncoding (U) exons are even more abundant, with >50% gene assimilation across all organisms except *C. elegans*. The genome comparison of T/U/D exons showed a general increase in the number of genes having U/D exons from lower to higher organisms, with the zebrafish genome being an exception. The amino acid sequence changes will be discussed in the next section. We will discuss the zebrafish genome being an exception in a later section.

3.3.4 Prevalence of alternate/constitute exons among U/T/D categories across genomes

As previously explained (section 2.2.3, subsection C, point 'b'), Block-II of ENACT furnishes the relative position of exon and its inclusion frequency in a gene that is employed to broadly classify exon as constitutive (G), alternate (A) and majorly constitutive (F). Constitutive (G) exons exist in all transcripts with identical GCs, whereas F exons exist in all transcripts but with alternate splice sites. On the contrary, alternate exons (A) are present in some transcripts. We combined the information from Block-I and Block-II to assess distribution of G/A exons among coding/noncoding and dual exons and their occurrences in genomes. The distribution of the F category is described in the next section.

First, we divided coding exons (T) into subgroups as TA (coding alternate), TA (aaChange/FSE), with amino acid sequence change, TG (coding constitutive), and TG (aaChange/FSE) to assess the distribution of exon fraction in these subgroups. As seen in Figure 3.4B (sub-panel (a)), the constitutive exons are more prevalent (>40%) than alternate exons across genomes, with the maximum fraction observed in the zebrafish genome. Among the distribution of alternate exons, we observed a general upward trend of these from zebrafish (~15%) to human (~30%) genome; concomitantly, the fraction of constitutive exons decreased from zebrafish to humans. Similarly, the gene fraction constituting alternate exons showed a similar upward trend (Figure 3.4B, panel (b)), whereas the gene fraction consisting of constitutive exons remains similar (>90%) across higher organisms suggesting most genes

harbor such exons. The zebrafish genome has a high occurrence of T exon, which sub-fractions into mostly TG group. This phenomenon might be related to widespread paralogue gene fractions in zebrafish exhibiting lower rates of alternative exon-mediated splicing that requires a separate study on AS in zebrafish (Lambert, Olsen, and Cooper 2014). Next, we examined the distribution of TG(FSE) and TA(FSE) exons across organisms showing their occurrences are <2% and <1%, respectively. However, the genes harboring TA(FSE) exons are the maximum in humans (16%), whereas TG(FSE) is abundant in the zebrafish genome (31%), as shown in Figure 3.4B (sub-panels a and b). These suggest that exons showing amino acid sequence category need detailed analyses to understand their impact on the structure/function of affected isoforms.

Subsequently, we evaluated UG (noncoding constitutive), UA (noncoding alternate), DG (dual constitutive), and DA (dual alternate) exon distributions and their abundance in genes. More than 90% of UTR exons are classified under UA, and their fraction increases from 3% to 9% from zebrafish to human genomes (Figure 3.4B, sub-panel c). Consequently, the fraction of genes harboring such exons also increased from zebrafish (~28%) to humans (~54%), as seen in Figure 3.4B (panel d). Interestingly, UG exons are found <1.5% across genomes, indicating the noncoding exons are more frequently spliced out during AS or undergo 5'/3' splice site changes. Those UG exons that remained constitutive and the role they may incorporate for their assimilating genes need separately designed study. Compared to U/T exons, the dual exons have similar fractions in constitutive and alternate categories, and their occurrences in genes are also comparable (Figure 3.4B, sub-panels e and f).

We analyzed the pattern of alternate/constitutive exons in the coding/noncoding/dual category by computing ratio of A/G in each of these categories. We observed that the A/G ratio of noncoding exons in the human genome is rather high, 14.5, and for coding exons, it is 0.69, suggesting a greater preference for alternate noncoding and constitutive coding exons in the human genome. This observation highlights the simultaneous introduction of alternative transcription and termination regions or their complicated interplay with the splicing process. Additionally, higher variation in the UTR region is a probable consequence of relaxed selection pressure than imposed by structural feature constraints in the CDS region.

3.3.5 Alternate splice site occurrences with their n/c/b subtypes

The Block-III of ENACT embeds the splice site variations and their occurrences of these events at 5th and 6th positions of EUID, respectively. The 5th position provides the type of splice site variations and is based on change(s) in the GCs of RSOEx. Based on it, the exons showing these variations are assigned 'n', 'c', and 'b' tags for its 5', 3', and both 5' and 3' splice site changes, respectively. These variations are identified in alternate exons, denoted as A(ss) category, and in constitutive exons, which are already defined as F category. Our nomenclature keeps track of exon through Block-II (relative position, 4th in EUID); however, it can have any combination of features of Block-I and Block-III. For instance, an exon with EUID: U.-2.A.4.0.0 is noncoding, and with possible splice site change (EUID: T.1.A.4.c.1) can become coding in another transcript. Thus, features of block-III of ENACT can be analyzed in combination with block-II features.

We analyzed the splice site changes in coding/noncoding and alternate/constitutive categories to examine the contribution of their occurrences across organisms. The A(ss) category of exons (Figure 3.5A, sub-panel a) comprises <5% of all exons, and their gene prevalence in genes increased from 25% (worm) to 40% (human), as shown in sub-panel b of Figure 3.5A. However, constitutive splice site variants (F category) show a decrease in exon fraction from worm (20%) to human (5%), and the same is also evident in their contribution to gene fraction (Figure 3.5A sub-panels a and b). Even though coding (G) exons fraction has decreased from zebrafish to human genome, these are not as drastic as in the F category. The A(ss) and F categories of UTR exons constitute <3% and <1.5%, respectively, across genomes. In comparison to F category, the alternate UTR exons with splice site changes show a greater contribution to increase transcript variation in most organisms except *C. elegans*. We found that constitutive coding splice site change (F) category shows decreasing usage in higher organisms compared to TG category. While alternate (only) exons are widely used in higher organisms and, to a similar extent, A(ss) category is also represented in gene fraction. This suggests that during evolution, there is probably a shift in the generation of splice site variants of an exon to using alternate exons. This also supports previously observed "Exon skipping" as an abundant splicing event in higher organisms (Kim, Yang, et al. 2020).

We analyzed details of splice site variations of both coding A(ss) and F categories to study the n/c/b occurrences and genes affected across genomes. Table 3.2 summarizes the prevalence of these variations in five genomes. As can be seen from the table, more than 70% of genes in

all organisms consist of at least one exon, which undergoes splice site changes (A(ss) or F). Interestingly, worms show the maximum fraction of genes affected by these variations. Among various splice site variations involved in alternative splicing, the 'n' and 'c' cases are most prevalent (~6% - 13%) across organisms. However, 'b' cases occur the least, indicating a possibility that exons tend to maintain at least either 5' or 3' splice sites. As organismal complexity increases, there is a considerable reduction in both 'n' and 'c' variations and their affected genes, where they are reduced from 78% (worm) to 51% (human) for 'n' cases and 82% (worm) to 44% (human) for 'c' cases. These suggest a possible mechanism of using exon skipping during AS rather than using splice site variations.

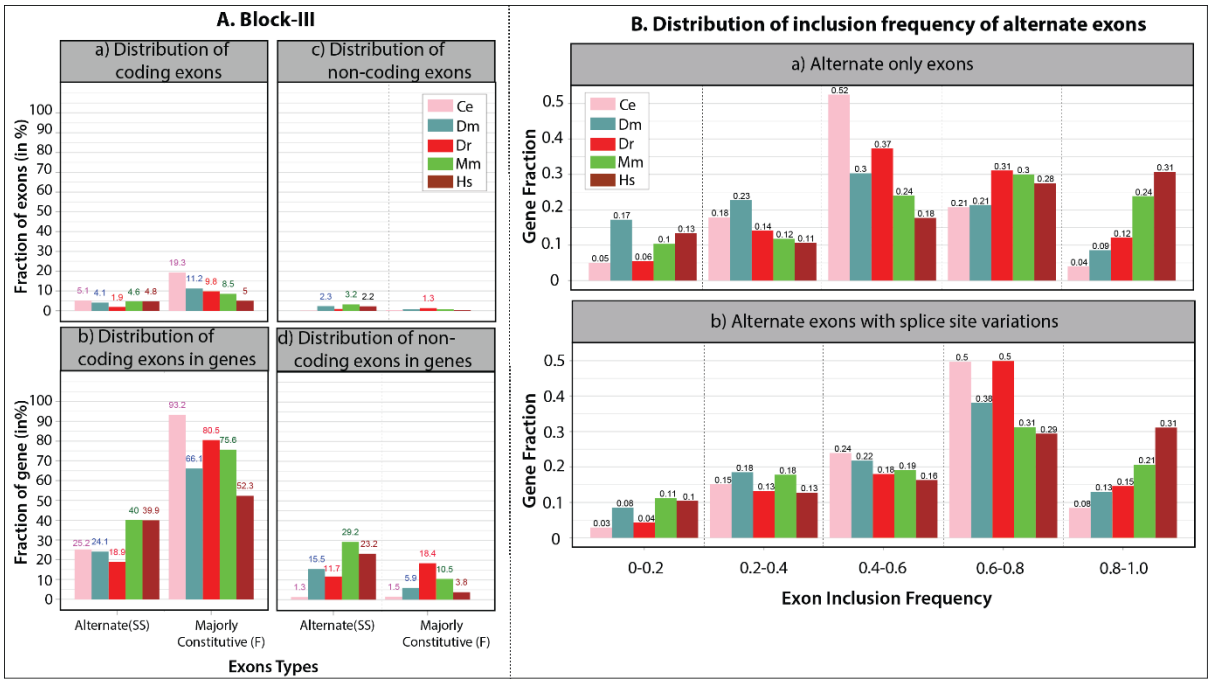


Figure 3.5: Comparative distribution of splice site variations in coding/noncoding exons and their inclusion frequency across five genomes. A) The bar chart describing the distribution of splice site variations in alternate and majorly coding (F) categories of exons as per Block-III is shown in sub-panel a, and their relative abundance in genes is shown in sub-panel (b). B) The inclusion frequency of alternate only exons and alternate with splice site variations (A(ss)) are shown as bar chart with a bin width of 0.2 inclusion frequency in five genomes.

The above analysis indicates that there is probably a shift from lower to higher organisms to involve more alternate exons for splicing rather than altering splice sites of constitutive exons. We hypothesize that this may be due to an evolutionary tendency to strictly regulate the splicing events considering their far greater role in providing phenotypic and proteome complexity in higher organisms while also encountering and strictly regulating their complex interplay with

several other processes, including but not limited to alternative transcription initiation and termination and co transcription splicing. Additionally, observation of length regulation in A, A(ss), and F cases supports dynamic changes in preference for splicing events.

3.3.6 Assessment of alternative exon's effect on protein isoforms

We found it quite puzzling that the fraction of alternate exons is similar between humans and worms despite having a large difference in the number of genes 2X2 dataset of these organisms. Even though alternate exons fraction can be similar, their impact on genes depends on many factors, such as their inclusion frequency, local variations in splice sites, differences in the length of exons, and change in sequence with frameshift events. The latter event can easily be analyzed as it is encompassed in Block-I, and we examined aaChange considering both A and A(ss) categories. It is observed that there is a greater contribution of alternate exons with amino acid sequence change in humans (1.6%) in comparison to all organisms (<1%). Irrespective of their small proportion, a corresponding rise in the gene fraction having such exons shows an increase from 2% (worm) to 16% (human), as shown in Figure 3.4B (panels a and b). This indicates disparate roles of alternate exon subpopulations that cannot be inferred solely from their overall gene fraction comparison; however, it could be facilitated by ENACT exon nomenclature. Next, we compared the similar contribution of G(aaChange) in genes across genomes. Despite similar exon G(aaChange) fraction, there is a modest rise in their contribution from 6% to 19% in the worm to human genome. It is pertinent to note that even though frameshift events are observed in both A and G exons, the reading frame is often restored in the successive exons so that such events marginally impact the protein sequence.

Subsequently, we also analyzed the inclusion frequency of alternate exons and A(ss) exons (Figure 3.5B). Considering the exon inclusion frequency of >60%, a greater fraction of human genes (~60%) harbor such exons than worms, which have only 25% of genes. Thus, it suggests that genes in humans with alternate exons tend to have higher inclusion rates than those from worms despite having similar alternate exon fractions (Figure 3.5A, sub-panel a). A similar comparison on A(ss) exons showed that with inclusion frequency of >60%, both worm and human gene fractions (~58% in worm and ~61%) are comparable despite that these are contributed by 5% of exons (Figure 3.5A, sub-panel b).

Table 3.2: Summary of splice site variants of coding A(ss) and F categories divided into ('n'/'c'/'b') types of variations

Organism	Total coding exons	'n' cases Total (Exon in %) [Gene frac in %]	'c' cases Total (Exon in %) [Gene frac in %]	'b' cases Total (Exon in %) [Gene frac in %]	Total genes in 2X2, (Genes in n/c/b) [Gene Frac %]
C. elegans	37834	5094 (13) [78]	4237 (11) [82]	548 (1.4) [12]	4413 (4253) [96]
D. melanogaster	23807	2300 (9.7) [50]	2250 (9.4) [50]	205 (0.8) [5]	3520 (2710) [77]
D. rerio	109646	7375 (6.7) [65]	5942 (5.4) [59]	857 (0.7) [10]	8111 (6931) [85]
M. musculus	134556	10976 (8.1) [64]	10325 (7.6) [67]	1787 (1.3) [14.6]	11494 (10076) [87]
H. sapiens	156417	10090 (6.4) [51]	8430 (5.4) [44.7]	1823 (1.2) [12.7]	13063 (9438) [72]

3.3.7 Comparative analysis of length from constitutive/alternate exons

The length distribution of coding exons and their various types (G, F, A, and A(ss)) from 5 organisms is shown in Figure 3.6. In general, the amino acid length of all exons in vertebrates is similar; however, invertebrates show variation, with worms having a small length compared to flies. The median length of the *D. melanogaster* exon is unusually longer than other genomes, and it is observed in all categories of coding exons. A recent study has shown that there are subpopulations of exons having unusual lengths in fruit fly genome separated by short introns as a consequence of co-transcriptional splicing while emphasizing intron definition mechanism of theirs like in *S. cerevisiae* (Prudêncio et al. 2022). Among different categories of coding exons, we observed significantly longer exons in G cases followed by F, where they may follow previously indicated literature report and hint towards possible elongation of first and last exon as driven by ATIT processes. In vertebrates, both constitutive and alternate exons

are smaller in length. Regarding length comparisons of alternate and constitutive exons, literature indicates conserved fraction of alternate exons tends to be smaller than their conserved constitutive counterparts (Keren, Lev-Maor, and Ast 2010). Our analysis confirms a smaller median length of alternate exons than constitutive among all organisms. Extending discussion on previously hypothesized regulated length fraction for A(ss) and F categories with organismal complexity, Figure 3.6 on careful visual inspection details that, not only median, but overall IQR showed constriction in A(ss) category when we move from zebrafish to human, and similar can be observed in F cases where mouse showed modestly broader IQR than zebrafish, but human has shortest IQR range among all, indeed confirming that for subpopulation of exons that persist in vertebrates, their lengths are under selection pressure and similar is also reflective from length distribution of A exons.

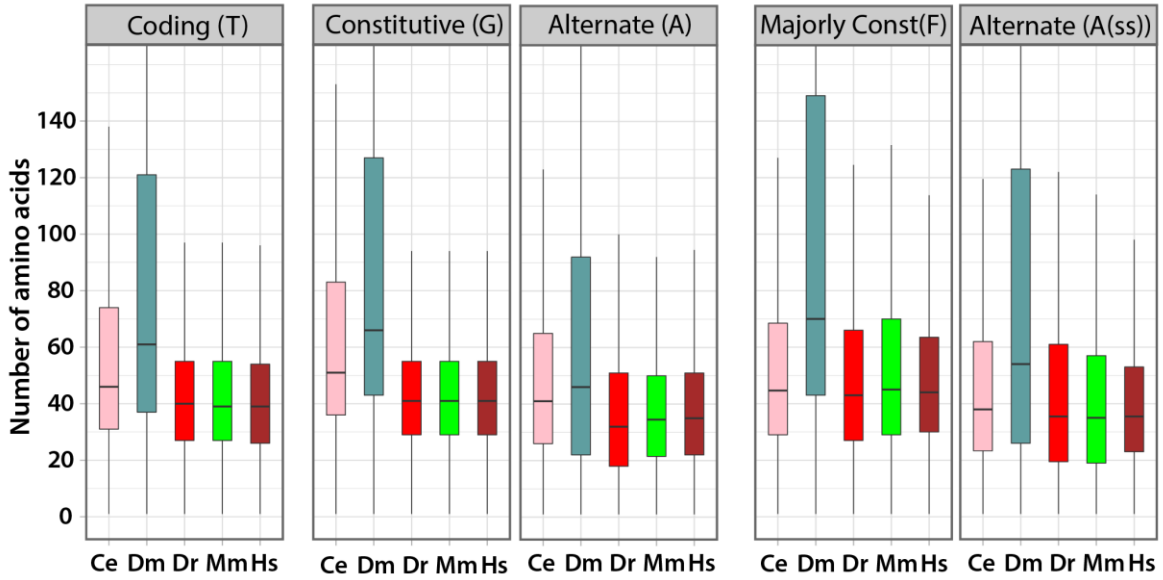


Figure 3.6: Distribution of amino acid length for coding exons classified in various categories as Block-II and Block-III. The box plot showing the distribution of amino acid length of coding exons and their sub-types as: alternate exons only (A), constitutive (G), majorly constitutive (F), and Alternate exons with splice site (A(ss)) are shown for exons from various organisms. The median is shown in the box, and whiskers are shown as lines above and below the boxes. (The upper whisker is truncated in some plots to make a compact representation).

3.3.8 Examples of genes annotated with ENACT

Having demonstrated the usefulness of ENACT framework in comparing and elucidating splicing extent among chosen distant model organisms, we have next chosen 4 sets of examples

from the human genome where we establish the utility of ENACT framework in comprehensively understanding and enhanced interpretation of intra-transcript variations.

3.3.8.1 Fragile X messenger ribonucleoprotein 1 (FMR1)

The FMR1 gene encodes fragile X mental retardation protein (FMRP), whose loss of function causes inheritable disorder fragile X syndrome and premature ovarian failure (Crawford, Acuña, and Sherman 2001). The protein regulates the translation of a subset of mRNAs and the shuttling of mRNA in the intracellular compartment and correspondingly localizes in polyribosomes or ribonucleoprotein complexes. Although FMR1 is expressed in almost all tissues, it is abundant in testes and brain cells. In neuronal cells, it regulates synaptic plasticity (Santoro, Bray, and Warren 2012). The gene has 17 coding exons, which undergo alternative splicing to result in multiple transcripts that could vary up to 49 isoforms (Zafarullah et al. 2020). Some human and mouse isoforms are well characterized by their sequence features and role in function or cellular localization (Fu et al. 2015; Sittler et al. 1996).

We analyzed representative ENACT annotated isoforms of FMR1, focusing on the AS events in the C-terminal to explore possible effects of alternative splicing on the protein function. The FMRP protein is an RNA binding protein consisting of RNA binding domain *viz.* hnRNP K Homology (KH) domains and RGG box motif. Two KH domains are in exons 3-8, and the RGG motif in exon 15. The reference isoform IS-1 (NP_002015.1) has all 17 coding exons. Two Pfam domains FXMRP1_C_Core (exons 13-15) and FXMR_C2 (exons 16-17), are associated with C-terminal region of the protein (Figure 3.7A). The exon 12 (T.1.A.12.0.0) is skipped without affecting the domain architecture in IS-2 (NP_001172005.1). Exon 15 shows the same 5’ss in IS-3, IS-4, and IS-5, however, with varying consequences of alternative splicing on the sequence of exon 15, evident from Exon IDs (EUIDs) in these isoforms. In IS-3 (NP_001172011.1), the 5’ss AS event of exon 15 is assigned with EUID: T.1.F.15.n.1, showing that its N-terminal sequence is altered, which results in truncation of FXMRP1_C_Core domain. However, in IS-4 (NP_001172004.1) and IS-5 (NP_001172010.1), the exon 14 skipping event leads to the reading frame shift resulting in the amino acid sequence change from exons 15-17. The amino acid change is evident from their nomenclature as exons 15, 16, and 17 are assigned T.2.F.15.n.1, T.2.G.16.0.0, and T.2.G.17.0.0 EUIDs, respectively. The second character (T.2.) shows that the amino acid sequence is different from the reference exon without any change in their genomic coordinates. Due to the frameshift of the reading frame and loss of exon 14 in IS-4 and IS5, the two Pfam domains are lost and may lead to

altered protein function. Moreover, exon 14 harbors a nuclear export signal, and in the absence of this, exon IS-4 and IS-5 will be unable to perform nucleocytoplasmic shuttling functions. Additionally, it should be emphasized that skipping exon 14 in this context also demonstrates the previously discussed subpopulation emergence of frame-altering alternative exons in higher eukaryotes.

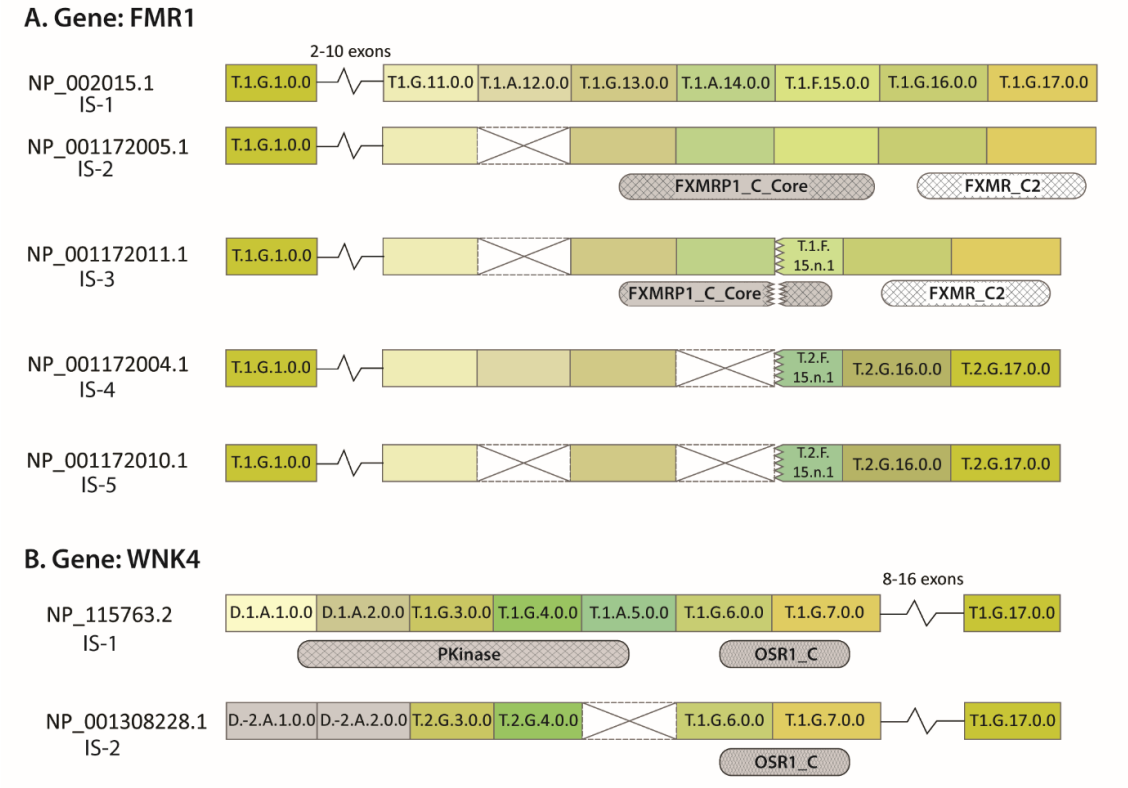


Figure 3.7: FMR1 and WNK4 exon annotations. Schematics show the exon organization in isoforms of human FMR1 and WNK4 genes in panels (A) and (B), respectively. The NCBI protein identifier is shown for each isoform along with its isoform number for reference purposes only. The color and grey color-filled rectangle boxes show coding/partially coding and noncoding exons, respectively. The absence or skipped exon is shown with crossed empty rectangle box. The jagged edges of a rectangle represent 5SS or 3SS alternate splice sites. The EUID of an exon is displayed on its rectangular box; however, EUID is not shown on those that do not change its attribute from its reference transcript. The break in transcript shows that exons lying in the intervening region do not undergo variation in the isoform. The isoforms sharing Pfam domains for a region are shown under it. In FMR1 gene, FXMRP1_C_Core and FXMR_C2 domains are assigned in the same region of IS-1 and IS-2. However, IS-4 and IS-5 lack these two domains. Similarly, IS-2 of WNK4 lacks Pkinase domain in the N-terminal region.

3.3.8.2 WNK4 gene

The WNK4 gene belongs to the conserved "With no lysine (WNK)" group of serine/threonine kinases (STK) in eukaryotic organisms. These have been named because of their atypical positioning of catalytic lysine in subdomain II instead of I, like in other STKs. The WNK4 is primarily expressed in the kidney, having a role with other family members in modulating the balance between sodium chloride reabsorption and renal potassium ion secretion (Murillo-de-Ozores et al. 2021) by regulating activities of cation coupled cotransporters (SLC12, NCC), ion channels (ENaC), and ion exchangers (Moriguchi et al. 2005; San-Cristobal et al. 2008). The WNK4 is linked with a rare genetic type of hypertension called pseudo-hypoaldosteronism type 2 (PHA2). The gene consists of 19 exons. The WNK4 protein has two Pfam domains (Protein kinase and Oxidative-stress-responsive Kinase1 C-terminal domain), with the rest of the sequence being intrinsically disordered region. The OSR1_C encompasses Pask-Fray 2 (PF2) domain, which is known to interact with RFX[VI] motif and is known to suppress the activity of kinase domain (Murillo-de-Ozores et al. 2021).

A total of 13 isoforms are listed in the NCBI RefSeq database. Of these, two isoforms are reviewed (Figure 3.7B). Among the 19 exons, the first two are 'Dual' as these are coding in some transcripts and non-coding in others. These two exons are coding in the reference isoform (NP_115763.2), resulting in Pfam Protein kinase domain being assigned to the N-terminal region of the protein. On the contrary, the first two exons are non-coding in IS-2 (NP_001308228.1) with alternate translation initiation in exon 3 (EUID: T.2.G.3.0.0), leading to amino acid sequence change of exons 3-4 due to frameshift (EUIDs are T.2.G.3.0.0 and T.2.G.4.0.0). Interestingly, exon 5 skipping restores the reading frame and the rest of protein sequence is maintained as in the reference isoform, which is also evident from the EUIDs of exons. The IS-2 shows characteristic kinase domain loss but supports OSR1_C (PF2) domain. The latter domain is known to interact with SPAK/OSR1 protein, suggesting that IS-2 may act as a sequestering factor for them and affect their biological function. Skipping of exon 5 in this context helped restore the frame. As previously highlighted, non-3n divisible exons may not always be disastrous to resulting proteins. They may also be an evolutionary compensated compensatory mechanism evolved in higher eukaryotes to rectify the frame changes when alternate translation initiation sites are picked and indicate recognition of such processes in nuclei.

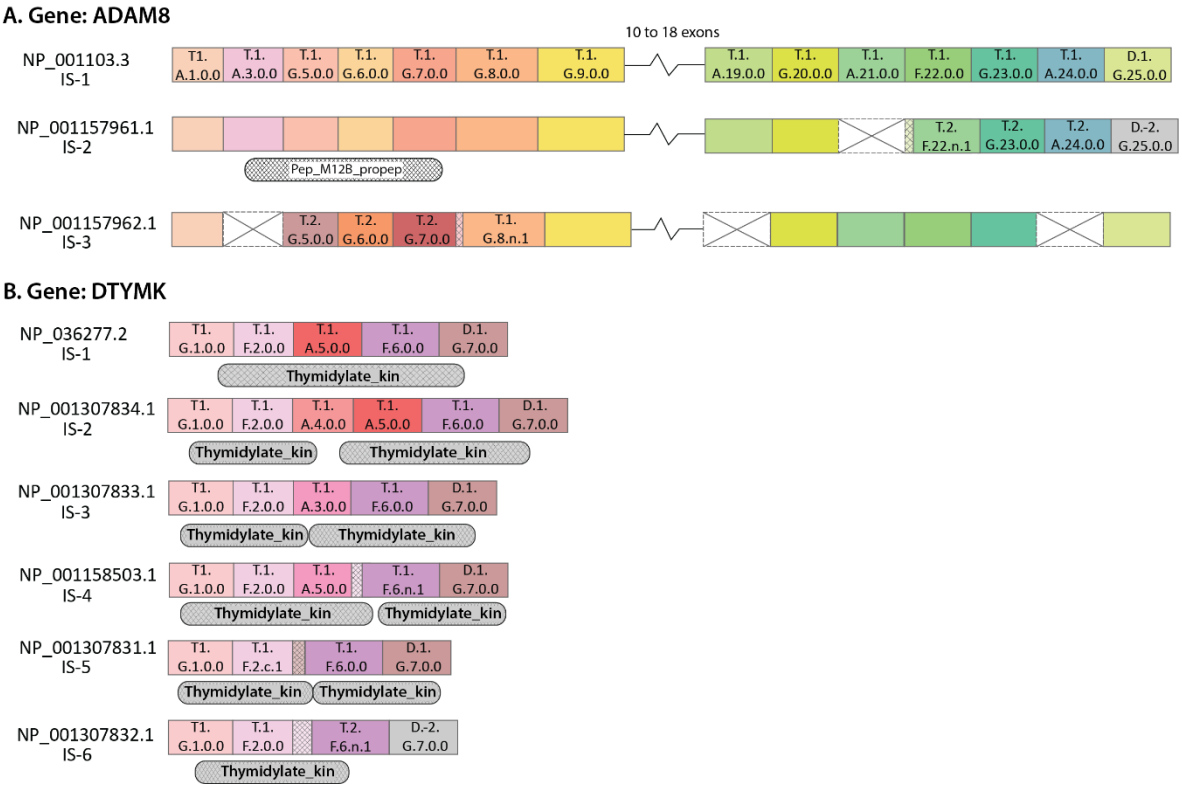


Figure 3.8: ENACT exon annotation of ADAM8 and DTYMK isoforms. Schematics show the exon organization for isoforms of ADAM8 and DTMYK genes in panels (a) and (b). The NCBI identifier is shown for each isoform, and exons are represented as colored rectangle boxes with their EUID. The absence or skipped exon is shown with crossed empty rectangle box. The small extension of exon rectangle boxes with crisscross filled lines represents extended exon boundaries due to alternate splice sites (5SS/3SS). If an exon does not show variation with exon(s) in previously shown isoform, then the EUID is not labeled. The break shows that exons intervening in the region do not change in the isoforms. The isoforms sharing Pfam domains for a region are shown under it. In ADAM8 isoforms, IS-3 lacks the Pep_M12B_propep Pfam domain. All isoforms of DTYMK consist of shortened or extended Thymidylate kinase domain.

3.3.8.3 ADAM8

ADAM8 gene encodes a protein belonging to the family of membrane anchored disintegrins and metalloproteases proteinases that cleaves extracellular domain of several cell surface proteins and receptors (Fourie et al. 2003). The ADAM8 protein is implicated in various cellular functions such as inflammation, immunomodulation, neutrophil activation/mobility, immune cell migration, osteoclast stimulating factor, and neurodegeneration (Yamamoto et al. 1999; Schlomann et al. 2000; Romagnoli et al. 2014). ADAM8 domain architecture consists of an N-terminal prodomain, a catalytic metalloproteinase domain, a disintegrin domain

involved in interaction with integrins, a cysteine-rich domain followed by a transmembrane region, and a C-terminal domain probably involved in protein-protein interaction through SH3 or proline-rich regions (Knolle and Owen 2009).

The ADAM8 gene comprises 23 coding exons, one noncoding, and one dual exon. Of coding exons, 17 are constitutive/constitutive-like, and the rest are alternate. We analyzed three well-annotated isoforms showing a combination of AS events (Figure 3.8A). The reference isoform (IS-1; NP_001100.3) has Pep_M12B_propep, Reprolysin (metalloproteinase), Disintegrin, and ADAM_CR (cysteine-rich domain) Pfam domains prior to transmembrane region (can be seen for isoform in ENACTdb). In IS-2 (NP_001157961.1), skipping of exon 21 is combined with 5' splice site change in exon 22 that leads to reading frame change in it and the rest of subsequent exons, as is also evident from their EUIDs: T.2.F.22.n.1, T.2.G.23.0.0, and T.2.A.24.0.0 along with premature termination in exon 24. The resulting isoform lacks the proline-rich region required for protein-protein interaction. The IS-2 isoform is expressed in metastatic lung cancer cell lines (Knolle and Owen 2009). The IS-3 (NP_001557962.1) shows skipping of exons 2 to 4, which affects local reading frame involving exons 5 to 7. However, the reading frame is restored from exon 8 onwards by 5'ss. These features are clearly interpreted from the exon nomenclature. Due to a change in amino acid sequence in IS-3, the pro-domain cannot be identified in this isoform's N-terminal region, suggesting that it may have constitutive metalloproteinase activity. In addition, 2 out of 4 glycosylation sites have also been lost as part of the pro-domain (Srinivasan et al. 2014). However, one of the conserved Glutamate (158E) essential for pro-domain's catalytic removal (Hall et al. 2009) is preserved in IS-3. Further experimental studies will provide insights into the enzymatic activity and biological role of IS-3.

Skipping exon 21 is a non-3n divisible instance as nucleotide length was 115 'nt', and selection of alternate splice site in succeeding exon leads to truncation of 52 'nt' in isoform NP_001157961.1. The above example additionally also demonstrates that more than 1 splice event occurred in the transcript, whether such splice site resulting in truncation of 52 'nt' was randomly chosen or was there an evolutionary selection pressure embedded which saved the transcript from truncating abruptly way before than it did way later in exon 24, in the absence of this 'n' side alteration otherwise. A detailed discussion of this is out of scope, and likley needs application of comparative genomics founded positive and negative selection pressure.

3.3.8.4 DTYMK

The deoxythymidylate kinase (DTYMK) gene encodes an enzyme essential for DNA synthesis, nuclear genome stability, and mitochondrial copy number maintenance (Hu Frisk et al. 2022). The enzyme catalyzes the transfer of γ-phosphate of ATP to dTMP in the presence of Mg2+ ion. The gene expression peaks during the S-phase and is low from mitosis to the early G-1 phase (Hu Frisk et al. 2022). The gene is upregulated in cancer cells (Liu et al. 2013) and tissues linked to developing severe microcephaly-like neurodegenerative diseases in humans like severe microcephaly (Löffler, Carrey, and Zameitat 2018). Previous studies on the sequence analysis of DTYMK identified three sequence motifs: a) lid region (residue 142 - 154) required for conformation change and P-loop motif; b) the DRX motif (X = Y/F, D) involved in catalysis and c) non-covalent interaction network formed by R76, D96 and π-π stacking between by F72, F105, and Y151 residues (Hu Frisk et al. 2022). These residue numbers correspond to the protein sequence of NCBI protein id NP_036277.2.

DTYMK gene consists of six coding and one dual exon. Exons 3, 4, and 5 of the six coding are alternate exons. We have associated the Thymidylate kinase Pfam domain in all six isoforms (Figure 3.8B). However, it was either an extended domain in some isoforms or atrophied in others. For instance, exon 4 (EUID: T.1.A.4.0.0) occurs only in IS-2 (NP_001307834.1) and introduces an insertion in the domain region. The shortest isoform (IS-6) is 113 amino acids long, primarily due to exons 3 to 5 skipping and the changed amino acid sequence of exon 6 with 5SS (EUID: T.2.F.6.n.1) and premature termination. Interestingly, the exon alignment view shows that exons 3 and 4 are mutually exclusive AS events as both do not co-occur in any isoform. Since the alternate exons are present in the middle of the DTYMK gene, we examined the conservation of structural and sequence features among isoforms. In recent work, Frisk et al. investigated changes in the domains of thymidylate kinase and found that isoforms lacking exons 3, 4, 5, or 6 with sequence variation lack crucial sequence motifs required for DTMYK activity (Hu Frisk et al. 2022). The reference isoform IS-2 (NP_001307834.1) has the insertion of 39 residues and weak enzymatic activity. It is interesting to note that IS-2 is expressed along with IS-1 (NP_036277.2) in fibroblast cell lines suggesting that IS-2 may impact the function of IS-1 by sequestering substrate for the enzymatic activity as it has conserved binding site residues. The other isoforms and their complex interactions, structural fate, and functional divergence are yet to be elucidated experimentally.

3.4 CONCLUSIONS

In the present chapter, we utilized the ENACT framework to gain insights into the evolution of various exon variations and performed a comparative analysis across five representative genomes. We restricted our dataset to genes for these analyses, which encode 2 or more distinct isoforms. In general, we observed an increase in exons during evolution with a relatively more increase in noncoding exons suggesting that variation in UTR regions may have a role in translational regulation in addition to translational diversity as AS is supposed to introduce. Interestingly, these are more prevalent in higher organisms.

To investigate exon types and their sub-types abundances across genomes, we restricted our analyses to genes having at least two coding exons and two distinct isoforms (2X2 dataset). We utilized ENACT nomenclature to investigate occurrences of coding/noncoding/dual exons across these evolutionary distant organisms. We observed that coding exons are the most prevalent, followed by noncoding and dual exons least common. However, dual exons are relatively commonly found in mouse and human genomes (>6%), and despite their low occurrences, these contribute to variation in 35% of their genes. As has been observed before, the prevalence of noncoding exons is increased in higher organisms indicating their role in increasing proteome expansion. We further divided each of the U/T/D exons as constitutive (present in all transcripts) and alternate exons that showed the former is most prevalent and alternate exons have an increasing trend from zebrafish to humans. Notably, the noncoding alternate exons are more abundant than their constitutive counterpart, and this trend is the opposite compared to constitutive coding exons going from fish to the human genome. This indicates that alternate UTR exons are evolutionary gained in higher organisms to regulate translation. The dual exons are found to have similar occurrences in higher organisms. On comparison of exons showing the change in amino acid sequence due to reading frame shift showed that constitutive (FSE) exons are more common than alternate (FSE), suggesting these bring diversity in the proteome of higher organisms.

We examined the contribution of splice site change variations from alternative/constitutive exons. The alternate with splice site changes has increased in higher organisms; however, the majorly constitutive exons have decreased, indicating the possibility that alternative(ss) exons are commonly utilized for proteome diversity. Moreover, it could also suggest that during evolution, there is a shift in relying on alternate exons rather than generating splice variants of constitutive exons supporting the “Exon skipping” observed as a commonly occurring AS

event in the higher organism. The detailed n/c/b events across organisms showed that these events have decreased from worm to human, indicating splice site variants are less preferred in higher organisms. The comparison of exon amino acid sequence length showed that higher organism exons are of shorter length and have a lower IQR. In general, constitutive exons are slightly higher than that of alternate exons. Importantly, most of the above inferences are correct even when on the 4X4 dataset (genes with 4 distinct protein isoforms and 4 different coding exons). Especially the rise in the fraction of Dual exons and alternate UTR exons was especially prominent.

The above inferences detailed the increasing extent of alternative exons-based splicing, with a fraction of those also harboring altered splice sites under length regulation. Many studies have focused on understanding the roles of alternate exons in tuning protein interactome [26, 27]. However, the corresponding indels association and impact of A(ss) and F cases have not been analyzed. Even molecular consequences of alternate exons have not been concluded convincingly for their impact on translated proteins, and more has been discussed only from the perspective of interaction rewiring. The emerging rise of UTR exons and that of Dual exons warrant their detailed analysis and indicate a strong presence of alternative transcription initiation and termination, which have also been emphasized in higher eukaryotes for establishing tissue-specific transcriptome signatures and identity in addition to splicing-based phenomenon (Reyes and Huber 2018; Shabalina et al. 2014). Their cumulative rise indicates hard to isolate regions and fingerprints on translated proteins; however, it can be easily facilitated using ENACT framework. Expanding on to the finding of this chapter and considering above unanswered impact elucidation and co-participation of alternative transcription initiation and termination with splicing, the next chapter will uncover the extent of those in humans and how much region per gene in 2X2 datasets are altered by those processed while they generate intra-gene transcript variability. Additionally, we uncover the indel impact elucidation of the A(ss) and F cases and the integrity of protein functional domains as they have been introduced in alternate transcripts from the perspective of their intersection with ATIT and splicing dominated regions.

Chapter 4

Detailed investigation into Alternative Splicing and Alternative Transcription induced proteome expansion in *Homo Sapiens*

4.1 INTRODUCTION

The process of alternative splicing introduces transcriptome diversity and contributes to the proteome diversity in eukaryotes (Nilsen and Graveley 2010). The differences in molecular abundances of mRNA and proteins have created a gap in the realization of transcriptome translation to the proteome; consequently have questioned the contribution of alternative splicing to proteome expansion (Tress et al. 2007; Tress, Abascal, and Valencia 2017). Consequentially, disagreement can be noted for AS contribution in creating global proteome-wide diversity (Blencowe 2017). Notably, previous studies have highlighted structural details of splicing-induced modifications and labeled them "drastic" due to perceived lack of potential to translate into physiologically relevant proteins (Tress et al. 2007). However, on close inspection of these studies, it can be realized that substantial changes were observed in the region proximal to N or C-terminal of proteins. (Wang et al. 2005; Tress et al. 2007). The extensive work of Koonin and others have elucidated the functional coupling and intricacies between alternative transcription initiation (ATI) and termination (ATT) with alternative splicing (AS) present among listed gene transcripts in annotated databases (Shabalina et al. 2014).

Detailed investigations into the extent of ATI/ATT (also called ATIT) and AS have been scarce. However, previous studies have deciphered the following: a) There is a 4-fold higher fraction of alternate regions in ATIT than in consistent CDS regions driven by splicing (Shabalina et al. 2014), b) ATIT region are enriched in IDRs and phosphorylation sites (Shabalina et al. 2014), c) Consistent CDS region are enriched in domains (Shabalina et al. 2014), d) There is issue-specific cassette alternative exons responsible for the proteome rewiring (Buljan et al. 2012), and e) UTR exons in the ATIT are shown to be responsible for

tissue-specific splicing identity signatures (Reyes and Huber 2018). The impact of AS events on protein tertiary structures have long been debated in the literature that have been suggested to range from subtle to drastic effect on stability of proteins (section 1.5.1 and section 1.5.2 in chapter 1). In the view of assessing the role of ATIT, these require revisiting previously demonstrated inference drawn from inclusion/exclusion/truncation of domain assessments that raised questions on resulting isoforms structural stability. Specifically, analyzing whether such domains are populated in the ATIT or consistent CDS regions and could these be consequences of alternate promoters or splicing. Considering 4-fold higher fraction of alternate nucleotides in ATI/ATT region (Shabalina et al. 2014), it seems more probable for such regions to often be 4 times populated in ATI/ATT region than in consistent CDS boundaries, provided they were domain or protein encoding regions. In addition, several other important questions can also be addressed to shed light on whether domains lying in the ATIT region or consistent CDS region are differentially regulated. One can study to find, is there one-to-one correspondence between alternate exons in ATI/ATT regions and the same in the CDS region? How many domains are covered by these regions, and how much of domain region truncation will compromise their sequence/structural integrity?

This chapter analyzed human data sourced from NCBI's RefSeq and annotated by ENACT framework. Using the features of ENACT, we differentiated the gene and their corresponding isoforms into ATIT and CORE regions (consistent CDS boundaries) and compared the general alterations introduced separately for those regions. Firstly, we explored the extent of ATIT in the listed human genome as represented in the most representative collection by isoform and coordinates in RefSeq, followed by the relatively underexplored impact of insertion/deletion (indel) introduced due to cryptic splice site choices. Subsequently, we examined the domain and exon boundaries to address whether there is a region bias in domain occurrences, if any for ATIT and CORE regions.

4.2 METHODOLOGY

4.2.1 2X2 set and RISO

We have sourced the data from the NCBI GeneTable resource. For multiple RefSeq entries listed in GeneTable, we selected the reference isoform with the largest number of coding exons, preferably from reviewed gene entries (see section 2.2.3.A). To analyze the prevalence of

splicing in the human proteome, we selected a subset with at least two different protein-coding isoforms and two different exons and termed it a 2X2 dataset (discussed in section 3.2.1).

4.2.2 Secondary structure

The secondary structure of isoforms was predicted by modified PSIPred (Madeira et al. 2022a; Jones 1999) from the I-TASSER package (Roy, Kucukural, and Zhang 2010).

4.2.3 Pfam methodology

Pfam domains were predicted using PfamScan (HMMERv3.2.1) (Madeira et al. 2022a) routine with Pfam 35.0 domain definitions (Mistry, Chuguransky, Williams, Qureshi, Salazar, Sonnhammer, Tosatto, Paladin, Raj, Richardson, et al. 2021). We filtered and considered domains with an e-value cutoff of 0.01 for each isoform and a model length cutoff of ≥ 0.7.

4.2.4 Determining the relative position of exon in an isoform

The procedure to determine relative position of an exon is demonstrated below by taking a hypothetical example. Consider an isoform of 100 amino acids generated from a gene with 10 exons each of 10 amino acid length. Assuming it consists of three exons, which undergo splice site variations such that exon-4, exon-5 and exon-6 shows, ‘n’, ‘b’, and ‘c’ variations, respectively. The definitions of n/c/b variations have already been described in previous chapters. To determine affected position of exon in the isoform, we calculated amino acid length contributions and normalized it by isoform length as follows:

a) For the exon-4 'n' splice site, we consider the cumulative amino acid length contributed by exons preceding exon-4, as is shows 'n' splice site that would affect the left region of the exon.
b) For the exon-6 'c' case, we calculate the cumulative amino acid length contributed by exons up to the length of exon-6, as the 'c' splice site would affect the region on the right of the exon.
c) For the exon-5 'b' case, which contains both splice site changes, we determine the position affected by considering the region affected until the middle of exon-5.

Following the above procedure, specific affected positions are determined based on their amino acid (aa) contributions. In the example, these are: 30 aa for exon-4, 45 aa for exon-5,

and 60 aa for exon 6. We normalize these positions by the amino acid contributions to simply interpretation. As a result, exons-4, 5, and 6 affected positions are 0.3, 0.45, and 0.6 of the isoforms, respectively. Additionally, we assign the 'N-Ter', 'Middle', and 'C-Ter' tags to the affected protein regions based on their relative affected position as follows: 'N-Ter' is assigned if the affected position lies within 0-0.3, 'Middle,' if it lines within 0.3 to 0.7 and 'C-Ter' if it is within 0.7-1.0.

4.2.5 Defining relevant n/c/b splice site variations for isoforms

We define n/c/b splice site change cases based on variation in genomic coordinates (GC). To assess their contribution on protein indels, we compared the assignment of 'aa' from coding genome coordinates and with that of RSOEx forms. The relevant cases to be considered are briefly mentioned below in the following procedure:

a) Coding genomic coordinates (CGC) should overlap between the n/c/b splice site choice in exon and in its RSOEx form.

b) **N case**:
   condLeft = abs(left(RSOEx [CGC]) - left(N splice site exon[CGC])) > 0
   condRight = abs(right(RSOEx [CGC]) - right(N splice site exon[CGC])) in range {0,1,2}

c) **C case**:
   condLeft = abs(left(RSOEx [CGC]) - left(N splice site exon[CGC])) in range {0,1,2}
   condRight = abs(right(RSOEx [CGC]) - right(N splice site exon[CGC])) > 0

d) **B case:**
   condLeft = abs(left(RSOEx [CGC]) - left(N splice site exon[CGC])) > 0
   condRight = abs(right(RSOEx [CGC]) - right(N splice site exon[CGC])) > 0

   here, *abs* is absolute value.

For the 'n' case qualifying above, we will have an anchor site as the right CGC coordinates and assess the feature introduced by querying the protein annotation properties. A similar will be for 'c' cases, where left CGC will be considered anchor and right CGC as variable indel. For the 'b' case, both the splice sites had to be variable but with overlap with both 5' and 3' CGC. Relaxation of 1,2 in CGC genomic coordinates was given above to accommodate 'nt' adjustments for non-3n divisible exons.

4.3 RESULTS AND DISCUSSIONS

4.3.1 The extent of transcript diversity driven by ATIT and AS and the mechanistic basis

In the previous chapter, we observed an increased fraction of alternate coding exons and the gradual emergence of the 'UTR alternate' and 'Dual' categories with increasing organismal complexity. Fraction rise of alternate exons was noted in all three regions of gene architecture (Block I's UTR, Dual and Coding exons) with a gradient preference over constitutive exons, making discernment of them being driven by AS, ATI/ATT or their co-action, difficult to infer. Imperative for such distinctions and their relative contributions in transcript makeup, we divided the coding region of genes and their corresponding isoforms into two different parts: a) CORE: region of the gene where exon dynamics are controlled by splicing and includes exon contributions within the first and the last constitutively coding exons (both the first and the last constitutive exons are considered to be part of CORE), and b) ATIT: region where exon contributions are affected by ATI/ATT processes that includes region outside of the first and the last constitutive coding exons. For this segregation, we employed our nomenclature and chose the 'T' tag from block-I and the 'G' tag from block-II to define the constitutive coding exons. The 'D' Tag from block-I and 'G' from block-II were not considered, as these exons are not coding in all transcripts. In Figure 4.1, we plot the gene fraction in our 2X2 dataset (see methods) with their linear sequence position (LSP) of first encountered TG exon to assess number of exons available for the changes in the ATIT region. We are interested in genes having at least 1 TG exon that will facilitate comparison of contributions between ATIT and CORE. Of 13063 human genes in 2X2 dataset, 11725 genes have at least 1 TG exon criterion. As evident from the Figure 4.1, the first TG exon is at ≥ 5 LSP in ~42% of genes and ~12% of genes encountered the first TG exons positioned at ≥ 10 indicating varying exon region available for ATIT processes to introduce diversity in the isoform. In $1/10^{th}$ of gene fraction lacks TG exons, where these genes either include alternate exons or they have a constitutive exon contributed by DG exons, which were not included in defining the CORE region. These 1338 genes were excluded from the analysis.

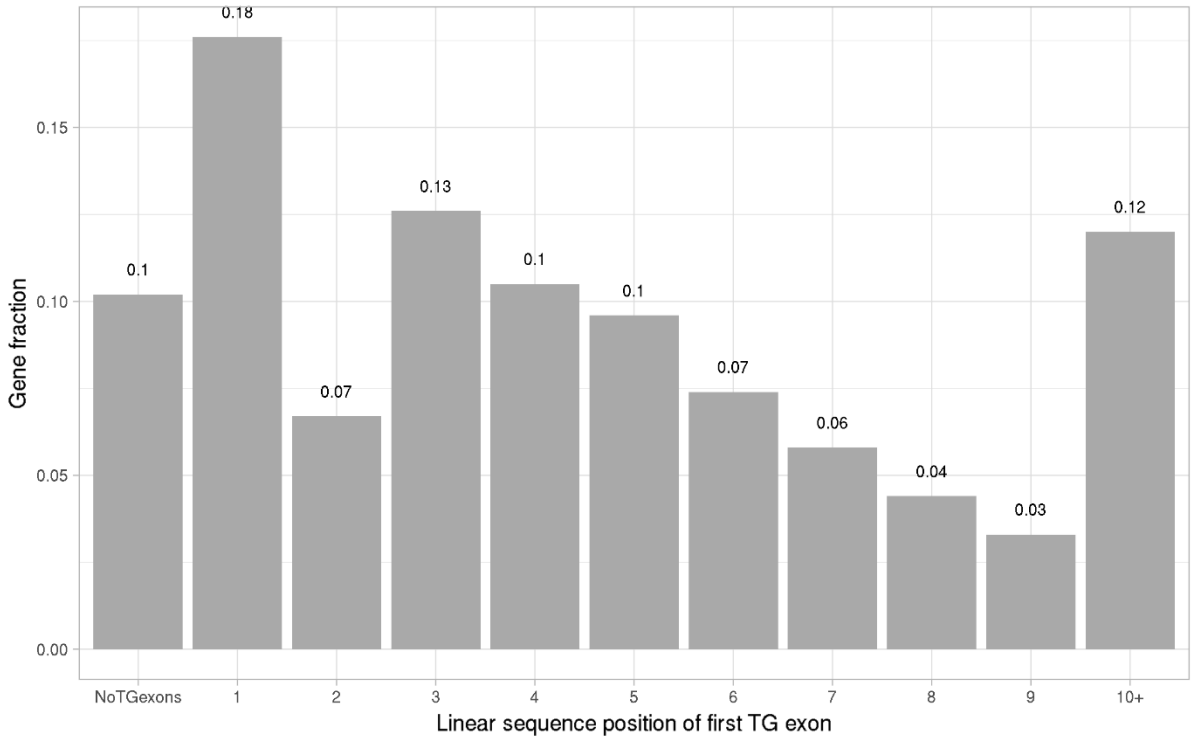


Figure 4.1: Distribution of relative gene fraction for position of the first TG exon. Bar chart showing the distribution of linear sequence position of the first TG exon.

4.3.1.1 Fraction contribution of ATIT rand CORE region in coding gene architecture

The segregation of protein isoforms for 11725 genes into ATIT/CORE region will assist us to analyze the scope of ATI/ATT and AS processes on individual genes and also assess relative contributions of alternate and constitutive fractions. The summary of exon and their contributions to RISO and their respective prevalence in genes is shown in Table 4.1. Among exon subtypes, 'A' exon fraction in ATIT comprises of dual and coding exons of A(ss) (**A**lternate with **s**plice **s**ite choice, 'n'/'c'/'b' fraction), F (constitutive exons but with at least one instance of alternate of 'n'/'c'/'b' splice site choice) and A (not present in all transcripts but has consistent splice sites), whereas CORE region have coding A(ss), F and A subtypes. As observed from Table 4.1, the CORE region (A+G) contributes ~2/3rd of total 'aa' in RISO, whereas rest (1/3rd) is ATIT region. This is consistent with the findings from Koonin's group (Shabalina et al. 2014), which found that the population of coding alternate exons exhibits distinct differences in the CORE and ATIT regions. In CORE, 'A' exons 'aa' (0.41 million) fraction has reduced to 83% than that in ATIT region (2.50 million), and comparably their

influence also reduces to approximately half of the genes (4790) to that in ATIT. It is also important to note that not all exons in ATIT will be coding.

Table 4.1: Summary statistics of exon type distribution in ATIT and CORE regions RISO in the 2X2 dataset.

Region	Exon Type	Fraction of genes (in %)	Fraction of exon numbers (in %)	Fraction of total 'aa' (in %)
CORE	A	4790 (40.85%)	8349 (5.6%)	415090 (6%)
	G	11725 (100%)	80531 (54.4%)	4255869 (58%)
ATIT	A	10072 (86%)	54076 (36.5%)	2501488 (34%)
	G	2234 (19%)	4962 (3.4%)	138203 (2%)
CORE	A+G	11725 (100%)	88880 (60%)	4670959 (64%)
ATIT	A+G	10242 (87%)	59038 (40%)	2639691 (36%)
ATIT+CORE	A+G	11725 (100%)	147918 (100%)	7310650 (100%)

To assess whether variable A/G exons contribution is a general phenomenon for RISO of all gene types or if there are distinct gene populations, which exhibits varying contributions to ATIT/CORE regions, we compared their fraction contribution per RISO in genes of 2X2 dataset (Figure 4.2A). The median contribution of the total ATIT region is 37%, similar to our previous observation. However, a positively skewed distribution can be noted where for 50% of genes have >37% ATIT region, within it 1/4th genes, this cumulatively contributes >66% region of isoforms (upper whisker). In contrast, the CORE region has a negative skewed distribution, with a median contribution of 68% of RISO and for 1/4th of genes encodes 89% to 100% of their sequence. From the exon type assessment perspective, the 'A' exon subtypes of ATIT and 'G' exon subtypes of CORE follow similar distribution to their 'Total' fractions, indicating their dominating roles in respective regions.

4.3.1.2 Roles of ATIT and CORE regions in generating transcript variability

The preceding analysis elucidates the contribution of RISO to different regions. However, a comprehensive assessment of the impact of these regions on the possible functional variability due to these among isoforms of a gene remains elusive. The assessment of functional differences among isoforms is not trivial, as detailed annotations of isoforms are scarce and,

primarily, exist for a handful of experimentally characterized genes. To estimate regional variation of ATIT and CORE and their functional implication within intra-gene isoforms, we utilized protein length variation as a suitable parameter, where length change within CORE and ATIT are expected to reflect changes in phosphorylation, IDR, and domain composition, thereby impacting overall protein function. (Shabalina et al. 2014) (Buljan et al. 2012).

We compared the distribution of length change between RISO and Maximally length Divergent Isoforms (MDI) per gene to gain insights into their possible function variability. The MDI is the isoform having maximum length difference with the RISO form of the gene. We computed the fold length change by normalizing length difference by the RISO length. In addition to overall fold length change within a gene, we assessed the length change of ATIT and CORE regions similarly by considering their respective lengths in MDI and RISO. A positive value would mean isoform (MDI) is shortened and the maximum value can be 1 indicating that MDI is almost negligible in size from RISO. Similarly, negative values indicate MDI being extended (we truncate the scale at -2, indicating a 2-fold length increase from RISO). As can be seen in Figure 4.2B, the median of fold length change is 0.26 for complete isoform suggesting that MDIs are usually shorter with respect to RISO. On comparing their respective ATIT and CORE regions, the distribution shows their contrasting contribution to gene isoforms. The ATIT region is usually truncated 64% of its length from RISO in MDI as indicated by median value and shorter to >95% for $1/4^{th}$ of genes (upper whisker). Additionally, the ATIT region is not always truncated but also extended, as inferred by a long whisker below the first quartile. Contrary to the ATIT region, changes in the CORE are minuscule for MDI and can be considered reasonably modest, as indicated by median value of 0.12 and 0.23 IQR. Extended lengths below zero can also be noted for CORE region but are less populated compared to ATIT region and are often results from less frequent instance of wider exon selection in MDI than in RISO. For instance, in gene ZNF268, skipping exon 6 changes the frame and skips the termination codon, adding 795 ‘aa’ in MDI in distinction to only 5 ‘aa’ in RISO.

4.3.1.3 Role of exon length and inclusion frequency in ATIT/CORE regional variation

Previous analysis of length contribution in generating variability showed presence of an overwhelming gene fraction causing truncation or extension in ATIT. This analysis extends previous studies and emphasizes coding footprint of ATI/ATT exons in generating inter-isoforms variability of genes. Having known this for ATIT, a crucial aspect remains elusive that if alternate exons within ATIT/CORE region are present in atleast ~40% of considered

2X2 gene set (Table 4.1), what differentiates their magnitude change in isoform makeup in addition to their marked different frequencies (Table 4.1). Specifically, we wanted to know whether marked infrequent population of theirs are also associated with sub population-based differences in length and inclusion frequency of alternate exons. To gain insights into it, we analyzed length distribution of alternate/constitutive from ATIT and CORE regions (Figure 4.2 C), which has the top two panels showing the coding and dual exons from ATIT region, and the bottom panel has coding exon length distribution from CORE region.

Additionally, instead of relying on exon length of only RISO, we included length of all exons (alternate exons from other transcripts) for analysis. If an exon has different coding variations in other isoforms, we computed their average amino acid lengths for the exon to better represent intra-exon length. Notably, such analysis is feasible because ENACT provides relative position of exon and facilitates tracking an exon by its relative position. The relative frequency of various types (A/A(ss)/G/F) exon in ATIT/CORE region and their occurrences in genes are summarized in Table 4.2. This differs from previous table (Table 4.1) regarding exon contribution in addition to RISO. As seen in Figure 4.2C, that the 'aa' length of coding 'A' in ATIT (IQR: 22-51, median: 35) is almost similar to distribution of 'A' in CORE region (IQR: 23-49, median: 34). In the coding category, only A(ss) exons were found 12% longer in CORE than ATIT, and other exons were almost of comparable lengths. The Dual ('D') exon subtypes are often shorter by $1/4^{th}$ to $1/3^{rd}$ than the lengths of exons in the CORE region. Analyzing this observation with Table 4.2 count of exons, it can be emphasized that coding 'A' exon subtype in alternate exon category of ATIT (A+A(ss)+F) shares similar length distribution with its counterpart in CORE region, indicating length is not associated with heightened variability of previously observed ATTT region.

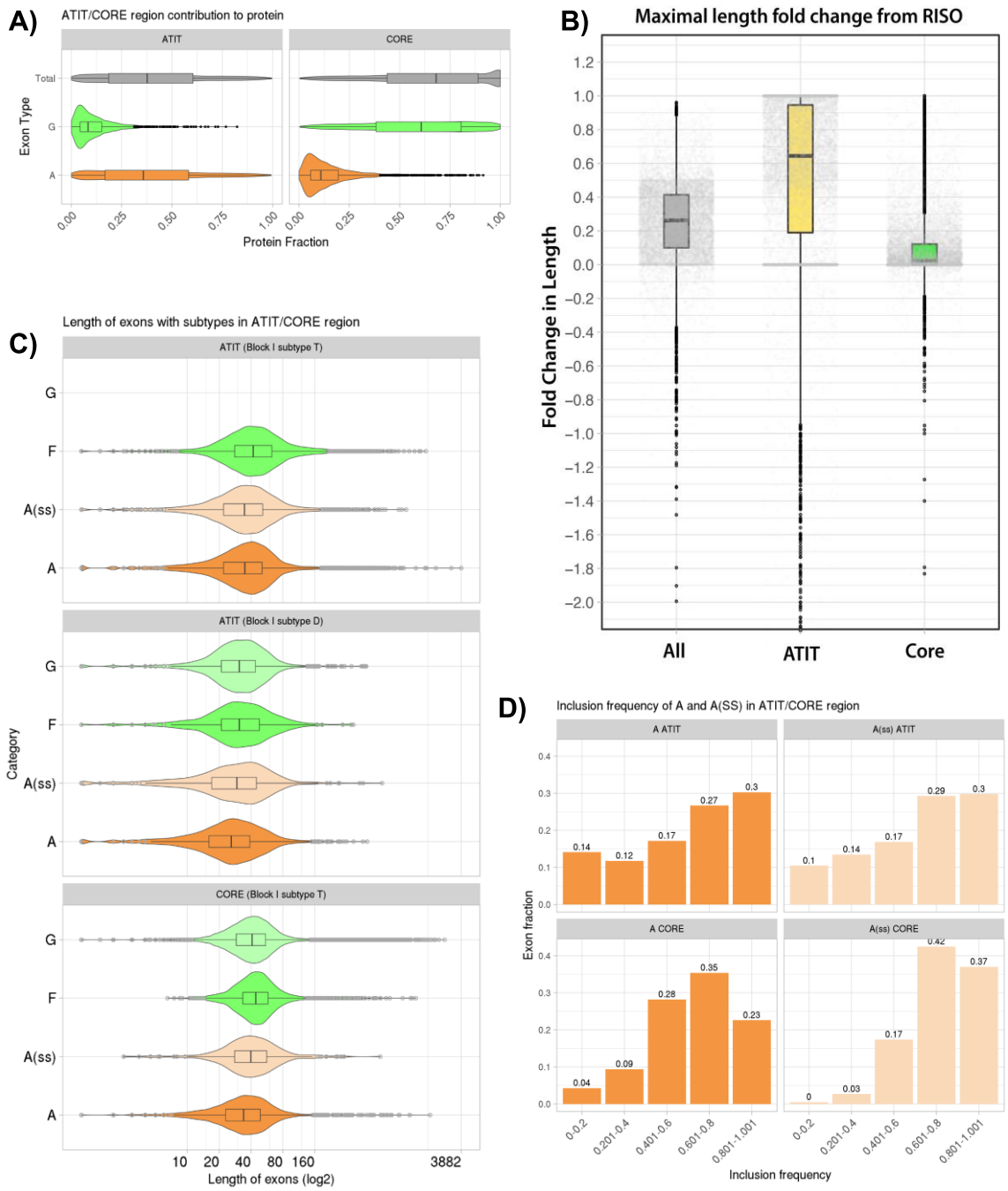


Figure 4.2: Assessment effect of ATIT and CORE region on 11725 gene subsets of 2X2 dataset having at least 1 constitutive coding exon. A) Box plot showing the contribution of A/G exon types in the ATIT and CORE region of RISO. B) Box showing the fold length change computed between maximally divergent isoform from RISO and segregated for ATIT and CORE regions. C) Length distribution of various exon subtypes in the CORE and ATIT regions shown as box plot overlaid with their distribution D) Bar chart showing the inclusion frequency of alternate (A) and alternate with splice site changes (A(aa)) from ATIT and CORE regions. The inclusion frequency is binned in 5 separate bins.

Table 4.2: Exon type prevalence in ATIT and CORE regions.

Region	Block I	Exon Type	Gene count (Fraction in %)	Exon count (Fraction in %)
ATIT	T	A	8505 (72.5%)	40478 (26.5%)
		A(ss)	4098(35%)	6734 (4.4%)
		F	5010 (42.7%)	6412 (4.2%)
	D	A	2473 (21.1%)	4225 (2.8%)
		A(ss)	806 (6.9%)	993 (0.7%)
		F	2303 (19.64%)	3852 (2.5%)
		G	506 (4.3%)	546 (0.4%)
CORE	T	A	3740 (32%)	6223 (4%)
		A(ss)	370 (3.16%)	414 (0.3%)
		G	11725 (100%)	80531 (52.7%)
		F	1884 (16.07%)	2305 (1.5%)

Another feature of alternate exons that may influence variability in ATIT region is their inclusion frequency of their sub-types. To assess this, we compared the inclusion rates for the A and A(ss) sub-categories within the ATIT and CORE regions as shown in Figure 4.2D. As can be seen, 26% of 'A' exon types have an inclusion frequency of <40% in the ATIT region, and for the same inclusion frequency (<40%) there are only 13% of 'A' exons in the CORE region. This indicates that there are differences in A exon sub-type, where in ATIT region these prefer to assimilating in isoform often a small number of times than in CORE region. On the contrary, a similar fraction (~57%) of 'A' exon is observed with inclusion frequencies ≥60% from both ATIT and CORE regions. The A(ss) exons, however, show trend in both directions, where these form part of CORE more frequently than ATIT as indicated by 15% of A(ss) fraction showing inclusion frequency <40% in ATIT and 3% in CORE. Similarly, 60% of exons has ≥60% inclusion frequency in ATIT in comparison to 80% in the CORE region. These differences in inclusion frequency may be associated with increasing variability of ATIT region, where exons often have a lower inclusion rate than in the CORE. While this analysis provides a fraction-wide comparison for highlighting the contrast between A and A(ss) exons

inclusion rates in ATIT and CORE region, it should be highlighted that the population of A exons in ATIT are at least twice the CORE and more than 10 times for A(ss) category (Table 4.2). Such proportion differences in synergy with distinct inclusion frequencies of theirs seem plausible in generating inter-transcript length variations, as observed in Figure 4.2B.

4.3.2 Molecular consequence of Alternative splice site choices

The results from previous last analysis detailed that the ATIT region is highly variable among length-divergent gene isoforms, and along with the fold length change count of Alternate exons (‘A’ and ‘A(ss)’) in those regions, their inclusion frequency differences make them susceptible to undergo length change. Continuing the extent exploration of their subtypes with possible molecular impact differences in ATIT/CORE region, we hereby focus first on the ‘A(ss)’ group. The ‘A(ss)’ exons showed greater inclusion frequency in CORE than ATIT, indicating their frequent assimilation within splicing driven regions. We extend this knowledge of their inclusion frequency differences to compare how they affect protein regions and how their indel introduction is compared between CORE and ATIT regions. Abundant literature reports the choice of alternative splice sites, their strength, and strict regulation by complex interplay of many cis and trans factors (Koren, Lev-Maor, and Ast 2007), but not how these alternate splice site choices affect the resulting protein isoform. This aspect has been overlooked for several reasons, one of them being data representation of splicing isoforms, where many previous studies relating splicing and its impact in proteome sourced data from UniProt (which generally lacked exon definitions) (Tress et al. 2007; Wang et al. 2005). The splicing isoforms in UniProt were compared in multiple sequence alignment (MSA) manner to infer events and lacked sensitivity to distinguish ‘A’ and ‘A(ss)’ events. The importance of ‘A(ss)’ can be realized by their gene prevalence and shall not be underestimated to their small exon contribution (Table 4.2). Implications of their dysregulation have been discussed in literature where they could activate cryptic splice sites, potentially leading to unhealthy transcriptomes (Divina et al. 2009). Considering their importance, we focused on the 2X2 dataset and quantified gene prevalence harboring (‘n’/’c’/’b’) splice site choices. We studied a) the positional prevalence of such regions to affect the protein transcript. b) How much indel change do A(ss) incorporate, c) How much secondary structure regions do they add or delete, and c) how does the region that anchors the indel (an unchanged portion of the exon) maintain its sequence integrity during these alterations?

We calculated the position of an exon in their participating transcripts (see methods). We considered unique exon pairs from the gene (parent form from RSOEx set and its 'n'/'c'/'b' form) for assessment and their repetitions if they affected different regions of isoforms. Among the pairs (Table 4.3), more than ½ are coding in all three 'n', 'c', and 'b' categories, and a minor fraction ($1/5^{th}$ to $1/4^{th}$) is part of non-coding region (Figure 4.3). Among 'n'/'c'/'b', category, 'n' has the highest fraction of coding-only pairs, and category 'b' has the highest fraction of noncoding-only pairs. The differences in the fraction of coding and non-coding pairs were accounted to regions where the choice of alternate splice site changes the coding nature of exon and makes them transition between coding and non-coding parts of the protein transcripts variably, and this ranges from 14-24% for all 'n', 'c' and 'b' cases. Such events may result from the interplay of ATIT driven processes with AS, as transition to non-coding region/UTR is observed, indicating exon definition influenced change in translation start sites. To check this, we segregated all 'n', 'c', and 'b' pairs in ATIT and CORE regions as previously described in section 4.2.1 and observed that non-coding and 'transitioning' pairs are localized to indeed ATIT region (outside the first and last coding constitutive exon), conforming them to have a role in translation initiation and termination (Figure 4.3).

Table 4.3: The n/c/b splice site pairs observed for 11725 genes and their segregation into ATIT and CORE regions.

Tag	'n'	'c'	'b'
Total	22216	16488	3407
ATIT	14497	11830	2689
Core	4169	1846	37

Additionally, ATIT region has more than $2/3^{rd}$ pairs collectively of all 'n'/'c'/'b' subtypes, and their frequency in CORE is notably scarce (Table 4.3). In the CORE region, the 'n' cases are $<1/3^{rd}$ of ATIT; 'c' cases are less than $1/5^{th}$ and 'b' cases <1/100. The differences in total pairs of ATIT+CORE (Table 4.3) from the total numbers are from genes neglected in our analysis, *i.e.* genes with no TG exons.

As our interest lies in the protein-coding subsets and their corresponding indel adjustment assessment in isoforms, we compared the coding only pairs after segregating them in ATIT and

CORE regions. The elucidation of their impact can only be examined after careful incorporation coding genomic coordinates (CGC) (subsets of genomic coordinates (GC)) in amino acid (‘aa’) assignment. We only considered indel region of pairs for whom non-varying sites remain intact in pair members, for example, the right CGC and left CGC should be identical in ‘n’ and ‘c’ case, respectively, and for ‘b’ case, both left and right CGC should be varying but with CGC overlap (see Figure 4.4A). To facilitate distinguishing these, we have listed criteria to filter cases not following the ‘n’, ’c’, and ’b’ splice site region for CGC coordinates in method section 4.2.5. The pairs that did not qualify for the above criteria had complex CGC and GC organization and provided a non-trivial challenges in assessing the indel contribution to isoforms, as shown in Figures 4.4B, 4.4C, and 4.4D. Their prevalence is quantified in Table 4.4. Briefly, they are classified into “ATIT: Same amino acids”, “ATIT: Different Amino Acids” and “Variation in CGC” tags for ‘n’/‘c’/’b’ splice site subtypes (Figure 4.4 panel B and C) and “Pseudo C”, “Pseudo N” and “No overlap exists”, tags for ‘b’ splice site cases (Figure 4.4 panel D).

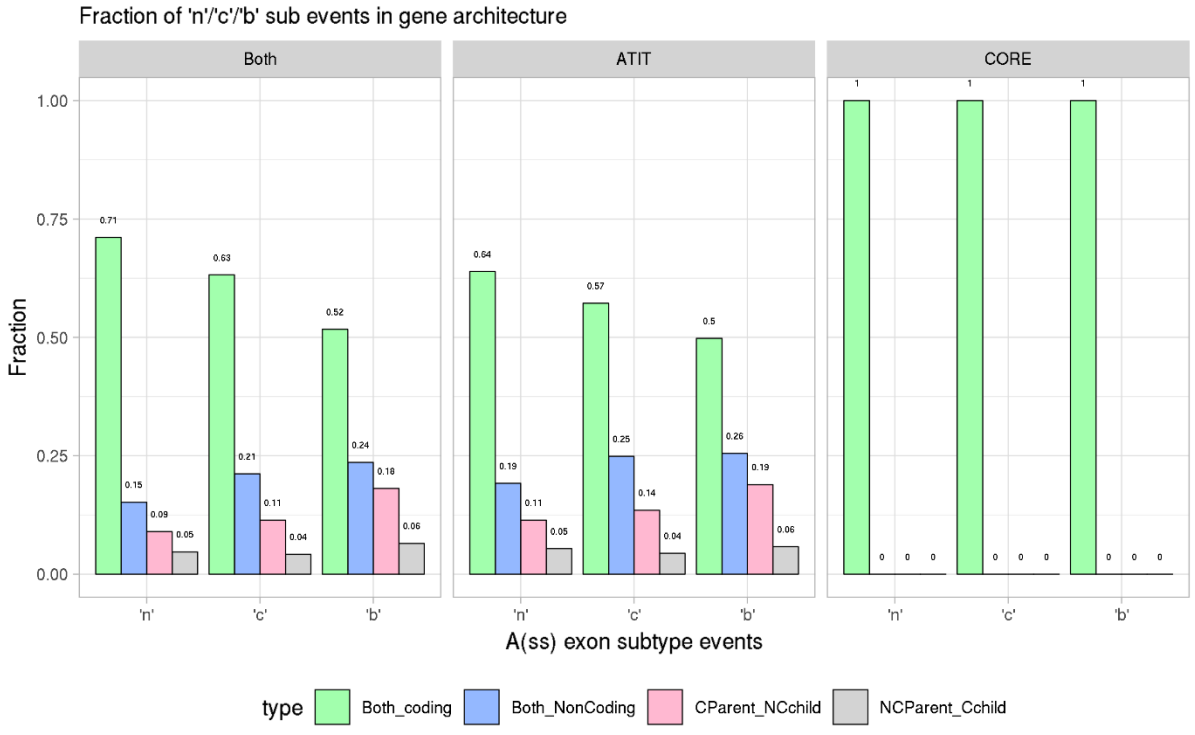


Figure 4.3: Relative fraction of coding, non-coding and transitioning exon pairs classified as ‘A(ss)’ pairs in genes. A) Bar showing the fraction contribution of exon pairs having variation of n/c/b in A(ss) exons for entire gene region without segregation into ATIT and CORE region. B) ATIT: Bar chart showing pairs which were present outside the 1st and last constitutive coding exon with tag T(coding, block I) and G (constitutive, block II) and C) CORE: exons pairs showing n/c/b/ variations defined the CORE regions.

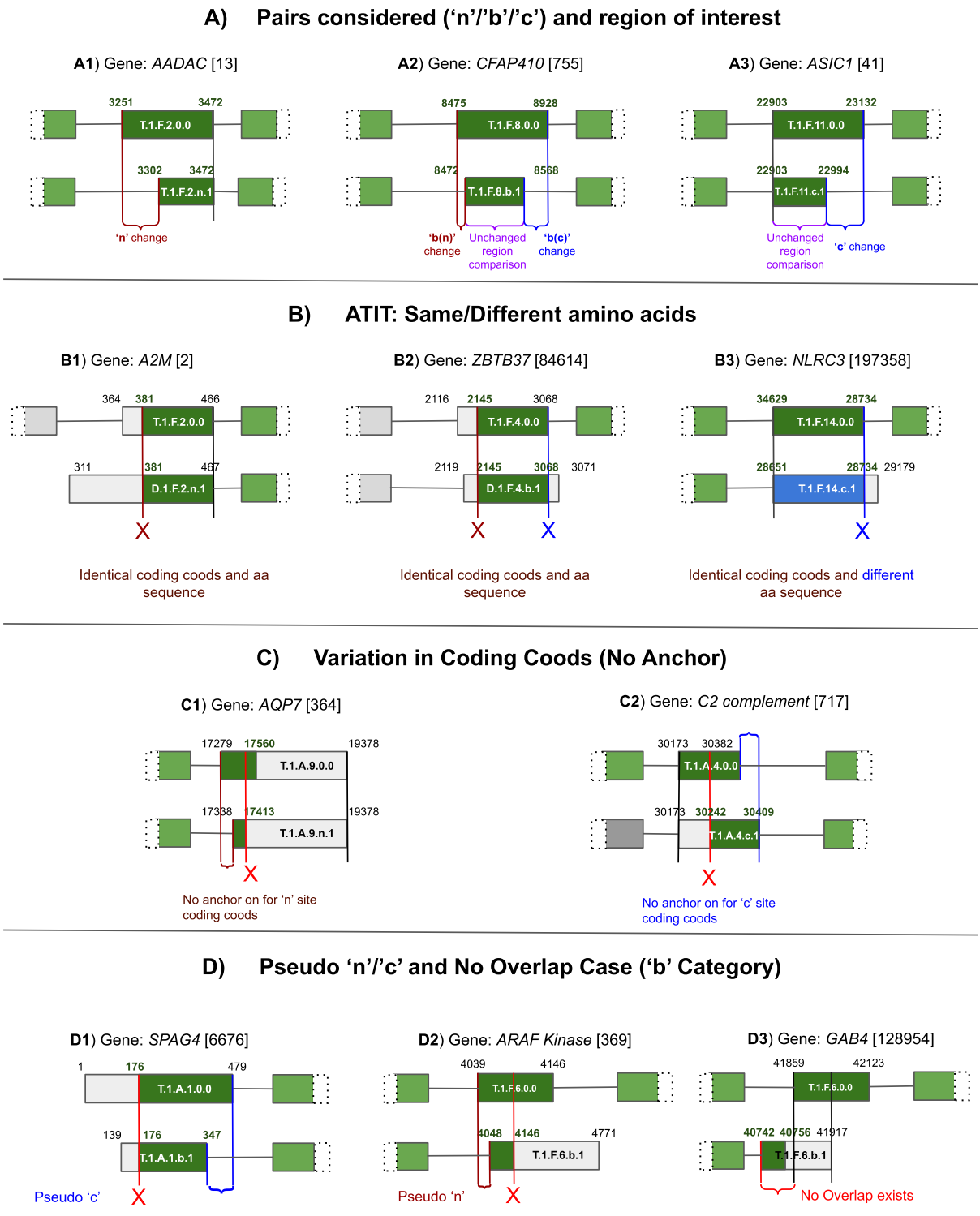


Figure 4.4: Concise representation of the complexities that emerged after considering the 'aa' assignment in genomic coordinates of exons for their splice site indel assessment. A) It shows the pairs, which we have considered for their indel assessment. The 'n', 'b', and 'c' splice site choice pairs are discussed from left to right. Brown colored region depicts the indel

of ‘n’, and blue colored region depicts the ‘c’ region as named in the figure for their respective indel contribution assessment, opposite of ‘n’ and ‘c’ splice site choices are considered anchored and unchanged with the relaxation of 1-2 nt as these may be added/removed during ‘aa’ assignment to exons. The ‘b’ site splice choice case will not have any anchor site in coding genomic coordinates and, if found, were neglected for the subsequent analysis, and separate N-Ter and C-Ter regions are chosen per ‘b’ site change as depicted by B(n) and B(c) region in the middle of panel A2). The Panels B and C show the cases that contribute to protein in regions but were not considered either because of a lack of anchor site or non-variable CGC (a subset of GC used to define the ‘n’/’b’/’c’ events). B) depicts event driven by alternative transcription initiation and termination and splicing’s interplay; those events have changed the exon definition enough for the definition of ‘n’/’b’/’c’ case, but with identical coding genomic coordinates and hence nonamenable to assess indel contribution. Some cases have identical ‘aa’ (B1 and B2), and some have different ‘aa’ (B3). C) Shows cases where the anchor site is questionable in events of interest and is likely to be driven by ATIT and AS’s interplay. These cases are also prevalent at the termini, or they create the termini by introducing premature termination codons. D) Illustrates cases having not-considering B site splice choice cases, where ‘Pseudo N’ and ‘Pseudo C’ category depicts ‘n’ and ‘c’ indels only without corresponding opposite sides of theirs. Another category in B cases includes ‘No overlap exists’, which, as displayed, depicts pairs whose definitions are as per ‘B’ event in GC but no overlap between CGC.

For qualifying pairs (Figure 4.4 A, or ‘Relevant’ Category in Table 4.4), we observed a considerable count of these have indel of length ≤2 aa (Table 4.5), which were neglected in subsequent analyses as they may be consequence of annotation mismatches by few nucleotides, or these can be reasonably accommodated in proteins in comparison to longer indels. The relevant pairs (>2aa indels) showed similar disproportional distribution between CORE and ATIT regions, where the latter has twice, and more overall pairs than CORE. A similar trend in their gene prevalence can also be noted where 1532 genes harbor splicing site choice within the first and last of constitutive coding exons (CORE region), in contrast to 3941 genes in ATIT (2.5 times of CORE) as shown in Table 4.5. The comparison of their subtypes ‘n’/’c’/’b’ shows that CORE region has lower proportion of ‘n’ cases by 2-fold and ‘c’ case by 3-fold relative to ATIT region, whereas ‘b’ splice site changes infrequent of all. For these pairs, we compared the subset of shared (unchanged fraction) and different (overhang indel) CGC and assessed predicted secondary structure features and region of the protein affected (see methods section 4.2.4).

4.3.2.1 Analysis of affected transcript regions

We investigated if the transcript region affected by n/c/b indels differs between ATIT and CORE. It is pertinent to note that the defined CORE region does not necessarily specify the middle of the protein region. Moreover, in Figure 4.1 we have observed a considerable number of genes having linear positioning of the first TG exon ≥10, indicating a large region available for ATIT exons that would correspond to the middle region of respective encoded isoforms. Figure 4.5 shows the transcript region affected by indels separately for the 'n'/'c'/'b' cases (see method section 4.2.4). It can be observed from ATIT (upper panel) and CORE (lower panel) that 'n' pairs introduce indel near N-terminal of the proteins in ATIT with a very small population affecting positions from 60-80% of transcript and close to C-terminal of protein. Similarly, the 'c' splice site pairs introduce indels close to C-terminal in ATIT, and the 'b' splice site pairs primarily affect the C-terminal with a noticeable population near the N-terminal for ATIT region. In contrast, exons pairs in CORE region affect close to the middle and C-terminal of proteins for all 'n'/'c'/'b' cases.

Table 4.4: Summary of relevant pairs having complex cases as shown in Figure 4.4 having in n/c/b exon splice site variation.

	CORE			ATIT		
Type	**Count 'n' (Freq in %)**	**Count 'c' (Freq in %)**	**Count 'b' (Freq in %)**	**Count 'n' (Freq in %)**	**Count 'c' (Freq in %)**	**Count 'b' (Freq in %)**
Relevant	4161 (99%)	1845 (99%)	37 (100%)	6013 (65%)	4730 (70%)	515 (38.5%)
ATIT: Same amino acid	8 (0.2%)	1 (0.1%)	0	1862 (20%)	1521 (22%)	1 (0.1%)
ATIT: Different amino acid	0	0	0	5 (0.1%)	67 (1%)	0
Variation in CGC	0	0	0	605 (6.5%)	218 (3%)	0
Pseudo 'c'	0	0	0	0	0	52 (3.9%)
Pseudo 'n'	0	0	0	0	0	127 (9.5%)
No overlap exists	0	0	0	783 (8.4%)	227 (3%)	644 (48%)
Total	4169 (100%)	1846 (100%)	37 (100%)	9268 (100%)	6763 (100%)	1339 (100%)

Observing their distinct difference in introducing indels by n/c/b in protein regions, we compared them for altered secondary structured regions as well as positional impact on the transcript by binning indels based on length. Such length-based binning would help assess and identify appropriate cohort discernment for a specific indel length range.

Table 4.5: Comparative count summary of exon pairs having indel length ≤ 2aa and >2aa in various genes.

Affected protein change length >2						
	CORE			**ATIT**		
A(ss)	**Gene count**	**Exon Positions**	**All Variations**	**Gene count**	**Exon Positions**	**All Variations**
All	1532	1563	3948	3941	4215	8769
‘n’	969	987	2352	2101	2251	4179
‘c’	721	735	1575	2363	2471	4218
‘b’	16	16	21	279	281	372
Affected protein change length ≤2 aa						
	CORE			**ATIT**		
A(ss)	**Gene (count)**	**Exon Positions**	**All Variations**	**Gene count**	**Exon Positions**	**All Variations**
All	841	859	2095	1098	1134	2489
‘n’	728	746	1809	771	794	1834
‘c’	133	133	270	299	305	512
‘b’	9	9	16	118	120	143

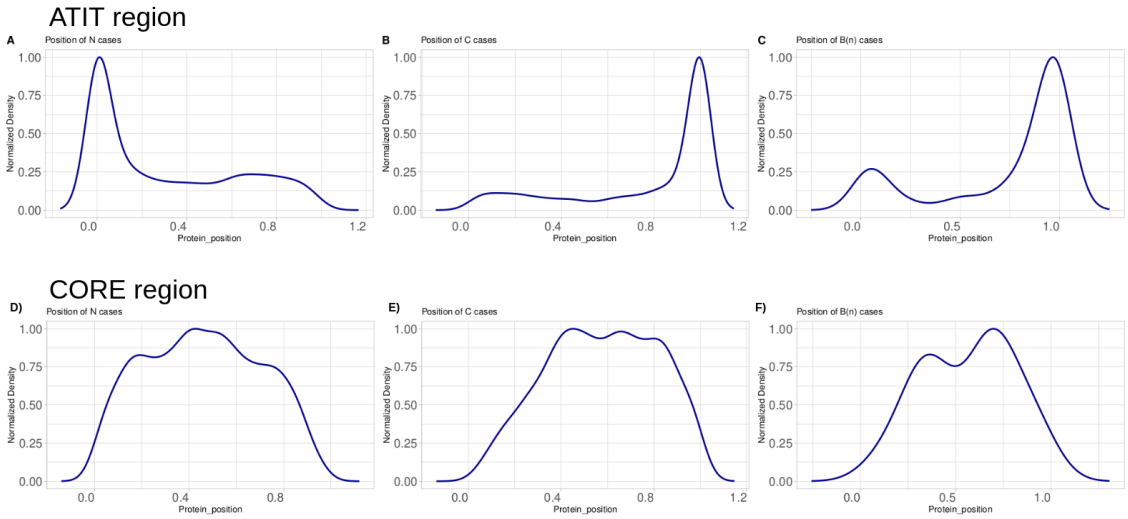


Figure 4.5: Density distribution of the affected protein region by introducing N, C, and B splice site regions. The top row depicts the density distribution of the ATIT region and the CORE region's bottom row.

4.3.2.2 Impact of 'n' subtype

The comparison of 'n' sub-type, which is the largest fraction of 'n'/'c'/'b' pairs showed ATIT region harbors comparatively longer indels than that of the CORE region. For instance, <30aa indels are 72% of all pairs in the ATIT, whereas, in CORE, they represent 80% of the fraction. Additionally, indels exceeding 50aa constitute 13% of the fraction in ATIT, whereas they comprise only 8% in CORE (Figures 4.6A and 4.6B). Though the ATIT region introduces larger indels, not all are localized to the N-Ter region. We observed that apart from 50% of all indels across various indel lengths affecting C-Ter and N-Ter regions, ~1/4th was also found lying in the middle of protein region, especially for <30aa lengths. For longer indels (>30aa), the proportion affecting the middle of the protein region decreases gradually to only <20 of pairs (Figure 4.6C). In contrast, ~50% of all pairs in the CORE region indels are localized mostly in the middle of the protein region (Figure 4.6D).

Subsequently, we analyzed secondary structure involving indels for ATIT/CORE regions. The secondary structure encompassing indels in the ATIT are predominantly coiled residues (affected more than 50% across all lengths). A gradual increase in the affected coil proportion with increasing indel length can also be observed where the fraction of coiled residues increased to almost 2/3rd of all affected secondary structure types, indicating their relatively easy accommodation with increasing length. In the CORE region, especially under 30aa indel lengths, more than 1/3rd of all pairs are helices, slightly higher than the ATIT region, and coiled residues are close to 50% in all bins. With the increasing indel size in CORE, a similar trend

was noticed in ATIT region, where introduction/removal of coiled states increased to 1/3rd of all pairs. Thus, suggesting the longer indels are easily accommodated in the unstructured (coil) region of the protein

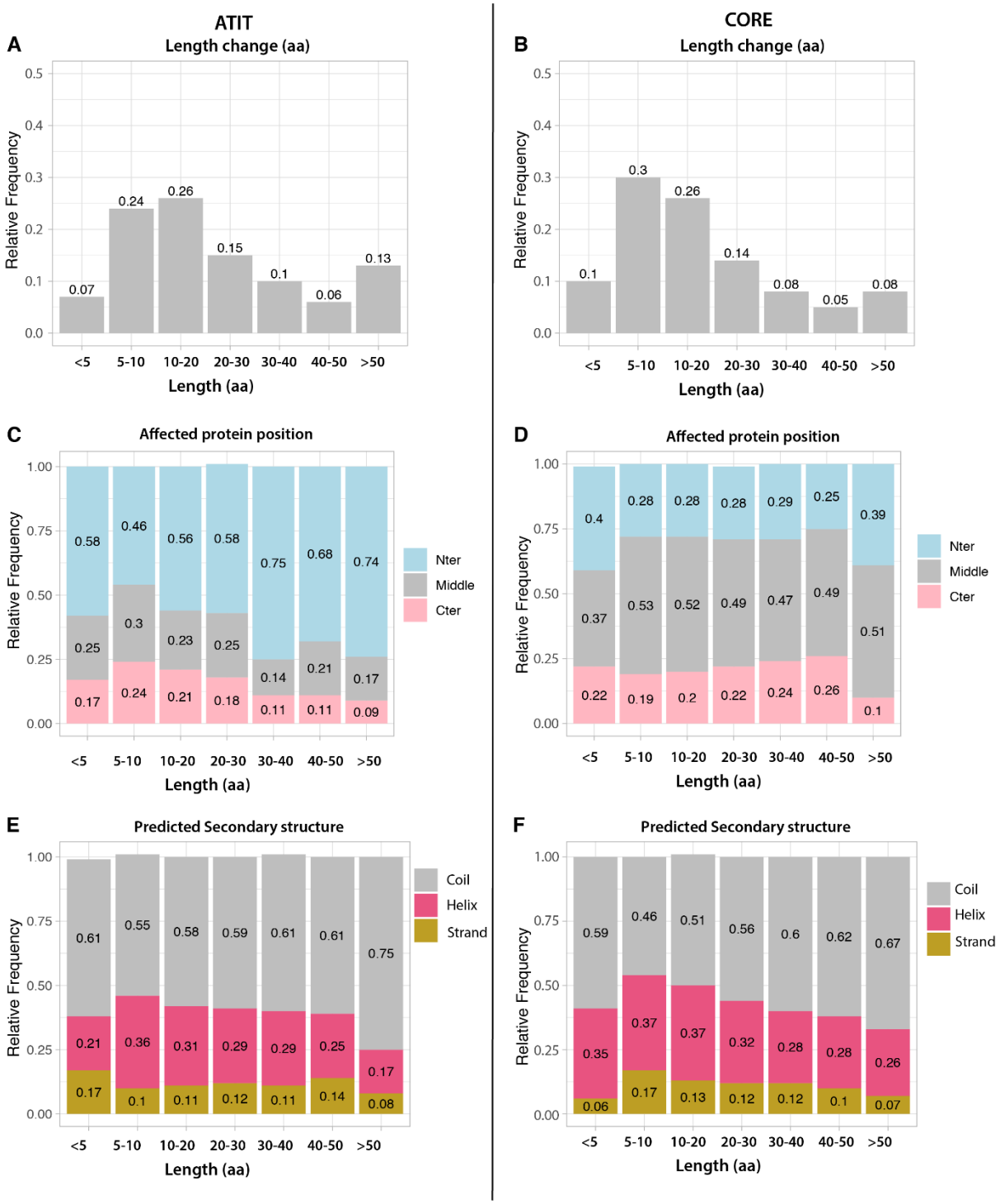


Figure 4.6: Impact assessment of 'n' splice site indels. The left column depicts the ATIT region, and the right column is the CORE region. A) and B) show the frequency distribution of absolute indel length change (from their RISO form). The panels C and D shows the region of protein affected by their introduction where N-Ter is the first 30% of the region, the middle is

30-70%, and C-Ter is changing in the end 70% to complete C-Ter of the transcript region; the bars depict the length range bins defined in panel A and B and range is also shown in the X axis. Similarly, panels E and F show the affected secondary structure content for those defined length range bins by their introduction.

4.3.2.3 Impact of 'c' Subtype

Comparing the indel length frequency ATIT and CORE region for 'c' site cases (Figure 4.7), a similar pattern of distribution like 'n' cases can be seen, where indels of smaller length are more prominent in the CORE (74%) than in ATIT (61%), and fraction of indels exceeding 50aa length are almost twice the fraction in ATIT (20%) than in CORE (12%). In context to affected position, 2/3rd of ATIT pairs are frequently localized to C-Ter, and with increasing length (>30aa bins), the fraction of pairs affecting regions in the middle reduces to only 1/10th of the pairs. This observation also agrees with the previously noted overwhelming population localizing to C-Ter region in the Figure 4.5. The indels in CORE region affect the middle of protein in >50% cases, and their tendency to affect coiled region increases with increasing length. There is no noticeable fraction change of secondary structure regions affected in different indel length bins for ATIT and CORE region, and both of those regions almost equally affect coiled/unstructured region more than half of the time with minor alterations in helix (1/3rd of times in all indel length bins but >50aa).

4.3.2.4 Impact of 'b' subtype

The number of 'b' cases with their termini regions introducing the 'n' and 'c' indels are only a handful of all cases, affecting 279 genes in ATIT and only 16 genes in CORE (Table 4.5). Their 'n' indel lengths are of comparable frequency in almost all bins for ATIT and CORE. In the ATIT region, they prefer affecting C-Ter for small indels and N-Ter for larger indels, whereas, in CORE, they often affect the middle region. Regarding secondary structure, significant alteration in helical regions is prominent in CORE and contrasts with ATIT, which prefers affecting coiled regions (Figure 4.8C). Irrespective of the small handful of 16 affected genes assimilating B genes in CORE, their significant introduction in secondary structural regions makes them a promising set for analyzing splicing extent and accommodation in future studies.

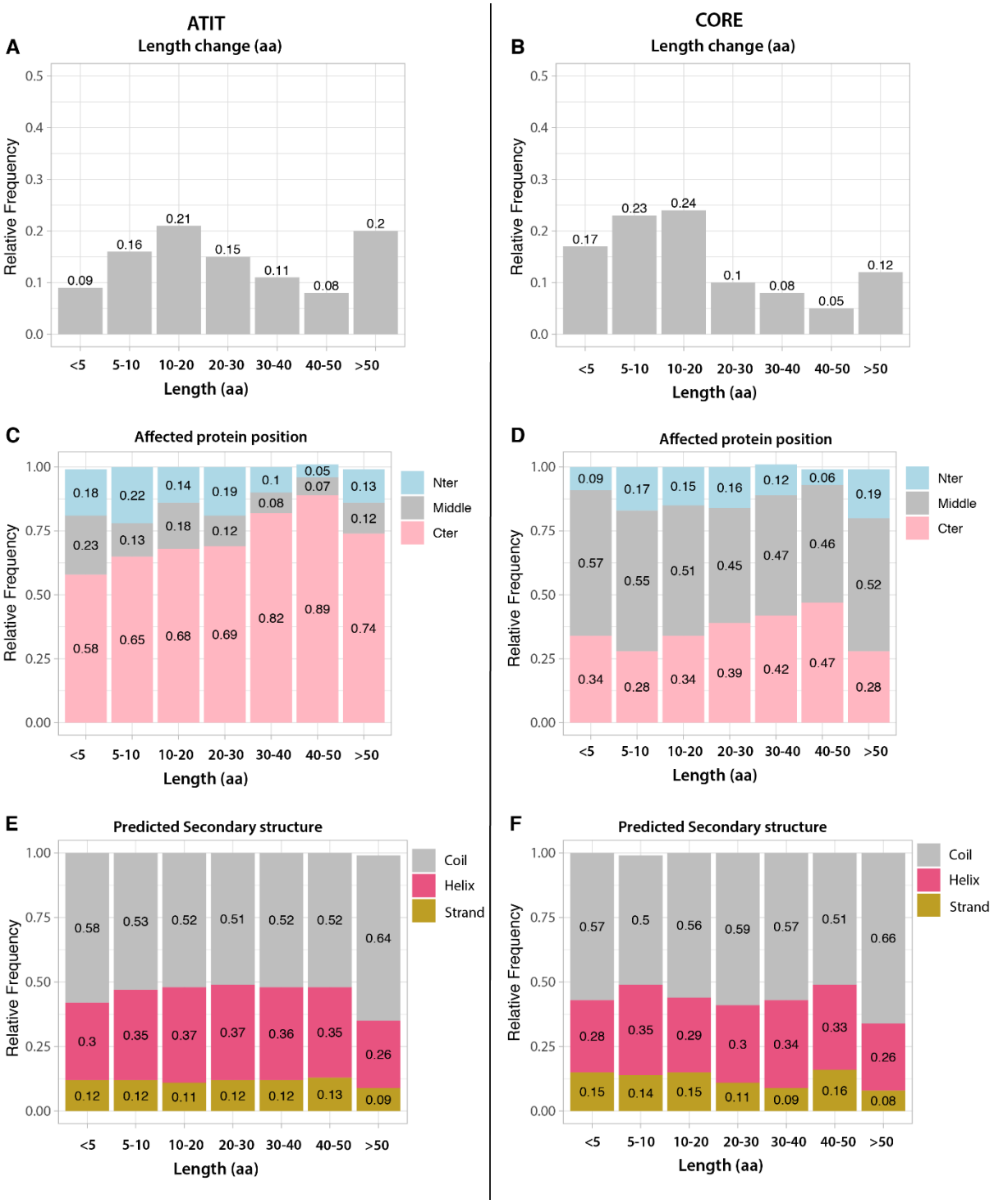


Figure 4.7: Impact assessment of 'c' splice site indels. The left column depicts the ATIT region, and the right column is the CORE region. Panels A and B show the frequency distribution of absolute indel length change (from their RISO form). Panel C and D enlist the region of protein affected by their introduction where N-Ter is the first 30% of the region, the middle is 30-70%, and C-Ter is changing in the end 70% to complete C-Ter of the transcript region; the bars depict the length range bins defined in panel A and B and range is also shown in the X axis. Similarly, panels E and F show the affected secondary structure content for those defined length range bins by their introduction.

For the ‘c’ indel changes, smaller lengths are apparent in CORE and larger in ATIT. Similar to B(n) regions affecting C-Ter often, their localization to N-Ter regions can be noticed, however, but more pronounced. In contrast to B(n) cases, N-Ter localization of theirs is preponderant for smaller lengths, and increased localization to C-Ter can be noticed with increasing indel length Figure 4.9C. As discussed before, their affected region in CORE has been sparse but often harbors the middle of the protein region. In the context of their secondary structure alterations, a trend similar to B(n) regions can be noticed in ATIT and CORE, where they prefer to affect helical regions comparatively higher in B(n) and B(c) for ATIT region.

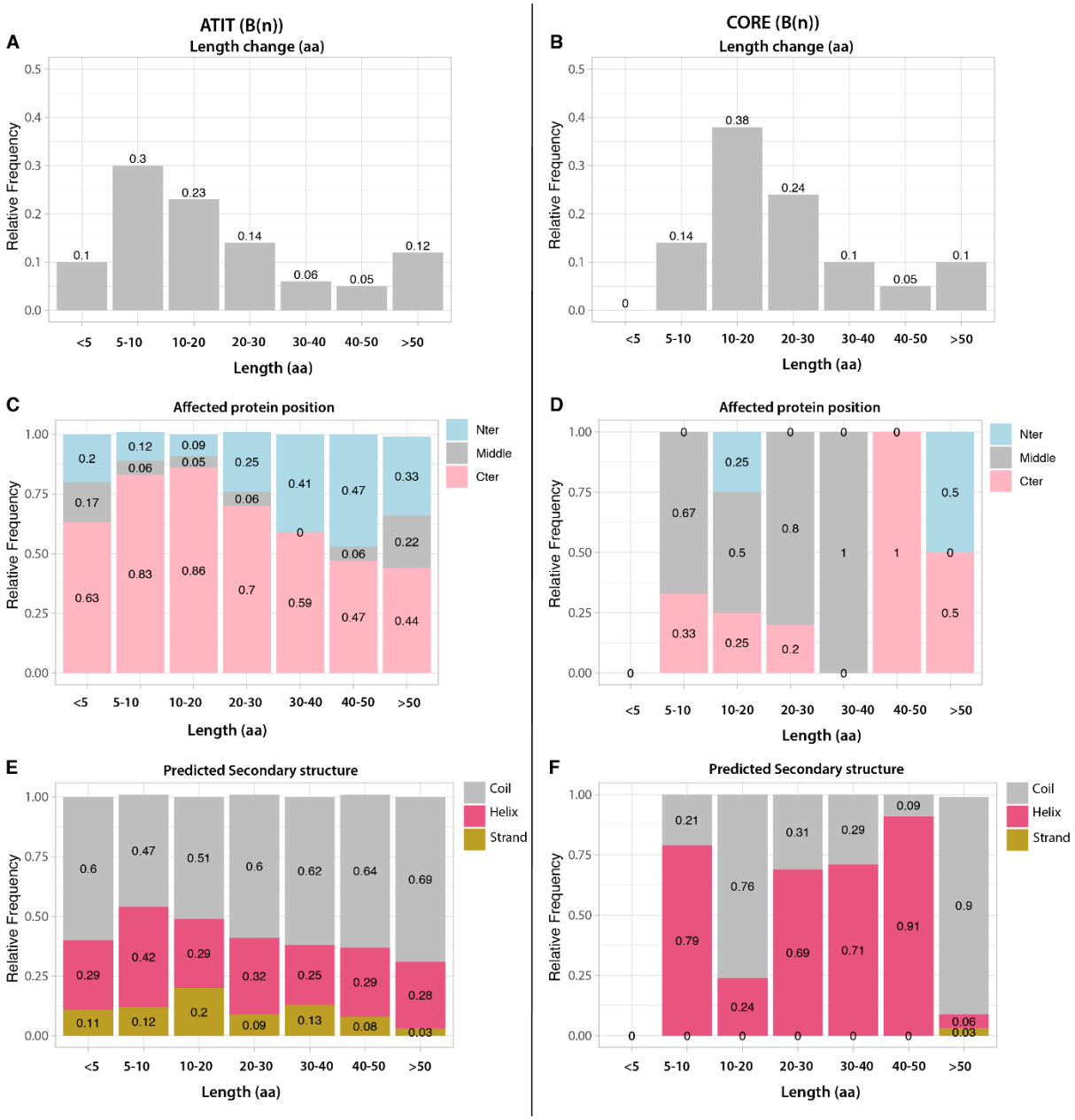


Figure 4.8: Impact assessment of B(n) splice site indels. The left column depicts the ATIT region, and the right column is the CORE region. Panels A and B show the frequency

distribution of absolute indel length change (from their RISO form). Panel C and D enlist the region of protein affected by their introduction where N-Ter is the first 30% of the region, the middle is 30-70%, and C-Ter is changing in the end 70% to complete C-Ter of the transcript region; the bars depict the length range bins defined in panel A and B and range is also shown in the X axis. Similarly, panels E and F show the affected secondary structure content for those defined length range bins by their introduction.

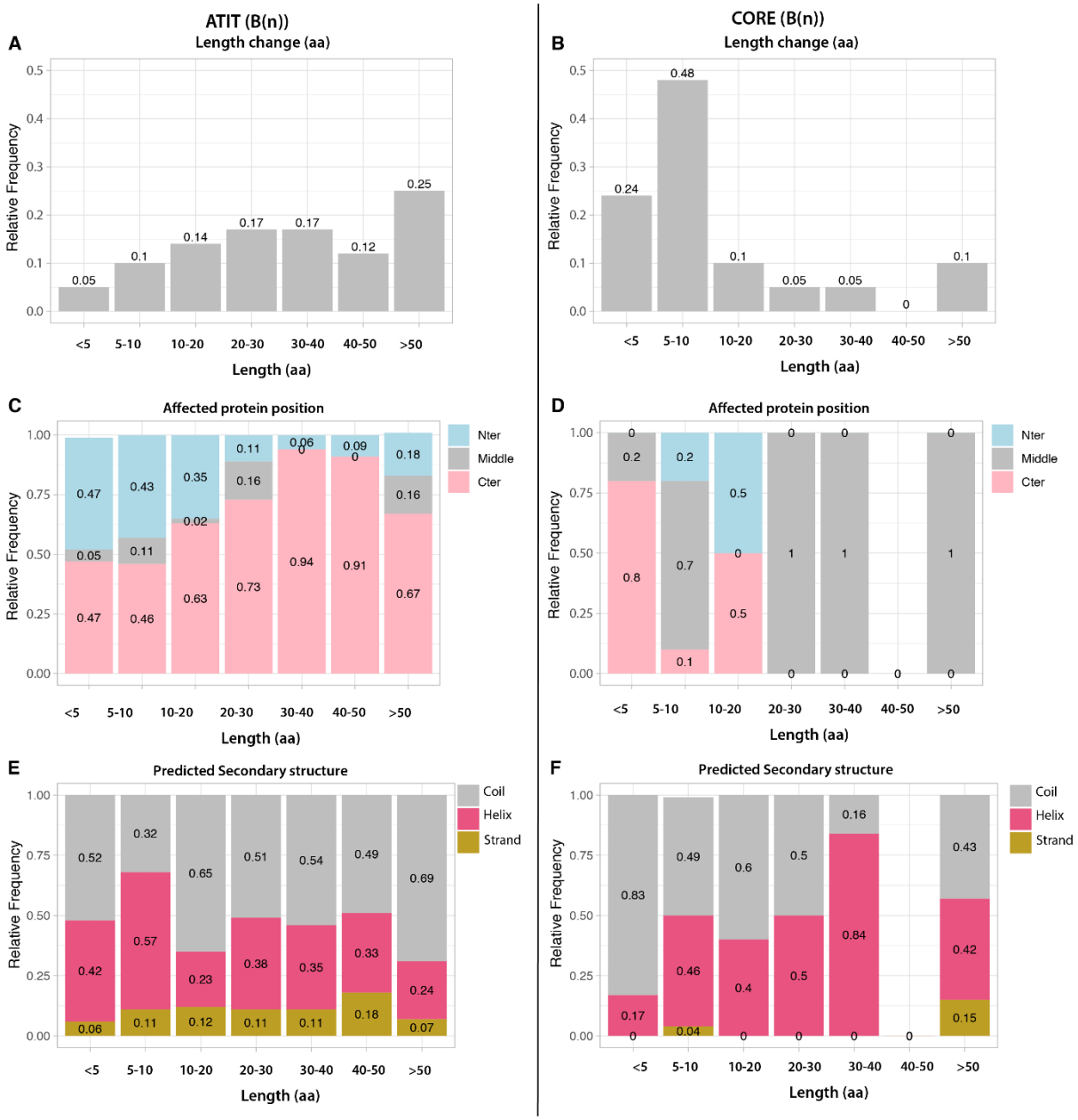


Figure 4.9: Impact assessment of B(c) splice site indels. The left column depicts the ATIT region, and the right column is the CORE region. Panels A and B show the frequency distribution of absolute indel length change (from their RISO form). Panel C and D enlist the region of protein affected by their introduction where N-Ter is the first 30% of the region, the middle is 30-70%, and C-Ter is changing in the end 70% to complete C-Ter of the transcript region; the bars depict the length range bins defined in panel A and B and range is also shown

in the X axis. Similarly, panels E and F show the affected secondary structure content for those defined length range bins by their introduction.

4.3.2.5 Sequence Integrity of unchanged fraction, anchoring the indel

Collectively from impact assessment of 'n'/'c'/'b' subtypes in protein-coding isoforms, we realized existence of length disparity-based cohorts in almost all splice site cases discussed, where often smaller indels are more prevalent in the CORE than in the ATIT. For less frequent longer lengths, pronounced population affects the N and C-Ter regions. Whether such longer indel lengths have been localized on termini for reasons other than increased length is not known, and to gain further details, we compared the sequence identity of the region, which has been maintained in pairs (non-indel) to evaluate whether indel contribution is indeed local or they introduce non-3n 'nt' indel that may disrupt the reading frame of the exon. Figure 4.10 shows the sequence identify frequency distribution of pairs in the range from 0-100 with a bin width of 25. The highly populated 'n' splice site variation maintains 100% identity for 81% cases in the ATIT region. This increases to 88% in the CORE region indicating 'n' splice site variation introduce indels, however, maintains the sequence mostly identical in the unchanged region.

Contrary to 'n' site changes, 26% of 'c' splice site changes pairs localization to 0-25% identity bin, indicating different protein segments introduced in ATIT. However, as 'c' splice site change does preponderantly affect the C-Ter of exon, it should ideally not have influenced the non-variable region, and likely these events may result from other non-3n exon skipping events before their occurrence. Contrary to ATIT, 93% of CORE cases maintain their identity 100% for non-variable regions (Figure 4.10C and 4.10D). The 'b' cases possibly have the most diverging impacts, wherein ATIT, 78% of total pairs have <25% identity to their parent form in the RSOEx. Such introduction of variability may explain their very least frequent population among all 'n'/'c'/'b' cases. Additionally, for 'c' site splice cases in ATIT and 'b' splice site cases, literature correlation of them as likely responsible drivers for substitution in the C-Ter region can be speculated. (Tress et al. 2007).

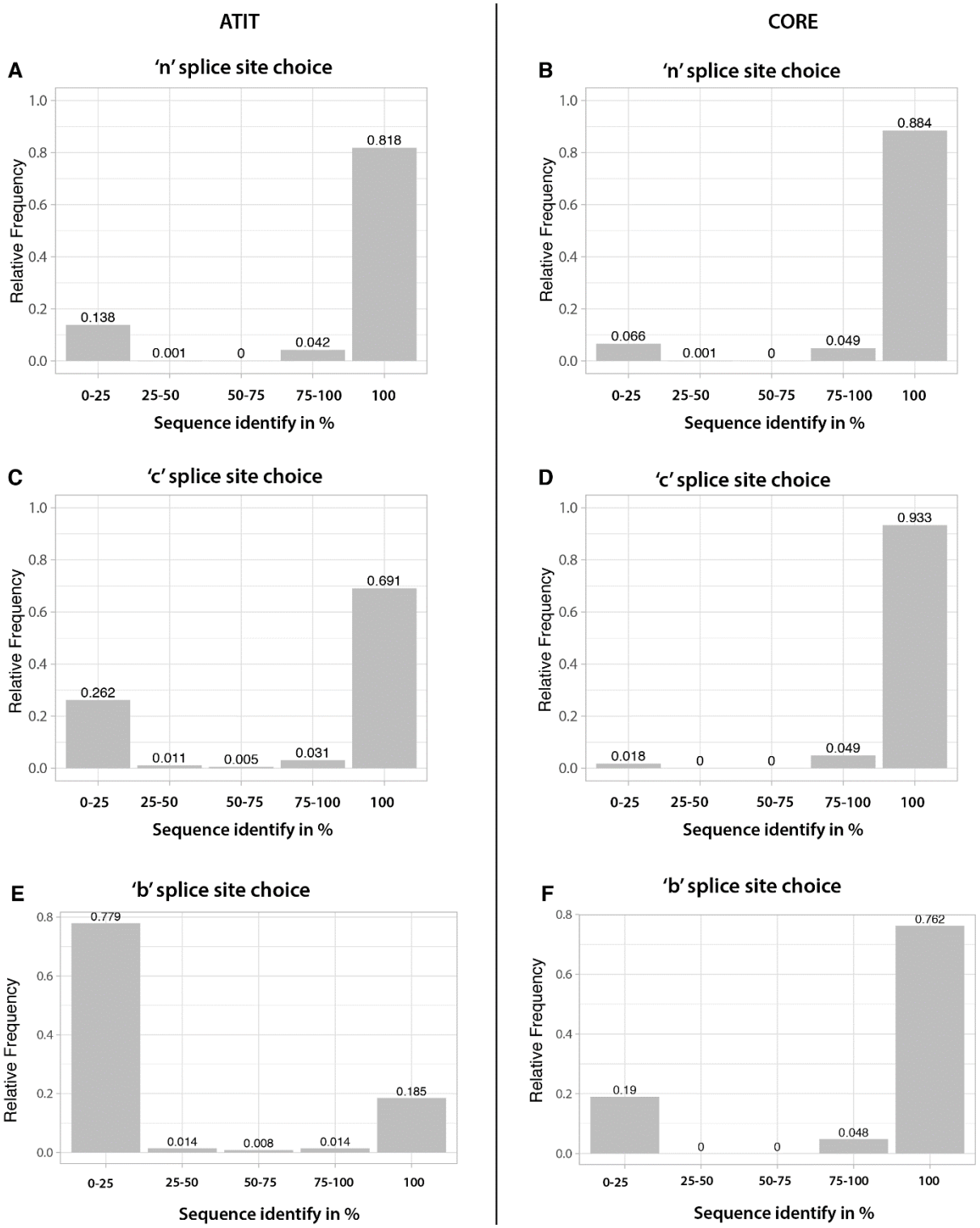


Figure 4.10: Sequence identity changes in the overlapping region of n/c/b variations. Histogram showing sequence identity computed for the overlapping region in the exon variation of n/c/b cases. The sequence identity is computed with respect to exon in RSOEx set. The three rows depicts the distribution of sequence identity with the 'n' variation shown in the

Top row, the ‘c’ shown in the middle, and ‘b’ shown in the bottom row. The ATIT region and CORE regions are shown in the left and right column respectively.

4.3.3 Correspondence of domain and exon boundaries in the ATIT and CORE regions

In the last section, we examined impacts of indels associated with A(ss) exons on proteins and compared them in respective ATIT and CORE regions. Building on these insights, we next assessed ‘A’ exons (without splice site variations) for their contribution to domains, especially given their significant variability in ATIT region (section 4.3.1.2). The influence of gene architecture and alternate/constitutive exons on domains is paramount to understand the functional variability, which can be included/excluded by ATIT regions within different transcripts of the genes. Previous studies indicated that CORE/CDS region is significantly enriched with domains than the ATIT and the gray region (Shabalina et al. 2014). The work of Buljian et al. also reported the details of domain trade-off and contrast preferences, by comparing constitutive and tissue-specific cassette with other cassette exons (Buljan et al. 2012). Despite their analysis, the domain composition of the ATIT regions remained unclear and sparsely studied. Considering ATIT contributes 1/3rd of the transcript region and were previously observed to increase gradually with ‘Dual’ exons from fish to humans (Chapter 3), we assessed their contributions to the protein sequence domains. We also compared ATIT domain composition with CORE region along with their exon subtypes.

4.3.3.1 Dataset and domain assignment

Continuing analysis from 2X2 dataset, we predicted Pfam domains with an e-value of ≤0.01 and model length ≥70% of domains. Out of 13063 genes in the 2X2 dataset of human genome, we could reliably associate at least one Pfam domain to the RISO of 11947 genes. In terms of amino acid coverage, domains covered 46.9% of RISO's total ‘aa’, which is 7,522,032. We excluded genes lacking any coding consecutive exons as it is important to differentiate ATIT and CORE region, thus, resulted in 10,808 genes in the final dataset for the present analysis. Table 4.6 summarizes gene prevalence in the ATIT and CORE regions with exon subtypes. We observed similar relative gene frequency of these regions as has been seen previously in Table 4.1. The coding alternate exons in ATIT and CORE regions are prevalent in 86% and 41%, respectively. Constitutive exons in ATIT (DG) are prevalent in 19% of genes.

Table 4.6: Gene prevalence for 10808 genes and their contribution from ATIT and CORE region with sub exons.

Region	Exon Type	Gene count	Gene Fraction
ATIT	A	9287	0.859
	G	2064	0.191
CORE	A	4418	0.409
	G	10808	1

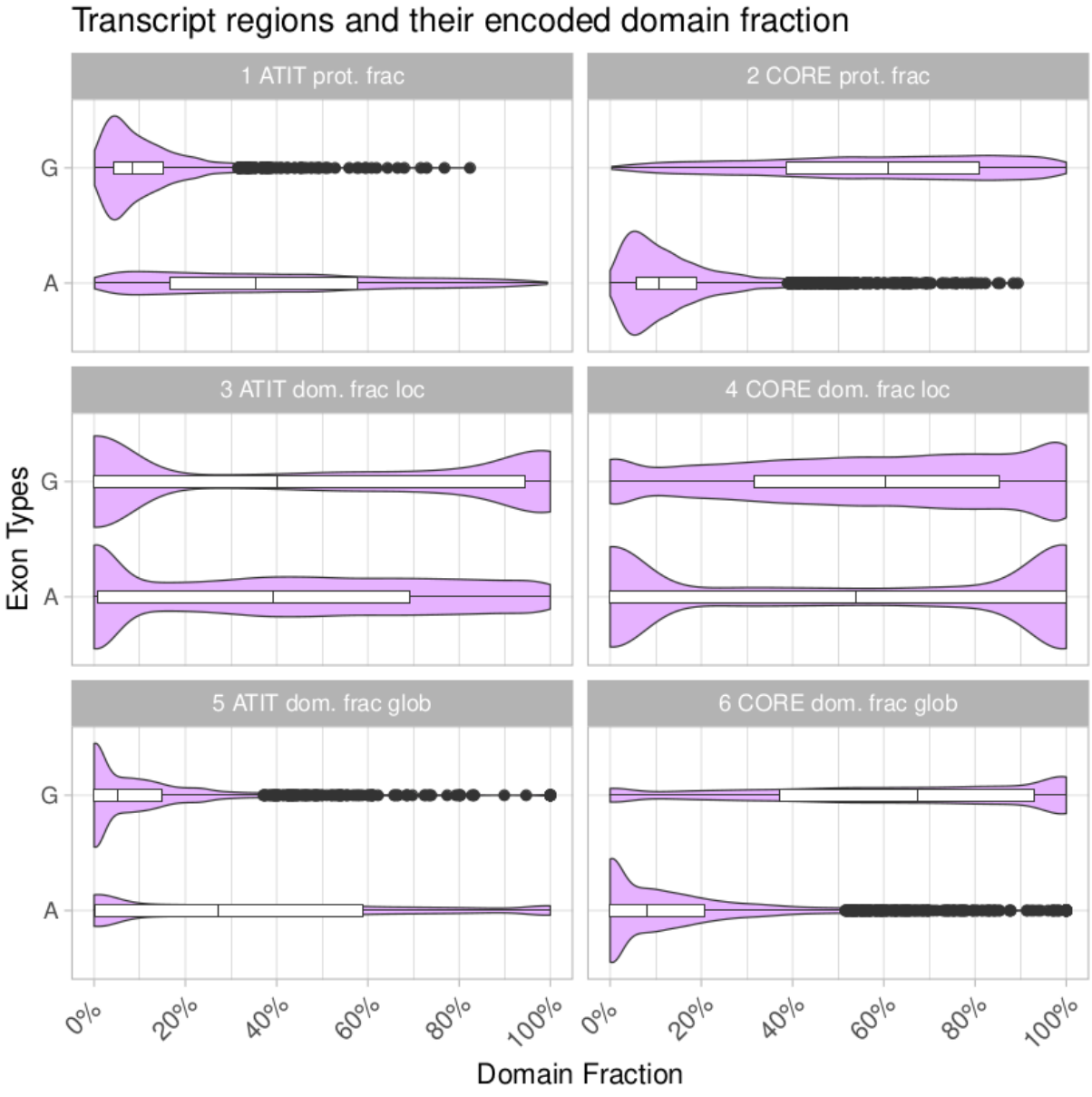


Figure 4.11: Domain and protein fraction encoded by ATIT and CORE regions and their subtype A and G exons. ATIT fraction is listed on the left and CORE on the right. The first row depicts the overall protein region contributed by such exons, and fraction distribution is normalized 'aa' count in those sub-regions by RISO protein length. The second/middle row depicts the local domain fraction per residue assigned to those regions, and the third row shows the cases akin to the second row; however, the complete domain length summation of the entire RISO was used for normalization instead of the local domain assignment normalization.

4.3.3.2 Protein fraction contribution of ATIT and CORE region

The fraction contribution of ATIT and CORE regions has been previously evaluated in Figure 4.2A, and the present analysis differs only in number of genes. Briefly, the first row of Figure 4.11 ('1 ATIT prot. frac' and '2 CORE prot. frac' panels) suggests that alternate exons in the ATIT region and constitutive exons in the CORE region contribute majorly to the protein. These exons are prevalent in 86% and 100% of the considered genes. There is a minor contribution from coding G exons in the ATIT and coding A exons from CORE are also noted, as they affect close to 20% and 40% of genes in the dataset, respectively. This is consistent with previous observations (section 4.3.1.1).

4.3.3.3 Domain fraction contribution of ATIT and CORE regions

We assessed domain contribution of ATIT and CORE region per gene emphasizing the alternate and constitutive exons in these regions. We have considered two approaches to quantify ATIT/CORE contribution to domain as discussed below and the calculation of the same is described using a hypothetical example.

Consider a hypothetical isoform of 100 residues long having the first and last 20 residue constituting ATIT region and middle 60 residues are the CORE. If assigned length of domain encompasses 60 residues such that 20 is from ATIT and 40 is part of CORE region. We computed fraction of ATIT/CORE region part of domain and composition of domain from these regions. The two measures are discussed below:

a) Region contributions to domain:

We calculated the fraction contributions of the ATIT and CORE regions to domain based on how many residues from the region participate in domain. In the above example, for the ATIT region, the value will be 0.5 as 20 residues of domain are coming from 40 residues of this region. Similarly, the CORE region contributes 2/3 as 40 residues are assigned to domains out of 60 residues in the CORE. We computed distribution of the dataset using this score shown in

the Figure 4.11 (panels C and D) labeled as "ATIT dom. frac. Loc" and "4 CORE dom. frac. Loc" for different sub-exon types.

b) Domain fraction from ATIT/CORE region:

We can also compute the relative ATIT/CORE region composition of domains by normalizing the domain assigned in these regions by the domain length. The fraction contribution of the ATIT region to the domains will be 1/3 as 20 residues of domains are part of the ATIT region, whereas for the CORE region, it will be 2/3 as 40 domain residues are part of CORE region. We computed distribution of the dataset using this score shown in the Figures 4.11E and 4.11F with labeled as "ATIT dom. frac. glob" and "4 CORE dom. frac. glob" for different sub-exon types.

In the analysis of overall ATIT/CORE region contribution to domains, we observed a broad distribution for both A and G types of exons in the ATIT and CORE (Figures 4.11C and 4.11D) suggesting varying exon residue contribution from different genes. The IQR of A exons is 67 whereas the same of G exon is 95 and both have median value around 40 suggesting that distribution of later type exons is broader in comparison to the A exon. Considering the upper whisker of box plot, we can see that 25% of the 2064 genes in which G exons are present contribute more than 90% of their fraction towards the Pfam domains in the ATIT region. This is exceptional as G exon is infrequent in comparison to other exons (Table 4.1) and may indicate specific roles in specific gene populations. Similar comparison of A exons in ATIT shows that >70% their fraction is contributed by 25% of 9287 genes. When we compared CORE region, it also shows a similar broad distribution for A/G exons. The IQR(median) of A and G exons are 100%(53%) and ~54% (60%), respectively. The A exons tends to exhibit broader distribution even in CORE region, though the median value is higher suggesting alternate exons in CORE region are more frequently part of domain. Importantly, in both ATIT and CORE regions, the constitutive exons have higher contribution towards domain. It is also pertinent to note that ATIT region could be longer in comparison to assigned domain sequence, which can also result in relatively lower contribution of these regions to domains.

Important insights can be gained when we examine A/G exon fraction contribution to the domain by normalizing the assigned region of ATIT/CORE by the domain length (approach *b*). Previously, we have observed that the maximal length-diverging isoform with respect to RISO shows truncation in the ATIT region, while maintaining the core fraction (Figure 4.2B). Here, we are analyzing the scope of ATIT and CORE region in contributing the domains to the

protein (Figure 4.11E and 4.11F). The IQR of G exons in the ATIT region is 15% with median of 5%. The upper whisker lies at around 40%, suggesting that contribution of constitutive exons to domain in ATIT is limited. Similarly, we also observed that A exons in the CORE has IQR (median) of 20% (8%), and upper whisker is at ~50%, suggesting limited contribution of A exons in domain of CORE region. Interestingly, the 'A' exons in the ATIT region and G of CORE are show opposite distribution pattern from one another. The former is skewed to lower fraction (median 27%) contribution to domains and latter has large contribution to domains (median 67%). Both of these exons affect more than 85% of the genes and considering their non-outlier upper whiskers, it can be noticed that for the quarter of genes harboring population A in ATIT region can indeed encode for 60% and more of the entire protein's domain constituents. To our knowledge, such contribution from A exons is not reported in the literature for ATIT regions. This ability to control >60% of domain constituents can be explored further in a separate study as it needs elaborated details, especially in gene architecture evolution.

4.3.3.4 Domain and exon boundary relevance

The above analysis only indicates the contribution of the regions towards the domain fraction but not address the integrity of domains. We asked several question regarding domains including: a) How many domains are in ATIT/CORE regions? b) Whether they are contained (intact) within an exon or need multiple exons for their formation?, c) If contained, are they part of the A exons or G in ATIT and CORE regions, and what are their relative contributions?, and d) Domains, which are split in exons, Are these split in exons of same nature (A/G) as well as ATIT/CORE region *i.e* lie at junction?

To investigate above questions, we compared the domain coordinates with exon junctions and considered a domain contained in an exon if ≥90% of domain fraction is covered by an exon, otherwise it is classified as split domain. We have 26,219 predicted domain regions with model length ≥0.7 in 10808 genes. Figure 4.12 shows the relative fraction of total (contained + split) in ATIT, CORE, and their junction. It can be observed that of all the domains, 55% occur in the CORE region affecting 58% of genes, 22% occur in the ATIT region of the protein (28% of genes), and 21% occur in the junction of both ATIT and CORE region (~50% of the genes). The domain frequency of these regions with their prevalence in genes is mentioned in Table 4.7. In ATIT and CORE, more than half of domains in these regions undergo a split. The domain split was comparatively more apparent in the CORE region (64% domains in 5169 genes) than ATIT (59% in 2246 genes), which conversely also means that more domain

fraction is contained in the ATIT than that of the CORE region and to best of our knowledge has not been presented in the literature and indicates that it likely plays essential role in diversifying transcript functionality while contributing only 1/3rd of the RISO (Figure 4.2A).

Regarding the nature of the exons that contain such domains, out of 41% of domains lying in ATIT, 40% (1143 genes) has contribution from A exons, and 1% was contributed by DG exons (48 genes) as mentioned in Table 4.8. In the CORE region, out of 35% contained domains, 33% were the G exons (1890 genes), and only 2% were A (213 genes) (Table 4.8).

Table 4.7: Domain occurrence prevalence in ATIT/CORE regions and their intersection with exon boundaries.

Region	Domain type as contained and split	Domain Count (fraction in %)	Gene Count (fraction in %)
CORE	All	14522 (55.4%)	6311 (58.4%)
	Contained	5210 (20%)	1965 (18.2%)
	Split	9312 (35.5%)	5169 (47.8%)
CORE ATIT Junction	All	5715 (22%)	5353 (49.5%)
	Contained	0	0
	Split	5715 (22%)	5353 (49.5%)
ATIT	All	5982 (23%)	3085 (28.5%)
	Contained	2447 (9.3%)	1168 (10.8%)
	Split	3535 (13.5%)	2246 (20.8%)

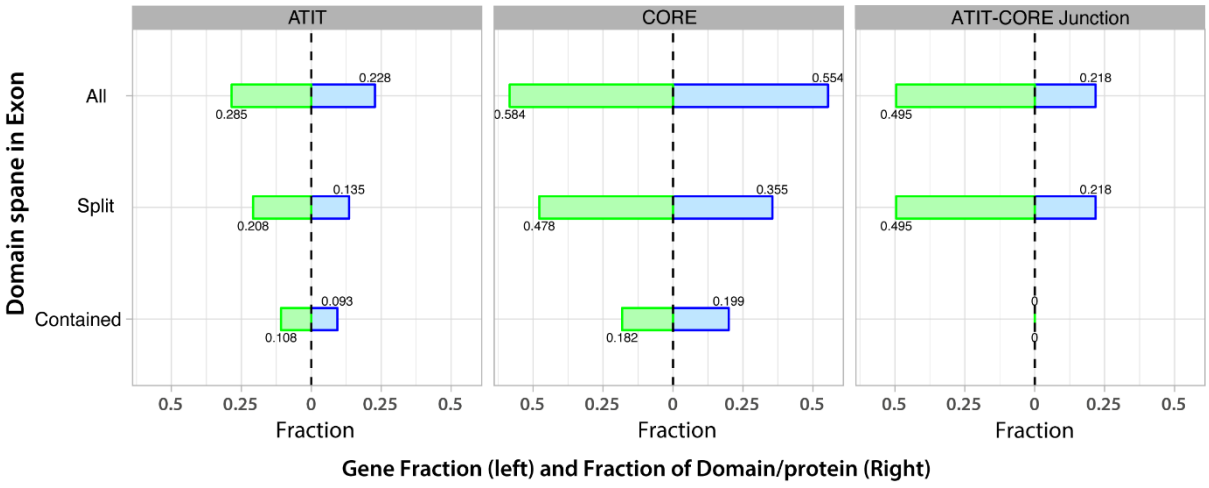


Figure 4.12: Assigned Pfam domain prevalence and overall distribution in ATIT, CORE, and their interface (ATIT + CORE junction). The 'All' row shows fraction of all domains (26219) without their subtype distinction. The second and third rows are 'Split' and 'Contained' and are subsets of 'All'. Domains will be classified as 'contained' when >90% of its region is covered by exon, otherwise split. The right bar (Blue colored) is the domain fraction, and the left bar is gene fraction (green). For domain fraction, the summation of split and contained in three cells equals 1 ('All' will sum to 1 independently).

4.3.3.4.1 Split and contained domains.

Having studies contribution of ATIT/CORE region in domains, we analyzed the domains that are not contained and need contribution from two or more exons and are considered as 'split' domains. Out of 64% of these cases in CORE affecting 47% of genes, 48% are split among constitutive exons in 40% of genes (Table 4.9) suggesting they are likely being maintained if their constitutive exon is not disrupted by insertion of alternative exons and do not undergo a change in reading frame because of non-3n nt driven splicing event in the vicinity. The remaining 15% domains were split among AG exons (20% genes) and 0.5% among A exons (0.8% genes). In the ATIT region having a total of 59% encoded domains undergoes split with 53% fraction is split among the alternate exons (Table 4.9). This raises questions about the integrity of such domains and also how possibly inclusion of such exons is regulated in the transcript. The remaining 4.3% of domains were split among A and G exons and 1% among G exons. The fraction of domains that undergo split among A exons, and AG exons in the ATIT region and CORE region, are close to half of the original domain count assigned and affect >50% of genes.

Table 4.8: Contained domain prevalence in ATIT and CORE region with the exon sub types

Region	Type	Contained domains (fraction in %)	Gene count (fraction in %)
ATIT	A	2400 (9%)	1143 (10.4%)
	G	64 (0.2%)	48 (0.4%)
CORE	A	300 (1.1%)	213 (1.9%)
	G	4933 (18.5%)	1890 (17.2%)

To analyze the integrity of such domains undergoing split cases between A and G exon types, we checked their fraction contribution of A and G exon subtypes for their occurrence in RISO. First, we analyzed the subsets undergoing split into CORE and ATIT regions, where 2203 domains in 1945 genes were affected in CORE region and 259 domains in 247 genes in ATIT region (Table 4.9). Their distribution comparison (Figure 4.13) shows that AG in CORE and AG of ATIT follow trends on similar lines to their domain fraction region contribution (Figure 4.11). The contribution of G exons in CORE and A exons in ATIT are noteworthy, and they contribute >60% domain fraction, indicating difficulty maintaining domain integrity in the ATIT region than in CORE.

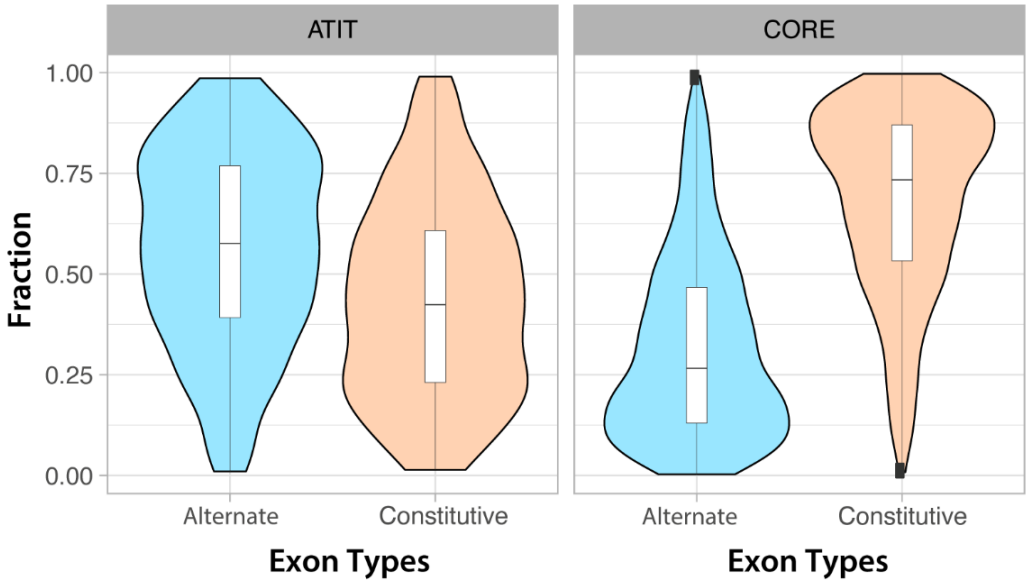


Figure 4.13: Fraction contribution of A and G subtypes when domains are split among these exons in ATIT and CORE region. The box plot shows the fraction contribution of alternate and constitutive exons to split domains lying in the ATIT and CORE regions.

In the split instance between ATIT and CORE junction (Table 4.9), the majority of domains (3990 domains in 3786 genes) span 'A' Exon in ATIT and 'G' Exon in CORE (Table 4.9, junction, 1CORE_G__2ATIT_A'). The distribution of their fraction contribution for each domain from A exon in ATIT and G exon in the core is shown in Figure 4.14 (leftmost cell in the bottom row, where aA is A exon of ATIT and cG is G exon of CORE). We observed that their medians are centered near 0.5, where G exon has a median slightly above and that of aA. The second large fraction overlaps with A exon in ATIT region and CORE region's G and A exons with junction heading '1CORE_AG__2ATIT_A'. For the A exons in CORE, we have previously seen their broad distribution in Figure 4.11 in the context of contributing to domains.

For their domain contribution in 759 domains of 753 genes, they only contribute a median of ~15%, indicating either A exons in CORE either lack domains or contribute to small embellishments to the domain. Most domains in this junction have their contributions from G of CORE and A of ATIT, where the former contributes large fraction to domain region. Above two categories (1CORE_G__2ATIT_A', ''1CORE_AG__2ATIT_A') contribute to more than ⅘th of cases when domains undergo a split between the CORE and ATIT interface. Additional categories with minor domains undergoing split were segregated to CORE G and ATIT's A and G; with junction headings, '1CORE_G__2ATIT_AG', '1CORE_AG__2ATIT_G', '1CORE_AG__2ATIT_AG'. In all those categories, G exons of CORE contribute the most domain, followed by A exon of ATIT and a minor fraction contribution of G in ATIT and A in CORE exons (Figure 4.14).

Table 4.9: Summary for split domains classified based on the region shared with A/G exon subtypes with their fraction occurrences in gene/domain.

Region	Junction*	Domains (fraction in %)	Genes (fraction in %)
CORE	A	86 (0.5%)	77 (0.8%)
CORE	AG	2203 (11.9%)	1945 (20%)
CORE	G	7023 (37.8%)	3959 (41%)
ATIT	A	3205 (17.3%)	2031 (21%)
ATIT	AG	259 (1.4%)	247 (2.5%)
ATIT	G	71 (0.4%)	58 (0.6%)
CORE ATIT Junction	1CORE_G__2ATIT_G	236 (1.3%)	236 (2.4%)
CORE ATIT Junction	1CORE_AG__2ATIT_A	759 (4.1%)	753 (7.8%)
CORE ATIT Junction	1CORE_G__2ATIT_AG	542 (2.9%)	540 (5.6%)
CORE ATIT Junction	1CORE_AG__2ATIT_G	62 (0.3%)	62 (0.6%)
CORE ATIT Junction	1CORE_G__2ATIT_A	3990 (21.5%)	3786 (39%)
CORE ATIT Junction	1CORE_AG__2ATIT_AG	126 (0.7%)	126 (1.3%)

*Junction types: 1) '1CORE_G__2ATIT_G': Domain spans the constitutive (G) exon in CORE and constitutive (G) exon in ATIT; 2) '1CORE_AG__2ATIT_A': Domain spans the constitutive (G) and alternate (A) exon in CORE and alternate (A) exon in ATIT; 3) '1CORE_G__2ATIT_AG': Domain spans the constitutive (G) exon in CORE and constitutive (G), alternate (A) exon in ATIT; 4) '1CORE_AG__2ATIT_G': Domain spans the constitutive

(G) and alternate (A) exon in CORE and constitutive (G) exon in ATIT; 5) ‘1CORE_G__2ATIT_A’: Domain spans the constitutive (G) exon in CORE and alternate (A) exon in ATIT; 6) ‘1CORE_AG__2ATIT_AG’: Domain spans the constitutive (G) and alternate (A) exons in both CORE and in ATIT regions.

Interesting are the cases when domains split over the ATIT’G and CORE’s G exons with junction heading ‘1CORE_G__2ATIT_G’, having 236 Domains from 236 genes (bottom right cell of Figure 4.14). This is the only category where G or DG exons of ATIT span the full spectrum of domains fraction with a median fraction of 25% but with comparatively higher IQR and third quartile higher than 50% and agrees with our previously extreme domain fraction contribution observed in Figure 4.11 for DG exons. Though the contribution of CORE G exons is still higher, a complimentary skewed distribution can be noted for the DG exons and the negatively skewed CORE G exon’s contribution.

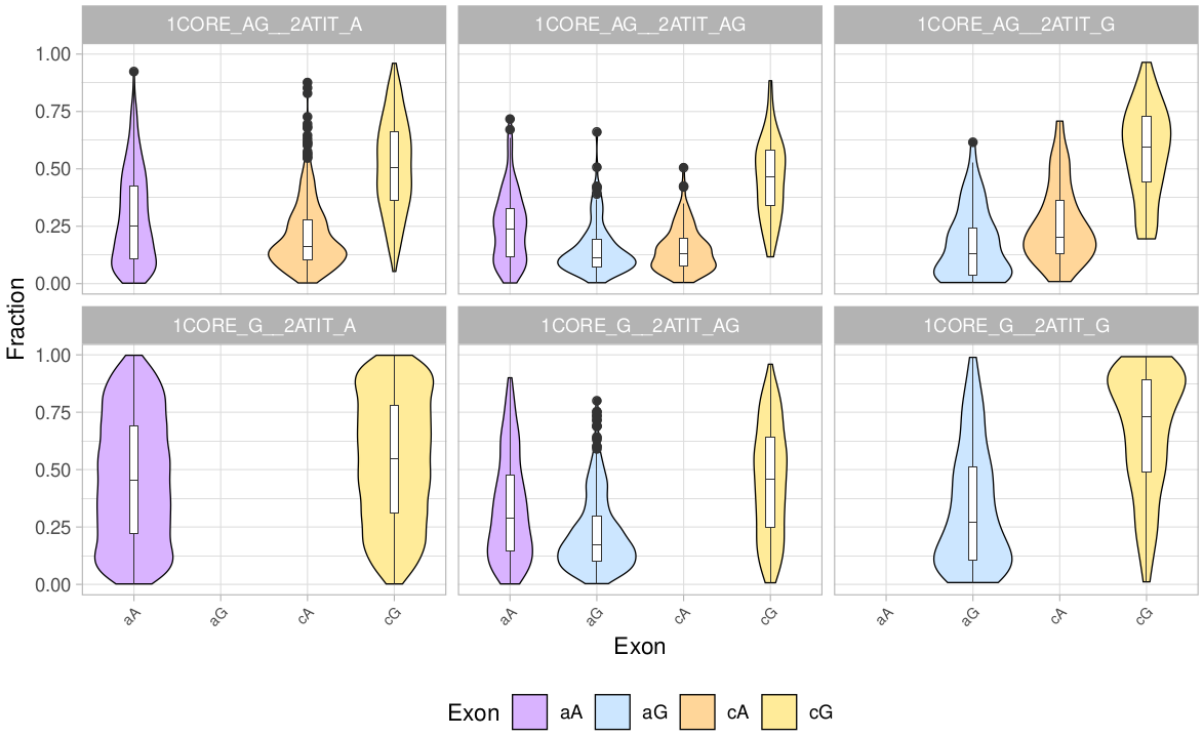


Figure 4.14: Domain fraction contribution from A and G subtypes of exons in the ATIT and CORE region junction. The box plot showing contribution of various subtypes of exon to various categories of split domains lying the junction. The categories of domains are discussed in Table 4.9 and are mentioned on top of each panel. The ‘aA’ is alternate exon in ATIT, ‘aG’ is constitutive exon in ATIT, ‘cA’ is ‘A’/Alternate exon in CORE and ‘cG’ is ‘G’/constitutive exon in CORE). The 6 cells' subtype architecture of exons, when domains are split among ATIT/CORE junction, their individual domain distribution per subtype of exons, are shown respectively in those cells.

4.4 CONCLUSIONS

In the present chapter we have used ENACT exon nomenclature to study the ATIT and AS driven variation in the protein variation in human genome. The analysis exemplifies the usage of the previously developed framework (chapter 1) and how that can be proposed to detail insights regarding the extent of splicing-driven and alternative transcription-driven changes. Our observations herein matched that of the previous (Shabalina et al. 2014) but extend further and detail the scarcely discussed in their report regarding domain prevalence in ATIT and its comparison with the CORE region. The main observations from the present studies are that ATIT region encodes 1/3rd of coding region in RISE and 2/3rd is encoded by CORE region. The comparison of RISO with Maximally length divergent isoform among RefSeq listed transcripts showed that ATIT region relatively undergoes truncation or extensive expansion in comparison to CORE region, which does not show these length variations and probable mechanisms of such changes are suggested by contrasting inclusion rate differences in their frequency between CORE and ATIT.

After studying the extent of changes in ATIT and CORE region, we analyzed in detail the impact of indel of alternate splice choices on protein isoforms. We compared their relative frequencies and impact assessment criteria for ATIT and CORE regions. We observed at least twice the reduction in the ‘n’/’c’/’b’ cases in the CORE than in the ATIT and realized they often affect the coiled residues, and trend increases for longer indels. Moreover, we have also observed that the region in CORE harbors comparatively smaller indels than that of ATIT. Regarding their scope of impact, the ‘n’ site unchanged fraction is relatively identical, with ‘c’ site changes in ATIT changing the sequence of unchanged fraction 25% of the time. The ‘b’ site changes impacting protein contribute only a minor fraction compared to ‘n’ and ‘c’ splice site cases but were most drastic, where only a handful of cases share considerable identity in region overlapping between pairs and more than ⅔ have <25% identity.

Subsequently, we analyzed the domain contributions from ATIT and CORE region. Interestingly, ATIT core region consists of relatively more contained domains in comparison to CORE domains suggesting that these region brings whole domains. However, it does have domains split between two alternate exons and it will be interesting to understand how these exons are regulated to be spliced in/out during splicing. Barring some handful of cases where the domain gets split among alternate-alternate (AA) exon junctions in ATIT, many split domains have considerably higher contributions from the G exons. Even for the domains,

which have undergone split in AA exon junction, a detailed analysis of their inclusion frequency and possible association and selection pressure co-occur together as have been discussed for protein units (Gelly et al. 2012) can be detailed in future studies.

Chapter 5

Optimal protein sequence design mitigates mechanical failure in silk β-sheet nanocrystal

"Reprinted (adapted) with permission from [Verma, P., Panda, B., Singh, K. P., & Pandit, S. B. (2021). Optimal protein sequence design mitigates mechanical failure in silk β-sheet nanocrystals. ACS Biomaterials Science & Engineering, 7(7), 3156-3165.].(Verma et al. 2021) Copyright [2021]. American Chemical Society". License details have been added at the end of this document.

5.1 INTRODUCTION

The silk fiber is a rare biomaterial that possesses exceptional mechanical properties such as high tensile strength with elasticity in tensile loading (Porter, Guan, and Vollrath 2013), torsional super-elasticity in cyclic loading (Kumar et al. 2013; Liu et al. 2017). It is also anti-bacterial, biodegradable, and anti-inflammatory making it suitable for green chemistry and materiomics (Römer and Scheibel 2008; Rising 2014; Brown et al. 2015). The extraordinary mechanical strength of silk fiber is mostly attributed to the hierarchical arrangement of laminated antiparallel β-sheet nanocrystals embedded in an amorphous matrix composed of predominantly short stretches of regular or non-regular secondary structures (Römer and Scheibel 2008). These two regions are known to have separate roles in response to external forces, where amorphous regions are known to provide elasticity and crystalline regions the ultimate strength (Termonia 1994; Work 1985). In addition, amorphous regions are also known to characterize the silkworm and spider dragline silks, as the former harbors tyrosine rich domains and have a comparatively small fraction of intramolecular β sheets (Du et al. 2011; Numata et al. 2015). This compromises the strain hardening feature in silkworm silk when subjected to stress/strain (Du et al. 2011; Numata et al. 2015). Nevertheless, the ultimate tensile strength is mostly governed by the well-conserved nanocrystalline antiparallel β sheets having sequence repeats of pAla/pAlaGly in both silkworm and spider dragline silks (Eisoldt, Smith, and Scheibel 2011; Römer and Scheibel 2008). The layered antiparallel β-sheet arrangements, besides silk fiber, are known occur in other fibrous or globular proteins having a role in the

mechanical function of the cell (Lu et al. 1998; Gao et al. 2003; Forman et al. 2005). Numerous previous studies have elucidated that the hierarchical arrangement of β-sheets in spider silk can resist forces up to ~3 nN (Keten et al. 2010). Other similar structures require lower yet comparable rupture forces, such as cross β-sheet structures of amyloid fibrils (Ndlovu et al. 2013; Ndlovu et al. 2012) and β-sheet arrangement found at protein-protein interface of bacterial adhesion (SdrG) and human fibrinogen proteins (Milles et al. 2018). There have been several attempts to mimic the molecular assembly of silk protein on an industrial scale, yet the performance of natural silk is not fully mimicked owing to a lack of a complete understanding of design principles relating silk sequence to their structure-property (Kluge et al. 2008; Blamires, Blackledge, and Tso 2017).

The β-sheet nanocrystal, the building block of the silk nanocrystalline region, has been an attractive system to understand and leverage Nature's design principles in constructing biomaterial with unusual mechanical properties. Consequently, there have been numerous computational simulation studies, mostly using Molecular Dynamics (MD) and Steered Molecular Dynamics (SMD), to decipher the mechanical behavior and fracture mechanism of nanocrystalline at the atomistic level (Yarger, Cherry, and Van Der Vaart 2018). These studies have revealed that the size of β-sheet nanocrystal confined to 2-4 nm having an optimal number of residues in a strand (4-8 amino acids) achieves higher strength and toughness than other large-sized structures (Keten et al. 2010; Bratzel and Buehler 2012; Buehler and Yung 2010). These studies also elucidated the importance of hydrogen bonds cooperativity in determining ultimate tensile strength and dissipative stick-slip mechanism (Buehler and Yung 2010; Keten et al. 2010). Their contributions towards the tensile strength were also evident from another studies and are function of a) hydration levels, where increasing hydration reduces the fracture point due to completion between solvent and intra-protein hydrogen bonds and hence compromise of latter leading to reduced hydrogen bond interaction energies (Cheng et al. 2014) b) and their ability to regenerate after yield fracture slip which can compromise stick slip mechanisms and hence affect overall toughness properties (Kim, Choi, et al. 2020). In addition to effects of hydrogen bonds ultimate pullout forces and tensile strength could also affected by a) side chain interactions, where switching them off, and dampening van der Waals (vdW) interaction potential reduces the ultimate pull force by 28% (Xiao et al. 2009) b) orientation/arrangement of β-strands and their solvent exposure, c) loading rate and device stiffness in pullout simulations, and d) pulling direction during strand pullout experiments. (Brockwell et al. 2003; Jahn et al. 2010; Cheng et al. 2014; Xu et al. 2015). In addition to the

above studies, in-silico pull-out experiments on the naturally occurring motifs $(GAGAGA)_n$/$(GAGAGS)_n$/$(GAGAGY)_n$ of silk fibroins showed that among these $(GAGAGA)_n$ has the maximum tensile elastic modulus (Kim et al. 2018).

Despite the above previous studies, the nanocrystal mechanical properties dependence on its β-sheet sequence composition has largely remained unexplored. To address this, we computationally investigated the effect on nanomechanical properties of nanocrystals modelled for representative amino acid repeat sequences, which formed extensive side chain interactions (hydrophobic/hydrogen bond). We modelled structures with enhanced side chain interactions as these have valuable contributions to increase tensile strength (Xiao et al. 2009). Moreover, we can also assess whether the β-sheet sequence occurring in nanocrystalline region of natural silk crystal is optimized for nanomechanical features. In materiomics spirit, these will provide a holistic understanding of natural/synthetic materials for designing potentially new material constructs (Buehler and Yung 2010; Cranford et al. 2013).

In the present study, we have performed *in silico* experiments to compare the mechanical properties of β-sheet nanocrystals composed of naturally occurring amino acid sequences in silk with structures of several designed sequence constructs. For the latter, we modelled mainly β-strand favoring hydrophobic (Isoleucine/Valine) and polar (Threonine/Asparagine for its ability to make multiple side chain hydrogen bonds) amino acid repeat sequences on the known topology of silk β-sheet to mimic their sequence and structure.

The multiple SMD pull-out simulations showed that Alanine homopolymer, one of the naturally occurring sequence motifs in silk, has the highest mechanical strength and toughness among modelled structures consisting of other amino acid repeats. Further detailed analysis of pull dynamics to understand the effect of side chain interactions illustrated that they effectively reduce rupture force and alter molecular stick-slip dynamics. Thus, providing insights into sequence dependent nanomechanical failure of silk β-sheet.

5.2 **MATERIALS AND METHODS**

5.2.1 Molecular modeling of nanocrystals

We modelled nanocrystal structures of various amino acid homopolymers based on the theoretical model structure of silk poly-(Ala-Gly) (accession code: ma-cs24y from ModelArchive, old pdb id: 2slk) (Fossey et al. 1991). In modelling homopolymer structures,

first, we extended the number of strands in the β-sheet from five to seven strands by adding requisite additional β-strands while maintaining antiparallel β-strand orientation, overall topology, and distance geometries as observed in the initial structure. Thus, obtained seven-stranded β-sheet nanocrystal structure was used as a template to model other poly-amino acid sequences. Figure 5.1A shows a representative structure of the nanocrystal. Next, we appropriately mutated residues in the template to Alanine, Glycine, Threonine, Asparagine, Isoleucine, or Valine to model layered β-sheet structures for various poly-amino acids. The optimal rotamer of the mutated residue was chosen using *swapaa* routine in UCSF Chimera (Pettersen et al. 2004). Although we selected rotamer based on the lowest clash score, we found a number of atomic steric clashes in the poly-polar/hydrophobic amino acid structures, especially between side chain atoms of β-sheet layers. To minimize these clashes, we moved the upper and/or lower layers away from the middle layer to an optimal inter-sheet distance with reduced steric clashes. To quantify this inter-atomic clash, we defined a score (clash-sc) given by the equation (1):

$$Clash - sc = \sum_{i=1}^{m} \sum_{j=1}^{n} \begin{cases} 1 \; if \; dist(i,j) < (rad(i) + rad(j) \; - k_{contact}) \\ 0, otherwise \end{cases} \qquad (1)$$

where, *m* and *n* are the total number of atoms in two β-sheets, and *dist (i, j)* is the distance between *i* and *j* atoms. The *rad (i) and rad (j) are* van der Waals (vdW) radii of atoms *i* and *j*, respectively. The $k_{contact}$ is the relaxation of distance cut-off to allow for the closest approach of atoms, which we set to 0.5 Å.

The optimal modelled structure of a designed sequence constructs was obtained by iteratively increasing the inter-sheet distance in a step size of 0.1 Å until the clash score was reduced to a value of one. Finally, we performed energy minimization of distance optimized modelled structure using CHARMM force field (Best et al. 2012; Huang and MacKerell 2013) in NAMD (Phillips, Braun, Wang, Gumbart, Tajkhorshid, Villa, Chipot, Skeel, Kalé, et al. 2005) program.

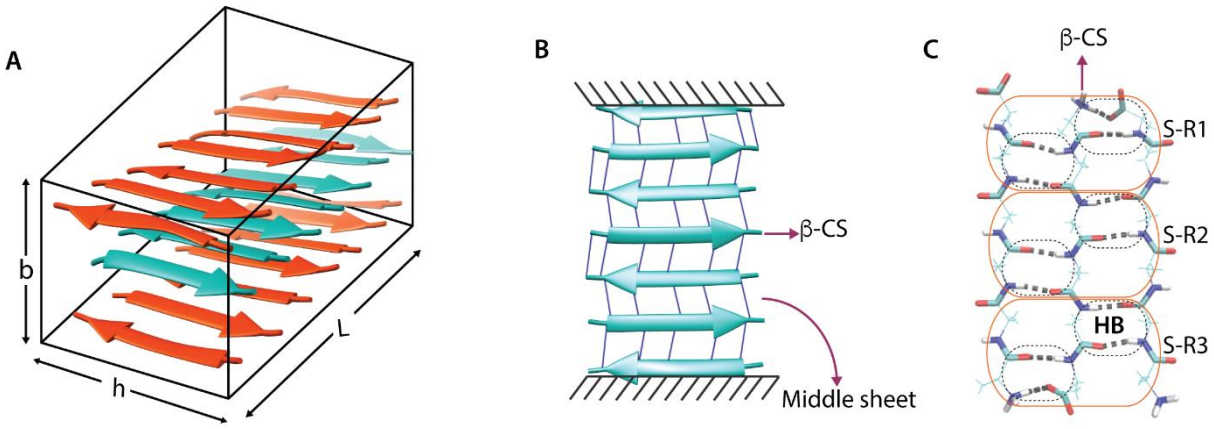


Figure 5.1: Overview of β-sheet nanocrystals shape and hydrogen bonds of the middle layer. A) Representative pAla nanocrystal model shows geometrical parameters to define their shape. The *h, L, and b* parameters correspond to the number of residues in a β-strand, the number of such strands in a β-sheet, and the number of such sheets respectively. B) SMD boundary condition of the middle sheet and pull direction of central β-strand, which is shown in cartoon representation. Thin blue lines between strands represent hydrogen bonds. C) A cartoon representation shows hydrogen bond (HB) rings and super-ring in an ideal antiparallel β-stand arrangement. The black dotted rectangle encloses the HB ring and orange rectangle surrounds the super ring, which is composed of two consecutive HB ring.

5.2.2 Molecular Dynamics simulation

We performed explicit water MD and SMD simulations for modelled nanocrystal polymorphs utilizing CHARMM36 topology and force field parameters for proteins having CMAP corrections (Huang and MacKerell 2013) using NAMD (v2.11) program (Phillips, Braun, Wang, Gumbart, Tajkhorshid, Villa, Chipot, Skeel, Kalé, et al. 2005). The energy minimized modelled structures were first solvated using TIP3P solvent water in a cuboid box such that solvent forms 30 Å thickness around the protein structure. The solvated system was energy minimized for 10,000 steps followed by ~400 ps of temperature equilibration at 300 K and pressure equilibration of ~1 ns or until average pressure reaches to ~1 bar (1 atm). During temperature equilibration, we gradually increased temperature in a step of 30 K/20 ps followed by equilibration for ~200 ps at 300 K. The system was simulated in periodic boundary conditions with electrostatics interactions computed using the Particle Mesh Ewald (PME) (Darden, York, and Pedersen 1993) method by specifying grid sizes. The vdW interaction involved switching functions with a cut-off distance of 12 Å (staring at 10 Å). The constant pressure of ~1 atm was maintained using the Nosé-Hoover Langevin piston method with a piston period of 100 fs, a damping timescale of 50 fs, and piston temperature of 300 K. A

constant temperature of 300 K was maintained using the Langevin dynamics, with the damping coefficient set to 5 ps^{-1} for all the heavy atoms. The time steps of 2 fs and 1 fs were used for the production and SMD simulations, respectively. We found the conservation of total energy in equilibration and production runs. The visual inspection of trajectories from the production run did not show protein unfolding events, except some loss of secondary structures in edge β-strands.

After pressure equilibration for ~1 ns, we divided simulation in the following two parallel systems: a) Constant velocity SMD pulling of central β-strand (Figure 5.1B) with a spring constant k = 10 kcal $mol^{-1}Å^{-2}$ and a displacement rate of $\dot{x}$ = 0.005Å ps^{-1} under a constant thermostat (pulling run); and b) production simulation under NPT condition for a similar time as SMD pull (production run). As a boundary condition, we positionally fixed two β-strands lying at both edges of the top and bottom β-sheet layers, and only one strand at both edges of the middle sheet layer during SMD simulations. We performed three replicates of SMD and production simulations for every poly-amino acid model that varies in their initial assigned velocities. The snapshots after every 0.5 ps were stored in trajectory for analysis.

5.2.3 Trajectory and hydrogen bond analysis

The solvent was removed from trajectory frames to facilitate analysis. We visually examined the SMD simulations to analyze stick-slip dynamics, twisting of central β-strand (β-CS), deformation of β-sheet assembly, and open/close of β-sheet layers. The pull force and corresponding displacement were calculated using the log file. We followed bin averaging approach to smoothened force-displacement profiles for their analysis. The force magnitudes were averaged over a displacement bin width of 0.1 Å intervals (average profile). The simulation time between the first force peak (rupture of hydrogen bonds) to the next step where the force increases (reformation of hydrogen bond) was empirically defined as slip time. We used both visualization and force-displacement curves to define the force peaks and subsequent increase in force.

The hydrogen bonds were identified using the *HBonds* plugin in VMD (Humphrey, Dalke, and Schulten 1996b) that is based on the angle (30°) and distance (3.5 Å) criteria. Using *HBonds*, we identified the number of hydrogen bonds in the frame at the peak force, relying on the criterion as described before. Additionally, we eliminated hydrogen bonds with a carbon atom as donor/acceptor in characterizing main-chain or backbone hydrogen bonds. The

hydrogen bonds were classified into types/sub-types depending on whether the donor/acceptor atom is from the main-chain or side-chain. We identified 10-membered hydrogen bond (HB) ring, based on previously described nomenclature (Cheng, Pham, and Nowick 2013), as shown in Figure 5.1C. Further, we consider two such consecutive HB rings to define a super-ring unit (Figure 5.1C).

5.3 **RESULTS**

We compared mechanical properties of β-sheet nanocrystals composed of naturally occurring sequence motifs in silk fibroin with homopoly-polar/hydrophobic amino acids sequences to find the optimally designed sequence with the best tensile strength/toughness. Furthermore, we investigated hydrogen bond and stick-slip dynamics to understand strand molecular failure mechanism in various modelled constructs.

5.3.1 Homopolymer nanocrystal models

We constructed β-sheet nanocrystal models of homopolymers composed of representative amino acids to study their tensile mechanical strength, toughness, and rupture behavior. The structures of homopoly-amino acids sequences were modelled on the topology of a standard theoretical β-sheet model of Bombyx mori silk (accession code: ma-cs24y from ModelArchive, old pdb id: 2slk) (Fossey et al. 1991), which is usually employed for studying silk mechanical properties using MD simulations. Since nanomechanical deformation and failure mechanisms depend on the critical size of the nanocrystal, we modelled homopolymer structures of L ~2.83 nm having the highest tensile strength (Buehler and Yung 2010; Keten et al. 2010). For this, we built structure assembly for various sequence constructs composed of three-layered antiparallel β-sheet with each sheet having seven β-strands (Figure 5.1A) and each of these β-strands consist of six residues (see Methods).

In our study, we mainly chose representative amino acids favoring β-strands for designing sequence constructs instead of building homopolymers for all amino acids considering computational simulation cost and feasibility of forming energetically stable three-layered β-sheets. Additionally, we chose Asparagine for building a homopolymer model to enhance side chain hydrogen bond interactions in nanocrystals as its amide side chain group can donate/accept two hydrogen bonds. We did not consider charged amino acids, as they will cause electrostatic repulsion among sheets, and bulky rigid aromatic ring containing amino

acids, as they may disrupt the overall integrity of nanocrystal structures. The selected representative amino acids are classified based on their size and physicochemical properties into three broad categories: (i) Small amino acids (SAA/small-AA): poly-Alanine (pAla), poly-Alanine-Glycine (pAlaGly), poly-Glycine (pGly); (ii) Polar amino acids (PAA/polar-AA): poly-Threonine (pThr) and poly-Asparagine (pAsn); and (iii) large hydrophobic amino acids (HAA/hydrophobic-AA): poly-Valine (pVal) and poly-Isoleucine (pIle). The steps in the modeling of β-sheet nanocrystals involved constructing an initial distance optimized model by reducing atomic steric clashes followed by energy minimization using NAMD (see Methods). Further, distance optimized structure was equilibrated in explicit solvent at 300 K temperature and ~1 bar pressure (see Methods).

We visually compared various features of modelled structures both before and after equilibration steps. Table 5.1 summarizes 'b' parameter/breadth of nanocrystal distances and the topology of minimized and final equilibrated structures. The 'b' parameter of the model is defined as the average distance between Cα atoms of the top and bottom layers (Figure 5.1A). As expected, pGly (no side-chain) models show the minimum breadth ('b' parameter) distance, and the maximum is observed for pIle amino acid models due to their bulky branched side-chain. Subsequent to equilibration, the β-sheet layers of polar/hydrophobic nanocrystals showed a slight change in their relative orientation and form twisted strands (more apparent in hydrophobic models) (Figure 5.2) to accommodate side chains while maintaining packing in β-sheet. These are accompanied by a slight increase in the 'b' parameter of nanocrystals from minimized models (Table 5.1).

5.3.2 Geometry and side-chain packing analysis of poly-amino acid β-sheet structures

We examined the overall topology, β-sheet organization, and side-chain packing of various modelled nanocrystals. These were compared based on the following features: a. overall shape and b. Face-to-Face side-chain packing arrangement of the middle strand with top and bottom β-sheets. The representative modelled nanocrystals structures of homopolymers are shown in Figure 5.2 and their features are summarized in Table 5.1. As is evident from the figure (Figure 5.2A, 5.2B and 5.2C), small amino acids (pAla, pGly, and pAlaGly) and polar-AA models (Figure 5.2D and 5.2E) have a relatively flat cuboidal geometry as observed in their starting structure of ma-cs24y (ModelArchive). In contrast, hydrophobic-AA models show a slightly twisted layer (Figure 5.2F and 5.2G), which resembles a cross-layer topology and gives a

twisted cuboidal geometry along the plane of the β-sheet layer. It has been suggested that such a twisted β-sheet arrangement facilitates a better packing of bulky side chain groups (Cheng, Pham, and Nowick 2013). The side-chain packing arrangement was analyzed using the nomenclature described for β-sheet face-to-face (FToF) interactions in a layered β sandwich-like structures (Cheng, Pham, and Nowick 2013). The FToF interactions of the central β-stand (β-CS) in the middle layer were compared among various nanocrystal models. In general, the overall orientation of β-CS and its FToF interactions are similar except that there is a slight variation in orientation observed in polar/ hydrophobic-AA models (Figure 5.3) Among small-AA, pAla and pAlaGly models show side chain inter-digitation packing arrangement, as has been previously observed (Bratzel and Buehler 2012), where the side chain of an amino acid (j) of the strand (β-CS) is well-packed against two side chains of residues (i and i+2) from either the top or bottom β- sheets (Figure 5.3A).The pAlaGly model shows that two Glycine residues of the top/bottom sheet flank the Alanine residue of β-CS (Figure 5.3A).

Table 5.1: Physical properties ('b' parameter (Figure 5.1A) and shape) of poly-amino acid nanocrystal models.

Category of model	Polymer model	Initial model (Distance in Å)	Minimized structure (Distance in Å)	Equilibrated structure Mean distance (in Å) ± Standard deviation	Shape
SAA/small-AA	pGly	8.4	7.7	8.2 ± 0.2	cuboidal
	pAlaGly	8.4	9.0	9.4 ± 0.06	cuboidal
	pAla	8.4	10.4	10.5 ± 0.03	cuboidal
PAA/ polar-AA	pThr	15	14.8	16.8 ± 0.3	cuboidal
	pAsn	18.5	18.9	17.8 ± 0.7	cuboidal
HAA/ hydrophobic-AA	pVal	15.2	16.2	19.0 ± 0.2	Twisted cuboidal
	pIle	17.2	19.2	22.2 ± 0.2	Twisted cuboidal

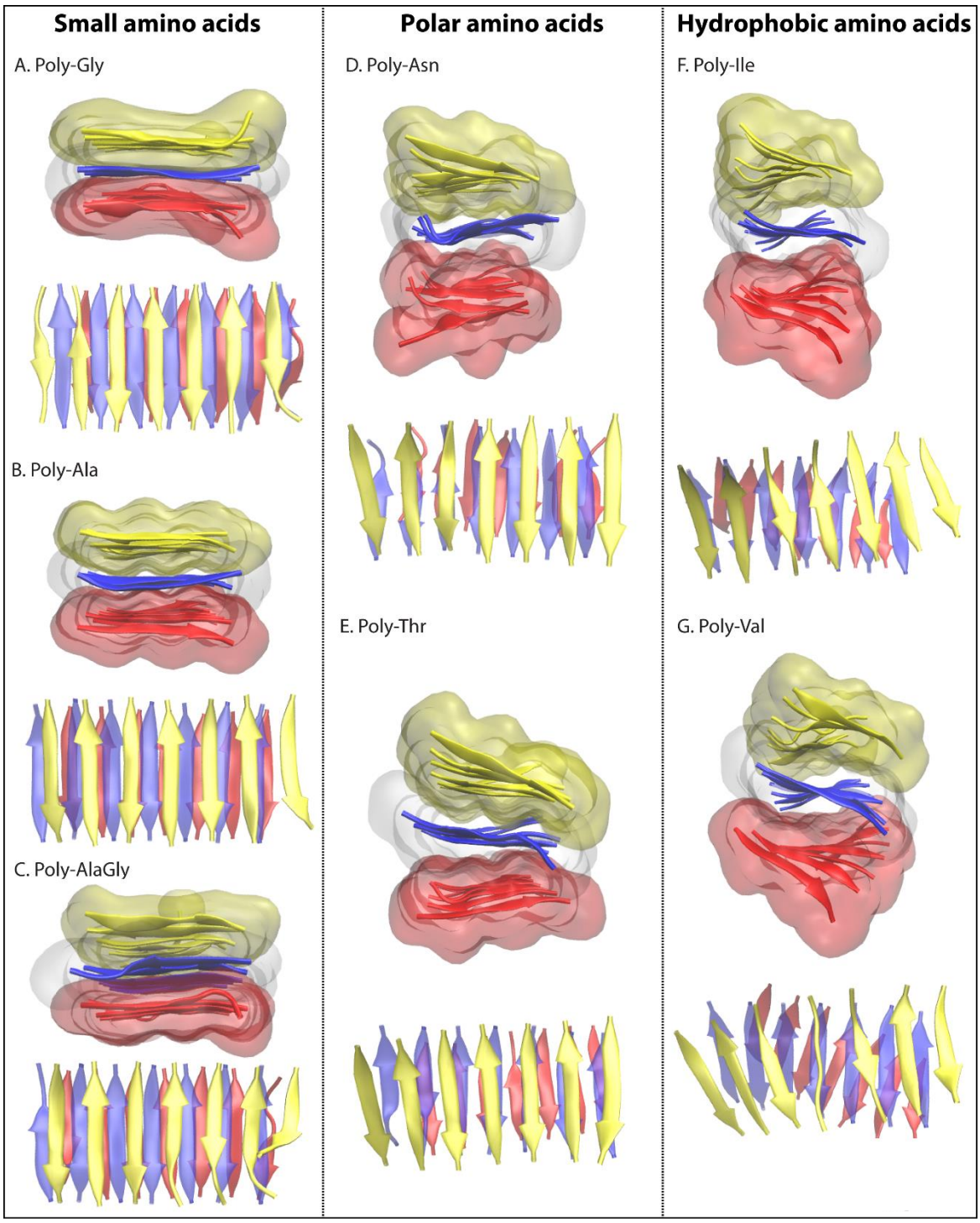


Figure 5.2: Shape and topology of homopolymers representative models. Figure shows side and top views of each homopolymer modeled nanocrystals post equilibration step. The side view shows β-strands in the new cartoon representation with upper/bottom sheet depicted in quicksurf representation and ghost rendered quicksurf view of the middle sheet. In the top view of nanocrystal, β-strands are shown in the new cartoon representation as viewed from above the model. In each view, the upper, middle, and bottom β-sheet layers are shown in yellow, blue, and magenta red colors respectively.

The FToF interaction of hydrophobic/polar amino acids shows an offset arrangement in the packing of two layers accommodating bulky side chain groups in a tightly packed environment without affecting the interlayer β-sheet interactions. In the case of polar side chain groups, we observed that side-chain hydrogen bonding reduces this offset between layers (Figure 5.3B) in comparison to hydrophobic-AA models (Figure 5.3C). The β-sheet arrangement of bulky amino acids shows a twisted cuboidal geometry in comparison to SAA models. Importantly, the side-chain packing, and edge-to-edge inter-strand hydrogen bond interactions were maintained in the modelled nanocrystals.

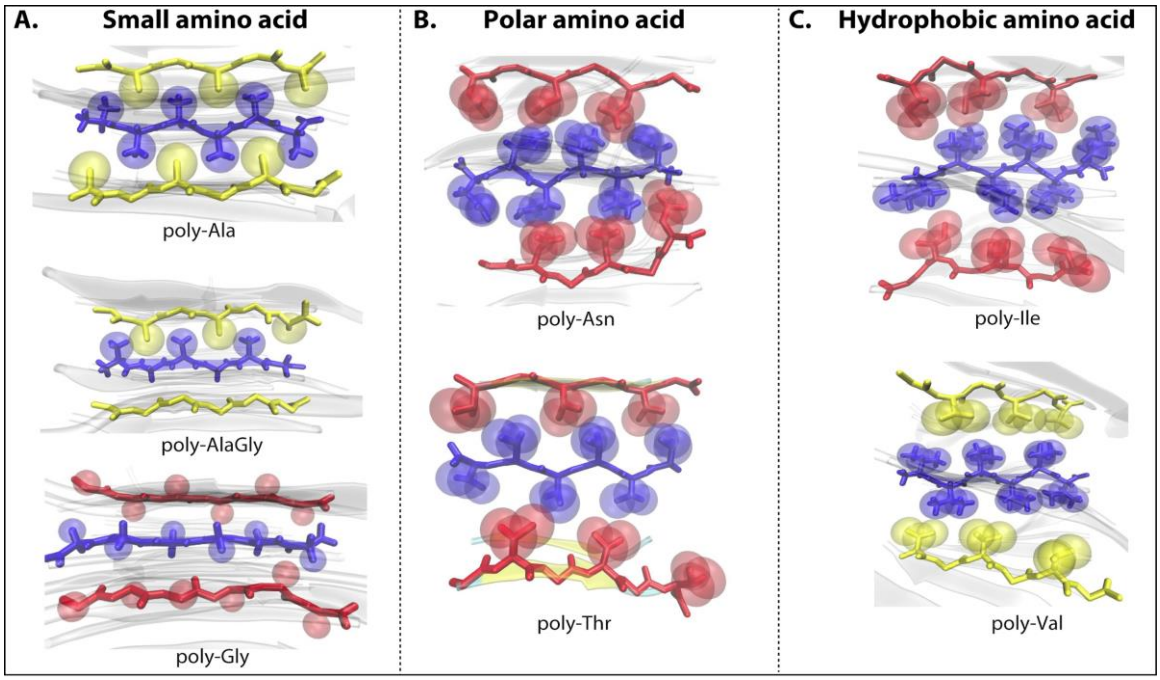


Figure 5.3: Face-To-Face side chain packing arrangement various models of hompolymer β-sheet nanocrystal. Transversal section of nanocrystal representative model of various homopolymers showing Face-To-Face side chain packing arrangement of β-CS with upper and lower layers. The β-CS is shown in blue color and other strands are shown either in yellow or red colors. The backbone and side-chain are shown in licorice and sphere representations, respectively.

5.3.3 Mechanical strength and toughness of homopolymer β-sheet nanocrystal

We investigated whether enhanced side chain interactions (hydrogen bond/hydrophobic) can increase the tensile strength and toughness of nanocrystals compared to that of naturally occurring silk sequences. As we are only interested in finding the ultimate tensile strength, we focused on the computation by pulling-out the central β-strand (β-CS) of the middle layer at a constant velocity using SMD in an explicit water simulation using boundary conditions

described in Materials and Methods. We have performed three replicates of SMD pulling for each homopolymer model to assess consistency of simulations. We evaluated the stability of the solvated nanocrystal system using Root Mean Square Deviation (RMSD) of the β-sheet structure in the production phase *i.e.*, after NPT equilibration (~1 ns) for the same simulation time as required for pulling β-CS. The β-sheet models of various amino acids do not show large variation from the equilibrated structure as evaluated using RMSD (Figure 5.4), suggesting that modelled nanocrystals form stable structures.

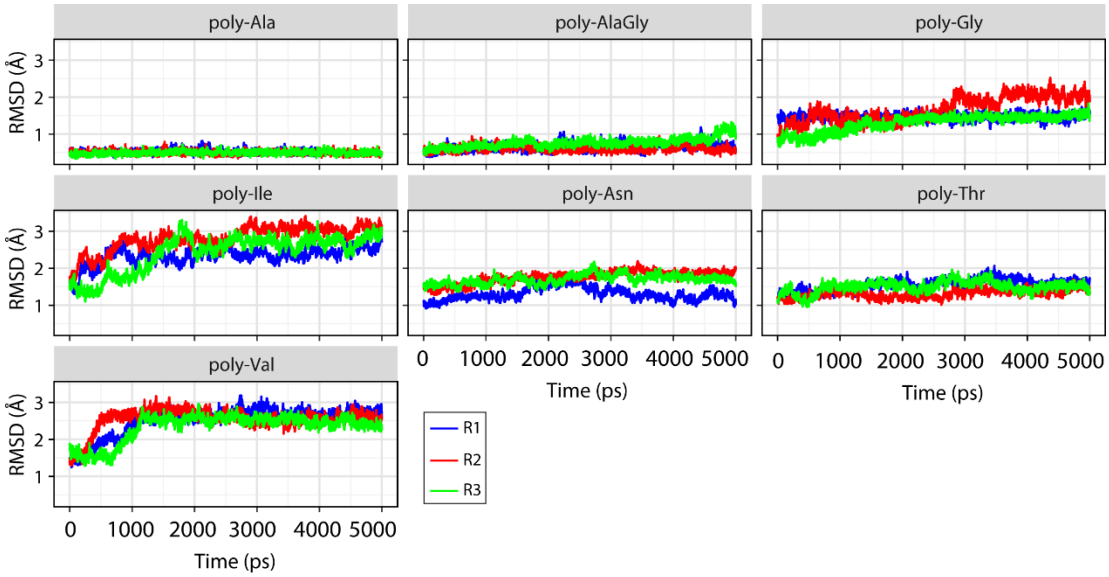


Figure 5.4: RMSD variation for various homolpolymers. Plot showing RMSD variation of homopolymers models from their energy minimized structure during equilibration and production runs. The SMD pull was initiated after equilibration of all systems (~1ns). The RMSD were calculated for strands, which were not fixed during SMD. The production phase constitutes ~4 ns after equilibration. Increase in RMSD for bigger systems especially for hydrophobic amino acids has been observed as they undergo twist from their minimized structure.

We analyzed force-displacement curves of β-CS pull-out simulations of all modelled homopolymers and compared their mechanical properties. The smoothened bin averaged force-displacement curves of representative replicate for homopolymers are shown in Figure 5.5 (force displacement profile behaviour in all replicates can be seen in Figure 5.6) . Evidently, small-AA models show multiple force peaks with decreasing magnitude at larger displacements giving it a characteristic saw-tooth pattern (Figure 5.5A). This pattern of force peak has been observed in a similar study on nanoconfined pAlaGly nanocrystal of L < 3 nm that suggested hydrogen bond cooperativity and stick-slip motion facilitates high rupture forces with enhancing energy dissipation (Xu and Buehler 2010). The rigid molecular assembly

(stick) is primarily due to cooperative rupture of backbone hydrogen bonds, which on its complete loss leads to a strand sliding (slip) before the reformation of hydrogen bonds (Figure 5.5B and 5.5C). Interestingly, both HAA and PAA nanocrystals show variable loss of molecular stick-slip motion even though these have intact backbone hydrogen bonds in their starting models. Some replicates of homopolymer models exhibited an extended slip motion without reformation of hydrogen bonds such as pIle. Importantly, the above characteristic features of β-strand pulling were consistent in other SMD replicates of various homopolymer structures, as is observed from their bin averaged force-displacement curves (Figure 5.6)

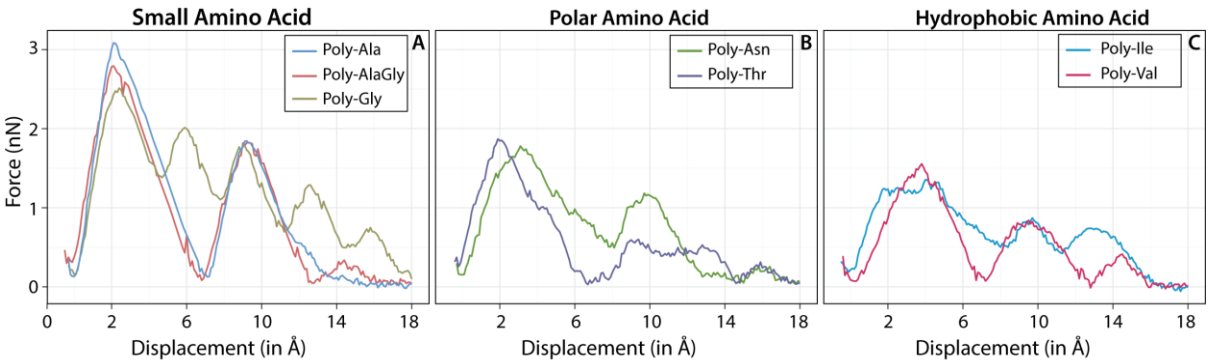


Figure 5.5: Force-displacement profiles of modeled homopolymers. Bin averaged force-displacement profile of a representative pulling β-CS SMD simulation replicate for various homopolymers. The line is mean force calculated for displacement bin width of 0.1 Å.

We compared mean of the maximum rupture force (at the first peak) over three SMD replicates for various modelled homopolymers to identify the β-sheet sequence motifs responsible for the best tensile strength. This could also assess the effect of increasing side chain interactions mechanical features of nanocrystal. As shown in Figure 5.5A, small-AA models showed a higher tensile strength, followed by polar-AA and hydrophobic-AA models. The pAla model requires the maximum β-CS pull-out force of 3.07 ± 0.10 nN, followed by pAlaGly (2.80 ± 0.02 nN) and pGly (2.27 ± 0.20 nN). In comparison to SAA, surprisingly, polar/ hydrophobic-AA models exhibit lower tensile strengths despite having enhanced side chain interactions (Figure 5.5A). Among PAA and HAA models, the pAsn model has the maximum pull-out force (1.80 ± 0.09 nN), which is only ~58% of the force magnitude in pAla models (Figure 5.7A).

Next, we compared the average toughness of homopolymer structures computed over three replicates. The toughness was calculated as the area under the force-displacement curve normalized by the volume of the nanocrystal. Similar to the observation in ultimate tensile strength, SAA models have a higher relative toughness than hydrophobic/polar-AA models

(Figure 5.5B). Among small-AA nanocrystals, we observed the maximum toughness of pAla model followed by pAlaGly and pGly models. In a related study, the toughness for pAla was found to be higher than pAlaGly (Xiao et al. 2009). In comparison to pAla, the toughness of pAsn was found to be lowered by ~34% among polar-AA models and pIle had ~33% lower toughness (Figure 5.5B). These results clearly show that β-sheet nanocrystals consisting of naturally occurring amino acids (pAla, or pAlaGly) have higher strength and toughness than other modelled polar/hydrophobic repeat sequences. Thus, suggesting that increasing side chain interactions (hydrogen bonds/hydrophobic) in nanocrystals do not increase their mechanical tensile strength or toughness. Notably, these side chain interactions have negatively affected material property as well as altered their rupture behavior. To study the effect of pulling speed, we performed SMD with 10 times slower pull velocity ($\dot{x}$ = 0.0005Å ps^{-1}) and observed that results (tensile strength/toughness) from both pull speeds are quantitatively comparable to each other. Importantly, pAla models showed the best tensile mechanical features.

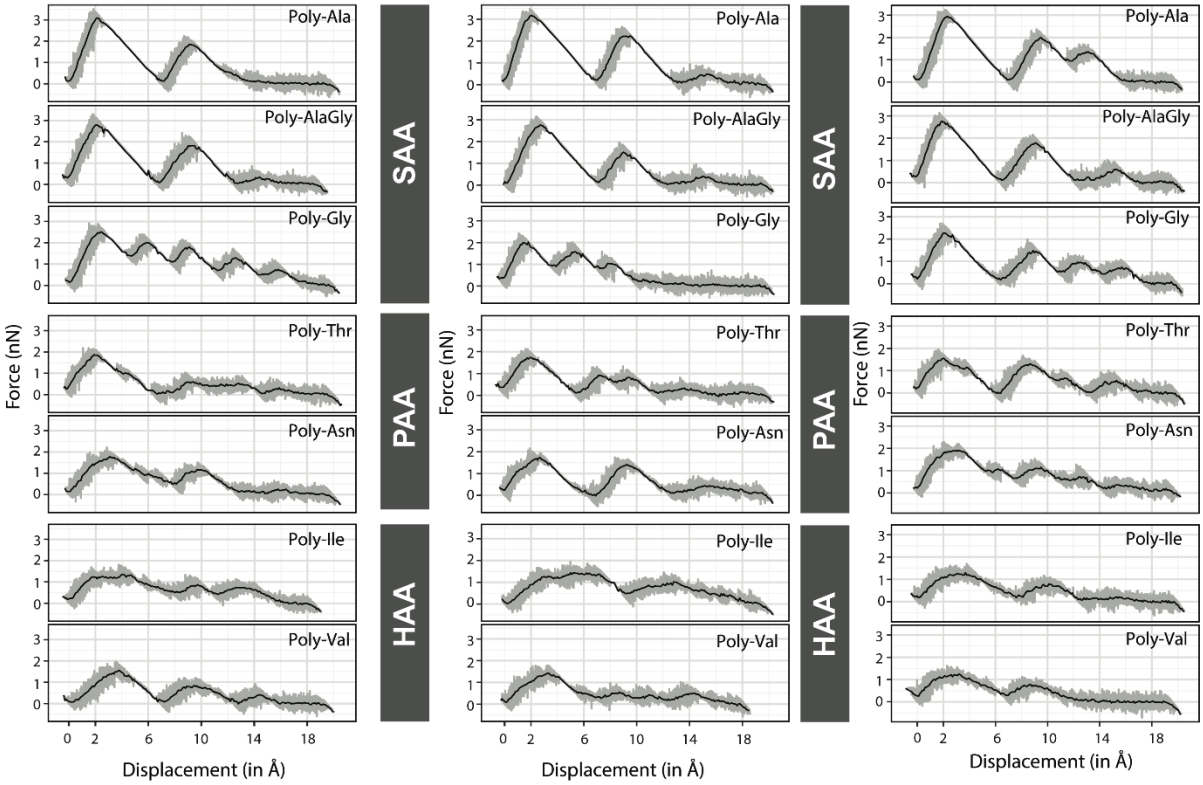


Figure 5.6: Force-displacement profiles for homopolymers of nanocrystal models. Bin average force-displacement profiles of all replicates for modeled homopolymer structures. The light gray colored is the raw data and black is the bin average fit line over displacement 0.1 Å bin width.

Further, we analyzed hydrogen bonds and β-strand dynamics during pull-out simulations to understand the contrary effect of increasing side chain interactions on mechanical features of nanocrystals.

5.3.4 Dynamics of hydrogen bonds

We carefully examined the intra-strand hydrogen bond interactions because these are essential for maintaining β-sheet topology and play a fundamental role in imparting tensile strength to nanocrystal (Keten et al. 2010). Instead of analyzing dynamics of individual main chain hydrogen bonds of β-CS, we discretized these empirically, by combining them into super-rings. Ideally, a super-ring is defined as composed of four consecutive hydrogen bonds from two consecutive 10-membered HB rings (Figure 5.1C). For analyzing interactions during pull simulations, we defined a relaxed criterion to identify super-rings; where even if at least one hydrogen bond is identified in both 10-membered HB rings such that it involves both adjacent strands. Such a criterion to identify super-ring ensures the inclusion of trajectory frames having weak β-CS interactions with both adjacent strands. Based on this definition, three super-rings can be defined for a β-CS that encompasses its interactions with both adjacent strands. We considered four consecutive hydrogen bonds with the perspective that a super-ring could define hydrogen bond cooperativity. We analyzed β-CS super-rings to investigate hydrogen bond dynamics during pull in various modelled nanocrystals. We identified super-rings in trajectory frames until the first force peak in SMD simulations and compared them among various homopolymer nanocrystal structures. It is evident from their timeline analysis of representative replicates (Figure 5.8) that small-AA models have a qualitatively higher density of super-rings, especially close to the peak force. On the contrary, polar/hydrophobic-AA models have relatively lesser density in frames close to the peak forces. Previously, it has been suggested that 3-4 hydrogen bonds show a cooperative rupture contributing to the mechanical strength (Buehler and Yung 2010; Keten et al. 2010), where these interactions not only act like clamps between intra-layer β strands but also helps to dissipate mechanical tension(Xu and Buehler 2010). If a super-ring represents the cooperativity of hydrogen bonds, the analysis indicates its loss in polar/hydrophobic-AA models might contribute to their lower mechanical strength. Among hydrophobic-AA, pVal shows relatively more super-rings close to the first force peak, mostly because of the adjacent strand being pulled along with the middle strand. Therefore, it is essential to analyze super-rings with other pull dynamic features.

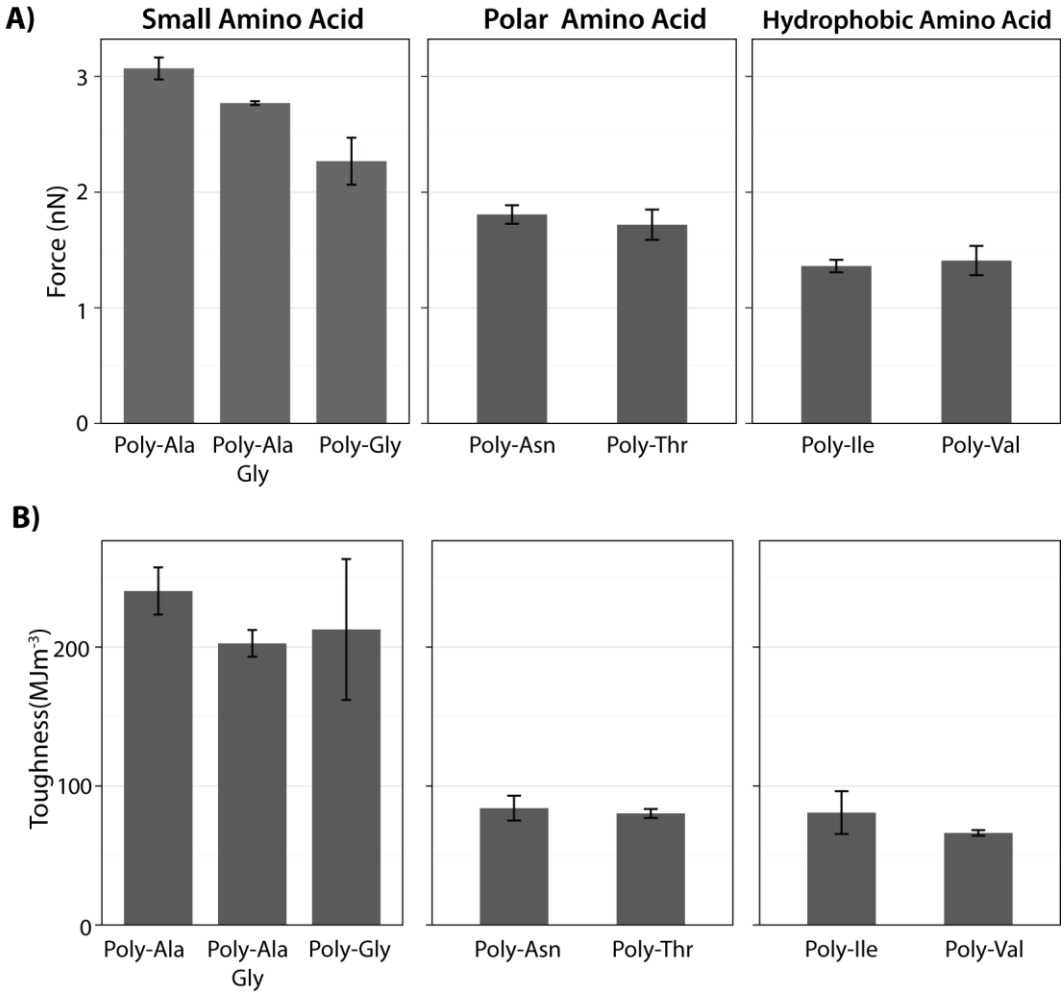


Figure 5.7: Mean ultimate tensile strength and toughness of various homopolymers nanocrystals. A) Histogram of mean rupture at the first peak force of three replicates is shown for small-AA, polar-AA and hydrophobic-AA models. B) Mean toughness calculated as area under the curve of bin averaged force-displacement curves shown for small-AA, polar-AA and hydrophobic-AA models. The error bars show standard deviation from the mean value.

Since side-chains of polar amino acids can form hydrogen bonds, we investigated whether these could alter the backbone hydrogen bond dynamics. To examine this, we categorized hydrogen bonds based on whether donor/acceptor atoms are from adjacent strands (Adj-hbonds) or between two consecutive layers (Int-hbonds). Further, these were sub-divided based on the interaction between main-chain and/or side-chain into: main-chain to main-chain (MCMC), main-chain to side-chain (MCSC) and side-chain to side-chain (SCSC). At the peak force, we identified hydrogen bonds in various structures and classified them into various types/sub-types. The distribution of the average number of hydrogen bonds with their types/sub-types is shown in supporting figure (Figure 5.9). As expected, polar-AA models (pAsn and pThr) have the maximum number of hydrogen bonds because their side-chain can

form one or more of these. For instance, pAsn and pThr have on average 4.2 and 3.0 number of SCSC hydrogen bonds. Notably, MCMC types are lower in polar-AA models suggesting that the side-chain atoms compete with main-chain atom for hydrogen bond interaction. Therefore, this likely affects the hydrogen bond cooperativity in polar-AA models.

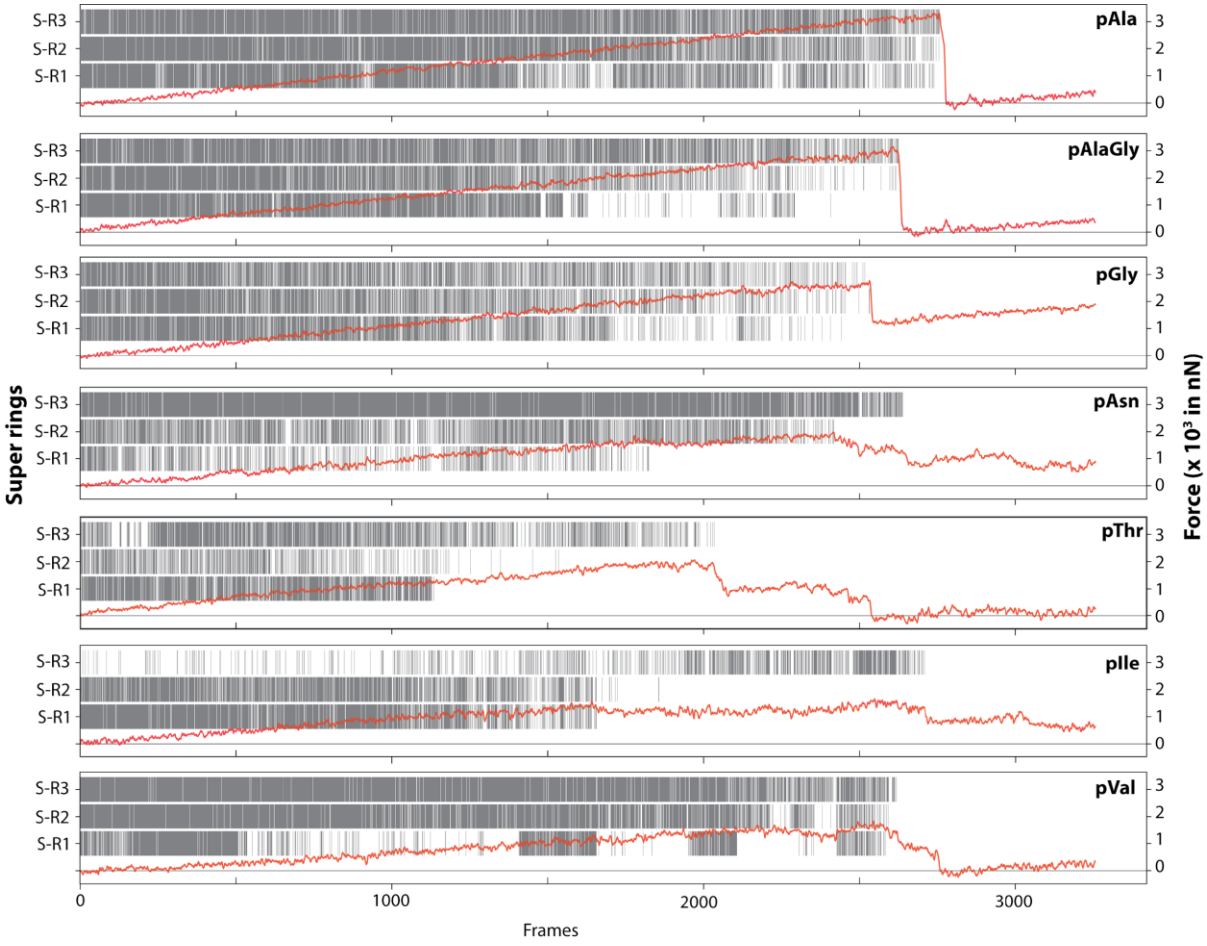


Figure 5.8: Timeline analysis of hydrogen bond super-rings. Heatmap showing timeline analysis of super-rings where vertical black tile represents presence of super-ring and no line is drawn for absence super-ring. The red line represents pull force (axis on right side). The S-R1, S-R2, S-R3 represent super-ring 1, 2 and 3 respectively. The S-R1 is the first super-ring from the first Cα of β-CS.

5.3.5 Dynamics in pull-out simulation

To further understand the mechanistic details of failure mechanism in various nanocrystals, we compared other characteristics of pull-out dynamics, such as interactions of β-CS, relative differences in stick-slip dynamics, twisting of β-CS and distance between the peak force in force-displacement profiles.

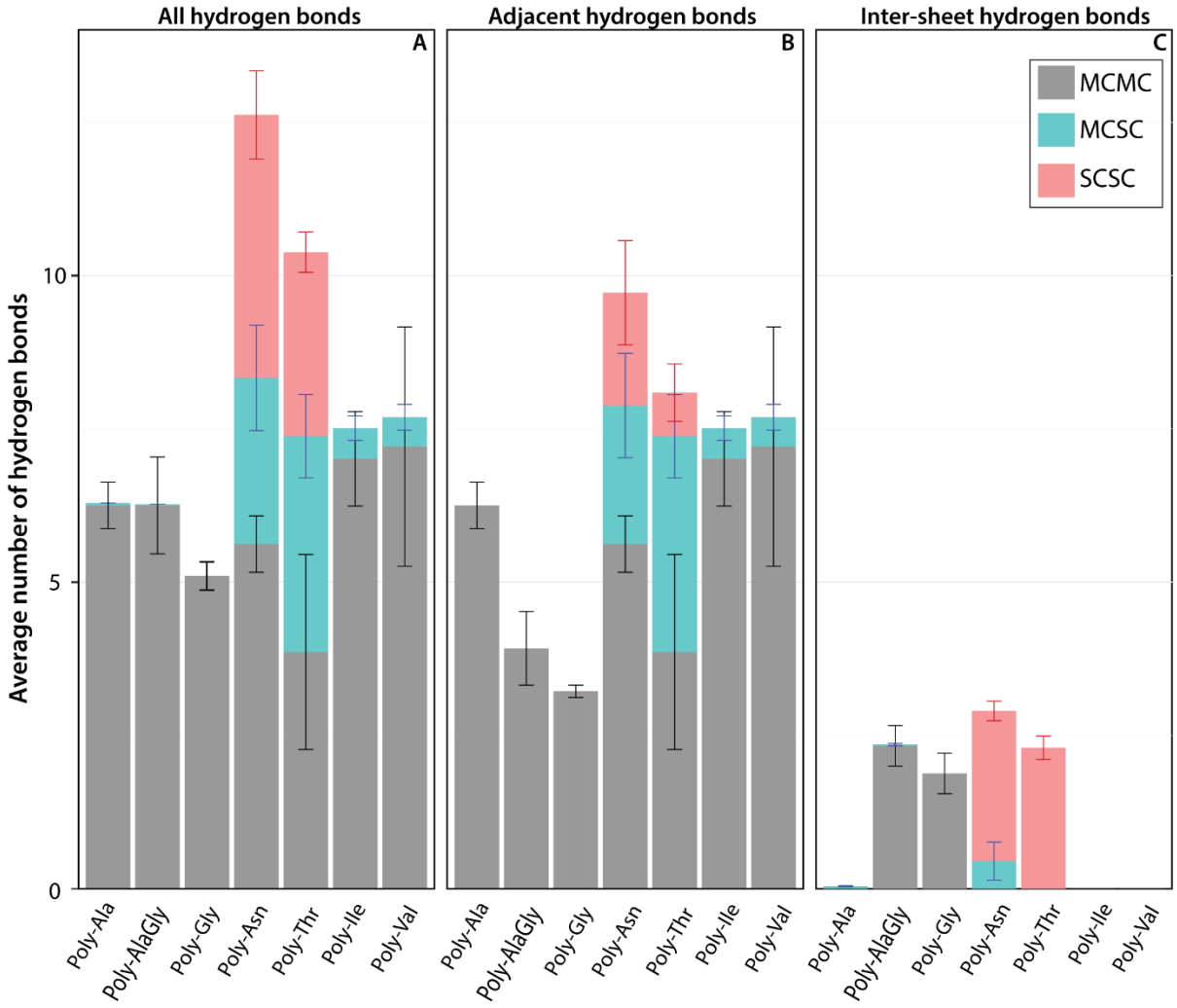


Figure 5.9: Mean hydrogen bonds and their subtypes determined at peak force. Stacked histograms of SMD replicate mean number of hydrogen bonds at the maximum peak force. The hydrogen bonds were classified into adjacent/inter-sheet and further into main-chain (MC)/side-chain (SC) based on atoms. The mean count of all, adjacent and inter-sheet hydrogen bonds are shown in panels figure A, B and C respectively. Various MCMC, MCSC and SCSC subtypes of hydrogen bonds are shown in gray, teal and orange red colors histogram respectively. Error bars are for the standard deviation over mean of three replicates.

To investigate β-CS interactions, we performed a qualitative analysis by visual inspection of SMD trajectory during β-CS pulling (Movies S1 to S7) combined with their force-displacement profile until rupture and hydrogen bond reformation after the first peak. The visual inspection of SMD trajectories showed that the polar-AA (Movies S4 and S5) and hydrophobic-AA models (Movies S6 and S7) does not show a distinct slip after an initial stick-like motion, and the same was evident from their force-displacement profiles. The extended slips are attributed mostly to the extensive side-chain (hydrogen bond or hydrophobic) interactions, which kept β-CS adhered to either/both adjacent strands after rupture of main-chain hydrogen bonds.

Moreover, this was quantitated from extended slip times of PAA/HAA models relative to pAla, which took on average ~130 ps for the slip. Relative to this, pThr, pAsn, pVal, and pIle showed increased slip times by a factor of 2, 2, 4, and 5, respectively. In polar-AA models, side-chains hydrogen bonds (SCSC/SCMC) quickly reform upon their initial rupture contributing to an extended slip along with a zigzag-like motion of β-CS (Movies S4 and S5). In HAA models, the hydrophobic sticky side chains keep the β-CS bound to adjacent strands prohibiting its release. These bulky side chain interactions can be overwhelming to the extent that adjacent strands are pulled along with β-CS in pVal (1 of 3 replicates, Movie S6C) and pIle (2 of 3 replicates, Movies S7A and S7C). Importantly, inter-layer side-chain packing interactions fail to prevent the pulling of three β strands together.

We examined differences in the displacement between peak forces in force-displacement curves (Figure 5.3). As evident, most nanocrystal structures have a typical distance of 7.6 Å between consecutive peaks, which coincides with a displacement of β-CS by two amino acids. However, pGly models showed a displacement of 3.8 Å accompanied by a full rotation of β-CS on its backbone axis. The discrete distances between peaks are observed primarily due to constraints imposed by hydrogen bond arrangement in an antiparallel β-sheet, where narrowly placed hydrogen bond pairs alternate with widely spaced ones (Figure 5.1C). Without breaking such an arrangement, the β-CS can move in the direction of pull either by displacement of two C-alpha (two amino acids) atoms or by one C-alpha atom accompanied by a rotation of the whole strand/peptide bond. In either of the above possibilities, the hydrogen bond would reform to facilitate locking strands in a thermodynamically favored position (38). In pull-out simulation, only pGly model showed strand rotation because the side-chain in the rest of amino acids would lead to steric clashes in a tightly packed β-sheet environment. Another distinguishing feature of the pGly model was that resistive force does not become negligible during slip in pGly models (Figure 5.3A), mostly because hydrogen bonds reform faster in strand displacement of one amino acid.

In pull-out simulations, we visually observed that slip is accompanied by its slight twist and marginal opening of β-sheet layers, which close as hydrogen bonds reform. This layer open/close motion appears as a pac-man like movement. Importantly, both β-strand twist and opening/closing of layers facilitate the release of β-CS by reducing steric hindrances in the compactly packed nanocrystal. The above motions were more prominently observed in SAA models. It is worth speculating that inhibiting the twist or opening of β-sheet structure can disrupt the nanocrystalline assembly leading to a permanent dissociation of non-covalent

supramolecular interactions, which can compromise the integrity of nanocrystal by rendering it soft and permeable to the solvent.

The present study evidently shows that commonly occurring sequences $(AGAGAG)_n/(AAAAAA)_n$ of β-sheet nanocrystal in silk have superior strength and toughness compared to homopolymers of other amino acids. The natural sequence motifs achieve the optimal compact packing and intra-strand hydrogen bond interactions to render the best tensile mechanical properties. Based on this study, it is appealing to suggest that commonly observed nanocrystal β-sheet sequence motifs in silk have probably evolved to maximize their strength/toughness.

5.3.5.1 Web Enhanced Objects

The pulling instance of nanocrystals reminiscent of their pulling dynamics in three replicates has been uploaded and can be accessed from https://pubs.acs.org/doi/10.1021/acsbiomaterials.1c00447. These animated movies of SMD pull trajectories were used to visualize stick-slip dynamics, hydrogen bond interactions, and β-sheet twisting for different amino acid instances. Each instance (S1 to S7) represents a distinct amino acid (S1: pGly, S2: pAla, S3: pAlaGly, S4: pThr, S5: pAsn, S6: pVal and S7: pIle), while replicates A to C represent three repetitions. There are a total of 21 files, where three replicates will be recorded for each 7 distinct nanocrystals modelled. Files will be names as S1A, S1B, S1C, S2A till S7C representing, pulling instance of pGly in repA, repB and repC; followed by pulling instance of pAla repA (S2A) and last (S7C) as pulling instance of pIle repC.

5.4 CONCLUSIONS

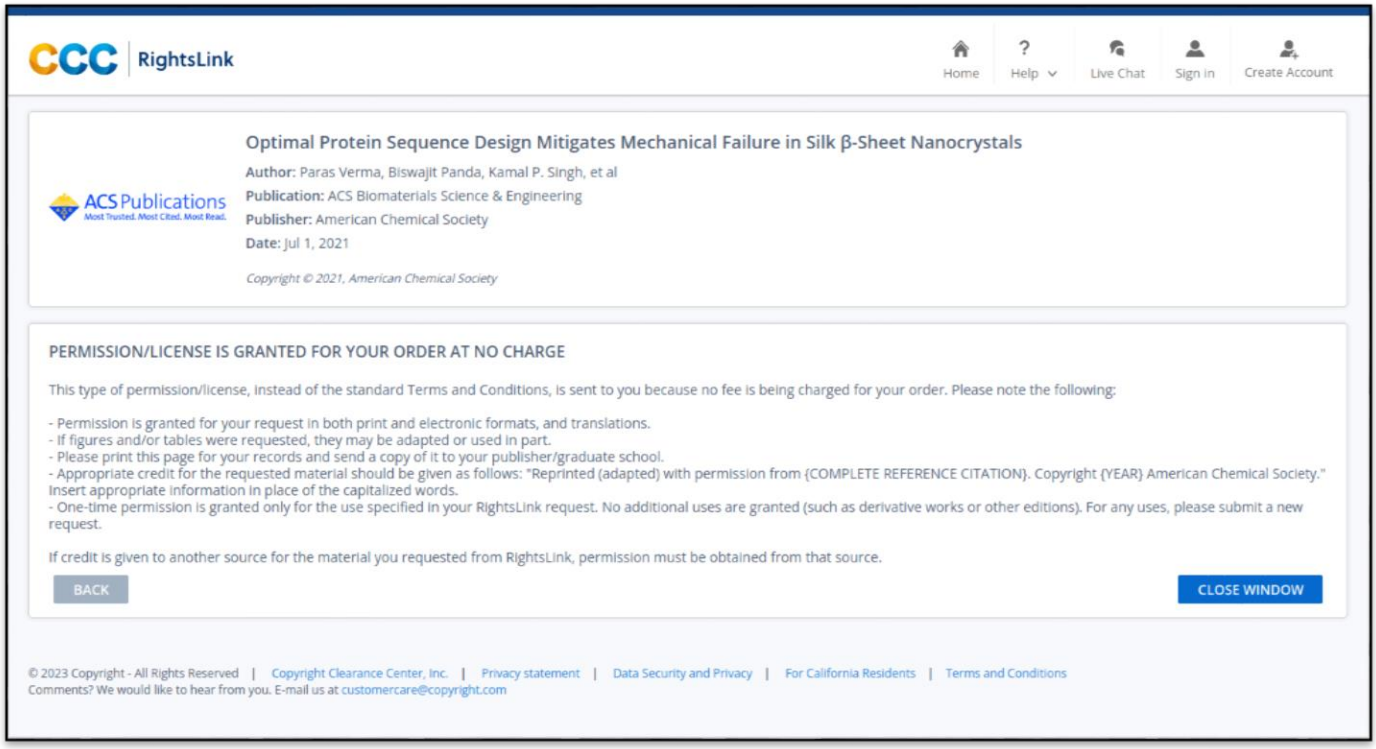
In the present study, we have delved into whether Nature's optimal sequence design (pAla or pAlaGly) of the β-sheet nanocrystal can be modified to increase its mechanical strength/toughness. To evaluate this, we modelled poly-polar/hydrophobic amino sequences to increase the side-chain interactions with an assumption that these could potentially supplement to improve the ultimate tensile strength. The computational investigation into the tensile strength using SMD pull-out simulations showed that all modelled homopoly-amino acid constructs could withstand forces greater than > 1.5 nN. Contrary to our expectations from designed homopolymer structures, these showed lower mechanical tensile strength than those

consisting of naturally occurring sequences (pAla or pAlaGly). Moreover, the pAla model has the maximum toughness. Relative to this, the best of homopoly-polar and homopoly-hydrophobic-AA models have lower toughness by 34% and 30%, respectively. It has been speculated that the addition of such repeats of amino acids especially of hydrophobic nature would improve the overall force resilience of nanocrystals (Johansson and Rising 2021). A previous study on pAlaGly and pAla system showed that side chain contributes to rupture forces as these are reduced by 28% when side-chain potentials are switched-off during SMD (Xiao et al. 2009). However, in our study we found that tensile strength does not increase by having additional side chain interactions. Moreover, we observe that having excessive interactions adversely affects the tensile mechanical properties of nanocrystal. Thus, this indicates a possibility that there is probably an inherent limit of improving the nanocrystal tensile strength by simply increasing side chain interactions among β-strands. Recently, studies on poly-(Gly), poly-(Gly-Ala), and poly-(Gly-Ser) models using empirical DFT and MD simulations have shown that in poly-Gly-Ala models, Alanine side-chain confers rigidity at the expense of destabilization interactions between layers (Mayen et al. 2015). A recent study on modelled uniform serine/ glycine/alanine nanocrystal with amorphous region showed that serine model has high tensile strength and lower toughness than other amino acids models (Kim, Choi, et al. 2020). The lower toughness is suggested to arise because of inability of hydrogen bond regeneration. However, it is not apparent whether the nanocrystalline region necessarily contributes to high tensile strength. In our study, we also observed the effect of backbone hydrogen bond regeneration in polar-AA models and its competition with side-chain H bonds during its extended slip after yield point.

The detailed analysis of the rupture mechanism in various homopolymers suggests that extensive side-chains interactions between β−strands within a layer have affected the molecular stick-slip motion and concomitantly affected the dissipative force owing to this motion. In polar-AA models, the potential competition of main-chain and side-chain atoms for hydrogen bonds probably weakens hydrogen bond cooperativity. Notably, nanocrystal require offset in β-sheet arrangement along with inter-layer twisting and increased spacing between layers for accommodating bulky side-chain groups (Cheng, Pham, and Nowick 2013). Overall, such changes might have affected the tightly packed environment as well as the nature of nanoconfinement. Thus, our study shows that a fine balance of side chain interactions, hydrogen bond cooperativity and β-sheet layer packing is essential for achieving the high tensile mechanical strength/toughness in silk.

We have studied homopoly-amino acid models in the context of their failure responses, which provided insights into shifts in nanomechanical properties and understanding the effect on incorporating bulkier amino acids in nanocrystal. The present atomistic understanding of pull dynamics in modelled structures could serve as a primer for designing other amino acid combinations in β-sheets models and designing globular protein with desired mechanical properties.

Permission to reproduce published work in thesis



Chapter 6

Unraveling the Functional Implications of Y321A mutation in the Vibrio cholerae cytolysin through MD Simulations and network analysis

Adapted from the computational analysis section discussing mechanistic insights from main article and supplementary material of: "Mondal, AK, Verma, P, Sengupta, N, Dutta, S, Bhushan Pandit, S, Chattopadhyay, K. Tyrosine in the hinge region of the pore-forming motif regulates oligomeric β-barrel pore formation by Vibrio cholerae cytolysin. Mol Microbiol. 2021; 115: 508– 525" https://doi.org/10.1111/mmi.14631 (Mondal et al. 2021). Copyright (2021). with permission from John Wiley and Sons. Details of the license are attached at the end of chapter.

6.1 INTRODUCTION

Pore-forming toxins (PFTs) are distinct membrane damaging protein toxins that usually form oligomeric pores to kill their target cells. Such cytotoxicity mediated by making pores is found to be conserved mechanism of killing cells in evolutionary time scale. Many pathogenic bacteria use PFTs as dominant mechanism to kill cells and act as virulent factors. Among various PFTs, Vibrio cholerae Cytolysin (VCC) from *Vibrio cholerae* is a widely studied protein toxin belonging to the bacterial β-PFT family. VCC forms pores in the red blood cells, leading to hemolytic activity and also show cytotoxicity thereby contributing to pathogenicity (Mondal and Chattopadhyay 2019).

VCC is secreted in a precursor inactive form as pro-VCC, which is activated upon cleavage of its N-terminal pro-domain by the action of bacterial proteases. The mature form of VCC is dimorphic protein, which is monomer in solution, and it assembles on interaction with membrane to form homo-oligomer heptameric β-barrel pores. The tertiary structure determination of monomer and oligomeric structures revealed that central scaffold domain of VCC referred to as cytolysin consists of pore-forming motif (pre-stem loop), which undergoes major structural transition during oligomerization. The pre-stem loop is packed against cytolysin domain in the monomer form that in oligomeric form creates stem of transmembrane β-barrel scaffold. This transition is depicted in Figure 6.1. Apart from cytolysin domain, mature

VCC consists of lectin-like domains viz. β-Trefoil and β-Prism. The transition of VCC from monomer to heptamer has been extensively studied to unravel the series of mechanistic steps involved in this structural transformation. Most of these studies relied on single residue mutants with varying degrees of impact ranging from partial to complete abolition of VCC hemolytic activity. Our collaborator (Prof. Kausik Chattopadhyay, DBS, IISER Mohali) performed single residue mutants of residues, which interact with membrane. Through computational work, we have obtained valuable insights into the possible effects of such mutations.

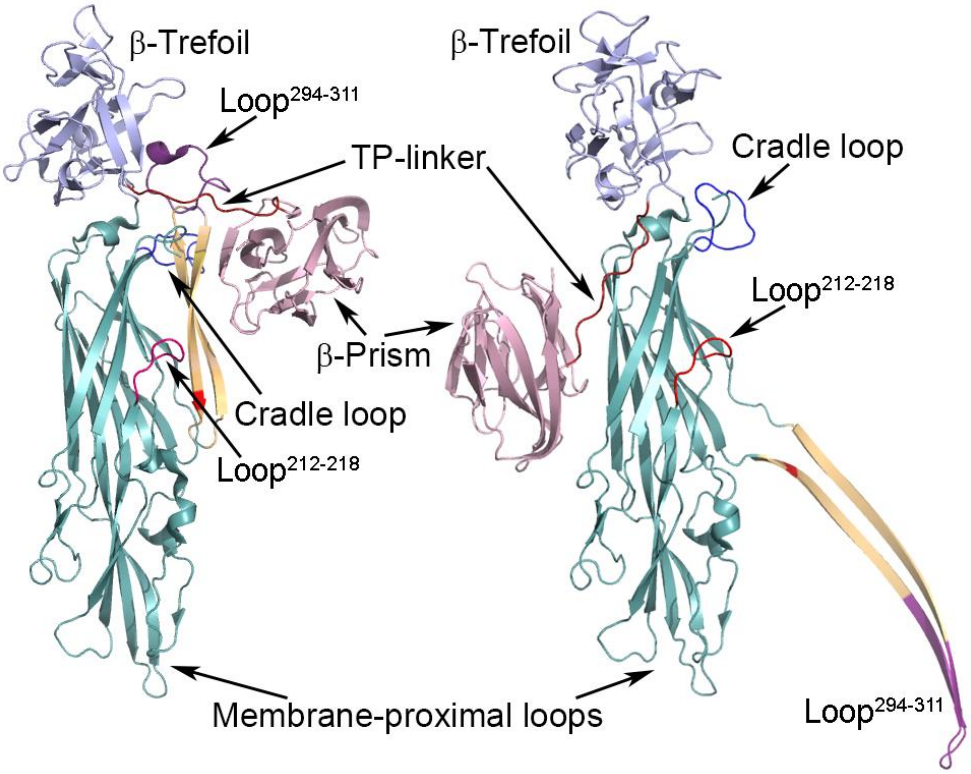


Figure 6.1: Structural domains/motifs in monomeric and oligomeric VCC. Structure of the monomeric state of VCC (PBD ID: 1XEZ) is shown on the left. Structure of a protomer (PDB ID: 3O44) unit from the oligomeric pore state of VCC is shown on the right. Location of the residue Y321 within the pore-forming motif (shown in light yellow color) is highlighted with red color in both the structures. Comparison of the two states clearly highlights prominent reorganization/rearrangement(s) of some of the major structural motifs/domains during the oligomeric pore-formation process of VCC. They include: (i) pore-forming pre-stem motif, (ii) β-Trefoil domain, (iii) β -Prism domain, (iv) TP-linker that connects the β-Trefoil and β-Prism domains, (v) cradle loop, and (vi) loop294-311.

As active β-barrel pore formation involves membrane interaction (Mondal and Chattopadhyay 2019), insights can be gained regarding detailed sequence of events leading to structural transition by examining aromatic residues that can have physicochemical interaction with membrane. Therefore, we investigated roles of residues F280, F288, Y313 and Y321,

which interact with membrane in mature functional pore. These were studied for their consequential impact due to mutation to elucidate their potential role in mediating physiochemically relevant interactions with the membrane lipid bilayer (Hong et al. 2007). The selection of such residues weighed their structural reorientation in addition to their membrane proximity and correspondingly can potentially provide insights on allosteric impact of their interfacial interactions mediating between protein surface and complementary lipid subgroups. These insights can be purposed to perform targeted drug discovery as VCC is one of major toxin in cholera pathogenesis (Saka et al. 2008).

Of the chosen 4 key aromatics residues, mutation Y321A drastically hampered the pore forming capability. Our detailed study using computational approaches of MD simulations and network approach indicated that the disruption of Y321 hydrophobic pocket in the monomeric structure compromises possible crucial intra-domain interactions affecting important conformational changes required to form interprotomer interfaces. By detailing these findings, our study sheds light on the critical functional implications of Y321 in the context of crucial structural reorganization and conformational change from monomeric to heptameric active state. Furthermore, our findings not only complement but also extend mechanistic aspects of conformational changes required in VCC protein during oligomerization. These also underscore allosteric sensitivity, which could potentially be modulated by appropriate ligands to disrupt long distance communications as has been observed in the case of Y321.

6.2 MATERIALS AND METHODS

6.2.1 Molecular dynamics (MD) simulation

The cartesian coordinates for MD simulation were obtained from monomeric crystal structure of VCC (PDB code: 1XEZ). Since the pro-domain of VCC is not present in protein used for experimental studies, we removed residues (residue id: -21 to 134) in the crystal structure corresponding to pro-domain. This truncated crystal structure of VCC without pro-domain was used for MD simulation and henceforth referred to as WT-VCC. The mutant structure Y321A was generated using *mutator* plugin of VMD (Humphrey, Dalke, and Schulten 1996a). In the present study, we performed simulations for two structures: WT-VCC and Y321A-VCC. The input protein structures and the coordinate files for simulations were created using *autopsfgen* plug-in of VMD. The prepared structures were solvated with TIP3P explicit water molecules within an approximate box size of 97x122x176Å using *solvate* VMD

plugin. This was neutralized using Na^+ ions using *autoionize* plugin. Both simulation systems have ~64000 water molecules with 8 Na^+ ions.

The MD simulations for both wild type and mutate VCC were performed using NAMD (v2.13) (Phillips, Braun, Wang, Gumbart, Tajkhorshid, Villa, Chipot, Skeel, Kale, et al. 2005) program with CHARMM22 topology and force field parameters for proteins having CMAP corrections (Huang and MacKerell 2013). The initial system was energy minimized for 10,000 steps followed by a ~500 ps of temperature equilibration at 300 K and pressure equilibration of variable times till average fluctuations minimized around at 1.01 bar (1 atm) respectively. During equilibration, the temperature gradually increased in steps of 30K/time step (2fs) to 300K. For both wild type and mutant proteins, we performed 3 equivalent simulations each of 100 ns at 300K and 1.01 bar pressure. These replicate simulations differed in the initial velocities of the atoms. The snapshots were taken at every 10 ps, which resulted in 10000 frames from each trajectory. The system was simulated in periodic boundary conditions with electrostatics interactions computed using Particle Mesh Ewald (PME) (Darden, York, and Pedersen 1993) method by specifying grid sizes. The van der Waals interaction involved switching function with cut-off distance of 12 Å (staring at 10 Å). The constant pressure at ~1 atm were maintained using Nosé-Hoover Langevin(Feller et al. 1995) piston method with piston period of 100 fs, a damping timescale of 50 fs and piston temperature of 300 K. A constant temperature of 300 K was maintained using the Langevin dynamics, with the damping coefficient set to 5 ps^{-1} for all the heavy atoms. A time-step of 2 fs was used in both equilibration and production runs.

The trajectory was visualized and mostly analyzed using the VMD program and Bio3D (Skjærven et al. 2014; Grant et al. 2006) package. The Root Mean Square Fluctuations (RMSF) was computed for the Cα-atoms after rigid body superposition of all the trajectory structures on the reference X-ray crystal structure. The B-factor was converted to RMSF using the equation:

$$RMSF_i = \sqrt{\frac{3B_i}{8\pi^2}}$$

6.2.2 Essential dynamics

The collective motions of wild type VCC and Y321A mutant were analyzed using principal component analysis (PCA) or essential dynamics (ED) on concatenated trajectory from three

simulations (~300 ns). The standard PCA was performed on Cα Cartesian coordinates using Bio3D package. The effects in components of PCA due to rotational and translational dynamics were eliminated by structural superposition of trajectory on initial crystal structure. The superposed Cartesian coordinates were used for the generation of covariance matrix consisting of elements, $\sum_{ij}$, given by:

$$\sum_{ij} = \langle (r_{i-}\langle r_i \rangle).(r_j - \langle r_j \rangle) \rangle$$

where i and j are coordinates of all possible Cα atoms, and <.> denotes mean value

The eigenvalue decomposition covariance matrix gives eigenvectors and eigenvalues or principal components (PCs). Since the first few principal modes usually have most of the variance, they describe the conformational space spanned during the MD simulation. The conformational variation can be studied by projecting onto the essential subspace or PC modes with most variance. Moreover, collective motions in essential subspace (PCs with high eigenvalues) were studied by reconstructing the Cartesian coordinates on the selected PC modes using *mktraj* module of Bio3D and VMD. The convergence of PC modes from simulation replicates were compared and evaluated using root mean squared inner product (RMSIP) (Amadei, Ceruso, and Di Nola 1999), which is widely used to assess the convergence of the main PC-modes of proteins, where value 1 indicates essential subspace encompassed by PC's being same, whereas value 0 indicates that they are orthogonal. In general, the RMSIP is greater than ~0.8 computed for the first 20 eigenvectors of wild type VCC and Y321A mutant.

6.2.3 Dynamic cross-correlation and correlation network analysis

The dynamic cross-correlations of Cα-atoms were computed for concatenated trajectory after superposing all frames on the initial crystal structure (PDB ID: 1XEZ). The dynamic cross-correlation matrix (DCCM) is given by:

$$C_{ij} = \frac{\langle \Delta r_i . \Delta r_j \rangle}{\sqrt[2]{\langle \Delta r_i^2 \rangle} \sqrt[2]{\langle \Delta r_j^2 \rangle}}$$

This is a normalized variance-covariance matrix (Lange and Grubmuller 2006; Ichiye and Karplus 1991). The C_{ij} values of 1, 0 and -1 show complete correlation, no-correlation, and anti-correlation respectively, between atoms *i* and *j*.

For identifying state specific residue couplings, the wild type DCCM*wt* was subtracted from mutant DCCM*mt*, and difference matrix ***S*** is obtained multiplying it by Kronecker delta function (δ_{ij}):

$$\Delta dccm_{ij} = DCCMwt_{ij} - DCCMmt_{ij}$$

$$S_{ij} = \left(\Delta dccm_{ij}\right) * \delta_{ij}$$

where,

$$\delta_{ij} = \begin{cases} 1 \; if \; \left(\left|DCCMwt_{ij}\right| \; or \; \left|DCCMmt_{ij}\right| \geq 0.4\right) \; and \; \Delta dccm_{ij} \geq 0.4 \\ 0 \; otherwise \end{cases}$$

Here, δ_{ij} function helps to filter strong correlations in either wild type or mutant proteins.

The residue wise correlation changes i.e., when $S_{ij} \neq 0$ were divided further into following main categories as gain/loss of correlation/anticorrelation in mutant proteins with respect to wild type protein:

a. Gain in residue correlation: $DCCMwt_{ij} \geq -0.4$ and $DCCMmt_{ij} \geq 0.4$
b. Gain in residue anti-correlation: $DCCMwt_{ij} \geq 0.4$ and $DCCMmt_{ij} < 0.4$
c. Loss in residue correlation: $DCCMwt_{ij} \geq 0.4$ and $DCCMmt_{ij} < 0.4$
d. Loss in residue anti-correlation: $DCCMwt_{ij} < -0.4$ and $DCCMmt_{ij} < 0.4$

6.2.4 Residue correlation network analysis (CNA)

The residue correlation network representing residues as nodes and the edge connecting nodes is the correlation observed between the residues. Following the approach of Yao et al. (Yao et al. 2016), we constructed residue correlation network. Briefly, the nodes (residues) are connected by edge if the correlation between nodes (residue) *i* and *j* (j=i+n, n>2) in all replicates is more than equal to 0.6; else if $|Cij| \geq 0.6$ in at least 1 replicate and the distance between corresponding Cα residues is ≤ 10Å in 75% of simulation frames. The edges are weighted by mean of $-\log(|\langle Cij \rangle|)$, across simulation (Sethi et al. 2009; Yao et al. 2016).

6.2.5 Community generation

The correlation network was partitioned according to the Girvan-Newman approach (Girvan and Newman 2002), which results in communities such that residues in a community are

connected densely inside, and sparsely to the residues of other communities according to the edge betweenness criteria. We constructed community networks at maximum modularity. The community edge weight is the maximum correlation of all edges connecting two communities.

Molecular graphics were generated using the VMD (Humphrey, Dalke, and Schulten 1996a) and plots were made with ggplot2 library (Wickham 2016). Most analyses were performed using Bio3D (Skjærven et al. 2014; Grant et al. 2006) package in R. The community clustering was obtained for the wild type VCC and Y321A mutant using the method as has been previously described (Yao et al. 2016).

6.3 RESULTS AND DISCUSSIONS

6.3.1 Experimental Results Elucidate Key Stages of Pore Formation Affected

Several experimental studies were performed on the single residue mutants *viz.* F280A, F288A, Y313A and Y321A to examine their role on pore forming activity of VCC. The experimental results clearly showed that mutants F280A, F288A and Y313A affected the VCC hemolytic activity from minimal to moderate, whereas Y321A exhibited profound impact on hemolytic activity and compromised VCC's cytotoxic activity against the T84 human intestinal epithelial cells (Mondal et al. 2021). Further, Y321A does not compromise the membrane-binding step in the pore-formation mechanism of VCC rather is arrests insertion of pore-forming motif into the membrane and probably blocks functional oligomeric pore formation (Mondal et al. 2021). Since Y321 is part of pre-stem loop, which undergoes conformation change to form the transmembrane β-barrel scaffold, we explored the possibility whether Y321 has role in stabilization of β-barrel. The analysis of heptameric crystal structure of VCC showed that R282 and E384 make side-chain interaction with Y321 apart from backbone interactions. The hemolytic activity of the mutants R282A and the double mutant (R282A and E384A) was found to be comparable to that of the wild type (WT). The comparison of Y321A with WT showed its differences in thermal denaturation profile, indicating loss in structural integrity and possible consequence on compromised SDS stable oligomers (Mondal et al. 2021). Thus, Y321A mutant is able to bind membrane efficiently but hampered at the pore formation stages, however, with little knowledge about stages in unlocking pre-stem (essential for barrel formation).

Since Tyrosine to Alanine (Y321A) is physiochemical drastic mutation, additionally we also performed conservative mutation of Tyrosine to Phenylalanine (Y321F) and found that Y321F has restored pore-forming ability of VCC.

6.3.2 In silico bioinformatics analyses of the structural models provide possible cues regarding the implication of Y321 for the pore-formation mechanism of VCC

With the results obtained from our present study so far, it still remains unclear how exactly the mutation of Y321A in VCC blocks oligomerization of the membrane-bound toxin molecules and abrogates pore formation. Based on the available crystal structure, Y321 is positioned way above the membrane plane in the final oligomeric pore state. Previous biochemical studies have shown that the three loop regions in the membrane-proximal rim domain of VCC mediate the key functional interaction of the toxin with the membrane lipid components to drive the subsequent steps of the oligomeric pore formation (Rai and Chattopadhyay 2015). The residue Y321 is positioned at a spatially distant location from these membrane-interacting motifs, both in the monomeric form as well as in the oligomeric pore state. Moreover, our result showed that the mutation of Y321A did not affect the membrane-binding efficacy of VCC. These observations altogether preclude any direct role of Y321 in mediating interaction of VCC with the membrane lipid bilayer. Also, based on our results so far, Y321 does not appear to mediate any crucial interaction between the pore-forming motifs, disruption of which could affect pore formation. Therefore, it still remains unclear how exactly the mutation of Y321A in VCC compromises the oligomeric pore-formation mechanism of the membrane-associated toxin molecules.

Interestingly, analysis of monomeric VCC found that the residue Y321 is located in the hinge region of the pre-stem motif, and it remains buried within a hydrophobic pocket lined by residues F271, P272, I276, F280, L319, I328, and W342. Moreover, side-chain hydroxyl group of Y321 is hydrogen-bonded to the main chain carbonyl group of the residue K384. Therefore, it is possible that the mutation of Y321A may cause local perturbation of this hydrophobic pocket, along with the loss of the hydrogen bond in the hinge region of the pre-stem motif. This, in turn, may disrupt the orchestrated rearrangement of the pore-forming pre-stem motif and other structural modules/domains (surrounding the pre-stem), which are the critical steps in the membrane pore-formation mechanism of VCC. Drawing from our observations, we postulated that the Y321A mutation might disrupt hydrophobic packing involving the hinge region of the pre-stem motif. This alteration probably affects coordinated movement of the

pore-forming pre-stem motif and hinders the reorganization of the essential domains of VCC during oligomerization. To explore this hypothesis, we used Molecular Dynamics (MD) followed by network analysis to compare changes in mutant (MT) community structure with respect to WT.

6.3.3 Analysis of MD simulations

We performed three independent explicit water MD simulations each of 100 ns for both WT and modelled mutant Y321A (see section 6.2.1) to compare conformational dynamics between them. Within 100 ns simulation, the structures, as evident from RMSD profiles (Figures 6.2A and 6.2B), remained stable with no major structural changes in WT or Y321A Next, we compared ensemble average residue fluctuations using Residue Mean Squared Fluctuations (RMSF) of mutant and wild type VCC concatenated trajectories. As is evident in Figure 6.3, residues fluctuation of Y321A mutant is marginally higher than WT and high fluctuations are mostly localized in the loop regions of the protein (Figure 6.3). Specifically, Y321A has relatively high RMSF in membrane-proximal loops, cradle loop, loop$^{212\text{-}218}$ (composed of the residues 212-218 of the cytolysin domain), TP-loop (linker region between the β-trefoil and β-prism domain, Figure 6.1), and residue(s) spatially proximal to the site of mutation. Notably, part of the membrane-proximal loops and the loop$^{294\text{-}311}$ (composed of the residues 294-311, connecting two anti-parallel β-strands in the pre-stem motif) show decrease in the fluctuations in the Y321A mutant and collectively indicating possible global impact on mutating Y321A on protein motion. Importantly, spatially distant residues show changes, although small, indicating that the mutation of Y321A has long-range effect on the protein structure.

6.3.4 Essential Dynamics of WT and Y321A trajectories

Subsequently, we used Essential Dynamics (ED) of trajectories to examine changes in functional dynamics or relevant motions in Y321A structure in comparison to WT. For this, we performed Principal Component Analysis (PCA) based dimensionality reduction of ensembles and compared their individual dominant modes of large variance. The first three eigenvectors from PCA of Cα can be used to define the essential subspace, as these together constitute the

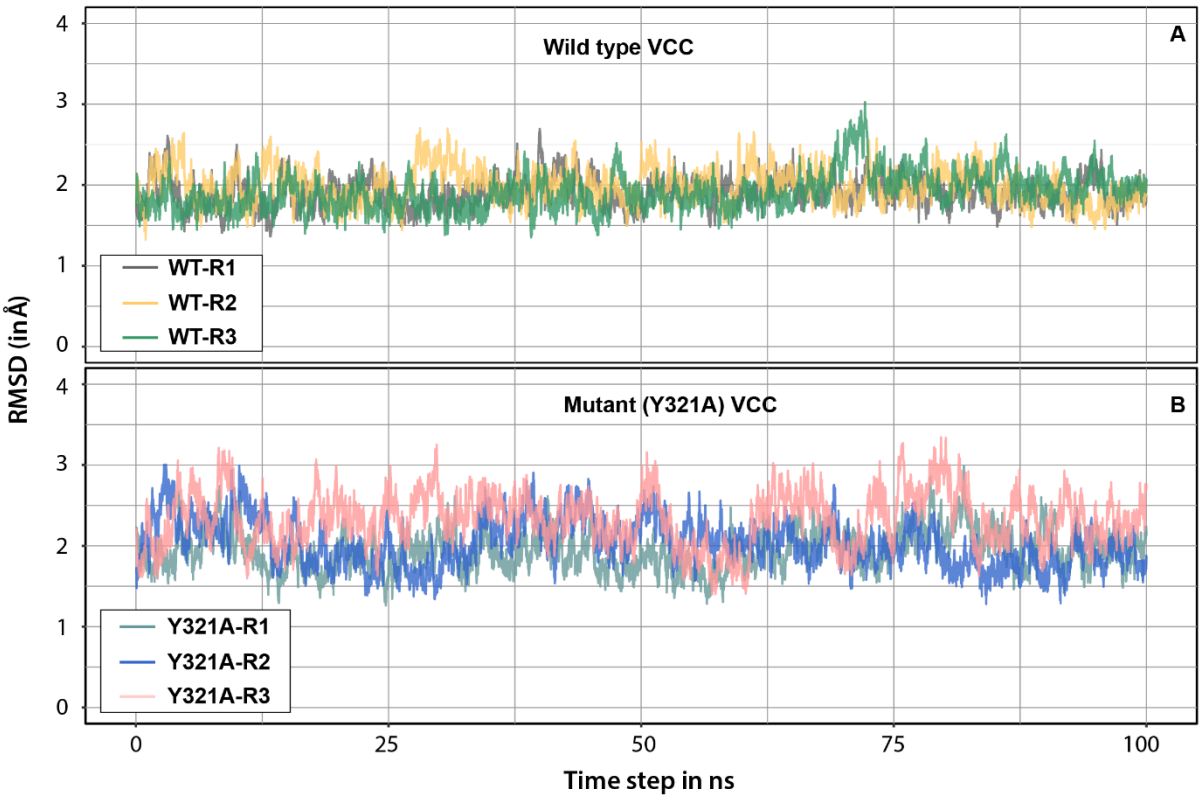


Figure 6.2. RMSD of WT and Y321A simulations during 100ns run. A) and B) shows RMSD of wild type VCC (A) and Y321A mutant (B) for 100 ns of production run in the MD simulations. The three replicates are marked as R1, R2, and R3.

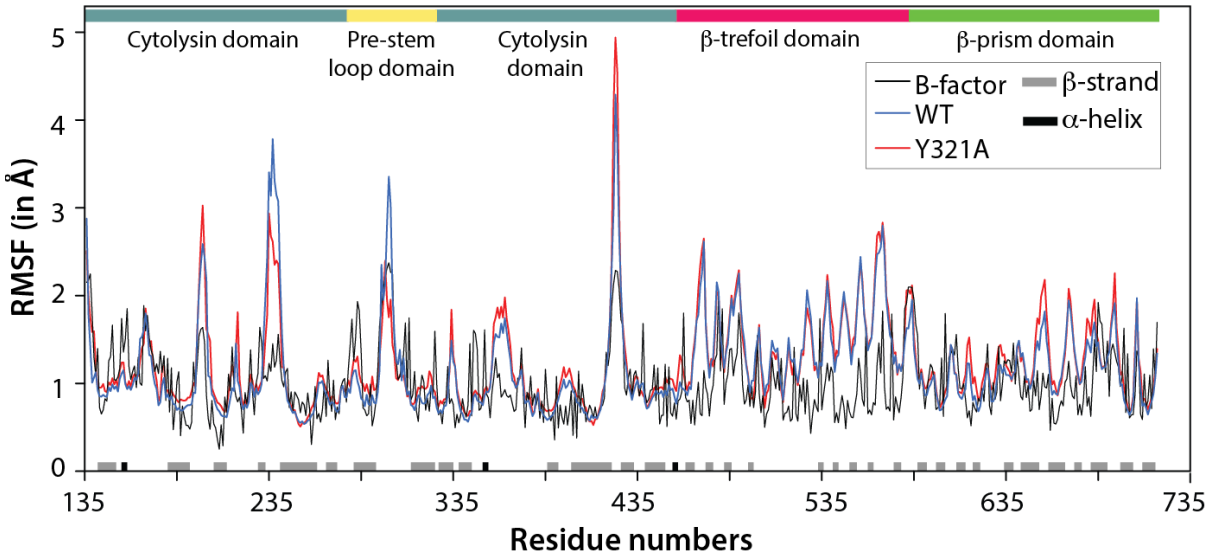


Figure 6.3: Root Mean Square Fluctuation (RMSF) between wild type VCC and Y321A mutant comparison. Additionally, B-factor of residues derived from the PDB structure (PDB ID: 1XEZ) is also shown. The secondary structures are represented above the x-axis. The panel has a color-bar drawn to represent the different structural domains.

total variance of 40% and 46% for wild type and Y321A mutant respectively (Figure 6.5A and 6.5C). Further, we delved into globally dominant motions and contributions of residues towards such functionally relevant motions by comparing the relative contribution of each residue towards the first two principal components (PC) in wild type VCC and Y321A shown in Figure 6.4. As can be seen (Figure 6.4), essential motions in PC1 for both WT and Y321A are mostly localized in β-trefoil domain and membrane binding loops, as these regions undergo conformational change during monomer to heptamer transition. In the mutant protein, there is slight change in the pattern of collective motion of the β-trefoil domain with reduced residue contribution of the 3_{10} helix residues of the loop$^{294\text{-}311}$, and the residues 579 to 583 in the linker joining β-Trefoil domain and β-Prism domains. In the second dominant motion (PC2), differences in Y321A to that of WT are apparent in subtle essential dynamics of domains. There are significant reduced residue contributions of residues from the omega loop (294-311), and TP-linker region. In the mutant protein, there are increased fluctuations in the β-prism domain, pre-stem motif and loop$^{212\text{-}218}$ (I-loop). In general, ED revealed that there is change in magnitude and residues contributing to the fluctuations of principal modes in the Y321A mutant, affecting the collective motions in the mutant protein. Moreover, collective motion variation localized to the linker between β-trefoil and β-prism domain indicates possible altered dynamics of the region that could hamper the carefully orchestrated conformational transition of β-prism domain affecting the oligomerization. These observations are also consistent with reduced RMSF of this region in mutant protein.

6.3.5 Rewired correlation couplings encompass inter protomer residues in Y321A

The observed differences in the second dominant motion prompted us to examine the role of Y321A and its mechanistic link. For this, we examined changes in the dynamic cross correlations of Cα residues between wild type and mutant protein to identify state specific residue dynamic couplings linking Y321. The absolute difference of dynamic cross-correlations between wild type and Y321A obtained from ensemble simulation is shown in the lower triangle on the right panel of Figure 6.6 and left panel shows line joining significantly varying correlated residue pairs of WT or mutant protein. This significant correlated residues pairs are those pairs having $|\Delta C_{ij}| \geq 0.4$ and $|C_{ij}| \geq 0.4$ either in the wild type VCC (red color) or in the mutant (blue color). Interestingly, most of changes in residue couplings are spatially distance from Y321 and residues near pre-stem region, as has been observed previously from RMSD/ED studies. Most of the wild type specific residue couplings are in β-trefoil domain

involving residue V580 and F581, which lie at domain boundary. The mutant specific dynamic residue correlations involve residues either within cytolysin domain, or it's coupling with β-trefoil and pre-stem domain. Interestingly, ~30% of residues involved in correlated motions in wild type or mutant protein lie at inter-protomer interaction interfaces. Thus, indicating mutation of Y321 leads to change in correlated motions, which might affect effective communication among domains as well as might affect protomer-protomer interactions. In order to further explore, how altered long distance interaction of pre-stem region in Y321A leads to affecting interactions of omega loop, linker between β-trefoil and β-prism, we combined dynamic couplings with network analysis to dissect apparent change in long range dynamics in mutant protein.

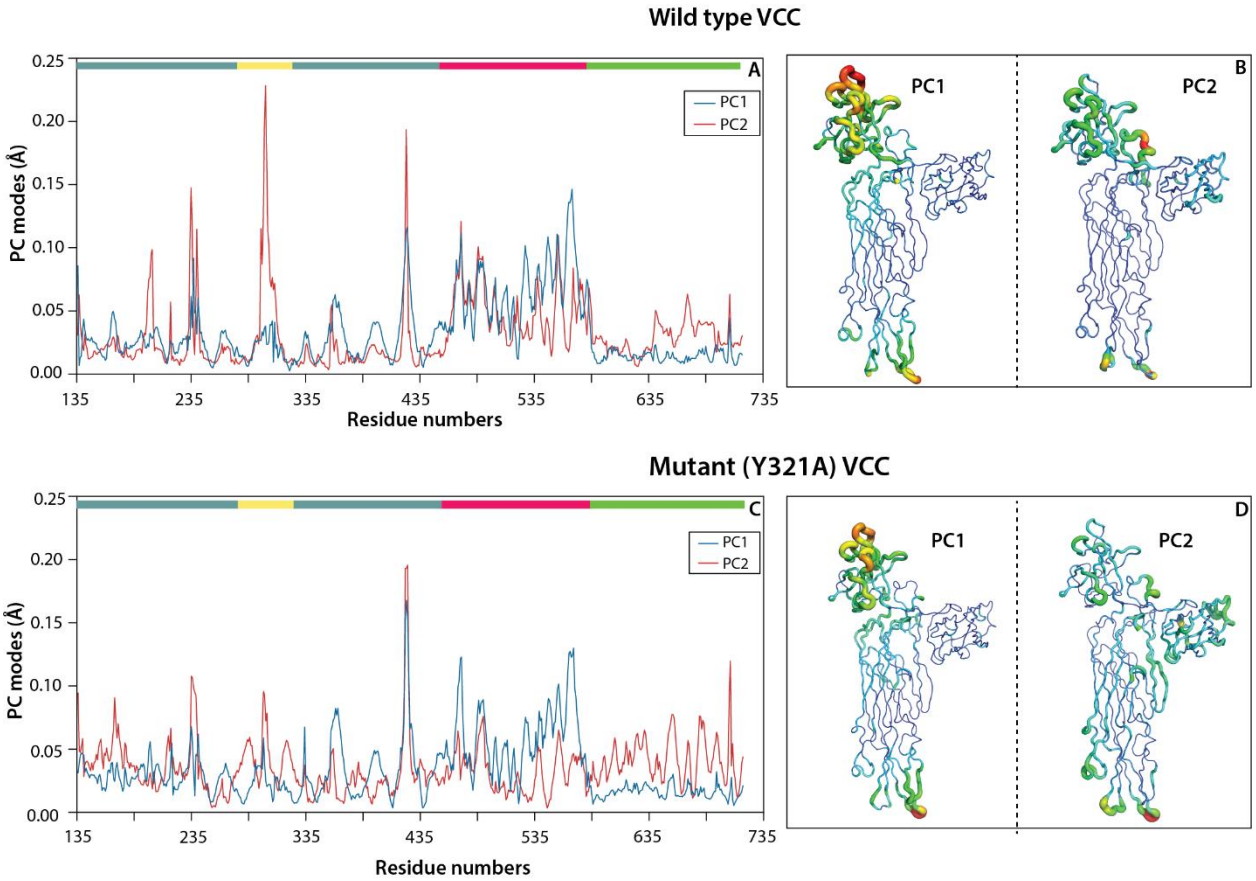


Figure 6.4: WT and Y321A's residue contribution to the first two principal components (PC1 and PC2). Top row represents wild type and Bottom row Y321A mutant; panel A) and C) represents WT and Y321A's residue contribution where PC1 and PC2 are colored blue and red respectively. Panel B) (Top row, WT) and D) (Bottom row, Y321A) show first (left) and second PCs (right) represented as residue displacement from the mean position, where tube thickness, along with the color from red-to-blue, shows the progressive decrease in fluctuations. The comparison of the PCs shows that there is change in the pattern of residue contribution between wild type VCC and Y321A mutant. The structural domains/motifs of VCC are shown as color bars in the various panels.

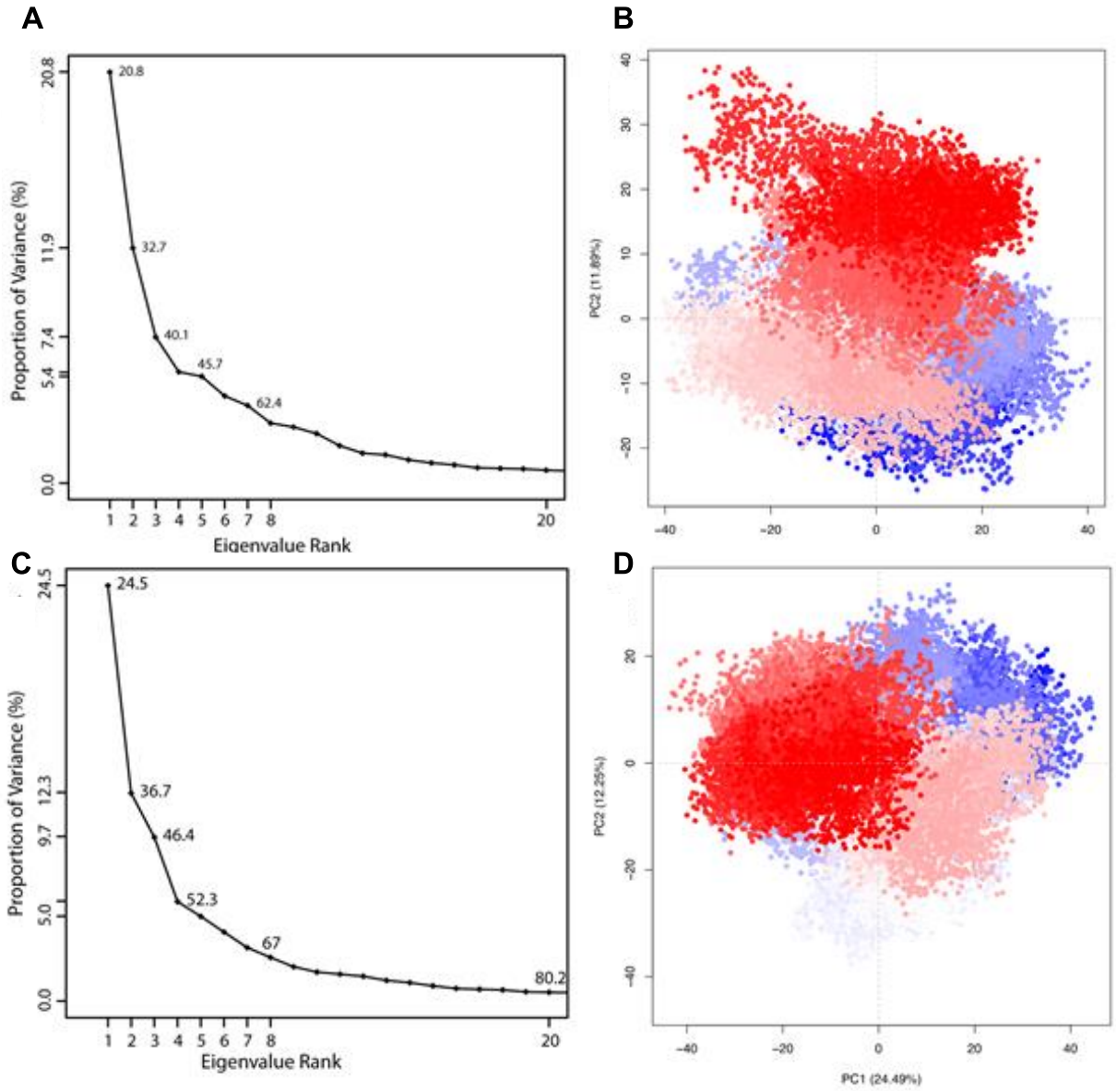


Figure 6.5: Essential dynamics of wild type VCC (Top row), and Y321A (bottom row) using PCA of Cα-atoms Cartesian coordinates. Panel A and C show the scree plot having cumulative contribution of the eigenvalues for WT and Y321A ensembles and Panel B and D show their trajectories projection on reduced dimension of first two principal components.

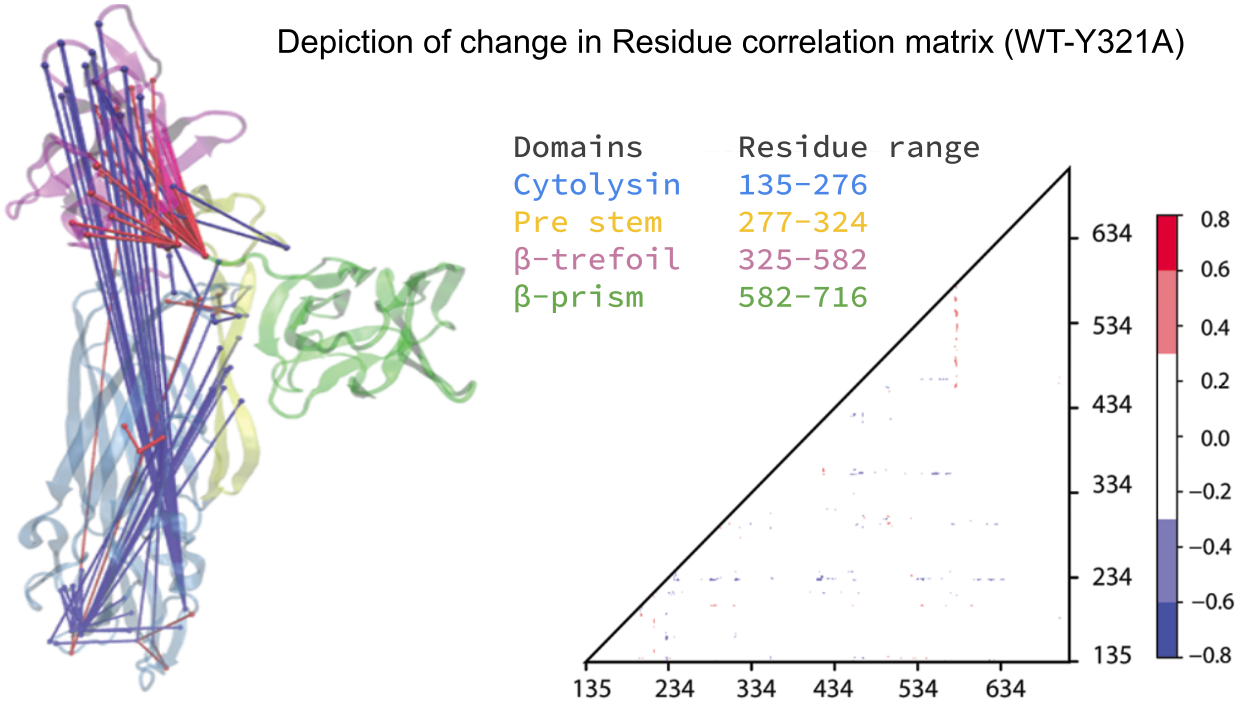


Figure 6.6: The difference in dynamic cross-correlations between wild type VCC and Y321A mutant. Difference shows the residues having predominant correlation in the wild type or the mutant states. These state-specific cross-correlations are shown in red for wild type VCC and blue for the mutant. Structure in left shows the dynamic couplings represented as lines connecting correlated residues having $|\Delta Cij| \geq 0.4$ and $|Cij| \geq 0.4$ either in the wild type VCC or in the mutant. The same color scheme as given above is followed to show the state-specific couplings.

6.3.6 Changes in structure of network community between WT and Y321A

In order to investigate change in communication dynamics affecting the phenotypic function of Y321A, we generated network correlation coupling derived networks for WT and Y321A ensembles as previously described method of Yao et al. (Yao et al. 2016), (see section 6.2.3, 6.2.4), wherein each residue is a node and edge between nodes is weight proportional to correlation observed in multiple replicates. Further, the network was partitioned into communities based on maximum modularity for wild type and mutant protein shown in Figures 6.7A and 6.7B respectively. Essentially, communities consist of correlated residues that represent highly connected substructures but loosely inter-connected substructures. The edge weight between communities shows strength of correlation. As shown in Figures 6.7A and 6.7B, there is remarkable change in community structure of Y321A in comparison to wild type. In the mutant protein the β-trefoil domain and β-prism domain become independent communities rather than contributing residues to other communities for maintaining dynamic

interactions (as these domains constitute of multiple communities in WT). Importantly, the cradle loop and the omega loop connecting antiparallel β-strands of pre-stem loops constitute a community (no. 2 in Figure 6.7A) and involved in community interactions with pre-stem motif, cytolysin, β-trefoil and β-prism domains in the WT. However, the same disintegrates in the mutant, with assimilation of most of cradle loop and omega loop in cytolysin domains and rest residues constitutes weak interacting communities (no. 2 and 7 in Figure 6.7B). Thus, indicating a significant loss of communicating networks in Y321A structure. The community partition of residues in WT and Y321A suggests that residues cradle loop and derived interactions probably plays important role in structural transition of VCC, because mutant protein shows a distinct loss of ability to make functional pore forming complex. This further also indicates alteration of carefully orchestrated allosteric mechanism required for structural transition of VCC as Y321A mutation also dissolved the distinct pre-stem communities and may have greater impact on its structural integrity loss, which could have been active with conformational change of β-prism domain on basis of their proximity and interface. We also performed MD simulations of Y321F mutant structure and showed that community structure is maintained like WT (Figure 6.7C) suggesting that interactions among residues required for conformational transition from monomer to heptamer is probably intact in Y321F.

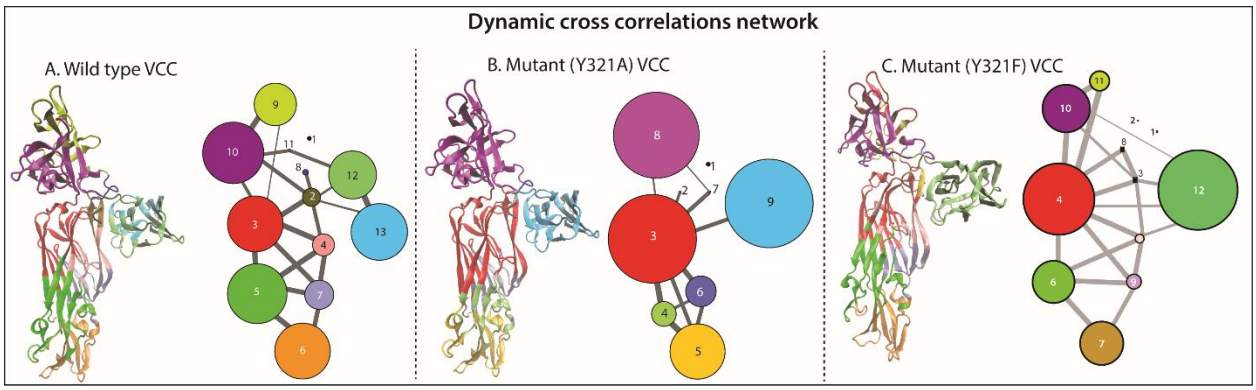


Figure 6.7: Community analysis of dynamic cross correlation network. The figure shows the community obtained at maximum modularity in the network constructed using cross-correlations (details are given in methods). Panels A and B show communities of wild type and mutant proteins respectively. Each panel has tertiary structure colored based on network community, which is shown in the right. The size of circle in network community represents number of members in the community. The thickness of edge in network corresponds to the maximum correlation observed between all edges between residues in two communities.

6.4 CONCLUSIONS

The computational analysis of WT and mutant (Y321A) ensembles provided insights into the understanding the role of Y321 residue in conformational rearrangement required during oligomerization of VCC. Initially, residue fluctuations measured using RMSF suggested noticeable changes in fluctuations of both spatially close and distant residues from the site of mutation. The same could be clearly observed in essential dynamics, which additionally showed the second largest eigenvector component encompassing motions in omega loop, β-trefoil and β-prism domains are greatly affected in Y321A in comparison to WT. Both of these observations were also seen in dynamic cross-correlation studies. Further using the correlations to perform network analysis showed a significant loss in the community structure of Y321A mutant suggesting a loss of community integrity of pre-stem region, omega, and cradle loops, which are assimilated in cytolysin or other domain communities. Our results of the MD simulations and associated analyses show the significance of residue interactions involving cradle loop, pre-stem, β-Trefoil domain and β-Prism domains. The mutation of Y321 residue, which lies in the hinge region leads to long-range defects in VCC. These probably affected communication between various structural components (domains) surrounding the pore-forming pre-stem motif, thereby compromising the coordinated reorganization of domains required during structural transition from monomer to oligomer. As this cytolysin and pre-stem domain architecture is conserved in beta pore forming toxins (Kaus et al. 2014; Olson and Gouaux 2005; Peraro and Van Der Goot 2016; Spaan, van Strijp, and Torres 2017), future studies will delve into hidden residue based allosteric networks orchestrating these domain conformational changes with aim to complementary design inhibitors to diminish their pathology.

License procured from self-co-authored publication (License Page (LP) 1-3, Figure license, LP 4-5 (Text License), LP 5-9 common "Terms and conditions")

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Aug 14, 2023

This Agreement between Mr. Paras Verma ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5607730161366
License date	Aug 14, 2023
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Molecular Microbiology
Licensed Content Title	Tyrosine in the hinge region of the pore-forming motif regulates oligomeric β-barrel pore formation by Vibrio cholerae cytolysin
Licensed Content Author	Kausik Chattopadhyay, Shashi Bhushan Pandit, Somnath Dutta, et al
Licensed Content Date	Nov 11, 2020
Licensed Content Volume	115
Licensed Content Issue	4

Licensed Content Pages	18
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	8
Will you be translating?	No
Title	Unraveling the Functional Implications of Y321 to A mutation in Vibrio cholerae cytolysin through Molecular Insights from MD Simulations and network analysis
Institution name	IISER Mohali
Expected presentation date	Nov 2023
Portions	8 figure, (7 from supplemenetary and 1 from main text)
Requestor Location	Mr. Paras Verma IISER Mohali Mohali, SASN, Punjab 140306 India Attn: Mr. Paras Verma
Publisher Tax ID	EU826007151

LP-2

Total 0.00 USD

Terms and Conditions

LP-3

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Jul 23, 2023

This Agreement between Mr. Paras Verma ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5594631501324
License date	Jul 23, 2023
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Molecular Microbiology
Licensed Content Title	Tyrosine in the hinge region of the pore-forming motif regulates oligomeric β-barrel pore formation by Vibrio cholerae cytolysin
Licensed Content Author	Kausik Chattopadhyay, Shashi Bhushan Pandit, Somnath Dutta, et al
Licensed Content Date	Nov 11, 2020
Licensed Content Volume	115
Licensed Content Issue	4
Licensed Content Pages	18
Type of use	Dissertation/Thesis

LP-4

Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Text extract
Number of Pages	9
Will you be translating?	No
Title	Unraveling the Functional Implications of Y321 to A mutation in Vibrio cholerae cytolysin through Molecular Insights from MD Simulations
Institution name	IISER Mohali
Expected presentation date	Nov 2023
Order reference number	THESIS_PH17060
Portions	Section 2.8, Figure 10, section 3, section 4.15 and supplementary material text involving details and analysis of MD simulations
Requestor Location	Mr. Paras Verma IISER Mohali Mohali, SASN, Punjab 140306 India Attn: Mr. Paras Verma
Publisher Tax ID	EU826007151
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.
- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the** STM Permissions Guidelines **only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.
- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or

to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.
- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.
- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.
- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.
- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.
- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com.

LP-9

Chapter 7

References

Agirre, E., A. J. Oldfield, N. Bellora, A. Segelle, and R. F. Luco. 2021. 'Splicing-associated chromatin signatures: a combinatorial and position-dependent role for histone marks in splicing definition', *Nat. Commun.*, 12: 1-16.

Alt, F. W., A. L. Bothwell, M. Knapp, E. Siden, E. Mather, M. Koshland, and D. Baltimore. 1980. 'Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends', *Cell*, 20: 293-301.

Amadei, A., M. A. Ceruso, and A. Di Nola. 1999. 'On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations', *Proteins*, 36: 419-24.

Aspden, Julie L, Edward WJ Wallace, and Nicola %J Cell Genomics Whiffin. 2023. 'Not all exons are protein coding: Addressing a common misconception', 3.

Baralle, Francisco E., and Jimena Giudice. 2017. 'Alternative splicing as a regulator of development and tissue identity', *Nat. Rev. Mol. Cell Biol.*, 18: 437-51.

Barbosa-Morais, Nuno L., Manuel Irimia, Qun Pan, Hui Y. Xiong, Serge Gueroussov, Leo J. Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Colak, Taehyung Kim, Christine M. Misquitta-Ali, Michael D. Wilson, Philip M. Kim, Duncan T. Odom, Brendan J. Frey, and Benjamin J. Blencowe. 2012. 'The evolutionary landscape of alternative splicing in vertebrate species', *Science*, 338: 1587-93.

Berget, S. M., C. Moore, and P. A. Sharp. 1977. 'Spliced segments at the 5' terminus of adenovirus 2 late mRNA', *Proc. Natl. Acad. Sci. U. S. A.*, 74: 3171-75.

Best, Robert B., Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. 2012. 'Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi 1$ and $\chi 2$ Dihedral Angles', *Journal of Chemical Theory and Computation*, 8: 3257-73.

Birzele, Fabian, Gergely Csaba, and Ralf Zimmer. 2008. 'Alternative splicing and protein structure evolution', *Nucleic Acids Res.*, 36: 550-58.

Black, Douglas L. 2003. 'Mechanisms of alternative pre-messenger RNA splicing', *Annu. Rev. Biochem.*, 72: 291-336.

Blamires, Sean J, Todd A Blackledge, and I-Min Tso. 2017. 'Physicochemical property variation in spider silk: ecology, evolution, and synthetic production', *Annual review of entomology*, 62: 443-60.

Blencowe, Benjamin. 2017. 'The Relationship between Alternative Splicing and Proteomic Complexity', *Trends in Biochemical Sciences*, 42: 407-08.

Blencowe, Benjamin J. 2006. 'Alternative splicing: new insights from global analyses', *Cell*, 126: 37-47.

Bratzel, Graham, and Markus J Buehler. 2012. 'Sequence-structure correlations in silk: Poly-Ala repeat of N. clavipes MaSp1 is naturally optimized at a critical length scale', *J Mech Behav Biomed Mater*, 7: 30-40.

Breitbart, R. E., A. Andreadis, and B. Nadal-Ginard. 1987. 'Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes', *Annu. Rev. Biochem.*, 56: 467-95.

Brockwell, David J., Emanuele Paci, Rebecca C. Zinober, Godfrey S. Beddard, Peter D. Olmsted, D. Alastair Smith, Richard N. Perham, and Sheena E. Radford. 2003. 'Pulling geometry defines the mechanical resistance of a beta-sheet protein', *Nat. Struct. Biol.*, 10: 731-37.

Brown, Cameron P, Alessandra D Whaite, Jennifer M MacLeod, Joanne Macdonald, and Federico Rosei. 2015. 'With great structure comes great functionality: Understanding and emulating spider silk', *Journal of Materials Research*, 30: 108-20.

Buehler, Markus J, and Yu Ching Yung. 2010. 'How protein materials balance strength, robustness, and adaptability', *HFSP journal*, 4: 26-40.

Buljan, Marija, Guilhem Chalancon, Sebastian Eustermann, Gunter P. Wagner, Monika Fuxreiter, Alex Bateman, and M. Madan Babu. 2012. 'Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks', *Mol. Cell*, 46: 871-83.

Bush, Stephen J., Lu Chen, Jaime M. Tovar-Corona, and Araxi O. Urrutia. 2017. 'Alternative splicing and the evolution of phenotypic novelty', *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 372.

Calarco, John A., Yi Xing, Mario Cáceres, Joseph P. Calarco, Xinshu Xiao, Qun Pan, Christopher Lee, Todd M. Preuss, and Benjamin J. Blencowe. 2007. 'Global analysis of alternative splicing differences between humans and chimpanzees', *Genes Dev.*, 21: 2963-75.

Castle, John C., Chaolin Zhang, Jyoti K. Shah, Amit V. Kulkarni, Auinash Kalsotra, Thomas A. Cooper, and Jason M. Johnson. 2008. 'Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines', *Nat. Genet.*, 40: 1416-25.

Cheng, Pin-Nan, Johnny D. Pham, and James S. Nowick. 2013. 'The supramolecular chemistry of β-sheets', *J. Am. Chem. Soc.*, 135: 5477-92.

Cheng, Yuan, Leng-Duei Koh, Dechang Li, Baohua Ji, Ming-Yong Han, and Yong-Wei Zhang. 2014. 'On the strength of β-sheet crystallites of Bombyx mori silk fibroin', *Journal of the Royal Society Interface*, 11: 20140305.

Churbanov, Alexander, Igor B. Rogozin, Vladimir N. Babenko, Hesham Ali, and Eugene V. Koonin. 2005. 'Evolutionary conservation suggests a regulatory function of AUG triplets in 5'-UTRs of eukaryotic genes', *Nucleic Acids Res.*, 33: 5512-20.

Climente-González, Héctor, Eduard Porta-Pardo, Adam Godzik, and Eduardo Eyras. 2017. 'The Functional Impact of Alternative Splicing in Cancer', *Cell Reports*, 20: 2215-26.

Cock, Peter JA, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, and Bartek %J Bioinformatics Wilczynski. 2009. 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', 25: 1422.

Consortium, Fantom, Riken Pmi the, Clst, Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, Michiel J. L. de Hoon, Vanja Haberle, Timo Lassmann, Ivan V. Kulakovskiy, Marina Lizio, Masayoshi Itoh, Robin Andersson, Christopher J. Mungall, Terrence F. Meehan, Sebastian Schmeier, Nicolas Bertin, Mette Jørgensen, Emmanuel Dimont, Erik Arner, Christian Schmidl, Ulf Schaefer, Yulia A. Medvedeva, Charles Plessy, Morana Vitezic, Jessica Severin, Colin A. Semple, Yuri Ishizu, Robert S. Young, Margherita Francescatto, Intikhab Alam, Davide Albanese, Gabriel M. Altschuler, Takahiro Arakawa, John A. C. Archer, Peter Arner, Magda Babina, Sarah Rennie, Piotr J. Balwierz, Anthony G. Beckhouse, Swati Pradhan-Bhatt, Judith A. Blake, Antje Blumenthal, Beatrice Bodega, Alessandro Bonetti, James Briggs, Frank Brombacher, A. Maxwell Burroughs, Andrea Califano, Carlo V. Cannistraci, Daniel Carbajo, Yun Chen, Marco Chierici, Yari Ciani, Hans C. Clevers, Emiliano Dalla, Carrie A. Davis,

Michael Detmar, Alexander D. Diehl, Taeko Dohi, Finn Drabløs, Albert S. B. Edge, Matthias Edinger, Karl Ekwall, Mitsuhiro Endoh, Hideki Enomoto, Michela Fagiolini, Lynsey Fairbairn, Hai Fang, Mary C. Farach-Carson, Geoffrey J. Faulkner, Alexander V. Favorov, Malcolm E. Fisher, Martin C. Frith, Rie Fujita, Shiro Fukuda, Cesare Furlanello, Masaaki Furino, Jun-Ichi Furusawa, Teunis B. Geijtenbeek, Andrew P. Gibson, Thomas Gingeras, Daniel Goldowitz, Julian Gough, Sven Guhl, Reto Guler, Stefano Gustincich, Thomas J. Ha, Masahide Hamaguchi, Mitsuko Hara, Matthias Harbers, Jayson Harshbarger, Akira Hasegawa, Yuki Hasegawa, Takehiro Hashimoto, Meenhard Herlyn, Kelly J. Hitchens, Shannan J. Ho Sui, Oliver M. Hofmann, Ilka Hoof, Furni Hori, Lukasz Huminiecki, Kei Iida, Tomokatsu Ikawa, Boris R. Jankovic, Hui Jia, Anagha Joshi, Giuseppe Jurman, Bogumil Kaczkowski, Chieko Kai, Kaoru Kaida, Ai Kaiho, Kazuhiro Kajiyama, Mutsumi Kanamori-Katayama, Artem S. Kasianov, Takeya Kasukawa, Shintaro Katayama, Sachi Kato, Shuji Kawaguchi, Hiroshi Kawamoto, Yuki I. Kawamura, Tsugumi Kawashima, Judith S. Kempfle, Tony J. Kenna, Juha Kere, Levon M. Khachigian, Toshio Kitamura, S. Peter Klinken, Alan J. Knox, Miki Kojima, Soichi Kojima, Naoto Kondo, Haruhiko Koseki, Shigeo Koyasu, Sarah Krampitz, Atsutaka Kubosaki, Andrew T. Kwon, Jeroen F. J. Laros, Weonju Lee, Andreas Lennartsson, Kang Li, Berit Lilje, Leonard Lipovich, Alan Mackay-Sim, Ri-Ichiroh Manabe, Jessica C. Mar, Benoit Marchand, Anthony Mathelier, Niklas Mejhert, Alison Meynert, Yosuke Mizuno, David A. de Lima Morais, Hiromasa Morikawa, Mitsuru Morimoto, Kazuyo Moro, Efthymios Motakis, Hozumi Motohashi, Christine L. Mummery, Mitsuyoshi Murata, Sayaka Nagao-Sato, Yutaka Nakachi, Fumio Nakahara, Toshiyuki Nakamura, Yukio Nakamura, Kenichi Nakazato, Erik van Nimwegen, Noriko Ninomiya, Hiromi Nishiyori, Shohei Noma, Shohei Noma, Tadasuke Noazaki, Soichi Ogishima, Naganari Ohkura, Hiroko Ohimiya, Hiroshi Ohno, Mitsuhiro Ohshima, Mariko Okada-Hatakeyama, Yasushi Okazaki, Valerio Orlando, Dmitry A. Ovchinnikov, Arnab Pain, Robert Passier, Margaret Patrikakis, Helena Persson, Silvano Piazza, James G. D. Prendergast, Owen J. L. Rackham, Jordan A. Ramilowski, Mamoon Rashid, Timothy Ravasi, Patrizia Rizzu, Marco Roncador, Sugata Roy, Morten B. Rye, Eri Saijyo, Antti Sajantila, Akiko Saka, Shimon Sakaguchi, Mizuho Sakai, Hiroki Sato, Suzana Savvi, Alka Saxena, Claudio Schneider, Erik A. Schultes, Gundula G. Schulze-Tanzil, Anita Schwegmann, Thierry Sengstag, Guojun Sheng, Hisashi Shimoji, Yishai Shimoni, Jay W. Shin, Christophe Simon, Daisuke Sugiyama, Takaai Sugiyama, Masanori Suzuki, Naoko Suzuki, Rolf K. Swoboda, Peter A. C. t Hoen, Michihira Tagami, Naoko Takahashi, Jun Takai, Hiroshi Tanaka, Hideki Tatsukawa, Zuotian Tatum, Mark Thompson, Hiroo Toyodo, Tetsuro Toyoda, Elvind Valen, Marc van de Wetering, Linda M. van den Berg, Roberto Verado, Dipti Vijayan, Ilya E. Vorontsov, Wyeth W. Wasserman, Shoko Watanabe, Christine A. Wells, Louise N. Winteringham, Ernst Wolvetang, Emily J. Wood, Yoko Yamaguchi, Masayuki Yamamoto, Misako Yoneda, Yohei Yonekura, Shigehiro Yoshida, Susan E. Zabierowski, Peter G. Zhang, Xiaobei Zhao, Silvia Zucchelli, Kim M. Summers, Harukazu Suzuki, Carsten O. Daub, Jun Kawai, Peter Heutink, Winston Hide, Tom C. Freeman, Boris Lenhard, Vladimir B. Bajic, Martin S. Taylor, Vsevolod J. Makeev, Albin Sandelin, David A. Hume, Piero Carninci, and Yoshihide Hayashizaki. 2014. 'A promoter-level mammalian expression atlas', *Nature*, 507: 462-70.

Contreras-Moreira, Bruno, Pall F. Jonsson, and Paul A. Bates. 2003. 'Structural context of exons in protein domains: implications for protein modelling and design', *J. Mol. Biol.*, 333: 1045-59.

Craik, C. S., S. Sprang, R. Fletterick, and W. J. Rutter. 1982. 'Intron-exon splice junctions map at protein surfaces', *Nature*, 299: 180-82.

Cranford, Steven W., Jan de Boer, Clemens van Blitterswijk, and Markus J. Buehler. 2013. 'Materiomics: an -omics approach to biomaterials research', *Adv. Mater.*, 25: 802-24.

Crawford, Dana C, Juan M Acuña, and Stephanie L Sherman. 2001. 'FMR1 and the fragile X syndrome: human genome epidemiology review', *Genetics in medicine*, 3: 359-71.

Cunningham, Fiona, James E. Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Olanrewaju Austine-Orimoloye, Andrey G. Azov, If Barnes, Ruth Bennett, Andrew Berry, Jyothish Bhai, Alexandra Bignell, Konstantinos Billis, Sanjay Boddu, Lucy Brooks, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Jose Gonzalez Martinez, Cristina Guijarro-Clarke, Arthur Gymer, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, José Carlos Marugán, Shamika Mohanan, Aleena Mushtaq, Marc Naven, Denye N. Ogeh, Anne Parker, Andrew Parton, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, José G. Pérez-Silva, William Stark, Emily Steed, Kyösti Sutinen, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Michal Szpak, Anja Thormann, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A. Walsh, Brandon Walts, Natalie Willhoft, Andrea Winterbottom, Elizabeth Wass, Marc Chakiachvili, Bethany Flint, Adam Frankish, Stefano Giorgetti, Leanne Haggerty, Sarah E. Hunt, Garth R. Iisley, Jane E. Loveland, Fergal J. Martin, Benjamin Moore, Jonathan M. Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J. Trevanion, Sarah Dyer, Peter W. Harrison, Kevin L. Howe, Andrew D. Yates, Daniel R. Zerbino, and Paul Flicek. 2022. 'Ensembl 2022', *Nucleic Acids Res.*, 50: D988-D95.

Darden, Tom, Darrin York, and Lee Pedersen. 1993. 'Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems', *The Journal of chemical physics*, 98: 10089-92.

Das, Debopriya, Tyson A. Clark, Anthony Schweitzer, Miki Yamamoto, Henry Marr, Josh Arribere, Simon Minovitsky, Alexander Poliakov, Inna Dubchak, John E. Blume, and John G. Conboy. 2007. 'A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing', *Nucleic Acids Res.*, 35: 4845-57.

De Conti, Laura, Marco Baralle, and Emanuele Buratti. 2013a. 'Exon and intron definition in pre-mRNA splicing', *Wiley Interdiscip. Rev. RNA*, 4: 49-60.

De Conti, Laura, Marco Baralle, and Emanuele %J Wiley Interdisciplinary Reviews: RNA Buratti. 2013b. 'Exon and intron definition in pre-mRNA splicing', 4: 49-60.

de Klerk, E , and P Hoen. 2015. 'Alternative mRNA transcription, processing, and translation: insights from RNA sequencing', *Trends Genet.*, 31: 128-39.

de Klerk, Eleonora, Andrea Venema, S. Yahya Anvar, Jelle J. Goeman, Ouhua Hu, Capucine Trollet, George Dickson, Johan T. den Dunnen, Silvère M. van der Maarel, Vered Raz, and Peter A. C. t Hoen. 2012. 'Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation', *Nucleic Acids Res.*, 40: 9089-101.

Derti, Adnan, Philip Garrett-Engele, Kenzie D. Macisaac, Richard C. Stevens, Shreedharan Sriram, Ronghua Chen, Carol A. Rohl, Jason M. Johnson, and Tomas Babak. 2012. 'A quantitative atlas of polyadenylation in five mammals', *Genome Res.*, 22: 1173-83.

Divina, Petr, Andrea Kvitkovicova, Emanuele Buratti, and Igor Vorechovsky. 2009. 'Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping', *Eur. J. Hum. Genet.*, 17: 759.

Du, Ning, Zhen Yang, Xiang Yang Liu, Yang Li, and Hong Yao Xu. 2011. 'Structural origin of the strain-hardening of spider silk', *Advanced Functional Materials*, 21: 772-78.

Dunker, A. Keith, Celeste J. Brown, J. David Lawson, Lilia M. Iakoucheva, and Zoran Obradović. 2002. 'Intrinsic disorder and protein function', *Biochemistry*, 41: 6573-82.

Eisoldt, Lukas, Andrew Smith, and Thomas Scheibel. 2011. 'Decoding the secrets of spider silk', *Materials Today*, 14: 80-86.

Erdős, Gábor, Mátyás Pajkos, and Zsuzsanna %J Nucleic acids research Dosztányi. 2021. 'IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation', 49: W297-W303.

Feller, Scott E, Yuhong Zhang, Richard W Pastor, and Bernard R Brooks. 1995. 'Constant pressure molecular dynamics simulation: the Langevin piston method', *The Journal of chemical physics*, 103: 4613-21.

Foissac, S., and M. Sammeth. 2007. 'ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets', *Nucleic acids research*, 35: W297-W99.

Forman, Julia R, Seema Qamar, Emanuele Paci, Richard N Sandford, and Jane Clarke. 2005. 'The remarkable mechanical strength of polycystin-1 supports a direct role in mechanotransduction', *Journal of molecular biology*, 349: 861-71.

Fossey, Stephen A, George Némethy, Kenneth D Gibson, and Harold A Scheraga. 1991. 'Conformational energy studies of β-sheets of model silk fibroin peptides. I. Sheets of poly (Ala-Gly) chains', *Biopolymers: Original Research on Biomolecules*, 31: 1529-41.

Fourie, Anne M., Fawn Coles, Veronica Moreno, and Lars Karlsson. 2003. 'Catalytic activity of ADAM8, ADAM15, and MDC-L (ADAM28) on synthetic peptide substrates and in ectodomain cleavage of CD23', *J. Biol. Chem.*, 278: 30469-77.

Fredericks, Alger M., Kamil J. Cygan, Brian A. Brown, and William G. Fairbrother. 2015. 'RNA-Binding Proteins: Splicing Factors and Disease', *Biomolecules*, 5: 893-909.

Fu, Xianguo, Dezhu Zheng, Juan Liao, Qingqin Li, Yuxiang Lin, Duo Zhang, Aizhen Yan, and Fenghua %J Molecular Medicine Reports Lan. 2015. 'Alternatively spliced products lacking exon 12 dominate the expression of fragile X mental retardation 1 gene in human tissues', 12: 1957-62.

Gao, Mu, David Craig, Olivier Lequin, Iain D Campbell, Viola Vogel, and Klaus Schulten. 2003. 'Structure and functional significance of mechanically unfolded fibronectin type III1 intermediates', *Proceedings of the National Academy of Sciences*, 100: 14784-89.

Gelly, Jean-Christophe, Hsuan-Yu Lin, Alexandre G. de Brevern, Trees-Juen Chuang, and Feng-Chi Chen. 2012. 'Selective constraint on human pre-mRNA splicing by protein structural properties', *Genome Biol. Evol.*, 4: 966-75.

Girvan, M., and M. E. Newman. 2002. 'Community structure in social and biological networks', *Proc Natl Acad Sci U S A*, 99: 7821-6.

Gracheva, Elena O., Julio F. Cordero-Morales, José A. González-Carcacía, Nicholas T. Ingolia, Carlo Manno, Carla I. Aranguren, Jonathan S. Weissman, and David Julius. 2011. 'Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats', *Nature*, 476: 88-91.

Grant, B. J., A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. Caves. 2006. 'Bio3d: an R package for the comparative analysis of protein structures', *Bioinformatics*, 22: 2695-6.

Grau-Bové, Xavier, Iñaki Ruiz-Trillo, and Manuel Irimia. 2018. 'Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture', *Genome Biol.*, 19: 135.

Grützmann, Konrad, Karol Szafranski, Martin Pohl, Kerstin Voigt, Andreas Petzold, and Stefan Schuster. 2014. 'Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study', *DNA Res.*, 21: 27-39.

Hahn, Matthew W., and Gregory A. Wray. 2002. 'The g-value paradox', *Evol. Dev.*, 4: 73-75.

Hall, Troii, Joseph W Leone, Joseph F Wiese, David W Griggs, Lyle E Pegg, Adele M Pauley, Alfredo G Tomasselli, and Marc D %J Bioscience reports Zack. 2009. 'Autoactivation of human ADAM8: a novel pre-processing step is required for catalytic activity', 29: 217-28.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. 2012. 'GENCODE: the reference human genome annotation for The ENCODE Project', *Genome Res.*, 22: 1760-74.

Hegyi, Hedi, Lajos Kalmar, Tamas Horvath, and Peter Tompa. 2011. 'Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder', *Nucleic Acids Res.*, 39: 1208-19.

Hong, Heedeok, Sangho Park, Ricardo H Flores Jiménez, Dennis Rinehart, and Lukas K %J Journal of the American Chemical Society Tamm. 2007. 'Role of aromatic side chains in the folding and thermodynamic stability of integral membrane proteins', 129: 8320-27.

Hong, Xin, Douglas G. Scofield, and Michael Lynch. 2006. 'Intron size, abundance, and distribution within untranslated regions of genes', *Mol. Biol. Evol.*, 23: 2392-404.

Howe, Kenneth James, Caroline M. Kane, and Manuel Ares, Jr. 2003. 'Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae', *RNA*, 9: 993-1006.

Hu Frisk, Junmei, Gunnar Pejler, Staffan Eriksson, and Liya Wang. 2022. 'Structural and functional analysis of human thymidylate kinase isoforms', *Nucleosides Nucleotides Nucleic Acids*, 41: 321-32.

Huang, Jing, and Alexander D. MacKerell, Jr. 2013. 'CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data', *J. Comput. Chem.*, 34: 2135-45.

Hughes, Thomas A., and Hugh J. M. Brady. 2005. 'Expression of axin2 is regulated by the alternative 5'-untranslated regions of its mRNA', *J. Biol. Chem.*, 280: 8581-88.

Humphrey, W., A. Dalke, and K. Schulten. 1996a. 'VMD: visual molecular dynamics', *J Mol Graph*, 14: 33-38.

Humphrey, William, Andrew Dalke, and Klaus Schulten. 1996b. 'VMD: visual molecular dynamics', *Journal of molecular graphics*, 14: 33-38.

Ichiye, T., and M. Karplus. 1991. 'Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations', *Proteins*, 11: 205-17.

Irvin-Wilson, Charletha V., and Gautam Chaudhuri. 2005. 'Alternative initiation and splicing in dicer gene expression in human breast cells', *Breast Cancer Res.*, 7: R563-9.

Jahn, Thomas R, O Sumner Makin, Kyle L Morris, Karen E Marshall, Pei Tian, Pawel Sikorski, and Louise C Serpell. 2010. 'The common architecture of cross-β amyloid', *Journal of molecular biology*, 395: 717-27.

James, Carissa C., and James W. Smyth. 2018. 'Alternative mechanisms of translation initiation: An emerging dynamic regulator of the proteome in health and disease', *Life Sci.*, 212: 138-44.

Jangi, Mohini, and Phillip A. Sharp. 2014. 'Building robust transcriptomes with master splicing factors', *Cell*, 159: 487-98.

Ji, Hong, Yinghua Zhang, Wei Zheng, Zheng Wu, Sunghou Lee, and Kathryn Sandberg. 2004. 'Translational regulation of angiotensin type 1a receptor expression and signaling by upstream AUGs in the 5' leader sequence', *J. Biol. Chem.*, 279: 45322-28.

Ji, Zhe. 2018. 'RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling', *Curr. Protoc. Mol. Biol.*, 124: e67.

Ji, Zhe, Ju Youn Lee, Zhenhua Pan, Bingjun Jiang, and Bin Tian. 2009. 'Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development', *Proc. Natl. Acad. Sci. U. S. A.*, 106: 7028-33.

Ji, Zhe, Wenting Luo, Wencheng Li, Mainul Hoque, Zhenhua Pan, Yun Zhao, and Bin Tian. 2011. 'Transcriptional activity regulates alternative cleavage and polyadenylation', *Mol. Syst. Biol.*, 7: 534.

Johansson, Jan, and Anna Rising. 2021. 'Doing What Spiders Cannot—A Road Map to Supreme Artificial Silk Fibers', *ACS nano*, 15: 1952-59.

Jones, D. T. 1999. 'Protein secondary structure prediction based on position-specific scoring matrices', *J. Mol. Biol.*, 292: 195-202.

Kaus, Katherine, Jeffrey W Lary, James L Cole, and Rich %J Journal of molecular biology Olson. 2014. 'Glycan specificity of the Vibrio vulnificus hemolysin lectin outlines evolutionary history of membrane targeting by a toxin family', 426: 2800-12.

Keren, Hadas, Galit Lev-Maor, and Gil Ast. 2010. 'Alternative splicing and evolution: diversification, exon definition and function', *Nat. Rev. Genet.*, 11: 345-55.

Keten, Sinan, Zhiping Xu, Britni Ihle, and Markus J. Buehler. 2010. 'Nanoconfinement controls stiffness, strength and mechanical toughness of β-sheet crystals in silk', *Nature Materials*, 9: 359-67.

Kim, Pora, Mengyuan Yang, Ke Yiya, Weiling Zhao, and Xiaobo %J Nucleic acids research Zhou. 2020. 'ExonSkipDB: functional annotation of exon skipping event in human', 48: D896-D907.

Kim, Yoonjung, Hyunsung Choi, Inchul Baek, and Sungsoo Na. 2020. 'Spider silk with weaker bonding resulting in higher strength and toughness through progressive unfolding and load transfer', *Journal of the mechanical behavior of biomedical materials*, 108: 103773.

Kim, Yoonjung, Myeongsang Lee, Hyunsung Choi, Inchul Baek, Jae In Kim, and Sungsoo Na. 2018. 'Mechanical features of various silkworm crystalline considering hydration effect via molecular dynamics simulations', *Journal of Biomolecular Structure Dynamics.*, 36: 1360-68.

Kluge, Jonathan A, Olena Rabotyagova, Gary G Leisk, and David L Kaplan. 2008. 'Spider silks and their applications', *Trends in biotechnology*, 26: 244-51.

Knolle, Martin D., and Caroline A. Owen. 2009. 'ADAM8: a new therapeutic target for asthma', *Expert Opin. Ther. Targets*, 13: 523-40.

Kochetov, Alex V. 2008. 'Alternative translation start sites and hidden coding potential of eukaryotic mRNAs', *Bioessays*, 30: 683-91.

Koren, Eli, Galit Lev-Maor, and Gil Ast. 2007. 'The emergence of alternative 3' and 5' splice site exons from constitutive exons', *PLoS Comput. Biol.*, 3: e95.

Kumar, Bhupesh, Ashish Thakur, Biswajit Panda, and Kamal P Singh. 2013. 'Optically probing torsional superelasticity in spider silks', *J Applied Physics Letters*, 103: 201910.

Lambert, Matthew J, Kyle G Olsen, and Cynthia D %J Gene Cooper. 2014. 'Gene duplication followed by exon structure divergence substitutes for alternative splicing in zebrafish', 546: 271-76.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-

Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and Consortium International Human Genome Sequencing. 2001. 'Initial sequencing and analysis of the human genome', *Nature*, 409: 860-921.

Landry, Josette-Renée, Dixie L. Mager, and Brian T. Wilhelm. 2003. 'Complex controls: the role of alternative promoters in mammalian genomes', *Trends Genet.*, 19: 640-48.

Lange, O. F., and H. Grubmuller. 2006. 'Generalized correlation for biomolecular dynamics', *Proteins*, 62: 1053-61.

Lee, Brian T., Galt P. Barber, Anna Benet-Pagès, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S. Hinrichs, Christopher M. Lee, Pranav Muthuraman, Luis R. Nassar, Beagan Nguy, Tiana Pereira, Gerardo Perez, Brian J. Raney, Kate R. Rosenbloom, Daniel Schmelter, Matthew L. Speir, Brittney D. Wick, Ann S. Zweig, David Haussler, Robert M. Kuhn, Maximilian Haeussler, and W. James Kent. 2022. 'The UCSC Genome Browser database: 2022 update', *Nucleic Acids Res.*, 50: D1115-D22.

Leppek, Kathrin, Rhiju Das, and Maria Barna. 2018. 'Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them', *Nat. Rev. Mol. Cell Biol.*, 19: 158-74.

Li, Yang I., Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. 2016. 'RNA splicing is a primary link between genetic variation and disease', *Science*, 352: 600-04.

Lin, Lan, Shihao Shen, Peng Jiang, Seiko Sato, Beverly L. Davidson, and Yi Xing. 2010. 'Evolution of alternative splicing in primate brain transcriptomes', *Hum. Mol. Genet.*, 19: 2958-73.

Liu, Dabiao, Longteng Yu, Yuming He, Kai Peng, Jie Liu, Juan Guan, and DJ Dunstan. 2017. 'Peculiar torsion dynamical response of spider dragline silk', *Applied Physics Letters*, 111: 013701.

Liu, Yan, Kevin Marks, Glenn S. Cowley, Julian Carretero, Qingsong Liu, Thomas J. F. Nieland, Chunxiao Xu, Travis J. Cohoon, Peng Gao, Yong Zhang, Zhao Chen, Abigail B. Altabef, Jeremy H. Tchaicha, Xiaoxu Wang, Sung Choe, Edward M. Driggers, Jianming Zhang, Sean T. Bailey, Norman E. Sharpless, D. Neil Hayes, Nirali M. Patel, Pasi A. Janne, Nabeel Bardeesy, Jeffrey A. Engelman, Brendan D. Manning, Reuben J. Shaw, John M. Asara, Ralph Scully, Alec Kimmelman, Lauren A. Byers, Don L. Gibbons, Ignacio I. Wistuba, John V. Heymach, David J. Kwiatkowski, William Y. Kim, Andrew L. Kung, Nathanael S. Gray, David E. Root, Lewis C. Cantley, and Kwok-Kin Wong. 2013. 'Metabolic and functional genomic studies identify deoxythymidylate kinase as a target in LKB1-mutant lung cancer', *Cancer Discov.*, 3: 870-79.

Löffler, M., E. A. Carrey, and E. Zameitat. 2018. 'New perspectives on the roles of pyrimidines in the central nervous system', *Nucleosides Nucleotides Nucleic Acids*, 37: 290-306.

Long, Manyuan, and Michael Deutsch. 1999. 'Intron—exon structures of eukaryotic model organisms', *Nucleic acids research*, 27: 3219-28.

Lu, Hui, Barry Isralewitz, Andre Krammer, Viola Vogel, and Klaus Schulten. 1998. 'Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation', *Biophysical journal*, 75: 662-71.

Lynch, Michael, Douglas G. Scofield, and Xin Hong. 2005. 'The evolution of transcription-initiation sites', *Mol. Biol. Evol.*, 22: 1137-46.

Macknight, Richard, Meg Duroux, Rebecca Laurie, Paul Dijkwel, Gordon Simpson, and Caroline Dean. 2002. 'Functional significance of the alternative transcript processing of the Arabidopsis floral promoter FCA', *Plant Cell*, 14: 877-88.

Madeira, Fábio, Matt Pearce, Adrian R. N. Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez. 2022a. 'Search and sequence analysis tools services from EMBL-EBI in 2022', *Nucleic Acids Res.*, 50: W276-W79.

Madeira, Fábio, Matt Pearce, Adrian RN Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo %J Nucleic acids research Lopez. 2022b. 'Search and sequence analysis tools services from EMBL-EBI in 2022', 50: W276-W79.

Mallarino, Ricardo, Tess A. Linden, Catherine R. Linnen, and Hopi E. Hoekstra. 2016. 'The role of isoforms in the evolution of cryptic coloration in Peromyscus mice', *bioRxiv*.

Marasco, Luciano E, and Alberto R %J Nature Reviews Molecular Cell Biology Kornblihtt. 2023. 'The physiology of alternative splicing', 24: 242-54.

Marasco, Luciano E., and Alberto R. Kornblihtt. 2022. 'The physiology of alternative splicing', *Nat. Rev. Mol. Cell Biol.*, 24: 242-54.

Martelli, Pier L., Mattia D'Antonio, Paola Bonizzoni, Tiziana Castrignanò, Anna M. D'Erchia, Paolo D'Onorio De Meo, Piero Fariselli, Michele Finelli, Flavio Licciulli, Marina Mangiulli, Flavio Mignone, Giulio Pavesi, Ernesto Picardi, Raffaella Rizzi, Ivan Rossi, Alessio Valletti, Andrea Zauli, Federico Zambelli, Rita Casadio, and Graziano Pesole. 2011. 'ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing', *Nucleic Acids Res.*, 39: D80-5.

Matera, A. Gregory, and Zefeng Wang. 2014. 'A day in the life of the spliceosome', *Nat. Rev. Mol. Cell Biol.*, 15: 108-21.

Mayen, Juan Francisco Carrascoza, Alexandru Lupan, Ciprian Cosar, Attila-Zsolt Kun, and Radu Silaghi-Dumitrescu. 2015. 'On the roles of the alanine and serine in the β-sheet structure of fibroin', *Biophysical Chemistry*, 197: 10-17.

Mayr, Christine %J Annual review of genetics. 2017. 'Regulation by 3′-untranslated regions', 51: 171-94.

Mignone, Flavio, Carmela Gissi, Sabino Liuni, and Graziano Pesole. 2002. 'Untranslated regions of mRNAs', *Genome Biol.*, 3: REVIEWS0004.

Milles, Lukas F., Klaus Schulten, Hermann E. Gaub, and Rafael C. Bernardi. 2018. 'Molecular mechanism of extreme mechanostability in a pathogen adhesin', *Science*, 359: 1527-33.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, and Lorna J %J Nucleic acids research Richardson. 2021. 'Pfam: The protein families database in 2021', 49: D412-D19.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. 2021. 'Pfam: The protein families database in 2021', *Nucleic Acids Res.*, 49: D412-D19.

Modrek, B., A. Resch, C. Grasso, and C. Lee. 2001. 'Genome-wide detection of alternative splicing in expressed sequences of human genes', *Nucleic Acids Res.*, 29: 2850-59.

Modrek, Barmak, and Christopher Lee. 2002. 'A genomic view of alternative splicing', *Nat. Genet.*, 30: 13-19.

Modrek, Barmak, and Christopher J. Lee. 2003. 'Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss', *Nat. Genet.*, 34: 177-80.

Mondal, Anish Kumar, and Kausik Chattopadhyay. 2019. 'Taking toll on membranes: curious cases of bacterial β-barrel pore-forming toxins', *Biochemistry*, 59: 163-70.

Mondal, Anish Kumar, Paras Verma, Nayanika Sengupta, Somnath Dutta, Shashi Bhushan Pandit, and Kausik %J Molecular microbiology Chattopadhyay. 2021. 'Tyrosine in the hinge region of the pore-forming motif regulates oligomeric β-barrel pore formation by Vibrio cholerae cytolysin', 115: 508-25.

Moriguchi, Tetsuo, Seiichi Urushiyama, Naoki Hisamoto, Shun-Ichiro Iemura, Shinichi Uchida, Tohru Natsume, Kunihiro Matsumoto, and Hiroshi Shibuya. 2005. 'WNK1 regulates phosphorylation of cation-chloride-coupled cotransporters via the STE20-related kinases, SPAK and OSR1', *J. Biol. Chem.*, 280: 42685-93.

Murillo-de-Ozores, Adrián Rafael, Alejandro Rodríguez-Gama, Héctor Carbajal-Contreras, Gerardo Gamba, and María Castañeda-Bueno. 2021. 'WNK4 kinase: from structure to physiology', *Am. J. Physiol. Renal Physiol.*, 320: F378-F403.

Nagasaki, Hideki, Masanori Arita, Tatsuya Nishizawa, Makiko Suwa, and Osamu Gotoh. 2006. 'Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns', *Bioinformatics*, 22: 1211-16.

Ndlovu, Hlengisizwe, Alison E. Ashcroft, Sheena E. Radford, and Sarah A. Harris. 2012. 'Effect of sequence variation on the mechanical response of amyloid fibrils probed by steered molecular dynamics simulation', *Biophys. J.*, 102: 587-96.

Ndlovu, Hlengisizwe., Alison E. Ashcroft, Sheena E. Radford, and Sarah A. Harris. 2013. 'Molecular dynamics simulations of mechanical failure in polymorphic arrangements of amyloid fibrils containing structural defects', *Beilstein Journal of Nanotechnology*, 4: 429-40.

Newton, Derek C., Sian C. Bevan, Stephen Choi, G. Brett Robb, Adam Millar, Yang Wang, and Philip A. Marsden. 2003. 'Translational regulation of human neuronal nitric-oxide synthase by an alternatively spliced 5'-untranslated region leader exon', *J. Biol. Chem.*, 278: 636-44.

Ni, Ting, Yanqin Yang, Dina Hafez, Wenjing Yang, Kurtis Kiesewetter, Yoshi Wakabayashi, Uwe Ohler, Weiqun Peng, and Jun Zhu. 2013. 'Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy', *BMC Genomics*, 14: 615.

Nilsen, Timothy W. 2003. 'The spliceosome: the most complex macromolecular machine in the cell?', *Bioessays*, 25: 1147-49.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. 'Expansion of the eukaryotic proteome by alternative splicing', *Nature*, 463: 457-63.

Numata, Keiji, Hiroyasu Masunaga, Takaaki Hikima, Sono Sasaki, Kazuhide Sekiyama, and Masaki Takata. 2015. 'Use of extension-deformation-based crystallisation of silk fibres to differentiate their functions in nature', *Soft Matter*, 11: 6335-42.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. 2016. 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Res.*, 44: D733-45.

Olson, Rich, and Eric %J Journal of molecular biology Gouaux. 2005. 'Crystal structure of the Vibrio cholerae cytolysin (VCC) pro-toxin and its assembly into a heptameric transmembrane pore', 350: 997-1016.

Ozsolak, Fatih, Philipp Kapranov, Sylvain Foissac, Sang Woo Kim, Elane Fishilevich, A. Paula Monaghan, Bino John, and Patrice M. Milos. 2010. 'Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation', *Cell*, 143: 1018-29.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat. Genet.*, 40: 1413-15.

Park, Eun-Hee, Joseph M. Lee, and Jerry Pelletier. 2006. 'The Tie2 5' untranslated region is inhibitory to 5' end-mediated translation initiation', *FEBS Lett.*, 580: 1309-19.

Peraro, Matteo Dal, and F Gisou %J Nature reviews microbiology Van Der Goot. 2016. 'Pore-forming toxins: ancient, but never really out of fashion', 14: 77-92.

Pettersen, Eric F, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. 2004. 'UCSF Chimera—a visualization system for exploratory research and analysis', *Journal of computational chemistry*, 25: 1605-12.

Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. 2005. 'Scalable molecular dynamics with NAMD', *J Comput Chem*, 26: 1781-802.

Phillips, James C., Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. 2005. 'Scalable molecular dynamics with NAMD', *J. Comput. Chem.*, 26: 1781-802.

Piwowar, Monika, Mateusz Banach, Leszek Konieczny, and Irena Roterman. 2013. 'Structural role of exon-coded fragment of polypeptide chains in selected enzymes', *J. Theor. Biol.*, 337: 15-23.

Porter, D, J Guan, and F Vollrath. 2013. 'Spider silk: super material or thin fibre?', *Advanced Materials*, 25: 1275-79.

Proudfoot, Nick J. 2016. 'Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut', *Science*, 352: aad9926.

Prudêncio, Pedro, Rosina Savisaar, Kenny Rebelo, Rui Gonçalo Martinho, and Maria %J Rna Carmo-Fonseca. 2022. 'Transcription and splicing dynamics during early Drosophila development', 28: 139-61.

Rai, Anand Kumar, and Kausik Chattopadhyay. 2015. 'Revisiting the membrane interaction mechanism of a membrane-damaging β-barrel pore-forming toxin Vibrio cholerae cytolysin', 97: 1051-62.

Reixachs-Solé, Marina, and Eduardo Eyras. 2022. 'Uncovering the impacts of alternative splicing on the proteome with current omics techniques', *Wiley Interdiscip. Rev. RNA*, 13: e1707.

Ren, Fanggang, Na Zhang, Lan Zhang, Eric Miller, and Jeffrey J. Pu. 2020. 'Alternative Polyadenylation: a new frontier in post transcriptional regulation', *Biomark Res*, 8: 67.

Resch, Alissa M., Aleksey Y. Ogurtsov, Igor B. Rogozin, Svetlana A. Shabalina, and Eugene V. Koonin. 2009. 'Evolution of alternative and constitutive regions of mammalian 5'UTRs', *BMC Genomics*, 10: 1-14.

Resch, Alissa, Yi Xing, Barmak Modrek, Michael Gorlick, Robert Riley, and Christopher Lee. 2004. 'Assessing the impact of alternative splicing on domain interactions in the human proteome', *J. Proteome Res.*, 3: 76-83.

Reyes, Alejandro, and Wolfgang Huber. 2018. 'Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues', *Nucleic Acids Res.*, 46: 582-92.

Reynolds, K., A. M. Zimmer, and A. Zimmer. 1996. 'Regulation of RAR beta 2 mRNA expression: evidence for an inhibitory peptide encoded in the 5'-untranslated region', *J. Cell Biol.*, 134: 827-35.

Rising, Anna. 2014. 'Controlled assembly: A prerequisite for the use of recombinant spider silk in regenerative medicine?', *Acta biomaterialia*, 10: 1627-31.

Rodriguez, Jose Manuel, Paolo Maietta, Iakes Ezkurdia, Alessandro Pietrelli, Jan-Jaap Wesselink, Gonzalo Lopez, Alfonso Valencia, and Michael L. Tress. 2013. 'APPRIS: annotation of principal and alternative splice isoforms', *Nucleic Acids Res.*, 41: D110-7.

Romagnoli, Mathilde, Nora D. Mineva, Michael Polmear, Catharina Conrad, Srimathi Srinivasan, Delphine Loussouarn, Sophie Barillé-Nion, Irene Georgakoudi, Áine Dagg, Enda W. McDermott, Michael J. Duffy, Patricia M. McGowan, Uwe Schlomann, Maddy Parsons, Jörg W. Bartsch, and Gail E. Sonenshein. 2014. 'ADAM8 expression in invasive breast cancer promotes tumor dissemination and metastasis', *EMBO Mol. Med.*, 6: 278-94.

Römer, Lin, and Thomas Scheibel. 2008. 'The elaborate structure of spider silk: structure and function of a natural high performance fiber', *Prion*, 2: 154-61.

Roy, Ambrish, Alper Kucukural, and Yang Zhang. 2010. 'I-TASSER: a unified platform for automated protein structure and function prediction', *Nat. Protoc.*, 5: 725-38.

Saka, Hector Alex, Carla Bidinost, Claudia Sola, Pablo Carranza, Cesar Collino, Susana Ortiz, Jose Ricardo Echenique, and José Luis %J Microbial pathogenesis Bocco. 2008. 'Vibrio cholerae cytolysin is essential for high enterotoxicity and apoptosis induction produced by a cholera toxin gene-negative V. cholerae non-O1, non-O139 strain', 44: 118-28.

Sammeth, Michael, Sylvain Foissac, and Roderic Guigó. 2008. 'A general definition and nomenclature for alternative splicing events', *PLoS Comput. Biol.*, 4: e1000147.

San-Cristobal, Pedro, José Ponce-Coria, Norma Vázquez, Norma A. Bobadilla, and Gerardo Gamba. 2008. 'WNK3 and WNK4 amino-terminal domain defines their effect on the renal Na+-Cl- cotransporter', *Am. J. Physiol. Renal Physiol.*, 295: F1199-206.

Santoro, Michael R, Steven M Bray, and Stephen T %J Annual Review of Pathology: Mechanisms of Disease Warren. 2012. 'Molecular mechanisms of fragile X syndrome: a twenty-year perspective', 7: 219-45.

Schad, Eva, Peter Tompa, and Hedi Hegyi. 2011. 'The relationship between proteome size, structural disorder and organism complexity', *Genome Biol.*, 12: R120.

Schlomann, U., S. Rathke-Hartlieb, S. Yamamoto, H. Jockusch, and J. W. Bartsch. 2000. 'Tumor necrosis factor alpha induces a metalloprotease-disintegrin, ADAM8 (CD 156): implications for neuron-glia interactions during neurodegeneration', *J. Neurosci.*, 20: 7964-71.

Sebestyén, Endre, Michał Zawisza, and Eduardo Eyras. 2015. 'Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer', *Nucleic Acids Res.*, 43: 1345-56.

Sethi, A., J. Eargle, A. A. Black, and Z. Luthey-Schulten. 2009. 'Dynamical networks in tRNA:protein complexes', *Proc Natl Acad Sci U S A*, 106: 6620-5.

Shabalina, Svetlana A., Aleksey Y. Ogurtsov, Nikolay A. Spiridonov, and Eugene V. Koonin. 2014. 'Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals', *Nucleic Acids Res.*, 42: 7132-44.

Shabalina, Svetlana A., Alexey N. Spiridonov, Nikolay A. Spiridonov, and Eugene V. Koonin. 2010. 'Connections between alternative transcription and alternative splicing in mammals', *Genome Biol. Evol.*, 2: 791-99.

Shepard, Peter J., Eun- A. Choi, Jente Lu, Lisa A. Flanagan, Klemens J. Hertel, and Yongsheng Shi. 2011. 'Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq', *RNA*, 17: 761-72.

Shyu, Ann-Bin, Miles F. Wilkinson, and Ambro van Hoof. 2008. 'Messenger RNA regulation: to translate or to degrade', *EMBO J.*, 27: 471-81.

Singh, Pooja, and Ehsan Pashay Ahi. 2022. 'The importance of alternative splicing in adaptive evolution', *Mol. Ecol.*, 31: 1928-38.

Sittler, Annie, Didier Devys, Chantal Weber, and Jean-Louis Mandel. 1996. 'Alternative Splicing of Exon 14 Determines Nuclear or Cytoplasmic Localisation of FMR1 Protein Isoforms', *Human Molecular Genetics*, 5: 95-102.

Skjærven, Lars, Xin-Qiu Yao, Guido Scarabelli, and Barry J Grant. 2014. 'Integrating protein structural dynamics and evolutionary analysis with Bio3D', *BMC Bioinformatics*, 15: 1-11.

Song, Kyu Young, Cheol Kyu Hwang, Chun Sung Kim, Hack Sun Choi, Ping-Yee Law, Li-Na Wei, and Horace H. Loh. 2007. 'Translational repression of mouse mu opioid receptor expression via leaky scanning', *Nucleic Acids Res.*, 35: 1501-13.

Spaan, András N, Jos AG van Strijp, and Victor J %J Nature Reviews Microbiology Torres. 2017. 'Leukocidins: staphylococcal bi-component pore-forming toxins find their receptors', 15: 435-47.

Srinivasan, Srimathi, Mathilde Romagnoli, Andrew Bohm, and Gail E Sonenshein. 2014. 'N-glycosylation regulates ADAM8 processing and activation', *Journal of Biological Chemistry*, 289: 33676-88.

Stamm, Stefan, Shani Ben-Ari, Ilona Rafalska, Yesheng Tang, Zhaiyi Zhang, Debra Toiber, T. A. Thanaraj, and Hermona Soreq. 2005. 'Function of alternative splicing', *Gene*, 344: 1-20.

Sterne-Weiler, Timothy, Rocio Teresa Martinez-Nunez, Jonathan M. Howard, Ivan Cvitovik, Sol Katzman, Muhammad A. Tariq, Nader Pourmand, and Jeremy R. Sanford. 2013. 'Frac-seq reveals isoform-specific recruitment to polyribosomes', *Genome Res.*, 23: 1615-23.

Talavera, David, Christine Vogel, Modesto Orozco, Sarah A. Teichmann, and Xavier de la Cruz. 2007. 'The (in)dependence of alternative splicing and gene duplication', *PLoS Comput. Biol.*, 3: e33.

Tamarkin-Ben-Harush, Ana, Jean-Jacques Vasseur, Françoise Debart, Igor Ulitsky, and Rivka Dikstein. 2017. 'Cap-proximal nucleotides via differential eIF4E binding and alternative promoter usage mediate translational response to energy stress', *Elife*, 6.

Tang, Jen-Yang, Jin-Ching Lee, Ming-Feng Hou, Chun-Lin Wang, Chien-Chi Chen, Hurng-Wern Huang, and Hsueh-Wei Chang. 2013. 'Alternative splicing for diseases, cancers, drugs, and databases', *ScientificWorldJournal*, 2013: 703568.

Tapial, Javier, Kevin C. H. Ha, Timothy Sterne-Weiler, André Gohr, Ulrich Braunschweig, Antonio Hermoso-Pulido, Mathieu Quesnel-Vallières, Jon Permanyer, Reza Sodaei, Yamile Marquez, Luca Cozzuto, Xinchen Wang, Melisa Gómez-Velázquez, Teresa Rayon, Miguel Manzanares, Julia Ponomarenko, Benjamin J. Blencowe, and Manuel Irimia. 2017. 'An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms', *Genome Research*, 27: 1759-68.

Tasic, Bosiljka, Christoph E. Nabholz, Kristin K. Baldwin, Youngwook Kim, Erroll H. Rueckert, Scott A. Ribich, Paula Cramer, Qiang Wu, Richard Axel, and Tom Maniatis. 2002. 'Promoter Choice Determines Splice Site Selection in Protocadherin α and γ Pre-mRNA Splicing', *Molecular Cell*, 10: 21-33.

Termonia, Yves. 1994. 'Molecular modeling of spider silk elasticity', *J Macromolecules*, 27: 7378-81.

Tilgner, Hagen, David G. Knowles, Rory Johnson, Carrie A. Davis, Sudipto Chakrabortty, Sarah Djebali, João Curado, Michael Snyder, Thomas R. Gingeras, and Roderic Guigó. 2012. 'Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs', *Genome Res.*, 22: 1616-25.

Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. 2010. 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nat. Biotechnol.*, 28: 511-15.

Tress, Michael L., Federico. Abascal, and Alfonso. Valencia. 2017. 'Alternative Splicing May Not Be the Key to Proteome Complexity', *Trends in Biochemical Sciences*, 42: 98-110.

Tress, Michael L., Pier Luigi Martelli, Adam Frankish, Gabrielle A. Reeves, Jan Jaap Wesselink, Corin Yeats, Páll Isólfur Olason, Mario Albrecht, Hedi Hegyi, Alejandro Giorgetti, Domenico Raimondo, Julien Lagarde, Roman A. Laskowski, Gonzalo López, Michael I. Sadowski, James D. Watson, Piero Fariselli, Ivan Rossi, Alinda Nagy, Wang Kai, Zenia Størling, Massimiliano Orsini, Yassen Assenov, Hagen Blankenburg, Carola Huthmacher, Fidel Ramírez, Andreas Schlicker, France Denoeud, Phil Jones, Samuel Kerrien, Sandra Orchard, Stylianos E. Antonarakis, Alexandre Reymond, Ewan Birney, Søren Brunak, Rita Casadio, Roderic Guigo, Jennifer Harrow, Henning Hermjakob, David T. Jones, Thomas Lengauer, Christine A. Orengo, László Patthy, Janet M. Thornton, Anna Tramontano, and Alfonso Valencia. 2007. 'The implications of alternative splicing in the ENCODE protein complement', *Proc. Natl. Acad. Sci. U. S. A.*, 104: 5495-500.

Trinklein, Nathan D., Shelley J. Force Aldred, Alok J. Saldanha, and Richard M. Myers. 2003. 'Identification and functional analysis of human transcriptional promoters', *Genome Res.*, 13: 308-12.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. 2001. 'The sequence of the human genome', *Science*, 291: 1304-51.

Verma, Paras, Biswajit Panda, Kamal P Singh, Shashi B %J ACS Biomaterials Science Pandit, and Engineering. 2021. 'Optimal protein sequence design mitigates mechanical failure in silk β-sheet nanocrystals', 7: 3156-65.

Vernia, Santiago, Yvonne Jk Edwards, Myoung Sook Han, Julie Cavanagh-Kyros, Tamera Barrett, Jason K. Kim, and Roger J. Davis. 2016. 'An alternative splicing program promotes adipose tissue thermogenesis', *Elife*, 5.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. 2008. 'Alternative isoform regulation in human tissue transcriptomes', *Nature*, 456: 470-76.

Wang, Peng, Bo Yan, Jun-Tao Guo, Chindo Hicks, and Ying Xu. 2005. 'Structural genomics analysis of alternative splicing and application to isoform structure modeling', *Proc. Natl. Acad. Sci. U. S. A.*, 102: 18920-25.

Wang, Xiaojing, Simona G. Codreanu, Bo Wen, Kai Li, Matthew C. Chambers, Daniel C. Liebler, and Bing Zhang. 2018. 'Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity', *Mol. Cell. Proteomics*, 17: 422-30.

Wang, Zefeng, and Christopher B. Burge. 2008. 'Splicing regulation: From a parts list of regulatory elements to an integrated splicing code', *RNA*, 14: 802-13.

Weatheritt, Robert J., Timothy Sterne-Weiler, and Benjamin J. Blencowe. 2016. 'The ribosome-engaged landscape of alternative splicing', *Nat. Struct. Mol. Biol.*, 23: 1117-23.

Wickham, Hadley. 2016. *ggplot2: elegant graphics for data analysis* (Springer).

Work, Robert. 1985. 'Viscoelastic behaviour and wet supercontraction of major ampullate silk fibres of certain orb-web-building spiders (Araneae)', *Journal of Experimental Biology*, 118: 379-404.

Xiao, Senbo, Wolfram Stacklies, Murat Cetinkaya, Bernd Markert, and Frauke Gräter. 2009. 'Mechanical response of silk crystalline units from force-distribution analysis', *Biophys. J.*, 96: 3997-4005.

Xin, Dedong, Landian Hu, and Xiangyin Kong. 2008. 'Alternative promoters influence alternative splicing at the genomic level', *PloS one*, 3: e2377.

Xing, Yi, and Christopher Lee. 2005. 'Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences', *Proc. Natl. Acad. Sci. U. S. A.*, 102: 13526-31.

Xing, Yi, Qiang Xu, and Christopher Lee. 2003. 'Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains', *FEBS Lett.*, 555: 572-78.

Xing, Yi., and Christopher. Lee. 2006. 'Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes', *Nat. Rev. Genet.*, 7: 499-509.

Xu, Changjian, Dechang Li, Yuan Cheng, Ming Liu, Yongwei Zhang, and Baohua Ji. 2015. 'Pulling out a peptide chain from beta-sheet crystallite: Propagation of instability of H-bonds under shear force', *Acta Mechanica Sinica*, 31: 416-24.

Xu, Zhiping, and Markus J. Buehler. 2010. 'Mechanical energy transfer and dissipation in fibrous beta-sheet-rich proteins', *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 81: 061910.

Yamamoto, Shunsuke, Yasunori Higuchi, Kazuhiro Yoshiyama, Eiichi Shimizu, Masashi Kataoka, Naoki Hijiya, and Keiko Matsuura. 1999. 'ADAM family proteins in the immune system', *Immunology Today*, 20: 278-84.

Yang, Xinping, Jasmin Coulombe-Huntington, Shuli Kang, Gloria M. Sheynkman, Tong Hao, Aaron Richardson, Song Sun, Fan Yang, Yun A. Shen, Ryan R. Murray, Kerstin Spirohn, Bridget E. Begg, Miquel Duran-Frigola, Andrew MacWilliams, Samuel J. Pevzner, Quan Zhong, Shelly A. Wanamaker, Stanley Tam, Lila Ghamsari, Nidhi Sahni, Song Yi, Maria D. Rodriguez, Dawit Balcha, Guihong Tan, Michael Costanzo, Brenda Andrews, Charles Boone, Xianghong J. Zhou, Kourosh Salehi-Ashtiani, Benoit Charloteaux, Alyce A. Chen, Michael A. Calderwood, Patrick Aloy, Frederick P. Roth, David E. Hill, Lilia M. Iakoucheva, Yu Xia, and Marc Vidal. 2016. 'Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing', *Cell*, 164: 805-17.

Yao, Chengguo, Jacob Biesinger, Ji Wan, Lingjie Weng, Yi Xing, Xiaohui Xie, and Yongsheng Shi. 2012. 'Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation', *Proc. Natl. Acad. Sci. U. S. A.*, 109: 18773-78.

Yao, X. Q., R. U. Malik, N. W. Griggs, L. Skjaerven, J. R. Traynor, S. Sivaramakrishnan, and B. J. Grant. 2016. 'Dynamic Coupling and Allosteric Networks in the alpha Subunit of Heterotrimeric G Proteins', *J Biol Chem*, 291: 4742-53.

Yarger, Jeffery L, Brian R Cherry, and Arjan Van Der Vaart. 2018. 'Uncovering the structure–function relationship in spider silk', *Nature Reviews Materials*, 3: 1-11.

Yates, Andrew D., Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Andrey G. Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon, Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N. Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E. Hunt, Garth Iisley, Myrto Kostadima, Nick Langridge, Jane E. Loveland, Fergal J. Martin, Joannella Morales, Jonathan M. Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J. Trevanion, Fiona Cunningham, Kevin L. Howe, Daniel R. Zerbino, and Paul Flicek. 2020. 'Ensembl 2020', *Nucleic Acids Res.*, 48: D682-D88.

Zafarullah, Marwa, Hiu-Tung Tang, Blythe Durbin-Johnson, Emily Fourie, David Hessl, Susan M Rivera, and Flora %J Scientific reports Tassone. 2020. 'FMR1 locus isoforms: potential biomarker candidates in fragile X-associated tremor/ataxia syndrome (FXTAS)', 10: 1-10.

Zhang, Hong, Yirong Wang, Xinkai Wu, Xiaolu Tang, Changcheng Wu, and Jian Lu. 2021. 'Determinants of genome-wide distribution and evolution of uORFs in eukaryotes', *Nat. Commun.*, 12: 1076.

Zhang, Theresa, Peter Haws, and Qiang Wu. 2004. 'Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation', *Genome Res.*, 14: 79-89.

Zhang, Xiao-Ou, Yu Fu, Haiwei Mou, Wen Xue, and Zhiping Weng. 2018. 'The temporal landscape of recursive splicing during Pol II transcription elongation in human cells', *PLoS Genet.*, 14: e1007579.

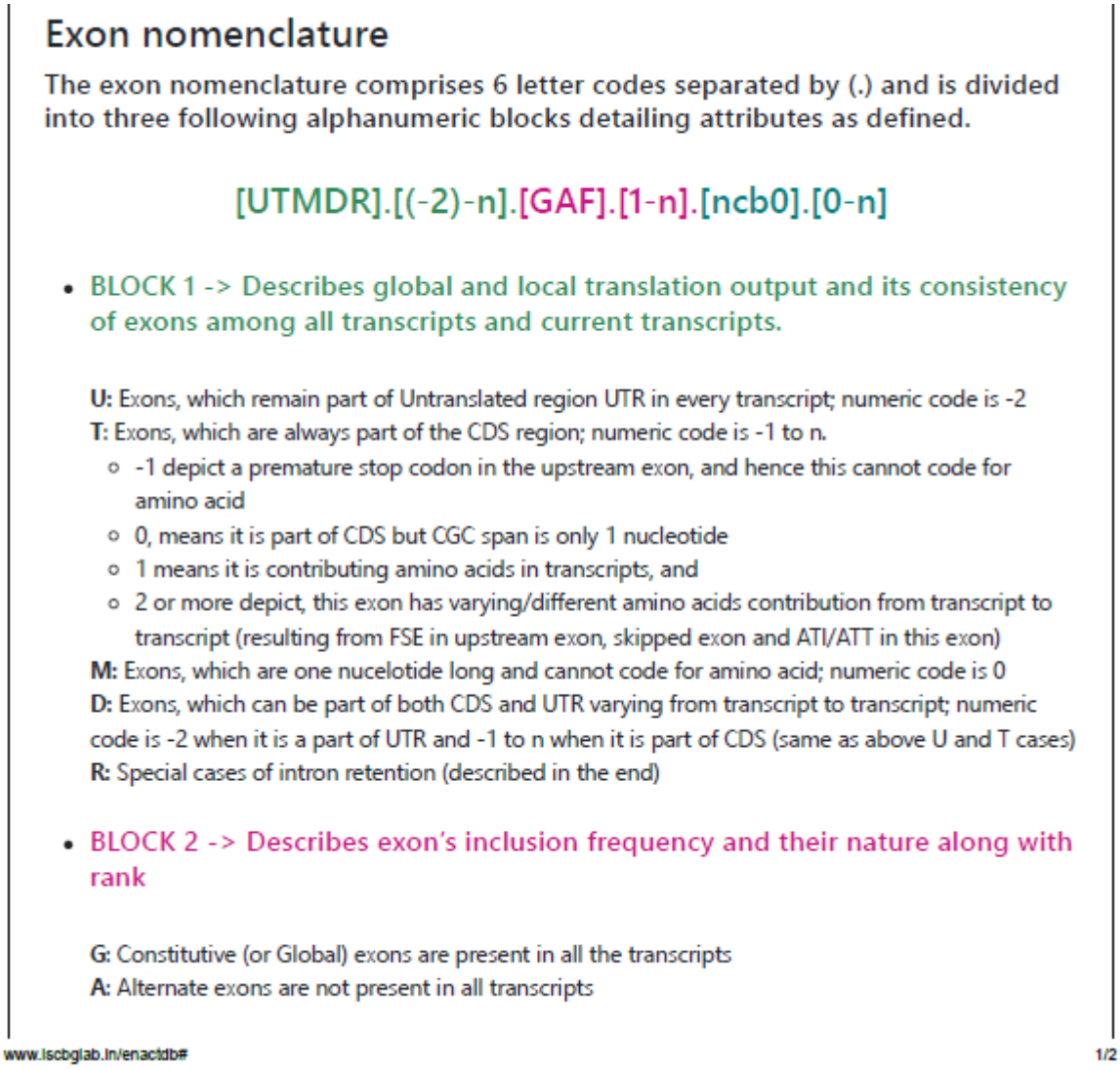
Zhang, Yuanjiao, Jinjun Qian, Chunyan Gu, and Ye Yang. 2021. 'Alternative splicing and cancer: a systematic review', *Signal Transduct Target Ther*, 6: 78.

Zhang, Yuanyuan, Wei Li, and Mary Vore. 2007. 'Translational regulation of rat multidrug resistance-associated protein 2 expression is mediated by upstream open reading frames in the 5' untranslated region', *Mol. Pharmacol.*, 71: 377-83.

Zhao, Shanrong, and Baohong Zhang. 2015. 'A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification', *BMC Genomics*, 16: 97.

Appendix A – Exon nomenclature description

Attached below is the snapshot of 'Nomenclature' section from the ENACTdb web utility.



Exon nomenclature

The exon nomenclature comprises 6 letter codes separated by (.) and is divided into three following alphanumeric blocks detailing attributes as defined.

[UTMDR].[(-2)-n].[GAF].[1-n].[ncb0].[0-n]

- BLOCK 1 -> Describes global and local translation output and its consistency of exons among all transcripts and current transcripts.

U: Exons, which remain part of Untranslated region UTR in every transcript; numeric code is -2
T: Exons, which are always part of the CDS region; numeric code is -1 to n.
  - -1 depict a premature stop codon in the upstream exon, and hence this cannot code for amino acid
  - 0, means it is part of CDS but CGC span is only 1 nucleotide
  - 1 means it is contributing amino acids in transcripts, and
  - 2 or more depict, this exon has varying/different amino acids contribution from transcript to transcript (resulting from FSE in upstream exon, skipped exon and ATI/ATT in this exon)

M: Exons, which are one nucelotide long and cannot code for amino acid; numeric code is 0
D: Exons, which can be part of both CDS and UTR varying from transcript to transcript; numeric code is -2 when it is a part of UTR and -1 to n when it is part of CDS (same as above U and T cases)
R: Special cases of intron retention (described in the end)

- BLOCK 2 -> Describes exon's inclusion frequency and their nature along with rank

G: Constitutive (or Global) exons are present in all the transcripts
A: Alternate exons are not present in all transcripts

www.iscbglab.in/enactdb#

1/2

F: Exons, which when considered with their alternate splicing sites, can be found in every transcript
Numeric code here gives a serial identifier (from 1 to n) to the exon within each of the above-described categories (Decided after sorting the RSOEx by genomic coordinates)

- BLOCK 3 -> Describes choice of alternate splice site and frequency of occurrence

All these modifications are assigned names by comparing them to the RSOEx.
n: 3′ splice site of upstream intron is changed, leading to extension or constriction of exon length from its 5′ end
c: 5′ splice site of downstream intron is changed, leading to extension or constriction of exon length from its 3′ end
b: 5′ splice site of downstream intron and 3′ splice site of upstream intron both are changed simultaneously
Numeric code here also gives a serial identifier (from 0 to n) to the exon within each of the above described categories.
0: denotes the original exon as in the RSOEx of that gene. Numeric code will be 0 in this case.

- Intron Retention Cases

They are enlisted as identifier having 5 separate parts (separated by ':')
  - R is the first letter detailing it as a retention case.
  - Numeric code (-2 to n) as previously described in block 1.
  - 6-letter identifier of the exon (separated by ',') from which the intron is retained
  - Serial number(0 to n) of the retention event from that exon
  - 6 letter identifier of exon upto which the intron has been retained.

Glossary: Untranslated Region(UTR), Coding Sequence(CDS), Coding Genomic Coordinates(CGC), Frame Shift Events(FSE), Alternative Transcription Initiation(ATI), Alternative Transcription Termination(ATT), Principal Isoform(PI), Reference Set of Exons (RSOEx)

Server developed and maintained by shashibp-lab: "https://shashibp-lab.github.io/"
© 2024 Copyright: Department of Biological Sciences, IISER Mohali, Sector 81, SAS Nagar, 140306, Punjab, India