



Protein diversification through post-translational modifications, alternative splicing, and gene duplication

Yonathan Goldtzvik¹, Neeladri Sen¹, Su Datt Lam^{1,2} and Christine Orengo¹

Abstract


Proteins provide the basis for cellular function. Having multiple versions of the same protein within a single organism provides a way of regulating its activity or developing novel functions. Post-translational modifications of proteins, by means of adding/removing chemical groups to amino acids, allow for a well-regulated and controlled way of generating functionally distinct protein species. Alternative splicing is another method with which organisms possibly generate new isoforms. Additionally, gene duplication events throughout evolution generate multiple paralogs of the same genes, resulting in multiple versions of the same protein within an organism. In this review, we discuss recent advancements in the study of these three methods of protein diversification and provide illustrative examples of how they affect protein structure and function.


Addresses

¹ Department of Structural and Molecular Biology, University College London, London, United Kingdom

² Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Corresponding author: Orengo, Christine (c.orengo@ucl.ac.uk)

 (Sen N.)

 (Orengo C.)

Current Opinion in Structural Biology 2023, **81**:102640

This review comes from a themed issue on **Sequences and Topology (2023)**

Edited by **Madan Babu** and **Rita Casadio**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online xxx

<https://doi.org/10.1016/j.sbi.2023.102640>

0959-440X/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Introduction

Traditionally, we think of protein variation in terms of genetic variation within a population, or between species. A protein appears in two different individuals or two different species, but with small changes to its

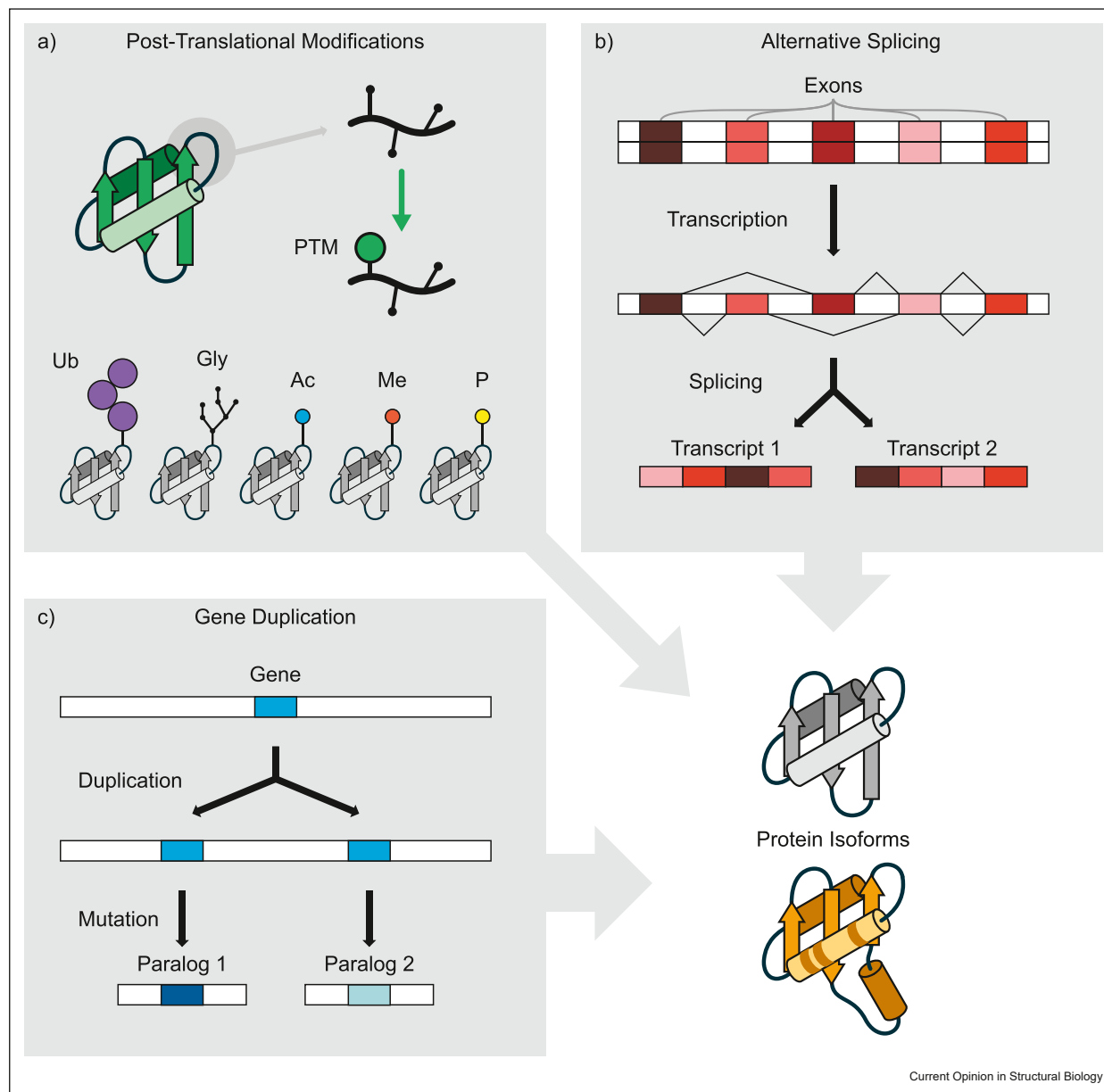
sequence, due to mutation. In contrast, there are many examples of how multiple versions, or protein species, of the same protein come about within an individual [1]. Three sources of such protein species are post-translational modifications (PTM), alternative splicing (AS), and gene duplication (GD) that result in gene paralogs (Figure 1).

PTMs refer to the post-translational chemical modification of protein residues by the covalent addition of chemical groups such as acetyl, glucosyl, methyl, phosphoryl, or ubiquitin. These modifications expand the repertoire of the standard 20 amino acids and lead to various effects on protein interactions, lifespan, folding, solubility, and localization. Hence, PTMs are important in various biological processes such as signal transduction, gene expression regulation, cell cycle, and DNA repair.

AS is the rearrangement and assembly of distinct exons (protein-coding sequences) from a single gene, resulting in multiple protein species called isoforms [2]. Multicellular creatures including humans, animals, and plants have been found to exhibit AS [3]. AS is a useful way of possibly increasing protein diversity and introducing additional levels of regulation, as different protein isoforms can be differentially expressed in various tissues and different developmental stages.

Finally, another evolutionary path for the formation of multiple protein species of the same protein within a species is GD [4]. The scale of GD events throughout evolution can range from duplication of single genes to the duplication of whole genomes. The resulting duplicates, called paralogs, accumulate mutations over time, resulting in evolutionary divergence in both sequence and function [5]. In order to study the structural diversity between paralogs within different CATH families, we compared the good quality protein domains containing high pLDDT, low unordered regions, higher secondary structures, high globularity, and packing density (for details about the structures being considered, please refer to Bordin et al. for the comparison) [6]. We were interested in looking at how much the structures

Figure 1



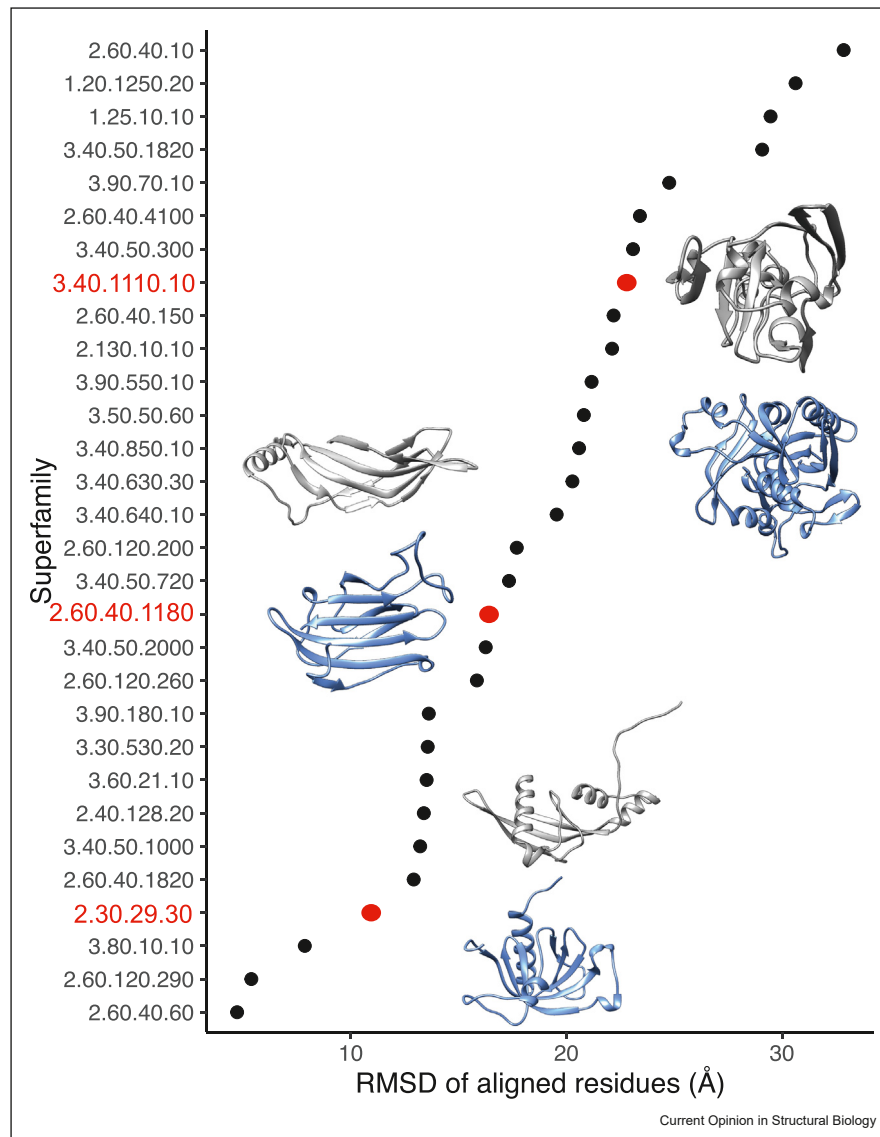
Different sources of protein species. **a)** Post-translational modifications (PTMs) are chemical groups that modify the protein by covalently binding to one or more of its amino acid residues. **b)** Alternative splicing (AS) is the formation of different protein isoforms from the same gene by alternative combinations of its exons during the splicing process. **c)** Gene duplication events result in multiple copies of a single gene, called paralogs. The different paralogs accumulate mutations over time, resulting in different versions of the same gene.

diverged and hence only considered the most diverse protein domain alignment in our study. We found that in some families paralogs had similar structures, while in others the structures diverged significantly (Figure 2) [6–8]. These findings further highlight how GD contributes to structural and functional diversity. GD is a useful evolutionary tool, as it increases the tolerance for the accumulation of deleterious mutations, as well as

potentially beneficial mutations, thus introducing opportunities to develop new functions.

In this review, we examine how these three forms of protein species contribute to protein functional diversity and evolutionary fitness of an organism. We also highlight how advances in protein structure determination and analysis, both experimental and computational, can

Figure 2



Structural diversity within paralogs - All against all superimposition of paralogs of protein domains (within a superfamily and species) were created for good quality AlphaFold domains [6,9] using FoldSeek [10]. The RMSD scores for the aligned residues for the 30 superfamilies with minimum TM-scores (calculated using FoldSeek) were calculated using the structure superimposition tool SSAP [11]. These maximal RMSD scores were plotted against the respective superfamily. Representative examples of protein domains belonging to diverse superfamilies (highlighted in red) have been shown in ribbon diagram. For the superfamilies 3.40.1110.10, 2.60.40.1180, and 2.30.29.30 representative domains in the figure belong to UniProt IDs G5EEK8 (residue nos 361–603) and A0A7I9BBC6 (residue no. 168–345), Q9UL18 (residue no. 34–164) and Q9BQ17 (residue no 508–632), and Q2QXF6 (residue no. 78–192) and Q2RAZ2 (residue no. 530–647), respectively.

contribute to our understanding of the functional diversity of protein species.

Post-translational modifications

PTMs contribute to protein species diversification dramatically as they provide a plethora of ways of modifying a protein. Large-scale, mass spectrometry-

based proteomics studies have identified tens of thousands of PTMs, a large majority of which have not been associated with functional relevance [12]. Their effect on the biophysical and, by extension, biological properties of proteins is complex. Analysis of X-ray structures in the PDB has shown that only 7% of glycosylated and 13% of phosphorylated proteins undergo changes >2 Å

[13]. The effects of PTMs on backbone conformation could be either stabilizing or destabilizing, depending on the type of PTM [14]. Recent studies involving molecular dynamics (MD) simulations have explored the effects of PTMs on protein–protein interactions [15]. In general, acetylation has been shown to have a stabilizing effect on such interactions, while phosphorylation tends to have a destabilizing effect, however, these effects are not additive. Furthermore, these PTMs affect the dynamics and allosteric interactions of proteins. Computational studies show that phosphorylation sites tend to be predominantly on solvent-exposed residues, or buried residues that are flexible enough to be exposed after modification [16–18]. Genomic analysis of these PTM sites in humans have shown them to be negatively selected based on percentage of rare substitutions and ratio of non-synonymous to synonymous mutations. In addition, these sites have a higher number of disease-associated mutations compared to other residues [19]. Conservation of PTMs across the tree of life has shown that the PTM sites have only weak evolutionary constraints [20]. However, clade specific studies on human PTMs in which they were compared to other ordered and disordered regions of the eukaryotes have shown strong conservation signals based on the PTM type and whether a PTM is in a structured or disordered region [21]. Indeed, PTMs are found in many intrinsically disordered proteins (IDPs)/intrinsically disordered regions (IDRs) which are also involved in various cellular regulation pathways. Additions/removals of the PTMs in these IDPs can lead to various structural changes and transitions between ordered and disordered states [22]. PTMs in IDPs can also lead to phase separation of these proteins leading to phase-separated droplets and

membrane-less compartments in the cell [23]. The protein huntingtin, involved in the Huntington disease, has an N-terminus that is intrinsically disordered and has PTM sites. MD studies of huntingtin have shown that phosphorylation leads to helix stabilization and charge neutralization by N-terminus acetylation [24]. Similarly, tau protein is also an IDP and has multiple PTM sites. Tau PTMs cause various structural changes leading to phase separation, aggregation, microtubule assembly, and degradation [25].

With the boom in the number of near-accurate computational predictions of protein structures, tools such as Privateer [26,28] have been developed to model the PTMs on the residues that undergo these modifications. Other tools such as StructureMap have been developed to map PTMs from proteomics studies onto these computational models [12]. Additionally, there are several databases with information on PTM sites, their function, mutations, and 3D structural context [27]. A summary of these can be found in Table 1.

Alternative splicing

AS and the resulting isoforms provide a useful mechanism for protein diversification and protein species generation within an individual. A fundamental question is the prevalence of isoforms that originate from AS in nature. The evidence from transcriptomics experiments does indicate that AS generates many transcripts. In contrast, proteomic studies have only been able to confirm the presence of a small number of isoforms that are generated by such transcripts. Transcripts may not be translated into proteins, may generate only small amounts of protein, or could be only expressed in

Table 1

Bioinformatics resources providing information on post-translational modifications (PTMs), alternative splicing (AS), and paralog structures.

Resource	Type	Description	Reference
Privateer	PTM	Conformational validation of carbohydrate structures	[28]
StructureMap	PTM	Python package for integration of AlphaFold data and proteomics and PTM data.	[12]
ActiveDriverDB	PTM	Using PTM sites to interpret genetic variation in humans	[29]
PTMD	PTM	A database of human disease-associated PTMs	[30]
Scop3P	PTM	A resource of human phosphosites within their full context, including structural context	[31]
APPRIS	AS	Protein isoform annotations for a range of species	[32]
Oncosplicing	AS	Database for clinically relevant alternative splicing in 33 human cancers	[33]
LncAS2Cancer	AS	A comprehensive database for alternative splicing of long non-coding RNAs across human cancers	[34]
DIGGER	AS	Database of functional roles of alternative splicing in protein interactions	[35]
ThorAxe	AS	Assessment of AS evolutionary conservation	[36]
MeDAS	AS	A Metazoan developmental AS database	[37]
PISE	AS	An AS database for several plant species	[38]
CATH	Structure	Classification of protein domains by structure	[8]
SCOPe	Structure	Classification of protein domains by structure	[39]
ECOD	Structure	Evolutionary classification of protein domains	[40]

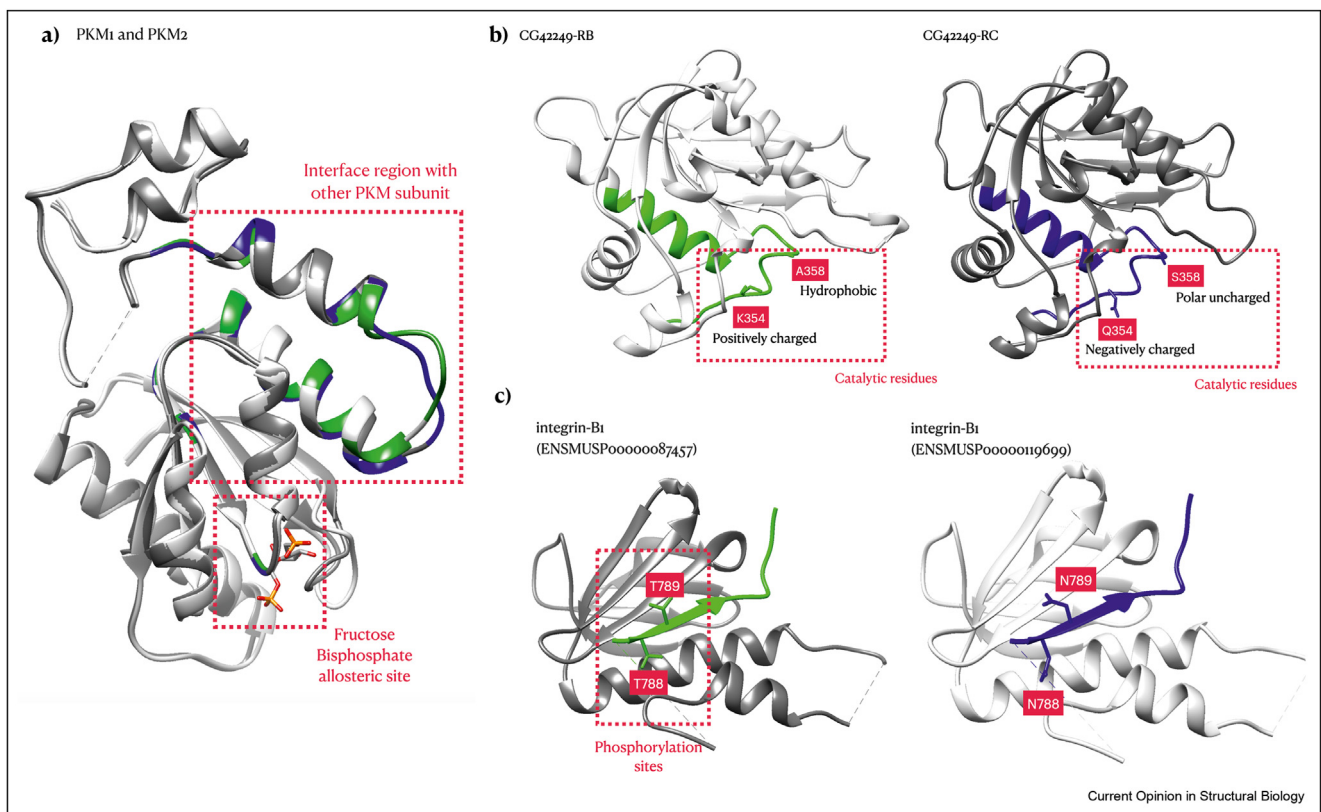
limited tissues and conditions [41]. This disparity results in an ongoing debate regarding the extent to which AS results in different protein isoforms [42–44].

While the extent of the contribution of AS to protein polymorphism in general is still being established, there is emerging evidence for AS playing a functional role in biology. Perhaps the most indicative evidence of AS functionalization is the tissue specificity of expression of different transcripts. A recent study showed that more than a third of splice events for which both proteomics and RNAseq evidence can be found are tissue-specific [41]. Furthermore, from an evolutionary perspective, the vast majority of such tissue-specific splice events are ancient, conserved over more than 400 million years. This indicates a correlation between functionalization and evolutionary conservation of AS. A recent study used evolutionary splicing graphs to investigate a set of 50 genes, and the findings suggest that AS may be conserved between amphibians and primates, providing additional evidence for its potential functional significance [36]. A wide range of studies resulted in AS data,

which can be found in a selection of curated databases. For a summary of several such resources, see Table 1.

An important question is how AS contributes to functional diversity of a protein at the structural level. There are many different types of AS, but one type, mutually exclusive exons (MXE), is a good example (Figure 3). The AS of proteins whose isoforms are confirmed by proteomics experiments tend to be enriched in MXE, which are less likely to disrupt the protein structural core [43,45,46]. Interestingly, structural analysis studies reveal that AS involving MXEs tend to affect surface-exposed residues at functional sites and lead to radical amino acid substitutions [47]. In the case of tandem duplicated exons, residue substitutions also tend to be at the protein surface, however, the nature of these substitutions is more conservative in terms of amino acid identities and may serve as a means of fine-tuning protein function [48]. Protein–protein networks may undergo tissue-specific rewiring as a result of AS [49,50]. It is important to note that MXEs represent only a small part of all AS.

Figure 3



Examples of MXE events altering protein functional sites. **a)** MXE varying residues were discovered at PKM1/PKM2 isoforms' allosteric site and tetramerization interface region. **b)** Key catalytic residue switches between CG42249 isoforms; Enzymes with different activities may be produced by such physicochemical changes. **c)** Two threonine residues that serve as phosphorylation sites were replaced in one integrin-B1 isoform, potentially causing a shift in downstream signaling.

Spliceosomes, which are RNA-protein complexes responsible for catalyzing the splicing process, and splicing factors, which interact with the spliceosome to regulate its activity and guide the selection of splice sites, have been found to undergo evolutionary changes that impact the diversity of splicing isoforms [51,52]. Splicing in prokaryotes occurs independently of a spliceosome, as this complex is exclusively present in eukaryotic cells. To trace the evolutionary origins of the spliceosome in the last eukaryotic common ancestor (LECA), Vosseberg et al. conducted homology searches using human spliceosomal proteins and identified 145 spliceosomal orthogroups [52]. Their analysis revealed that the prokaryote-derived core of the spliceosome was supplemented with an excess of proteins associated with ribosome-related processes, which underwent extensive duplications, leading to increased complexity in the evolving spliceosome.

A discussion of AS and its effect on biological function would be incomplete without concrete examples. There are several cases of proteins with key biological roles that undergo AS that impact their function. One such example is that of the histone core component H2A and its two isoforms: macroH2A1.1 and macroH2A1.2 [53]. The isoform macroH2A1.1 contains features that allow it to accommodate ADP-ribose within the binding pocket, while macroH2A1.2 lacks these features. Therefore, whichever isoform is expressed may affect ADP-ribose signaling and NAD⁺ metabolism. The AS of H2A appears to be a recent addition in the evolution of histones and is only observed in jawed vertebrates [54].

Another important example is that of G Protein-Coupled Receptors (GPCR). A recent study demonstrated the functional divergence of different isoforms of a single GPCR gene, with varied signaling capabilities [55]. The study highlights how the expression of different unique isoform combinations in different tissues activates distinct signaling mechanisms. Some isoforms may alter cellular responses to drugs and provide novel targets for treatments with greater tissue selectivity.

Gene duplication

While GD provides a way for proteins to acquire new functions without sacrificing their original role in the organism, the constraints affecting the evolutionary pathways of paralogs are all but simple. This is particularly the case when paralogs share interactions with other proteins, resulting in trigenic interactions. In such cases, the sequence and structural divergence of one paralog can affect its counterpart via evolutionary changes in their shared partner [56]. Indeed, the more entangled the two paralogs are in their interactions, the more they tend to retain functional redundancy [57]. A

particularly interesting example of how protein interactions can influence the evolution of paralogs is the case of homo-oligomeric proteins. When genes of proteins that form homo-oligomeric assemblies undergo duplication and divergence, two possible assembly outcomes emerge. The first outcome is the formation of two different sets of homo-oligomers, each corresponding to a different paralog, where the paralogs do not mix. In the second outcome the two paralogs form hetero-oligomers [58]. In eukaryotes, hetero-oligomeric complexes appear to be more common [59–61]. Nevertheless, paralogs can evolve to avoid heteromeric assembly in certain cases [62].

While a full survey of the functions of paralogs is beyond the scope of this work, we present several illustrative examples of biological interest. Proteins with regulatory functions provide a good starting point, and in particular, there are examples of transcription factors with paralogs of varying degrees of functional redundancy, as well as tissue-specificity in their expression [63–67]. Transcription factors are a good example of how paralogs evolve new functions. Because paralogs of the same transcription factor compete for the same DNA binding sites with different affinities, their divergence is a means of tweaking gene networks [68]. Interestingly, it has been shown in yeast that the way in which the paralog DNA binding affinities are modified is not through changes in the DNA binding domains, but rather changes to parts of the protein sequence that affect interactions with secondary factors, which in turn affect the affinity [69].

We find interesting examples of paralogs of proteins involved in fundamental cellular processes. Several proteins involved in DNA repair have associated paralogs. In humans, topoisomerase II has two paralogs, TOP2A and TOP2B, whose sequences are similar (70–80% sequence homology) but they differ in function [70]. Another example is that of RAD51, a protein involved in DNA repair and homologous recombination. The RAD51 paralogs have been shown to be involved in the formation of different protein complexes with different roles [71–75]. There are also paralogs found in the transcription machinery. Examples of such genes with paralogs are GPN, a crucial biogenesis factor of RNA polymerase II, and the paralogs POLR3G and POLR3L, which are subunits of RNA polymerase III [76,77].

Interestingly, genes coding for the components of the fundamental molecular complexes charged with protein production and degradation, the ribosome and proteasome, have paralogs as well. An emerging field of research that is attracting much attention and debate is that of ribosome heterogeneity [78–80]. The paralogs of the ribosomal protein genes bL31 and bL36 in bacteria, and RPL8 in yeast, have been shown to be

involved in response to changes in the environment of the organism [81–83]. In mammals, ribosomal protein paralogs such as RPL10L, RPL39L, and RPL22L have been shown to be important in fertility, cell proliferation, and development, and some of them are involved in certain types of cancer [84–88].

Finally, in addition to the standard proteasome, there are three more versions of the proteasome with high tissue expression and functional specificity: immunoproteasome, thymoproteasome, and spermatoproteasome [89,90]. These are defined by specific paralogs that are incorporated into the subunits of the proteasome and expressed in the respective tissues. An example is PSMA8, a paralog of PSMA7, that codes for the $\alpha 4$ s subunit, a component of the spermatoproteasome [91,92]. Additionally, PA28 is a proteasome activator for which there are 3 paralogs in jawed vertebrates: PA28 α , PA28 β , and PA28 γ . These paralogs assemble into either a hetero-heptameric ring (PA28 $\alpha\beta$) or a homo-heptameric ring (PA28 γ) [93–95].

Conclusions

It is clear, given these different forms of protein species (PTM, AS, GD), that protein diversity within an individual is prevalent and that protein species play an important functional role in biology. Furthermore, while substantial progress has been made in mapping different protein species and analyzing their function, it is also likely that there may be many more protein species that contribute to biological functions that have not been properly assessed yet.

The explosion in the number of solved protein structures in recent years, partially due to advances in cryo-EM techniques, together with the remarkable computational progress in structure determination and prediction, primarily due to AlphaFold2, presents a fantastic opportunity in this context [96]. The vast amounts of structural data can be invaluable in analyzing protein species and provide a lens with which we can determine the effects of sequence variation among protein species on their function.

CRedit author contributions

Yonathan Goldtzvik: Writing – original draft, review, and editing, Visualization; **Neeladri Sen:** Writing – original draft, Formal analysis, Visualization; **Su Datt Lam:** Writing – original draft, Visualization; **Christine Orengo:** Writing – review and editing.

Declaration of competing interest

The authors declare no competing interests.

Data availability

Data will be made available on request.

Acknowledgements

This work was funded by the Biotechnology and Biological Sciences Research Council (grant numbers: BB/V014722/1 and PO number 60175814), and the Fundamental Research Grant Scheme from the Ministry of Higher Education Malaysia (grant number: FRGS/1/2020/STG01/UKM/02/3).

References

Papers of particular interest, published within the period of review, have been highlighted as:

- * of special interest
- ** of outstanding interest

1. Schlüter H, Apweiler R, Holzhütter H-G, Jungblut PR: **Finding one's way in proteomics: a protein species nomenclature.** *Chem Cent J* 2009, **3**:1–10.
 2. Wright CJ, Smith CWJ, Jiggins CD: **Alternative splicing as a source of phenotypic diversity.** *Nat Rev Genet* 2022, **23**:697–710.
 3. Chaudhary S, Khokhar W, Jabre I, Reddy ASN, Byrne LJ, Wilson CM, Syed NH: **Alternative splicing and protein diversity: plants versus animals.** *Front Plant Sci* 2019, **10**.
 4. Zhang J: **Evolution by gene duplication: an update.** *Trends Ecol Evol* 2003, **18**:292–298.
 5. Koonin EV: **An apology for orthologs - or brave new memes.** *Genome Biol* 2001, **2**:2.
 6. Bordin N, Sillitoe I, Nallapareddy V, Rauer C, Lam SD, Waman VP, Sen N, Heinzinger M, Littmann M, Kim S, *et al.*: **AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms.** *Commun Biol* 2023, **6**:1–12.
 7. Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J: **Cath – a hierarchic classification of protein domain structures.** *Structure* 1997, **5**:1093–1109.
 8. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, *et al.*: **CATH: increased structural coverage of functional space.** *Nucleic Acids Res* 2021, **49**:D266–D273.
 9. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, *et al.*: **AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.** *Nucleic Acids Res* 2022, **50**:D439–D444.
 10. Kempen M van, Kim SS, Tumescheit C, Mirdita M, Gilchrist CLM, Söding J, Steinegger M: **Foldseek: fast and accurate protein structure search.** 2022, <https://doi.org/10.1101/2022.02.07.479398>.
 11. Orengo CA, Taylor WR: **[36] SSAP: sequential structure alignment program for protein structure comparison.** In *Methods in enzymology.* Academic Press; 1996:617–635.
 12. Bludau I, Willems S, Zeng W-F, Strauss MT, Hansen FM, ****** Tanzer MC, Karayel O, Schulman BA, Mann M: **The structural context of posttranslational modifications at a proteome-wide scale.** *PLoS Biol* 2022, **20**, e3001636.
- The study uncovers the structural context of PTM occurrence in both folded and intrinsically disordered regions. 3D proximity analysis has also identified PTM crosstalk and spatial coregulation. They have developed python packages to visualize proteomics data on predicted proteins.
13. Xin F, Radivojac P: **Post-translational modifications induce significant yet not extreme changes to protein structure.** *Bioinformatics* 2012, **28**:2905–2913.
 14. Craveur P, Narwani TJ, Rebehmed J, de Brevem AG: **Investigation of the impact of PTMs on the protein backbone conformation.** *Amino Acids* 2019, **51**:1065–1079.
 15. Šoštarić N, Noort V van: **Molecular dynamics shows complex interplay and long-range effects of post-translational modifications in yeast protein interactions.** *PLoS Comput Biol* 2021, **17**, e1008988.

This study uses molecular dynamics simulations and free energy calculations to understand the modulation of protein interactions due to acetylation and phosphorylation in yeast. They have also analysed conformational changes due to PTM and its conservation in paralogues.

16. Henriques J, Lindorff-Larsen K: **Protein dynamics enables phosphorylation of buried residues in cdk2/cyclin-A-bound p27**. *Biophys J* 2020, **119**:2010–2018.
 17. Orioli S, Hansen CGH, Lindorff-Larsen K: **Transient exposure of a buried phosphorylation site in an autoinhibited protein**. *Biophys J* 2022, **121**:91–101.
 18. Ramasamy P, Vandermarliere E, Vranken WF, Martens L: **Panoramic perspective on human phosphosites**. *J Proteome Res* 2022, **21**:1894–1915.
 19. Reimand J, Wagih O, Bader GD: **Evolutionary constraint and disease associations of post-translational modification sites in human genomes**. *PLoS Genet* 2015, **11**, e1004919.
 20. Beltrao P, Bork P, Krogan NJ, van Noort V: **Evolution and functional cross-talk of protein post-translational modifications**. *Mol Syst Biol* 2013, **9**:714.
 21. Narasumani M, Harrison PM: **Discerning evolutionary trends in post-translational modification and the effect of intrinsic disorder: analysis of methylation, acetylation and ubiquitination sites in human proteins**. *PLoS Comput Biol* 2018, **14**, e1006349.
 22. Bah A, Forman-Kay JD: **Modulation of intrinsically disordered protein function by post-translational modifications**. *J Biol Chem* 2016, **291**:6696–6705.
 23. Owen I, Shewmaker F: **The role of post-translational modifications in the phase transitions of intrinsically disordered proteins**. *Int J Mol Sci* 2019, **20**:5501.
 24. Yalinca H, Gehin CJC, Oleinikovas V, Lashuel HA, Gervasio FL, Pastore A: **The role of post-translational modifications on the energy landscape of huntingtin N-terminus**. *Front Mol Biosci* 2019, **6**.
 25. Ye H, Han Y, Li P, Su Z, Huang Y: **The role of post-translational modifications on the structure and function of tau protein**. *J Mol Neurosci* 2022, **72**:1557–1571.
 26. Bagdonas H, Fogarty CA, Fadda E, Agirre J: **The case for post-predictive modifications in the AlphaFold protein structure database**. *Nat Struct Mol Biol* 2021, **28**:869–870.
 27. Ramazi S, Zahir J: **Post-translational modifications in proteins: resources, tools and prediction methods**. *Database* 2021. 2021. baab012.
- This article reviews the major PTM databases, tools for PTM predictions, association of PTM with diseases and biological processes.
28. Agirre J, Iglesias-Fernández J, Rovira C, Davies GJ, Wilson KS, Cowtan KD: **Privateer: software for the conformational validation of carbohydrate structures**. *Nat Struct Mol Biol* 2015, **22**: 833–834.
 29. Krassowski M, Pellegrina D, Mee MW, Fradet-Turcotte A, Bhat M, Reimand J: **ActiveDriverDB: interpreting genetic variation in human and cancer genomes using post-translational modification sites and signaling networks (2021 update)**. *Front Cell Dev Biol* 2021, **9**.
 30. Xu H, Wang Y, Lin S, Deng W, Peng D, Cui Q, Xue Y: **PTMD: a database of human disease-associated post-translational modifications**. *Dev Reprod Biol* 2018, **16**:244–251.
 31. Ramasamy P, Turan D, Tichshenko N, Hulstaert N, Vandermarliere E, Vranken W, Martens L: **Scop3P: a comprehensive resource of human phosphosites within their full context**. *J Proteome Res* 2020, **19**:3478–3486.
 32. Rodríguez JM, Pozo F, Cerdán-Vélez D, Di Domenico T, Vázquez J, Tress ML: **APPRIS: selecting functionally important isoforms**. *Nucleic Acids Res* 2022, **50**:D54–D59.
 33. Zhang Y, Yao X, Zhou H, Wu X, Tian J, Zeng J, Yan L, Duan C, Liu H, Li H, et al.: **OncoSplicing: an updated database for clinically relevant alternative splicing in 33 human cancers**. *Nucleic Acids Res* 2022, **50**:D1340–D1347.
 34. Deng Y, Luo H, Yang Z, Liu L: **LncAS2Cancer: a comprehensive database for alternative splicing of lncRNAs across human cancers**. *Briefings Bioinf* 2021, **22**:bbaa179.
 35. Louadi Z, Yuan K, Gress A, Tsou O, Kalinina OV, Baumbach J, Kacprowski T, List M: **DIGGER: exploring the functional role of alternative splicing in protein interactions**. *Nucleic Acids Res* 2021, **49**:D309–D318.
 36. Zea DJ, Laskina S, Baudin A, Richard H, Laine E: **Assessing conservation of alternative splicing with evolutionary splicing graphs**. *Genome Res* 2021, **31**:1462–1473.
 37. Li Z, Zhang Y, Bush SJ, Tang C, Chen L, Zhang D, Urrutia AO, Lin J, Chen L: **MedAS: a metazoan developmental alternative splicing database**. *Nucleic Acids Res* 2021, **49**:D144–D150.
 38. Zhang H, Jia J, Zhai J: **Plant Intron-Splicing Efficiency Database (PISE): exploring splicing of ~1,650,000 introns in Arabidopsis, maize, rice, and soybean from ~57,000 public RNA-seq libraries**. *Sci China Life Sci* 2022, <https://doi.org/10.1007/s11427-022-2193-3>.
 39. Chandonia J-M, Fox NK, Brenner SE: **SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database**. *Nucleic Acids Res* 2019, **47**:D475–D481.
 40. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H, Grishin NV: **ECOD: an evolutionary classification of protein domains**. *PLoS Comput Biol* 2014, **10**, e1003926.
 41. Rodríguez JM, Pozo F, Domenico T di, Vázquez J, Tress ML: **An analysis of tissue-specific alternative splicing at the protein level**. *PLoS Comput Biol* 2020, **16**, e1008287.
 42. Blencowe BJ: **The relationship between alternative splicing and proteomic complexity**. *Trends Biochem Sci* 2017, **42**: 407–408.
 43. Tress ML, Abascal F, Valencia A: **Alternative splicing may not be the key to proteome complexity**. *Trends Biochem Sci* 2017, **42**:98–110.
 44. Bhuiyan SA, Ly S, Phan M, Huntington B, Hogan E, Liu CC, Liu J, Pavlidis P: **Systematic evaluation of isoform function in literature reports of alternative splicing**. *BMC Genom* 2018, **19**: 1–12.
 45. Abascal F, Ezkurdia I, Rodríguez-Rivas J, Rodríguez JM, Pozo A del, Vázquez J, Valencia A, Tress ML: **Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level**. *PLoS Comput Biol* 2015, **11**, e1004325.
 46. Martínez Gomez L, Pozo F, Walsh TA, Abascal F, Tress ML: **The clinical importance of tandem exon duplication-derived substitutions**. *Nucleic Acids Res* 2021, **49**:8232–8246.
 47. Lam SD, Babu MM, Lees J, Orengo CA: **Biological impact of mutually exclusive exon switching**. *PLoS Comput Biol* 2021, **17**, e1008708.
- This structural analysis of mutually exclusive exons (MXEs) in the genomes of five metazoan species revealed that MXE-specific residues are highly enriched in surface-exposed residues and cluster at/near protein functional sites, demonstrating the ability of mutually exclusive exons to fine-tune the function of proteins.
48. Martínez-Gomez L, Cerdán-Vélez D, Abascal F, Tress ML: **Origins and evolution of human tandem duplicated exon substitution events**. *Genome Biology and Evolution*. 2022, <https://doi.org/10.1093/gbe/evac162>.
- This analysis of human tandem duplicated exon substitutions revealed that despite being highly enriched in surface-exposed residues, these residues have undergone more moderate changes. Three-quarters of tandem duplicated exon events are tissue-specific and are enriched in terms of functionality pertaining to brain and skeletal muscle structures.
49. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM: **Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks**. *Mol Cell* 2012, **46**:871–883.

50. Ellis JD, Barrios-Rodiles M, Çolak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, *et al.*: **Tissue-specific alternative splicing remodels protein-protein interaction networks.** *Mol Cell* 2012, **46**:884–892.
51. Will CL, Lührmann R: **Spliceosome structure and function.** *Cold Spring Harbor Perspect Biol* 2011, **3**:a003707.
52. Vosseberg J, Stolker D, von der Dunk SHA, Snel B: **Integrating phylogenetics with intron positions illuminates the origin of the complex spliceosome.** *Mol Biol Evol* 2023, **40**, msad011.
53. Guberovic I, Farkas M, Corujo D, Buschbeck M: **Evolution, structure and function of divergent macroH2A1 splice isoforms.** *Semin Cell Dev Biol* 2023, **135**:43–49.
- This manuscript demonstrated structural features found in histone core component H2A macroH2A1.1 isoform that allow it to accommodate ADP-ribose within the binding pocket, while macroH2Z1.2 lacks these features and can't bind ADP-ribose.
54. Guberovic I, Hurtado-Bagès S, Rivera-Casas C, Knobloch G, Malinverni R, Valero V, Leger MM, García J, Basquin J, Gómez de Cedrón M, *et al.*: **Evolution of a histone variant involved in compartmental regulation of NAD metabolism.** *Nat Struct Mol Biol* 2021, **28**:1009–1019.
55. Marti-Solano M, Crilly SE, Malinverni D, Munk C, Harris M, Pearce A, Quon T, Mackenzie AE, Wang X, Peng J, *et al.*: **Combinatorial expression of GPCR isoforms affects signaling and drug responses.** *Nature* 2020, **587**:650–656.
56. Teufel AI, Johnson MM, Laurent JM, Kachroo AH, Marcotte EM, Wilke CO: **The many nuanced evolutionary consequences of duplicated genes.** *Mol Biol Evol* 2019, **36**:304–314.
57. Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O, *et al.*: **Exploring whole-genome duplicate gene retention with complex genetic interaction analysis.** *Science* 2020, **368**:eaaz5667.
58. Mallik S, Tawfik DS, Levy ED: **How gene duplication diversifies the landscape of protein oligomeric state and function.** *Curr Opin Genet Dev* 2022, **76**, 101966.
59. Mallik S, Tawfik DS: **Determining the interaction status and evolutionary fate of duplicated homomeric proteins.** *PLoS Comput Biol* 2020, **16**, e1008145.
60. Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW: **Evolution of increased complexity in a molecular machine.** *Nature* 2012, **481**:360–364.
61. Pillai AS, Chandler SA, Liu Y, Signore AV, Cortez-Romero CR, Benesch JLP, Laganowsky A, Storz JF, Hochberg GKA, Thornton JW: **Origin of complexity in haemoglobin evolution.** *Nature* 2020, **581**:480–485.
62. Hochberg GKA, Shepherd DA, Marklund EG, Santhanagoplan I, Degiacomi MT, Laganowsky A, Allison TM, Basha E, Marty MT, Galpin MR, *et al.*: **Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions.** *Science* 2018, **359**:930–935.
63. Bridoux L, Zarrineh P, Mallen J, Phuycharoen M, Latorre V, Ladam F, Losa M, Baker SM, Sagerstrom C, Mace KA, *et al.*: **HOX paralogs selectively convert binding of ubiquitous transcription factors into tissue-specific patterns of enhancer activation.** *PLoS Genet* 2020, **16**, e1009162.
64. Gerner-Mauro KN, Akiyama H, Chen J: **Redundant and additive functions of the four Lef/Tcf transcription factors in lung epithelial progenitors.** *Proc Natl Acad Sci USA* 2020, **117**:12182–12191.
65. Miramón P, Pountain AW, van Hoof A, Lorenz MC: **The paralogous transcription factors Stp1 and Stp2 of *Candida albicans* have distinct functions in nutrient acquisition and host interaction.** *Infect Immun* 2020, **88**, e00763. 19.
66. Wu C-J, Liu Z-Z, Wei L, Zhou J-X, Cai X-W, Su Y-N, Li L, Chen S, He X-J: **Three functionally redundant plant-specific paralogs are core subunits of the SAGA histone acetyltransferase complex in *Arabidopsis*.** *Mol Plant* 2021, **14**:1071–1087.
67. Wu Y, Wu J, Deng M, Lin Y: **Yeast cell fate control by temporal redundancy modulation of transcription factor paralogs.** *Nat Commun* 2021, **12**:3145.
- This study highlights functionally redundant transcription factor paralogs in yeast, which allow for modulation of temporal response to stress. This may explain why such redundant paralogs may be retained throughout evolution.
68. Zhang Y, Ho TD, Buchler NE, Gordân R: **Competition for DNA binding between paralogous transcription factors determines their genomic occupancy and regulatory functions.** *Genome Res* 2021, **31**:1216–1229.
69. Gera T, Jonas F, More R, Barkai N: **Evolution of binding preferences among whole-genome duplicated transcription factors.** *Elife* 2022, **11**, e73225.
- A study of transcription factor duplicate genes in budding yeast reveals that approximately 60% of transcription factor paralogs have evolved differential DNA binding preferences. Surprisingly, these preferences emerge primarily due to mutations outside the DNA binding domain.
70. Moreira F, Arenas M, Videira A, Pereira F: **Evolutionary history of TOP2A topoisomerases in animals.** *J Mol Evol* 2022, **90**:149–165.
71. Berti M, Teloni F, Mijic S, Ursich S, Fuchs J, Palumbieri MD, Krietsch J, Schmid JA, Garcin EB, Gon S, *et al.*: **Sequential role of RAD51 paralog complexes in replication fork remodeling and restart.** *Nat Commun* 2020, **11**:3531.
72. Bonilla B, Hengel SR, Grundy MK, Bernstein KA: **RAD51 gene family structure and function.** *Annu Rev Genet* 2020, **54**:25–46.
73. Halder S, Ranjha L, Tagliatalata A, Ciccio A, Cejka P: **Strand annealing and motor driven activities of SMARCAL1 and ZRANB3 are stimulated by RAD51 and the paralog complex.** *Nucleic Acids Res* 2022, **50**:8008–8022.
74. Rein HL, Bernstein KA, Baldock RA: **RAD51 paralog function in replicative DNA damage and tolerance.** *Curr Opin Genet Dev* 2021, **71**:86–91.
75. Roy U, Kwon Y, Marie L, Symington L, Sung P, Lisby M, Greene EC: **The Rad51 paralog complex Rad55-Rad57 acts as a molecular chaperone during homologous recombination.** *Mol Cell* 2021, **81**:1043–1057.e8.
76. Liu X, Xie D, Hua Y, Zeng P, Ma L, Zeng F: **Npa3 interacts with Gpn3 and assembly factor Rba50 for RNA polymerase II biogenesis.** *Faseb J* 2020, **34**:15547–15558.
77. Wang X, Gerber A, Chen W-Y, Roeder RG: **Functions of paralogous RNA polymerase III subunits POLR3G and POLR3GL in mouse development.** *Proc Natl Acad Sci USA* 2020, **117**:15702–15711.
78. Barna M, Karbstein K, Tollervey D, Ruggero D, Brar G, Greer EL, Dinman JD: **The promises and pitfalls interestized ribosomes.** *Mol Cell* 2022, **82**:2179–2184.
79. Genuth NR, Barna M: **The discovery of ribosome heterogeneity and its implications for gene regulation and organismal life.** *Mol Cell* 2018, **71**:364–374.
80. Norris K, Hopes T, Aspden JL: **Ribosome heterogeneity and specialization in development.** *WIREs RNA* 2021, **12**:e1644.
- A study characterizing ribosomal protein heterogeneity across 4 different tissues in *Drosophila*. The heterogeneity comes about by means of paralog switching and is particularly found in the testes and ovaries.
81. Lilleorg S, Reier K, Volónkin P, Remme J, Liiv A: **Phenotypic effects of paralogous ribosomal proteins bL31A and bL31B in *E. coli*.** *Sci Rep* 2020, **10**, 11682.
82. Lilleorg S, Reier K, Pulk A, Liiv A, Tammsalu T, Peil L, Cate JHD, Remme J: **Bacterial ribosome heterogeneity: changes in ribosomal protein composition during transition into stationary growth phase.** *Biochimie* 2019, **156**:169–180.
83. Samir P, Browne CM, Rahul, Sun M, Shen B, Li W, Frank J, Link AJ: **Identification of changing ribosome protein compositions using mass spectrometry.** *Proteomics* 2018, **18**, 1800217.
84. Fahl SP, Sertori R, Zhang Y, Contreras AV, Harris B, Wang M, Perrigoue J, Balachandran S, Kennedy BK, Wiest DL: **Loss of ribosomal protein paralog rpl22-like1 blocks lymphoid**

- development without affecting protein synthesis. *J Immunol* 2022, **208**:870–880.
85. Tu C, Meng L, Nie H, Yuan S, Wang W, Du J, Lu G, Lin G, Tan Y-Q: **A homozygous RPL10L missense mutation associated with male factor infertility and severe oligozoospermia.** *Fertil Steril* 2020, **113**:561–568.
 86. Wong QW-L, Li J, Ng SR, Lim SG, Yang H, Vardy LA: **RPL39L is an example of a recently evolved ribosomal protein paralog that shows highly specific tissue expression patterns and is upregulated in ESCs and HCC tumors.** *RNA Biol* 2014, **11**: 33–41.
 87. Zhang D, Zhou Y, Ma Y, Jiang P, Lv H, Liu S, Mu Y, Zhou C, Xiao S, Ji G, *et al.*: **Ribosomal protein L22-like1 (RPL22L1) mediates sorafenib sensitivity via ERK in hepatocellular carcinoma.** *Cell Death Dis* 2022, **8**:1–9.
 88. Zou Q, Qi H: **Deletion of ribosomal paralogs Rpl39 and Rpl39l compromises cell proliferation via protein synthesis and mitochondrial activity.** *Int J Biochem Cell Biol* 2021, **139**, 106070.
 89. Kniepert A, Groettrup M: **The unique functions of tissue-specific proteasomes.** *Trends Biochem Sci* 2014, **39**:17–24.
 90. Murata S, Takahama Y, Kasahara M, Tanaka K: **The immunoproteasome and thymoproteasome: functions, evolution and human disease.** *Nat Immunol* 2018, **19**:923–931.
 91. Gómez-H L, Felipe-Medina N, Condezo YB, Garcia-Valiente R, Ramos I, Suja JA, Barbero JL, Roig I, Sánchez-Martin M, Rooij DG de, *et al.*: **The PSMA8 subunit of the spermatoproteasome is essential for proper meiotic exit and mouse fertility.** *PLoS Genet* 2019, **15**, e1008316.
 92. Zhang Z-H, Jiang T-X, Chen L-B, Zhou W, Liu Y, Gao F, Qiu X-B: **Proteasome subunit $\alpha 4$ s is essential for formation of spermatoproteasomes and histone degradation during meiotic DNA repair in spermatocytes.** *J Biol Chem* 2021:296.
 93. Cascio P: **PA28 γ : new insights on an ancient proteasome activator.** *Biomolecules* 2021, **11**:228.
 94. Chen D-D, Hao J, Shen C-H, Deng X-M, Yun C-H: **Atomic resolution Cryo-EM structure of human proteasome activator PA28 γ .** *Int J Biol Macromol* 2022, **219**:500–507.
 95. Chen J, Wang Y, Xu C, Chen K, Zhao Q, Wang S, Yin Y, Peng C, Ding Z, Cong Y: **Cryo-EM of mammalian PA28 $\alpha\beta$ -iCP immunoproteasome reveals a distinct mechanism of proteasome activation by PA28 $\alpha\beta$.** *Nat Commun* 2021, **12**:739.
 96. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold.** *Nature* 2021, **596**:583–589.