Evolution of Virus-like Features and Intrinsically Disordered Regions in Retrotransposon-derived Mammalian Genes

Rachele Cagliani [iD],[1,*] Diego Forni [iD],[1] Alessandra Mozzi [iD],[1] Rotem Fuchs [iD],[2] Dafna Tussia-Cohen [iD],[2] Federica Arrigoni [iD],[3] Uberto Pozzoli [iD],[1] Luca De Gioia [iD],[3] Tzachi Hagai [iD],[2] Manuela Sironi [iD][1]

[1]Scientific Institute IRCCS E. MEDEA, Computational Biology Unit, Bosisio Parini 23842, Italy

[2]Shmunis School of Biomedicine and Cancer Research, George S Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

[3]Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan 20126, Italy

*Corresponding author: E-mail: rachele.cagliani@lanostrafamiglia.it.

Associate editor: Yong Zhang

Abstract

Several mammalian genes have originated from the domestication of retrotransposons, selfish mobile elements related to retroviruses. Some of the proteins encoded by these genes have maintained virus-like features; including self-processing, capsid structure formation, and the generation of different isoforms through −1 programmed ribosomal frameshifting. Using quantitative approaches in molecular evolution and biophysical analyses, we studied 28 retrotransposon-derived genes, with a focus on the evolution of virus-like features. By analyzing the rate of synonymous substitutions, we show that the −1 programmed ribosomal frameshifting mechanism in three of these genes (*PEG10*, *PNMA3*, and *PNMA5*) is conserved across mammals and originates alternative proteins. These genes were targets of positive selection in primates, and one of the positively selected sites affects a B-cell epitope on the spike domain of the PNMA5 capsid, a finding reminiscent of observations in infectious viruses. More generally, we found that retrotransposon-derived proteins vary in their intrinsically disordered region content and this is directly associated with their evolutionary rates. Most positively selected sites in these proteins are located in intrinsically disordered regions and some of them impact protein posttranslational modifications, such as autocleavage and phosphorylation. Detailed analyses of the biophysical properties of intrinsically disordered regions showed that positive selection preferentially targeted regions with lower conformational entropy. Furthermore, positive selection introduces variation in binary sequence patterns across orthologues, as well as in chain compaction. Our results shed light on the evolutionary trajectories of a unique class of mammalian genes and suggest a novel approach to study how intrinsically disordered region biophysical characteristics are affected by evolution.

Key words: positive selection, intrinsically disordered regions, retrotransposon, domesticated gene, conformational features.

Introduction

Transposable elements (TEs) are mobile genetic units that are able to move and self-amplify within host cells. TEs have successfully populated eukaryotic and prokaryotic genomes and ~50% of the human genome is derived from TEs (Venter et al. 2001; De Koning et al. 2011). Although TEs behave as selfish elements and their replication at the genomic level may be detrimental, increasing evidence indicates that they also contribute significantly to genome evolution (Doolittle and Sapienza 1980; Jangam et al. 2017; Joly-Lopez and Bureau 2018; Schrader and Schmitz 2019; Almojil et al. 2021; Almeida et al. 2022). Indeed, TEs provide the raw material for evolutionary innovation and an extremely wide source of genetic variation. Thus, TEs were shown to contribute gene regulatory elements, serve as seeds for small RNAs and long noncoding RNAs, generate alternative splicing signals, and give rise to novel coding genes or exons. The acquisition of host functions by TEs is referred to as domestication or exaptation (Doolittle and Sapienza 1980; Gould and Vrba 1982; Jangam et al. 2017; Joly-Lopez and Bureau 2018; Schrader and Schmitz 2019; Almojil et al. 2021; Almeida et al. 2022).

In mammals, the most abundant TEs are retrotransposons, which use a "copy and paste" strategy for expansion (Burns and Boeke 2012; Finnegan 2012). Retrotransposons are categorized into two groups: Long terminal repeat (LTR) and nonLTR retrotransposons. LTR retrotransposons

Open Access

contain two identical LTRs, flanking an internal region that typically encodes the group-specific antigen (GAG), polymerase (POL), and envelope (ENV) proteins of retroviral origin (Schrader and Schmitz 2019). LTR retrotransposons were shown to represent an exceptional source of novel coding genes. For instance, syncytin genes, which are essential in many mammals for normal placenta development, originated through multiple domestications of retrotransposon-derived *ENV* genes (Kim et al. 2004; Lavialle et al. 2013). Other mammalian genes instead originated from retrotransposon-derived GAG or POL sequences. These include aspartic peptidase retroviral-like 1 (*ASPRV1*) and activity-regulated cytoskeleton (*ARC*-associated protein), as well as the members of two gene families usually referred to as retrotransposon Gag-like/sushi-ichi retrotransposon homologs (RTL/SIRH) family and paraneoplastic antigen MA (PNMA) family (Campillos et al. 2006; Kaneko-Ishino and Ishino 2023). The RTL/SIRH family comprises *GAG*-derived genes (e.g. *RTL3* to *RTL8*), as well as *RTL1* and *PEG10* (paternally expressed 10), which also retain part of the *POL* open reading frame (ORF). The PNMA family includes several members, some of which are implicated in the development of paraneoplastic syndromes, as well as Coiled-Coil Domain Containing 8 (*CCDC8*).

Some of these genes have maintained retroviral-like features that are unique among mammalian genes. For example, *PEG10*, *PNMA3*, and *PNMA5* produce two alternative transcripts as the result of a −1 programmed ribosomal frameshifting (−1 PRF) mechanism, as observed in retroviruses (Jacks et al. 1988; Wills et al. 2006; Clark et al. 2007). It is however unclear whether the longer isoforms of *PNMA3* and *PNMA5* are translated. Also, the PEG10 protein is capable of self-processing in a way that is reminiscent of retrotransposons, and the protein products of some *GAG*-derived genes, including *ARC*, *PEG10*, *PNMA2*, *PNMA3*, and *PNMA5*, retain the ability to form virus-like particles that traffic between cells (Ashley et al. 2018; Pastuzyn et al. 2018; Segel et al. 2021; Black et al. 2023; Madigan et al. 2024; Xu et al. 2024). In particular, the ARC and PEG10 proteins form capsid structures that are released in extracellular vesicles (Pastuzyn et al. 2018; Pandya et al. 2021). Conversely, PNMA2 is secreted by human cells as a nonenveloped icosahedral capsid (Madigan et al. 2024; Xu et al. 2024). Notably, the PNMA2 capsid is immunogenic and epitopes on the exposed "spike" structure are recognized by antibodies in the cerebrospinal fluid of patients with paraneoplastic syndromes (Xu et al. 2024).

Finally, as is the case of retroviral capsid proteins, several *GAG*-derived mammalian genes encode proteins with intrinsically disordered regions (IDRs) (Goh et al. 2015; Monette et al. 2020, 2022; Beckwith et al. 2023; Kaddis Maldonado et al. 2023). These regions do not adopt a stable 3D structure but rather exist in a collection of structurally distinct conformers known as an ensemble (Holehouse and Kragelund 2024). These ensembles can be considered as the landscape of different conformations that the IDR can adopt and its physicochemical features include conformational entropy (the exploration by the

IDR of multiple structures, where greater values denote a higher number of structures) and chain compaction (how extended or compact the accessible structures are) (Holehouse and Kragelund 2024). In addition, some IDRs can adopt stable folds upon binding with their partners. Despite the lack of a defined 3D structure, IDRs are increasingly recognized as important players in multiple cellular functions and, in the human proteome, about 35% of the residues are embedded within IDRs (Holehouse and Kragelund 2024; Tesei et al. 2024).

In general, IDRs are known to be fast-evolving and tend to be less conserved than structured domains (Brown et al. 2011; Afanasyeva et al. 2018; Holehouse and Kragelund 2024). However, recent evidence has indicated that, while sequence conservation is often lower in IDR than in structured regions, specific IDR features are instead conserved across orthologues. For instance, parameters such as overall amino acid composition, net charge per residue, and binary sequence patterns (the segregation or mixing of specific residues or types of residues in the sequence), which may be related to functional properties, are often conserved across orthologous IDRs (Mao et al. 2010; Moesa et al. 2012; Das et al. 2015; Holehouse et al. 2017; Zarin et al. 2017, 2019). Additionally, short linear motifs are often found in IDRs, where they act as sub-cellular localization signals, as posttranslational modification and cleavage sites, or mediate domain−motif interactions, important for cellular regulation (Fuxreiter et al. 2007; Davey et al. 2012; Van Roey et al. 2014). These motifs were found to be conserved and occur across orthologous proteins in related species; however, their exact location within the IDR may shift between species (Nguyen Ba and Moses 2010; Hagai et al. 2012). To date, most studies on IDR evolution have focused on comparing relatively distant taxa (Zarin et al. 2017; Afanasyeva et al. 2018; Fahmi and Ito 2019; Zarin et al. 2019; Hsu et al. 2021; González-Foutel et al. 2022; Shinn et al. 2022), which makes it difficult to estimate fine scale and site-wise evolutionary rates. Even in the case of studies that analyzed IDRs in relatively closely related taxa, the forces underlying the fast evolution of these regions have remained unknown (Afanasyeva et al. 2018). However, a recent study of SARS-CoV-2 variants found that their genetic differences impact the physicochemical parameters of the nucleocapsid protein IDR (Nguyen et al. 2024).

Herein, we analyzed the evolution of retrotransposon-derived proteins, including their IDRs. Our data expand previous analyses (Henriques et al. 2024) by focusing on the evolution of virus-like features maintained by these genes and by developing a new approach to study IDR evolution. Our findings show that (i) the −1 PRF mechanism in three of these genes is conserved in mammals and originates two alternative proteins; (ii) retrotransposon-derived proteins vary in their IDR content and this is directly associated with their evolutionary rates; (iii) most positively selected sites are in IDRs; (iv) positively selected sites in IDRs are linked to functional regions such as cleavage or phosphorylation sites and tend to be found in

regions with specific biophysical characteristics pertaining to their functions, such as lower conformational entropy. Our work thus provides novel insights into the evolution of retrotransposon-derived proteins as well as a new approach to studying the evolution of IDRs. Namely, by focusing on macro-level biophysical characteristics of IDRs (conformational entropy and chain compaction) and interaction modules embedded within IDRs (linear motifs and cleavage sites), and by investigating whether and how they change across orthologous sequences, we link functions of IDRs with their evolution.

Results

Gene set Assembly and Analysis of Programmed Ribosomal Frameshifting

We assembled a list of 28 GAG-and/or POL-containing coding genes from previous works (Campillos et al. 2006; Kokošar and Kordiš 2013; Segel et al. 2021; Wang and Han 2021). Specifically, from individual studies we retained genes deriving from Metaviridae retrotransposons that contain GAG-derived (ARC, RTL4–RTL10, CCDC8, ZCCHC12, ZCCHC18, MOAP1, PNMA1–PNMA3, PNMA5, PNMA6A/E/F, and PNMA8A/B/C) or POL-derived sequences (ASPRV1) or both (RTL1 and PEG10) (Campillos et al. 2006; Kokošar and Kordiš 2013) (Fig. 1). We focused on the RTL/SIRH and PNMA families, plus ARC and ASPRV1. The proteins encoded by these genes have different domain architectures and many of them encompass IDRs (Fig. 1). As mentioned above, PEG10 was previously reported to encode two proteins of different sizes because of −1 PRF. The same mechanism was shown to operate for the PNMA3 and PNMA5 transcripts. Specifically, the function of the −1 PRF signal from human PEG10, PNMA3, and PNMA5 was validated in vitro by measuring the percentage of frameshifted transcripts, which were higher for PEG10 and PNMA3 (8.5% to 11%) than for PNMA5 (4.2% to 4.9%) (Khan et al. 2022) (supplementary fig. S1a, Supplementary Material online). To assess whether the frameshifted transcripts are translated, we inspected ribosome footprint density in human cells using GWIPS-viz (Michel et al. 2014, 2018). In particular, we inspected global aggregate footprints of elongating ribosomes, which were obtained in several different studies. A clear ribosome footprint signal was evident downstream the termination codon of the GAG-derived ORF for PEG10 and PNMA3 (Fig. 2a). This signal was less clear for PNMA5, which was characterized by an overall lower footprint density. As a control, we analyzed PNMA1, which is not known to undergo −1 PRF: The signal in the 3′ untranslated region (UTR) was not comparable to the one in the translated region (supplementary fig. S1b, Supplementary Material online). Overall, these data are in agreement with those obtained from in vitro transcript analysis and indicate that a minor fraction of PEG10, PNMA3 and, possibly, PNMA5 transcripts are translated from −1 PRF products (supplementary fig. S1a, Supplementary Material online)

(Khan et al. 2022). Notably, although the PNMA3 and PNMA5 sequences downstream of the −1 PRF signals are expected to be POL-derived, they are fully disordered (Fig. 1) and no similarity is observed with any known POL protein.

We next sought to determine whether −1 PRF is conserved in mammals. The mechanism of −1 PRF is mediated by a slippery sequence (5′-GGGAAAC-3′ in all these transcripts) and by an RNA secondary structure element, usually a pseudoknot or a stem-loop, located 5–8 nt downstream of the slippery sequence (Caliskan et al. 2015). The presence of noncoding functional elements within coding sequences can be inferred by searching for regions with a statistically significant reduction in the degree of variability at synonymous sites (Firth 2014). We thus analyzed the sequences of representative placentals that have conserved the three genes and span ~99 million years of mammalian evolution, from armadillos to humans (Kumar et al. 2017; Henriques et al. 2024). We found that the slippery sequence is highly conserved among orthologues in all three genes (Fig. 2b). We next measured the rate of synonymous substitutions (dS) in sliding windows along alignments that encompass the putative frameshifted transcripts from representative mammals (supplementary Additional file S1, Supplementary Material online). In all three gene alignments, a peak of statistically significant reduction in dS was observed exactly in the region that surrounds the slippery sequence (Fig. 2a). This reduction in dS is indicative of the conservation of these specific codons, irrespective of the encoded amino acids (AA), most likely to enable translation from both reading frames. This mechanism is conserved throughout the studied species, suggesting that mammals that retained these tree genes also retained the −1 PRF mechanism.

A recent study (Henriques et al. 2024) confirmed a previous analysis (Brandt et al. 2005) whereby the rodent Rtl3 gene (also known as Zcchc5) also undergoes −1 PRF, whereas this mechanism is lost in primates. We thus retrieved sequence information for six representative species that are predicted to produce the −1 frameshifted transcript (Henriques et al. 2024) (supplementary Additional file S1, Supplementary Material online). For all of them, the same slippery sequence as in the above-analyzed transcripts was found to be fully conserved (Fig. 2b) and a peak of dS reduction was observed in the corresponding position (Fig. 2a). Inspection of footprints of elongating ribosomes for the mouse genome identified a clear signal in the 3′UTR or Zcchc5 (Fig. 2a). Thus, in rodents and other mammals, Zcchc5 is likely to produce two alternative transcripts, both of which are translated into proteins. This underscores the plasticity of these retrotransposon-derived genes, which have maintained distinct virus-like features in different mammals.

Evolutionary Rates Correlate With IDR Fraction

We next sought to gain insight into the evolution of the sequences encoded by the 28 retrotransposon-derived
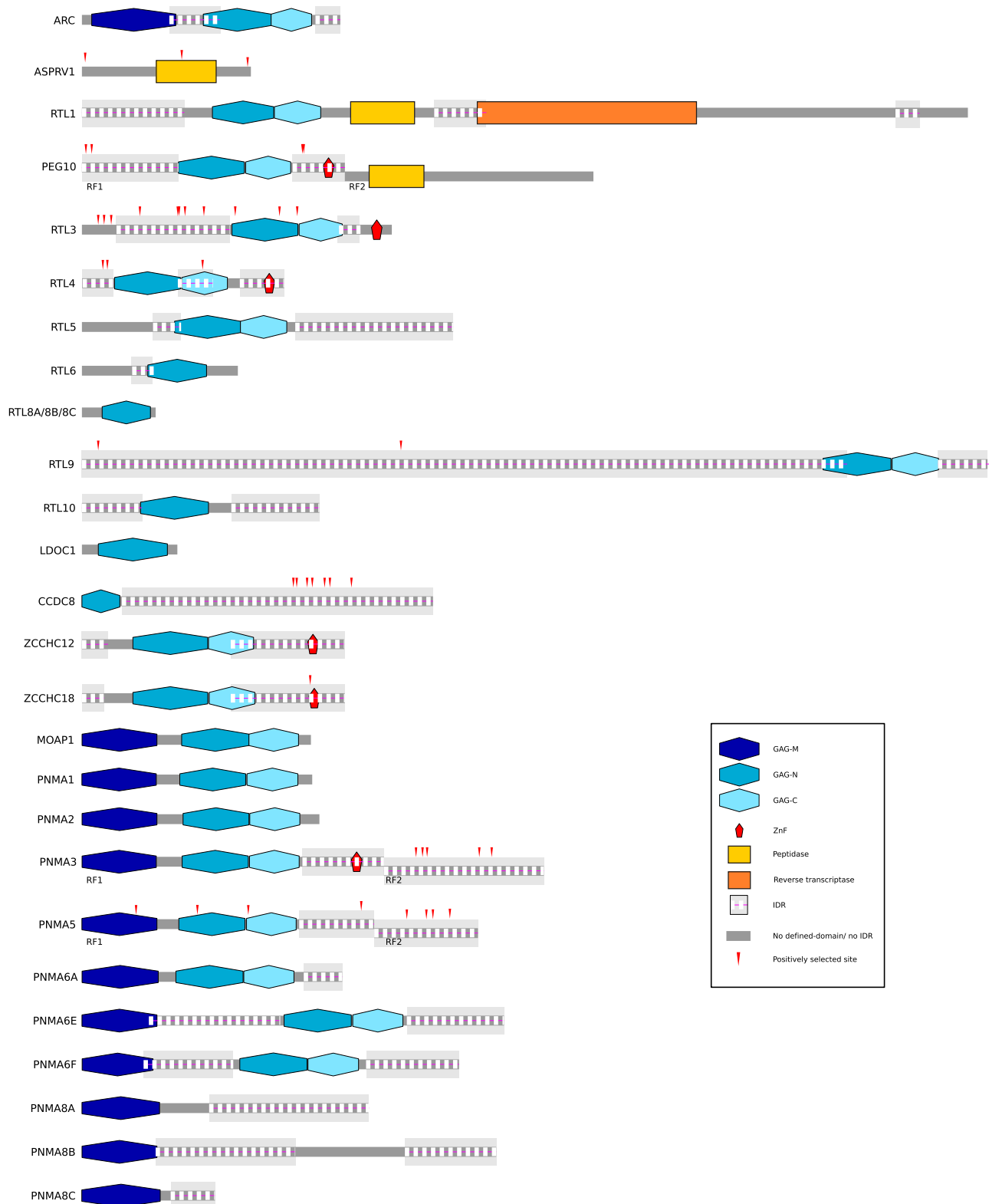
Fig. 1. Domain structures of retrotransposon-derived proteins. The domain structure of the 28 proteins we analyzed is schematically shown. The shaded areas represent IDRs, as per legend. For proteins resulting from −1 PRF products, the 2 regions corresponding to different Reading frames (RF1 and RF2) are shown. The red arrows denote positively selected sites as obtained from positive selection analysis.

genes, including those deriving from translation of the −1 PRF transcripts. Because these retrotransposon-derived genes were domesticated in an early mammalian ancestor, they are common to all primates (with some losses occurring in specific lineages) (Henriques et al. 2024). We thus retrieved sequence information of their coding sequences from available primate genomes, with the aim to analyse evolutionary patterns (supplementary Additional
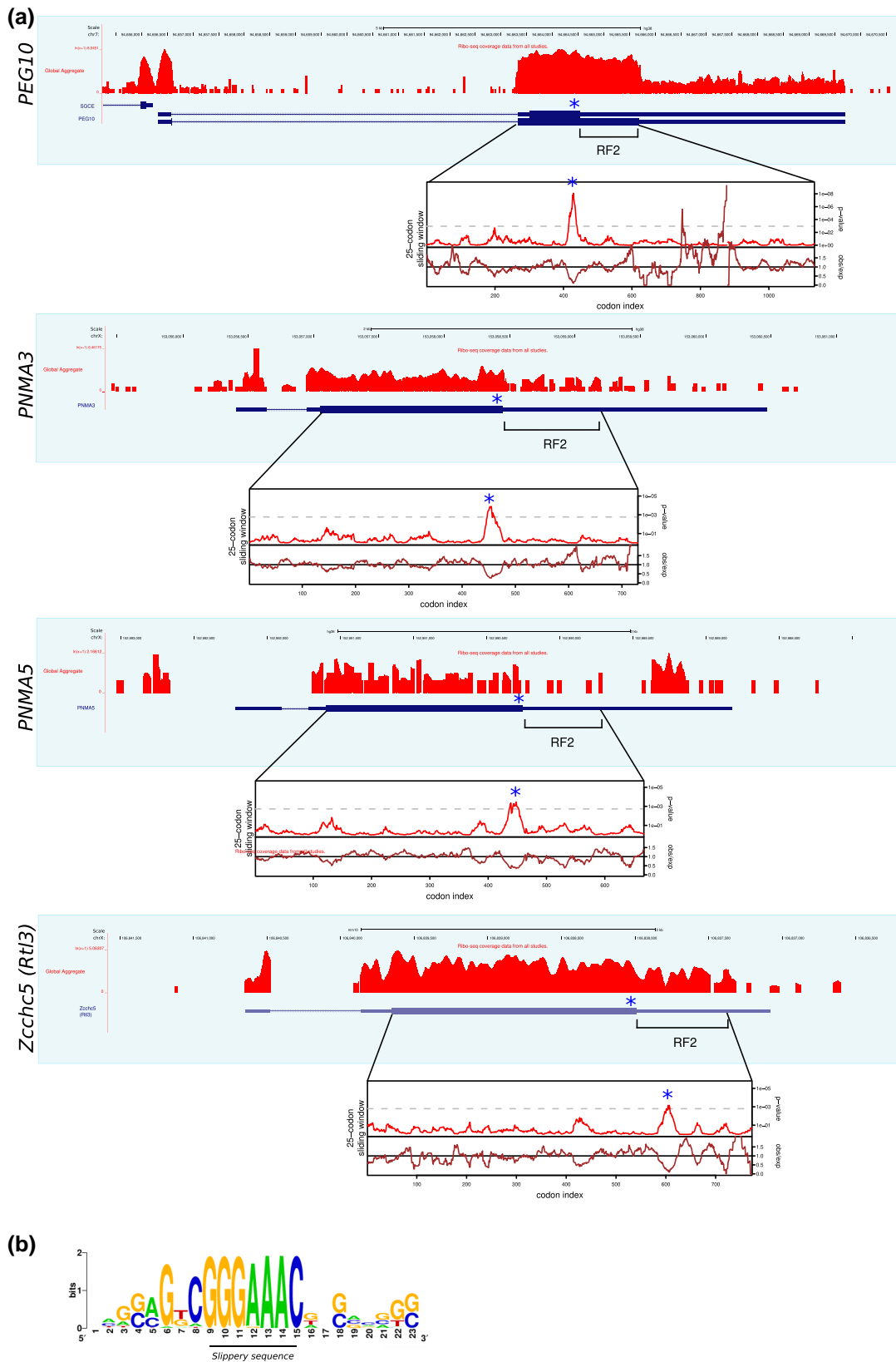
Fig. 2. Analysis of −1 PRF. a) GWIPS-viz visualizations of aggregated ribosome footprint data are shown for *PEG10*, *PNMA3*, *PNMA5*, and *Rtl3*. The region downstream of the putative frameshift site (marked by an asterisk) up to the next −1 frame stop codon is denoted as RF2. For the 3 genes, the enlargements show the distribution of synonymous site variation as obtained from Synplot. The brown line indicates relative dS variability calculated as the ratio of the observed over the expected values of dS in a sliding window of 25 codons. The red line shows the corresponding *P*-value and the dashed line represents the *P*-value cutoff. The position of the slippery sequence is marked with an asterisk. b) Sequence conservation of the slippery sequence and its flanking bases. The letter size represents the normalized frequency of each base calculated for the *PEG10*, *PNMA3*, *PNMA5*, and *Rtl3* sequences.
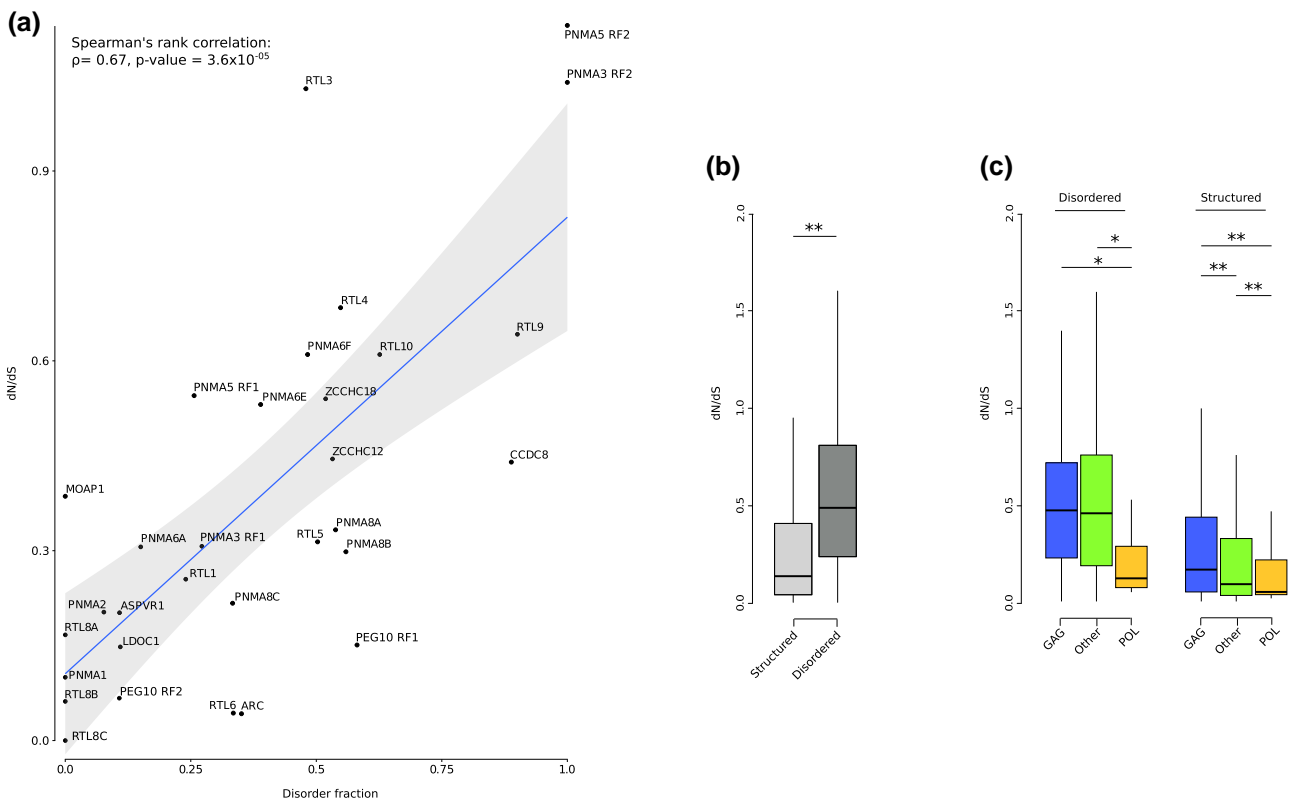
Fig. 3. Evolutionary rates in structured regions and IDRs. a) Correlation between average dN/dS and disorder fraction. b) Codon-wise dN/dS computed for codons in structured regions and for disordered codons. Statistical significance was assessed by the Wilcoxon rank-sum test. c) Codon-wise dN/dS computed for structured regions and disordered codons in different domains. Codons in disordered regions, whether or not overlapping with GAG or other domains, were considered in the disordered fraction. Statistical significance was assessed by Kruskal–Wallis rank sum tests followed by pairwise Nemenyi post hoc tests. *P-value < 0.05; **P-value < 0.01.

file S1, Supplementary Material online). We first calculated the average nonsynonymous substitution/synonymous substitution rate ratio (dN/dS). This metric reflects how rapidly a protein's AA changes relative to synonymous changes. It is commonly used to measure the selective pressure acting on coding sequences: dN/dS < 1 indicates purifying selection, dN/dS around 1 is indicative of neutrality, and dN/dS > 1 reflects positive diversifying selection.

For transcripts that undergo −1 PRF, dN/dS was calculated separately for the two regions with different reading frames (hereafter referred to as RF1 and RF2), and the sequence around the slippery signal was masked. Results indicated that retrotransposon-derived genes evolve at very different rates in primates, with dN/dS ranging from 0 to 1.13 (Fig. 3a). The two fully disordered PNMA3 and PNMA5 RF2 regions had dN/dS higher than 1, whereas the PEG10 RF2 region had lower dN/dS than the GAG-derived RF1 portion. Because IDRs are known to be fast evolving (Holehouse and Kragelund 2024), this is most likely a consequence of the different IDR coverage in the two regions. We thus tested whether, in all retrotransposon-derived genes, dN/dS values were influenced by the fraction of disordered codons. We found a very strong correlation between the evolutionary rates and the percentage of disordered sequence

(Spearman's rank correlation, rho = 0.67, P-value = 3.6×10^{-5}) (Fig. 3a).

To gain further insight into the evolutionary pattern of retrotransposon-derived genes, a codon-wise measure of dN/dS was calculated using Selecton (Doron-Faigenboim et al. 2005; Stern et al. 2007). Unsurprisingly, we found that disordered residues evolve much faster than those in structured regions (Fig. 3b). Among the latter, though, a difference was observed depending on sequence origin, with GAG-derived regions evolving faster than other structured regions, and POL-derived domains having the lowest average dN/dS. POL-derived disordered codons, that are notably fewer than in GAG-derived regions, also had significantly lower dN/dS than other regions (Fig. 3c).

Positive Selection Drives the Evolution of Several Retrotransposon-derived Genes

On one hand, high dN/dS values can result from either positive selection or relaxation of functional constraints. On the other, positive selection can occur even in genes showing, on average, low dN/dS (when only a minority of sites are targeted) (Sironi et al. 2015). To formally test the hypothesis that positive selection is driving the evolution of some retrotransposon-derived genes, we applied likelihood ratio tests implemented in the phylogenetic

MBE

Table 1 Likelihood ratio test statistics for models of variable selective pressure among sites (F3 × 4 codon frequency model)

Gene	No. of species	dN/dS (SLAC)	M8a versus M8		M7 versus M8		Positive selected sites[c]
			$-2\Delta\ln L$[a]	P-value[b] (df: 1)	$-2\Delta\ln L$[a]	P-value[b] (df: 2)	
ASPRV1	31	0.202	**8.619**	**0.0033**	**9.811**	**0.0074**	G5, Q153, R254
PEG10							
ORF1	34	0.151	**5.544**	**0.0185**	**9.239**	**0.00986**	V6, G15, I338, S340
ORF2	34	0.067	0	1	−0.00082	0.9995	
RTL3	21	1.03	**31.922**	**1.60×10^{-8}**	**37.658**	**6.649×10^{-9}**	W25, Q34, A45, P89, A147, I149, P158, P187, D235, E303, H330
RTL4	28	0.684	**5.6713**	**0.0172**	**7.254**	**0.0266**	P32, K39, E185
RTL9	31	0.642	**34.049**	**5.374×10^{-9}**	**45.254**	**1.49×10^{-10}**	Q25, K489
CCDC8	31	0.44	**43.535**	**4.164×10^{-11}**	**53.547**	**2.357×10^{-12}**	P324, A329, A345, A353, P372, T380, H413
ZCCHC18	28	0.54	**7.882**	**0.00499**	**7.885**	**0.0194**	R348
PNMA3							
ORF1	31	0.307	0.4106	0.5216	2.457	0.292693	S512, S522, T529, P609, G628
ORF2	26	1.04	**18.774**	**1.47×10^{-5}**	**18.963**	**7.624×10^{-5}**	
PNMA5							
ORF1	30	0.545	**13.81**	**0.00020**	**18.453**	**9.838×10^{-5}**	P83, D177, I255, G428, A498, R528, A538, T564
ORF2	26	1.13	**5.933**	**0.0149**	**6.054**	**0.0485**	

Statistically significant comparisons are shown in bold.
[a]Twice the difference of likelihood for the 2 models compared.
[b]P-value of rejecting the neutral models (M8a and M7) in favor of the positive selection model (M8). df, degree of freedom.
[c]Positively selected sites detected by at least 2 methods among BEB, MEME, and FUBAR (see Methods).

analysis by maximum likelihood (PAML) suite (Yang 1997, 2007). All genes were screened for recombination and split into different regions if necessary. The neutral models (M7 and M8a) were rejected in favor of the positive selection model (M8) for nine genes (Table 1, supplementary table S1, Supplementary Material online).

To identify positively selected sites in these genes we used three methods: The Bayes empirical Bayes (BEB) analysis from M8, Mixed Effects Model of Evolution (MEME), and Fast, Unconstrained Bayesian AppRoximation (FUBAR). To be conservative, a site was defined as positively selected if it was detected by at least two methods. The average fraction of positively selected sites across the tested genes was 1.3%, with the highest proportion being observed for the PNMA5 RF2 region (2.5%) (Fig. 1). These data are in good agreement with a previous study (Henriques et al. 2024) that found evidence of positive selection in PEG10, RTL3, RTL4, RTL9, PNMA5, ZCCHC18, and CCDC8. This previous work found PNMA6E and PNMA8A to be positively selected, which was not the case in our analyses. Conversely, we detected evidence of selection in ASPRV1 and in the RF2 regions of PNMA3 and PNMA5 (these three genes/regions were not included in the previous analysis) (Henriques et al. 2024). Some discrepancies between the two studies are expected as a different number of primate genomes was included in the analyses. In fact, the power to detect positive selection is influenced by different factors, including the number of taxa and their divergence (Anisimova et al. 2002; Wong et al. 2004).

Positively Selected Sites Are Located in Potentially Functional Regions

We next aimed to investigate the potential functional effects of positive selection. By looking at the positions of the positively selected sites within the proteins, we observe that most of them (70.45%) are found in IDRs (Fig. 1). A notable exception was accounted for by sites in PNMA5, which were located in a structured region of the GAG-derived protein. Recently, PNMA2 was shown to assemble into icosahedral capsids, with the N-terminal residues forming an immunogenic spike on the capsid exterior (Xu et al. 2024). We thus used the structure of PNMA2 to model PNMA5 (RF1 region) and to map positively selected sites onto the protein structure (Fig. 4a). Results indicated that two of the positively selected sites (D177 and I255) are within the capsid interior. Based on the PNMA2 structure, one of them (D177) mediates the interaction between the N-terminal capsid domains of different monomers (Fig. 4c). In contrast, another selected site (P83) is located on the spike region and is surface exposed (Fig. 4a). We thus wondered whether it is part of a B-cell epitope. To assess this possibility, we employed DiscoTope (Kringelum et al. 2012) and BepiPred (Jespersen et al. 2017) tools to predict conformational and linear epitopes. Both methods predicted the selected site to occur within an epitope (Fig. 4b). These observations are reminiscent of findings in viruses, whereby the host immune system often
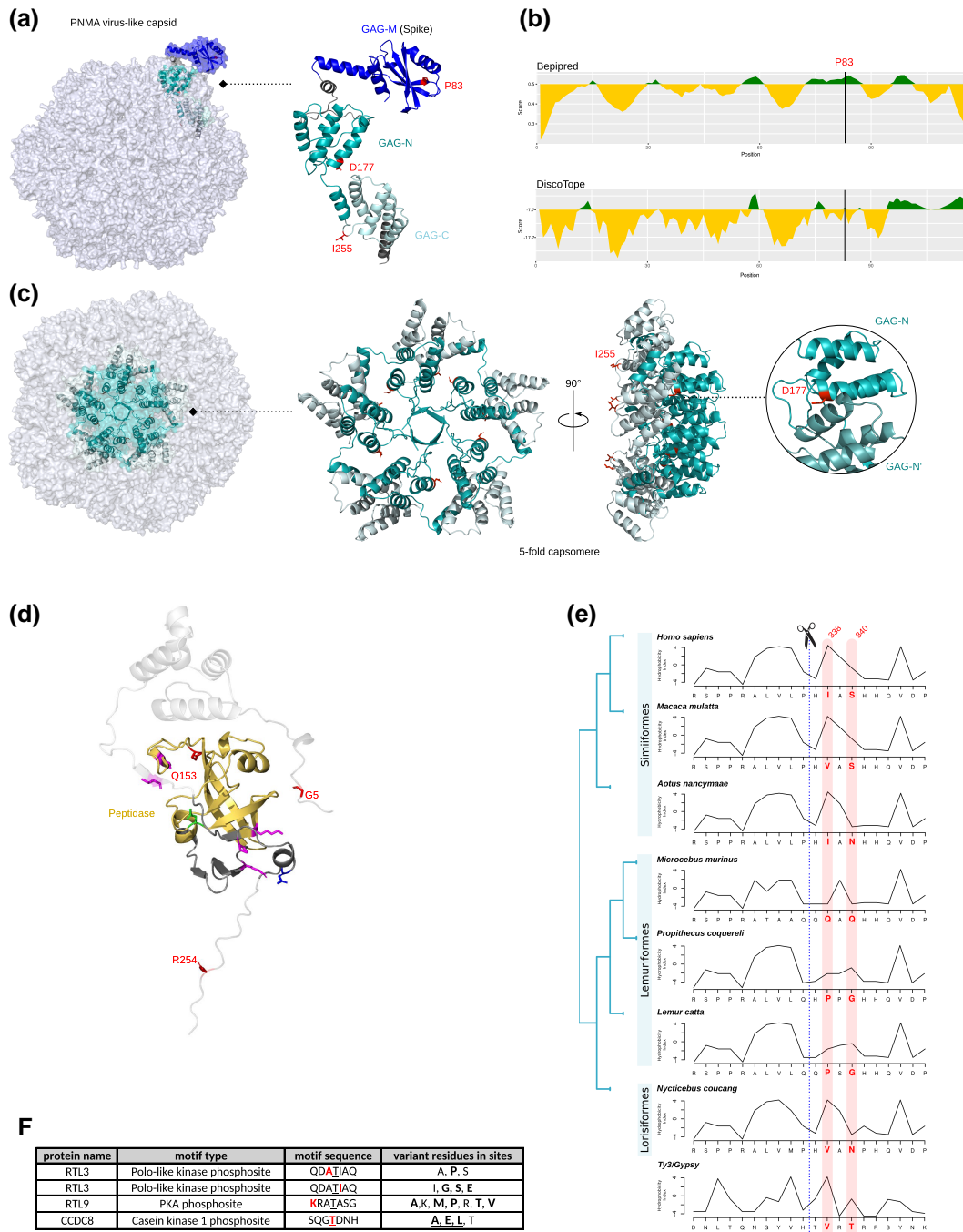
Fig. 4. Functional effects of positively selected sites. a) Molecular model of the virus-like capsid assembly of human PNMA5. The structural model of a PNMA5 monomer from the AlphaFold protein structure database (Q96PV4, aa 1-328) was imposed onto the virus-like capsid assembly and color-coded as in Fig. 1. This monomer is reported as an enlarged ribbon representation on the side. Positively selected sites are in red. b) Prediction score plots of linear and discontinuous epitopes determined for the GAG-N spike domain of PNMA5. Positive prediction is in green; the positively selected site is shown with a black line. c) An isolated 5-fold capsomere is highlighted onto the PNMA5 virus-like capsid model. PNMA5 structural domains are color-coded as in Fig. 1. The 5-fold capsomere is also presented as ribbon in both front and side views. Positively selected sites of each monomer are presented as red sticks. In the enlargement, the CA_GAG-N:CA_GAG-N" interface of 2 different PNMA5 molecules is shown with the D177 marked in red. d) Ribbon representation of the molecular structure of the ASPRV1 model from AlphaFold (Q53RT3 in the AlphaFold protein structure database, aa 84-341). Domain are color-coded as in Fig. 1 and numbering refers to the reference protein sequence (NP_690005). The auto-cleaved mature protein is highlighted, whereas the propeptide regions are in transparency. The catalytic site is in green, positively selected sites in red, residues that when mutated cause ichthyosis or alter protease activity are in magenta, a mutation found in a dog with ichthyosis is in blue. e) Hydrophobicity profiles of the PEG10 autocleavage site in representative primates. The phylogenetic relationships and taxonomic classification are reported on the left. As a comparison the cleavage site of the Ty3/Gypsy retro-transposon (GenBank: CAA97115.1) is also reported. The positively selected sites are marked in red and their position is shaded. f) A table showing the four occurrences of phosphorylation sites that include positively selected sites whose substitutions (in bold) impact the motif and can disrupt the phosphorylation. The phosphorylated residue itself is underlined. The motifs match known regular expression patterns from the ELM database, whose motif IDs are: ELME000442, ELME000008, ELME000063.
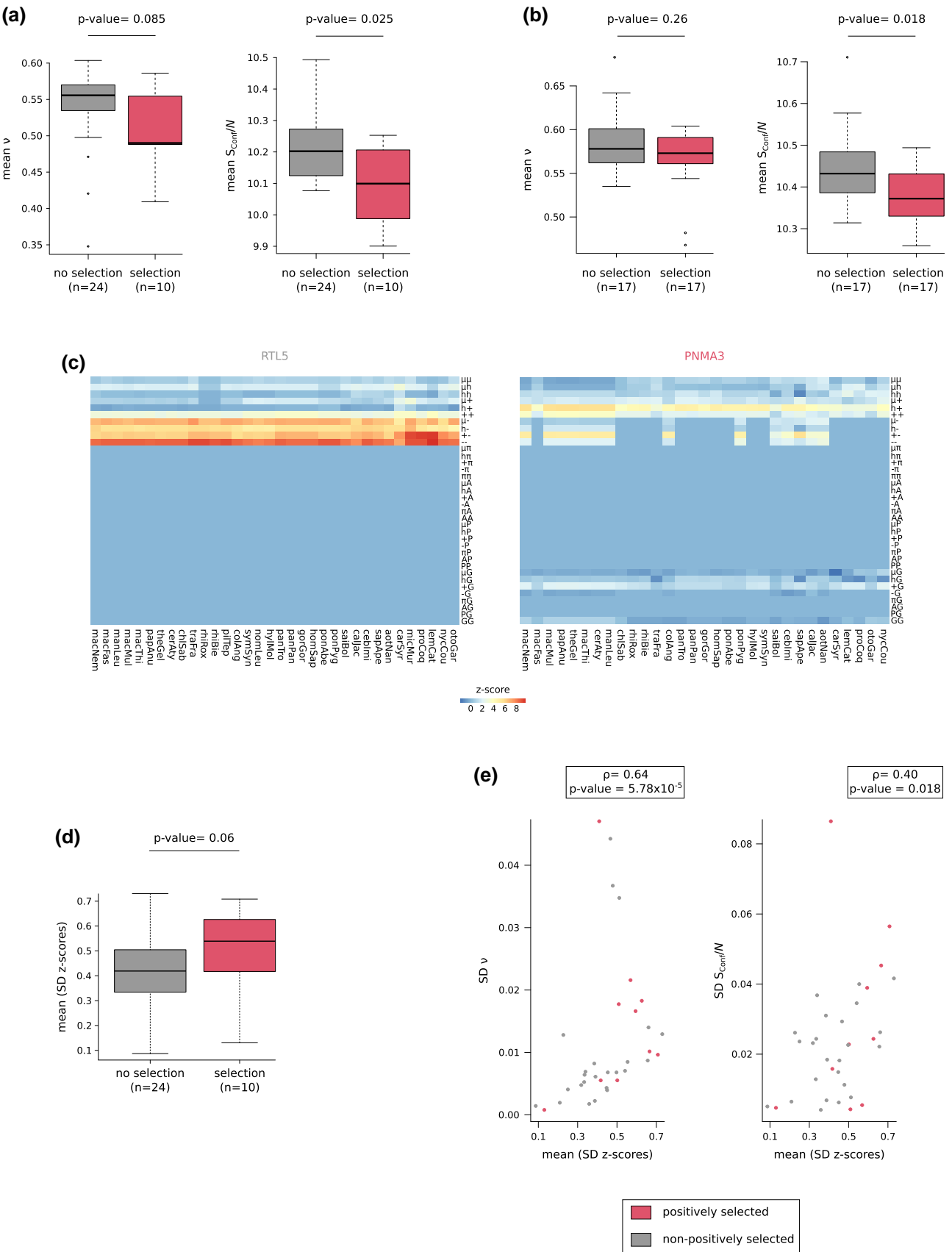
Fig. 5. Comparison of binary patterns and ensemble features across orthologous IDRs. a) The ensemble conformational properties were calculated for IDRs that are (red) or are not (gray) targeted by positive selection. Boxplots show the mean values (among orthologs) of the Flory scaling exponent (ν) and of conformational entropy per residue (S_{conf}/N). Statistical significance was assessed using Wilcoxon rank-sum tests.

(continued)

9

contributes to exert selective pressure. The identification of a similar pattern for a mammalian endogenous gene is quite unusual, and to the best of our knowledge has not been observed previously, adding to the unique behavior of retrotransposon-derived genes (see Discussion).

Three positively selected sites in RTL3 were also located in the structured portion of the GAG-derived domain. Mapping of these sites on the model structure of RTL3 and super-imposition with the PNMA5 monomer indicated that sites H330 in RTL3 and I255 in PNMA5 involve almost corresponding protein regions. Conversely, the other two RTL3 sites are in regions distinct from the ones where PNMA5 selected residues are located (supplementary fig. S2, Supplementary Material online).

We also mapped the three positively selected sites we detected in ASPRV1 onto the 3D model structure (Fig. 4d). Mutations in ASPRV1 cause ichthyosis, a skin disorder, and most changes that alter protease activity occur in the catalytic domain, where one of the positively selected sites we detected (Q153) is also located (Matsui et al. 2011; Boyden et al. 2020) (Fig. 4d). Conversely, the two other positively selected sites are located in the unstructured propeptide regions (which we did not consider as IDRs because they are shorter than 30 AA, see Materials and Methods).

Positive Selection in IDRs May Impact Protein Posttranslational Modifications

We next focused on positively selected sites located in IDRs. IDRs often include motifs that are important for protein regulation, mediating protein–protein interactions, affecting protein cellular localization, and acting as cleavage sites or for posttranslational modifications (Fuxreiter et al. 2007; Davey et al. 2012; Van Roey et al. 2014). In the case of PEG10, as well as of several other retrotransposon-derived proteins, the positively selected sites are located in IDRs. We noticed that two of the positively selected sites are located immediately downstream of the proteolytic cleavage site where PEG10 cleaves itself to generate the nucleocapsid fragment (Black et al. 2023) (Fig. 4e). The proteolytic cleavage sites of

retrotransposons and retroviruses are not conserved in sequence, but they are known to preferentially occur within hydrophobic contexts (specifically, hydrophobic residues are particularly common in the −2, +1 and +2 positions) (Kirchner and Sandmeyer 1993; Black et al. 2023). We thus assessed whether changes at the positively selected residues might change the hydrophobic properties at the cleavage site. To this end, we calculated hydropathy profiles for representative primate proteins that carry different AA at the positively selected sites. A clear drop in hydropathy was observed for the PEG10 proteins encoded by lemurs as compared to the simian counterparts (Fig. 4e). Overall, this suggests that positive selection might have acted to modulate self-processing efficiency at the nucleocapsid cleavage site.

Next, we asked whether some of the positively selected sites in IDRs are located within short linear motifs that are important for protein regulation. Such motifs are often embedded within IDRs and can be identified by searching for sequences that match known motif patterns (regular expressions) (Davey et al. 2012; Kumar et al. 2020). By employing a method we previously used that searches for motif occurrences and links them with relevant protein interaction data (Shuler and Hagai 2022), we detected regions within IDRs that include motif-matching sequences that are likely to be functional (see Materials and Methods). Among these inferred functional motifs, we found nine motifs that include positively selected sites—all of which are phosphorylation sites, recognized by various kinases, including Polo-like kinase, Protein Kinase A, and Casein kinase 1 (Kumar et al. 2020) (supplementary fig. S3a, Supplementary Material online). In four of these cases, the substitutions in the positively selected sites are predicted to compromise the motif and decrease phosphorylation in comparison with the human protein (Fig. 4f, supplementary fig. S3b, Supplementary Material online). In one of these motifs, the positively selected site overlaps the residue that is phosphorylated.

In summary, in the above analyses we have identified several positively selected sites in IDRs with potential importance for protein regulation through self-cleavage or phosphorylation.

Fig. 5. (Continued)

n, number of IDRs b) The same as in (a) but using 30 AA regions that contain (red) or do not contain (gray) one or more positively selected sites. *n*, number of IDRs c) NARDINI analysis of 2 representative IDRs. The *z*-score matrices are shown for all available orthologs. Negative *z*-scores imply that the original sequence is more well mixed with respect to the residue groups compared to the scrambled sequences. Positive *z*-scores indicate nonrandom segregation between 2 types of residues or a blocky distribution of 1 type of residue. *z*-scores close to 0 indicate random patterning. A pattern is considered to be nonrandom if the associated *z*-score is lower than −1.5 or higher than 1.5. Types of residues are categorized as follows: Polar (μ), hydrophobic (h), positively charged (+), negatively charged (−), aromatic (π), alanine (A), proline (P), and glycine (G). Species abbreviations are as follows: *Aotus nancymaae* (aotNan), *Callithrix jacchus* (calJac), *Carlito syrichta* (carSyr), *Cebus imitator* (cebImi), *Cercocebus atys* (cerAty), *Chlorocebus sabaeus* (chlSab), *Colobus angolensis* (colAng), *Gorilla gorilla* (gorGor), *Homo sapiens* (homSap), *Hylobates moloch* (hylMol), *Lemur catta* (lemCat), *Macaca fascicularis* (macFas), *Macaca mulatta* (macMul), *Macaca nemestrina* (macNem), *Macaca thibetana* (macThi), *Mandrillus leucophaeus* (manLeu), *Microcebus murinus* (micMur), *Nomascus leucogenys* (nomLeu), *Nycticebus coucang* (nycCou), *Otolemur garnettii* (otoGar), *Pan paniscus* (panPan), *Pan troglodytes* (panTro), *Papio anubis* (papAnu), *Piliocolobus tephrosceles* (pilTep), *Pongo abelii* (ponAbe), *Pongo pygmaeus* (ponPyg), *Propithecus coquereli* (proCoq), *Rhinopithecus bieti* (rhiBie), *Rhinopithecus roxellana* (rhiRox), *Saimiri boliviensis* (saiBol), *Sapajus apella* (sapApe), *Symphalangus syndactylus* (symSyn), *Theropithecus gelada* (theGel), *Trachypithecus francoisi* (traFra). d) Comparison of variation in binary patterns in positively selected and nonpositively selected IDRs. The boxplot shows the average standard deviation of *z*-scores from NARDINI analysis. Statistical significance was assessed using the Wilcoxon rank-sum test. *n*, number of IDRs e) Correlation between variance in binary patterns (average standard deviation of *z*-scores) and ensemble features (standard deviations of ν and S_{conf}/N). Red dots correspond to positively selected IDRs, gray dots to IDRs that are not positively selected.

Positively Selected IDRs Differ in Ensemble Biophysical Features

We next aimed to investigate several biophysical properties of the positively selected sites located in IDRs and how evolutionary changes may impact these properties. The difficulty in predicting structural features of IDRs has previously hampered efforts to understand their functional roles and evolutionary trajectories (Lindorff-Larsen and Kragelund 2021; Tesei et al. 2024). However, some ensemble properties are quantifiable and can provide information on 3D features and IDR functions. These include the conformational entropy per residue (S_{conf}/N) and the Flory scaling exponent (v), a measure of chain compactness. These features are clearly nonindependent, as IDRs with low S_{conf}/N tend to be more compact (Tesei et al. 2024). Importantly, we reasoned that an analysis in primates was very well suited for analyzing IDR evolution in parallel to assessing these biophysical measures. In fact, the evolutionary distance among taxa is relatively small, making it possible to reliably align the IDRs, enabling to rigorously measure evolutionary rates and infer positively selected sites. We thus used predictors based on support vector regression models to calculate S_{conf}/N and v for all orthologous IDRs in the retrotransposon-derived proteins. For all IDRs, the average dN/dS value was also calculated. We found no significant correlation between dN/dS and mean S_{conf}/N or v (Spearman's rank correlation, both P-values > 0.05). This suggests that the overall evolutionary rate of IDRs is not related to their ensemble conformational features. We next compared ensemble features between IDRs that were targeted by positive selection (i.e. display at least one positively selected site) with those that have no such signatures. Notably, we found that positively selected IDRs have significantly lower S_{conf}/N and tend to be more compact (although in the case of v we did not reach full statistical significance, possibly due to the small number of comparisons: Number of IDRs $= 34$) (Fig. 5a). Given these results, we sought to analyze the relationship between local conformational properties and the action of positive selection. To this end, we focused on human proteins and selected regions of 30 AA in length that contain positively selected sites. When possible, the region was centered around the selected site, after collapsing sites closer than 30 AA. For sites located at the N- or C-termini of IDRs, 30 AA were selected irrespective of site centering (see Materials and Methods). This resulted in 17 regions and, for comparison, an equal number of 30 AA regions was selected from IDRs that do not contain positively selected sites. Calculation of conformational properties for these regions indicated that positively selected regions have significantly lower S_{conf}/N values than the nonselected ones (Fig. 5b). This difference was even more significant than when whole IDRs were analyzed (Fig. 5a). No difference in v values was instead observed between IDRs with positive selections and their control regions (Fig. 5b).

Finally, we tested whether the observed low conformational entropy in positively selected IDRs may be related to conditional folding. For this, we obtained Alphafold2 per-residue predicted local difference test (pLDDT) scores (Jumper et al. 2021). Recently, high pLDDT scores were shown to reflect a region's conditional folding upon binding or after posttranslational modifications (Piovesan et al. 2022; Alderson et al. 2023). No difference in pLDDT scores was observed between positively selected and nonpositively selected IDRs (Wilcoxon rank sum test, $P = 0.81$). Likewise, no difference was detected when 30 AA regions were analyzed (Wilcoxon rank-sum test, $P = 0.51$). Overall, these findings indicate that positive selection in these retrotrasposon-derived proteins preferentially targeted compact IDRs with low conformational entropy, but this may not be due to their propensity to fold upon binding.

Positively Selected IDRs Differ in Binary Patterns Related to Sequence-ensemble Relationship

We next sought to investigate how sequence changes may impact IDR function and ensemble properties. Specific binary sequence patterns, such as the linear clustering or the mixture of specific residue types with respect to one another, were shown to be important determinants of sequence-ensemble relationships in IDRs (Das and Pappu 2013; Das et al. 2015; Martin et al. 2016; Holehouse et al. 2017; Sherry et al. 2017; Zarin et al. 2017, 2019; Beveridge et al. 2019). We thus asked whether there are nonrandom sequence patterns within the IDRs of retrotransposons-derived proteins and to which degree patterns differ across orthologues. To address these questions, we utilized the recently developed NARDINI (nonrandom arrangement of residues in disordered regions inferred using numerical intermixing) algorithm. NARDINI quantifies the extents of mixing or segregation of different pairs of amino acid types and computes a z-score for each binary pattern. NARDINI is well suited to study binary patterns across orthologs, irrespective of the level of sequence conservation (Cohan et al. 2022). Results indicated that distinct IDRs from different proteins have specific nonrandom patterns with variable conservation across orthologous sequences (supplementary fig. S4, Supplementary Material online). For instance, in the IDR region in RTL5 (that has no positively selected residues), several binary patterns were statistically significant across orthologues (e.g. segregation of negatively charged residues into clusters) (Fig. 5c). In the case of the PNMA3 IDR (positively selected), instead, the most prominent nonrandom patterns included the segregation of hydrophobic and positively charged residues and of positively and negatively charged residues. However, especially for the latter pattern, z-scores differed remarkably among orthologues (Fig. 5c).

We thus aimed to determine whether binary patterns differ in conservation among orthologues for IDRs that were or were not targeted by positive selection. For this purpose, we calculated the standard deviations of the z-scores for each pattern among orthologues and then averaged them. In so doing, we obtained a single measure

of pattern conservation for each IDR. A comparison between positively selected and nonpositively selected IDRs indicates that the former has more variability than the latter. Although the P-value is not considered significant, most likely because of the limited sample size ([Fig. 5d](#)), this result may suggest that positive selection modulates IDR functional properties by introducing changes in binary patterns, which are thought to be related to their functions. Future analyses on a larger sample size, when such becomes available, can test the significance of these results.

Finally, we asked whether the degree of binary pattern conservation is related to the variability of ensemble features across orthologues. We thus calculated the standard deviation for conformational entropy and compactness. The results indicated that changes of binary patterns calculated from NARDINI have a greater influence on v compared to S_{conf}/N ([Fig. 5e](#)). In line with these results, we found that the standard deviation of v is higher in positively selected regions compared to nonselected regions, while no difference was observed for S_{conf}/N ([supplementary fig. S5, Supplementary Material](#) online).

Overall, these results indicate that positive selection introduces variation in binary patterns across orthologs, as well as in sequence compaction, whereas it exerts weaker effects on conformational entropy.

Discussion

In this work, we investigated the evolutionary history and protein structural features of retrotransposon-derived genes in primates. Such genes have been maintained in several mammalian lineages for almost 100 million years and were recently shown to be the targets of purifying and positive selection ([Henriques et al. 2024](#)). Whereas these observations suggest that they are beneficial to their hosts, the selective advantages ensuing from their domestication, as well as their functions, are largely unknown. Interestingly, several of these retrotransposon-derived genes have preserved virus-like characteristics, making them unique in the repertoire of mammalian genes. Thus, the primary aim of our work was to assess how virus-like features have been maintained and evolved during the domestication process. Specifically, we (i) focused on the conservation of the −1 PRF mechanism and its impact at the protein level, (ii) investigated selection signatures that involve the capsid-like structures formed by these proteins and their autocatalytic processing, and (iii) used different approaches to gain insights into the evolution of IDRs, which are common in these endogenous proteins as in their retrotransposon/retrovirus ancestors.

Our results indicate that *PEG10*, *PNMA3*, and *PNMA5* produce two transcripts as a result of −1 PRF, both of which are translated into proteins. The −1 PRF signals are conserved in mammals, indicating that the longer products may play some functional role, which is presently unknown. This conclusion is also supported by the fact that we detected positive selection signals in the C-terminal regions of the PNMA3 and PNMA5 long isoforms. In

PNMA5, we also found evidence of positive selection in the GAG-derived region, which was shown to form virus-like capsids. Structural modeling revealed that one of the positively selected sites is exposed on the spike domain and may be part of a B-cell epitope. PNMA2 and other PNMA proteins have been implicated in the development of paraneoplastic syndromes, which occur when solid tumors expressing PNMA proteins elicit autoantibodies, leading to encephalitis ([Schüller et al. 2005](#)). A recent study of PNMA2 showed that the capsid form is antigenic, whereas a capsid-assembly-defective PNMA2 protein is not, and that antibodies preferentially bind to spike capsid epitopes ([Xu et al. 2024](#)). This behavior, which is clearly reminiscent of observations in infectious viruses, has no parallels among other autoantigens. The fact that we detected a positively selected site in a PNMA5 epitope on the spike domain further contributes to the peculiarity of PNMA antigen biology. Infectious viruses and their hosts are expected to be engaged in genetic conflicts that result in the rapid evolution of interacting molecules, with viruses often evolving variants within epitopes to elude humoral or cellular immune responses ([Sironi et al. 2015](#); [Daugherty and Malik 2012](#); [Tenthorey et al. 2022](#); [Crespo-Bellido and Duffy 2023](#); [Markov et al. 2023](#)). In these cases, the host immune response has clearly the purpose of clearing or controlling the virus, whereas the latter aims to escape immune surveillance and infect its host. The reasons why an endogenous protein should form antigenic capsids that carry signatures of positive selection within a B-cell epitope are presently unknown but deserve further exploration. PNMA5 was shown to form capsids, but it is not known to be a target of autoantibodies in paraneoplastic syndromes ([Rosenfeld et al. 2001](#); [Graus and Dalmau 2019](#)). In any case, paraneoplastic syndromes are rare diseases that can hardly be considered as a target of selective pressure. Both PNMA5 and PNMA3 are highly expressed in primate brain, and PNMA5 is also the most abundant member of the PNMA family in unfertilized oocytes, where it contributes to oocyte maturation and fertilization ([Gallardo et al. 2007](#); [Takaji et al. 2009](#); [Zhang et al. 2017](#)). Moreover, Human Protein Atlas data indicate that both PNMA3 and PNMA5 are also expressed in the testis (https://www.proteinatlas.org/, last accessed 2024 May, 29). It is thus possible that the selective pressure acting on PNMA3 and PNMA5 is related to their function in cognition or fertility. Whether their ability to form virus-like capsid has a role in their function and evolution remains a fascinating question for further investigation. Besides PNMA5, other retrovirus-derived genes have very important roles in fertility and reproduction. In particular, *PEG10*, which also forms virus-like particles, undergoes −1 PRF, and is positively selected, is an essential gene in mice. *Peg10* knockout animals show early embryonic lethality owing to defects in placentation ([Ono et al. 2006](#)). This clearly underscores how the exaptation of these retrotransposon-derived elements has provided essential functions to mammalian biology. However, increasing evidence also links *PEG10* with neuropathology. *PEG10* is

highly expressed in the human brain (Pandya et al. 2021). Its function in neurons is unknown, but the PEG10 protein is upregulated in Angelman syndrome (AS), a neurodevelopmental disorder (Pandya et al. 2021). Furthermore, PEG10 is one of the most upregulated proteins in the spinal cord of patients suffering from amyotrophic lateral sclerosis (ALS) (Whiteley et al. 2021; Black et al. 2023). Intriguingly, in both AS and ALS the pathological increase of PEG10 results from the mutational loss of endogenous restrictors (UBE3A and UBQLN2, respectively) that target the retrotransposon-derived protein for proteasome degradation (Pandya et al. 2021; Black et al. 2023). It is also worth noting that PEG10 self-processing generates a nucleocapsid fragment, responsible for the deregulation of axon remodeling genes (Black et al. 2023). We found that two of the positively selected sites in PEG10 are located at the nucleocapsid cleavage site, suggesting that selection might operate to reduce the generation of the potentially toxic nucleocapsid fragment. Overall, these observations indicate that the domestication of these retrotransposons, while providing important functions, has also exposed mammalian cells to dangerous endogenous molecules. We thus hypothesize that natural selection is targeting the virus-like behavior of these genes to limit their potential to cause disease, although experimental analyzes will be necessary to test this possibility.

Two of the retrotransposon-derived genes that were targets of positive selection, *CCDC8* and *ASPRV1*, have been associated to human diseases. *ASPRV1* encodes a mammalian-specific protease which is highly expressed in stratified epithelia and hair follicles. Mutations in *ASPRV1* cause a skin disorder characterized by lamellar ichthyosis (Boyden et al. 2020). One of the positively selected sites we detected is located in the catalytic domain, whereas the other two map to unstructured regions in the N- and C-terminal propeptides. Autoinihibitory modules are often intrinsically disordered (Trudeau et al. 2013; Fenton et al. 2023). In turn, IDRs are inherently sensitive to changes in their physicochemical environment, a feature that may help fine-tuning of transitions from an inhibited to an active state (Nussinov et al. 2020; Moses et al. 2023). Whereas, as is the case for IDRs, autoinhibitory modules are known to be less conserved than their cognate structured domain, previous descriptions of positive selection in their sequences are scant. An interesting possibility is that evolutionary changes in the ASPRV1 propeptide sequences modulated protease activity in response to specific stimuli, eventually influencing skin phenotypes in primates.

CCDC8 is a poorly characterized gene expressed in many tissues and cell types (www.proteinatlas.org/ENSG00000169515-CCDC8/tissue, last accessed 2024 May, 29), where it participates in a pathway that controls growth. In fact, mutations in the gene cause 3-M syndrome, an autosomal-recessive condition characterized by severe postnatal growth defects that result in significantly short stature (Hanson et al. 2011). It is possible that the positively selected sites modulate growth phenotypes and physical dimensions because primates differ

widely in body size. The protein encoded by CCDC8 is almost fully disordered, with only a short structured GAG-derived region at the N-terminus. In CCDC8, as well as in two other IDR-containing retrotransposon-derived proteins, we found that positively selected sites affect phosphorylation sites. Phosphorylation, which is one of the most common posttranslational modification in IDRs (Wright and Dyson 2015), can play key regulatory roles and influence properties such as homomeric or heteromeric interactions, ability to phase separate, and nucleic acid binding affinity and specificity (Wright and Dyson 2015; Modic et al. 2024). In fact, recent evidence indicated that human pathogenic mutations that ablate phosphorylation sites within short linear motifs alter the interactomes of the affected proteins (Rrustemi et al. 2024). Thus, these observations are in line with the notion that positively selected sites are likely functional and exert and modulate certain protein properties.

In general, we found that IDRs are very common in the retrotransposon-derived genes we analyzed. IDRs pose challenges in their analysis due to the generally lower conservation compared to structured regions, as we also show here by the analysis of dN/dS. Moreover, because these regions do not fold into stable secondary or tertiary structures, it is difficult to infer their biological and biochemical functions. As a consequence, IDRs and their evolutionary trajectories remain under-studied (Tesei et al. 2024). To fill this knowledge gap, we thus applied recently developed methods to study binary patterns and ensemble properties of IDRs in relation to evolutionary parameters.

Growing evidence indicates that IDRs with different compaction or conformational entropy preferentially perform specific functions or localize to specific cellular compartments (Sherry et al. 2017; González-Foutel et al. 2022; Tesei et al. 2024). For instance, the IDRs of proteins that undergo homotypic phase separation tend to be compact, whereas those in proteins that function as signaling receptors are more expanded (Sherry et al. 2017; González-Foutel et al. 2022; Ibrahim et al. 2023; Holehouse and Kragelund 2024; Tesei et al. 2024). We thus first asked whether IDRs that were targeted by positive selection differed from the nonselected ones in terms of conformational ensemble features. We found this to be the case, as selected IDRs are, on average, more compact and have lower conformational entropy. Although this is not sufficient to assign them a function, our data represent one of the first descriptions of a link between selective pressure and ensemble properties. This is also interesting because, a recent work showed that human pathogenic variants are more likely to be located in IDRs with low conformational entropy compared to benign variants (Tesei et al. 2024), since such IDRs are less tolerant to change. This observation implicitly confirms that our inference of positive selection is reliable and not the result of constraint relaxation. It also implies that amino acid changes in regions with low conformational entropy are more likely to be functional, as we expect in the case of positively selected sites.

We next aimed to determine how positive selection influences IDR properties. Several studies have demonstrated,

even across highly divergent orthologues with limited identity in primary sequences, that specific features may be conserved, in a manner that may be related to function. These include amino acid composition, sequence length, and net charge per residue (Strickfaden et al. 2007; Mao et al. 2010; Schlessinger et al. 2011; Moesa et al. 2012; Zarin et al. 2017; Bremer et al. 2022). Most of these features can be assessed by calculation of binary patterning parameters that quantify the spatial arrangement of residues, or residue types, within the sequence with respect to other residues (Das et al. 2015; Holehouse et al. 2017; Sherry et al. 2017; Zarin et al. 2017, 2019; Beveridge et al. 2019). We thus quantified the extent to which binary patterns are conserved across orthologous regions. Our hypothesis was that IDRs targeted by positive selection had less conserved binary patterns than those showing no evidence of selection. We verified this hypothesis, although with borderline statistical significance, most likely due to the small number of comparisons. Because changes in binary patterns have previously been associated with functional differences (Cohan et al. 2022; Shinn et al. 2022; Patil et al. 2023), it is conceivable that, through alteration of such patterns, positively selected sites modulate some functional properties of the IDRs where they are located. Indeed, we found significant correlations between the across-orthologues variance in z-scores and the variance of conformational entropy and of the Flory scaling exponent, in line with the view that binary patterns contribute to sequence-ensemble-function relationships (Das et al. 2015; Holehouse et al. 2017; Sherry et al. 2017; Zarin et al. 2017, 2019; Beveridge et al. 2019). Clearly, experimental investigation will be required to determine the effect of positive selection on IDR function in these genes. Nonetheless, our data show that, by leveraging different inferences on sequence and structural features, important insight can be gained on IDR evolutionary trajectories. These approaches are applicable to other proteins with IDRs and hold great potential to expand our understanding of the principles driving IDR evolution.

Materials and Methods

Sequence Retrieval and Alignment

Genes were included based on previous works (Campillos et al. 2006; Kokošar and Kordiš 2013; Segel et al. 2021; Wang and Han 2021). *PNMA6B* was excluded from the analyzes as it is annotated as a pseudogene in humans. For all genes, coding sequence information was retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/, last accessed 2024 May, 29) (supplementary Additional file S1, Supplementary Material online). The RevTrans 2.0 utility was used to generate multiple sequence alignments (MSAs) using MAFFT v6.240 as an aligner (Wernersson and Pedersen 2003). Phylogenetic trees were reconstructed using the phyML program (version 3.1) with a General Time Reversible model plus gamma-distributed rates and 4 substitution rate categories with a fixed proportion of invariable sites (Guindon et al. 2009).

Detection of Synonymous Substitution Reduction and Ribosomal Footprint Data

Synonymous substitution (dS) reduction was calculated using synplot2 program (Firth 2014). This tool is designed to identify overlapping or structural-functional elements along coding sequence alignments. A peculiar signature of these elements is a reduction of dS variability. We used windows of 25 codons, as suggested (Firth 2014), and a P-value threshold was calculated on the basis of the number of windows (i.e. 0.05/number of windows). The analysis was run on the MSAs of the putative frameshifted transcripts from representative mammals (supplementary Additional file S1, Supplementary Material online) (Henriques et al. 2024).

Elongating ribosome footprint data were obtained from the GWIPS-viz browser (https://gwips.ucc.ie/, last accessed 2024 June, 26) (Michel et al. 2014, 2018). For both mouse (GRCm38/mm10) and human (GRCh38/hg38), data from all available studies were used ("global aggregate" track). The following parameters were applied to generate the plots: Overlay method, solid; track height, 128; data view scaling, auto-scale to data view; transform function, LOG [ln(1 + x)]; windowing function, mean; smoothing window, 16 pixels.

The sequence logo was generated using the WebLogo 3 online tool (Schneider and Stephens 1990; Crooks et al. 2004), using the MSA of the regions around the slippery sequence from all representative mammals for *PEG10*, *PNMA3*, *PNMA5*, and *Rtl3* (supplementary Additional file S1, Supplementary Material online) (Henriques et al. 2024). Default parameters were applied.

Evolutionary Analysis in Primates

Primate coding sequences carrying stop codons or with a low sequence coverage were excluded. A list of species and of GenBank accession number for each gene is reported in supplementary Additional file S1, Supplementary Material online. In the case of RTL8A/B/C, because it was difficult to reconstruct orthology due to extensive sequence similarity, we resolved to a phylogenetic approach. Specifically, the transcript sequences of the three genes were obtained from GenBank, aligned, and used to generate a phylogenetic tree which clearly identified three major clusters corresponding to orthologous genes (supplementary fig. S6, Supplementary Material online).

MSAs were analyzed for the presence of recombination signals using Genetic Algorithm Recombination Detection (GARD) (Pond et al. 2006). GARD is a genetic algorithm, which uses phylogenetic incongruence among segments in the alignment to detect the best-fit number and location of recombination breakpoints. When evidence of recombination was detected, the coding alignment was split on the basis of the recombination breakpoints and subregions were used as the input for subsequent molecular evolution analyses. We identified 2 gene alignments showing one recombination event (supplementary table S1, Supplementary Material online). In the case of *PEG10*,

PNMA3, and PNMA5 the region overlapping RF1 and RF2 was masked.

The average nonsynonymous substitution/synonymous substitution rate ratio (dN/dS) parameter for each gene/-gene region was calculated using the single-likelihood ancestor counting (SLAC) method (Kosakovsky Pond and Frost 2005). Inputs were the MSAs and phyML trees.

Positive Selection Analysis

To detect positive selection, the codon-based codeml program implemented in the PAML suite was applied (Yang 1997). Using F3 × 4 codon frequencies model (codon frequencies estimated from the nucleotide frequencies in the data at each codon site) (Yang 1997, 2007), a model (M8, positive selection model) that allows a class of sites to evolve with dN/dS > 1 was compared to two models (M7 and M8a, neutral models) that do not allow dN/dS > 1. To assess statistical significance, twice the difference of the likelihood (ΔlnL) for the models (M8a versus M8 and M7 versus M8) is compared to a χ^2 distribution (1 or 2 degrees of freedom for M8a versus M8 and M7 versus M8 comparisons, respectively). To be conservative and to obtain robust results, we called a gene as a target of positive selection only if it was detected by both M7 versus M8 and M8a versus M8 comparisons. In order to identify specific sites subject to positive selection, we applied 3 different methods: (i) the BEB analysis (with a cutoff of 0.90), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) (Anisimova et al. 2002); (ii) FUBAR (with a cutoff of 0.90), an approximate hierarchical Bayesian method that generates an unconstrained distribution of selection parameters to estimate the posterior probability of positive diversifying selection at each site in a given alignment (Murrell et al. 2013); (iii) MEME (with a cutoff of 0.1), which allows the distribution of dN/dS to vary from site to site and from branch to branch at a site (Murrell et al. 2012). Again, to be conservative and to limit false positives, only sites detected using at least two methods were considered positive selection targets.

GARD, FUBAR, MEME, and SLAC analyses were run locally through the HyPhy suite V 2.5.29 (Pond et al. 2005).

A codon-wise measure of dN/dS was obtained with Selecton (version 2.4) (Stern et al. 2007). For each MSA-tree pair, the M8 model was applied with default parameters.

Analysis of Positively Selected Sites

Protein structures were obtained from previous works (Xu et al. 2024) or from the AlphaFold Protein Structure database (https://alphafold.ebi.ac.uk/, last accessed 2024 May, 29) (Jumper et al. 2021; Varadi et al. 2022). The virus-like capsid model of PNMA5 was obtained by homology through the SWISS-MODEL server (Waterhouse et al. 2018). The recently solved 3D structure of mouse PNMA2 (Xu et al. 2024) (Protein Data Bank, PDB, ID: 8rb3_1) was used as the template. Models were built based

on the target-template alignment using ProMod3 (Studer et al. 2021). The GAG-N and GAG-C domains of human PNMA5 were modeled with quite good accuracy (QMEANDisCo Global: 0.69 ± 0.5). 3D molecular structures were visualized, analyzed, and superimposed using Pymol (PyMOL[TM] Molecular Graphics System, Version 2.4.0, Schrödinger, LLC).

Epitope predictions were obtained with different tools from The Immune Epitope Database (IEDB; https://www.iedb.org/, last accessed 2024 May, 29). In particular, for linear B-cell epitope prediction, we used the Bepipred Linear Epitope Prediction 2.0 tool (Jespersen et al. 2017) with default parameters. Conformational B epitopes were predicted using DiscoTope 1.1 (Kringelum et al. 2012) with a threshold of −7.7 (default value) and using the modeled structure of the PNMA5 spike region (Q96PV4 in the AlphaFold protein structure database).

Hydropathy profiles were generated using the Kyte and Doolittle scale (Kyte and Doolittle 1982), as suggested (Kirchner and Sandmeyer 1993; Black et al. 2023).

Identification of IDRs

IDRs were identified by the Metapredict V2 tool (Emenecker et al. 2021, 2022). This tool defines IDRs by applying a deep-learning algorithm based on a consensus score calculated from eight different disorder predictors (Emenecker et al. 2021). Metapredict V2 was run using default parameters and we considered residues disordered if they were labeled as disordered by the tool. IDRs were instead defined as consecutive disordered stretches longer than 30 residues.

Analysis of Motifs Within IDRs

We studied whether substitutions in positively selected sites may impact the functionality of short linear motifs, by using the defined regular expressions (RegEx) in the Eukaryotic Linear Motif (ELM) database (Kumar et al. 2020) for such motifs. We searched for matches within the proteins' sequences and considered only motif matches that were embedded within the predefined IDRs (since matches embedded in ordered regions are unlikely to be functional).

We next partitioned these motif occurrences into motifs that are likely to be functional and nonfunctional, based on an approach we previously used (Shuler and Hagai 2022). Briefly, we used an existing human protein–protein interaction (PPI) database, taking all interactions that are defined "physical interactions" in the STRING database version 11.5 (Szklarczyk et al. 2021). Across this PPI dataset, we searched for cases where one interactor has a domain and the other interactor has a matching motif (e.g. SH3-domain and SH3-domain binding motif that are found in the first and the second interacting partners, respectively). Thus, matches to motif RegEx patterns were classified as "functional" if we could find a match between this motif sequence and a matching domain in at least one of its known interacting partners. All other matches to

motif patterns were defined as "non functional motifs", under the assumption that many of them represent random matches to the motif sequence.

Analysis of IDR Conformational Properties and alphafold2 Scores

The conformational entropy per residue (S_{conf}/N) and the Flory scaling exponent (v) were calculated for all orthologous IDRs identified by Metapredict (Emenecker et al. 2021, 2022). S_{conf}/N is a measure of the landscape of different structures accessible to an IDR. v is a length-independent measure of chain compaction. It derives from the scaling laws of polymers that describe how chain dimensions vary as a function of chain length (Flory and Volkenstein 1969).

Using a Colab notebook (https://colab.research.google.com/github/KULL-Centre/_2023_Tesei_IDRome/blob/main/IDR_SVR_predictor.ipynb, last accessed 2024 May, 29), S_{conf}/N and v were estimated by a support vector regression model, which was trained on simulations performed using the CALVADOS model (Tesei and Lindorff-Larsen 2022; Tesei et al. 2024). Then, for each IDR we calculated the mean and standard deviation among orthologues.

S_{conf}/N and v were also calculated for 30 AA regions from human proteins selected to contain or not to contain positively selected sites, and for being located in IDRs. Specifically, to obtain an independent set of sequences, we first extracted 30 AA regions for sites that are more than 30 AA apart. For selected sites that are closer than 30 AA, we randomly selected one of them and the region was extracted. When the selected site was located at the N- or C- terminus of the IDR, a 30 AA sequence was selected to include all N- or C- terminal AA and to be 30 AA long. Using these criteria, we obtained a list of 17 regions. The same number of control sequences was randomly selected to be located in IDRs and to be free of positive selection signals.

Binary sequence patterns were calculated by the NARDINI tool (Cohan et al. 2022). NARDINI groups residues in eight different groups: Polar (S, T, N, Q, C, and H), hydrophobic (I, L, M, and V), positive (R and K), negative (E and D), aromatic (F, W, and Y), alanine, proline, and glycine and then it assesses the distribution along the IDR of each group with respect to one another (Cohan et al. 2022). In particular, using 10^5 shuffled sequences and z-scores, nonrandom segregation between two types of groups was evaluated: Positive z-scores indicate that the distribution of the two residue groups is clustered along the IDR, whereas negative z-scores suggest a nonrandom mixing between two residue groups or a uniform distribution along the sequence.

To compare all 36 patterns among orthologues for each IDR but also among IDRs, we calculated the standard deviations of the z-scores of each pattern among orthologues and then we averaged them to generate a single measure for each IDR. NARDINI sets z-scores to 0 if the amino acid sequence contains < 10% of the residue type(s)

contributing to specific patterns. In the calculation of the standard deviation, z-scores equal to 0 were removed.

pLDDT scores were retrieved from the Alpha Fold structure database (https://alphafold.ebi.ac.uk/) (Jumper et al. 2021). We downloaded the scores for the 28 human proteins. In the case of PNMA3 and PNMA5, only the RF1 region was available.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Conflict of Interest

The authors declare no competing interests.

Data Availability

The lists of species analyzed in this study are provided in supplementary Additional file S1, Supplementary Material online.

References

Afanasyeva A, Bockwoldt M, Cooney CR, Heiland I, Gossmann TI. Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res.* 2018:**28**(7):975–982. https://doi.org/10.1101/gr.232645.117.

Alderson TR, Pritišanac I, Kolarić Đ, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc Natl Acad Sci USA.* 2023:**120**(44): e2304302120. https://doi.org/10.1073/pnas.2304302120.

Almeida MV, Vernaz G, Putman ALK, Miska EA. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends Genet.* 2022:**38**(6):529–553. https://doi.org/10.1016/j.tig.2022.02.009.

Almojil D, Bourgeois Y, Falis M, Hariyani I, Wilcox J, Boissinot S. The structural, functional and evolutionary impact of transposable elements in eukaryotes. *Genes (Basel).* 2021:**12**(6):918. https://doi.org/10.3390/genes12060918.

Anisimova M, Bielawski JP, Yang Z. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol.* 2002:**19**(6):950–958. https://doi.org/10.1093/oxfordjournals.molbev.a004152.

Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T. Retrovirus-like gag protein arc1 binds RNA and traffics across synaptic boutons. *Cell.* 2018:**172**(1-2):262–274.e11. https://doi.org/10.1016/j.cell.2017.12.022.

Beckwith SL, Nomberg EJ, Newman AC, Taylor JV, Guerrero-Ferreira RC, Garfinkel DJ. An interchangeable prion-like domain is required for ty1 retrotransposition. *Proc Natl Acad Sci USA.* 2023:**120**(30): e2303358120. https://doi.org/10.1073/pnas.2303358120.

Beveridge R, Migas LG, Das RK, Pappu RV, Kriwacki RW, Barran PE. Ion mobility mass spectrometry uncovers the impact of the

patterning of oppositely charged residues on the conformational distributions of intrinsically disordered proteins. *J Am Chem Soc.* 2019:**141**(12):4908–4918. https://doi.org/10.1021/jacs.8b13483.

Black HH, Hanson JL, Roberts JE, Leslie SN, Campodonico W, Ebmeier CC, Holling GA, Tay JW, Matthews AM, Ung E, et al. UBQLN2 restrains the domesticated retrotransposon PEG10 to maintain neuronal health in ALS. *eLife.* 2023:**12**:e79452. https://doi.org/10.7554/eLife.79452.

Boyden LM, Zhou J, Hu R, Zaki T, Loring E, Scott J, Traupe H, Paller AS, Lifton RP, Choate KA. Mutations in ASPRV1 cause dominantly inherited ichthyosis. *Am J Hum Genet.* 2020:**107**(1):158–163. https://doi.org/10.1016/j.ajhg.2020.05.013.

Brandt J, Veith AM, Volff J-N. A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. *Cytogenet Genome Res.* 2005:**110**(1-4):307–317. https://doi.org/10.1159/000084963.

Bremer A, Farag M, Borcherds WM, Peran I, Martin EW, Pappu RV, Mittag T. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat Chem.* 2022:**14**(2):196–207. https://doi.org/10.1038/s41557-021-00840-w.

Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. *Curr Opin Struct Biol.* 2011:**21**(3):441–446. https://doi.org/10.1016/j.sbi.2011.02.005.

Burns KH, Boeke JD. Human transposon tectonics. *Cell.* 2012:**149**(4):740–752. https://doi.org/10.1016/j.cell.2012.04.019.

Caliskan N, Peske F, Rodnina MV. Changed in translation: mRNA recoding by −1 programmed ribosomal frameshifting. *Trends Biochem Sci.* 2015:**40**(5):265–274. https://doi.org/10.1016/j.tibs.2015.03.006.

Campillos M, Doerks T, Shah P, Bork P. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* 2006:**22**(11):585–589. https://doi.org/10.1016/j.tig.2006.09.006.

Clark MB, Jänicke M, Gottesbühren U, Kleffmann T, Legge M, Poole ES, Tate WP. Mammalian gene PEG10 expresses two Reading frames by high efficiency −1 Frameshifting in Embryonic-associated Tissues. *J Biol Chem.* 2007:**282**(52):37359–37369. https://doi.org/10.1074/jbc.M705676200.

Cohan MC, Shinn MK, Lalmansingh JM, Pappu RV. Uncovering non-random binary patterns within sequences of intrinsically disordered proteins. *J Mol Biol.* 2022:**434**(2):167373. https://doi.org/10.1016/j.jmb.2021.167373.

Crespo-Bellido A, Duffy S. The how of counter-defense: viral evolution to combat host immunity. *Curr Opin Microbiol.* 2023:**74**:102320. https://doi.org/10.1016/j.mib.2023.102320.

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004:**14**(6):1188–1190. https://doi.org/10.1101/gr.849004.

Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci USA.* 2013:**110**(33):13392–13397. https://doi.org/10.1073/pnas.1304749110.

Das RK, Ruff KM, Pappu RV. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol.* 2015:**32**:102–112. https://doi.org/10.1016/j.sbi.2015.03.008.

Daugherty MD, Malik HS. Rules of engagement: molecular insights from host-virus arms races. *Annu Rev Genet.* 2012:**46**(1):677–700. https://doi.org/10.1146/annurev-genet-110711-155522.

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. Attributes of short linear motifs. *Mol Biosyst.* 2012:**8**(1):268–281. https://doi.org/10.1039/C1MB05231D.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 2011:**7**(12):e1002384. https://doi.org/10.1371/journal.pgen.1002384.

Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature.* 1980:**284**(5757):601–603. https://doi.org/10.1038/284601a0.

Doron-Faigenboim A, Stern A, Mayrose I, Bacharach E, Pupko T. Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics.* 2005:**21**(9):2101–2103. https://doi.org/10.1093/bioinformatics/bti259.

Emenecker RJ, Griffith D, Holehouse AS. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys J.* 2021:**120**(20):4312–4319. https://doi.org/10.1016/j.bpj.2021.08.039.

Emenecker RJ, Griffith D, Holehouse AS. Metapredict V2: an update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *bioRxiv 494887.* https://doi.org/10.1101/2022.06.06.494887, 9 June 2022, preprint: not peer reviewed.

Fahmi M, Ito M. Evolutionary approach of intrinsically disordered CIP/KIP proteins. *Sci Rep.* 2019:**9**(1):1575. https://doi.org/10.1038/s41598-018-37917-5.

Fenton M, Gregory E, Daughdrill G. Protein disorder and autoinhibition: the role of multivalency and effective concentration. *Curr Opin Struct Biol.* 2023:**83**:102705. https://doi.org/10.1016/j.sbi.2023.102705.

Finnegan DJ. Retrotransposons. *Curr Biol.* 2012:**22**(11):R432–R437. https://doi.org/10.1016/j.cub.2012.04.025.

Firth AE. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res.* 2014:**42**(20):12425–12439. https://doi.org/10.1093/nar/gku981.

Flory PJ, Volkenstein M. Statistical mechanics of chain molecules. *Biopolymers.* 1969:**8**(5):699–700. https://doi.org/10.1002/bip.1969.360080514.

Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics.* 2007:**23**(8):950–956. https://doi.org/10.1093/bioinformatics/btm035.

Gallardo TD, John GB, Shirley L, Contreras CM, Akbay EA, Haynie JM, Ward SE, Shidler MJ, Castrillon DH. Genomewide discovery and classification of candidate ovarian fertility genes in the mouse. *Genetics.* 2007:**177**(1):179–194. https://doi.org/10.1534/genetics.107.074823.

Goh GK-M, Dunker AK, Uversky VN. Shell disorder, immune evasion and transmission behaviors among human and animal retroviruses. *Mol Biosyst.* 2015:**11**(8):2312–2323. https://doi.org/10.1039/C5MB00277J.

González-Foutel NS, Glavina J, Borcherds WM, Safranchik M, Barrera-Vilarmau S, Sagar A, Estaña A, Barozet A, Garrone NA, Fernandez-Ballester G, et al. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat Struct Mol Biol.* 2022:**29**(8):781–790. https://doi.org/10.1038/s41594-022-00811-w.

Gould SJ, Vrba ES. Exaptation—a missing term in the science of form. *Paleobiology.* 1982:**8**(1):4–15. https://doi.org/10.1017/S0094837300004310.

Graus F, Dalmau J. Paraneoplastic neurological syndromes in the era of immune-checkpoint inhibitors. *Nat Rev Clin Oncol.* 2019:**16**(9):535–548. https://doi.org/10.1038/s41571-019-0194-4.

Guindon S, Delsuc F, Dufayard J-F, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol.* 2009:**537**:113–137. https://doi.org/10.1007/978-1-59745-251-9_6.

Hagai T, Tóth-Petróczy Á, Azia A, Levy Y. The origins and evolution of ubiquitination sites. *Mol Biosyst.* 2012:**8**(7):1865. https://doi.org/10.1039/c2mb25052g.

Hanson D, Murray PG, O'Sullivan J, Urquhart J, Daly S, Bhaskar SS, Biesecker LG, Skae M, Smith C, et al. Exome sequencing identifies CCDC8 mutations in 3-M syndrome, suggesting that CCDC8 contributes in a pathway with CUL7 and OBSL1 to control human growth. *Am J Hum Genet.* 2011:**89**(1):148–153. https://doi.org/10.1016/j.ajhg.2011.05.028.

Henriques WS, Young JM, Nemudryi A, Nemudraia A, Wiedenheft B, Malik HS. The diverse evolutionary histories of domesticated metaviral capsid genes in mammals. *Mol Biol Evol.* 2024:**41**(4):msae061. https://doi.org/10.1093/molbev/msae061.

Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu RV. CIDER: resources to analyze sequence-ensemble relationships of

intrinsically disordered proteins. *Biophys J*. 2017:**112**(1):16–21. https://doi.org/10.1016/j.bpj.2016.11.3200.

Holehouse AS, Kragelund BB. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol*. 2024:**25**(3):187–211. https://doi.org/10.1038/s41580-023-00673-0.

Hsu IS, Strome B, Lash E, Robbins N, Cowen LE, Moses AM. A functionally divergent intrinsically disordered region underlying the conservation of stochastic signaling. *PLoS Genet*. 2021:**17**(9): e1009629. https://doi.org/10.1371/journal.pgen.1009629.

Ibrahim AY, Khaodeuanepheng NP, Amarasekara DL, Correia JJ, Lewis KA, Fitzkee NC, Hough LE, Whitten ST. Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. *J Biol Chem*. 2023:**299**(1):102801. https://doi.org/10.1016/j.jbc.2022.102801.

Jacks T, Power MD, Masiarz FR, Luciw PA, Barr PJ, Varmus HE. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*. 1988:**331**(6153):280–283. https://doi.org/10.1038/331280a0.

Jangam D, Feschotte C, Betrán E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet*. 2017:**33**(11):817–831. https://doi.org/10.1016/j.tig.2017.07.011.

Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*. 2017:**45**(W1):W24–W29. https://doi.org/10.1093/nar/gkx346.

Joly-Lopez Z, Bureau TE. Exaptation of transposable element coding sequences. *Curr Opin Genet Dev*. 2018:**49**:34–42. https://doi.org/10.1016/j.gde.2018.02.011.

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**(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kaddis Maldonado R, Lambert GS, Rice BL, Sudol M, Flanagan JM, Parent LJ. The Rous sarcoma virus Gag polyprotein forms biomolecular condensates driven by intrinsically-disordered regions. *J Mol Biol*. 2023:**435**(16):168182. https://doi.org/10.1016/j.jmb.2023.168182.

Kaneko-Ishino T, Ishino F. Retrovirus-Derived RTL/SIRH: their diverse roles in the current eutherian developmental system and contribution to eutherian evolution. *Biomolecules*. 2023:**13**(10):1436. https://doi.org/10.3390/biom13101436.

Khan YA, Loughran G, Steckelberg A-L, Brown K, Kiniry SJ, Stewart H, Baranov PV, Kieft JS, Firth AE, Atkins JF. Evaluating ribosomal frameshifting in CCR5 mRNA decoding. *Nature*. 2022:**604**(7906): E16–E23. https://doi.org/10.1038/s41586-022-04627-y.

Kim FJ, Battini J-L, Manel N, Sitbon M. Emergence of vertebrate retroviruses and envelope capture. *Virology*. 2004:**318**(1):183–191. https://doi.org/10.1016/j.virol.2003.09.026.

Kirchner J, Sandmeyer S. Proteolytic processing of Ty3 proteins is required for transposition. *J Virol*. 1993:**67**(1):19–28. https://doi.org/10.1128/jvi.67.1.19-28.1993.

Kokošar J, Kordiš D. Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. *Mol Biol Evol*. 2013:**30**(5):1015–1031. https://doi.org/10.1093/molbev/mst014.

Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*. 2005:**22**(5):1208–1222. https://doi.org/10.1093/molbev/msi105.

Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol*. 2012:**8**(12):e1002829. https://doi.org/10.1371/journal.pcbi.1002829.

Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, Diakogianni A, Valverde JA, Bukirova D, Čalyševa J, *et al*. ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res*. 2020:**48**(D1):D296–D306. https://doi.org/10.1093/nar/gkz1030.

Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 2017:**34**(7):1812–1819. https://doi.org/10.1093/molbev/msx116.

Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982:**157**(1):105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Phil Trans R Soc B*. 2013:**368**(1626):20120507. https://doi.org/10.1098/rstb.2012.0507.

Lindorff-Larsen K, Kragelund BB. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J Mol Biol*. 2021:**433**(20):167196. https://doi.org/10.1016/j.jmb.2021.167196.

Madigan V, Zhang Y, Raghavan R, Wilkinson ME, Faure G, Puccio E, Segel M, Lash B, Macrae RK, Zhang F. Human paraneoplastic antigen ma2 (PNMA2) forms icosahedral capsids that can be engineered for mRNA delivery. *Proc Natl Acad Sci USA*. 2024:**121**(11):e2307812120. https://doi.org/10.1073/pnas.2307812120.

Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci USA*. 2010:**107**(18): 8183–8188. https://doi.org/10.1073/pnas.0911107107.

Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, Katzourakis A. The evolution of SARS-CoV-2. *Nat Rev Microbiol*. 2023:**21**(6):361–379. https://doi.org/10.1038/s41579-023-00878-2.

Martin EW, Holehouse AS, Grace CR, Hughes A, Pappu RV, Mittag T. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *J Am Chem Soc*. 2016:**138**(47):15323–15335. https://doi.org/10.1021/jacs.6b10272.

Matsui T, Miyamoto K, Kubo A, Kawasaki H, Ebihara T, Hata K, Tanahashi S, Ichinose S, Imoto I, Inazawa J, *et al*. SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing. *EMBO Mol Med*. 2011:**3**(6):320–333. https://doi.org/10.1002/emmm.201100140.

Michel AM, Fox G, Kiran AM, De Bo C, O'Connor PBF, Heaphy SM, Mullan JPA, Donohue CA, Higgins DG, Baranov PV. GWIPS-viz: development of a ribo-seq genome browser. *Nucl Acids Res*. 2014:**42**(D1):D859–D864. https://doi.org/10.1093/nar/gkt1035.

Michel AM, Kiniry SJ, O'Connor PBF, Mullan JP, Baranov PV. GWIPS-viz: 2018 update. *Nucleic Acids Res*. 2018:**46**(D1): D823–D830. https://doi.org/10.1093/nar/gkx790.

Modic M, Adamek M, Ule J. The impact of IDR phosphorylation on the RNA binding profiles of proteins. *Trends Genet*. 2024:**40**(7): 580–586. https://doi.org/10.1016/j.tig.2024.04.004.

Moesa HA, Wakabayashi S, Nakai K, Patil A. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol Biosyst*. 2012:**8**(12):3262. https://doi.org/10.1039/c2mb25202c.

Monette A, Niu M, Chen L, Rao S, Gorelick RJ, Mouland AJ. Pan-retroviral nucleocapsid-mediated phase separation regulates genomic RNA positioning and trafficking. *Cell Rep*. 2020:**31**(3):107520. https://doi.org/10.1016/j.celrep.2020.03.084.

Monette A, Niu M, Nijhoff Asser M, Gorelick RJ, Mouland AJ. Scaffolding viral protein NC nucleates phase separation of the HIV-1 biomolecular condensate. *Cell Rep*. 2022:**40**(8):111251. https://doi.org/10.1016/j.celrep.2022.111251.

Moses D, Ginell GM, Holehouse AS, Sukenik S. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. *Trends Biochem Sci*. 2023:**48**(12):1019–1034. https://doi.org/10.1016/j.tibs.2023.08.001.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, Scheffler K. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. *Mol Biol Evol*. 2013:**30**(5): 1196–1205. https://doi.org/10.1093/molbev/mst030.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*. 2012:**8**(7):e1002764. https://doi.org/10.1371/journal.pgen.1002764.

Nguyen A, Zhao H, Myagmarsuren D, Srinivasan S, Wu D, Chen J, Piszczek G, Schuck P. Modulation of biophysical properties of nucleocapsid protein in the mutant spectrum of SARS-CoV-2. *eLife*. 2024:**13**:RP94836. https://doi.org/10.7554/eLife.94836.3.

Nguyen Ba AN, Moses AM. Evolution of characterized phosphorylation sites in budding yeast. *Mol Biol Evol*. 2010:**27**(9): 2027–2037. https://doi.org/10.1093/molbev/msq090.

Nussinov R, Tsai C-J., Jang H. Autoinhibition can identify rare driver mutations and advise pharmacology. *FASEB J*. 2020:**34**(1):16–29. https://doi.org/10.1096/fj.201901341R.

Ono R, Nakamura K, Inoue K, Naruse M, Usami T, Wakisaka-Saito N, Hino T, Suzuki-Migishima R, Ogonuki N, Miki H, *et al*. Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet*. 2006:**38**(1): 101–106. https://doi.org/10.1038/ng1699.

Pandya NJ, Wang C, Costa V, Lopatta P, Meier S, Zampeta FI, Punt AM, Mientjes E, Grossen P, Distler T, *et al*. Secreted retrovirus-like GAG-domain-containing protein PEG10 is regulated by UBE3A and is involved in Angelman syndrome pathophysiology. *Cell Rep Med*. 2021:**2**(8):100360. https://doi.org/10.1016/j.xcrm.2021.100360.

Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, *et al*. The neuronal gene arc encodes a repurposed retrotransposon gag protein that mediates intercellular RNA transfer. *Cell*. 2018:**172**(1-2): 275–288.e18. https://doi.org/10.1016/j.cell.2017.12.024.

Patil A, Strom AR, Paulo JA, Collings CK, Ruff KM, Shinn MK, Sankar A, Cervantes KS, Wauer T, St. Laurent JD, *et al*. A disordered region controls cBAF activity via condensation and partner recruitment. *Cell*. 2023:**186**(22):4936–4955.e26. https://doi.org/10.1016/j.cell.2023.08.032.

Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci*. 2022:**31**(11): e4466. https://doi.org/10.1002/pro.4466.

Pond SLK, Frost SDW, Muse SV. Hyphy: hypothesis testing using phylogenies. *Bioinformatics*. 2005:**21**(5):676–679. https://doi.org/10.1093/bioinformatics/bti079.

Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol*. 2006:**23**(10):1891–1901. https://doi.org/10.1093/molbev/msl051.

Rosenfeld MR, Eichen JG, Wade DF, Posner JB, Dalmau J. Molecular and clinical diversity in paraneoplastic immunity to Ma proteins. *Ann Neurol*. 2001:**50**(3):339–348. https://doi.org/10.1002/ana.1288.

Rrustemi T, Meyer K, Roske Y, Uyar B, Akalin A, Imami K, Ishihama Y, Daumke O, Selbach M. Pathogenic mutations of human phosphorylation sites affect protein–protein interactions. *Nat Commun*. 2024:**15**(1):3146. https://doi.org/10.1038/s41467-024-46794-8.

Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder—a breakthrough invention of evolution? *Curr Opin Struct Biol*. 2011:**21**(3):412–418. https://doi.org/10.1016/j.sbi.2011.03.014.

Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucl Acids Res*. 1990:**18**(20):6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. *Mol Ecol*. 2019:**28**(6):1537–1549. https://doi.org/10.1111/mec.14794.

Schüller M, Jenne D, Voltz R. The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. *J Neuroimmunol*. 2005:**169**(1-2):172–176. https://doi.org/10.1016/j.jneuroim.2005.08.019.

Segel M, Lash B, Song J, Ladha A, Liu CC, Jin X, Mekhedov SL, Macrae RK, Koonin EV, Zhang F. Mammalian retrovirus-like protein

PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. *Science*. 2021:**373**(6557):882–889. https://doi.org/10.1126/science.abg6155.

Sherry KP, Das RK, Pappu RV, Barrick D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the notch receptor. *Proc Natl Acad Sci USA*. 2017:**114**(44):E9243–E9252. https://doi.org/10.1073/pnas.1706083114.

Shinn MK, Cohan MC, Bullock JL, Ruff KM, Levin PA, Pappu RV. Connecting sequence features within the disordered C-terminal linker of *Bacillus subtilis* FtsZ to functions and bacterial cell division. *Proc Natl Acad Sci USA*. 2022:**119**(42): e2211178119. https://doi.org/10.1073/pnas.2211178119.

Shuler G, Hagai T. Rapidly evolving viral motifs mostly target biophysically constrained binding pockets of host proteins. *Cell Rep*. 2022:**40**(7):111212. https://doi.org/10.1016/j.celrep.2022.111212.

Sironi M, Cagliani R, Forni D, Clerici M. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat Rev Genet*. 2015:**16**(4):224–236. https://doi.org/10.1038/nrg3905.

Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res*. 2007:**35**(Web Server):W506–W511. https://doi.org/10.1093/nar/gkm382.

Strickfaden SC, Winters MJ, Ben-Ari G, Lamson RE, Tyers M, Pryciak PM. A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell*. 2007:**128**(3):519–531. https://doi.org/10.1016/j.cell.2006.12.032.

Studer G, Tauriello G, Bienert S, Biasini M, Johner N, Schwede T. ProMod3-A versatile homology modelling toolbox. *PLoS Comput Biol*. 2021:**17**(1):e1008667. https://doi.org/10.1371/journal.pcbi.1008667.

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, *et al*. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021:**49**(D1):D605–D612. https://doi.org/10.1093/nar/gkaa1074.

Takaji M, Komatsu Y, Watakabe A, Hashikawa T, Yamamori T. Paraneoplastic antigen-like 5 gene (PNMA5) is preferentially expressed in the association areas in a primate specific manner. *Cereb Cortex*. 2009:**19**(12):2865–2879. https://doi.org/10.1093/cercor/bhp062.

Tenthorey JL, Emerman M, Malik HS. Evolutionary landscapes of host-virus arms races. *Annu Rev Immunol*. 2022:**40**(1):271–294. https://doi.org/10.1146/annurev-immunol-072621-084422.

Tesei G, Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res Europe*. 2022:**2**:94. https://doi.org/10.12688/openreseurope.14967.2.

Tesei G, Trolle AI, Jonsson N, Betz J, Knudsen FE, Pesce F, Johansson KE, Lindorff-Larsen K. Conformational ensembles of the human intrinsically disordered proteome. *Nature*. 2024:**626**(8000): 897–904. https://doi.org/10.1038/s41586-023-07004-5.

Trudeau T, Nassar R, Cumberworth A, Wong ETC, Woollard G, Gsponer J. Structure and intrinsic disorder in protein autoinhibition. *Structure*. 2013:**21**(3):332–341. https://doi.org/10.1016/j.str.2012.12.013.

Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, Gibson TJ, Davey NE. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev*. 2014:**114**(13):6733–6778. https://doi.org/10.1021/cr400585q.

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**(D1):D439–D444. https://doi.org/10.1093/nar/gkab1061.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. The sequence of the

human genome. *Science*. 2001:**291**(5507):1304–1351. https://doi.org/10.1126/science.1058040.

Wang J, Han G-Z. Unearthing LTR retrotransposon gag genes co-opted in the deep evolution of eukaryotes. *Mol Biol Evol*. 2021:**38**(8): 3267–3278. https://doi.org/10.1093/molbev/msab101.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, *et al*. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018:**46**(W1):W296–W303. https://doi.org/10.1093/nar/gky427.

Wernersson R, Pedersen AG. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res*. 2003:**31**(13):3537–3539. https://doi.org/10.1093/nar/gkg609.

Whiteley AM, Prado MA, De Poot SAH, Paulo JA, Ashton M, Dominguez S, Weber M, Ngu H, Szpyt J, Jedrychowski MP, *et al*. Global proteomics of Ubqln2-based murine models of ALS. *J Biol Chem*. 2021:**296**: 100153. https://doi.org/10.1074/jbc.RA120.015960.

Wills NM, Moore B, Hammer A, Gesteland RF, Atkins JF. A functional −1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. *J Biol Chem*. 2006:**281**(11):7082–7088. https://doi.org/10.1074/jbc.M511629200.

Wong WSW, Yang Z, Goldman N, Nielsen R. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*. 2004:**168**(2):1041–1051. https://doi.org/10.1534/genetics.104.031153.

Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*. 2015:**16**(1):18–29. https://doi.org/10.1038/nrm3920.

Xu J, Erlendsson S, Singh M, Holling GA, Regier M, Ibiricu I, Einstein J, Hantak MP, Day GS, Piquet AL, *et al*. PNMA2 forms immunogenic non-enveloped virus-like capsids associated with paraneoplastic neurological syndrome. *Cell*. 2024:**187**(4):831–845.e19. https://doi.org/10.1016/j.cell.2024.01.009.

Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*. 1997:**13**(5):555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007:**24**(8):1586–1591. https://doi.org/10.1093/molbev/msm088.

Zarin T, Strome B, Nguyen Ba AN, Alberti S, Forman-Kay JD, Moses AM. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife*. 2019:**8**:e46883. https://doi.org/10.7554/eLife.46883.

Zarin T, Tsai CN, Nguyen Ba AN, Moses AM. Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc Natl Acad Sci USA*. 2017:**114**(8):E1450–E1459. https://doi.org/10.1073/pnas.1614787114.

Zhang X-L, Liu P, Yang Z-X, Zhao J-J, Gao L-L, Yuan B, Shi L-Y, Zhou C-X, Qiao H-F, Liu Y-H, *et al*. Pnma5 is essential to the progression of meiosis in mouse oocytes through a chain of phosphorylation. *Oncotarget*. 2017:**8**(57):96809–96825. https://doi.org/10.18632/oncotarget.18425.