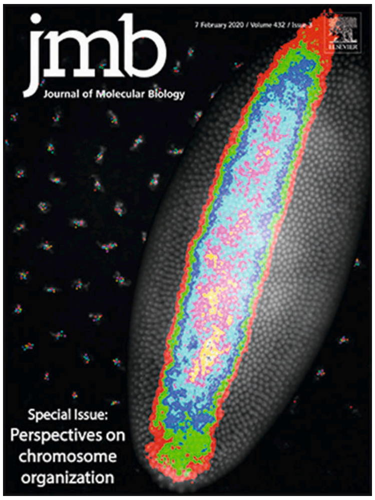


Research Article

Mapping the Constrained Coding Regions in the human genome to their corresponding proteins

Marcia A. Hasenahuer, Alba Sanchis-Juan, Roman A. Laskowski, James A. Baker, James D. Stephenson, Christine A. Orengo, F. Lucy Raymond, Janet M. Thornton

PII: S0022-2836(22)00512-5
DOI: https://doi.org/10.1016/j.jmb.2022.167892
Reference: YJMBI 167892

To appear in: *Journal of Molecular Biology*

Received Date: 10 September 2022
Revised Date: 8 November 2022
Accepted Date: 14 November 2022

Please cite this article as: M.A. Hasenahuer, A. Sanchis-Juan, R.A. Laskowski, J.A. Baker, J.D. Stephenson, C.A. Orengo, F. Lucy Raymond, J.M. Thornton, Mapping the Constrained Coding Regions in the human genome to their corresponding proteins, *Journal of Molecular Biology* (2022), doi: https://doi.org/10.1016/j.jmb.2022.167892

Title

Mapping the Constrained Coding Regions in the human genome to their corresponding proteins

Authors

Marcia A. Hasenahuer[1,2,5], Alba Sanchis-Juan[3,4,6], Roman A. Laskowski[1], James A. Baker[1], James D. Stephenson[1], Christine A. Orengo[5], F. Lucy Raymond[2,4], Janet M. Thornton[1]

Affiliations

1. European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

2. Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK

3. Department of Haematology, NHS Blood and Transplant Centre, University of Cambridge, Cambridge CB2 0XY, UK

4. NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK

5. Institute of Structural and Molecular Biology, University College London, London WC1E 6BT, UK

Present address

6. Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA; Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; Department of Neurology, Harvard Medical School, Boston, MA

Corresponding author

Marcia A. Hasenahuer

Emails: hasenahuer@ebi.ac.uk , m.hasenahuer@ucl.ac.uk , marcia.hasenahuer@gmail.com

Phone: +44 07305719847

Address: European Molecular Biology Laboratory – European Bioinformatics Institute
(EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.

Abstract

Constrained Coding Regions (CCRs) in the human genome have been derived from DNA sequencing data of large cohorts of healthy control populations, available in the Genome Aggregation Database (gnomAD) [1]. They identify regions depleted of protein-changing variants and thus identify segments of the genome that have been constrained during human evolution. By mapping these DNA-defined regions from genomic coordinates onto the corresponding protein positions and combining this information with protein annotations, we have explored the distribution of CCRs and compared their co-occurrence with different protein functional features, previously annotated at the amino acid level in public databases. As expected, our results reveal that functional amino acids involved in interactions with DNA/RNA, protein-protein contacts and catalytic sites are the protein features most likely to be highly constrained for variation in the control population. More surprisingly, we also found that linear motifs, linear interacting peptides (LIPs), disorder-order transitions upon binding with other protein partners and liquid-liquid phase separating (LLPS) regions are also strongly associated with high constraint for variability. We also compared intra-species constraints in the human CCRs with inter-species conservation and functional residues to explore how such CCRs may contribute to the analysis of protein variants. As has been previously observed, CCRs are only weakly correlated with conservation, suggesting that intraspecies constraints complement interspecies conservation and can provide more information to interpret variant effects.

Keywords

Population variability; disease-related variants; constrained coding regions; protein functional features; protein structure; protein intrinsic disorder; liquid-liquid phase separation

Introduction

Predicting the impact of variants on protein function has traditionally been based on combining information derived from protein sequences, inter-species conservation and knowledge of the structure and function of the protein. With the emergence of multiple human genome sequences from different populations, observed variation patterns in humans provide an orthogonal source of information for assessing the impact of variants. The comprehensive catalogues of genetic variations, compiled from many human population sequencing projects, have fuelled the development of different metrics that measure the general tolerance of genes to variation. Metrics like the probability of being loss-of-function intolerant (pLI) and missense Z-scores are extensively used to prioritise genes in genome interpretation of individuals [2]. However, it is well established that ‘all parts of a protein are not equal’. The modular presence of different domains and folds can endow different regions with different functions and structural constraints [3,4]. Hence, one would expect that regions which are extremely important for the function of a protein would be depleted of protein changing variants in healthy individuals.

In 2019 Havrilla et al. defined the Constrained Coding Regions (CCRs) as regions in the human coding genome where the following protein-changing variants were depleted:

missense, stop gained, stop lost, start lost, frameshift variant, initiator codon variant, rare amino acid variant, protein altering variant, inframe insertion, inframe deletion, and splice donor variant or splice acceptor variant when affecting the protein sequence. These regions were identified in whole exome and genome sequencing data from large cohorts of healthy control populations from the Genome Aggregation Database (gnomAD2.0.1), which is currently the largest and most widely used publicly available collection of data on population variation from harmonised sequencing data [2,5]. The premise of Havrilla et al. was that the data in gnomAD2.0.1 came from individuals who were either healthy or did not have early onset developmental abnormalities ('healthy control populations') and therefore their variant loci could be considered as 'tolerated'.

In the CCRs model, each of the variant-depleted (constrained) regions is weighted based on i) its length in base pairs, and ii) the fraction of individuals (above 50% of total individuals) having at least a 10x sequencing coverage at each bp of the region. A linear regression is then calculated comparing the weights and the CpG density of the regions, as an indicator of the region mutability upon spontaneous deamination of methylated cytosines. Regions with a greater weighted distance between protein-changing variants than expected based upon their CpG density (residual from the linear regression), are predicted to be under the greatest constraint. The residuals of the regression are ranked in CCRs percentiles (CCRpct) from 0 to 100, with 0 signifying unconstrained (i.e. having 'tolerated' variants in gnomAD) and 100 being the most highly constrained regions. Put simply, the longer a constrained region and the larger its CpG content, in general, the higher its CCRpct will be. Havrilla *et al.* observed that only ~1% of the highly constrained regions were found to be enriched with known pathogenic variants and associated with developmental disorders, and that 72% of the genes harbouring a CCR in the 99th percentile or higher were not linked yet to any disease, suggesting that CCRs could be used to reveal regions of protein coding genes that are likely to be under potentially purifying selection.

Given their relevance, it has been proposed that analysing the presence of regions like these can complement the classical procedures of phylogenetic conservation, amino acid substitution scores, and three-dimensional protein structural characterization and aid in the process of variant interpretation [1,6]. Although recent studies have used CCRpct as an extra score for assessing the pathogenicity of variants [7–13], there has been no large-scale attempt to map the distribution of these constrained genomic regions to amino acids and to analyse their co-occurrence with different protein functional features.

The human proteome is a continuum where proteins can be fully ordered, intrinsically disordered (ID) or flexible/mobile, have a mixture of folded and ID regions or even exert transitions between both states upon binding with other proteins [14,15]. These ID proteins and regions, can perform important and diverse functions in the cell from displaying sites for post-translational modifications (PTMs) to assembling molecular complexes that promote the phenomena of liquid-liquid phase separation (LLPS) and formation of membraneless organelles in the cell, amongst others [16–18]. Disease-causing mutations can occur in both ordered and ID regions [19,20], and recently the focus has turned towards variants predisposing to disease in LLPS regions, mostly related to autism spectrum disorders (ASD), cancer, neurodegeneration, and infectious diseases [21–23]. However, ID regions are usually not well conserved, lack a stable protein three-dimensional structure, sequence alignments have poor accuracy and most studies and tools focus on ordered regions, making it a challenge to interpret the molecular mechanism behind disease-related variants in these regions. Hence, observing CCRs in these regions may provide some insight into their constraints during human evolution.

Herein we map the CCRs onto their corresponding protein sequences and 3D structure, by assigning the CCRpct to each amino acid site (residue) spanned by each CCR. This process has the potential to highlight key functional amino acids in both ordered and disordered proteins, lying in regions of the protein which are strongly constrained. We explore the distribution of these regions across human proteins and compare their co-occurrence with different protein functional features annotated at the amino acid level. We then perform an enrichment test of Gene Ontology (GO) terms to explore which protein classes and cellular pathways are more frequently associated with genes harbouring regions with high CCRpct.

Results

Mapping the CCRs to amino acids

Our aim was to explore how CCRs are distributed across the human canonical protein sequences as defined by UniProtKB/Swiss-Prot [24]. For this purpose, first, we ran the CCRs model pipeline (available in the repository accompanying the work of Havrilla et al., 2019: https://quinlan-lab.github.io/ccr/examples/updates) to obtain the genomic coordinates of the CCRs, but using the gnomAD3.0 dataset of variants, which aggregates 76,156 whole genomes using coordinates from the human GRCh38 genome assembly. The resulting file with the genomic coordinates of the CCRs can be obtained from our GitHub repository (https://github.com/marciaah/CCRStoAAC/blob/main/data/rawCCRs/gnomad3_0/vep101/sor t_weightedresiduals-cpg-synonymous-novariant.txt.gz). Then, we mapped the genomic coordinates of the CCRs to amino acids in UniProtKB proteins, via the Ensmebl transcripts in the GENCODE basic set (see Figure 9 A for further details), the corresponding code of the pipeline is available in our repository https://github.com/marciaah/CCRStoAAC, and the output of the mapping to amino acids can be found here https://github.com/marciaah/CCRStoAAC-output.

The use of GRCh38 gives a more accurate cross-mapping of genes, transcripts and proteins in the Ensembl [25] and UniProtKB databases, while using the Swiss-Prot canonical set of proteins ensures the availability of functional annotations for further analysis. From a total of 18,583 human UniProtKB/Swiss-Prot canonical proteins that matched the Ensembl protein sequences, we were able to map CCRpct to at least one region in 17,366 of them (Figure 1 A). 6,608 of the 17,366 proteins had partial coverage of CCRpct for their amino acid sites (residues) and 1,217 (from the expected 18,583) completely lacked CCRpct, as a consequence of low quality conflicted genomic regions (see Methods) that prevented the identification of CCRs. In total, about 9.8 million amino acid sites (Figure 1 B) are represented with CCRpct, out of an expected 10.7 million from the total 18,583 sequences.

68.8% of the 9.8 million amino acids sites that we could map are in constrained regions, i.e. no protein changing variants are reported in gnomAD3.0. The remaining 3.06 million (31.2% of the mapped residues) contain at least one variant with a minimum allele count of one (Figure 1 B).

We categorised the CCRpct into different groups, as shown in Figure 1 B and C, based on the original considerations proposed by Havrilla et al, 2019, i.e., percentiles in the top 1% [99,100] for the most highly constrained regions down to 0 for unconstrained (i.e. sites having variants in gnomAD3.0). As expected, mapping from CCRs to amino acid sites gives approximately 10% of residues in each group (Figure 1 C). The exception is for the 0-10% group, which is underpopulated at the residue level, since these CCRs are the shortest regions, with an average length of only 1.05 amino acids without variants.

To summarise, we were able to assign CCRs to 93.5% of human UniProt/SwissProt canonical proteins, equating to 91.6% of the expected residues; about two thirds of residues are constrained (i.e. without protein changing variants in gnomAD 3.0); of these sites only 0.24% residues are assigned to the top [99,100] CCRpct bin (i.e. most highly constrained) and are exclusively from 839 regions in 751 proteins.

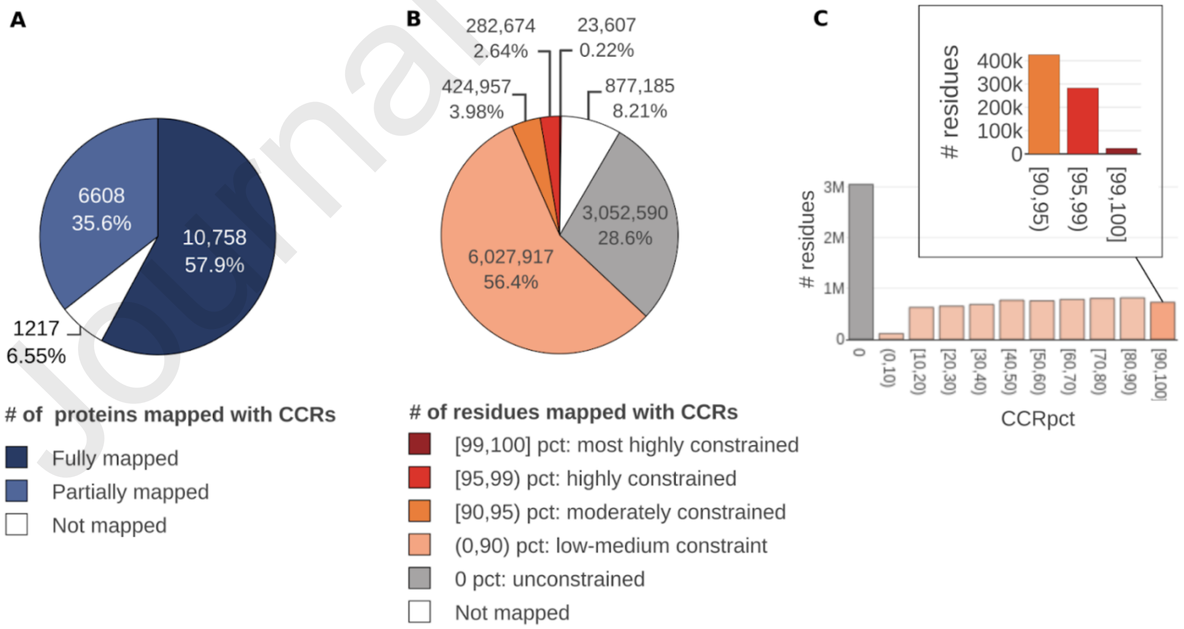


Figure 1: 68.8% of the 9.8 million mapped amino acid sites correspond to constrained regions, ranked by different constraint percentiles. The moderately, highly, and most highly constrained positions ([90,100] CCRpct bin) represent only 7.6% of these positions. Charts show: (A) the coverage of UniProtKB/SwissProt canonical proteins with the mapping of CCRpct, (B) the number of residues covered by the different percentiles, grouped into 5 categories, and the proportion of residues without CCRpct (not covered), and (C) the distribution of sites among the different percentiles grouped by tens, with a call-out showing the number of positions in the [90,100] CCRpct. Interval boundary numbers should be interpreted as follows: [ ]=included, ( )=excluded, or combinations of both.

Comparing CCRs percentiles with other whole gene scores (pLI and missense OEUF)

In clinical genomics and population genetics a number of metrics for assessing the overall intolerance to variability for a given gene or protein have become popular. The latest version of these scores is based on gnomAD2.1.1 [2,26]. One, the pLI score, represents the probability of a gene being intolerant to heterozygous putative loss of function (pLoF) variants: nonsense (stop-gained), frameshift, splice acceptor, and splice donor variants. A pLI ≥ 0.9 has been used to highlight “essential” transcripts/genes/proteins. For missense and synonymous variants, Z-scores are used, measuring how far from the mean a gene is in terms of observed/expected (o/e) missense or synonymous variants. Accompanying these metrics, the authors recommend the use of upper bound fraction of the 90% confidence intervals around the o/e ratios (OEUF) for the different types of variants. For further details, please refer to Supplementary Methods.

Here, we assigned pLI and OEUF scores to UniProtKB/Swiss-Prot canonical proteins with calculated CCRs, via their Ensembl transcript identifier. The mapping table, based on Ensembl (version 101) and UniProtKB (version 10-2020) can be downloaded from our repository (https://github.com/marciaah/CCRStoAAC/blob/main/data/mapping_tables/ensembl_uniprot_MANE_metrics_07102020.tsv.gz). For our analysis, we used the recommended thresholds:

pLI ≥ 0.9 and missense OEUF ≤ 0.35, as a simple way to define a gene/protein as highly constrained for pLoF and missense variants, respectively.

We observed 2,916 ‘essential’ proteins with pLI ≥ 0.9, and 75% of these have at least one region that scores with very high CCRpct in the range [95,100]. Also, only 113 proteins presented missense OEUF ≤ 0.35, and 93% of them have CCRpct in the [95,100] group. However, when we looked at proteins with pLI < 0.9 or OEUF > 0.35, we found that about a quarter or more of these more “variant-tolerant” proteins also include highly constrained regions, which are distributed across the proteins with all values of missense OEUF and pLI (Supplementary Figure 1). These observations highlight the importance of looking at local constraint scores, using CCRpct, in order to understand more deeply the impact of variants in protein coding genes.

The correlation between CCRs percentiles, interspecies conservation and length of the regions

We next explored how the CCRpct (based on intra-human variation) correlates with interspecies amino acid conservation for each amino acid position in the human proteome. The average interspecies conservation increases with increasing CCRs percentile (Supplementary Figure 2 A), however, there is a surprisingly large variability within each percentile category (Supplementary Figure 2 A and B) and the overall correlation is very low (Pearson=0.11).

In a similar way, for all amino acid positions we compared the length of the CCRs (in number of amino acids) against their percentiles (Figure 2 C and D). The average length of regions increases with CCRs percentile, which is expected given that CCRs are prioritised by region length. Nevertheless, there is a high variability within each CCRs percentile category. The most highly constrained regions (percentiles [99,100]) are only present in proteins of at least 100 amino acids in length (Supplementary figure 3).

To explore the numbers of amino acid sites having different combinations of CCRpct and conservation scores, we stratified both measures into ten groups and built a 2D matrix counting the numbers of residues in each cell. The majority of both constrained and unconstrained positions have conservation scores >0.4 and distribute evenly in a plateau up to a conservation of 0.95 (Figure 2). Above this level, the counts increase, in particular for the two extremes of more constrained (percentiles [90,100]) and unconstrained sites (percentile 0). The 3D surface shown in Figure 2 highlights the disparity between these CCRpct and conservation scores and the high frequency (importance) of residues which are completely conserved (ScoreCons=1)[28]. CCRpct are able to differentiate between such residues, according to observed variation and length of conserved regions, providing a valuable score for analysis.

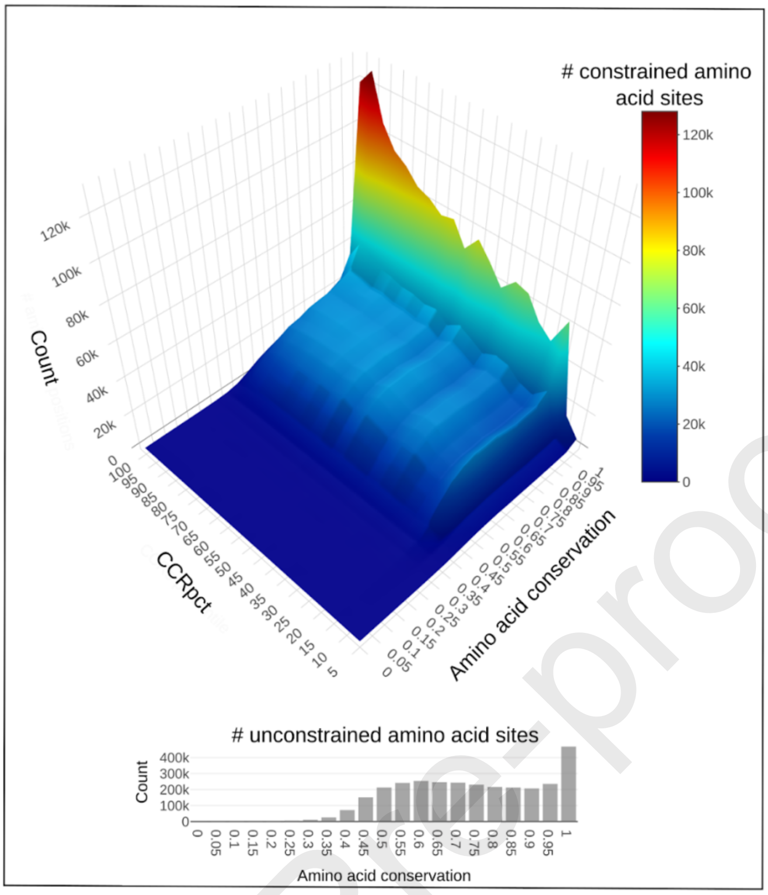


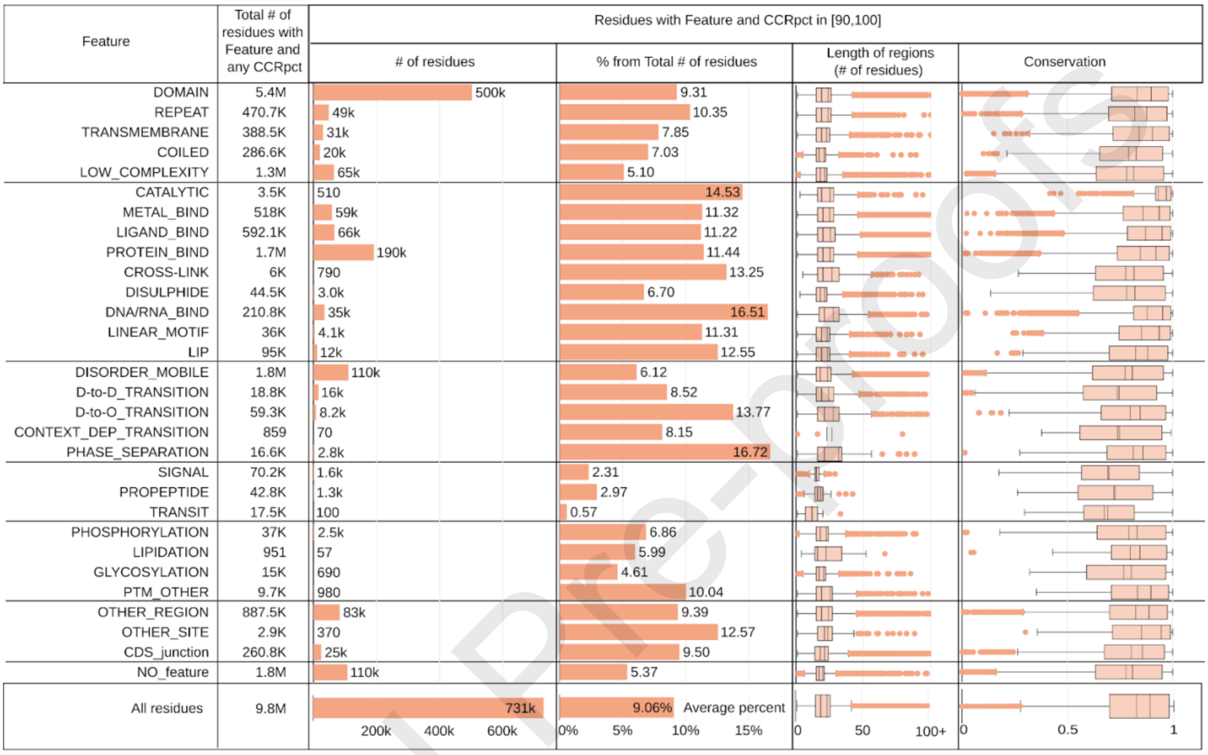
Figure 2: Distribution of counts of amino acid sites in different bins of CCRpct and conservation scores. The 3D heat map shows counts for constrained amino acid sites, while the histogram shows the unconstrained ones.

Protein features and CCRs percentiles

Protein annotations from UniProtKB/Swiss-Prot [24], PDBe [24,29], VarSite [30], M-CSA [31], BioLip [32], MobiDB [33], ELM [34], Ensembl [35] and ClinVar [36] databases were obtained and aggregated for each amino acid site in the human canonical and annotated proteins of UniProtKB. 9.8 million sites were annotated in this way, assigning CCRpct and the 30 protein annotations listed in Figure 9 (Methods).

Figure 3 A and B present a broad overview of the distribution of total sites annotated with different protein features, with the corresponding distributions of conservation scores and

length of regions for such sites. For simplicity, we only present this information for the two extremes of CCRpct: the more constrained sites in percentiles [90,100] and the variable or unconstrained sites with percentile 0.



Feature	Total # of residues with Feature and any CCRpct	Residues with Feature and CCRpct in [90,100]			
		# of residues	% from Total # of residues	Length of regions (# of residues)	Conservation
DOMAIN	5.4M	500k	9.31		
REPEAT	470.7K	49k	10.35		
TRANSMEMBRANE	388.5K	31k	7.85		
COILED	286.6K	20k	7.03		
LOW_COMPLEXITY	1.3M	65k	5.10		
CATALYTIC	3.5K	510	14.53		
METAL_BIND	518K	59k	11.32		
LIGAND_BIND	592.1K	66k	11.22		
PROTEIN_BIND	1.7M	190k	11.44		
CROSS-LINK	6K	790	13.25		
DISULPHIDE	44.5K	3.0k	6.70		
DNA/RNA_BIND	210.8K	35k	16.51		
LINEAR_MOTIF	36K	4.1k	11.31		
LIP	95K	12k	12.55		
DISORDER_MOBILE	1.8M	110k	6.12		
D-to-D_TRANSITION	18.8K	16k	8.52		
D-to-O_TRANSITION	59.3K	8.2k	13.77		
CONTEXT_DEP_TRANSITION	859	70	8.15		
PHASE_SEPARATION	16.6K	2.8k	16.72		
SIGNAL	70.2K	1.6k	2.31		
PROPEPTIDE	42.8K	1.3k	2.97		
TRANSIT	17.5K	100	0.57		
PHOSPHORYLATION	37K	2.5k	6.86		
LIPIDATION	951	57	5.99		
GLYCOSYLATION	15K	690	4.61		
PTM_OTHER	9.7K	980	10.04		
OTHER_REGION	887.5K	83k	9.39		
OTHER_SITE	2.9K	370	12.57		
CDS_junction	260.8K	25k	9.50		
NO_feature	1.8M	110k	5.37		
All residues	9.8M	731k	9.06% Average percent		

A

B

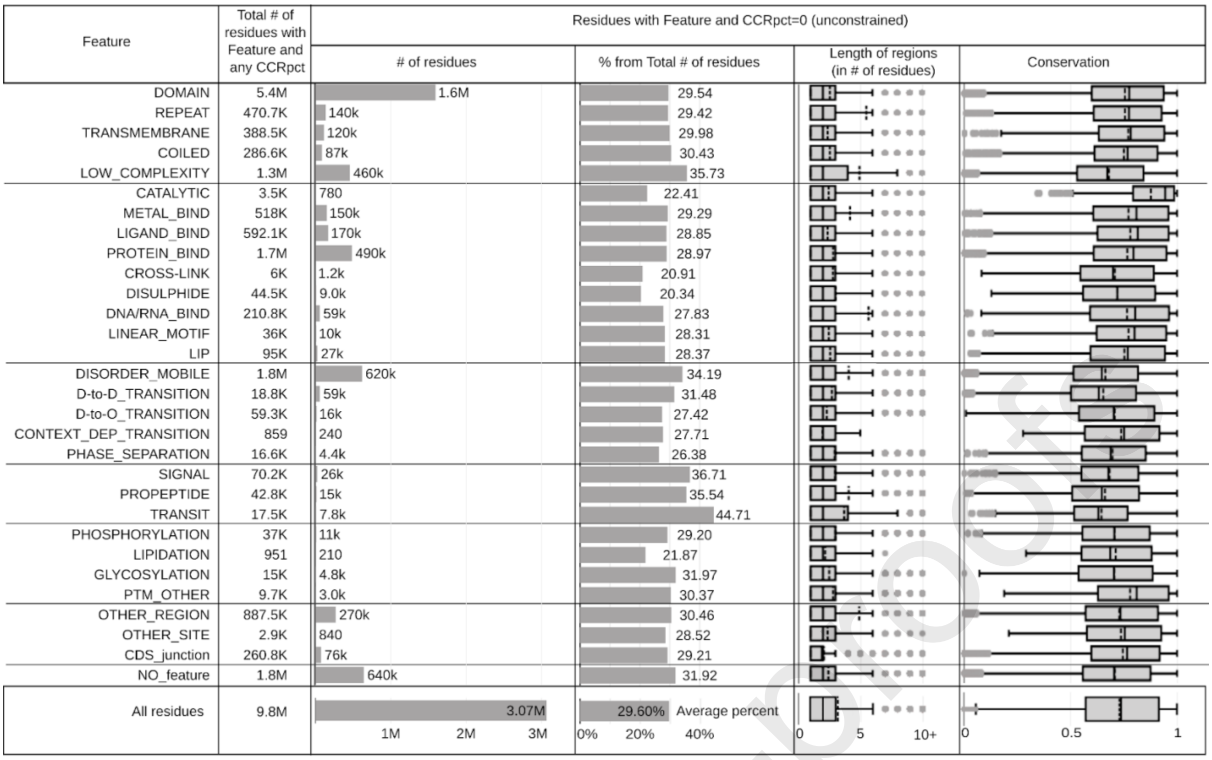


Figure 3: Distribution of amino acid sites corresponding to A) highly constrained regions with percentiles in the interval [90,100] and B) unconstrained regions harbouring tolerated variants in gnomAD3.0, and in coincidence with the different protein features as listed in the first column. “All sites” panels at the bottom correspond to counting all the sites without distinction of protein features. Average percent= average of all the % from Total # of residues.

In order to investigate in detail how different protein features are constrained across human populations, we calculated odds ratios (OR) to measure the enrichment of residues in CCRpct categorised in 7 groups for each of the 30 protein feature annotations, compared to a random distribution (see in Methods, *Odds ratios tests for enrichment: I. CCRpct and presence of each one of the 30 protein features*). Figure 4 presents forest plots for comparing the resulting OR (listed in Supplementary Table 1).

Additionally, we performed OR tests for assessing the overall enrichment of sites with the different protein features and their co-occurrence with inter-species conservation scores and CCRpct. We did this by defining 6 different groups, as described in Methods, *Odds ratios tests for enrichment: II. CCRpct and conservation with the presence of protein features*. Put simply, we divided the cells of the heatmap in Figure 2 into 4 quadrants and the histogram for unconstrained sites into 2 halves, counted the numbers of residues and calculated the ORs. Supplementary Table 2 shows the resulting OR and Table 1 summarises the features enriched in each group combining CCRpc and conservation score.

	Lower conservation (ScoreCons ≤0.5)	Higher conservation (ScoreCons >0.5)
Higher CCRpct (50,100]	**D_to_D (1,90***)** **PROPEP (1,50***)** DISORDER_MOBILE (1,46***) CONTEXT_DEP (1,45*) D_to_O (1,44***) PHOSPHO (1,25***) NO_feature (1,15***) LLPS (1,13***)	**CATALYTIC (1,90***)** **CROSSLINK (1,63***)** **LLPS (1,50***)** DNA/RNA (1,46***) DOMAIN (1,42***) DISULPHIDE (1,35***) MOTIF (1,34***) LIP (1,33***) PROTEIN (1,32***) LIGAND (1,31***) D_to_O (1,30***) METAL (1,28***) REPEAT (1,25***) CONTEXT_DEP (1,25**) SITE (1,22***) TRANSMEM (1,20***) LIPID (1,19**) CDSjunction (1,16***) PTM_OTHER (1,12***) REGION (1,11***) COILED (1,10***)
Lower CCRpct (0,50]	**TRANSIT (1,83***)** **PROPEP (1,79***)** **D_to_D (1,71***)** **DISORDER_MOBILE (1,62***)** LOW_COMPLEXITY (1,39***) PHOSPHO (1,22***) NO_feature (1,17***) SIGNAL (1,11***)	**DISULPHIDE (1,74***)** **LIPID (1,56***)** SIGNAL (1,27***) TRANSIT (1,16***) CARBOHYD (1,15***) PTM_OTHER (1,14***) CDSjunction (1,10***)

Unconstrained CCRpct [0]	**TRANSIT (1,86***)** **PROPEP (1,81***)** **D_to_D (1,69***)** **DISORDER_MOBILE (1,67***)** **LOW_COMPLEXITY (1,61***)** SIGNAL (1,23***) NO_feature (1,10***)	**TRANSIT (1,58***)** SIGNAL (1,24***) LOW_COMPLEXITY (1,14***) CARBOHYD (1,11***)

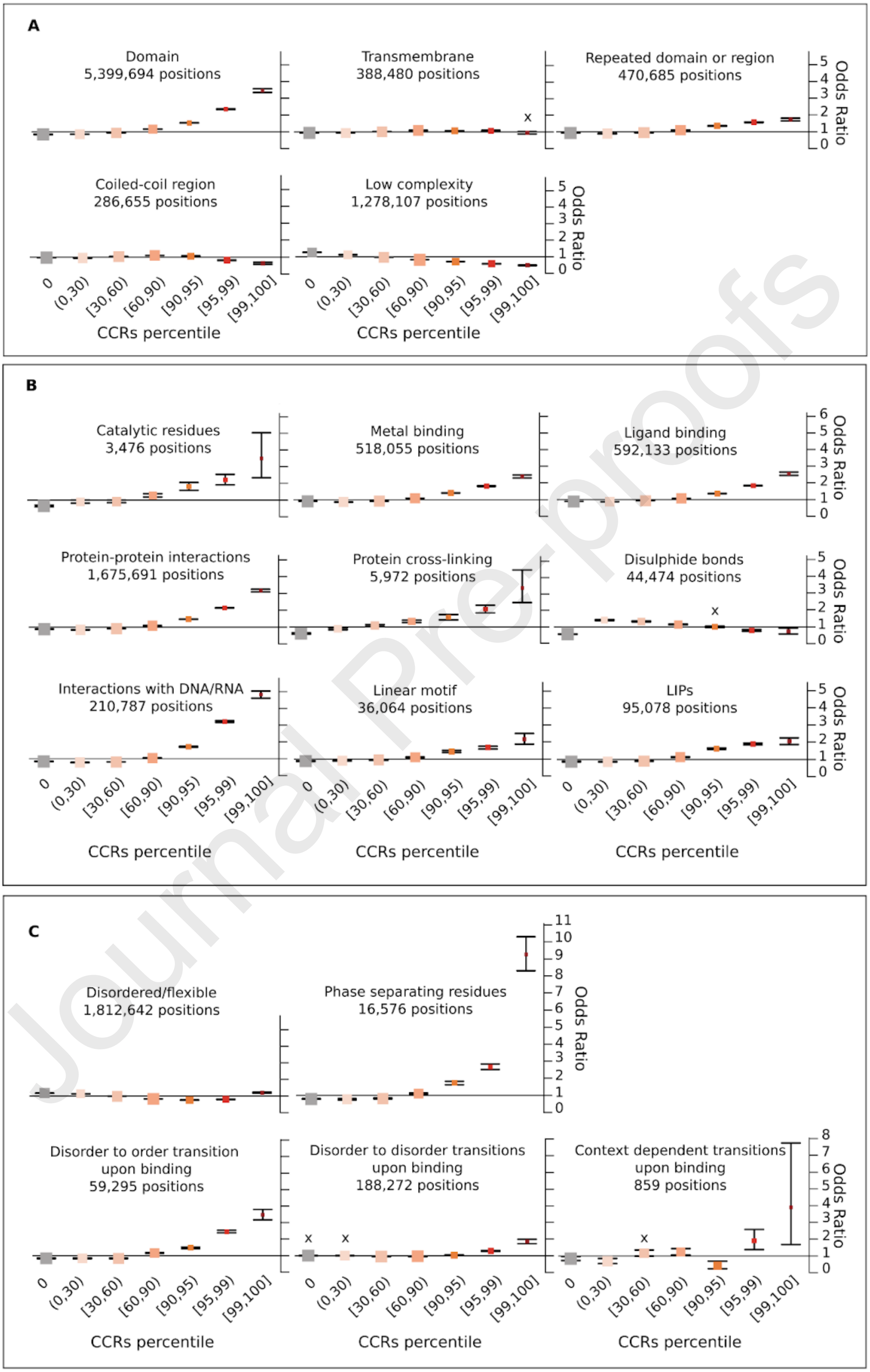
Table 1: List of features that are more enriched in the different combinations of CCRpct (row-wise) and conservation (column-wise), sorted by OR in each cell (only showing OR≥1.1). OR and p-values are in parenthesis, with stars depicting significant Fisher p-values: ‘*’ ≤ 0.05 , ‘**’ ≤ 0.01, ‘***’ ≤0.001. The full list of features and values is in Supplementary Table 2. In bold we highlight the most enriched features, with OR≥1.5.

For capturing the strongest associations, we set a threshold of OR≥1.5 for our analysis and grouped the different features for discussing their enrichments with CCRpct and with CCRpct and conservation, as is summarised below:

I. ***Domains and compositionally-biassed protein regions.*** This group includes large features (e.g. structural domains) and biassed sequences (e.g. coiled-coils) (Figure 4 A). The most striking ORs distribution occur for domain regions (i.e. those regions of a protein classified as being in a domain, according to Pfam or CATH) compared to those lying outside such a domain. There is a clear indication that amino acid sites within domains are more constrained than those outside. The repeated domains show a similar tendency. Surprisingly, the transmembrane regions showed little if any enrichment for highly constrained regions, perhaps reflecting the lipid environment where variants between the hydrophobic amino acids are common. The residues in low complexity and coiled-coil regions are preferentially unconstrained and rarely show the highest levels of constraint.

II. Interactions and catalytic residues. Residues involved in catalytic sites, binding to metals and/or ligands, protein-protein interactions, protein-protein cross-linking, interactions with DNA/RNA, linear motifs, and linear Interacting peptides (LIPs) are all more likely associated with medium to high percentiles of constraint (in the range [60,100]) (Figure 4 B).

Most of these residues are also less likely to be associated with unconstrained regions. Catalytic sites, in particular, presented the highest odds of having high CCRpct and high conservation (OR=1.9, in Table 1), and the average residue conservation is consistently high in combination with all the CCRpct, including the unconstrained sites where gnomAD tolerated variants are located (Supplementary Figure 5). Disulphide bonds are an interesting exception - showing no preference to lie in a highly constrained region, but also they are rarely unconstrained. In particular, catalytic sites, disulphide bonds and cross-linking covalent linkages have ORs that suggest they are of the order of 0.5-0.6 times as likely to have tolerated variants, while for sites involved in other interactions the ORs are between 0.7 and 0.91.


A
Domain
5,399,694 positions
Transmembrane
388,480 positions
x
Repeated domain or region
470,685 positions
Odds Ratio
0
1
2
3
4
5
Coiled-coil region
286,655 positions
Low complexity
1,278,107 positions
0
(0,30)
[30,60)
[60,90)
[90,95)
[95,99)
[99,100]
CCRs percentile
B
Catalytic residues
3,476 positions
Metal binding
518,055 positions
Ligand binding
592,133 positions
6
Protein-protein interactions
1,675,691 positions
Protein cross-linking
5,972 positions
Disulphide bonds
44,474 positions
Interactions with DNA/RNA
210,787 positions
Linear motif
36,064 positions
LIPs
95,078 positions
C
Disordered/flexible
1,812,642 positions
Phase separating residues
16,576 positions
7
8
9
10
11
Disorder to order transition
upon binding
59,295 positions
Disorder to disorder transitions
upon binding
188,272 positions
Context dependent transitions
upon binding
859 positions

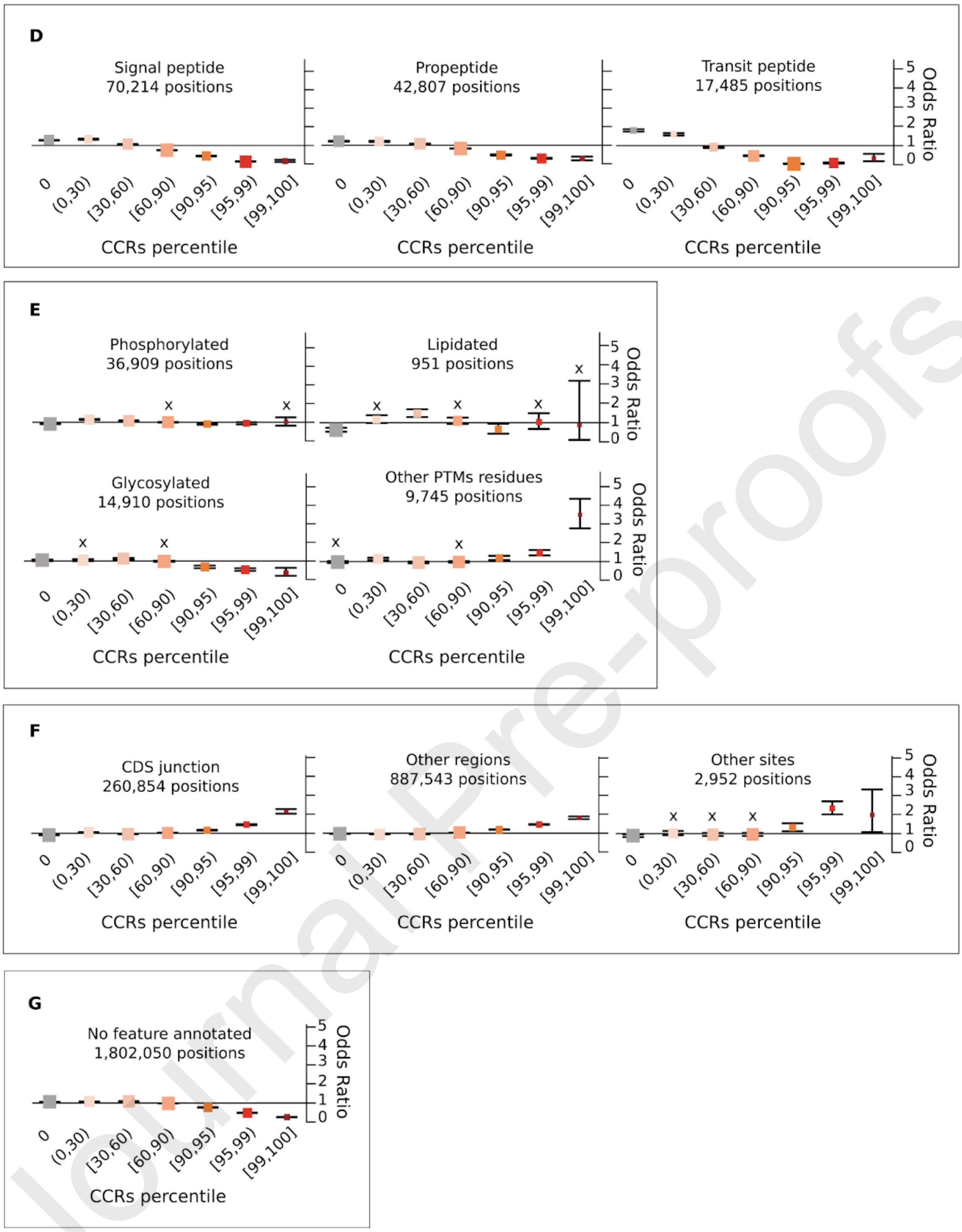


Figure 4: Propensity of co-occurrence of amino acid positions in specific protein features or functional sites with the different CCRpct. Odds ratios (OR) forest plots with 95% confidence intervals (CI) based on two-tailed Fisher's exact test for amino acid sites in (A) domains, globular, non-globular and compositionally-biased protein regions, (B) interactions and catalysis, (C) disordered/mobile residues, structural transitions between disorder/flexibility and order upon binding and regions driving LLPS, (D) different signalling regions, (E) post translationally modified sites, (F) coding DNA sequence (CDS) junctions and other functionally relevant regions annotated in UniProtKB, and (G) residues lacking any of the features or functional annotation considered in this

work. The vertical lines through the boxes illustrate the length of the CI. The line at OR=1 is the line of no clear difference, boxes and intervals above this represent co-occurrences more likely to happen, while boxes and intervals under the line represent the contrary. A cross (X) above a box depicts an OR where the association is statistically not significant (i.e. p-value > 0.05 or the 95% CI crosses over OR=1). See Supplementary Table 1 for all the p-values corresponding to the two and one tailed Fisher's exact tests.

III. Disorder related features. The 1.8M amino acids that were annotated in our dataset as intrinsically disordered (ID) or mobile and with CCRpct assigned (about 18.5% from the total of 9.8M residues) showed two contrasting tendencies. Although with weaker OR, these sites were more likely to coincide with unconstrained and very lowly constrained sites (percentiles [0,30)) and also with the top most highly constrained percentiles [99,100] (Figure 4 C). ID regions/proteins tend to evolve faster than structured proteins at the sequence level [37,38], this is reflected in the enrichment of lower mean interspecies amino acid conservation across all levels of constraint, even for the high CCRpct residues (Table 1). High percentiles of constraint ([95,100]) were strongly associated with residues in disorder to order (D-to-O), disorder to disorder (D-to-D) and context dependent transitions upon binding with other protein partners, with higher OR values for those undergoing D-to-O and context dependent transitions.

Residues in regions driving LLPS present the highest association we observed, with OR 9.26 times more likely to be in the most highly constrained regions (percentiles [99,100]) and with the longest average length of 75 amino acids (Figure 3). Furthermore, 29 (53%) of the 54 LLPS proteins in our dataset have LLPS driving amino acid sites with percentiles in [95,100] (Supplementary Table 5).

When considering gnomAD2.1.1 per gene variant intolerance metrics, 34 out of these 54 (63%) LLPS driving proteins have pLI >= 0.9 (i.e. are 'essential' genes, extremely intolerant to pLoF variants in heterozygosity), while only 7 of the 54 are highly intolerant to missense changes (missense OEUF<=0.35) (Supplementary Table 5).

IV. Signalling regions. Residues in propeptides, signal peptides and transit signalling regions tend to be in regions more likely to have unconstrained to medium constrained percentiles in the interval [0,60] (Figure 4 D). The very few highly constrained sites in pro-peptides and signal peptides show much lower conservation and shorter region length (Figure 3 and Table 1).

V. Post translationally modified positions (PTMs). Although with weak OR, the overall tendency is for sites that are post translationally modified to be more likely in constrained regions at lower percentiles (Figure 4 E). Given that these are specific sites with, in general, very short flanking motifs, it is expected that shorter regions, and therefore lower CCRpct, are associated with these positions. Also, some PTMs may not be functionally relevant and represent false positives [39].

Glycosylated sites tend to be more associated with being unconstrained or at low constraint (percentiles [0,60]). The number of lipidated sites is very low, therefore statistically not significant for most of the CCRs percentile categories. However, they are less likely to be unconstrained (OR=0.62, 95% CI: 0.53-0.72).

Other types of PTMs are more likely to be associated with highly constrained regions (percentiles [95,99]). There are 482 highly constrained sites, 302 (62%) correspond to N-acetylations, and 79% are N-acetyl lysine. This is not surprising given the abundance of the latter modification. When comparing conservation and constraint for these sites, the behaviour is mixed and ORs are overall weak. Phosphorylation sites tend to be less conserved in general for all constraint levels (Table 1), and a slightly higher prevalence is for sites with high constraint and low conservation. Lipidated sites have a stronger association with being more conserved and with low percentiles. Glycosylation and other PTMs are more associated with high conservation for all constraints.

VI. Other relevant sites and regions. The localisation of "other sites" and "other regions" was obtained from UniProt. This category corresponds to regions/positions of functional relevance for proteins, identified mostly from experimental evidence, that cannot be described by other feature annotations of UniProt. We also recorded coding DNA sequence (CDS) junctions, by translating the genomic coordinates of these sites obtained from Ensembl onto the corresponding amino acids in the UniProt sequences. As suspected, given their relevance, the protein sites in these three categories are more likely to be constrained at high percentiles (Figure 4 F) and also mostly related to high percentiles and high conservation, although with a weak OR (Table 1).

VI. Residues without annotations. 1.8M amino acids did not have any of the functional sites or regions or domains that we aggregated in the present study. These were very weakly associated with unconstrained and low constraint regions and particularly less associated with higher percentiles (Figure 4 G). They were slightly associated with all combinations of constraints and low conservation (Table 1), and slightly more with low CCRpct and low conservation. The 96.4K and 658.7K residues that are in CCRpct [50,100] with low and high conservation, could be explained by functional features that still remain to be discovered and annotated for some proteins, or some domains that were difficult to delimit, creating fuzziness at their boundaries.

In summary, when bringing together the classifications regarding CCRpct and protein functional features, for the 9.8M positions that can be assigned to a CCRs percentile, it is possible to observe how the co-occurrence with certain functional features becomes more evident at higher percentiles (Figure 4 A vs B and Figure 3). The 23.6K most highly constrained residues in the human genome (CCRpct in [99,100], 0.24% of the total mapped residues) correspond to, on average, the longest linear stretches depleted of tolerated variability, and strongly highlight positions involved in DNA/RNA binding, protein binding, catalytic sites and in driving LLPS.

Amino acid sites with clinically interpreted variants, their CCR percentiles and co-occurrence with protein features

We next investigated how residues with different types of clinically interpreted variants correlate with the different percentiles of constraint. For this purpose, we employed variants from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar), and classified the amino acid positions in our dataset as pathogenic (including pathogenic/likely_pathogenic), benign (including benign/likely_benign) and/or VUS/conflicting (including variants of uncertain significance or with conflicting interpretations of pathogenicity), and followed a similar methodology for performing OR tests (Supplementary Figures 5 and 7) Pathogenic missense variants account for only 28,398 protein sites, with the majority of missense variants in ClinVar corresponding to VUS/conflicting interpretations, affecting 194,508 residues. Residue sites with pathogenic missense variants were observed as strongly associated with higher percentiles of constraint, in the intervals between [90,100]. Surprisingly, these types of variants were also associated with unconstrained regions (percentile [0,0]), although with weaker odds. Sites with benign and VUS/conflicting missense variants had higher odds of being in unconstrained regions, although a few benign variants, affecting 476 amino acid sites, were also present in the top most highly constrained regions.

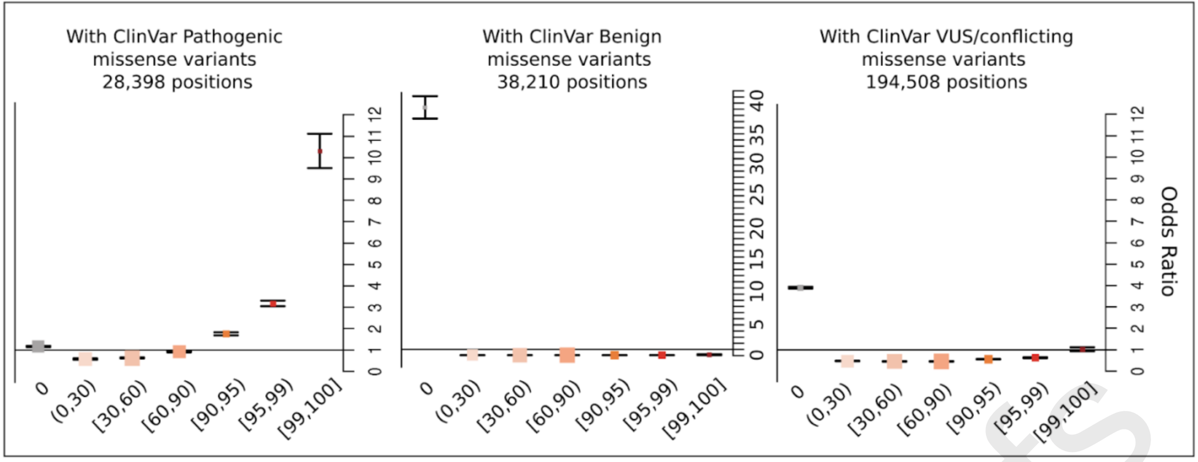


Figure **5: Propensity of co-occurrence of amino acid positions having clinically interpreted missense variants with the different CCRpct.** The ORs with 95% CI based on two-tailed Fisher's exact test, represent the odds that amino acid sites with a particular type of variant will co-occur in combination with one of the CCRs percentile categories, compared to the odds of having such a type of variant but with any of the other percentile categories. The vertical lines through the boxes give the length of the CI. The line at OR=1 is the line of no clear difference; boxes and intervals above this represent more likely co-occurrences, while boxes and intervals under the line represent the converse. See Supplementary Table 3 for all the p-values corresponding to the two and one tailed Fisher's exact tests.

When comparing the distribution of residues with the three groups of variants with CCRpct and conservation scores, we observed that they are spread across all categories of conservation and CCRpct (Supplementary Table 4 and Supplementary Figure 4), but only the amino acid sites affected by pathogenic/likely_pathogenic were OR=1.63 times more likely to be more conserved and with high CCRpct.

Additionally, we assessed the co-occurrence of ClinVar variants with the 30 protein feature annotations (Supplementary Figures 7 and 8). Sites with missense pathogenic variants are mostly associated with being in domains, transmembrane regions, catalytic sites, metal binding, ligand binding, protein binding, disulphide bonds, DNA/RNA binding, linear motifs, LIPs, D-to-O, lipidation, other regions, and CDS junctions. Sites with missense benign variants were more likely to be disordered/mobile, low complexity, LLPS, signal, propeptide and transit peptides, and sites without any protein features. Sites with missense VUS/conflicting variants, were mostly associated with those regions generally more difficult to characterise: repeats, coiled-coils, LIPS, D-to-O and D-to-D transitions, phosphorylation sites, other regions of biological relevance annotated in UniProt, and sites without any feature.

Over-representation of regions with high percentiles in GO protein classes and Reactome pathways

We investigated whether there was an enrichment of specific types of proteins and biological pathways in the proteins having regions with high CCRpct. For this purpose, we submitted a ‘query list’ of 6,402 protein identifiers of sequences with percentiles in the interval [95,100] to the PANTHER Classification System [40] to perform a Gene Ontology (GO) Over-representation Test for two categories of terms: ‘protein class’ and ‘Reactome pathways’. We used as the ‘reference list’ the 17,366 genes/proteins for which we have CCRs estimates; however, for only 17,022 PANTHER was able to assign GO terms.

Genes with [95,100] CCRpct were enriched in 14 protein classes out of a total of 196 that were assigned to our lists of proteins. Genes with [95,100] CCRpct were enriched in 70 Reactome pathways, out of a total of 2.482. Figure 8 A and B list the statistically relevant over- and under-represented protein classes and Reactome Pathways, respectively.

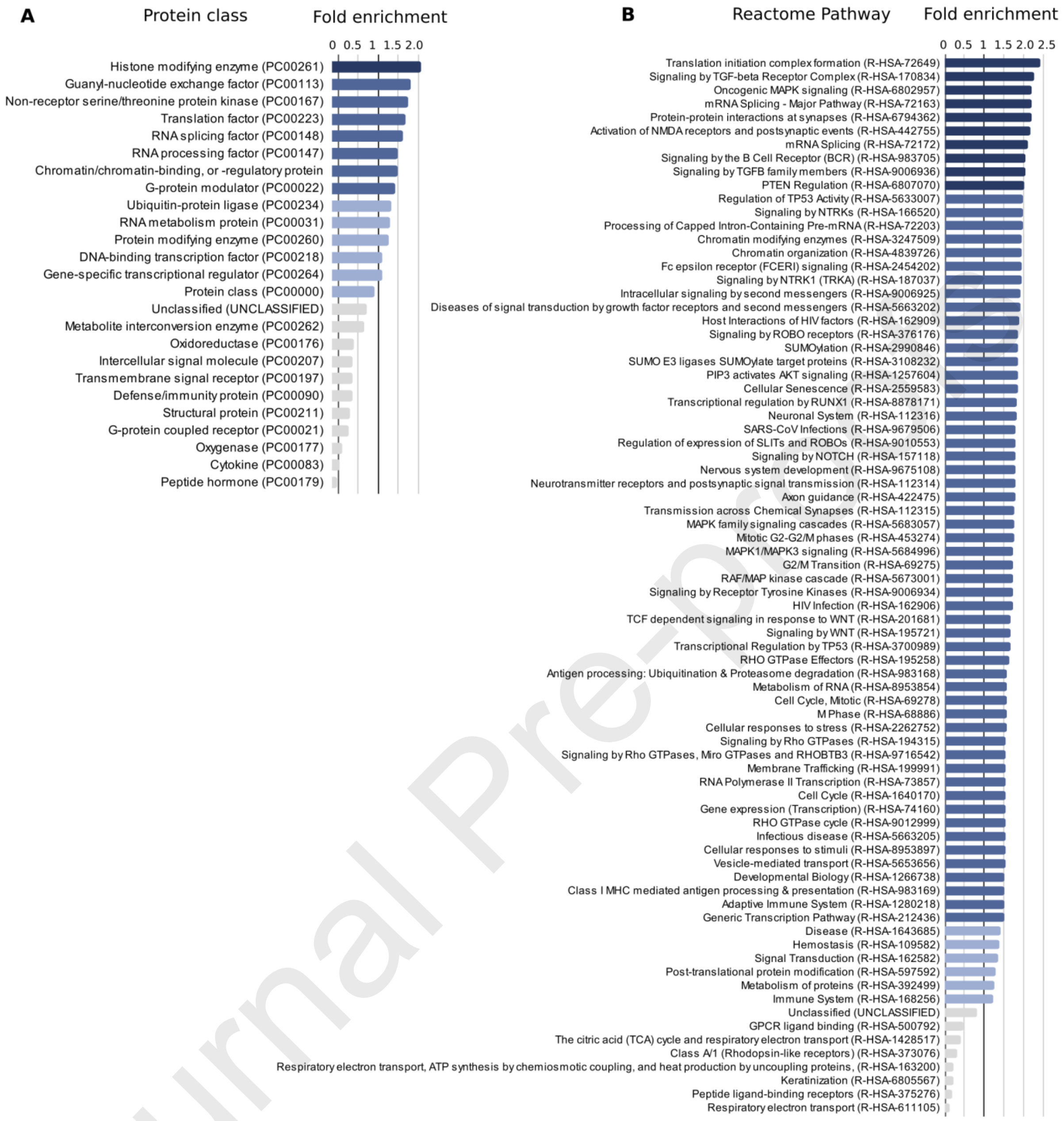


Figure 6: Gene Ontology enrichment test for A) protein class, and B) pathways annotated in Reactome, for proteins harbouring residues in highly constrained regions with percentiles in [95,100]. The over-representation tests are based on multiple Fisher tests with Bonferroni correction, and only significant (p-values<0.05) terms are listed and ordered by fold of enrichment. Bars in different shades of blue correspond to over represented terms (>1 fold of enrichment). Darker blues highlight ⩾ 1.5 and ⩾2.0 folds of enrichment.

Liquid-Liquid phase separation: biological processes, variability constraints and related diseases

LLPS driving regions presented a low number of sites with Pathogenic variants, while being highly associated with high CCRpct (see Figures 3 A and 4 E and Supplementary Figure 7). This motivated us to further investigate the distribution of Pathogenic variants in the corresponding proteins, the types of diseases they associate with and the biological processes where such proteins are involved. The Supplementary Table 5 presents this information, and Supplementary Table 6 summarises the list of clinical conditions and number of LLPS driving genes/proteins associated with them.

We observed the majority of LLPS driving genes (35 out of the 54, 63%) act in key biological processes that facilitate DNA damage repair, epigenetic gene repression and RNA metabolism (transcription, splicing, polyadenylation, transport and translation, see column ‘Main biological process groups’ and the corresponding genes in Supplementary Table 5). The remaining 19 genes are involved in many different processes, including neuron cell growth, adhesion, axonogenesis and synaptogenesis, development, synaptic plasticity and regulation of neurotransmitter vesicles release; signal transduction pathways for cell survival, migration, proliferation, differentiation and apoptosis; protein degradation/recycling; cell cycle regulation; immune responses; nuclear transport; elasticity of organs and tissues; muscle structure and function and glomerular filtration in kidney.

34 (64%) of the 54 proteins driving LLPS had pLI>=0.9 (i.e. essential genes highly intolerant to loss of function in heterozygosity), in particular this is the case for proteins involved in RNA metabolism. 43 of the 54 LLPS proteins (79.6%) have amino acid positions which drive phase separation and are highly constrained for variation (CCRpct in [90,100]).

30 LLPS proteins (56%) had variants related to at least one disease: seventeen (31.5%) of the 54 genes were associated with severe early onset developmental disorders, including different organ malformations, oestrogen resistance with absence of sexual maturation or severe early onset immunodeficiency, with 10 of them in particular linked to neurodevelopmental disorders (Supplementary Table 5, column ‘Disease group’). 10 genes

(18.5% of the 54) were associated with later onset diseases, mostly neurodegenerative but also affecting muscles and bone and triggering earlier menopause. 5 genes (9%) were associated with different cancers of pancreas, breast, uterus and prostate, lung and leukaemia. There was also a high incidence of associations with conditions not yet described (see Supplementary Table 6, ‘not provided’ or ‘not specified’).

The remaining 24 LLPS driving proteins (44% out of the total 54) have not yet been associated to any disease by protein changing variants by the time we consulted ClinVar. Fourteen of these 24 (58%) have pLI⩾0.9 (i.e. extremely intolerant to LoF) and are associated to RNA metabolism (10 genes), DNA damage repair (1 gene), immune response (1 gene) and signal transduction for cell proliferation and differentiation (1 gene), all of them presenting multiple regions of high constraint (CCRpct [90,100]).

In summary, the LLPS driving regions are clearly biologically important, often related to disease and very constrained for variability in the human genome.

Exploring some examples: U2AF2, the splicing factor U2AF 65 kDa subunit, and SLC12A2, the solute carrier family 12 member 2

Aggregating different protein annotations such as inter-species conservation, human variability and constraint, functional features, 3D structure and presence of clinically interpreted variants can help understand why variants have different propensities in different contexts. We have chosen two examples of proteins for illustration, the first of which illustrates a protein with highly disordered/mobile regions which have nevertheless been constrained during human evolution, although the conservation across species is patchy. The second example is a transmembrane protein in which the functional ion channel residues are highlighted as highly constrained by the CCRs and also a disordered region of

20 amino acids which are variable across species but depleted of tolerated variants and include a cluster of pathogenic variants.

The first example is *U2AF2*, the 65 kDa subunit of the U2 auxiliary splicing factor U2AF (Figure 7), a protein that is highly constrained against variability. This is an essential splicing factor that recognizes the polypyrimidine-tract (Py) 3' splice-site signal in pre-mRNA and initiates spliceosome assembly in the nucleus [41]. *U2AF2* is highly constrained for missense and LoF variability in gnomAD2.1 (missense OEUF=0.31, missense Z-score=4.21, pLI=1, LOEUF=0.133), suggesting essentiality for humans. There is partial structural data for this protein, derived from 3 separate PDB files, each containing a different part of the protein. The protein contains a low complexity arginine-serine rich motif (RS) at positions 27–62, which has been proposed to initiate liquid-liquid phase separation (LLPS) to form nuclear speckle drops in the nucleus, bringing together pre-mRNAs and the proteins of the spliceosome [42]. Further along the sequence, the region 85-112 is a UHM ligand motif (ULM) that has been shown responsible for the interactions with U2AF1 (also known as U2AF35) the 35kDa subunit of the splicing factor *U2AF* [43]. The two central RNA recognition motifs (RRM) are shown in the Pfam domains panel and central dashed box of Figure 7, with protein structures on top. These regions bind to the Py-tract signal in the pre-RNA, are highly mobile (light grey shading regions in the Pfam domains panel) [44,45], and are connected by a flexible/disordered linker region (231–258) that modulates the binding specificity for the proper Py-tracts in pre-mRNA [46]. The third RRM domain, known as U2AF Homology Motif, UHM, is atypical and has lost its RNA-binding ability, but interacts with splicing factor 1 (SF1) (right dashed box with protein structures on top) [46,47].

The protein has long, moderate to highly constrained regions (CCRs panels) that co-localize with the U2AF2-U2AF1 binding interface, the three RRMs, the flexible linker and also with regions involved in LLPS and/or linear interacting peptides (pink and violet rectangles in the "Other protein features" panel). A few variants have been reported in the literature as associated with a developmental disorder and different types of cancer (represented with lollipops above Pfam domains in Figure 7) and only a VUS is reported in ClinVar. All the pathological missense variants represent drastic physiochemical changes, and affect highly constrained and highly conserved sites.

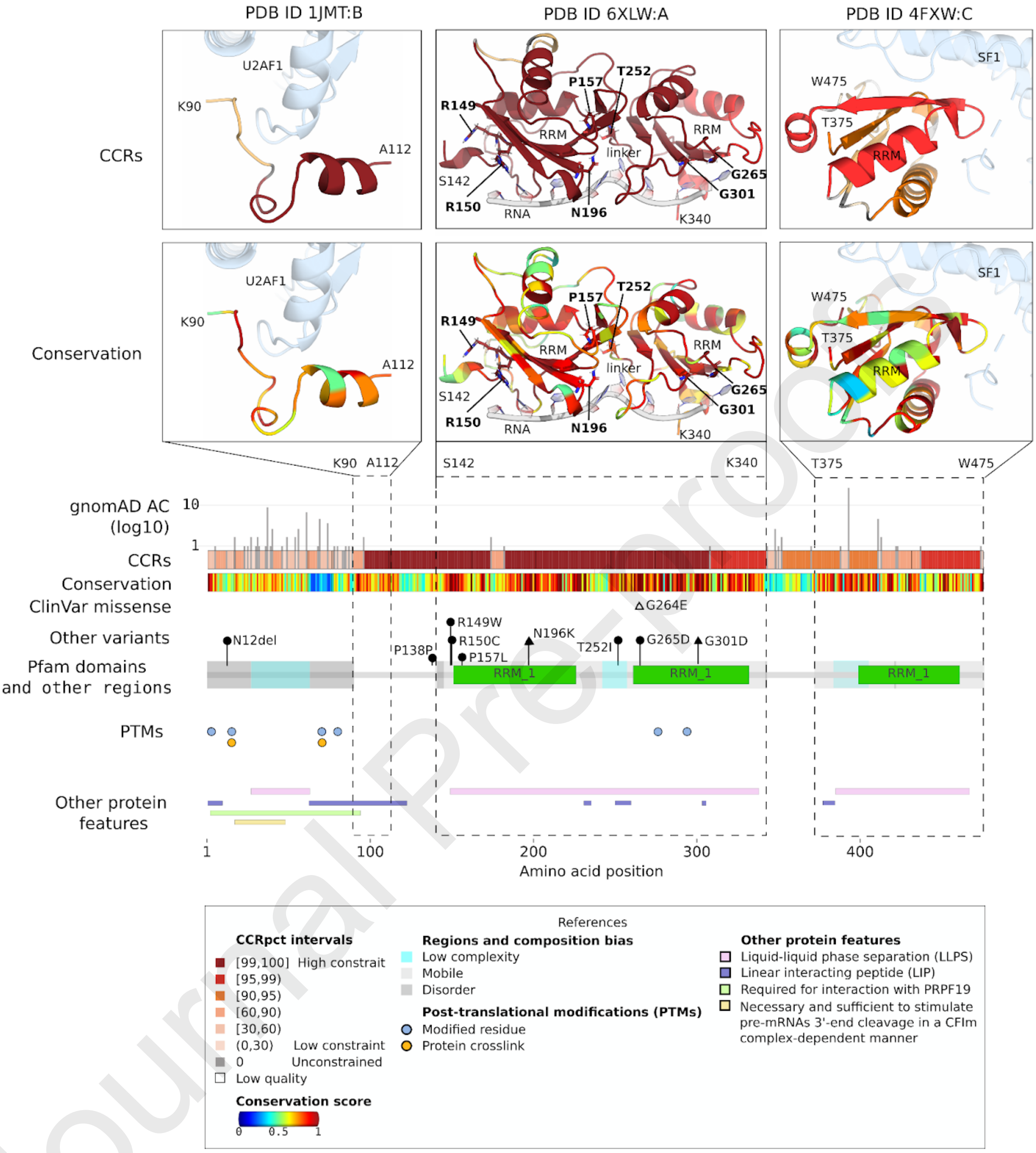


Figure 7: An example of a protein with long regions highly constrained for variability: *U2AF2,* the 65 kDa subunit of the U2 auxiliary splicing factor U2AF *(*also known as *U2AF65*, UniProtKB: P26368, Ensembl: ENST00000308924*)*. From middle to bottom the different panels represent, by amino acid positions, gnomAD3.0 allele counts (AC), CCRpct, species conservation score, Pfam domain with disorder/mobility and low-complexity regions, post-translationally modified sites (PTMs), other protein features, as listed in the Figure. Pfam domains correspond to: 'RRM_1'=RNA recognition motif, RNP-1. On top of these panels, the dashed rectangles call out regions of the protein with available PDB structures.

The lollipops above the Pfam domains depict positions with *de novo* variants reported in bibliography as associated to developmental disorders (N12del, P138P, R149W, R150C, P157L, T252I and G265G with black circles) [48], acute myeloid leukaemia (N196K), colon adenocarcinoma and castration-resistant prostate carcinoma (G301D) (black triangles) [49]. The unfilled triangle represents a variant of uncertain significance reported in ClinVar (G264E). By-residue ScoreCons conservation scores were obtained from VarSite. Low-complexity, mobility, disorder LLPS, and LIP annotations were obtained from the MobiDB database. PTMs and other interacting regions were obtained from UniProtKB. Interacting proteins in the PDB structures are shown in pale blue, with their interacting side chain shown in stick representation.

U2AF2 illustrates how CCRpct and conservation both highlight that this protein is not only essential, but also has functional protein sites along its length. The clinical data supports this hypothesis, with variants associated with developmental disorders and cancers.

The second example is shown in Figure 8, which illustrates how CCR data, combined with protein function information, can highlight regions with potential functions that have yet to be determined. The gene is *SLC12A2*, which encodes for the solute carrier family 12 member 2 protein, a Na+, K+ and 2Cl– cotransporter 1 (NKCC1), which plays a critical role in the homeostasis of K+ enriched endolymph in the membranous labyrinth of the inner ear. NKCC1 subunits are ion channels that work as homodimers, and each protein monomer comprises a transmembrane domain (TMD), cytosolic N- and C-terminal domains (NTD and CTDs, respectively) and extracellular flexible loops stabilised by disulphide bonds. The dimeric interface involves interactions between TMDs and CTDs [50,51]. *SLC12A2* is overall highly constrained for missense and LoF variability in gnomAD2.1 (missense OEUF=0.78, missense Z-score=2.4, pLI=0.96, LOEUF=0.31), suggesting essentiality for humans. The protein structures (Figure 8), reveal that the highest CCRpct in this protein (in the range [90,99)) corresponds to functionally important regions: residues in pore-lining helices (involved in the ion flow through the channel), the dimer interface, and also a ‘not so obviously relevant’ disordered and lowly conserved (scores < 0.4) region in the C-terminal domain. Intriguingly, this last region, comprising amino acids 977 to 993, is fully encoded by exon 21 of *SLC12A2* and coincides with the boundaries of a CCR ranked with a moderate percentile of 94 (small dashed rectangle in Figure 8). It has previously been noted that this region, whose functionality still remains unclear, is unique to SLC12A2 and is not shared with the other proteins in the SLC12 family [52], suggesting that it might confer a specific functional characteristic to this protein.

Seven pathogenic missense variants have been reported in ClinVar for this protein. Two of them, associated with Delpire-McNeill neurodevelopmental syndrome, are in the TMD in highly conserved and low-medium constrained residues: N376I lining the pore (conservation=1, CCRs pct=37.68) and A327V adjacent to a pore-lining helix (conservation=0.72, CCRs pct=74.95). The other five: E979K, E980K, D981Y, P988T and P988S cause deafness and sensorineural hearing loss, and cluster in the moderately constrained region encoded by exon 21. Furthermore, functional assays in cultured cells showed that applying the variants E979K, D981Y and P988T, or skipping exon 21, significantly decreases chloride influx mediated by the SLC12A2 protein [53]. All this evidence and the moderately high CCRpct for this disordered and lowly conserved region, highlight its putative relevance for the function of this protein.

Both the U2AF2 and SLC12A2 proteins described above also serve to exemplify different scenarios for amino acid sites where high CCRs percentiles go hand in hand with high conservation, and the converse where high CCRs percentiles go with low conservation, and vice versa.

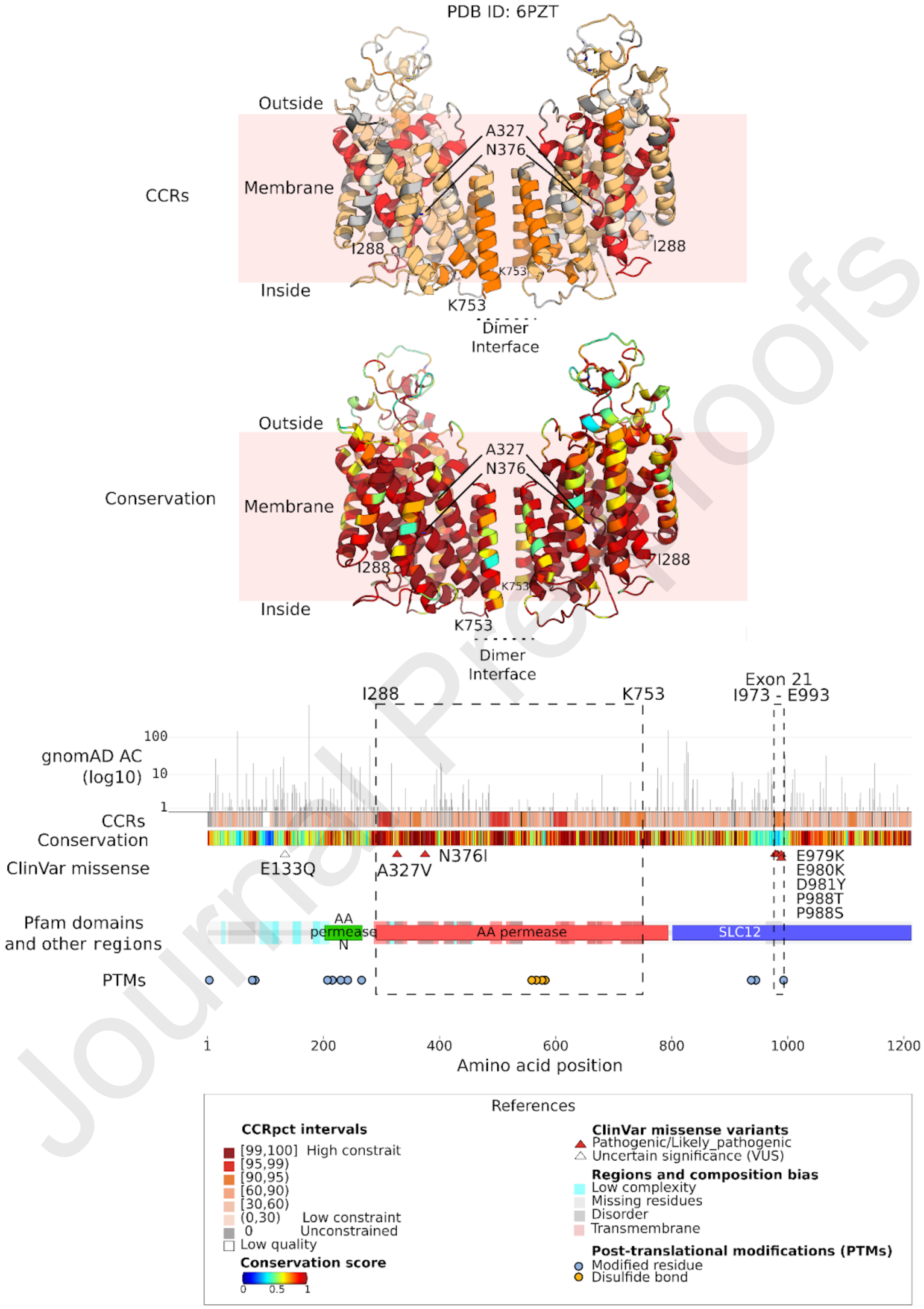


Figure 8: An example where CCRs highlight regions important for protein function: the pore lining helices and dimer interface in the transmembrane region of *SLC12A2* (NKCC1) solute carrier

family 12 member 2 *(*UniProtKB: P55011-1, Ensembl: ENST00000262461*)* and a peculiar disordered region of this protein encoded by its exon 21 and with a cluster of pathogenic/likely_pathogenic variants related to deafness and hearing loss [53]. From middle to bottom the plots in the horizontal panels represent, by amino acid position, gnomAD3.0 allele counts (AC), CCRpct, species conservation scores, sites with ClinVar missense variants, Pfam domains with disorder/mobility, low-complexity and transmembrane regions and post-translationally modified sites (PTMs). Pfam domains correspond to: 'AA permease N'=Amino acid permease N-terminal, 'AA permease'= Amino acid permease, 'SLC12'=Solute carrier family 12. Over these domains, the dashed rectangles call out the transmembrane region of the protein characterised in the 6PZT PDB structure as a homodimer. This structure is coloured by CCRpct and by conservation and displayed on the top panels. The location in the structure of the pathogenic/likely_pathogenic ClinVar variants A327V and N376I is depicted with triangles and squares, respectively. The small dashed rectangle fully encloses exon 21 (amino acids 977-993). Residue conservation scores, as calculated by ScoreCons, were obtained from VarSite, domains from Pfam, low-complexity, mobility and disorder from MobiDB, transmembrane regions and PTMs from UniProtKB.

Discussion

In the present work we extended the characterisation of constrained coding regions in the human genome, by accurately fine-mapping these regions and their level of constraint from the Human Build 38 genomic coordinates to protein sequence coordinates in 17,366 human UniProt canonical sequences, totalling about 9.8 million amino acid positions. Furthermore, aggregating protein functional annotations, available for these positions, allowed us to analyse the distribution and correlation of the different levels of constraint and inter-species conservation with different protein features.

Overall, our results agreed with the previous observations of Havrilla et al. 2019 that the correlation between the CCRpct and the average nucleotide GERP++ conservation scores [27] for the regions is very low and hence the intra-species conservation in humans complements the interspecies conservation.

For the catalytic sites and interactions with different partners (small molecules, proteins, DNA/RNA, metals), we observed the expected associations between high percentiles of constraint and high conservation scores. This is in concordance with the observations of

Havrilla et al. [1] that domains enriched with the most highly constrained regions were involved in ion transport and in different DNA/RNA interactions (like zinc fingers, helicases and translation factors). Additionally, we observed that the unconstrained (i.e. with gnomAD3.0 variants) or lowly constrained (i.e. average shorter regions depleted of variants) regions were mostly associated with signalling regions (signal, propeptides and transit peptides), low complexity, glycosylation sites, and with more mixed inter-species conservation scores.

Surprisingly, the transmembrane regions showed little if any enrichment for highly constrained regions, but slightly higher enrichment for medium constraint (CCRpct [60,90)), i.e. on average shorter regions. Perhaps, this reflects the lipid environment where variants between the hydrophobic amino acids are common.

Among the unexpected results, we observed that disulphide bond cysteines were more prone to lie within regions with low to medium percentiles of constraint (CCRpct in (0,90)). Disulphide bonds are covalent tertiary interactions important for stabilising protein folds and/or performing physiologically relevant redox activity and hence highly conserved in evolution [54]. We hypothesise that the association with lower-medium CCRpct (i.e. average shorter regions depleted of variants, with mean length=20 amino acidos) reflects the fact that the formation of such bonds requires only the presence of short motifs involving only the cysteines and their immediate flanking residues [55]).

Perhaps the most unexpected results we observed were related to disordered and mobile regions in proteins, showing dual enrichment for unconstrained/lowly constrained and also for highly constrained percentiles, mostly in sites with low conservation, and this might relate to the multiplicity of functions, or “flavours”, of disorder that such regions can present, which depend on their length, composition and location in proteins [14,56]. Disordered proteins and regions are able to fulfil a variety of tasks: they can serve as flexible linkers between structured regions or flexible binding sites for ligands, they can undergo disorder-order transitions upon binding to other proteins through specific molecular recognition features (MoRFs) within longer disordered regions, they can also have short linear motifs that work as targets for post-translational modifications or cell signalling, or longer regions which promote molecular recognition and protein-protein interactions. The characterisation of the dynamic of IDPs/IDRs has led to the identification of their plausible role in regulating enzymatic activity [57] and has also been useful to investigate ligand selection for developing drugs [58]. This motivates the necessity of characterising disordered proteins and regions, for discovering the function and relevant mechanism where they are involved.

In the present work, in particular, residues involved in D-to-O, context dependent transitions and in driving LLPS showed association with high constraints, for conserved and also unconserved sites. In the case of residues in order-disorder transitions, our observations align with what has been proposed in terms of the binding mechanisms. D-to-O are defined by a single, well-defined, fully ordered binding configuration, mediated by a unique well-defined contact pattern that excludes ambiguities and is determined by the presence of binding motifs. Context dependent transitions involve alternative binding configurations, which change with the cellular conditions and different partners. Conversely, D-to-D transitions are defined by many different binding configurations, including alternative contact patterns, often with weak or redundant motifs [15]. For amino acids driving LLPS, our results suggest a strong association with constrained regions ranked with the highest percentiles ([95,100]) and that such regions are, on average, the longest stretches (75 amino acids in length) depleted of protein changing variants across the human coding genome.

The scientific community is beginning to untangle the complexity of interactions and regulations involved in LLPS, and evidence shows that these protein condensates do not follow classical rules of molecular recognition. LLPS regions are generally mobile/disordered and involve long sequence stretches that orchestrate multiple and multivalent interactions with proteins and RNA for the formation of membrane-less organelles in the cell. They are important for organising and regulating key cellular processes such as transcription, splicing, translation, chromosome condensation, synapsis and downstream signalling, all essential for tightly regulating the differential expression of genes, ensuring cell survival, correct differentiation into different tissues, and for the development and function of the neuronal and immune systems [59–63]. However, it was remarkable that amino acid positions in regions driving LLPS were not observed as significantly associated with Pathogenic protein altering variants (Supplementary Figure 7), considering that previous works have observed these genes frequently related to cancer, autism spectrum disorders, neurodegeneration, and infectious diseases [21–23]. Here, apart from these associations, we noticed that 31.5% of these proteins were involved in diseases with profound impact in the normal human postnatal and early development, with a high prevalence of neurodevelopmental disorders. In addition, 64% of the LLPS driving genes were highly constrained for loss-of-function in heterozygosity and also, 44% of the 54 genes were not associated with any disease with 14 of them, mostly associated to RNA metabolism being highly constrained for loss-of-function variation. We hypothesise that variants in such genes could possibly cause severe phenotypes and affect embryonic viability.

Most of the significantly enriched GO terms for proteins with highly constrained regions (CCRpct in [95,100]) were related to RNA-processing, DNA binding, protein-protein interactions, and enzymatic activities. This is in concordance with our observations that amino acid sites functionally annotated as binding DNA/RNA and/or proteins, in catalytic sites and in LIPs, and/or driving LLPS, are among the ones with the greatest odds of being highly constrained, i.e. in the, on average, longest regions in the human genome intolerant to variability.

Our results also complement and extend what the authors of the CCRs model [1] have derived before by analysing the co-occurrence with Pfam domains and observing that the highly constrained regions are involved in ion transport and in different DNA/RNA interactions (like zinc fingers, helicases and translation factors), but also that about 30% of these highly constrained regions did not correspond to any protein Pfam [3] domain.

Undoubtedly, the sequencing of genetically more diverse human populations will refine some CCRs further, but the data presented here has significant clinical utility. The key challenge for clinical genomics is interpreting the pathogenicity of rare variants. Identifying whether a rare variant lies within a defined constrained region of the protein facilitates consequence interpretation especially for novel variants absent from existing genomic databases.

We emphasise that combining interspecies and intraspecies (human population) conservation can help to highlight regions of individual genes that have appeared more recently in evolution or confer some degree of uniqueness/specificity to an individual paralogue. This data has the potential to facilitate the discovery of new associations between genes/variants with previously unknown phenotypes. CCRs also highlight many highly constrained regions currently not linked to any Mendelian disease. This may indicate mutations in these regions are lethal to humans or are sufficiently rare that they have not yet been identified [1].

The current tools that attempt to predict the clinical relevance of a specific sequence variant have been developed mostly based on the characteristics of folded protein regions [64] making it difficult to understand the effect of variants affecting intrinsically disordered/mobile and liquid-liquid phase separating regions. Furthermore, many protein functional sites still remain to be characterised and currently lack sufficient functional annotations, in particular the difficult cases, where flexible linkers/disordered regions are poorly characterised and/or can be poorly conserved across species while being constrained, at different levels, in human populations. Our mapping of CCRs to amino acids helps to define these regions in proteins more accurately and could contribute to the further annotation of these challenging regions.

Our approach, however, is limited to the analysis of single features, we are aware of the possibility that multiple features can co-occur for the same protein site or that other confounding biological factors can be present. Furthermore, because the CCRs were defined from sequences, i.e. in a uni-dimensional or linear space, and the weighting of the regions takes into consideration their length, the model does not consider the possibility of having short lowly constrained regions coming together in the three-dimensional space to define a larger structural cluster constrained for variability. Looking forward, we believe that all these current limitations will open new avenues for further research and refinement.

Methods

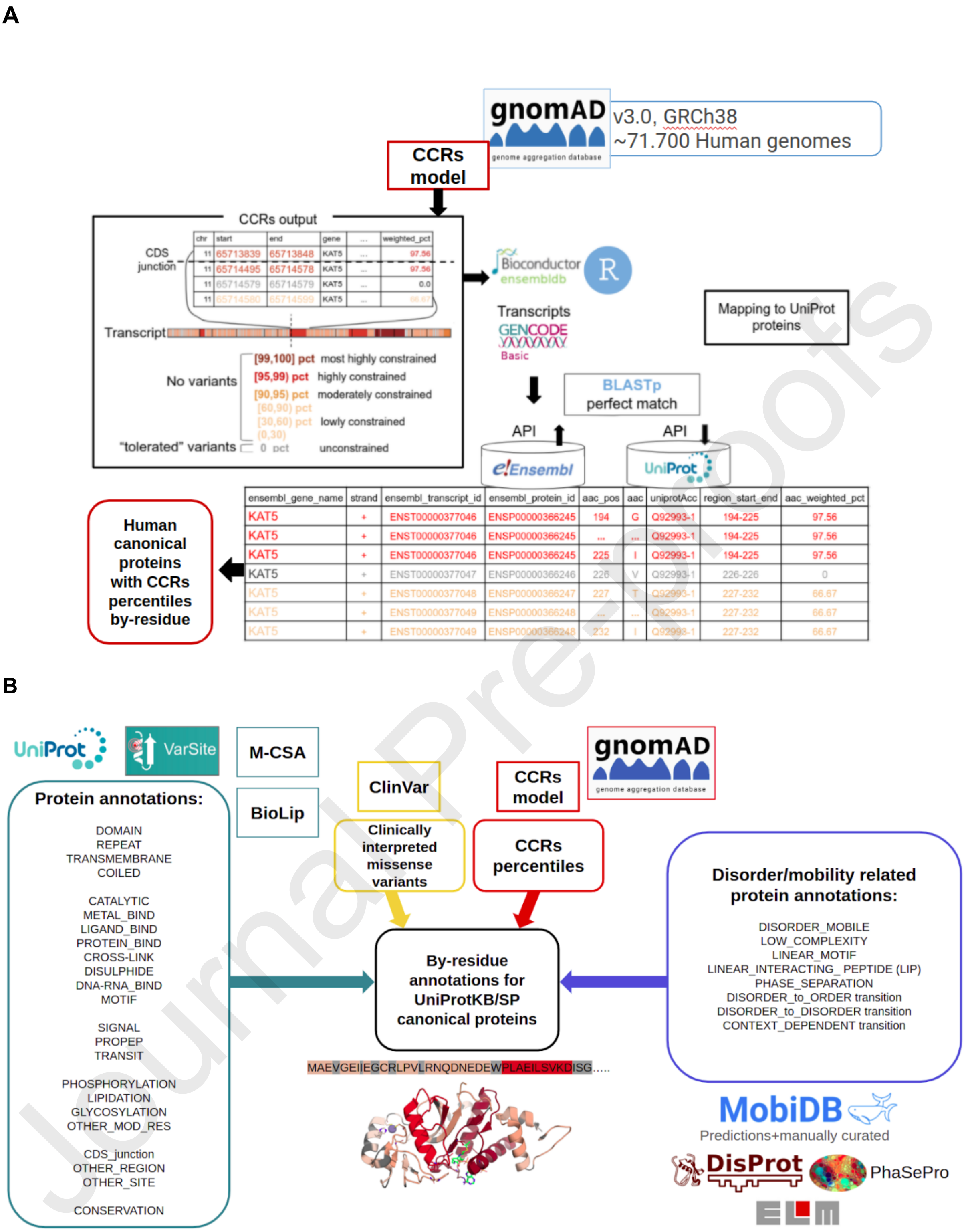
Generation of the CCRs based on gnomAD3

To obtain the CCRs we ran the pipeline developed by [1] (https://quinlan-lab.github.io/ccr/examples/updates) but employing the dataset of gnomAD [26] version 3.0 and the corresponding files for the coordinates of the human genome in version GRCh38 [65] (https://www.ncbi.nlm.nih.gov/grc). We used the Variant Effect Predictor (VEP) [66] of Ensembl [25] version 101. We also followed the recommendations of the authors to only consider genetic variants from autosomes and chromosome X, and avoid those in conflicting genomic regions - i.e. where there are segmental duplications and/or high identity with other genomic regions (>=90% identity) or with low sequencing coverage. In the same line of recommendations, we ran the weighting of the regions for autosomes and X chromosome separately, but merged both output files into one for performing the mapping of the coordinates of the regions to protein amino acids.

Mapping the CCRs to protein amino acids

We developed an in-house pipeline in R that uses the '*ensembldb*' Bioconductor R package [67] to map the genomic coordinates of CCRs boundaries, and all the coding bases in between, to the Ensembl v101 transcripts which are part of the GENCODE [68] basic set version 35. This was to ensure we were including complete and well annotated relevant transcripts. For those amino acid sites where the corresponding codon had constrained and unconstrained bases, we assigned such amino acid sites as unconstrained. We then obtained the sequence identifiers that crosslink Ensembl transcripts and proteins and

UniProtKB proteins [24] by querying the APIs (application programming interfaces) of both databases. Finally, the CCRpct were accurately transferred to the amino acids in UniProtKB sequences by downloading the corresponding protein sequences from Ensembl and UniProtKB and performing Blastp local alignments [69] requesting 100% sequence identity (perfect match). The workflow is summarised in Figure 9 A, explained more in detail in Supplementary Methods and the corresponding scripts of the pipeline are available in this repository https://github.com/marciaah/CCRStoAAC.



chr	start	end	gene	...	weighted_pct
11	65713839	65713848	KAT5	...	97.56
11	65714495	65714578	KAT5	...	97.56
11	65714579	65714579	KAT5	...	0.0
11	65714580	65714599	KAT5	...	66.67

ensembl_gene_name	strand	ensembl_transcript_id	ensembl_protein_id	aac_pos	aac	uniprotAcc	region_start_end	aac_weighted_pct
KAT5	+	ENST00000377046	ENSP00000366245	194	G	Q92993-1	194-225	97.56
KAT5	+	ENST00000377046	ENSP00000366245	...	...	Q92993-1	194-225	97.56
KAT5	+	ENST00000377046	ENSP00000366245	225	I	Q92993-1	194-225	97.56
KAT5	+	ENST00000377047	ENSP00000366246	226	V	Q92993-1	226-226	0
KAT5	+	ENST00000377048	ENSP00000366247	227	T	Q92993-1	227-232	66.67
KAT5	+	ENST00000377049	ENSP00000366248	...	...	Q92993-1	227-232	66.67
KAT5	+	ENST00000377049	ENSP00000366248	232	I	Q92993-1	227-232	66.67

Figure 9: A) Flowchart showing the different databases and tools employed for mapping the CCRs in genomic coordinates to the amino acid coordinates in UniProtKB protein sequences. B) Flowchart presenting the different resources and databases employed for aggregating 30 general protein feature annotations and conservation score (blue-green boxes), clinically interpreted variants (yellow boxes), CCRpct (red boxes) and

disorder/mobile related protein feature annotations (blue boxes). UniProtKB/SP= UniProtKB/SwissProt

Aggregation of protein features annotations and clinically interpreted variants

We developed our own pipeline in R for fetching different protein features and functional annotations from multiple resources. For this purpose, we captured annotations and based our analysis only for the 9.8 million protein sites in UniProtKB/SwissProt canonical sequences because such sequences are the main references for annotations in the databases we employed. An overview of the databases and features that we included are presented in Figure 9 B, and obtained as described more in detail in Supplementary Methods.

Gene Ontology enrichment tests

The GO statistical over-representation tests were performed using the PANTHER classification system [40] (http://www.pantherdb.org/tools/index.jsp, PANTHER version 17.0 release 22-02-2022, with Reactome version 65), submitting the list of genes of interest (e.g. those presenting regions with percentiles in [95,100]) and using as "reference list' only those genes for which we were able to map CCRpct.

Odds ratios tests for enrichment

We performed four different Odds ratio (OR) test analyses to measure the enrichment of amino acid sites presenting different combinations of CCRpct, conservation, protein features and ClinVar variants:

I. CCRpct and presence of each one of the 30 protein features: we binary assigned whether or not an amino acid site had any of the protein features (see Figure 11 for full list) and a CCRpct in any of the 7 bins: unconstrained=[0]; low-medium constraint= (0,30), [30,60) and

[60,90); moderately constrained= [90,95), highly constrained= [95,99) and most highly constrained= [99,100].

II. CCRpct and conservation with the presence of protein features: we binary classified the amino acid sites as having or not any of the 30 protein features and any of the 6 combinations: a) CCRs unconstrained (0 pct) and conservation score ≤0.5, b) CCRs unconstrained (0 pct) and conservation score >0.5, c) CCRs in (0,50] pct and conservation score ≤0.5, d) CCRs in (0,50] pct and conservation score >0.5, e) CCRs in (50,100] pct and conservation score ≤0.5, f) CCRs in (50,100] pct and conservation score >0.5.

III. CCRpct and presence of ClinVar variants: amino acid sites were binary assigned whether or not they had "pathogenic/likely_pathogenic", "benign/likely_benign" or "VUS/conflicting interpretations of pathogenicity" variants and a CCRpct in any of the 7 bins mentioned in (I).

IV. *CCRpct and conservation with the presence of ClinVar variants:* we classified residues according to whether they had or not any of the 3 groups of variants as described in (III) and any of the 6 combinations of CCRpct and conservation as described in (II).

For the four enrichment analyses, contingency tables were constructed counting amino acid sites with the different classifications (See Supplementary Methods: *OR tests for enrichment* for further details) and the OR were calculated using two-tailed and one-tailed Fisher's exact tests [70] for obtaining the corresponding P-values and 95% confidence intervals (CI 95%). It is worth clarifying that when counting residues we did not request exclusivity in the intersections, i.e. a residue with a given CCRpct can intersect with being in DOMAIN, DOSORDER_MOBILE and DNA-RNA_BIND and hence will contribute to the cells in the three corresponding contingency tables.

Acknowledgments

This work was funded by the Cambridge NIHR Biomedical Research Centre (BRC-1215-200014) and the 3D-FunSites Project (Wellcome Trust Ref. number 221327/Z/20/Z). The

Wellcome Trust, the European Bioinformatics Institute (EMBL-EBI) and the Medical Research Council have also funded research infrastructure. We thank Prof. Raymond's and Prof. Thornton's group members for their valuable comments and suggestions.

The code and data used for the present analysis is provided in GitHub repositories, as mentioned in the Methods section. The authors welcome requests for additional information regarding the material presented in this paper.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Appendix: Supplementary Materials

I. Supplementary tables:

Supplementary Table 1: OR and Fisher Exact Tests assessing the associations between protein features and the CCRpct groups.

Supplementary Table 2: OR and Fisher Exact Tests assessing the associations between protein features and each one of the 6 groups combining different CCRpct and inter-species conservation.

Supplementary Table 3: OR and Fisher Exact Tests assessing the co-occurrence of amino acid positions affected by ClinVar missense variants with the different categories of CCRs percentiles.

Supplementary Table 4: Odd ratios (OR) of co-occurrence of amino acid positions affected by ClinVar missense variants with the different categories of CCRpct and inter-species conservation.

Supplementary Table 5: List of human proteins that harbour LLPS driving regions, and that have CCRpct assigned and their association with different clinical conditions.

*Supplementary Table 5: Lis*t of human clinical conditions (ClinVar) and associated proteins which are involved in driving LLPS.

II. Supplementary Figures:

Supplementary Figure 1: comparison of pLI and missense OEUF and the maximum CCRs percentile by protein

Supplementary figure 2: The higher the constrained percentile the larger the mean length of the regions and the higher the mean conservation of the amino acids within them. However, each percentile category exhibits a high variability

Supplementary Figure 3: Comparison of full length of proteins and the number of regions in each CCRs percentile that they have.

Supplementary Figure 4: Distribution of the raw count of amino acid positions grouped by different categories of CCRs percentiles and conservation scores.

Supplementary figure 5: Comparison of OR tests, conservation and length of regions by protein feature.

Supplementary figure 6: Comparison of OR tests, conservation and length of regions by type of ClinVar variants.

Supplementary Figure 7: Amino acid positions affected with clinically interpreted variants assessed by their location in the different protein features or functional sites

III. Supplementary Methods

References

[1] J.M. Havrilla, B.S. Pedersen, R.M. Layer, A.R. Quinlan, A map of constrained coding regions in the human genome, Nat. Genet. 51 (2019) 88–95.

[2] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A.H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, T. Tukiainen, D.P. Birnbaum, J.A. Kosmicki, L.E.

Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D.N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M.I. Kurki, A.L. Moonshine, P. Natarajan, L. Orozco, G.M. Peloso, R. Poplin, M.A. Rivas, V. Ruano-Rubio, S.A. Rose, D.M. Ruderfer, K. Shakir, P.D. Stenson, C. Stevens, B.P. Thomas, G. Tiao, M.T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D.M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J.C. Florez, S.B. Gabriel, G. Getz, S.J. Glatt, C.M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M.I. McCarthy, D. McGovern, R. McPherson, B.M. Neale, A. Palotie, S.M. Purcell, D. Saleheen, J.M. Scharf, P. Sklar, P.F. Sullivan, J. Tuomilehto, M.T. Tsuang, H.C. Watkins, J.G. Wilson, M.J. Daly, D.G. MacArthur, Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans, Nature. 536 (2016) 285–291.

[3] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman, Pfam: The protein families database in 2021, Nucleic Acids Res. 49 (2021) D412–D419.

[4] I. Sillitoe, N. Bordin, N. Dawson, V.P. Waman, P. Ashford, H.M. Scholes, C.S.M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S.D. Lam, K. Berka, I.H. Varekova, R. Svobodova, J. Lees, C.A. Orengo, CATH: increased structural coverage of functional space, Nucleic Acids Res. 49 (2021) D266–D273.

[5] S. Gudmundsson, M. Singer-Berk, N.A. Watts, W. Phu, J.K. Goodrich, M. Solomonson, Genome Aggregation Database Consortium, H.L. Rehm, D.G. MacArthur, A. O'Donnell-Luria, Variant interpretation using population databases: Lessons from gnomAD, Hum. Mutat. (2021). https://doi.org/10.1002/humu.24309.

[6] K.E. Samocha, J.A. Kosmicki, K.J. Karczewski, A.H. O'Donnell-Luria, E. Pierce-Hoffman, D.G. MacArthur, B.M. Neale, M.J. Daly, Regional missense constraint improves variant deleteriousness prediction, (n.d.). https://doi.org/10.1101/148353.

[7] Y.-F. Huang, Unified inference of missense variant effects and gene constraints in the human genome, PLoS Genet. 16 (2020) e1008922.

[8] M. Zhao, J.M. Havrilla, L. Fang, Y. Chen, J. Peng, C. Liu, C. Wu, M. Sarmady, P. Botas, J. Isla, G.J. Lyon, C. Weng, K. Wang, Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases, NAR Genom Bioinform. 2 (2020) lqaa032.

[9] D. Šimčíková, P. Heneberg, Refinement of evolutionary medicine predictions based on clinical evidence for the manifestations of Mendelian diseases, Sci. Rep. 9 (2019) 18577.

[10] P. Evans, C. Wu, A. Lindy, D.A. McKnight, M. Lebo, M. Sarmady, A.N. Abou Tayoun, Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets, Genome Res. 29 (2019) 1144–1151.

[11] F.K. Satterstrom, J.A. Kosmicki, J. Wang, M.S. Breen, S. De Rubeis, J.-Y. An, M. Peng, R. Collins, J. Grove, L. Klei, C. Stevens, J. Reichert, M.S. Mulhern, M. Artomov, S. Gerges, B. Sheppard, X. Xu, A. Bhaduri, U. Norman, H. Brand, G. Schwartz, R. Nguyen, E.E. Guerrero, C. Dias, Autism Sequencing Consortium, iPSYCH-Broad Consortium, C. Betancur, E.H. Cook, L. Gallagher, M. Gill, J.S. Sutcliffe, A. Thurm, M.E. Zwick, A.D. Børglum, M.W. State, A.E. Cicek, M.E. Talkowski, D.J. Cutler, B. Devlin, S.J. Sanders, K. Roeder, M.J. Daly, J.D. Buxbaum, Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism, Cell. 180 (2020) 568–584.e23.

[12] A. Sanchis-Juan, M.A. Hasenahuer, J.A. Baker, A. McTague, K. Barwick, M.A. Kurian, S.T. Duarte, NIHR BioResource, K.J. Carss, J. Thornton, F.L. Raymond, Structural analysis of pathogenic missense mutations in GABRA2 and identification of a novel de novo variant in the desensitization gate, Mol Genet Genomic Med. 8 (2020) e1106.

[13] C. Rodger, E. Flex, R.J. Allison, A. Sanchis-Juan, M.A. Hasenahuer, S. Cecchetti, C.E. French, J.R. Edgar, G. Carpentieri, A. Ciolfi, F. Pantaleoni, A. Bruselles, Genomics England Research Consortium, R. Onesimo, G. Zampino, F. Marcon, E. Siniscalchi, M. Lees, D. Krishnakumar, E. McCann, D. Yosifova, J. Jarvis, M.C. Kruer, W. Marks, J. Campbell, L.E. Allen, S. Gustincich, F.L. Raymond, M. Tartaglia, E. Reid, De Novo VPS4A Mutations Cause Multisystem Disease with Abnormal Neurodevelopment, Am. J. Hum. Genet. 107 (2020) 1129–1148.

[14] R. van der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, P.M. Kim, R.W. Kriwacki, C.J. Oldfield, R.V. Pappu, P. Tompa, V.N. Uversky, P.E. Wright, M.M. Babu, Classification of intrinsically disordered regions and proteins, Chem. Rev. 114 (2014) 6589–6631.

[15] M. Fuxreiter, Classifying the Binding Modes of Disordered Proteins, Int. J. Mol. Sci. 21 (2020). https://doi.org/10.3390/ijms21228615.

[16] S. Brocca, R. Grandori, S. Longhi, V. Uversky, Liquid-Liquid Phase Separation by Intrinsically Disordered Protein Regions of Viruses: Roles in Viral Life Cycle and Control of Virus-Host Interactions, Int. J. Mol. Sci. 21 (2020). https://doi.org/10.3390/ijms21239045.

[17] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29.

[18] G. Fusco, S. Gianni, Function, Regulation, and Dysfunction of Intrinsically Disordered Proteins, Life. 11 (2021) 140. https://doi.org/10.3390/life11020140.

[19] V. Vacic, L.M. Iakoucheva, Disease mutations in disordered regions--exception to the rule?, Mol. Biosyst. 8 (2012) 27–32.

[20] V.N. Uversky, C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human diseases: introducing the D2 concept, Annu. Rev. Biophys. 37 (2008) 215–246.

[21] B. Tsang, I. Pritišanac, S.W. Scherer, A.M. Moses, J.D. Forman-Kay, Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations, Cell. 183 (2020) 1742–1756.

[22] J. Li, Y. Zhang, X. Chen, L. Ma, P. Li, H. Yu, Protein phase separation and its role in chromatin organization and diseases, Biomed. Pharmacother. 138 (2021) 111520.

[23] B. Wang, L. Zhang, T. Dai, Z. Qin, H. Lu, L. Zhang, F. Zhou, Liquid-liquid phase separation in human health and diseases, Signal Transduct Target Ther. 6 (2021) 290.

[24] UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (2021) D480–D489.

[25] K.L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, A.G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. Garcia Giron, T. Grego, C. Guijarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O.G. Izuogu, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. Gonzalez Martinez, J.C. Marugán, T. Maurel, A.C. McMahon, S. Mohanan, B. Moore, M. Muffato, D.N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M.P. Sakthivel, A.I.A. Salam, B.M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A.

Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. De Silva, B. Flint, A. Frankish, S.E. Hunt, G.R. IIsley, N. Langridge, J.E. Loveland, F.J. Martin, J.M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S.J. Trevanion, F. Cunningham, A.D. Yates, D.R. Zerbino, P. Flicek, Ensembl 2021, Nucleic Acids Res. 49 (2021) D884–D891.

[26] K.J. Karczewski, L.C. Francioli, G. Tiao, B.B. Cummings, J. Alföldi, Q. Wang, R.L. Collins, K.M. Laricchia, A. Ganna, D.P. Birnbaum, L.D. Gauthier, H. Brand, M. Solomonson, N.A. Watts, D. Rhodes, M. Singer-Berk, E.M. England, E.G. Seaby, J.A. Kosmicki, R.K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J.X. Chong, K.E. Samocha, E. Pierce-Hoffman, Z. Zappala, A.H. O'Donnell-Luria, E.V. Minikel, B. Weisburd, M. Lek, J.S. Ware, C. Vittal, I.M. Armean, L. Bergelson, K. Cibulskis, K.M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M.E. Talkowski, Genome Aggregation Database Consortium, B.M. Neale, M.J. Daly, D.G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans, Nature. 581 (2020) 434–443.

[27] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, S. Batzoglou, Identifying a high fraction of the human genome to be under selective constraint using GERP++, PLoS Comput. Biol. 6 (2010) e1001025.

[28] W.S.J. Valdar, Scoring residue conservation, Proteins. 48 (2002) 227–241.

[29] S. Velankar, Y. Alhroub, A. Alili, C. Best, H.C. Boutselakis, S. Caboche, M.J. Conroy, J.M. Dana, G. van Ginkel, A. Golovin, S.P. Gore, A. Gutmanas, P. Haslam, M. Hirshberg, M. John, I. Lagerstedt, S. Mir, L.E. Newman, T.J. Oldfield, C.J. Penkett, J. Pineda-Castillo, L. Rinaldi, G. Sahni, G. Sawka, S. Sen, R. Slowley, A.W. Sousa da Silva, A. Suarez-Uruena, G.J. Swaminathan, M.F. Symmons, W.F. Vranken, M. Wainwright, G.J. Kleywegt, PDBe: Protein Data Bank in Europe, Nucleic Acids Res. 39 (2011) D402–10.

[30] R.A. Laskowski, J.D. Stephenson, I. Sillitoe, C.A. Orengo, J.M. Thornton, VarSite: Disease variants and protein structure, Protein Sci. 29 (2020) 111–119.

[31] A.J.M. Ribeiro, G.L. Holliday, N. Furnham, J.D. Tyzack, K. Ferris, J.M. Thornton, Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites, Nucleic Acids Res. 46 (2018) D618–D623.

[32] J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions, Nucleic Acids Res. 41 (2013) D1096–103.

[33] D. Piovesan, M. Necci, N. Escobedo, A.M. Monzon, A. Hatos, I. Mičetić, F. Quaglia, L. Paladin, P. Ramasamy, Z. Dosztányi, W.F. Vranken, N.E. Davey, G. Parisi, M. Fuxreiter, S.C.E. Tosatto, MobiDB: intrinsically disordered proteins in 2021, Nucleic Acids Res. 49 (2021) D361–D367.

[34] M. Kumar, S. Michael, J. Alvarado-Valverde, B. Mészáros, H. Sámano-Sánchez, A. Zeke, L. Dobson, T. Lazar, M. Örd, A. Nagpal, N. Farahi, M. Käser, R. Kraleti, N.E. Davey, R. Pancsa, L.B. Chemes, T.J. Gibson, The Eukaryotic Linear Motif resource: 2022 release, Nucleic Acids Res. 50 (2022) D497–D508.

[35] F. Cunningham, J.E. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, O. Austine-Orimoloye, A.G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S.

Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C.G. Giron, T. Genez, J.G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J.C. Marugán, S. Mohanan, A. Mushtaq, M. Naven, D.N. Ogeh, A. Parker, A. Parton, M. Perry, I. Piližota, I. Prosovetskaia, M.P. Sakthivel, A.I.A. Salam, B.M. Schmitt, H. Schuilenburg, D. Sheppard, J.G. Pérez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M.-M. Suner, M. Szpak, A. Thormann, F.F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T.A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S.E. Hunt, G.R. IIsley, J.E. Loveland, F.J. Martin, B. Moore, J.M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S.J. Trevanion, S. Dyer, P.W. Harrison, K.L. Howe, A.D. Yates, D.R. Zerbino, P. Flicek, Ensembl 2022, Nucleic Acids Res. 50 (2022) D988–D995.

[36] M.J. Landrum, J.M. Lee, M. Benson, G.R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J.B. Holmes, B.L. Kattman, D.R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (2018) D1062–D1067.

[37] C.J. Brown, S. Takayama, A.M. Campen, P. Vise, T.W. Marshall, C.J. Oldfield, C.J. Williams, A.K. Dunker, Evolutionary rate heterogeneity in proteins with long disordered regions, J. Mol. Evol. 55 (2002) 104–110.

[38] J.W. Chen, P. Romero, V.N. Uversky, A.K. Dunker, Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder, J. Proteome Res. 5 (2006) 888–898.

[39] P. Beltrao, P. Bork, N.J. Krogan, V. van Noort, Evolution and functional cross-talk of protein post-translational modifications, Mol. Syst. Biol. 9 (2013) 714.

[40] P.D. Thomas, M.J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, A. Narechania, PANTHER: a library of protein families and subfamilies indexed by function, Genome Res. 13 (2003) 2129–2141.

[41] R. Singh, H. Banerjee, M.R. Green, Differential recognition of the polypyrimidine-tract by the general splicing factor U2AF65 and the splicing repressor sex-lethal, RNA. 6 (2000) 901–911.

[42] M. Tari, V. Manceau, J. de Matha Salone, A. Kobayashi, D. Pastré, A. Maucuer, U2AF assemblies drive sequence-specific splice site recognition, EMBO Rep. 20 (2019) e47604.

[43] C.L. Kielkopf, N.A. Rodionova, M.R. Green, S.K. Burley, A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer, Cell. 106 (2001) 595–605.

[44] J.L. Jenkins, K.M. Laird, C.L. Kielkopf, A Broad range of conformations contribute to the solution ensemble of the essential splicing factor U2AF(65), Biochemistry. 51 (2012) 5223–5225.

[45] J.-R. Huang, L.R. Warner, C. Sanchez, F. Gabel, T. Madl, C.D. Mackereth, M. Sattler, M. Blackledge, Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study, J. Am. Chem. Soc. 136 (2014) 7068–7076.

[46] H.-S. Kang, C. Sánchez-Rico, S. Ebersberger, F.X.R. Sutandy, A. Busch, T. Welte, R. Stehle, C. Hipp, L. Schulz, A. Buchbender, K. Zarnack, J. König, M. Sattler, An autoinhibitory intramolecular interaction proof-reads RNA recognition by the essential splicing factor U2AF2, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 7140–7149.

[47] W. Wang, A. Maucuer, A. Gupta, V. Manceau, K.R. Thickman, W.J. Bauer, S.D. Kennedy, J.E. Wedekind, M.R. Green, C.L. Kielkopf, Structure of phosphorylated SF1 bound to U2AF65 in an essential splicing factor complex, Structure. 21 (2013) 197–208.

[48] J. Kaplanis, K.E. Samocha, L. Wiel, Z. Zhang, K.J. Arvai, R.Y. Eberhardt, G. Gallone, S.H. Lelieveld, H.C. Martin, J.F. McRae, P.J. Short, R.I. Torene, E. de Boer, P. Danecek, E.J. Gardner, N. Huang, J. Lord, I. Martincorena, R. Pfundt, M.R.F. Reijnders, A. Yeung, H.G. Yntema, Deciphering Developmental Disorders Study, L.E.L.M. Vissers, J. Juusola, C.F. Wright, H.G. Brunner, H.V. Firth, D.R. FitzPatrick, J.C. Barrett, M.E. Hurles, C. Gilissen, K. Retterer, Evidence for 28 genetic disorders discovered by combining healthcare and research data, Nature. 586 (2020) 757–762.

[49] D. Maji, E. Glasser, S. Henderson, J. Galardi, M.J. Pulvino, J.L. Jenkins, C.L. Kielkopf, Representative cancer-associated U2AF2 mutations alter RNA interactions and splicing, J. Biol. Chem. 295 (2020) 17148–17157.

[50] T.A. Chew, B.J. Orlando, J. Zhang, N.R. Latorraca, A. Wang, S.A. Hollingsworth, D.-H. Chen, R.O. Dror, M. Liao, L. Feng, Structure and mechanism of the cation-chloride cotransporter NKCC1, Nature. 572 (2019) 488–492.

[51] X. Yang, Q. Wang, E. Cao, Structure of the human cation-chloride cotransporter NKCC1 determined by single-particle electron cryo-microscopy, Nat. Commun. 11 (2020) 1016.

[52] K.B. Gagnon, E. Delpire, Physiology of SLC12 transporters: lessons from inherited human genetic mutations and genetically engineered mouse knockouts, Am. J. Physiol. Cell Physiol. 304 (2013) C693–714.

[53] H. Mutai, K. Wasano, Y. Momozawa, Y. Kamatani, F. Miya, S. Masuda, N. Morimoto, K. Nara, S. Takahashi, T. Tsunoda, K. Homma, M. Kubo, T. Matsunaga, Variants encoding a restricted carboxy-terminal domain of SLC12A2 cause hereditary hearing loss in humans, PLoS Genet. 16 (2020) e1008643.

[54] I. Bošnjak, V. Bojović, T. Šegvić-Bubić, A. Bielen, Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank, Protein Eng. Des. Sel. 27 (2014) 65–72.

[55] F. Ferrè, P. Clote, DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification, Nucleic Acids Res. 34 (2006) W182–5.

[56] M. Necci, D. Piovesan, S.C.E. Tosatto, Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe, Protein Sci. 25 (2016) 2164–2174.

[57] M. Palombo, A. Bonucci, E. Etienne, S. Ciurli, V.N. Uversky, B. Guigliarelli, V. Belle, E. Mileo, B. Zambelli, The relationship between folding and activity in UreG, an intrinsically disordered enzyme, Sci. Rep. 7 (2017) 5977.

[58] B.K. Maity, V. Vishvakarma, D. Surendran, A. Rawat, A. Das, S. Pramanik, N. Arfin, S. Maiti, Spontaneous Fluctuations Can Guide Drug Design Strategies for Structurally Disordered Proteins, Biochemistry. 57 (2018) 4206–4213.

[59] S. Gueroussov, R.J. Weatheritt, D. O'Hanlon, Z.-Y. Lin, A. Narula, A.-C. Gingras, B.J. Blencowe, Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing, Cell. 170 (2017) 324–339.e23.

[60] D. Hnisz, K. Shrinivas, R.A. Young, A.K. Chakraborty, P.A. Sharp, A Phase Separation Model for Transcriptional Control, Cell. 169 (2017) 13–23.

[61] X. Su, J.A. Ditlev, E. Hui, W. Xing, S. Banjade, J. Okrut, D.S. King, J. Taunton, M.K. Rosen, R.D. Vale, Phase separation of signaling molecules promotes T cell receptor signal transduction, Science. 352 (2016) 595–599.

[62] B. Tsang, J. Arsenault, R.M. Vernon, H. Lin, N. Sonenberg, L.-Y. Wang, A. Bah, J.D. Forman-Kay, Phosphoregulated FMRP phase separation models activity-dependent translation through bidirectional control of mRNA granule formation, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 4218–4227.

[63] S.E. Reichheld, L.D. Muiznieks, F.W. Keeley, S. Sharpe, Direct observation of structure and dynamics during phase separation of an elastomeric protein, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) E4408–E4415.

[64] S. Stefl, H. Nishi, M. Petukh, A.R. Panchenko, E. Alexov, Molecular mechanisms of disease-causing missense mutations, J. Mol. Biol. 425 (2013) 3919–3936.

[65] V.A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P.A. Kitts, T.D. Murphy, K.D. Pruitt, F. Thibaud-Nissen, D. Albracht, R.S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K.M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J.T. Simpson, G. Threadgold, J. Torrance, J.M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C.-S. Chin, A.M. Phillippy, R. Durbin, R.K. Wilson, P. Flicek, E.E. Eichler, D.M. Church, Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly, Genome Res. 27 (2017) 849–864.

[66] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor, Genome Biol. 17 (2016) 122.

[67] J. Rainer, L. Gatto, C.X. Weichenberger, ensembldb: an R package to create and use Ensembl-based annotation resources, Bioinformatics. 35 (2019) 3151–3153. https://doi.org/10.1093/bioinformatics/btz031.

[68] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J.E. Loveland, J.M. Mudge, C. Sisu, J.C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I.T. Fiddes, C. García Girón, J.M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K.L. Howe, T. Hunt, O.G. Izuogu, R. Johnson, F.J. Martin, L. Martínez, S. Mohanan, P. Muir, F.C.P. Navarro, A. Parker, B. Pei, F. Pozo, F.C. Riera, M. Ruffier, B.M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M.Y. Wolf, J. Xu, Y.T. Yang, A. Yates, D. Zerbino, Y. Zhang, J.S. Choudhary, M. Gerstein, R. Guigó, T.J.P. Hubbard, M. Kellis, B. Paten, M.L. Tress, P. Flicek, GENCODE 2021, Nucleic Acids Res. 49 (2021) D916–D923.

[69] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[70] S.R.A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, 1970.

Marcia A. Hasenahuer: Conceptualization , Methodology, Software, Investigation, Data Curation, Formal analysis, Writing – Original Draft

Alba Sanchis-Juan: Conceptualization, Software, Writing – Review & Editing

Roman A. Laskowski: Resources, Writing – Original Draft

James A. Baker: Writing – Review & Editing

James D. Stephenson: Writing – Review & Editing

Christine A. Orengo: Writing – Review & Editing, Funding acquisition

F Lucy Raymond: Conceptualization, Writing – Original Draft, Supervision, Funding acquisition

Janet M. Thornton: Conceptualization, Writing – Original Draft, Supervision, Funding acquisition

Title

Mapping the Constrained Coding Regions in the human genome to their corresponding proteins

Marcia A. Hasenahuer[1,2,5], Alba Sanchis-Juan[3,4,6], Roman A. Laskowski[1], James A. Baker[1], James D. Stephenson[1], Christine A. Orengo[5], F. Lucy Raymond[2,4], Janet M. Thornton[1]

Highlights

- CCRs are based on human conservation and complement inter-species conservation

- CCRs assist in variant interpretation, here we mapped them onto proteins sites

- The most constrained coding sites correspond to protein sites in interactions

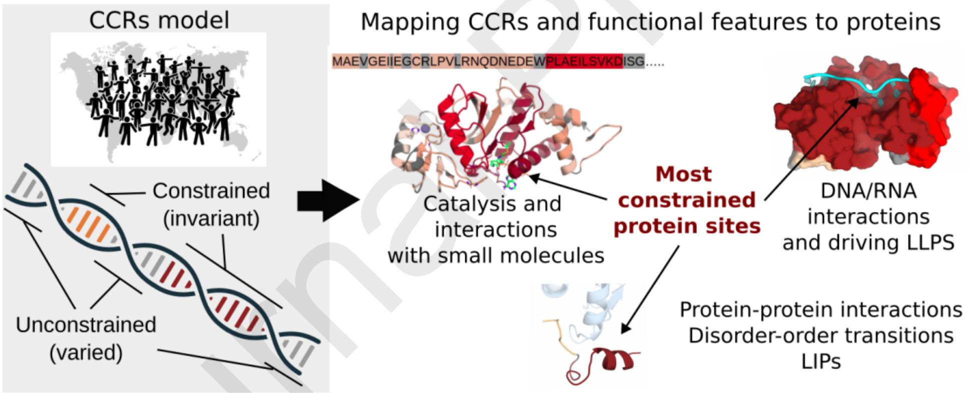
- These interactions include those with DNA/RNA, proteins and in catalytic active sites

- Those driving LLPS, in LIPs and in disorder-order transitions are also highly constrained

Title

Mapping the Constrained Coding Regions in the human genome to their corresponding proteins

Marcia A. Hasenahuer[1,2,5], Alba Sanchis-Juan[3,4,6], Roman A. Laskowski[1], James A. Baker[1], James D. Stephenson[1], Christine A. Orengo[5], F. Lucy Raymond[2,4], Janet M. Thornton[1]


CCRs model
Constrained (invariant)
Unconstrained (varied)
Mapping CCRs and functional features to proteins
MAEVGEIIEGCRLPVLRNQDNEDEWPLAEILSVKDISG.....
Catalysis and interactions with small molecules
Most constrained protein sites
DNA/RNA interactions and driving LLPS
Protein-protein interactions
Disorder-order transitions
LIPs

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: