Title: Proteins encoded by Novel ORFs have increased disorder but can be biochemically regulated and harbour pathogenic mutations

Authors: N. Suhas Jagannathan[1], Narendra Meena[2], Kethaki Prathivadi Bhayankaram[3] & Sudhakaran Prabakaran[2,3,4]*

[1]Cancer and Stem Cell Biology Programme, and Centre for Computational Biology, Duke-NUS Medical School, Singapore 169857

[2]Department of Biology, Indian Institute of Science Education and Research, Pune, Maharashtra, 411008, India

[3]Department of Genetics, University of Cambridge, Downing Site, CB2 3EH, UK.

[4]St Edmund's College, University of Cambridge, CB3 0BN, UK

*Corresponding author, email: sp339@cam.ac.uk.

Abstract

Recent evidence has suggested that protein or protein-like products can be encoded by previously uncharacterized Open Reading Frames (ORFs) that we define as Novel Open Reading Frames (nORFs)[1,2]. These nORFs are present in both coding and non coding regions of the human genome and the novel proteins that they encode have increased the number and complexity of cellular proteome from bacteria to humans. It is a conundrum whether these protein or protein-like products could play any significant functional biological role. But hopes have been raised to target them for anticancer and antimicrobial therapy [3,4]. To infer whether these novel proteins can perform biological functions, we used computational predictions to systematically investigate whether their amino acid sequences can form ordered or disordered structures. Our results indicated that that these novel proteins have significantly higher predicted disorder structure compared to all known proteins, yet we do not find any correlation between the pathogenicity of the mutations and whether they are present in the ordered and disordered regions of these novel proteins. This study reveals that we should investigate these novel proteins more systematically as they may be important to understand complex diseases.

Main text

Over 500 million years of evolution from hydra to humans, the total number of ORFs have been thought to remain the same at around 30,000. With the advent of deep sequencing strategies in both genomics and proteomics fields, we are now discovering nORFs that have remained undiscovered or 'hidden' [1,2,5,6]. These nORFs are pervasive throughout the genome and are observed in both the coding and non coding regions [1,7]. They are variously classified as small ORFs (sORFs) [8,9] which are 1-100 amino acids in length, altORFs [10], which are proteins in alternate frames to known proteins, some are called Denovogenes[11] or Orphan genes [12], some Pseudogenes have been shown to be translated [1,13], and many ncRNAs have been shown to have coding potential [14–18]. These new discoveries pose the question whether our definition of an ORF encoding a protein has been very conservative in the past [19].

While these new discoveries have illuminated that the cellular proteome is much more complex than our current understanding, there is a huge knowledge gap on the putative functions of nORFs. There have been two lines of speculations about it, on one side some have dismissed the novel proteins as mere biological noise [20], and on the other side some have proposed that such nORFs give evolutionary advantage [21,22,23] to an organism. Without a systematic structural investigation of these novel proteins, it is difficult to ascertain which of these two hypotheses might be true.

To investigate these conflicting hypotheses, we first curated a list of all nORFs that have been identified with evidence of translation. We obtained sequences for known and verified human proteins from NextProt (https://www.nextprot.org/) [24], sequences for sORFs from sORF database (http://sorfs.org/database) [8], sequences for altORFs from Roucou's lab [10]

(http://haltorf.roucoulab.com/), sequences of Pseudogenes with evidence of translation from Xu et al [13]. For Denovogenes, we manually curated a list of 42 protein sequences through literature search. For conservative measurements of disorder scores, we discarded protein sequences less than 30 amino acids in length from all these datasets, since these were likely to be enriched for disorder. Non coding RNA sequences were downloaded from RNACentral database (http://rnacentral.org) [25]. While all the other datasets contained translated protein sequences, the downloaded RNA central dataset contained 9,386,637 nucleotide transcript sequences. We did the following to identify potential ORFs from these transcripts. Each sequence was subjected to three-frame translation using the EMBOSS *transeq* program provided as a standalone utility by EMBL-EBI. From the output protein sequences, putative translated ORFs were obtained by identifying all possible subsequences (>30 residues in length) beginning with a Methionine and ending at a STOP codon. This was done using a combination of the EMBOSS *checktrans* program and Matlab scripts to parse the output text files. After removing redundant sequences from the extracted list, we obtained a unique set of 5,185,186 protein sequences, which we used as putative transcripts from the RNAcentral database for disorder prediction. From here on we will call the amino acid sequences encoded by the nORFs as novel proteins.

To investigate the structure forming capabilities of the novel proteins we first plotted their respective length distributions. Although **Figure 1** shows that the lengths of novel proteins are shorter compared to the lengths of known proteins, shorter proteins are known to form elementary structures [26]. We then employed two disorder prediction algorithms, PONDR (*http://www.pondr.com*) and IUPred (*https://iupred2a.elte.hu/*), to assess whether the novel

proteins are predominantly ordered or disordered, which will directly indicate whether they can form structures.

PONDR is an algorithm that uses feedforward neural networks and sequence attributes of amino acid windows (9-21 AAs) to predict disorder. Among PONDR-based algorithms, we used the VSL2 algorithm that was optimized and trained using both short and long sequences. IUPred identifies Intrinsically Disordered Protein Regions (IDPRs, i.e. regions that lack a stable monomeric structure under native conditions) based on a biophysics-based model. Among the three IUPred-based algorithms, we performed separate predictions with IUPred 'long disorder' and IUPred 'short disorder'. For both PONDR and IUPred, the output consisted of average disorder score (in the range 0-1) (**Fig. 2A-C**) for a protein sequence, and the percentage of each sequence that was predicted to be disordered (**Supplementary Fig. 1A-C**). A protein sequence was considered "Disordered" when the predicted average disorder score was greater than 0.5. The computed bootstrap confidence intervals of mean (and median) average disorder scores showed that the nORF datasets (sORFs, altORFs, RNACentral, Pseudogenes and Denovogenes) had higher mean (and median) values of disorder, compared to known proteins in NextProt (**Fig. 2D**). Using Fisher's exact test and Chi-square test, we also concluded that each of the nORF groups (except Denovogenes) was enriched for disordered sequences, in comparison to proteins in NextProt (**Fig. 2E**). All statistical tests were corrected for multiple hypothesis testing, using FDR values computed by the Benjamini Hochberg method.

Although the average disorder score and the percentage disorder scores are higher in nORFs, they are not drastically high to completely negate or rule out the structure forming capability of

novel proteins. Supplementary Fig. 1D shows the final number of amino sequences in each novel protein category used in the above analysis.

Some disordered regions show increased propensity to undergo disorder-to-order transitions upon binding to substrates. We used the Anchor program (*http://anchor.enzim.hu*) [27] to investigate whether the novel proteins show increased propensity to form structures. Matlab-based scripts were written to automate and batch process sequences for both Anchor program and above disorder prediction tools. The scripts were also used to parse the resulting webpages and extract useful summary statistics such as percentage of disorder and prediction score for each sequences. The results of this analysis indicate that novel proteins, except for Denovogenes, show increased disorder-to-order transition anchor scores compared to Next Prot proteins (**Figure 3A**), but we also found a strong positive correlation between average anchor score and average disorder score for most data sets, which is to be expected, since the prediction of binding sites uses biophysical parameters similar to those involved in disorder prediction. (**Figure 3B**) .

Because we have not been able to ascertain whether the novel proteins can indeed form structures or not, we investigated whether they may be biologically regulated by biochemical mechanisms such as Post Translational Modifications (PTM). Previous evidence suggested that novel proteins are indeed biologically regulated [1] and they may be enriched for regulatory sites but the observation was not done at a global scale. Hence to predict possible PTM sites in amino acid sequences from all novel proteins, we used the ModPred stand-alone software [28] . For each sequence, we counted the number of predicted sites for nine PTMs - Phosphorylation,

Acetylation, Methylation, Sulfation, SUMOylation, Ubiquitination, C-linked, O-linked and N-linked glycosylation (**Figure 4**)

To compare if each of the datasets have higher or lower predicted PTM sites than expected at random, we generated 500,000 random protein sequences of 100 amino acids each, from a uniform distribution of 20 amino acid propensities. ModPred was used to predict PTM sites in these random sequences for the same list of nine modifications. The number of predicted PTM sites were normalized to sequence length in all the other datasets. We compared the distribution of predicted PTMs in each of the nORF datasets against the predicted PTM distribution from the dataset of random sequences. Methylation, O-linked glycosylation, and Phosphorylation were found to be significantly enriched in both novel proteins and known proteins, compared to the random dataset ($p < 0.01$, Wilcoxon rank sum test). Overall, for most PTM types, the densities of predicted PTMs was comparable, if not higher in the novel proteins, compared to the NextProt database (**Figure 4**). This indicates that the novel proteins could be subjected to any biochemical regulation just as all known proteins.

To investigate whether the novel protein regions could harbour disease-associated mutations, we mapped mutations from the Catalogue of Somatic Mutations in Cancer (COSMIC) and Human Gene Mutation Database (HGMD) databases to the novel proteins. **Figure 5A-D** shows examples of COSMIC or HGMD mutations mapped to sORFs, Denovogens, and Pseudogenes demonstrating that these regions do indeed harbour mutations. Finally, we investigated whether the pathogenicity scores of these mutations assessed as Combined Annotation Dependent Depletion (CADD) [29] and Functional Analysis through Hidden Markov Models (FATHMM) [30] scores had any correlation with disorder scores at that region of the novel proteins (amino-acid

specific disorder score and average disorder score for a 7-aa window around the mutated residue). This analysis (**Figure 6A-D**) did not reveal any correlation between low pathogenicity and higher disorder scores.

Taken together our results demonstrate that although the novel proteins have increased disorder they can be biochemically regulated by PTMs and more importantly harbour deleterious mutations that is not correlated with whether those regions are ordered or disordered, suggesting that the novel proteins could be involved in biochemical processes in cells that may be dysregulated in diseases. Hence we should investigate the functional role these novel proteins in depth.

Acknowledgements: We would like to thank Asst Prof. Lisa-Tucker Kellogg (Duke-NUS Medical School, Singapore) for kindly supporting NSJ to work on this project

Funding: SP is funded by the Cambridge-DBT lectureship

Author Contributions: NSJ performed most of the analysis, participated in writing the manuscript. NM performed the mutational analysis. KB did the Denovogenes and Pseudogenes mutational analysis. SP designed and supervised the project, analysed the data, and wrote the manuscript; Competing interests: SP is a co founders of NonExomics; and Data and materials availability: Almost all processed data is in the main text or in the supplementary materials.

References

1. Prabakaran, S. *et al.* Quantitative profiling of peptides from RNAs classified as noncoding. *Nat. Commun.* **5**, 5429 (2014).
2. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X. & Albà, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol* **2**, 890–896 (2018).
3. Miravet-Verde, S. *et al.* Unraveling the hidden universe of small proteins in bacterial genomes. *Mol. Syst. Biol.* **15**, e8290 (2019).
4. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).
5. Pertea, M. *et al.* Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* 332825 (2018). doi:10.1101/332825
6. Brunet, M. A. *et al.* OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky936
7. Zhu, Y. *et al.* Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat. Commun.* **9**, 903 (2018).
8. Olexiouk, V. & Menschaert, G. Using the sORFs.Org Database. *Curr. Protoc. Bioinformatics* e68 (2018).
9. Plaza, S., Menschaert, G. & Payre, F. In Search of Lost Small Peptides. *Annu. Rev. Cell Dev. Biol.* (2017). doi:10.1146/annurev-cellbio-100616-060516
10. Vanderperre, B. *et al.* Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* **8**, e70698 (2013).

11. Mc Lysaght, A. & Hurst, L. D. Open questions in the study of denovo genes: what, how and why.
12. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
13. Xu, J. & Zhang, J. Are Human Translated Pseudogenes Functional? *Mol. Biol. Evol.* **33**, 755–760 (2016).
14. Yeasmin, F., Yada, T. & Akimitsu, N. Micropeptides Encoded in Transcripts Previously Identified as Long Noncoding RNAs: A New Chapter in Transcriptomics and Proteomics. *Front. Genet.* **9**, 144 (2018).
15. Smith, J. E. *et al.* Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Rep.* **7**, 1858–1866 (2014).
16. Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8**, 1365–1379 (2014).
17. Stein, C. S. *et al.* Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Rep.* **23**, 3710–3720.e8 (2018).
18. Cohen, S. M. Everything old is new again: (linc)RNAs make proteins! *EMBO J.* **33**, 937–938 (2014).
19. Steward, C. A. *et al.* Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med.* **9**, 49 (2017).
20. Jungreis, I. *et al.* Nearly all new protein-coding predictions in the CHESS database are not protein-coding. *bioRxiv* 360602 (2018). doi:10.1101/360602
21. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
22. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**, 117 (2013).

23. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).

24. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.* **43**, D764–70 (2015).

25. The RNAcentral Constortium. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky1034

26. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).

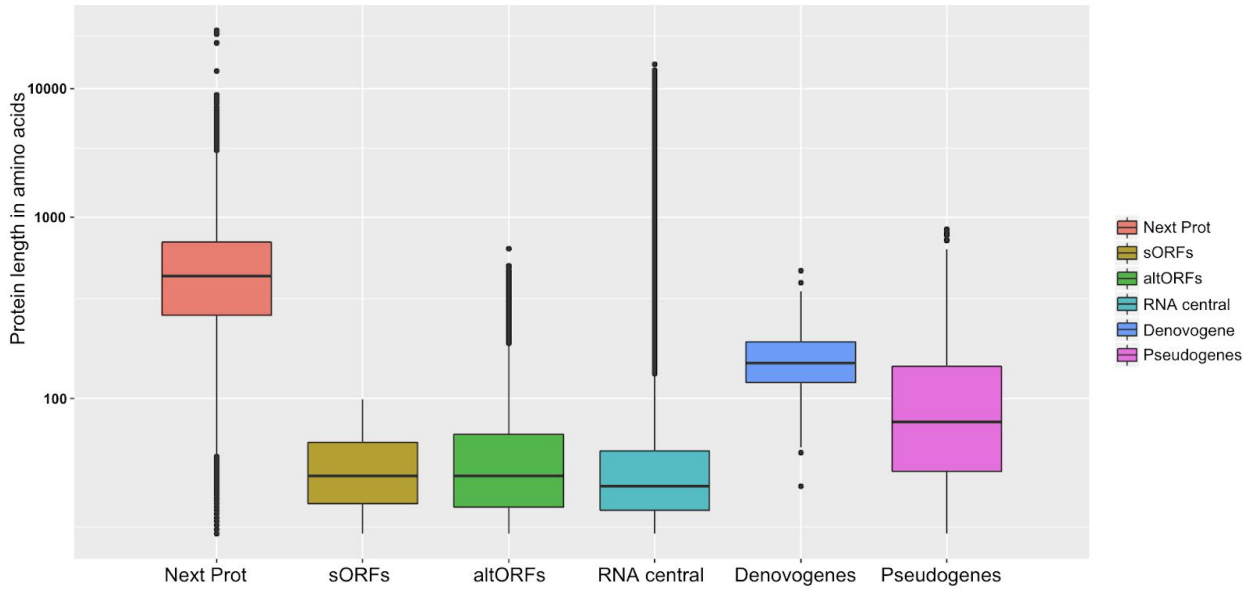
27. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746 (2009).

28. Pejaver, V. *et al.* The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.* **23**, 1077–1093 (2014).

29. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky1016

30. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).

Figure 1. Amino acid length distribution of novel proteins encoded by nORFs



Amino acid sequences of proteins from NeXtProt, and potential proteins encoded by sORFs, altORFs, Pseudogenes, Denovogenes along with all possible translated amino acids sequences from RNA central were retrieved as described in the in main text and the distribution of their lengths are plotted. It can be seen that the nORFs on average have shorter sequences, compared to the structured and known proteins in the NeXTProt database.

Figure 2. Average disorder scores of proteins encoded by novel ORFs compared to proteins in NeXtProt.

Figure 2A

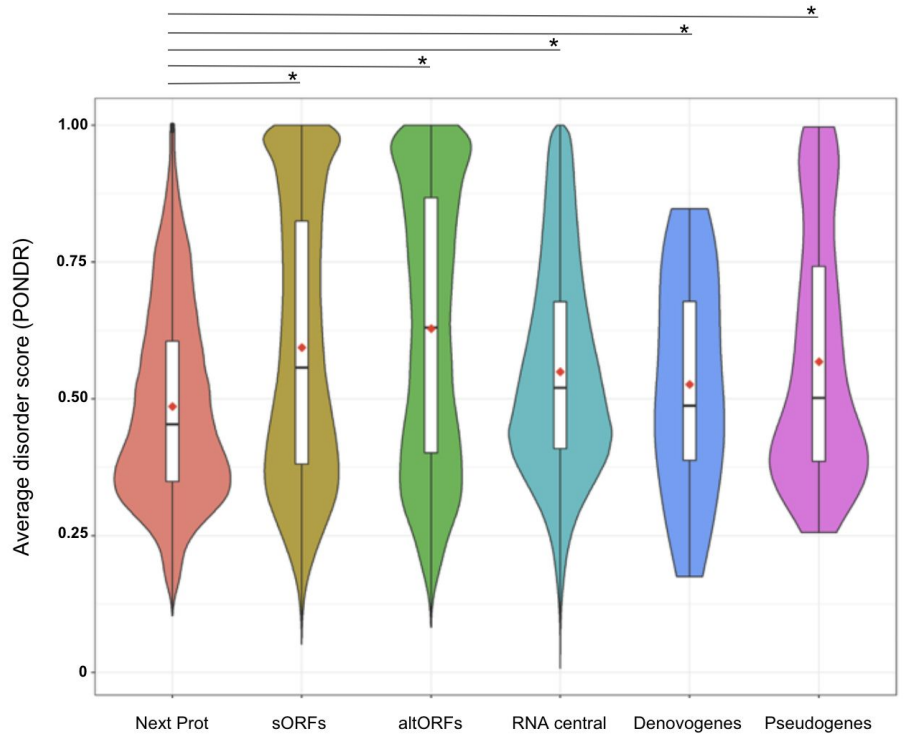


Figure 2B

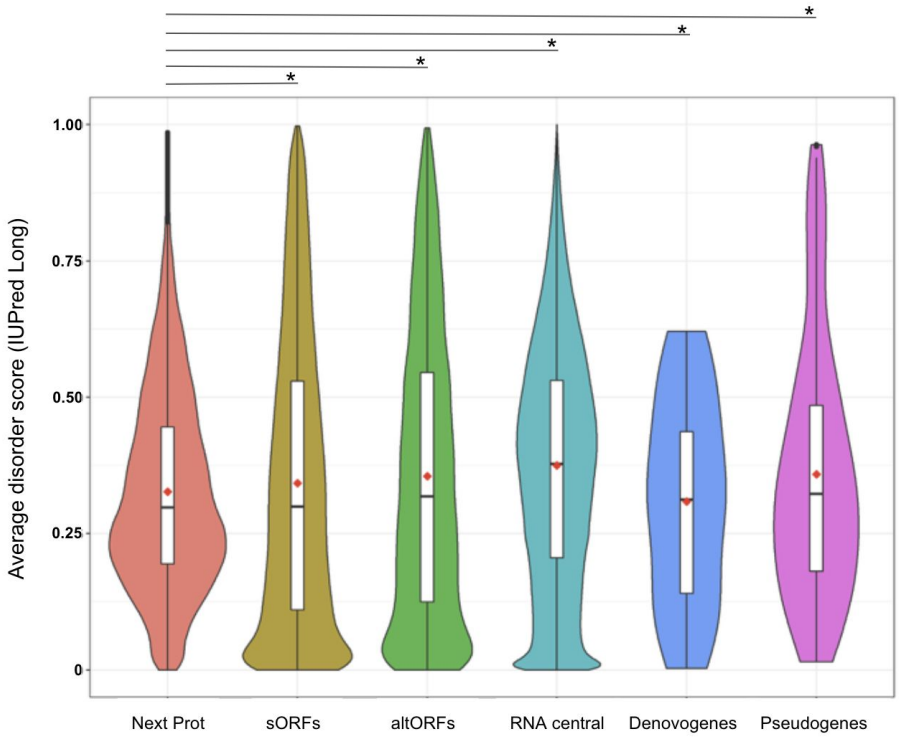


Figure 2C

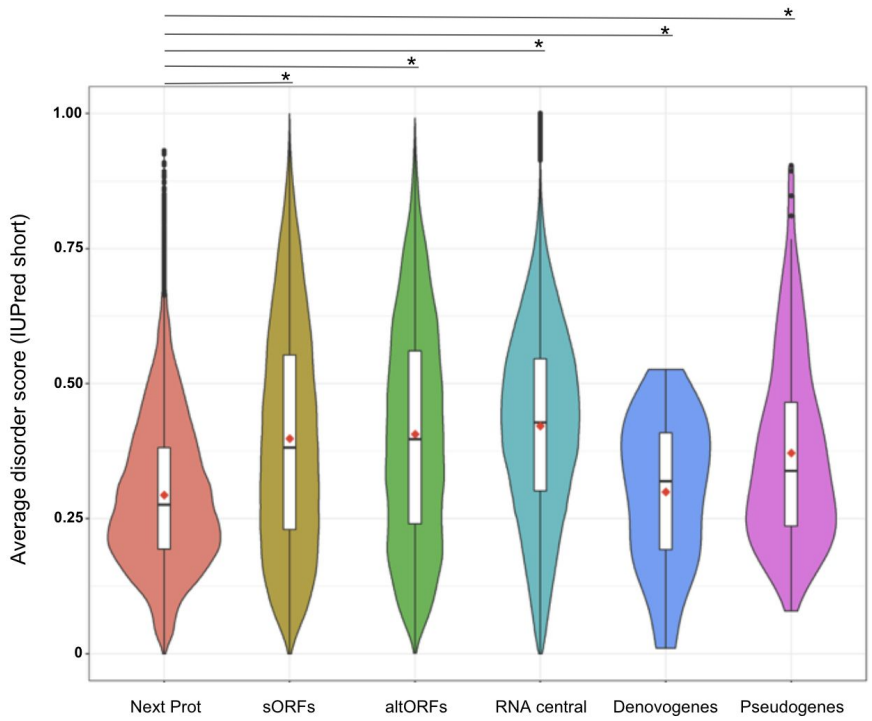


Figure 2D

	Pondr		IUPred-Long		IUPred-Short	
	Mean	Median	Mean	Median	Mean	Median
NextProt	[0.486,0.489]	[0.452,0.457]	[0.326,0.329]	[0.295,0.300]	[0.293,0.296]	[0.274,0.279]
sORFs	[0.593,0.596]	[0.554,0.560]	[0.341,0.344]	[0.297,0.302]	[0.397,0.399]	[0.380,0.384]
altORFs	[0.626,0.632]	[0.622,0.639]	[0.352,0.359]	[0.311,0.326]	[0.404,0.41]	[0.392,0.402]
RNA Central	[0.549,0.550]	[0.520,0.520]	[0.375,0.375]	[0.378,0.378]	[0.422,0.422]	[0.428,0.428]
Pseudogenes	[0.535,0.601]	[0.44,0.564]	[0.325,0.393]	[0.273,0.361]	[0.346,0.396]	[0.310,0.370]
Denovo genes	[0.457,0.601]	[0.440,0.638]	[0.240,0.377]	[0.175,0.421]	[0.249,0.352]	[0.208,0.396]

Figure 2E

	p-value - Pondr		p-value - IU L		p-value IU-S	
Nextprot vs	Fisher test	Chi-square	Fisher test	Chi-square	Fisher test	Chi-square
sORF	0.000	0.000	0.000	0.000	0.000	0.000
altORFs	0.000	0.000	0.000	0.000	0.000	0.000
RNA Central	0.000	0.000	0.000	0.000	0.000	0.000
Pseudogenes	0.013	0.021	0.062	0.000	0.000	0.000
Denovo genes	0.400	0.658	0.560	0.963	0.897	0.405

Average disorder scores of proteins in NeXtProt compared to average disorder scores of proteins encoded by novel ORFs - sORFs, altORFs, Pseudogene, Denovogenes and all possible translated amino acids sequences from RNA central - predicted by **A.** PONDR, **B**. IUPred Long, and **C.** IUPred short disorder predictors. **D.** 95% Bootstrap confidence Interval of the mean and median of disorder scores predicted by PONDR, IUPred Long and IUPred short for the Next Prot, sORFs, altORFs, Pseudogenes, Denovogenes and all possible translated amino acid sequences from RNA Central. **E.** Statistical significance of enrichment of disordered sequences in sORFs, altORFs, Pseudogenes, Denovogenes, and all possible translated amino acid sequences from RNA central, in comparison to the NeXtProt dataset.

Figure 3. Prediction of potential disorder to order transition upon binding regions of known proteins and proteins encoded by novel ORFs

Figure 3A

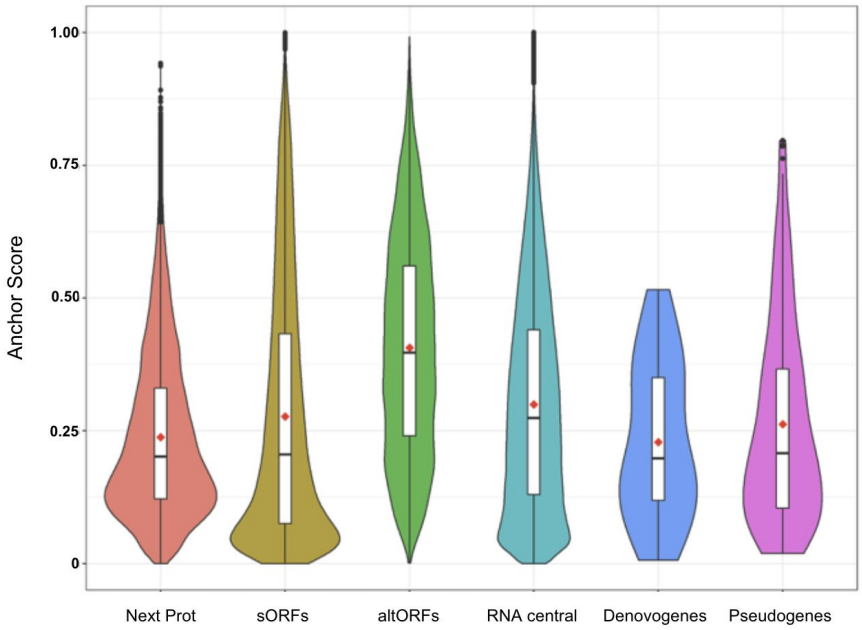
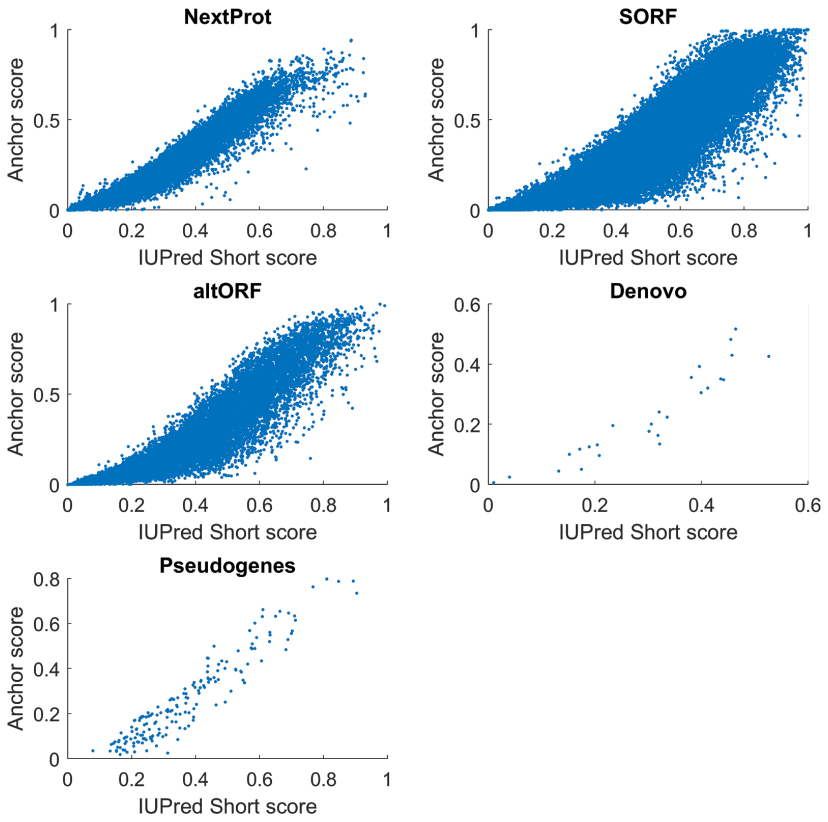
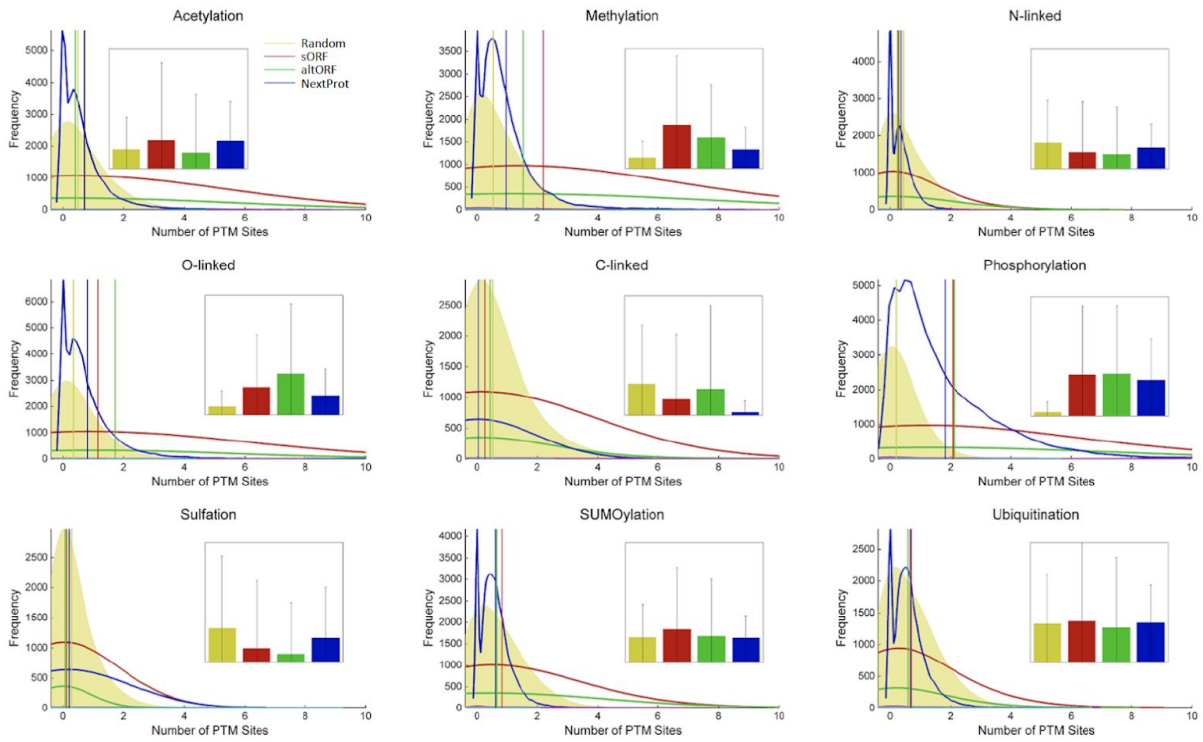


Figure 3B



Anchor predictions of disordered binding regions in the NeXTProt database and the novel ORFs. **A.** Anchor score represents the average propensity of an amino acid in a disordered region to be part of a protein-protein binding site. As expected, the novel ORFs have higher mean Anchor scores in comparison to the NeXTProt database. **B.** Observed correlation between average Anchor score and IUPred-Short disorder prediction score for individual datasets.

Figure 4. Prediction of PTMs in disordered regions of known proteins and proteins from novel ORFs



Predicted densities of nine PTM modifications for sORFs, altORFs, and NeXtProt sequences were compared with predicted densities of a pseudo-dataset composed of 500,000 randomly generated amino acid sequences. In each panel, the shaded yellow histogram represents the predicted distribution of number of PTM sites in the random dataset. The curves (red and green, blue) represent the predicted density distributions for sORF, altORF, and NeXtProt respectively. The vertical lines show the mean of each distribution. For many cases (e.g., Methylation), it can be clearly seen that the average number of predicted PTMs is higher in all three datasets compared to the random set (all three lines are to the right of yellow). Shown in insets are bar plots of the same data to show the difference in the distributions of the random data set (yellow bar) vs sORF (red), altORF (green) or NeXtProt (blue).

Figure 5. Mutations from COSMIC and HGMD were mapped to nORFs.

Figure 5A

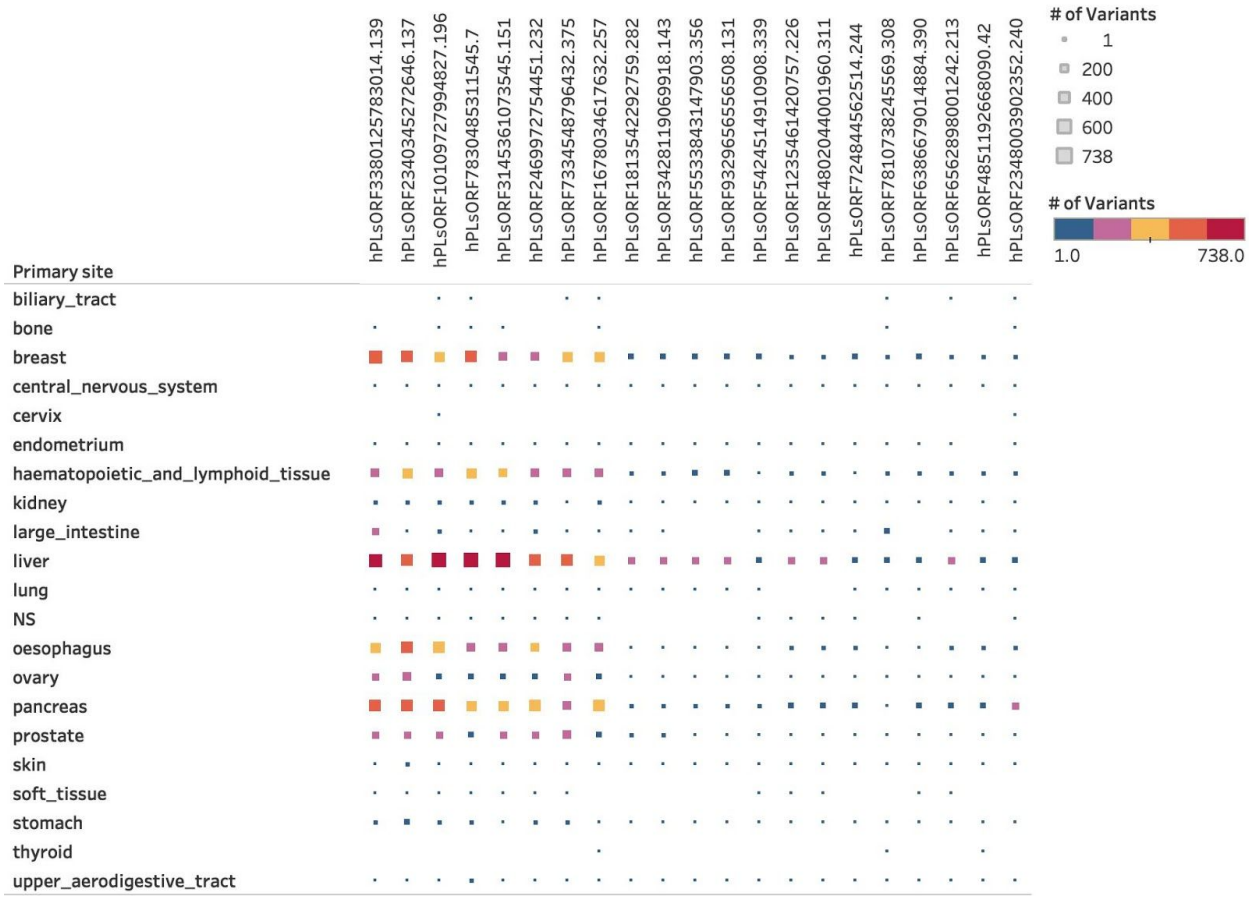


Figure 5B

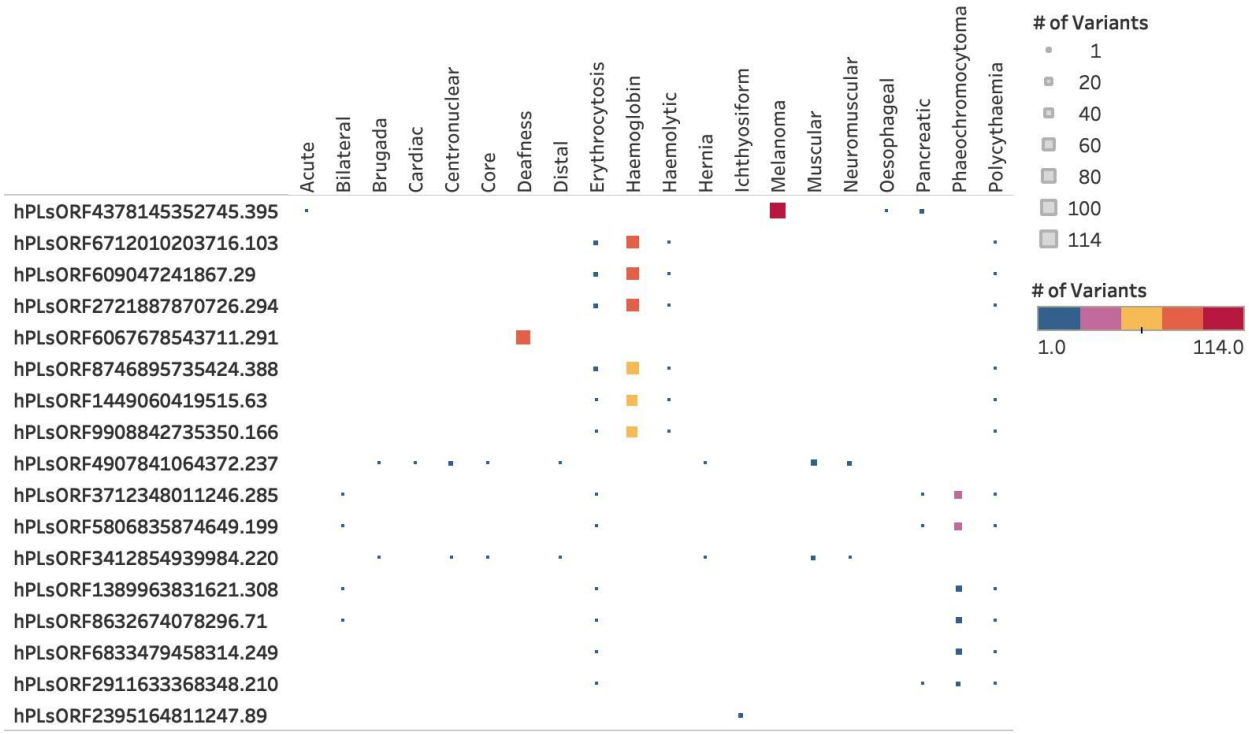

Acute
Bilateral
Brugada
Cardiac
Centronuclear
Core
Deafness
Distal
Erythrocytosis
Haemoglobin
Haemolytic
Hernia
Ichthyosiform
Melanoma
Muscular
Neuromuscular
Oesophageal
Pancreatic
Phaeochromocytoma
Polycythaemia
hPLsORF4378145352745.395
hPLsORF6712010203716.103
hPLsORF609047241867.29
hPLsORF2721887870726.294
hPLsORF6067678543711.291
hPLsORF8746895735424.388
hPLsORF1449060419515.63
hPLsORF9908842735350.166
hPLsORF4907841064372.237
hPLsORF3712348011246.285
hPLsORF5806835874649.199
hPLsORF3412854939984.220
hPLsORF1389963831621.308
hPLsORF8632674078296.71
hPLsORF6833479458314.249
hPLsORF2911633368348.210
hPLsORF2395164811247.89
# of Variants
1
20
40
60
80
100
114
# of Variants
1.0
114.0

Figure 5C

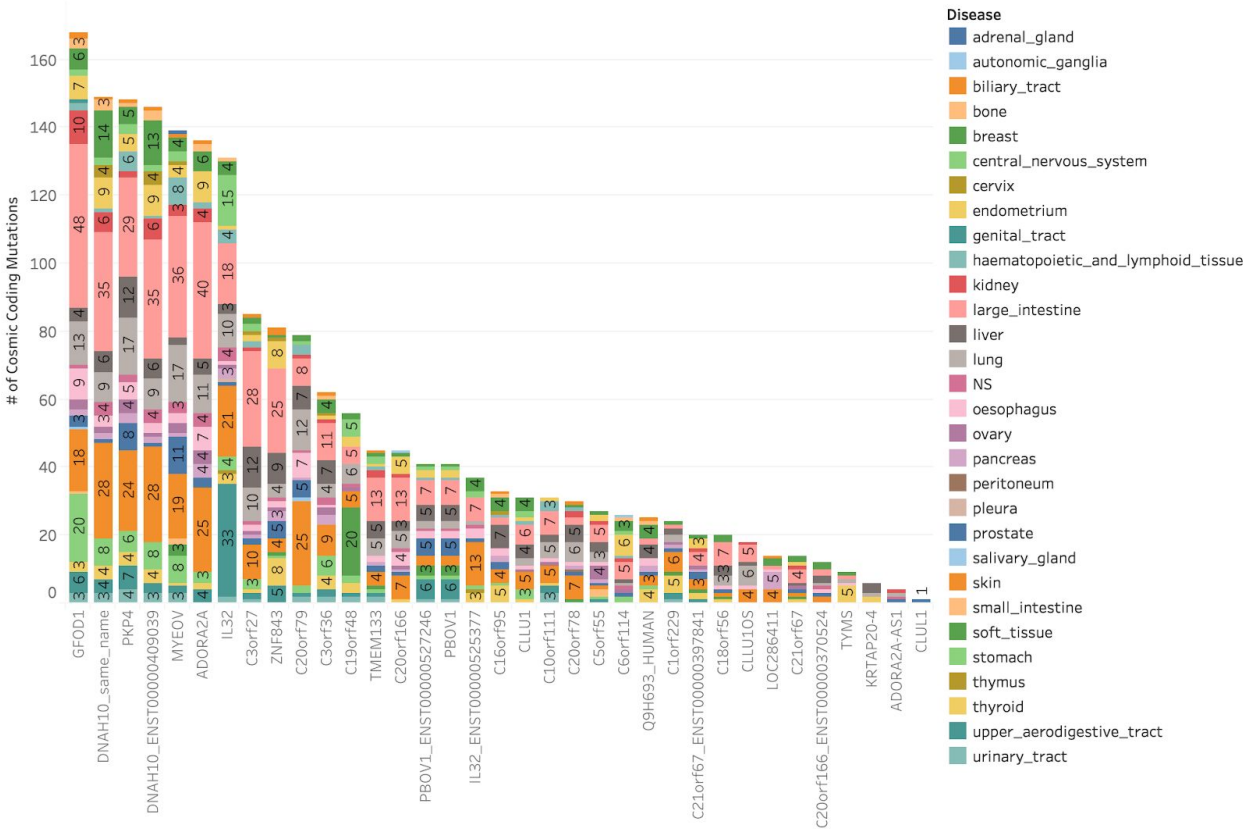
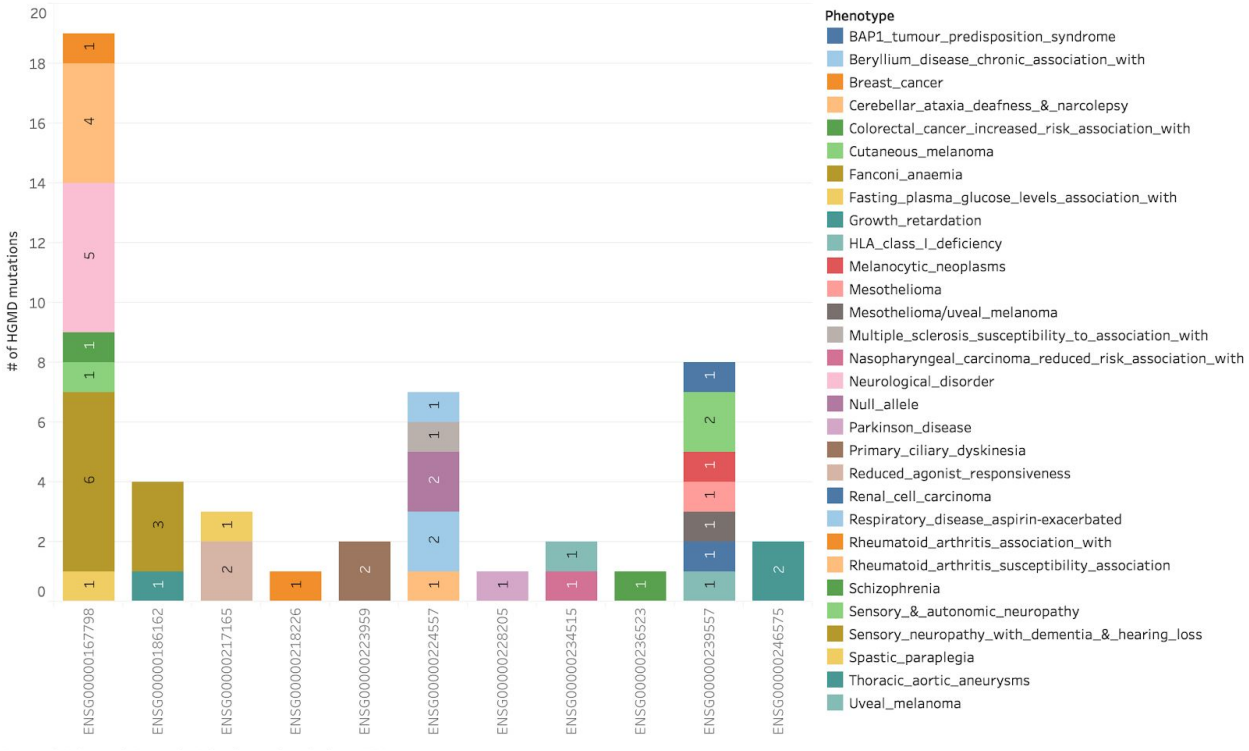

Disease
adrenal_gland
autonomic_ganglia
biliary_tract
bone
breast
central_nervous_system
cervix
endometrium
genital_tract
haematopoietic_and_lymphoid_tissue
kidney
large_intestine
liver
lung
NS
oesophagus
ovary
pancreas
peritoneum
pleura
prostate
salivary_gland
skin
small_intestine
soft_tissue
stomach
thymus
thyroid
upper_aerodigestive_tract
urinary_tract
# of Cosmic Coding Mutations
0
20
40
60
80
100
120
140
160
GFOD1
DNAH10_same_name
PKP4
DNAH10_ENST00000409039
MYEOV
ADORA2A
IL32
C3orf27
ZNF843
C20orf79
C3orf36
C19orf48
TMEM133
C20orf166
PBOV1_ENST00000527246
PBOV1
IL32_ENST00000525377
C16orf95
CLLU1
C10orf111
C20orf78
C5orf55
C6orf114
Q9H693_HUMAN
C1orf229
C21orf67_ENST00000397841
C18orf56
CLLU1OS
LOC286411
C21orf67
C20orf166_ENST00000370524
TYMS
KRTAP20-4
ADORA2A-AS1
CLUL1

Figure 5D



Mutations from COSMIC and HGMD databases were mapped to both entire nORFs genomic and specifically to novel protein amino acid sequence coordinates and represented as described below. **A.** Noncoding mutations from COSMIC were mapped to all sORFs genomic coordinates and only the top 21 sORFs with the highest number of mutations are represented here. Size of the squares against each sORF and each disease indicates the number of mutations and the color indicates the number of pathogenic mutations in that region. **B**. Mutations from HGMD database were mapped to sORFs and only the top 17 mutations are represented here. **C.** Coding mutations from COSMIC are mapped to all Denovogenes and are represented. **D.** Mutations from HGMD are mapped to pseudogenes that are known to be translated and presented.

Figure 6. Correlation plots of CADD and FATHMM pathogenicity scores with disorder scores of nORFs

Figure 6A

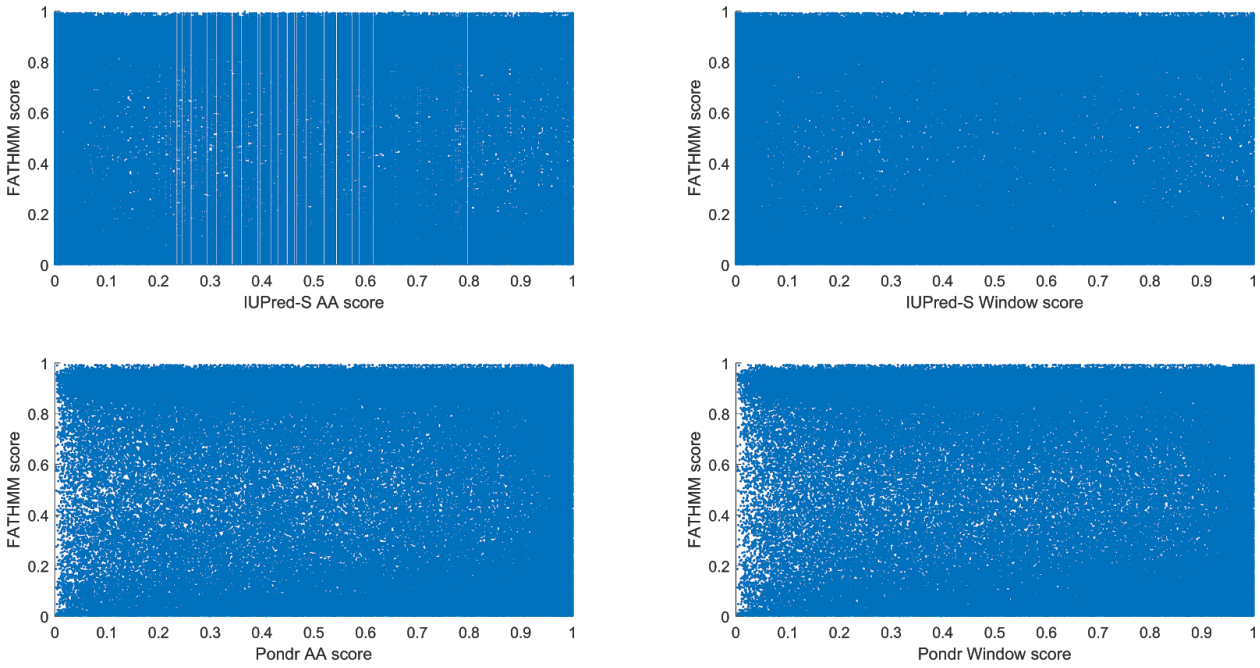


Figure 6B

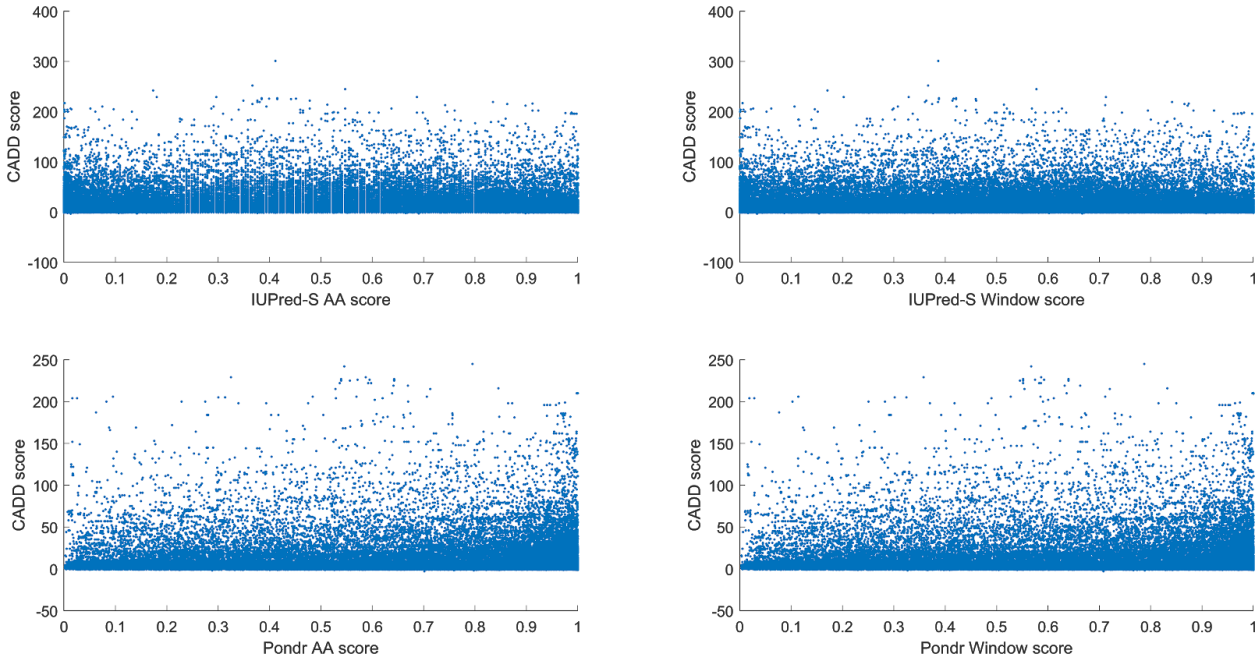


Figure 6C

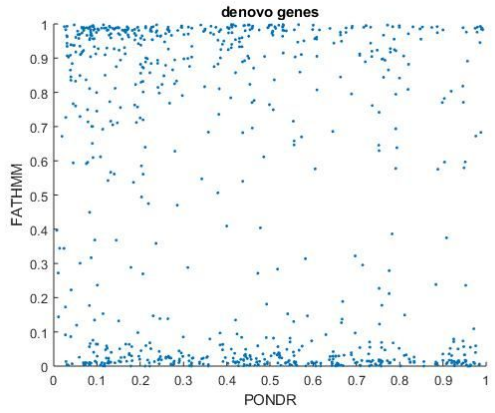
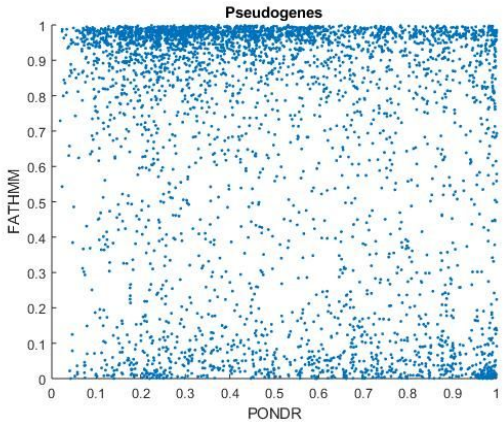


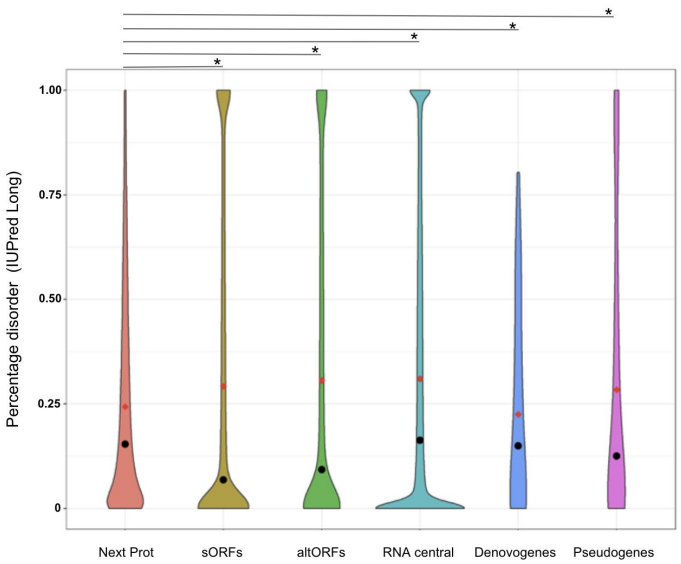
Figure 6D



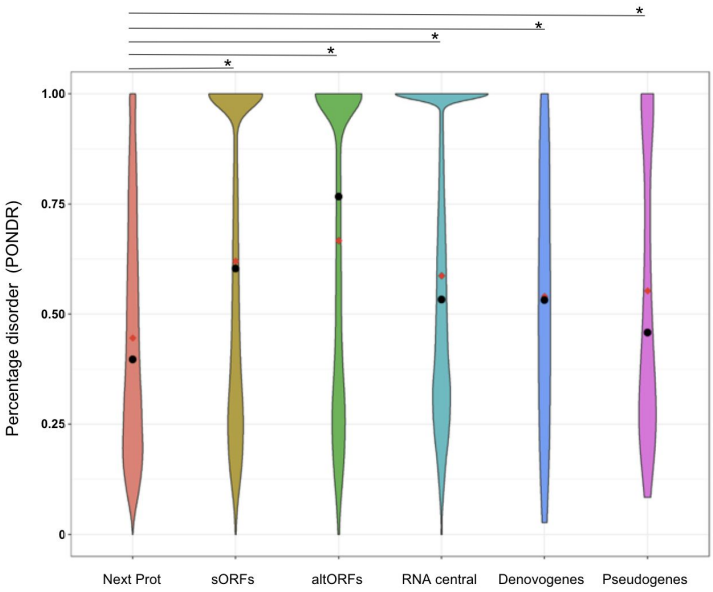
Correlation analysis of CADD or FATHMM pathogenicity scores for mutations in the amino acid sequence of the novel proteins and their corresponding disorder scores predicted with both PONDR and IUPred were performed. The analysis did not reveal any correlations between **A.** FATHMM scores and disorder scores for sORFs, **B.** CADD scores and disorder scores for sORFs, **C.** FATHMM scores and disorder scores for Denovogenes, or **D.** FATHMM scores and disorder scores for Pseudogenes.

Supplementary Figure

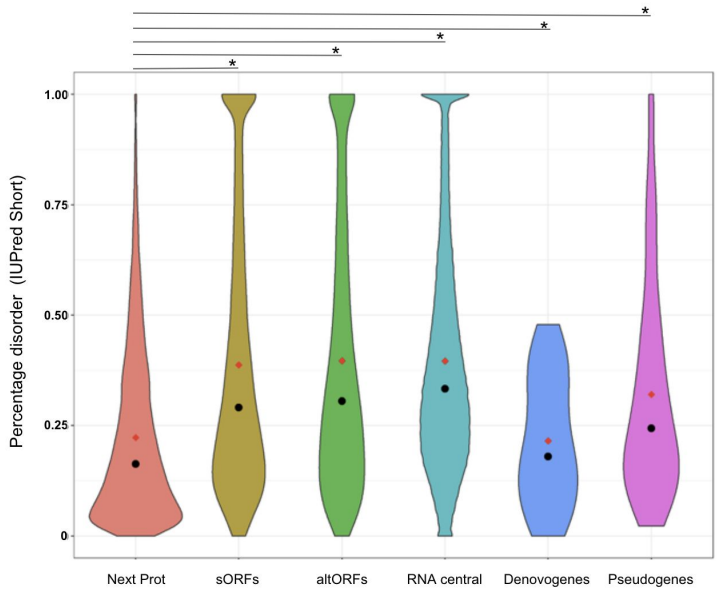
Supplementary Figure 1A



Supplementary Figure 1B



Supplementary Figure 1C



Supplementary Fig. 1D

Protein sequences	Overall	Num > 30aa	Pondr	IUPred Long	IUPred Short
NextProt	42024	41952	17449	7589	3363
sORF	190195	92176	52213	25537	29543
altORF	24676	19281	12037	5725	6523
RNACentral	5185186	5185186	2807585	1534132	1778105
Pseudogenes	172	172	87	41	36
Denovo genes	26	26	12	5	1

A-C. Percentage of protein sequence identified to be disordered (amino-acid disorder score > 0.5) in each of the nORF datasets, for three prediction algorithms. Shown in red and black within each distribution are the mean and median respectively. **D.** Table showing the number of protein sequences identified in each nORF dataset, number of sequences used for further analysis (Sequence length > 30) and the number of predicted disordered sequences (average disorder score > 0.5) from PONDR, IUPred-Long and IUPred-short algorithms.