Proteins encoded by Novel ORFs have increased disorder but can be biochemically regulated and harbour deleterious 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 advances in proteogenomics indicate 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]. Although it is yet unclear if these protein or protein-like products could possess any significant biological function hopes have been raised to target them for anticancer and antimicrobial therapy [3,4]. In this study, we used computational tools to systematically investigate these novel protein sequences for their propensities toward structural disorder, post-translational modifications (PTM) and mutational densities. We found that these novel proteins have significantly higher disorder and similar PTM frequencies compared to known proteins. Although these regions were found to harbour deleterious mutations, we did not observe any correlation between the pathogenicity of mutations and their location (ordered/disordered) within these novel proteins. This study suggests that these nORFs encode an important class of proteins, that could undergo sequence, structural or regulatory changes during complex diseases, and hence warrant further study.

Significance:

Certain noncoding regions of the genome are known to make protein-like products. These protein-like products have significantly expanded the number of the cellular proteome but we dont know whether they are capable of performing biological functions. Here in this study, we have investigated all such putative protein-like products and analyzed whether they can form structures, whether hey harbour mutations, and whether they can be regulated by biochemical regulatory processes. Results from this study indicate that indeed these protein-like proteins can perform and be involved in all the above processes.

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, Denovogenes[11] or Orphan genes [12], Pseudogenes [1,13], and many ncRNAs have been shown to have coding potential [14–18]. These new discoveries challenge traditionally held conservative definitions of ORF, as used until the recent 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], while on the other side some have proposed that such nORFs confer 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 the sORF database (http://sorfs.org/database) [8], sequences for altORFs from Roucou's lab [10] (http://haltorf.roucoulab.com/), and 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 protein sequences whose translation has been experimentally verified in literature, the downloaded RNA central dataset contained 9,386,637 nucleotide transcript sequences. We identified potential ORFs from these transcripts, using the following workflow. 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 (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. Since the size of the RNACentral dataset far exceeded that of the four other novel datasets, we decided to keep the datasets segregated for future analysis. From here on we will call the amino acid sequences encoded by the nORFs as novel proteins. **Figure 1** shows the protein length distributions for each of the studied datasets. We found that the novel proteins are shorter in length compared to known proteins. Next, we investigated the structure-forming capabilities of these novel proteins, since shorter proteins are known to form elementary structures [26]. We employed two disorder prediction algorithms, PONDR (*http://www.pondr.com*) and IUPred (*https://iupred2a.elte.hu/*), to assess whether these novel proteins are predominantly ordered or disordered, which would directly correlate with their ability to 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 protein 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'. Matlab-based scripts were written to automate and batch process protein sequences for disorder prediction, and parse the output. 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**). We also checked to see if these novel protein datasets were enriched for disordered proteins using Fisher's exact test and the Chi-square test. Our tests showed that each of the nORF groups (except Denovogenes) was enriched for disordered sequences in comparison to proteins in NeXtProt (**Fig. 2E**). Supplementary Fig. 1D shows the final number of amino sequences in each novel protein category used in the above analysis. 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. Some disordered regions have been known 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. The results of this analysis indicate that novel proteins, except for Denovogenes, show increased anchor scores compared to NeXtProt proteins (**Figure 3A**). However, we also found a strong positive correlation between average anchor score and average disorder score for most data sets, which is not surprising, since the prediction of binding sites uses biophysical parameters similar to those involved in disorder prediction (**Figure 3B**) .

Given that the novel proteins were found to be enriched for disorder, we investigated whether they might 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, we predicted possible PTM sites in the amino acid sequences from all novel proteins, using the ModPred stand-alone software [28] . For each sequence, we predicted amino acid 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 site-densities 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 then used to predict PTM sites in these random sequences for the same list of nine modifications. The number of predicted PTM sites in all datasets were normalized to account for variable sequence length . We compared the distribution of predicted PTMs in each of the nORF datasets, and

NeXtProt against the predicted PTM distributions corresponding to the dataset of random sequences. Methylation, O-linked glycosylation, and Phosphorylation were found to be significantly enriched in both novel proteins and known NeXtProt proteins, compared to the random dataset ($p < 0.0001$, Wilcoxon rank sum test). Overall, for most PTM types, the densities of predicted PTMs was comparable, if not higher, in the novel proteins versus the NeXtProt database (**Figure 4**). This indicates that the novel proteins could be subjected to any biochemical regulation just as much 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 the mutated region of the novel proteins (both amino-acid specific disorder score, and average disorder score for a 7-aa window around the mutated residue). This analysis (**Figure 6A-D** and Supplementary Fig. 2) 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 the proportion of increase is not substantial to affect their structure forming capabilities. In addition 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 and nORFs 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

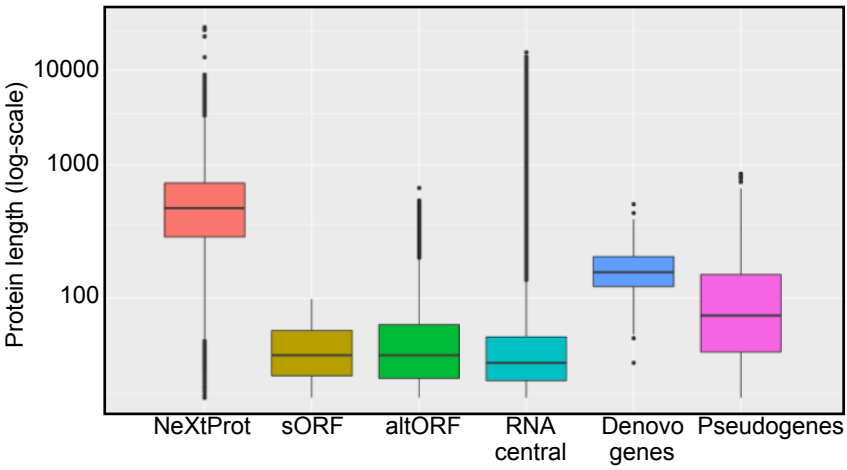
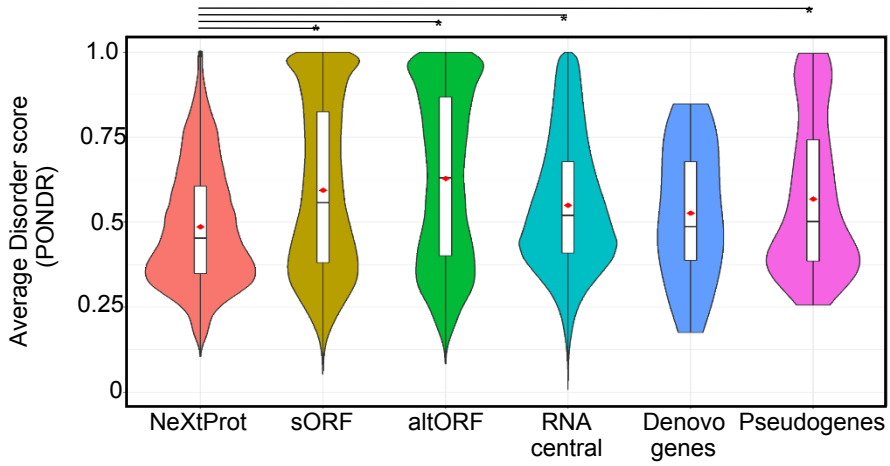


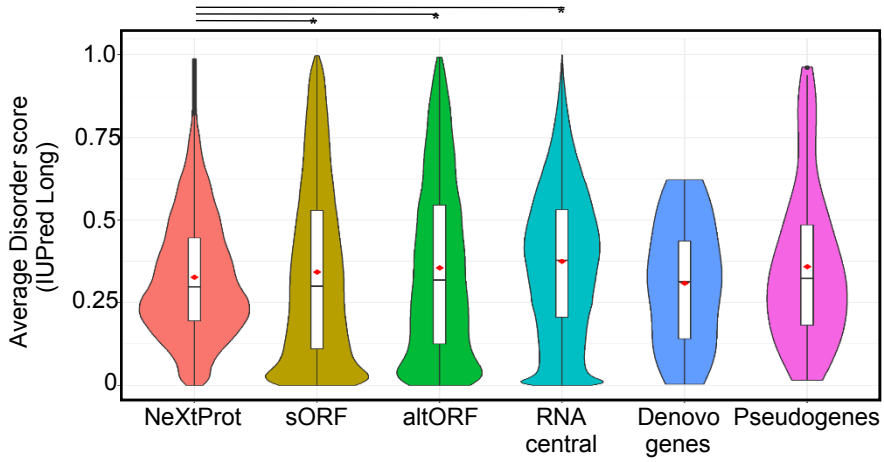
Figure 1. Amino acid length distribution of novel proteins encoded by nORFs. Amino acid sequences of known human proteins from NeXtProt, and potential proteins encoded by nORFs - sORFs, altORFs, Pseudogenes, Denovogenes, and all possible translated amino acids sequences from RNA central, were retrieved as described in the main text. Shown are the distributions of protein length in each dataset. The nORFs on average have shorter sequences compared to the structured proteins in the NeXTProt database.

Figure 2

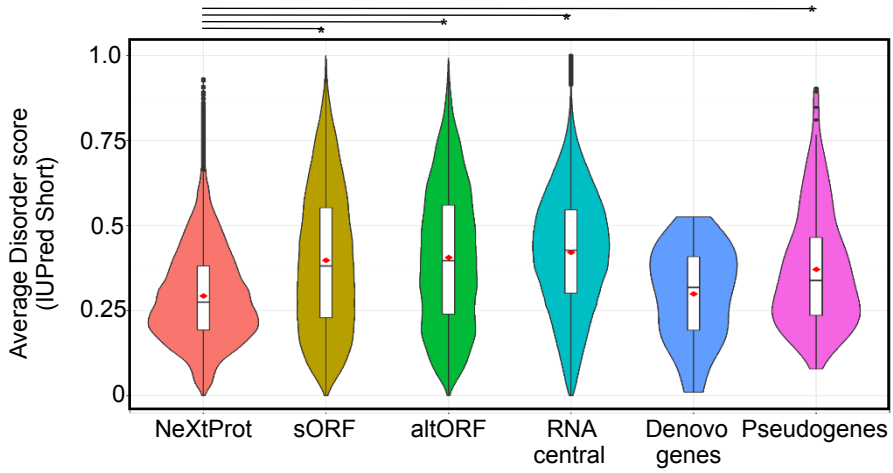
A


Average Disorder score
(PONDR)
1.0
0.75
0.5
0.25
0
NeXtProt
sORF
altORF
RNA
central
Denovo
genes
Pseudogenes

B


Average Disorder score
(IUPred Long)
1.0
0.75
0.5
0.25
0
NeXtProt
sORF
altORF
RNA
central
Denovo
genes
Pseudogenes

C


Average Disorder score
(IUPred Short)
1.0
0.75
0.5
0.25
0
NeXtProt
sORF
altORF
RNA
central
Denovo
genes
Pseudogenes

D

Dataset	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
sORF	0.593-0.596	0.554-0.560	0.341-0.344	0.297-0.302	0.397-0.399	0.380-0.384
altORF	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
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
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

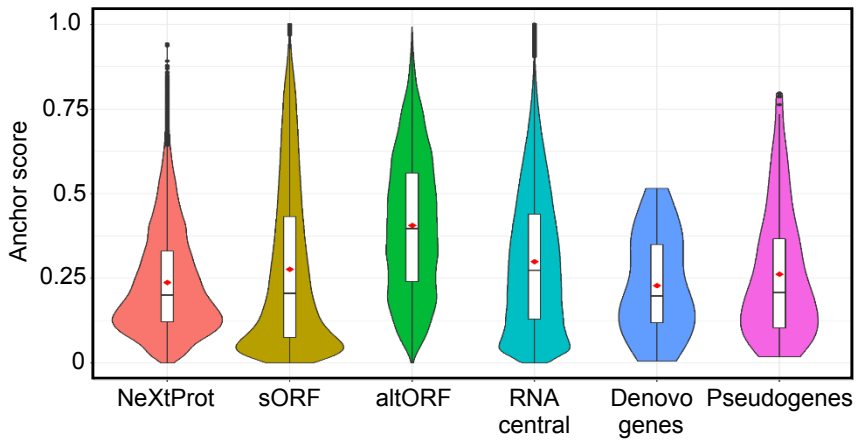
E

Disordered sequences in NeXtProt <	p-value PONDR		p-value IUPred-Long		p-value IUPred-Short	
	Fisher test	Chi-square	Fisher test	Chi-square	Fisher test	Chi-square
sORF	0.00	0.00	0.00	0.00	0.00	0.00
haltORF	0.00	0.00	0.00	0.00	0.00	0.00
RNA Central	0.00	0.00	0.00	0.00	0.00	0.00
Pseudogenes	0.01	0.02	0.06	0.00	0.00	0.00
Denovo genes	0.40	0.66	0.56	0.96	0.90	0.41

Figure 2. Prediction of disorder in proteins encoded by nORFs. Average disorder scores of proteins in NeXtProt compared to average disorder scores of proteins encoded by nORFs, 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 using PONDR, IUPred-Long and IUPred-Short for each of the nORF datasets **E.** Statistical significance (uncorrected p-values) for the enrichment of disordered sequences in nORF datasets, in comparison to NeXtProt.

Figure 3

A



B

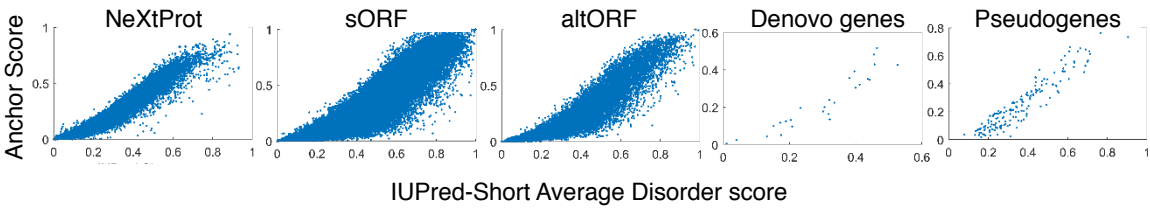


Figure 3. Anchor predictions of binding regions in proteins encoded by nORFs. Anchor scores represents the average propensity of an amino acid in a disordered region to be part of a protein-protein binding site. **A.** Shown are the distributions of the predicted Anchor scores for known proteins in NeXtProt, and those encoded by the nORF datasets. As expected, the nORF proteins 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

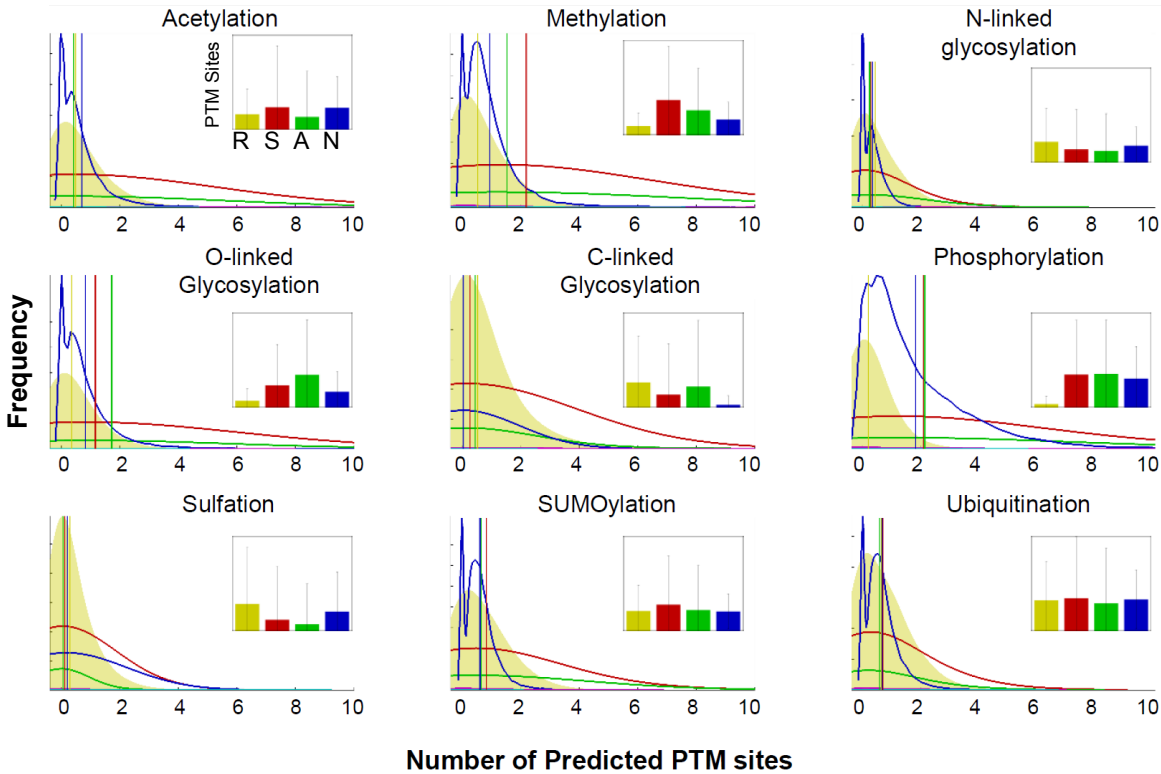


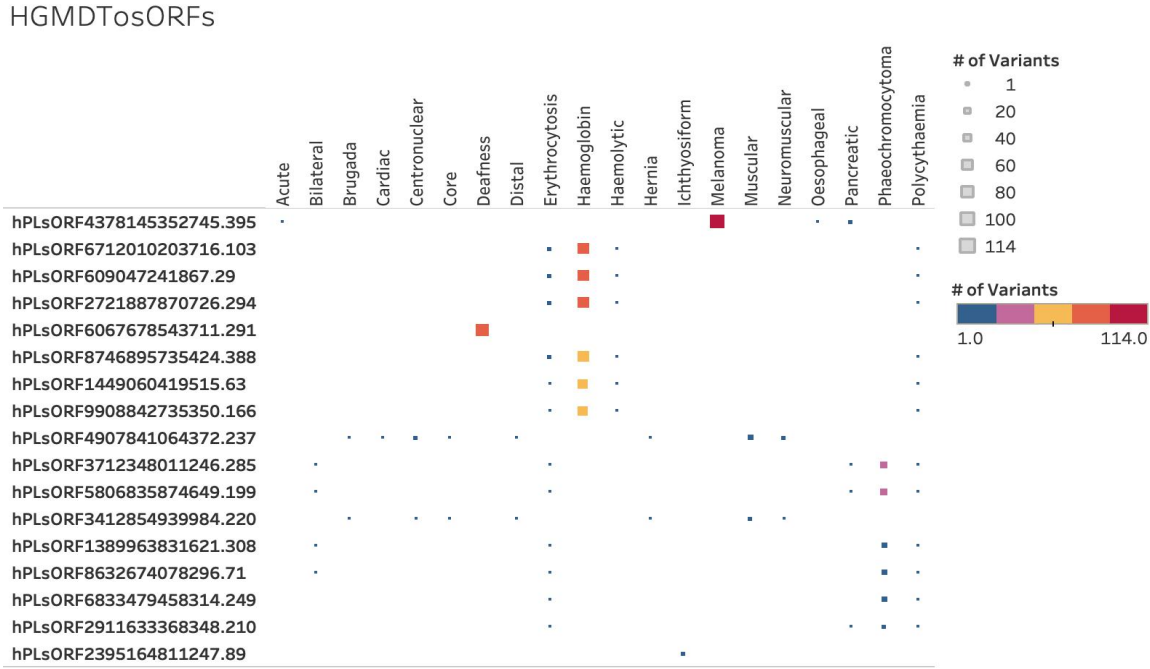
Figure 4. Predictions of PTM site-densities in the proteins encoded by nORFs. PTM sites in protein sequences were predicted using the ModPred tool. Predicted densities of nine PTM modifications for sORFs, altORFs, and NeXtProt sequences were compared with the predicted PTM densities for a pseudo-dataset composed of 500,000 randomly generated amino acid sequences (100-aa long). In each panel, the shaded yellow histogram represents the predicted distribution of the 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 overlaid 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 indicating the PTM density distributions of the random data set (R, yellow bar) vs sORF (S, red), altORF (A, green) or NeXtProt (N, blue).

Figure 5

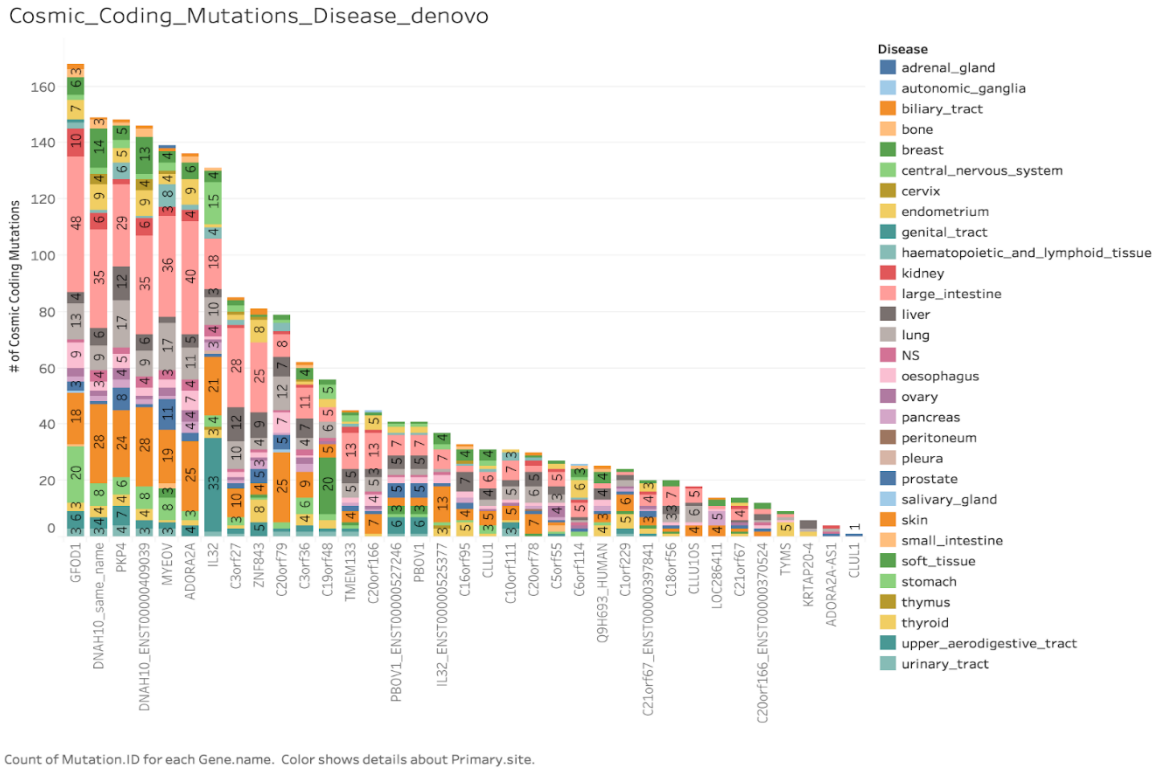
A

CosmicNonCodingTosORFs
hPLsORF3380125783014.139
hPLsORF2340345272646.137
hPLsORF1010972799482 7.196
hPLsORF7830485311545.7
hPLsORF3145361073545.151
hPLsORF2469972754451.232
hPLsORF7334548796432.375
hPLsORF1678034617632.257
hPLsORF1813542292759.282
hPLsORF3428119069918.143
hPLsORF5533843147903.356
hPLsORF9329656556508.131
hPLsORF5424514910908.339
hPLsORF1235461420757.226
hPLsORF4802044001960.311
hPLsORF7248445562514.244
hPLsORF7810738245569.308
hPLsORF6386679014884.390
hPLsORF6562898001242.213
hPLsORF4851192668090.42
hPLsORF2348003902352.240
Primary site
biliary_tract
bone
breast
central_nervous_system
cervix
endometrium
haematopoietic_and_lymphoid_tissue
kidney
large_intestine
liver
lung
NS
oesophagus
ovary
pancreas
prostate
skin
soft_tissue
stomach
thyroid
upper_aerodigestive_tract
# of Variants
1
200
400
600
738
# of Variants
1.0
738.0

B


HGMDTosORFs
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

C



D

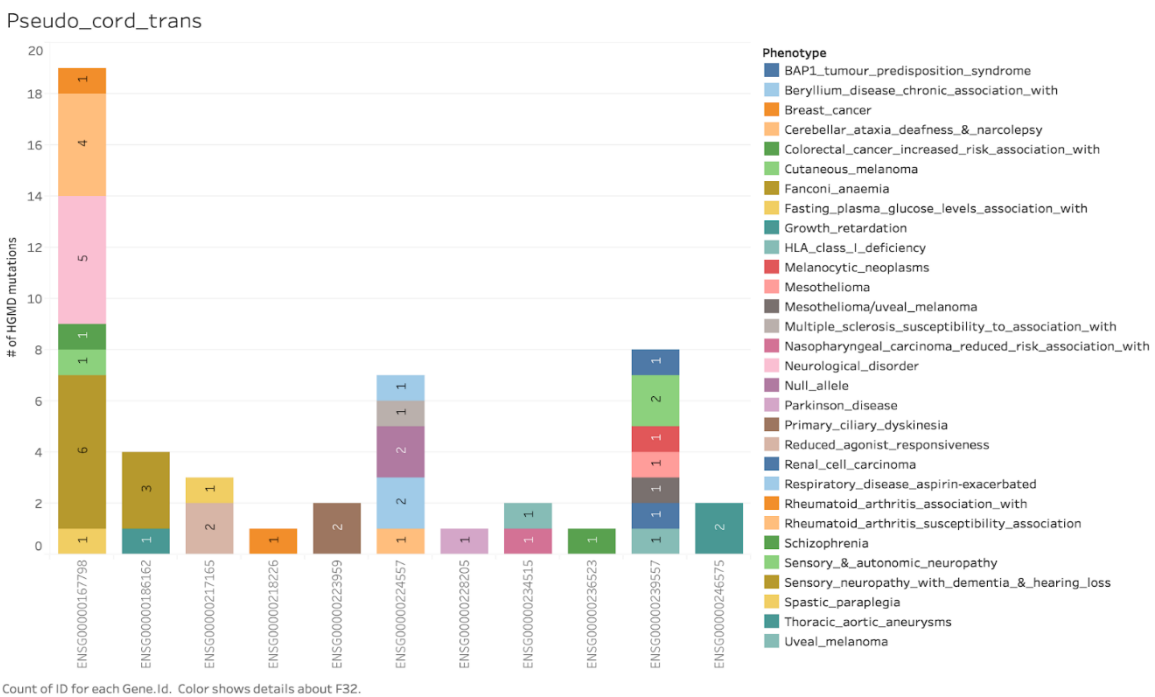


Figure 5. Known mutations in the proteins encoded by the nORFs. Mutations from COSMIC and HGMD databases were mapped to both entire nORFs genomic and specifically to novel protein amino acid sequence coordinates and represented as described below. **A.** Noncoding mutations from COSMIC were mapped to 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 Denovo genes and are represented. **D.** Mutations from HGMD are mapped to pseudogenes that are known to be translated and presented.

Figure 6

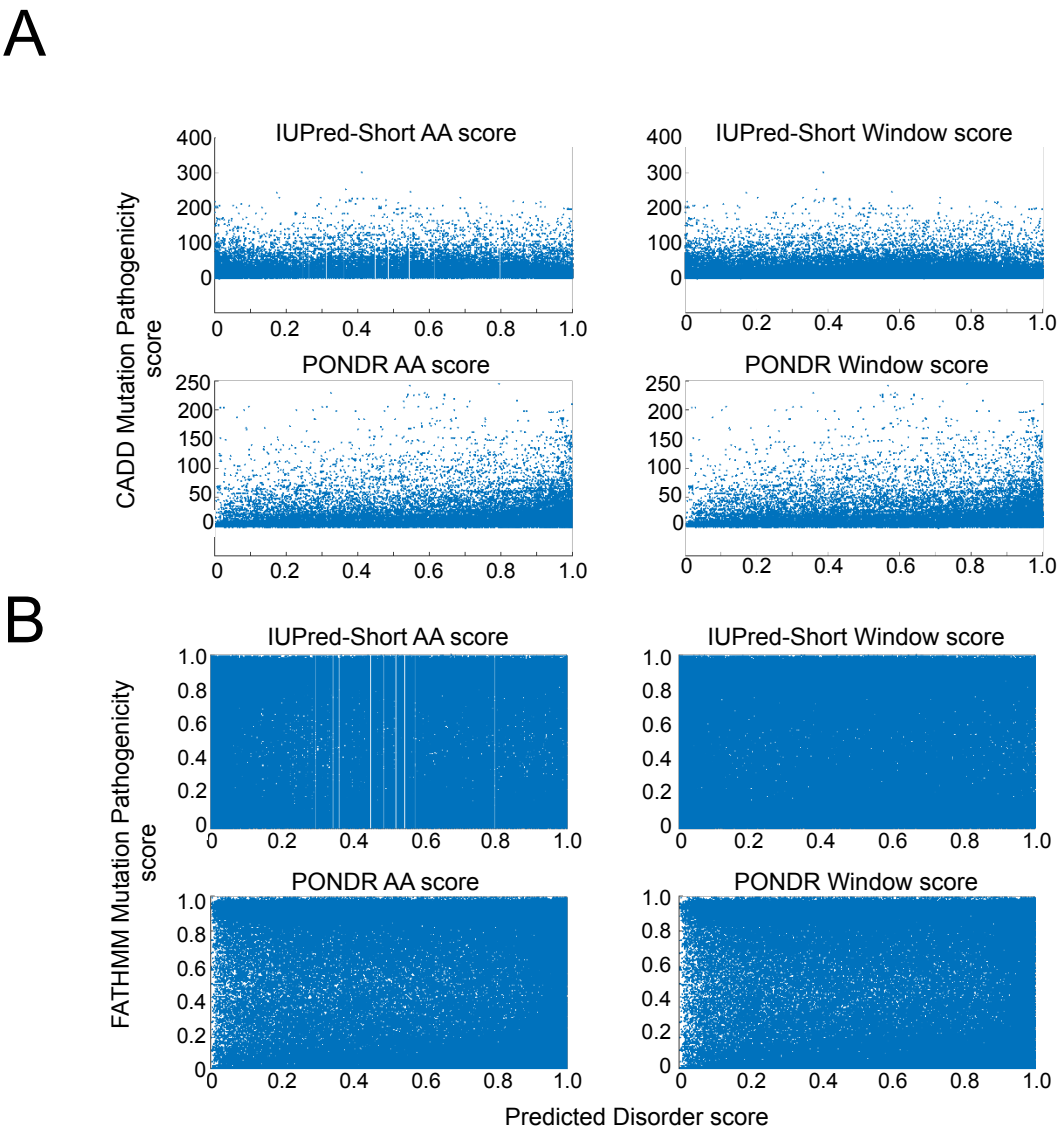
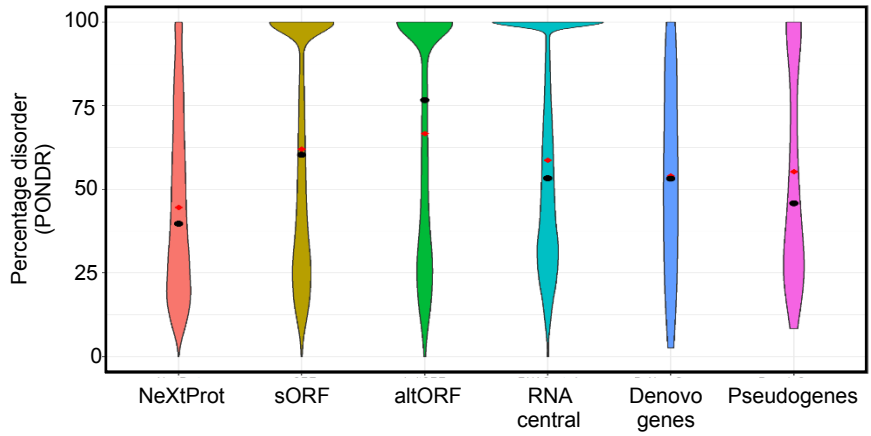


Figure 6. CADD and FATHMM pathogenicity scores vs predicted disorder scores for proteins encoded by sORFs. We plotted the CADD or FATHMM mutation pathogenicity scores for the proteins encoded by sORFs, against their corresponding disorder scores predicted using either PONDR and IUPred. Disorder scores were computed at either amino-acid resolution, or for a 7-AA window around the mutated residue. The analysis did not reveal any correlations between **A.** FATHMM scores and predicted disorder scores, or **B.** CADD scores and predicted disorder scores.

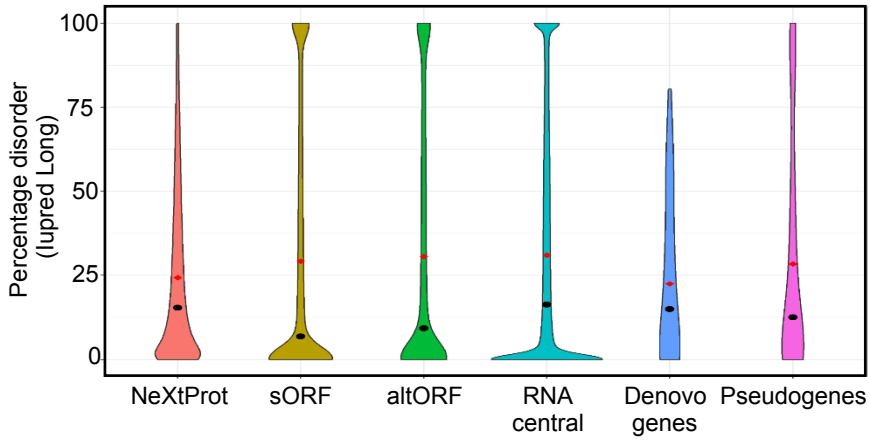
SUPPLEMENTARY INFO

Supplementary Figure 1

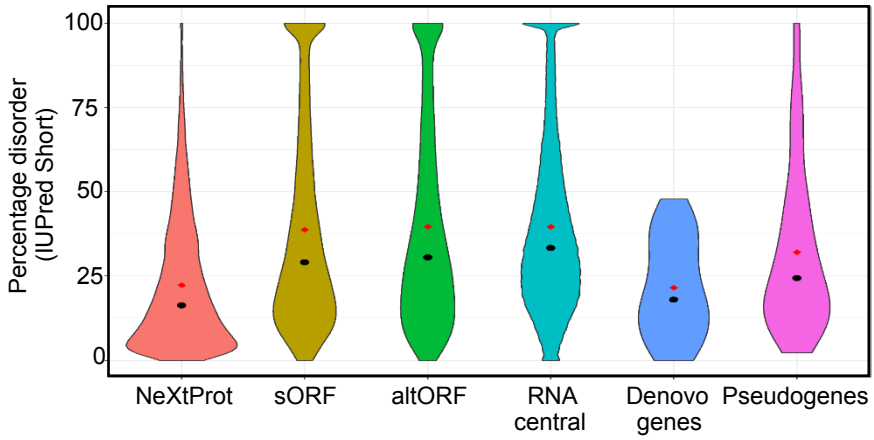
A



B



C



D

Dataset	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
Denovo genes	26	26	12	5	1
Pseudogenes	172	172	87	41	36

Figure S1 A-C. Percentage of protein sequence identified to be disordered (amino-acid disorder score > 0.5) in NeXtProt, and each of the nORF datasets, for three prediction algorithms PONDR, IUPred-Long and IUPred-Short. 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) obtained using PONDR, IUPred-Long and IUPred-short algorithms.

Supplementary Figure 2

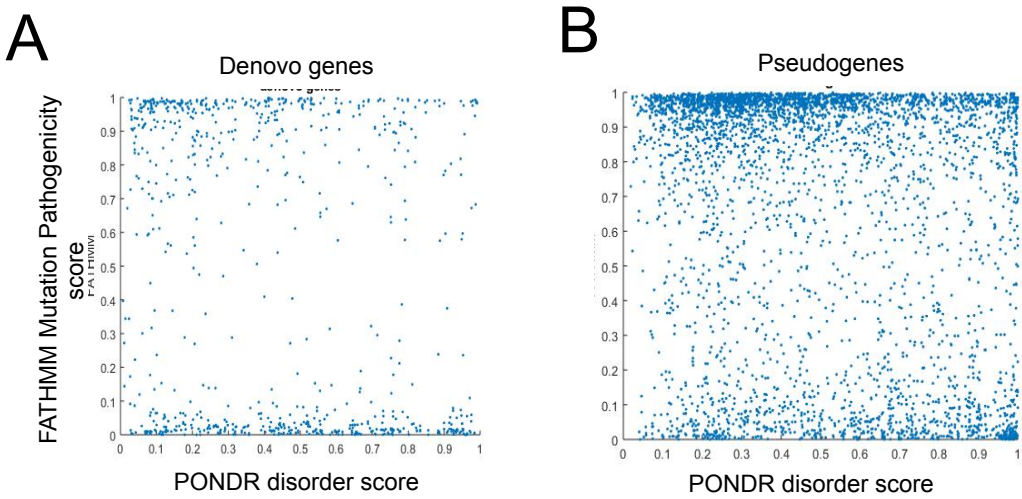


Figure S2. Computed FATHMM scores vs PONDR-predicted disorder scores for **A.** Denovo genes or **B.** Pseudogenes. No correlation between mutation pathogenicity and disorder score was observed in either case.