Unexpected features of the 'dark' proteome

Authors: Nelson Perdigão[1,2], Julian Heinrich[3], Christian Stolte[3], Kenneth S. Sabir[4,5], Michael J. Buckley[3], Bruce Tabor[3], Beth Signal[4], Brian S. Gloss[4], Christopher J. Hammang[4], Burkhard Rost[6], Andrea Schafferhans[6], Seán I. O'Donoghue[3,4,7,*]

Affiliations:

[1]Instituto Superior Técnico, Universidade de Lisboa, Portugal.

[2]Instituto de Sistemas e Robótica, Lisboa, Portugal.

[3]Digital Productivity, CSIRO, Sydney, Australia.

[4]Garvan Institute of Medical Research, Sydney, Australia.

[5]School of Information Technology, The University of Sydney, Australia.

[6]Department for Bioinformatics and Computational Biology, Technische Universität München, Germany.

[7]School of Molecular Bioscience, The University of Sydney, Australia.

*Correspondence to: sean@odonoghuelab.org.

Abstract: We surveyed the 'dark' proteome, i.e., regions of proteins that remain stubbornly inaccessible to both experimental structure determination and modeling. Building upon a recent structural modeling study covering 546,000 proteins across many organisms, we find 44–54% of the proteome in eukaryotes and viruses is dark, compared with only 14% for archaea and bacteria. This includes 68,621 dark proteins, in which the entire sequence lacks reliable similarity to any known structure. Surprisingly, most dark proteins cannot be accounted for by conventional explanations (e.g., intrinsic disorder, transmembrane regions, or compositional bias). Dark proteins are most strongly associated with secretion, the endoplasmic reticulum, specific tissues, disulfide bonding, proteolytic cleavage, and shorter sequence length. They also have surprisingly few interactions with other proteins, and, in humans, some association with cancer and retroviruses. Our results suggest new research directions in structural and computational biology.

One Sentence Summary: We surveyed regions of proteins with unknown 3D structure and found these regions exhibit surprising features that cannot be accounted for by conventional explanations.

Main Text:

Knowledge of protein three-dimensional (3D) structure can be highly valuable, and has led to key discoveries in the life sciences. The PDB, or Protein Data Bank (*1*), that accumulates experimental structures recently past 100,000 entries – a landmark in our understanding of the molecular processes of life. This lags far behind the growth in DNA sequencing; however – since evolution conserves structure more than sequence – computational modeling can leverage the PDB to provide accurate structural predictions for many proteins (*2*). Recently, we created Aquaria (*3*), a resource built by systematically comparing all PDB proteins against 546,000 SwissProt sequences (*4*), i.e., essentially all well-described protein sequences across a wide range of organisms. This comparison resulted in 46 million sequence-to-structure alignments (*3*),

a depth not available from other resources. In this study, we used Aquaria to survey the 'dark' proteome, i.e., regions of protein sequence, or whole sequences, stubbornly inaccessible to either experimental structure determination or modeling, and hence where 3D conformation is completely unknown. The dark proteome has often been overlooked – however, Aquaria allows it to be studied in unprecedented depth, and we were motivated by the hope that surveying the unknown may set future research directions, as dark matter has done in physics.

Although experimental structures cover only 2–4% of SwissProt (fig 1B), our survey revealed that for archaea and bacteria the dark proteome is strikingly small (13–14%), while in eukaryotes and viruses about half of the proteome is dark (fig. 1B). We also found that 40–55% of the dark proteome is comprised of dark proteins, in which the entire sequence has no reliable similarity to any known 3D structure (fig. 1B). Dark proteins tend to be short, and so comprise a substantial fraction of the total number of proteins: 20% for eukaryotes (fig. 2), 8% for archaea (fig. S1), 7% for bacteria (fig. S2), and 44% for viruses (fig. S3).

Conventional explanations for the dark proteome involve factors known to confound experimental structure determination; these include intrinsic disorder (*5*), transmembrane regions (*6*), compositional bias (*7*), and short (<50) or long (>700) sequence length (*8*). We examined these factors in eukaryotes (fig. 2), archaea (fig. S1), bacteria (fig. S2), and viruses (fig. S3) and saw - as expected – that, in many proteins, as these factors increase there is a corresponding increase in 'darkness' (i.e., in the percentage of dark residues). However, we also found unexpected features: firstly, the percentage of transmembrane residues shows an inverse correlation with darkness (figs. 2F, S1F, S2F, and S3F); secondly, most dark proteins have low disorder (figs. 2B, S1B, S2B, and S3B), no compositional bias (figs. 2D, S1D, S2D, and S3D), no transmembrane regions (figs. 2G, S1G, S2G, and S3G), and are slightly shorter than non-dark proteins (figs. 2H, S1H, S2H, and S3H). Thirdly, 45-70% of dark proteins are ordered, globular, and have low compositional bias (figs. 2J, S1J, S2J, and S3J) – and therefore cannot readily be accounted for by conventional explanations. These dark proteins also show highly significant differences in amino acid composition compared to non-dark proteins (figs. 2K, S1K, S2K, and S3K).

Compared to non-dark proteins, we found that dark proteins have surprisingly few interactions with other proteins (figs. 2L, S1L, and S2L). Also, less is known about their functions and subcellular or tissue locations (fig. 3A).

Examining functional annotations enriched amongst eukaryotic dark proteins (fig. 3C and Table S1) we found few under-represented annotations (figs. 3A) – implying that dark proteins fulfill a wide variety of functions – and many over-represented annotations (figs. 3A). As expected, many dark proteins are transmembrane; unexpectedly, however, dark proteins are most strongly over-represented amongst secreted proteins, followed by the endoplasmic reticulum (fig. 3B) – and they are under-represented in only one subcellular location: the cytoplasm (fig. 3C). Interestingly, they are over-represented in specific tissues – e.g., secretary glands and blood – and many are designed to cope with harsh, exterior environments – e.g., skin, saliva, shells, and spores. The only tissue-related annotations where they are under-represented are 'Red blood cells' (consisting largely of cytoplasm), 'Ubiquitous', and 'Widely expressed'. Dark proteins often have domains containing disulfides, and have posttranslational modifications such disulfide bonding, cleavage, phosphorylation, and palmitoylation (fig. 3C). Similar patterns were seen for bacterial and archaeal dark proteins (Table S1 and online resource).

We next considered the human proteome, finding that over half of it is dark (fig. 4A) and less is known about the function, subcellular location, and tissue distribution of dark proteins compared to non-dark proteins (fig. 4C), as in eukaryotes. Similarly, human dark proteins are associated with secretion, transmembrane regions, and cleavage; in addition, we see associations with cancer and endogenous retroviral proteins (fig. 4D).

For human dark proteins, we assessed whether they arise from sequential genes. We found seven such 'dark' gene clusters – including one not previously characterized (Table S2); proteins from these clusters echo features described above as typical for dark proteins. Finally, we tested whether some human dark 'proteins' may be non-coding, finding that this accounts for at most 14 dark proteins (Supplementary Methods).

Which new insights do these results provide? Firstly, the strikingly small dark proteome in archaea and bacteria implies that protein structural knowledge for these small single-celled organisms approaches a level of completeness. The cytoplasm – where most proteins in these organisms are located — is also the only subcellular location in eukaryotes where dark proteins are under-represented, implying that structural knowledge of all cytoplasmic proteins is also relatively complete.

In contrast, the structure of much of the proteome in eukaryotes and viruses remains unknown, although likely for different reasons. In viruses, rapid mutation (*9*) directly undermines sequence-based structure prediction (*2, 3*). In eukaryotes, darkness is commonly accounted for by the greater occurrence of disorder, low complexity (*7*) and transmembrane regions (*10*) – however these factors account for only about half of dark proteins, and are also present in about a third of non-dark proteins (figs. 2J).

Regarding disorder – i.e., intrinsically unstructured regions – we believe that a key contribution of this work is to distinguish this concept from darkness. Previous studies have conflated these concepts (*11, 12*), but they are quite distinct, since: (a) of the 3,141 proteins with 100% disorder, 1,150 have < 25% darkness – of those, 536 have 0% darkness; (b) of the 38,624 proteins with 0% darkness, 3,594 have ≥ 0.25 disorder – many determined from NMR (*13*); and (c) 75% of dark proteins have < 25% disorder. Clarifying this distinction will likely help further the understanding of protein disorder.

How do we account for dark proteins that are ordered, globular, and have low compositional bias (figs. 2J, S1J, S3J, and S3J)? Their distinct amino acid composition (figs. 2G, S1G, S3G, and S3G) suggests they occur in specific subcellular locations (*14*) or have specific functional roles. They may be partly accounted for by other factors known to confound structure determination – e.g., isoelectric point, hydrophobicity, coiled-coil regions, or irregular secondary structure (*8*); however, here we focused on characterizing their biological roles.

Unexpectedly, darkness was found to decrease as the fraction of transmembrane residues increases (figs. 2F, S1F, S2F, and S3F); perhaps the most straightforward explanation is that the prediction methods used to detect transmembrane regions progressively fail with increasing darkness. This explanation fits the data well, and has interesting implications, suggesting the existence of a novel type of transmembrane region, and that transmembrane regions – and proteins – are more common than currently believed.

The prominence of darkness amongst secreted proteins suggests that cleavage, disulfide bonding, and other processing that prepares a protein for the challenges of extracellular and harsh, exterior

environments also confounds structure determination. Secretion may also explain the low number of interaction partners (figs. 2L, S1L, and S2L), since many secreted proteins are 'autonomous', in the sense that they function via few interactions with other proteins.

In conclusion, the dark proteome is a key remaining barrier to understanding biological systems. We completed our survey as many 'structural genomics' initiatives come to a close (*15*); they focused on novel proteins folds, resulting in many landmark structures – including transmembrane proteins – greatly reducing the dark proteome. Similarly, we hope this work will focus structural and computational biology efforts to shed light on the remaining dark proteome.

References and Notes:

1. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235 (2000).
2. J. Haas *et al.*, The Protein Model Portal - a comprehensive resource for protein structure and model information. *Database* **2013**, bat031 (2013).
3. S. I. O'Donoghue *et al.*, Aquaria: Simplifying discovery and insight from protein structures. *Nat. Methods* **in press**, (2015).
4. The UniProt Consortium, Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42**, D191 (2014).
5. C. J. Oldfield *et al.*, Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim. Biophys. Acta* **1834**, 487 (2013).
6. E. P. Carpenter, K. Beis, A. D. Cameron, S. Iwata, Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **18**, 581 (2008).
7. M. A. Huntley, G. B. Golding, Simple sequences are rare in the Protein Data Bank. *Proteins Struct. Funct. Genet.* **48**, 134 (2002).
8. L. Slabinski *et al.*, The challenge of protein structure determination--lessons from structural genomics. *Protein Sci.* **16**, 2472 (2007).
9. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation. *Genetics* **148**, 1667 (1998).
10. J. Liu, B. Rost, Comparing function and structure between entire proteomes. *Protein Sci.* **10**, 1970 (2001).
11. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635 (2004).
12. A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, B. Rost, Improved disorder prediction by combination of orthogonal approaches. *PLoS One* **4**, e4433 (2009).
13. M. Ota *et al.*, An assignment of intrinsically disordered regions of proteins based on NMR structures. *J. Struct. Biol.* **181**, 29 (2013).
14. M. A. Andrade, S. I. O'Donoghue, B. Rost, Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.* **276**, 517 (1998).
15. R. L. Marsden, T. A. Lewis, C. A. Orengo, Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics* **8**, 86 (2007).
16. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173 (2012).
17. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. (Chapman and Hall, London, UK, 1986).
18. B. Rost, R. Casadio, P. Fariselli, C. Sander, Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521 (1995).

19. H. Bigelow, B. Rost, PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Res.* **34**, W186 (2006).

20. Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433 (2005).

21. A. Franceschini *et al.*, STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808 (2013).

22. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289 (1995).

23. S. Durinck, P. T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184 (2009).

24. L. Kong *et al.*, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345 (2007).

25. B. Rhead *et al.*, The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* **38**, D613 (2010).

26. N. E. Davey, G. Trave, T. J. Gibson, How viruses hijack cell regulation. *Trends Biochem. Sci.* **36**, 159 (2011).

27. S. Gibbs *et al.*, Molecular characterization and evolution of the SPRR family of keratinocyte differentiation markers encoding small proline-rich proteins. *Genomics* **16**, 630 (1993).

28. Y. Liang, T. R. Buckley, L. Tu, S. D. Langdon, T. F. Tedder, Structural organization of the human MS4A gene cluster on Chromosome 11q12. *Immunogenetics* **53**, 357 (2001).

29. M. A. Rogers *et al.*, Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein genes embedded in the type I keratin gene domain on chromosome 17q12-21. *J. Biol. Chem.* **276**, 19440 (2001).

30. M. A. Rogers *et al.*, Characterization of a first domain of human high glycine-tyrosine and high sulfur keratin-associated protein (KAP) genes on chromosome 21q22.1. *J. Biol. Chem.* **277**, 48993 (2002).

31. P. Dobrynin, E. Matyunina, S. V. Malov, A. P. Kozlov, The novelty of human cancer/testis antigen encoding genes in evolution. *Int. J. Genomics* **2013**, 105108 (2013).

Acknowledgments: This work is accompanied by an online resource that provides an up-to-date documentation of the dark proteome, and exposes most of the data and analysis results obtained, as well providing additional facilities to interactively explore the data. During the review process, a preliminary version of this resource is available at http://odonoghuelab.org:8030 using the user name 'darkproteome' and password 'IAmTheReviewerOK!'. Shortly prior to publication, the resource will be available without passphrase at http://DarkProteins.org. We would like to thank Profs. David James, Lars Juhl Jensen, Glenn F. King, and Justin Cooper-White for helpful discussions. This work was supported by CSIRO's OCE Science Leader program and its Computational and Simulation Sciences platform, as well as the Alexander von Humboldt Foundation. The authors declare no conflicts of interest.

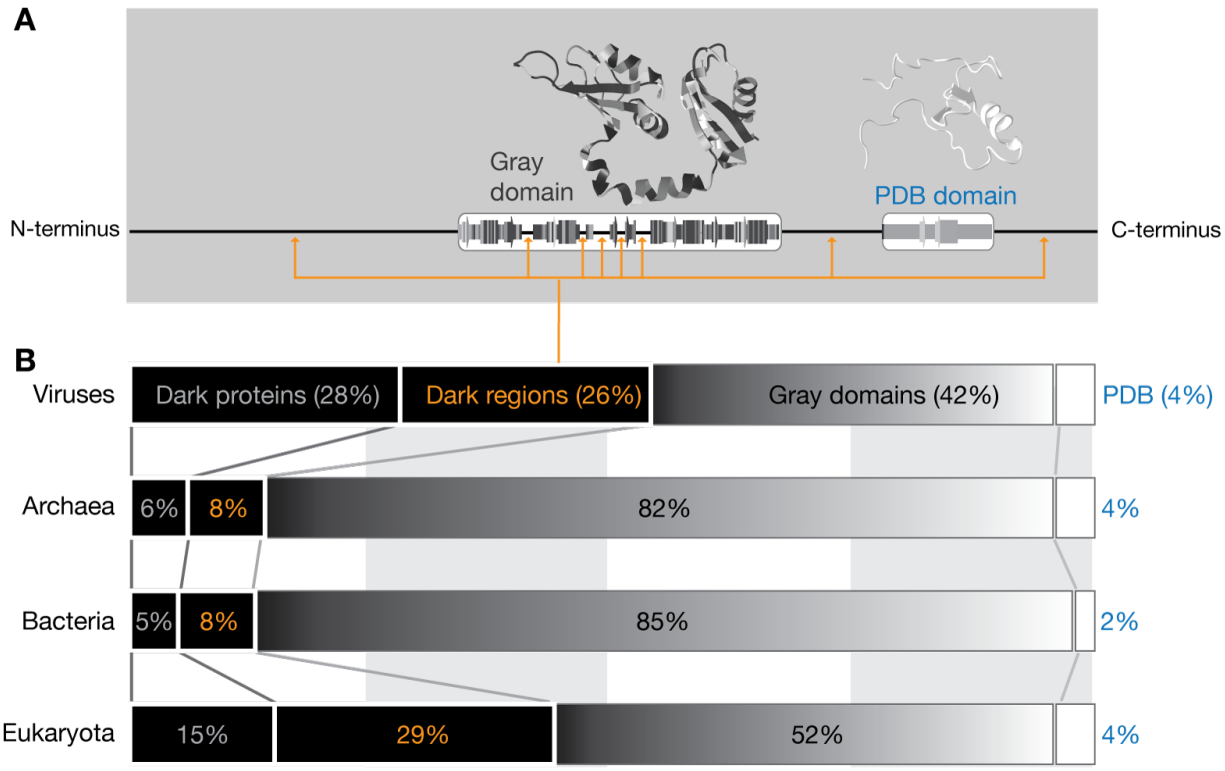

Fig. 1. Dark proteome overview. (**A**) For 546,000 SwissProt sequences we classified each residue into four categories: (1) *PDB domains*: aligns with exact match to at least one PDB entry, (2) *Gray domains*: aligns with reliable similarity to at least one PDB entry, (3) *Dark regions*: no reliable similarity to any PDB entry, and (4) *Dark proteins*: where a single dark region spans the entire sequence. On average, eukaryotic proteins contain eight dark regions, many very short; some are *dark domains*, i.e., conserved dark regions that evolved independently. (**B**) We pooled sequences by organism group and calculated the total fractions of amino acids in the above categories.

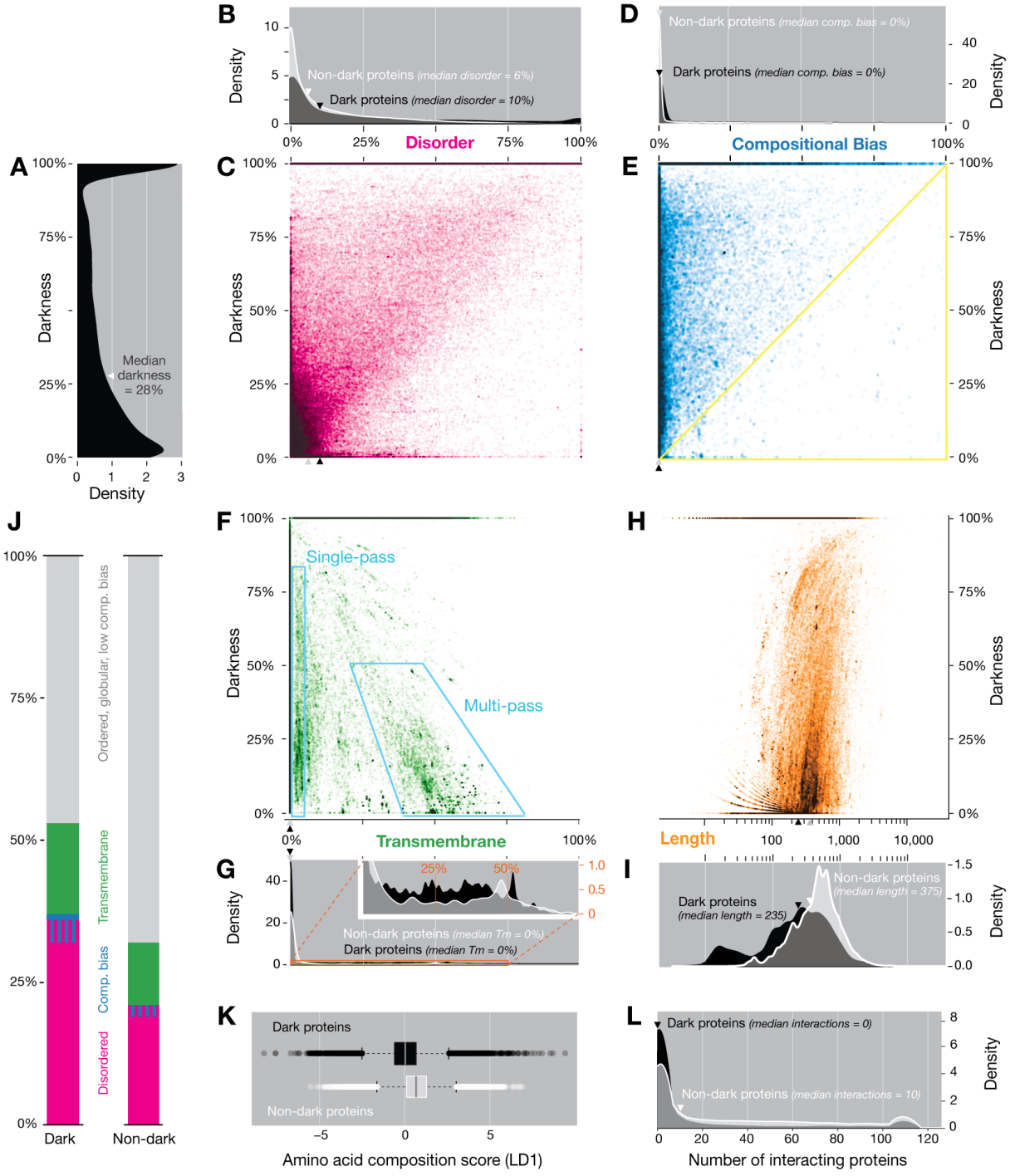

Fig. 2. Darkness vs. other properties for 178,692 eukaryotic proteins. (**A**) The distribution of darkness (i.e., the fraction of dark residues per protein) is bimodal; 50% of proteins have ≤ 28% darkness, while 20% (36,153) have 100% darkness. (**B**) The distribution of disorder (i.e., the fraction of disordered residues per protein) shows that disorder is slightly more prevalent amongst dark proteins, but most dark proteins have low disorder. (**C**) Darkness tends to increase

with disorder, and the majority of highly disordered proteins are dark. (**D**) Compositional bias is low for all proteins, but slightly more prevalent for dark. (**E**) Very few proteins occur in the indicated triangular region, suggesting that most compositionally biased regions are dark. (**F**) Multi-pass transmembrane proteins become unexpectedly rare at ≥ 25% darkness. (**G**) Proportionally more dark proteins are multi-pass transmembrane proteins (zoomed-in insert); however, most dark proteins have no transmembrane residues. (**H**) Darkness tends to increase with sequence length (note the log scale). (**I**) In contrast, dark proteins tend to be shorter. **(J)** About half of dark proteins have ≥ 25% of residues either disordered, transmembrane, or compositionally biased, compared with about one third of non-dark proteins – a smaller difference than expected. (**K**) We used linear discriminate analysis to compare the amino acid composition of dark and non-dark proteins that were ordered, globular and have low compositional bias (i.e., grey regions in **J**). Highly significant differences were found ($p \leq 10^{-10}$) along the first linear discriminant coefficient (LD1). (**L**) Dark proteins have fewer interactions with other proteins.

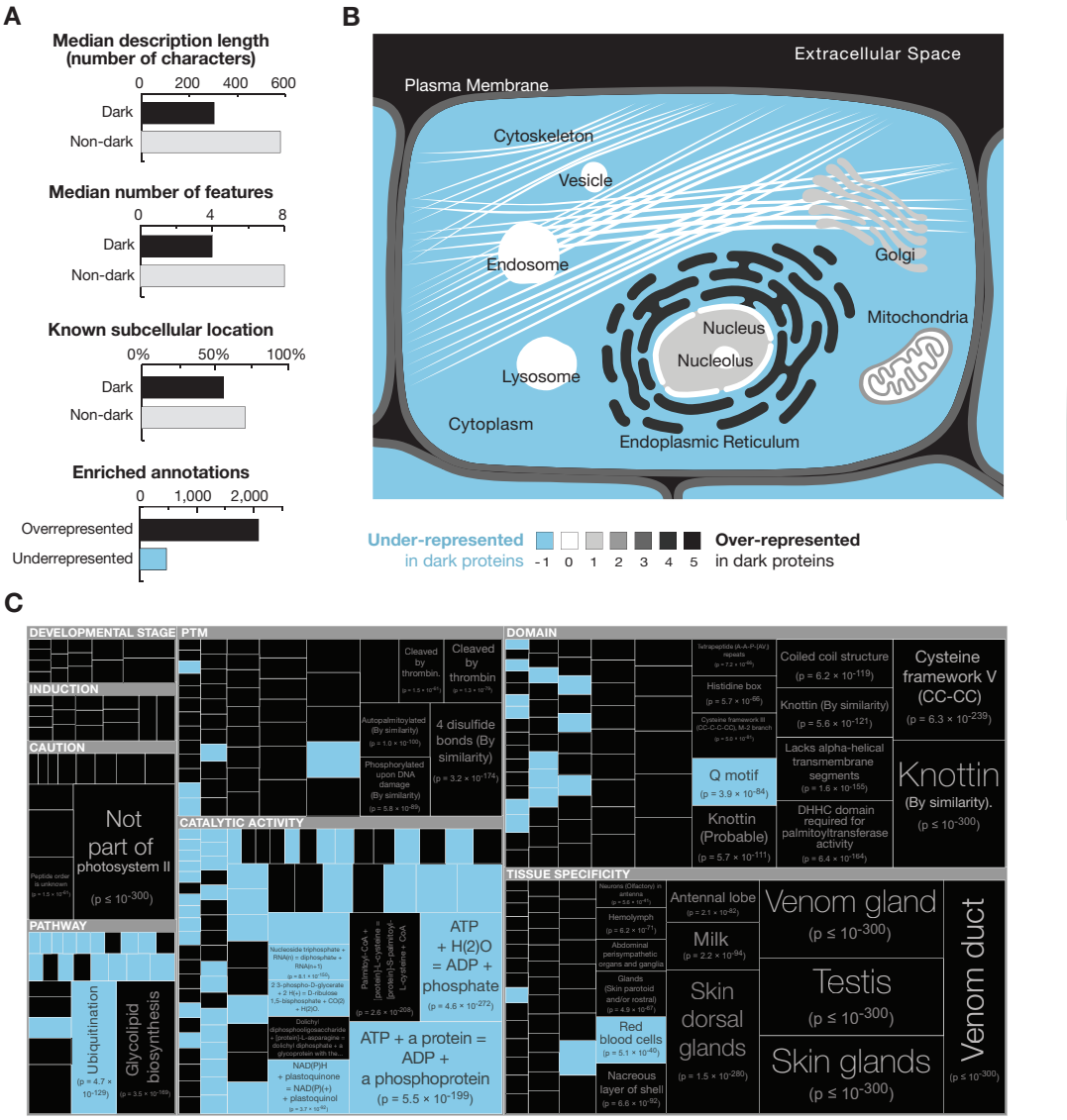

Fig. 3. Dark vs. non-dark proteins in eukaryotes. (**A**) Dark proteins have shorter text describing their function, fewer sequence-specific features, and less complete annotation of subcellular location. Enrichment analysis of dark proteins found four times more over-represented annotations than under-represented. (**B**) Shows cellular regions under- or over-represented in dark proteins. (**C**) Tree map showing under- (blue) or over-represented annotations (black); the area of each cell is proportional to $-\log_{10}(p_j)$, where p_j is the probability associated with the annotation in the cell. Dark proteins are under-represented only in the 'Catalytic site' and 'Pathway' subcategories, where annotations generally require similarity to a PDB structure. Complete enrichment results are in Table S1.

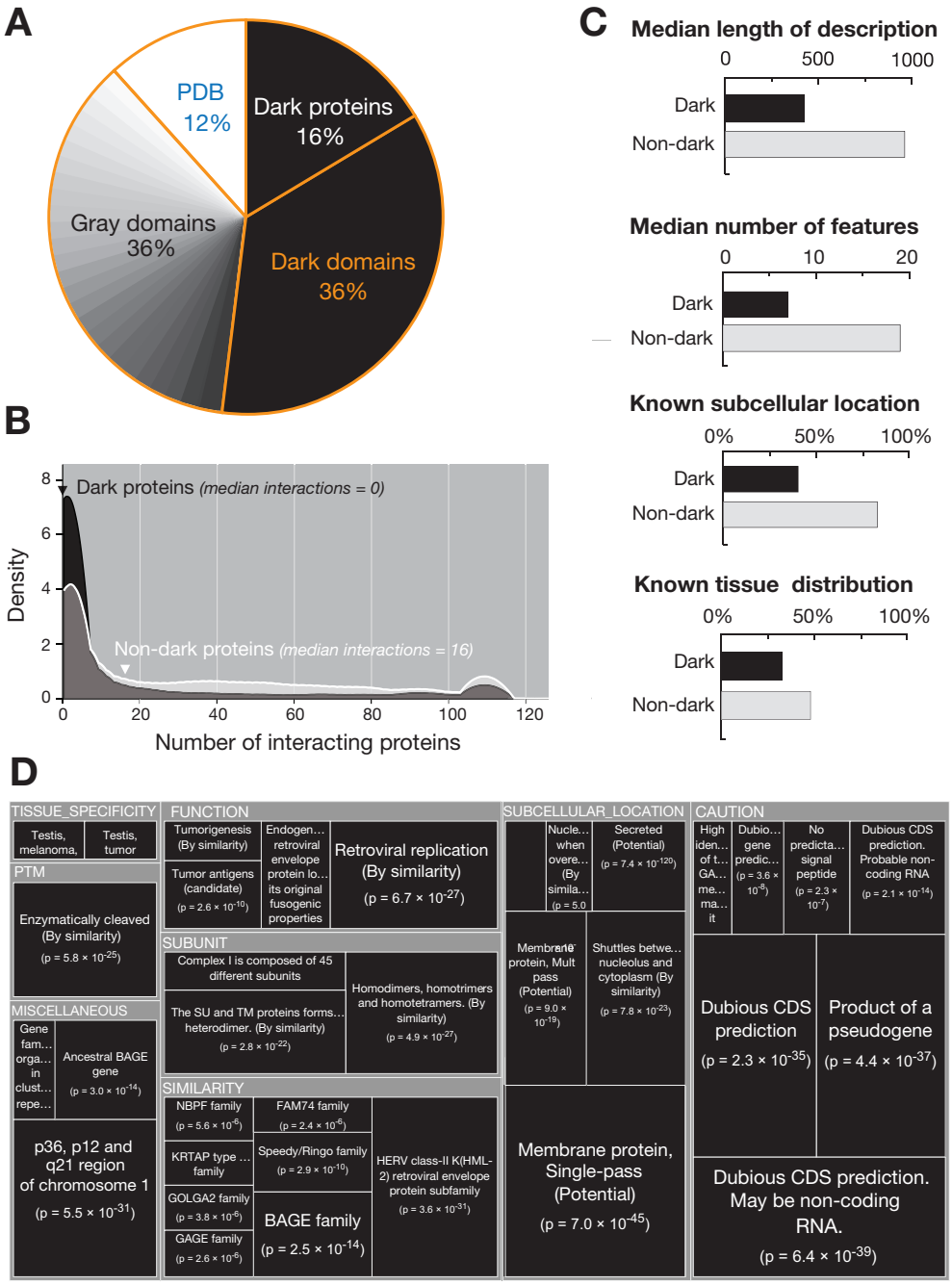

Fig. 4. Dark vs. non-dark proteins in human. (**A**) Shows the fractions of amino acids across all 20,209 human proteins assigned to PDB domains, gray domains, dark regions, and 4,382 dark proteins. (**B**) Dark proteins have fewer interaction partners. (**C**) Dark proteins have shorter functional descriptions, fewer sequence-specific features, and less complete annotation about subcellular location and tissue distribution. (**D**) Tree map showing all annotations over-represented in dark proteins (details are in Table S1). 'Caution' annotations seen for 215 dark 'proteins' indicate they may be long non-coding RNA or arise from pseudogenes; further tests suggest only 14 are non-coding (Supplementary Methods).

Supplementary Materials:

Materials and Methods

Figures S1-S5

Tables S1-S2