Unexpected features of the dark proteome

Nelson Perdigão[a,b], Julian Heinrich[c], Christian Stolte[c], Kenneth S. Sabir[d,e], Michael J. Buckley[c], Bruce Tabor[c], Beth Signal[d], Brian S. Gloss[d], Christopher J. Hammang[d], Burkhard Rost[f], Andrea Schafferhans[f], and Seán I. O'Donoghue[c,d,g,1]

[a]Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; [b]Instituto de Sistemas e Robótica, 1049-001 Lisbon, Portugal; [c]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, NSW 1670, Australia; [d]Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [e]School of Information Technology, The University of Sydney, Sydney, NSW 2006, Australia; [f]Department for Bioinformatics and Computational Biology, Technische Universität München, 80333 Munich, Germany; and [g]School of Molecular Bioscience, The University of Sydney, Sydney, NSW 2006, Australia

Edited by Alan R. Fersht, Medical Research Council Laboratory of Molecular Biology, Cambridge, United Kingdom, and approved October 13, 2015 (received for review April 29, 2015)

We surveyed the "dark" proteome–that is, regions of proteins never observed by experimental structure determination and inaccessible to homology modeling. For 546,000 Swiss-Prot proteins, we found that 44–54% of the proteome in eukaryotes and viruses was dark, compared with only ∼14% in archaea and bacteria. Surprisingly, most of the dark proteome could not be accounted for by conventional explanations, such as intrinsic disorder or transmembrane regions. Nearly half of the dark proteome comprised dark proteins, in which the entire sequence lacked similarity to any known structure. Dark proteins fulfill a wide variety of functions, but a subset showed distinct and largely unexpected features, such as association with secretion, specific tissues, the endoplasmic reticulum, disulfide bonding, and proteolytic cleavage. Dark proteins also had short sequence length, low evolutionary reuse, and few known interactions with other proteins. These results suggest new research directions in structural and computational biology.

structure prediction | protein disorder | transmembrane proteins | secreted proteins | unknown unknowns

The Protein Data Bank (PDB) (1) of experimentally determined macromolecular structures recently surpassed 110,000 entries—a landmark in understanding the molecular machinery of life. Structure determination lags far behind DNA sequencing, but high-throughput computational modeling (2, 3) can leverage the PDB to provide accurate structural predictions for a large fraction of protein sequences. Thus, structural data now scale with sequencing, providing a wealth of detail about molecular functions.

Previous studies have surveyed all sequence and structure data to characterize the "protein universe" [i.e., all proteins from all organisms (4–8)]; from such surveys, we know much of the proteome comprises evolutionarily conserved domains matching relatively few 3D folds (4, 5). These surveys have focused on the "known" and on extrapolating progress toward complete knowledge of all folds in the protein universe. Such studies have guided structural genomics initiatives aimed at determining at least one PDB structure for each distinct fold (8–10).

Our work focuses on the structurally "unknown" (i.e., the fraction of the proteome with no detectable similarity to any PDB structure). We call this fraction the "dark proteome"; we believe that studying the dark proteome will clarify future research directions, as studies of dark matter have done in physics (11).

The analogy to dark matter has inspired surveys of other "unknown" properties of proteins; for example, Levitt (6) examined "orphan" protein sequences that do not match to known sequence profiles, which he termed the "dark matter of the protein universe," and Taylor et al. (12) investigated the "dark matter of protein fold space" (i.e., theoretically plausible folds that have not been observed in native proteins). The same analogy has been used in studies of so-called "junk DNA" (13), which revealed a "hidden layer" of noncoding RNAs (14). Could surveying the dark proteome also reveal undiscovered biological systems?

In fact, discoveries have already resulted from studying regions of unknown structure, namely, intrinsically disordered regions. Long known to confound structure determination (15)—thus forming part of the dark proteome—disorder was largely ignored until recently (16) and yet is now known to play key functional roles, especially in eukaryotes (17). In addition, there is a second type of region that often has unknown structure and is associated with specific biological functions, namely, transmembrane segments (18). Thus, both disorder and transmembrane regions are "known unknowns" (i.e., we know that they are often "dark"). Could the dark proteome contain "unknown unknowns" (i.e., regions with specific functions that confound structure determination and that we are unaware of)?

To address this question, we need to map the dark proteome (i.e., determine all protein regions that cannot be modeled onto any PDB structure). Most available modeling datasets—collected in the Protein Model Portal (PMP) (2)—are not well suited because they aim for breadth of coverage, typically providing only a few PDB matches per protein. Mapping the dark proteome requires depth of coverage, such as the survey of Khafizov et al. (8). (Unfortunately, however, Khafizov et al. used only a few model organisms.) We recently announced Aquaria (19), which provides a median of 35 sequence-to-structure alignments for each Swiss-Prot sequence, a depth of structural data not available from other resources.

Significance

A key remaining frontier in our understanding of biological systems is the "dark proteome"—that is, the regions of proteins where molecular conformation is completely unknown. We systematically surveyed these regions, finding that nearly half of the proteome in eukaryotes is dark and that, surprisingly, most of the darkness cannot be accounted for. We also found that the dark proteome has unexpected features, including an association with secretory tissues, disulfide bonding, low evolutionary conservation, and very few known interactions with other proteins. This work will help future research shed light on the remaining dark proteome, thus revealing molecular processes of life that are currently unknown.

Author contributions: S.I.O. designed research; N.P., J.H., K.S.S., M.J.B., B.T., B.S., B.S.G., C.J.H., and A.S. performed research; N.P., J.H., C.S., B.R., A.S., and S.I.O. analyzed data; and S.I.O. wrote the paper with contributions from N.P.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: This work is accompanied by an online resource (darkproteins.org) that provides periodically updated versions of Datasets S1 and S2, and provides facilities to interactively explore these data.

[1]To whom correspondence should be addressed. Email: sean@odonoghuelab.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1508380112/-/DCSupplemental.

In this work, we used Aquaria to survey the dark proteome in unprecedented depth. We found most of the dark proteome cannot be readily accounted for and shows unexpected features.

Results and Discussion

Mapping the Dark Proteome. We based our survey on 546,000 Swiss-Prot sequences (20). Although smaller than other databases [e.g., TrEMBL (21), which has >50 million sequences], Swiss-Prot is meticulously curated; each entry has many annotations and a high likelihood that it represents a native protein.

Fig. 1*A* shows how we mapped the dark proteome: for each Swiss-Prot sequence, each residue was categorized as "not dark" if it was aligned to a PDB entry in Aquaria (19) and as "dark" otherwise (*SI Methods*). This definition partly underestimates the dark proteome, because Aquaria includes very remote homologies [found using HHblits (22)] and uses all PDB entries, including low-quality structures from electron microscopy (EM) or NMR spectroscopy. We deliberately chose this stringent definition of "darkness," so we can be confident that the dark proteome has completely unknown structure.

Most dark residues occurred in contiguous "dark regions" (Fig. 1); on average, eukaryotic proteins contained eight dark regions, many very short. In many cases, a single dark region covered the entire sequence; we call these "dark proteins" (Fig. 1*B*). Most nondark residues also occurred in continuous regions: some, called "PDB regions," exactly matched to a PDB entry—these residues accounted for only 2–4% of all Swiss-Prot residues (Fig. 1*B*). The remaining nondark residues occurred in "gray regions" (Fig. 1*B*), where 3D structure could be predicted based on similarity to at least one PDB entry.

We found that the dark proteome (i.e., the fraction of residues in dark proteins or dark regions) for archaea and bacteria was strikingly small (13–14%; Fig. 1*B*), implying that structural knowledge for these organisms approaches a level of completeness. In contrast, in eukaryotes and viruses, about half (44–54%) of the proteome was dark (Fig. 1*B*). Of the total dark proteome, nearly half (34–52%) comprised dark proteins.

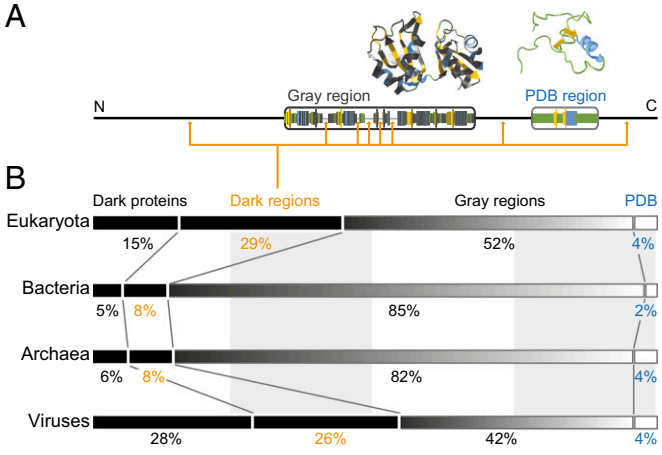

Fig. 1. Mapping the dark proteome. (*A*) For all proteins in Swiss-Prot, each residue was classified into one of four categories: (*i*) PDB regions—residues exactly matched to a PDB entry in Aquaria; (*ii*) gray regions—residues aligned to at least one PDB entry in Aquaria but always with amino acid substitutions (dark gray); (*iii*) dark regions—residues with no matching PDB entry in Aquaria; and (*iv*) dark proteins, where a single dark region spans the entire sequence. (*B*) We then calculated the total fraction of residues in each of the above four categories for all proteins in eukaryotes, bacteria, archaea, and viruses. The dark proteome (i.e., the fraction of residues in dark proteins or dark regions) varies from 13% (bacteria) to 54% (viruses).

We repeated the above analysis using an even more stringent definition for darkness—combining PMP (2) and Aquaria (*SI Methods*)—but this had little effect (Fig. S1).

We also calculated a darkness score for each protein, defined as the percentage of dark residues (Dataset S1). Thus, dark proteins have 100% darkness, whereas proteins with 0% darkness are those where the entire sequence is detectably similar to one or more PDB entries. The distribution of darkness scores was strongly bimodal; most proteins had either low or 100% darkness (density plots in Fig. 2*A* and Figs. S2*A* and S3). For brevity in this work, we use the term "nondark proteins" to refer to those with <100% darkness (noting that a small fraction had high darkness scores).

Dark Proteome Is Mostly Not Disordered. Intrinsically disordered regions are believed to account for much of the dark proteome, especially in eukaryotes (15). To explore this hypothesis, for each protein we calculated the percentage of residues predicted to be disordered [using IUPred (23); *SI Methods*]. Viewing these disorder and darkness scores on a 2D scatter plot, we see that darkness was greater than disorder for almost all eukaryotic proteins (most proteins above the diagonal in Fig. 2*C*), implying that many dark residues were not disordered. In this 2D plot, dark proteins are difficult to resolve because they cluster on a line at the top; thus, we made density plots comparing the disorder distribution for dark vs. nondark proteins (Fig. 2*B*). Surprisingly, most dark proteins had low disorder (median, 10% disorder), not greatly different from nondark proteins (median, 6% disorder); because both of these medians were less than half of the median darkness score (28%; Fig. 2*A*), this finding implies that most of the dark proteome in eukaryotes was not disordered.

In bacteria, archaea, and viruses, nondark proteins, surprisingly, had higher median disorder than dark proteins (Fig. S3). However, the median darkness was always higher still, implying that in these organisms as well, much of the dark proteome was not disordered.

For eukaryotic proteins, the pattern seen in the 2D plot (Fig. 2*C*) also implies that, as expected, most disordered residues were dark. However, a fraction of proteins occur below the diagonal, implying that many disordered residues were not dark. In the corresponding plots for bacteria, archaea, and viruses, this fraction is even larger (Fig. S3), implying that as much as half of all disordered residues were not dark. Many of our colleagues found this last result confusing, often because the distinction between disorder and darkness was unclear. Thus, to clarify: disordered regions are those with evidence of structural heterogeneity (23)—but some become well structured in particular contexts (e.g., most of the 536 Swiss-Prot proteins with 100% disorder and 0% darkness were ribosomal and are presumably well structured within the ribosomal complex). To clarify darkness: these are regions that do not match any PDB entry—but some PDB entries are highly disordered [often these are from EM or NMR (24)], and any sequence aligned to a PDB entry was classified as "not dark" using our stringent definition, because some structural information is known.

Dark Proteome Is Mostly Not Compositionally Biased. Compositional bias is also known to confound structure determination (25). To explore this idea, for each protein we calculated the percentage of compositionally biased residues (*SI Methods*). Viewing these compositional bias and darkness scores on 2D scatter plots, we see that darkness was greater than compositional bias for almost all proteins (Fig. 2*E* and Fig. S3), implying that, as expected, most compositionally biased residues were dark. Together with the density plots for compositional bias (Fig. 2*D* and Fig. S3), it is clear that most dark residues were not compositionally biased and that most dark proteins had very low compositional bias.

Dark Proteome Is Mostly Not Transmembrane. Transmembrane regions are also known to confound structure determination (15, 18). To explore this concept, for each protein we calculated the

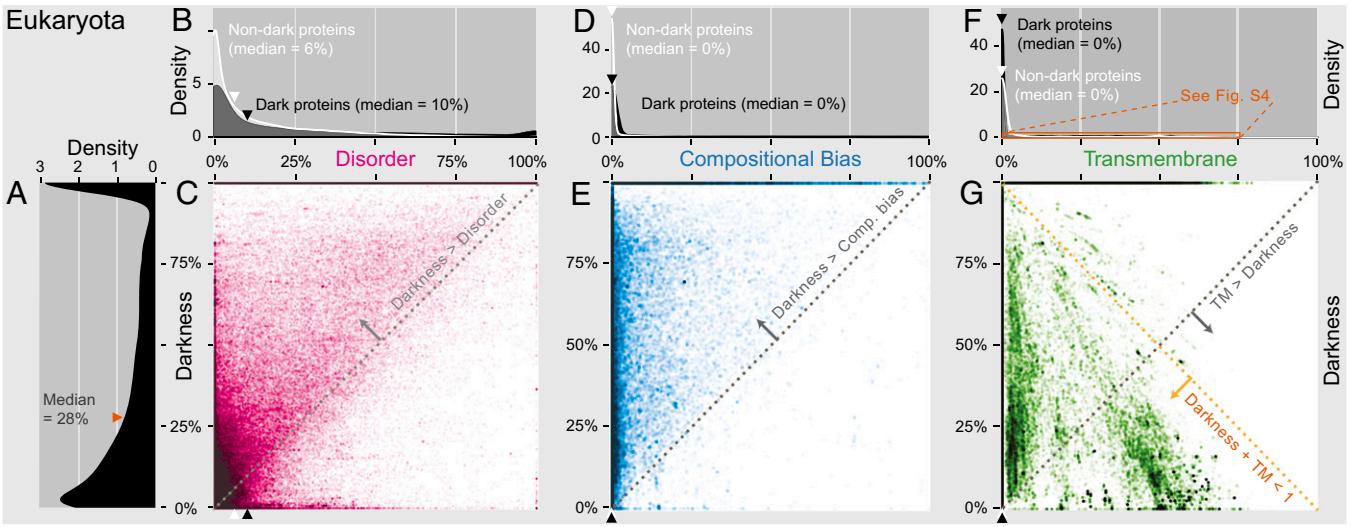

Fig. 2. Darkness vs. disorder, compositional bias, and transmembrane fraction for 178,692 eukaryotic proteins. Overall, these three factors explain only a small part of the dark proteome. Corresponding plots for bacteria, archaea, and viruses are in Fig. S3. In each 2D plot, dark proteins cluster on the line at darkness = 100%. Density plots *A*, *B*, *D*, and *F* are shown in more detail in Fig. S2. (*A*) The distribution of darkness was bimodal: 50% of proteins had ≤28% dark residues; 20% had 100% darkness. (*B*) The distribution of disorder was also bimodal: 50% of dark proteins had ≤10% disordered residues, whereas 4% had 100% disorder; for nondark proteins, 50% had ≤6% disorder, whereas 1% had 100% disorder. Median disorder was much less than median darkness (28%), implying that most of the dark proteome was not disordered. (*C*) Two-dimensional plot shows that darkness > disorder for most proteins (dotted line), implying that most disordered residues were dark and many dark residues were not disordered. (*D*) Compositional bias was 0% in most proteins and slightly more prevalent in dark proteins. (*E*) Two-dimensional plot shows that darkness > compositional bias for most proteins (dotted line), implying that most compositionally biased residues were dark and many dark residues were not compositionally biased. (*F*) Most dark proteins had no transmembrane residues (see Fig. S4 for details). (*G*) Two-dimensional plot shows that darkness > transmembrane fraction for many proteins (gray dotted line), implying that many dark residues were not transmembrane. Most proteins occur in the region where darkness + transmembrane ≤ 1 (orange dotted line), implying that dark and transmembrane regions were mostly disjoint.

percentage of transmembrane residues (*SI Methods*). Viewing these transmembrane and darkness scores on 2D scatter plots, we see that a surprisingly large fraction of transmembrane residues were not dark (Fig. 2*G* and Fig. S3). From the transmembrane density plots (Fig. 2*F* and Fig. S3), we also see that most dark proteins had no transmembrane residues; zooming into these plots shows (as expected) that dark proteins were strongly overrepresented among integral transmembrane proteins in bacteria and archaea but (unexpectedly) not so in eukaryotes and viruses (Fig. S4). Also unexpected was that the transmembrane fraction tended to decrease with increasing darkness in eukaryotes and, across all organisms, was unexpectedly low in proteins with 75% ≤ darkness < 100% (Fig. S5). These results suggest that knowledge of eukaryotic transmembrane protein structures may be more complete than commonly believed, thanks to an ongoing focus on membrane protein structures (26). Alternatively, these results may suggest that the methods used to predict transmembrane regions in this work progressively fail with increasing darkness [i.e., there may be transmembrane regions that are currently undetectable via PROF (27), PROFTMB (28), and other similar methods].

Dark Proteins Are Mostly Unknown Unknowns. To determine the fraction of dark proteins that could be accounted for by a combination of disorder, transmembrane regions, or compositional bias, we categorized each protein as having either a "high" (≥25%) or "low" (<25%) value for each score (Fig. 3). Most of the known unknown (colored fraction) is accounted for by disorder in eukaryotes and viruses and by transmembrane regions in bacteria and archaea (consistent with Figs. S4 and S5). However, a surprisingly large fraction of dark proteins (45–70%) were unknown unknowns (gray fraction) in that they cannot be easily accounted for by these conventional explanations (Fig. 3). This fraction was largest for viral dark proteins, possibly because of their rapid mutation rates (29), which would tend to increase darkness by undermining the sequence-based structure prediction used in this work (2, 19).

To further characterize unknown dark proteins, we next compared them to nondark proteins that were also ordered, globular, and had low compositional bias (i.e., Fig. S6, gray fraction). We found highly significant differences in amino acid composition across all organisms (Fig. S7), suggesting that these dark proteins have distinct functional roles or subcellular locations (30, 31). The largest difference seen was a ∼25% increase in cysteine in dark proteins, consistent with greater prevalence of disulfide bonds

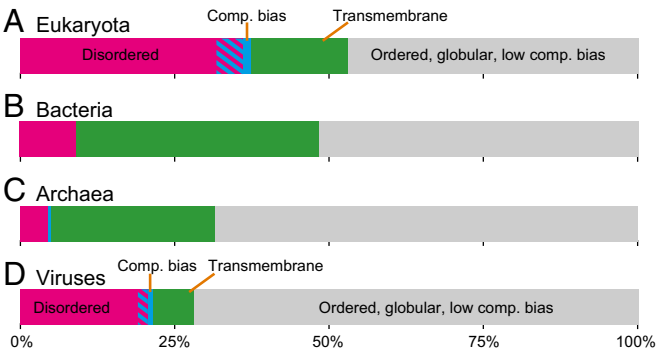

Fig. 3. Known vs. unknown dark proteins. Each linear diagram (38) shows known dark proteins [i.e., those with ≥25% of residues disordered (magenta), compositionally biased (blue), transmembrane (green), or both disordered and compositionally biased (stripes)]. The remaining fraction (gray) are unknown unknowns (i.e., dark proteins predominately ordered, globular, and low in compositional bias). (*A*) In eukaryotes, high disorder accounted for most of the known dark proteins. Most dark proteins with high compositional bias were also highly disordered. (*B* and *C*) In bacteria and archaea, highly transmembrane proteins accounted for most of the known dark proteins (consistent with Figs. S4 and S5). (*D*) Viruses had the largest unknown unknown fraction and, like eukaryotes, had a large fraction of highly disordered dark proteins.

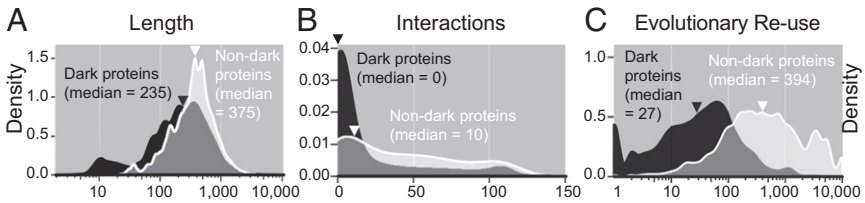

Fig. 4. Length, interactions, and evolutionary reuse for dark vs. nondark eukaryotic proteins. In each case, dark proteins had significantly lower values overall compared with nondark proteins (signed Kolmogorov–Smirnov test, $P \ll 10^{-4}$). Corresponding plots for bacteria, archaea, and viruses are in Fig. S8. (*A*) Dark proteins had shorter sequence length (median of 140 fewer amino acids, or 37% shorter). (*B*) Dark proteins had fewer interactions with other proteins. Note that the small peaks at ∼110 interactions arise from ribosomal proteins. (*C*) Dark proteins had lower evolutionary reuse. In *A* and *C*, note that to interpret the *y* axes values as true density scores, *x* values must be transformed using log base 10 (i.e., 100 becomes 2, etc.).

(*Functions of Dark Proteins*). The next largest differences were increases in both phenylalanine and tryptophan; these amino acids have also been reported to be most increased in transmembrane vs. nontransmembrane proteins (30). This result is consistent with greater prevalence of dark proteins in the range ∼10% < transmembrane < 25% (especially in bacteria and archaea; Fig. S4) but, partly surprising, because most dark proteins have no transmembrane residues (Fig. 2*F* and Fig. S3); a possible explanation could be undetected transmembrane regions (*Dark Proteome Is Mostly Not Transmembrane*).

Shorter Sequence Length. Very short or long sequence length sometimes confounds structure determination (32). We found that dark proteins had 26–50% shorter median length (Fig. 4*A* and Fig. S8) and 16% had a length of <50 aa or a length of >700 aa, compared with 11% of nondark proteins. So, extreme length may explain some dark proteins but not most.

Because dark proteins are shorter, their abundance is underestimated in Fig. 1, which is based on the fraction of dark residues. The fractions for dark proteins were 20% for eukaryotes, 7% for bacteria, 8% for archaea, 44% for viruses, and 13% for all Swiss-Prot proteins.

Fewer Known Protein–Protein Interactions. For each protein, we used STRING (33) to count how many other proteins it interacts with. We found that dark proteins had surprisingly few known interactions (Fig. 4*B* and Fig. S8). Although this observation may arise because dark proteins are not as well studied, the finding is, nonetheless, somewhat surprising because STRING aggregates multiple types of evidence, including high-throughput "omics" experiments and inference via homology.

Lower Evolutionary Reuse. For each protein, we calculated how frequently any part of its sequence has been reused across all other known proteins (*SI Methods*). Dark proteins were reused much less frequently than nondark proteins (Fig. 4*C* and Fig. S8), suggesting that dark proteins may be newly evolved proteins or rare proteins adapted to specific functional niches. This result was partly expected, given how darkness was defined and given the progress of structural genomics in targeting large protein families with unknown structure (8). Low evolutionary reuse also partly explains why dark proteins have few known interactions (Fig. 4*B* and Fig. S8), because many interactions are inferred by homology (33).

Subcellular Location of Dark Proteins. For each protein, we used UniProt annotations to determine its subcellular location; these data were missing for 44% of eukaryotic dark proteins compared with 29% of nondark proteins, consistent with lower evolutionary reuse (because location is often inferred via homology). These location data were used in an enrichment analysis (*SI Methods*) finding that, unexpectedly, eukaryotic dark proteins were most strongly overrepresented in the extracellular space, followed by the endoplasmic reticulum (Fig. 5). This observation partly explains why dark proteins had few interactions (Fig. 4*C* and Fig. S8); compared with intracellular proteins, secreted proteins are often "autonomous," fulfilling their functions via fewer interactions with other proteins. Interestingly, the only subcellular location where dark proteins were underrepresented was the cytoplasm (Fig. 5), and the only tissue where they were underrepresented was red blood cells (Fig. 6), which are mostly cytoplasm; this finding suggests that knowledge of cytoplasmic protein structures approaches a level of completeness—similar to bacterial and archaeal proteins (Fig. 1), most of which are also cytoplasmic.

Functions of Dark Proteins. For each protein, we extracted functional descriptions from the UniProt "CC" annotation; the median length of text in this field was 47% shorter for dark proteins, indicating that less is known about them (again, consistent with lower evolutionary reuse). The resulting set of 242,064 distinct functional annotation terms was used in an enrichment analysis (*SI Methods*), finding that only 2,098 were underrepresented in dark proteins, whereas 3,566 were overrepresented (Dataset S2). This finding implies that, overall, dark proteins fulfill a wide variety of functions, but, nevertheless, a subset have distinct biological functions.

Eukaryotic dark proteins were overrepresented in specific secretory tissues and exterior environments (Fig. 6), consistent with the result that many were secreted (Fig. 5). Eukaryotic dark proteins were also overrepresented in disulfide-rich domains and in disulfide bonds (Fig. 6 and Dataset S2), consistent with increased cysteine (Fig. S7). Additionally, eukaryotic dark proteins were overrepresented in cleavage and other posttranslational modifications known to prepare proteins for harsh environments and to confound experimental structure determination (Fig. 6).

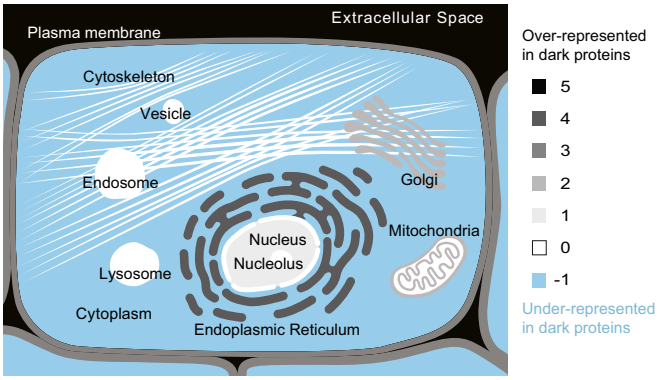

Fig. 5. Cellular locations over- and underrepresented in dark proteins. Pooling annotations for all eukaryotic proteins, we determined which subcellular compartments were enriched in dark proteins; these proteins were most strongly overrepresented in the extracellular space, followed by the endoplasmic reticulum and then the plasma membrane. Dark proteins were underrepresented among cytoplasmic proteins.

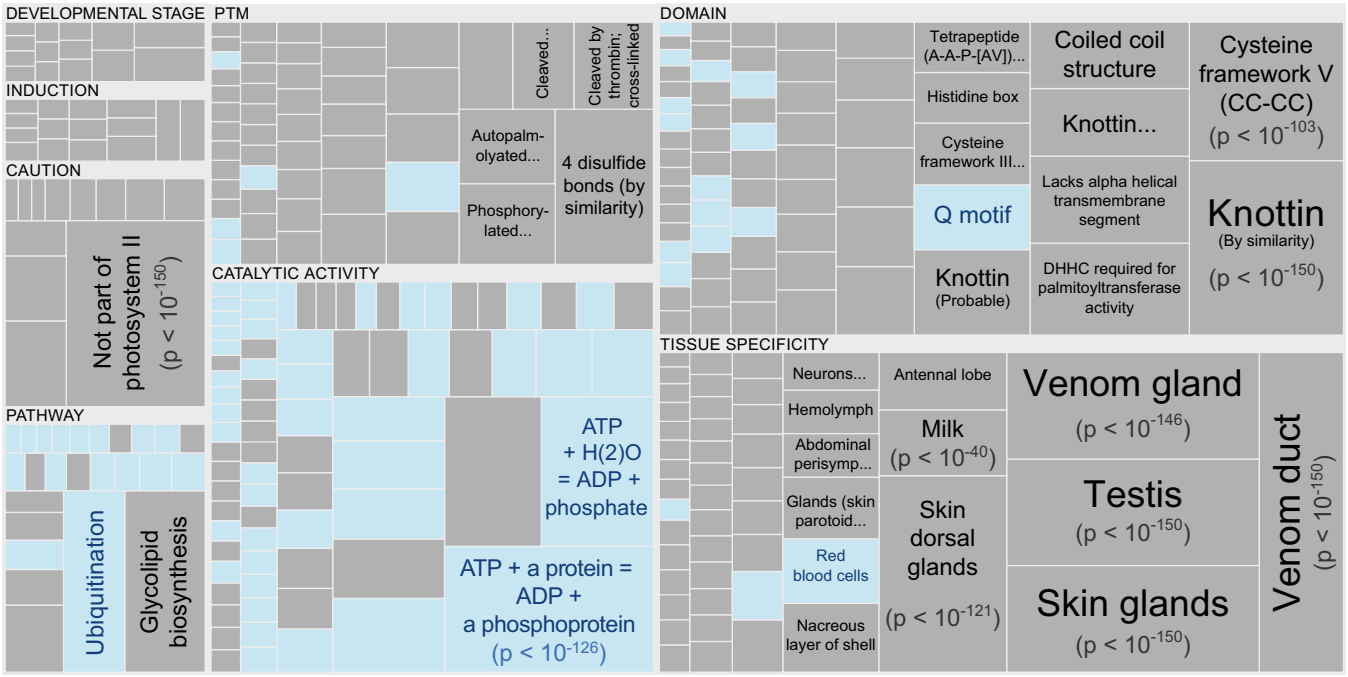

Fig. 6. Functional annotations over- or underrepresented in dark proteins. Pooling annotations for all eukaryotic proteins, we used enrichment analysis to find biological functions associated with dark proteins (Dataset S2). The tree map shows all over- and underrepresented annotations (dark gray and blue, respectively) in eight functional categories; cell area indicates annotation significance [scaled to $-\log_{10}(P)$, using the adjusted P value from Fisher's exact test]. Dark proteins were overrepresented in many specific secretory tissues and underrepresented only in three "tissue" annotations: "Red blood cells," "Ubiquitous," and "Widely expressed" (text not shown). Dark proteins were also overrepresented in cysteine-rich domains and disulfide bonds (of all dark proteins with annotated posttranslational modifications, 16% had disulfide bonds compared with 6.4% for nondark proteins). Dark proteins were underrepresented in many "Catalytic site" and "Pathway" annotations, where inference often requires similarity to a PDB structure.

Coding Potential. The unexpected features of dark proteins may raise the following question: Are they really proteins? Indeed, some overrepresented Swiss-Prot annotations suggest that a fraction of dark proteins are noncoding (Dataset S2); to examine this, we calculated a coding potential score for each human protein [using CPC (34); *SI Methods*]. We found that, of the 4,403 human dark proteins, 2 were likely noncoding and 48 were weakly noncoding; thus, noncoding accounted for only ∼1% of dark proteins. By comparison, of the 15,806 human nondark proteins, ∼0.14% were noncoding or weak noncoding. Thus, as expected, only a very small fraction of Swiss-Prot entries are likely noncoding; although this fraction was enhanced in human dark proteins, it seems likely that most dark proteins really are proteins.

Implications. Mapping the dark proteome has revealed many unexpected features; however, more analyses remain to be done—for example, examining physiochemical properties also known to confound structure determination [e.g., isoelectric point, hydrophobicity, or irregular secondary structure (32)]. Thus, we provide our data for use by others (Dataset S1). In this work, we focused primarily on dark proteins, which account for ∼42% of the dark proteome (Fig. 1); dark regions account for the remaining 58%.

Several insights can be gained from the dark protein features revealed in this work. (*i*) The observation that most dark proteins had low disorder (and many highly disordered proteins are not dark) helps clarify the distinction between darkness and disorder; this clarification in turn will help further studies into protein intrinsic disorder. (*ii*) The observation that transmembrane regions were rare among proteins with 75% ≤ darkness < 100% (especially in eukaryotes) may indicate the existence of transmembrane regions undetected by current prediction methods. (*iii*) The observation that many dark proteins are secreted and posttranslationally modified may help focus development of experimental and bioinformatics methods to better manage such cases. (*iv*) The combination of low evolutionary reuse (Fig. 4*B* and Fig. S8) with high occurrence of disulfide bonds is a signature, suggesting that many dark proteins are newly evolved folds (35) exploring the dark matter of protein folding space (12).

Mostly, however, dark proteins are a mystery; in addition to unknown structure, many have unknown location, unknown function, and no known interactions with other proteins. This is partly accounted for by low evolutionary reuse and by expression in specific tissues and developmental stages. Ultimately, many dark proteins are simply not as well studied as nondark proteins; this work will contribute by highlighting them for subsequent experimental and bioinformatics studies, which may reveal further unknown unknowns.

Future Perspectives. The dark proteome is a moving target, changing as the PDB grows. However, as sequence databases grow at much faster rates, will the dark proteome expand or contract? The current work cannot answer this directly, but earlier surveys have concluded that the number of folds is $\lesssim$10,000 (36), suggesting that the dark proteome will eventually contract if improvements in detection methods [e.g., HHblits (22)] keep pace with the rate of new sequence families. However, those surveys used databases (PDB, Swiss-Prot, etc.) with historical bias toward model organisms; newer experimental approaches are reducing this bias [e.g., structural genomics (8), DNA sequencing of environmental samples (37)]. A recent survey of 8 million protein sequences by Levitt (6) concluded that, eventually, the number of folds may increase linearly with sequences. However, uncertainty in this conclusion arose because ∼22% of the proteins surveyed were "uncharacterized" (i.e., orphans not matching any known sequence family); many of these uncharacterized

proteins may arise from errors in predicting genes from whole genomes.

In the current survey of half a million carefully curated Swiss-Prot sequences, we found that ~13% are dark proteins; although some of these dark proteins were not orphans (just hard to determine folds), most were, as evidenced by low evolutionary reuse scores. Although we used a very different approach from Levitt (6) (a focus on structure versus sequence and very different methods, thresholds, and cutoff values), both of our studies are in broad agreement. Thus, our results suggest that many of the uncharacterized orphan sequences reported by Levitt (or the dark matter of the protein universe) are indeed real proteins; this possibility strengthens the suggestion that folds will eventually increase linearly with sequences (6) and implies that dark proteins may remain a sizeable and irreducible feature of the protein universe.

Conclusions

The dark proteome is a key remaining frontier in the understanding of biological systems. This work will help focus future structural genomics and computational biology efforts to shed light on the remaining dark proteome, thus revealing currently unknown molecular processes of life.

Methods

In each subsection of *Results and Discussion*, we briefly outline the bioinformatics methods used to derive the presented results. *SI Methods* gives further details on how we derived the scores used in the work (darkness, disorder, coding potential, etc.), the statistics used to analyze the scores, and the density plots, 2D plots, linear diagrams, cell map, and tree maps used to visualize the scores. This work is accompanied by an online resource (darkproteins.org) that provides periodically updated versions of Datasets S1 and S2, and provides facilities to interactively explore these data.

ACKNOWLEDGMENTS. We thank Drs. David James, Lars Juhl Jensen, Glenn F. King, William John Wilson, and Justin Cooper-White for helpful discussions. This work was supported by Commonwealth Scientific and Industrial Research Organisation's Office of the Chief Executive Science Leader program and Computational and Simulation Sciences platform, as well as the Alexander von Humboldt Foundation, and the Fundação para a Ciência e Tecnologia.

1. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.
2. Haas J, et al. (2013) The Protein Model Portal–A comprehensive resource for protein structure and model information. *Database (Oxford)* 2013:bat031.
3. Petrey D, et al. (2015) Template-based prediction of protein function. *Curr Opin Struct Biol* 32:33–38.
4. Chothia C (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357(6379):543–544.
5. Holm L, Sander C (1996) Mapping the protein universe. *Science* 273(5275):595–603.
6. Levitt M (2009) Nature of the protein universe. *Proc Natl Acad Sci USA* 106(27): 11079–11084.
7. Nepomnyachiy S, Ben-Tal N, Kolodny R (2014) Global view of the protein universe. *Proc Natl Acad Sci USA* 111(32):11691–11696.
8. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proc Natl Acad Sci USA* 111(10):3733–3738.
9. Burley SK, et al. (1999) Structural genomics: Beyond the human genome project. *Nat Genet* 23(2):151–157.
10. Marsden RL, Lewis TA, Orengo CA (2007) Towards a comprehensive structural coverage of completed genomes: A structural genomics viewpoint. *BMC Bioinformatics* 8:86.
11. Bertone G, Hooper D, Silk J (2005) Particle dark matter: Evidence, candidates and constraints. *Phys Rep* 405(5-6):279–390.
12. Taylor WR, Chelliah V, Hollup SM, MacDonald JT, Jonassen I (2009) Probing the "dark matter" of protein fold space. *Structure* 17(9):1244–1252.
13. Travis J (2002) Biological Dark Matter: Newfound RNA suggests a hidden complexity inside cells. *Sci News* 161(2):24–25.
14. Mattick JS (2003) Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* 25(10):930–939.
15. Oldfield CJ, et al. (2013) Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim Biophys Acta* 1834(2):487–498.
16. Dunker AK, et al. (2001) Intrinsically disordered protein. *J Mol Graph Model* 19(1): 26–59.
17. Oldfield CJ, Dunker AK (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* 83:553–584.
18. Carpenter EP, Beis K, Cameron AD, Iwata S (2008) Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol* 18(5):581–586.
19. O'Donoghue SI, et al. (2015) Aquaria: Simplifying discovery and insight from protein structures. *Nat Methods* 12(2):98–99.
20. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42(Database issue):D191–D198.
21. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1):45–48.
22. Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9(2):173–175.
23. Dosztányi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21(16):3433–3434.
24. Ota M, et al. (2013) An assignment of intrinsically disordered regions of proteins based on NMR structures. *J Struct Biol* 181(1):29–36.
25. Huntley MA, Golding GB (2002) Simple sequences are rare in the Protein Data Bank. *Proteins* 48(1):134–140.
26. Punta M, et al. (2009) Structural genomics target selection for the New York consortium on membrane protein structure. *J Struct Funct Genomics* 10(4):255–268.
27. Rost B, Casadio R, Fariselli P, Sander C (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci* 4(3):521–533.
28. Bigelow H, Rost B (2006) PROFtmb: A web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Res* 34(Web Server issue):W186–W188.
29. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148(4):1667–1686.
30. Cedano J, Aloy P, Pérez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. *J Mol Biol* 266(3):594–600.
31. Andrade MA, O'Donoghue SI, Rost B (1998) Adaptation of protein surfaces to subcellular location. *J Mol Biol* 276(2):517–525.
32. Slabinski L, et al. (2007) The challenge of protein structure determination–lessons from structural genomics. *Protein Sci* 16(11):2472–2482.
33. Franceschini A, et al. (2013) STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41(Database issue):D808–D815.
34. Kong L, et al. (2007) CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 35(Web Server issue):W345–W349.
35. Edwards H, Abeln S, Deane CM (2013) Exploring fold space preferences of new-born and ancient protein superfamilies. *PLOS Comput Biol* 9(11):e1003325.
36. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. *Nature* 420(6912):218–223.
37. Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6(11):805–814.
38. Chapman P, Stapleton G, Rodgers P, Micallef L, Blake A (2014) Visualizing Sets: An Empirical Comparison of Diagram Types. *Visualizing Sets: An Empirical Comparison of Diagram Types*, eds Dwyer T, Purchace H, Delaney A (Springer, Berlin), pp 146–160.
39. Davey NE, Travé G, Gibson TJ (2011) How viruses hijack cell regulation. *Trends Biochem Sci* 36(3):159–169.
40. Silverman BW (1986) *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London).
41. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337(3):635–645.
42. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One* 4(2):e4433.
43. Hauser M, Mayer CE, Söding J (2013) kClust: Fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics* 14:248.
44. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57(1): 289–300.
45. Shneiderman B (1992) Tree visualization with Tree-Maps: 2-D space-filling approach. *ACM T Graphic* 11(1):92–99.
46. Binder JX, et al. (2014) COMPARTMENTS: Unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014:bau012.
47. Durinck S, Spellman PT, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 4(8):1184–1191.