Protein Domains of Unknown Function Are Essential in Bacteria

Norman F. Goodacre,[a] Dietlind L. Gerloff,[b] Peter Uetz[c]

Georgetown University, Washington, DC, USA[a]; Foundation for Applied Molecular Evolution, Gainesville, Florida, USA[b]; Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, USA[c]

ABSTRACT More than 20% of all protein domains are currently annotated as "domains of unknown function" (DUFs). About 2,700 DUFs are found in bacteria compared with just over 1,500 in eukaryotes. Over 800 DUFs are shared between bacteria and eukaryotes, and about 300 of these are also present in archaea. A total of 2,786 bacterial Pfam domains even occur in animals, including 320 DUFs. Evolutionary conservation suggests that many of these DUFs are important. Here we show that 355 essential proteins in 16 model bacterial species contain 238 DUFs, most of which represent single-domain proteins, clearly establishing the biological essentiality of DUFs. We suggest that experimental research should focus on conserved and essential DUFs (eDUFs) for functional analysis given their important function and wide taxonomic distribution, including bacterial pathogens.

IMPORTANCE The functional units of proteins are domains. Typically, each domain has a distinct structure and function. Genomes encode thousands of domains, and many of the domains have no known function (domains of unknown function [DUFs]). They are often ignored as of little relevance, given that many of them are found in only a few genomes. Here we show that many DUFs are essential DUFs (eDUFs) based on their presence in essential proteins. We also show that eDUFs are often essential even if they are found in relatively few genomes. However, in general, more common DUFs are more often essential than rare DUFs.

Received 4 September 2013 **Accepted** 21 November 2013 **Published** 31 December 2013

Citation Goodacre NF, Gerloff DL, Uetz P. 2013. Protein domains of unknown function are essential in bacteria. mBio 5(1):e00744-13. doi:10.1128/mBio.00744-13.

Editor Claire Fraser, University of Maryland, School of Medicine

Address correspondence to Peter Uetz, peter@uetz.us.

Most proteins are built of one or several domains that serve as the key mediators for their function(s). Given the ease of sequence acquisition today, the classic definition of a domain as an independently folding, and largely independent, tertiary structural unit is often replaced by a sequence-based "domain" concept, outside structural biology (1, 2). Segmenting proteins based on homology alone (3) is powerful because it does not require a representative with a known structure, and the initial predictions are largely automatable. Over time, structure determination can refine the domain boundaries. However, a large proportion of protein functional insights today are derived experimentally before three-dimensional (3D) structural information becomes available.

A variety of sequence-based domain collections exists; however, there is substantial overlap among databases (3). InterPro, which integrates Pfam as well as other sequence signatures, covers a large proportion of the protein sequences in the UniProt database and offers a good initial understanding of domain diversity (see Table S1A in the supplemental material). The Pfam database (if one includes Pfam B which contains automatically generated domain annotations) currently lists about 15,000 protein families (4). For example, the genome of *Escherichia coli* K-12 (MG1655) encodes 5,475 recognizable domains that are classified into 2,407 families in Pfam 26.0. The most highly represented domain in *E. coli*, the ATP-binding domain of the ABC transporter family (Pfam accession no. PF00005), is detected in 78 proteins with a total of 95 copies in the K-12 strain.

Sequence-based domain assignment requires detectable homology between several protein fragments. However, very few proteins or domains are universally conserved across all species. In 2010 (Pfam release 23.0), only 16% of all characterized domains were found in all kingdoms of life (but not necessarily in all species) (5). The number of recurrent domains by the sequence-based definition is about 3 orders of magnitude smaller than the number of species (thousands versus millions). A majority of these recurrent domains can be presumed to correspond to independently folding fragments that are more likely tractable in the laboratory than full-length proteins, especially in medium- or high-throughput experiments (6).

Domain assignments have become an effective starting point for studying and understanding molecular biology across the bewildering multitude of species. However, despite decades of research, more than 20% of all domains in the Pfam database, the ~3,600 so-called domains of unknown function (DUFs) (4) remain poorly understood (5). Pfam's DUF families are composed entirely of functionally uncharacterized protein fragments when they are assigned by the curators. New information about individual members may emerge before the next time assignment is reconsidered. However, in most instances, DUFs are in need of further study before they can be as informative as other Pfam domains. Taxonomically, about 9% of the DUFs in Pfam release 23.0 spanned all domains of life (*Bacteria*, *Archaea*, and *Eukarya*), while nearly half (43%) had been detected only in bacteria. An-

other 19% were only found in eukaryotes, and 3% were restricted to archaea (5).

The importance of prioritizing DUFs has been recognized in various experimental and/or computational characterization efforts (4, 5, 7). Bateman et al. (5) discussed DUFs from a structural perspective without providing specific information or prioritization for experimental study. In contrast, Dessailly et al. (8) prioritized the most phylogenetically common domains for crystallization but did not focus on DUFs in their approach. While many conserved domains have been preferred targets in previous studies (7), there has been no global attempt to provide a priority list for bacterial proteins. Related projects such as CALIPHO (Computer and Laboratory Investigation of Proteins of Human Origin) (9) focus on the approximately 5,000 human proteins with unknown function. However, highly conserved proteins may yield insights into the biology of many processes and species. Here we examine DUFs from a microbiological perspective and focus on the prospects of targeting DUFs found in bacteria. Not only is sequence information from culturable and unculturable bacterial isolates increasing faster than for other taxonomic groups, but bacterial proteins are also more tractable by high-throughput experiments in the laboratory, not the least because of their availability as complete clone sets (10). The bacterial kingdom also makes a substantive contribution to human infectious disease burden and death (11), which calls for a better understanding of the protein complement of pathogenic species. Here we attempted to identify DUFs that should be rewarding targets for experimental analysis in bacteria, and bacterial pathogens in particular. We identified DUFs that are not only highly conserved but that are essential in at least one species. Many of these uncharacterized bacterial domains are also found in eukaryotes, hence experimental analysis of these prokaryotic representatives should also shed light on the biology of higher life forms.

Results. For the domain survey presented here, we focused on Pfam even though we have used several other databases (see Table S1A in the supplemental material). The phylogenetic diversity of domains was investigated using the NCBI taxonomy, iToL, and the PATRIC database (see Methods for details).

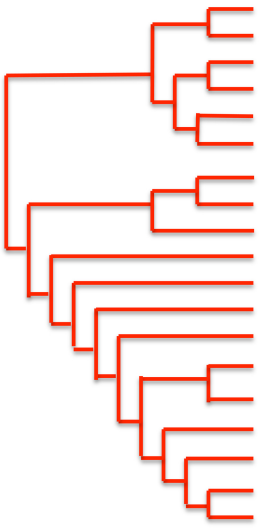
(i) Phylogenetic diversity of DUFs. Domains of unknown function occur in large numbers in all kingdoms of life, ranging from about just over 1,500 in eukaryotes to 2,704 in bacteria (see Table S1B and Fig. S1 in the supplemental material). However, DUFs represent a much greater proportion of domains in bacteria than they do in other kingdoms with about one-third of all detected domains being DUFs. This is surprising, as a large number of DUFs are shared between bacteria and other domains of life (Fig. S1). There are nearly 900 DUFs in common between bacteria and eukaryotes. In fact, more than 300 DUFs are found in all three kingdoms of life (Fig. S1). It is noteworthy that over three times as many DUFs have been defined in bacteria as in plants, the kingdom with the next-highest DUF count, although the difference may be explained largely by the different numbers of completely sequenced genomes from the two kingdoms. According to InterPro (36.0) (12), 2,702 bacterial domains are also present in animals, including 311 DUFs (Table S1B).

Among bacterial phyla, we observe a trend for larger phyla to have proportionally more DUFs, reflecting their larger genetic diversity (see Fig. S2A in the supplemental material). For example, 31% of proteobacterial domains are DUFs, while this fraction is only 25% for *Actinobacteria* and 21% for *Spirochaetes*. For these three phyla, the total numbers of domains annotated in Pfam 26.0 are 6,203, 4,029, and 2,966, respectively. This trend is evident despite the fact that the discovery of new domains and DUFs inevitably tapers off as more strains in a phylum are sequenced (Fig. S2B). It is also curious that DUFs tend to occur in relatively larger proteins in eukaryotes but relatively smaller proteins in bacteria (Fig. S3). The size distribution of DUF-containing proteins in bacteria appears skewed away from larger proteins compared to other bacterial proteins (Fig. S3, top left); therefore, this observation is not merely a reflection of generally different protein lengths in eukaryotes and bacteria. Moreover, the majority of eukaryotic proteins contain numbers of annotated domains that are comparable to those found in bacterial proteins (Fig. S3, right).

We have compiled 3,427 DUFs in Table S1C in the supplemental material, ranked by the number of fully sequenced bacterial genomes in which they are present. The first 24 DUFs are present in 500 or more species and usually in both eukaryotes and prokaryotes, as well as distributed over the great majority of bacterial phyla. While these protein domains are less common across archaea or fungi, most of them are present in more than 20% of all genomes. Distribution of the top 50 DUFs across taxa is rather variable, often as high as 80 to 90%, but occasionally as low as 15% of bacterial families are represented (where representation is defined as at least one genome in the family possessing the particular DUF). For example, the top-ranked DUF, DUF933, is present in 1,000 species represented by 1,495 completely sequenced genomes (Table S1C and S1D). In contrast, DUF177, ranked 5th, is missing in archaea and fungi and present in eukaryotes in only a few instances. Nevertheless, the domain is present in most bacteria, including 206 bacterial families and 859 completely sequenced bacterial species.

(ii) Structural representation of DUFs. Currently, structures of about 5,000 (36%) of the nearly 15,000 Pfam domains have been characterized, including 379 (10.5%) of the ~3,600 Pfam DUFs. A table of the top 20 most common DUFs (ranked by the number of sequenced bacterial genomes) for which a structure has been deposited in Protein Data Bank (PDB) (13) is provided in Table S1E in the supplemental material.

(iii) Many DUFs are essential. Across the 19 bacterial species represented in the Database of Essential Genes (DEG) (14), more than 10,000 essential genes have been identified (including redundancies). We found 393 of these proteins to contain at least one of 255 different DUFs (see Table S1F in the supplemental material). While 83 of those proteins contain multiple domains, the remainder appears to contain only the DUF. This clearly establishes these DUFs as essential DUFs (eDUFs). All model organisms that have been analyzed this way contain eDUFs (Fig. 1). Although the total number of domains for these model organisms has slightly decreased over the past 5 years (Pfam v23 versus v26), the numbers of DUFs and eDUFs have markedly increased (from 282 and 77 to 359 and 89, respectively, in *E. coli* [data not shown]). We explain the substantial increase in DUF numbers by the dramatic increase in available genome sequences which allowed new domains to be recognized by Pfam's comparative approach. Interestingly, we found three domains that occur both in essential multidomain proteins as well as essential single-domain proteins so that domains are likely to be essential in the multidomain configuration as well: DUF31 (Pfam accession no. PF01732) is a predicted peptidase domain that is also found in two essential *Mycoplasma* proteins together with another peptidase domain (Pfam accession no.



Species	Proteins	Domains	DUFs	eDUFs
Mycoplasma pulmonis	778	498	19	8
Mycoplasma genitalium	475	437	24	15
Streptococcus pneumoniae	2180	1112	133	25
Streptococcus sanguinis	2269	1176	169	24
Staphylococcus aureus	2934	1362	184	33
Bacillus subtilis	1541	1048	96	13
Porphyromonas gingivalis	2064	1089	124	35
Bacteroides thetaiotaomicron	4785	1470	247	57
Mycobacterium tuberculosis	8115	1417	260	47
Helicobacter pylori	2409	997	58	22
Caulobacter crescentus	3796	1633	253	40
Burkholderia thailandensis	5571	1982	309	42
Francisella novicida	1742	1149	82	23
Acinetobacter baylyi	3292	1535	205	44
Pseudomonas aeruginosa	5886	2046	400	59
Vibrio cholerae	3366	1734	294	86
Haemophilus influenzae	1709	1283	116	54
Salmonella enterica	5186	2169	378	82
Escherichia coli	4210	2048	359	89

FIG 1 Essential domains of unknown function (eDUFs) are common among bacteria. The table shows species for which essential genes have been determined. All numbers were derived using the reference proteome of either the DEG strain or a common (fully sequenced) strain. Different strains may have different numbers. Domains are all Pfam domains that are not DUFs, while eDUFs are a subset of DUFs. Many essential genes encode DUFs as their only domain. This table is based on Pfam v26 (2012). For a complete list of eDUFs, see Table S1F in the supplemental material.

PF00949). Similarly, DUF59 (Pfam accession no. PF01883) is found in a series of proteins of various functions in *Mycobacterium* and *Caulobacter*, but usually in combination with PF10609, an ATPase-like domain. Finally, DUF161 (Pfam accession no. PF02588) is found as an essential single-domain and multidomain protein in combination with DUF2179, another DUF (Pfam accession no. PF10035).

Interestingly, there does not seem to be a strong correlation between phylogenetic conservation and essentiality (Fig. 2). While highly conserved DUFs are more likely to be essential, poorly conserved DUFs (as measured by the number of genomes they are found in) are still essential in many cases. Our data set contains essential proteins that contain both known and unknown domains, but surprisingly, the majority of essential proteins containing DUFs contains only the eDUF (see Table S1F in the supplemental material).

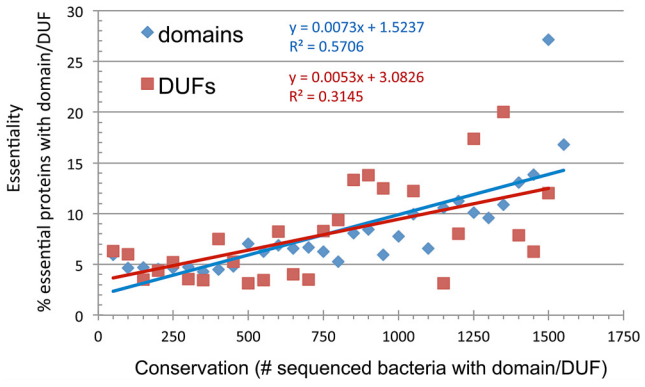


FIG 2 Many essential domains of unknown function (eDUFs) are not highly conserved. Although eDUFs tend to be better conserved (as measured by the number of genomes they are encoded in), the correlation is weak. Even poorly conserved DUFs are often essential. The linear fit was performed using simple linear regression. The figure uses data from DEG version 8.5.

(iv) Functional clues from attributes of DUF-containing proteins. This study is not primarily concerned with protein function prediction but rather relies on existing database annotations. However, we tried to obtain rough estimates of what functions might be associated with specific DUFs, using a simple subtractive protocol on transferred annotations (see Methods). Briefly, we collected potential functional attributes for the DUF-containing proteins found in 10 model bacteria (1,786 of all 3,601 known DUFs). Then, we derived very preliminary speculative clues for 31 of the top 50 bacterial DUFs as a starting point for experimental research by comparing full-length protein annotations in UniProt with domain-specific curated annotations in the Pfam2GO list (15) for all known domains in each protein. Additionally, we extended our view by considering STRING database (version 9.0) (16) predictions. As one has to expect, for functionally uncharacterized families, most predicted attributes remain relatively general and include "functions" such as ATP binding or relate to rather broad biological processes such as transcription. The potential attributes of nine DUFs indicate an integral membrane subcellular location, which may partly explain why the functions of these domains have remained unknown, given the difficulty of studying membrane proteins. Many of the top 50 bacterial DUFs also have functional associations that point to metabolic pathways. Since deeper functional predictions are beyond the scope of this paper, we refrain from a more detailed discussion of the clues shown in Table S1G in the supplemental material and refer the reader to more specialized studies (17–20).

(v) Domains in bacterial pathogens and model organisms. Interestingly, all of the top 50 DUFs (by the number of sequenced bacterial genomes) are found in at least one functionally annotated protein in 13 model organisms, 10 bacterial organisms and 3 eukaryotic organisms (see Table S1H in the supplemental mate-

rial). In 41 cases, a DUF is found in an annotated protein in more than one of these organisms. Seventeen of the top 50 DUFs occur in 32 proteins as the only identified domains—i.e., proteins that consist entirely of DUFs (data not shown). We speculate that these proteins may well be some of the most interesting targets for future research in this field.

We have compiled domain and DUF counts for 13 model organisms (including *Homo sapiens*) and important pathogens for which complete open reading frame (ORF) clone sets are available (see Table S1H in the supplemental material). Studies of these few selected organisms will allow researchers to extrapolate functional data to a large number of other genomes and organisms. We have also included *Homo sapiens* as the target of those pathogens and as a model for a higher eukaryote. As stated above, all species encode dozens, and more frequently, hundreds of DUFs awaiting functional characterization.

Discussion and conclusion. Independently of our study, the Protein Structural Initiative (PSI) has pointed out (8) that many of the domains currently without known function may be widespread in the tree of life, even if found predominantly in bacteria. This survey supports this view, as many of the most prevalent DUFs in bacteria are also found in animals, plants, and other phyla. Thus, studying bacterial DUFs is important for understanding not only microbiology but also molecular biology in general.

Many of the widespread DUFs must have important functions, even if they are not essential in standard mutant screens. For instance, DUF143, one of the most common DUFs, occurring in both bacteria and eukaryotes (but not in archaea), has been placed in the top 10 list of "unknown" proteins by Galperin and Koonin (21). Its deletion in *E. coli* showed no obvious phenotype (22). However, we recently showed that this protein is essential when cells are starved (7), a situation that is not commonly used in mutant screens in the laboratory. In fact, this function is probably conserved in all bacteria, although its role may be different in eukaryotes (where it is localized to mitochondria) (23, 24).

The functional analysis of DUFs will require concerted efforts, including crystallization, protein interaction screens, phenotyping of mutants, and more-specific functional assays. General predictions should also allow us to determine the experimental direction required to find the precise function of DUFs. For instance, DUFs predicted to be enzymes can be screened for potential substrates or activities while protein interaction domains need to be screened for interaction partners. We hope that our ranking list of DUFs will help the scientific community to find the most interesting, most important, and taxonomically most widespread DUFs to be identified and analyzed.

Methods. (i) Data sources. Domain, protein, and phylogenetic information for all kingdoms of life was obtained from the databases listed in Table S1A and Fig. S4A in the supplemental material. We specifically focused on the 1,540 bacterial, 290 eukaryotic, and 120 archaeal organisms with completely sequenced genomes represented in UniProt (version 2012_06) (25). Domains named DUFxxx where xxx is the number for the DUF or containing "unknown function" in the name were collected from Pfam (version 26.0) and make up the list of DUFs considered in this study. NCBI taxonomic identifiers associated with DUFs versus non-DUFs were obtained from UniProt. Identifiers for strains and species were mapped to higher taxonomic taxa, particularly phyla and kingdoms, for analysis and visualization (Fig. S4B). Essentiality information was obtained from the Database of Essential Genes version 8.5 (14).

(ii) Phylogenetic analysis. DUF and all-domain lists were generated for all kingdoms and phyla. Phylogenetic membership for each protein was defined by strain-specific taxonomic identifiers assigned in UniProt. DUFs/domains found in proteins belonging to a particular (sequenced) bacterial strain were said to be present in the phylum/kingdom containing the strain. Strain to phylum mapping was performed according to the NCBI hierarchy (a summary sheet for this hierarchy can be found on the NCBI taxonomy site ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/). Domain and DUF representation among 1,123 pathogenic bacterial strains recorded in PATRIC (26) was also calculated. This was achieved by adding a filtering step in the script described above, by which only proteins belonging to these PATRIC strains were used to count domains/DUFs. The results were then ranked by prevalence among sequenced bacterial genomes. Subsequent analyses focused on the top 50 DUFs according to this ranking. Representation by total genome count, total bacterial pathogen count, total protein count, structure (PDB), and protein length, was measured. A local version of the UniProt database, consisting of both Swiss-Prot and TrEMBL (UniProt releases from 3 October 2012 and 25 January 2012, respectively), was used. A bacterial pathogen was defined as a member of the 1,123 PATRIC bacterial strains that were linked to at least one disease (May 2012 release). For the PDB analysis, both Pfam-A.full and Pfam-A.seed of the Pfam database version 26.0 were used. Finally, data relating to domain and DUF counts for bacterial phyla were mapped onto pie chart data types on the iTOL website (27) using a definition file with one representative organism (selected somewhat arbitrarily) per phylum.

(iii) Functional clues. For a sample of 13 model organisms (10 bacterial and 3 eukaryotic organisms), any proteins containing one or more of the top 50 DUFs (ranked again as described above) in UniProt with functional annotation were collected. For the same proteins, the functions of partner proteins recorded in the STRING database (version 9.0) (16), a resource of experimentally or highly confidently predicted interactions, were used as a second (indirect) source of functional attributes that might possibly be associated with these DUF-containing proteins. Only STRING partners with at least a score of 700 (of the maximum 1,000) were considered. Our specifically DUF-focused analyses used proteins from only the 10 bacterial model organisms. All Gene Ontology (GO) terms accompanying each of the (full-length) proteins in UniProt were collected, then we removed GO terms associated with any non-DUF domains according to the (largely manually curated) Pfam2GO mapping on the Gene Ontology Consortium website (http://www.geneontology.org/external2go/pfam2go) (28). Any remaining GO terms were considered to be functional clues for the DUF(s) in the protein. The coverage of the Pfam2GO file was limited (~4,000 domains or ~25% of Pfam). Therefore, to avoid ambiguity of GO term assignment, no inferences were drawn from proteins with non-DUF domains not in the mapping file. This inference protocol for DUF-associated GO terms is illustrated in Fig. S5 in the supplemental material.

For the STRING-based contribution to the analysis, GO terms were collected from STRING version 9.0 for predicted functional partners of all proteins containing a DUF; no removal of non-DUF-specific GO terms was performed. GO terms found in at least 50% of all functional partners of all proteins with a particular

DUF were included as hypothetical functions for that DUF, if they were not too general. To avoid overly general GO term functions (e.g., "molecular function" or "binding"), only GO terms at a depth greater than 3 in the GO hierarchy were included. This functional inference method is illustrated in Fig. S5 in the supplemental material.

(iv) Essentiality analysis—eDUFs. The Database of Essential Genes (DEG) version 8.5 (last updated July 2013) was used to define essential proteins. Entrez GI numbers from DEG were mapped to UniProt accession numbers. UniProt was also used to provide a list of domains/DUFs for each DEG protein. Pfam annotation from a recent Pfam release (v26; September 2012) as well as an earlier release (v23; July 2008) was used to investigate how the numbers of essential DUFs change over time. The 355 DEG proteins with DUFs were analyzed to define essential DUFs. This combinatorial analysis was carried out using the following definitions for cases of essential and nonessential domains. Essential domains were defined using three cases: single-domain essential proteins, unique domains in multiple essential proteins (e.g., cases of the form A-B-C and C-D-E, where C is the inferred essential domain), and by comparison with nonessential proteins of similar domain membership (i.e., cases of the form A-B-C essential, where A and B are nonessential proteins). Nonessential domains were also defined as those that are not present in any essential proteins (case 1) or those in essential proteins only when all other domains are essential (case 2). Because defining nonessential domains helps define essential domains by removing potentially essential domains from each protein's domain composition, these 5 cases were identified iteratively until no further essential domains could be found. Finally, the DUFs among the essential domains were labeled eDUFs (see Table S1F in the supplemental material).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00744-13/-/DCSupplemental.

Figure S1, PDF file, 0.1 MB.
Figure S2, PDF file, 0.1 MB.
Figure S3, PDF file, 0.1 MB.
Figure S4, PDF file, 0.1 MB.
Figure S5, PDF file, 0.1 MB.
Table S1, XLS file, 0.8 MB.

ACKNOWLEDGMENTS

We acknowledge Ivica Letunic of EMBL Heidelberg for his assistance with iTOL.

P.U. conceived the study. N.G. and D.L.G. carried out the bioinformatics analysis. P.U., N.G., and D.L.G. wrote the manuscript.

REFERENCES

1. **Kessel A, Ben-Tal N.** 2011. Introduction to proteins. CRC Press, Boca Raton, FL.
2. **Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N.** 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. **38:** D161–D166.
3. **Mulder NJ, Kersey P, Pruess M, Apweiler R.** 2008. In silico characterization of proteins: UniProt, InterPro and Integr8. Mol. Biotechnol. **38:** 165–177.
4. **Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD.** 2012. The Pfam protein families database. Nucleic Acids Res. **40:**D290–D301.
5. **Bateman A, Coggill P, Finn RD.** 2010. DUFs: families in search of function. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. **66:** 1148–1152.
6. **Littler E.** 2010. Combinatorial domain hunting: solving problems in protein expression. Drug Discov. Today **15:**461–467.
7. **Häuser R, Pech M, Kijek J, Yamamoto H, Titz B, Naeve F, Tovchigrechko A, Yamamoto K, Szaflarski W, Takeuchi N, Stellberger T, Diefenbacher ME, Nierhaus KH, Uetz P.** 2012. RsfA (YbeB) proteins are conserved ribosomal silencing factors. PLoS Genet. **8:**e1002815. http://dx.doi.org/10.1371/journal.pgen.1002815.
8. **Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C.** 2009. PSI-2: structural genomics to cover protein domain family space. Structure **17:**869–881.
9. **Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, Gateau A, Gaudet P, Gleizes A, Masselot A, Zwahlen C, Bairoch A.** 2012. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res. **40:**D76–D83.
10. **Rajagopala SV, Yamamoto N, Zweifel AE, Nakamichi T, Huang HK, Mendez-Rios JD, Franca-Koh J, Boorgula MP, Fujita K, Suzuki K, Hu JC, Wanner BL, Mori H, Uetz P.** 2010. The Escherichia coli K-12 ORFeome: a resource for comparative molecular microbiology. BMC Genomics **11:**470. http://dx.doi.org/10.1186/1471-2164-11-470.
11. **Fonkwo PN.** 2008. Pricing infectious disease. The economic and health implications of infectious diseases. EMBO Rep. **9**(Suppl 1):S13–S17.
12. **Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY.** 2012. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. **40:**D306–D312.
13. **Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G.** 2011. A series of PDB related databases for everyday needs. Nucleic Acids Res. **39:**D411–D419.
14. **Zhang R, Lin Y.** 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. **37:**D455–D458.
15. **Schlicker A, Huthmacher C, Ramírez F, Lengauer T, Albrecht M.** 2007. Functional evaluation of domain–domain interactions and human protein interaction networks. Bioinformatics **23:**859–865.
16. **Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C.** 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. **39:**D561–D568.
17. **Lopez D, Pazos F.** 2009. Gene ontology functional annotations at the structural domain level. Proteins Struct. Funct. Bioinformatics **76:** 598–607.
18. **Fang H, Gough J.** 2013. A domain-centric solution to functional genomics via dcGO Predictor. BMC Bioinformatics **14**(Suppl 3):S9. http://dx.doi.org/10.1186/1471-2105-14-S3-S9.
19. **Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, Sangrador-Vegas A, Yong S-Y, Mulder N, Hunter S.** 2012. Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database 2012:bar068.
20. **de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, Chothia C, Gough J.** 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res. **39:**D427–D434.
21. **Galperin MY, Koonin EV.** 2010. From complete genome sequence to "complete" understanding? Trends Biotechnol. **28:**398–406.
22. **Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H.** 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. **2:**2006.0008. http://dx.doi.org/10.1038/msb4100050.
23. **Rorbach J, Gammage PA, Minczuk M.** 2012. C7orf30 is necessary for biogenesis of the large subunit of the mitochondrial ribosome. Nucleic Acids Res. **40:**4097–4109.
24. **Wanschers BF, Szklarczyk R, Pajak A, van den Brand MA, Gloerich J, Rodenburg RJ, Lightowlers RN, Nijtmans LG, Huynen MA.** 2012. C7orf30 specifically associates with the large subunit of the mitochondrial ribosome and is involved in translation. Nucleic Acids Res. **40:**4040–4051.
25. **UniProt Consortium.** 2012. Reorganizing the protein space at the Uni-

versal Protein Resource (UniProt). Nucleic Acids Res. 40:D71–D75. doi: 10.1093/nar/gkr981.

26. **Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, Nordberg EK, Scott M, Schulman JR, Snyder EE, Sullivan DE, Wang C, Warren A, Williams KP, Xue T, Yoo HS, Zhang C, Zhang Y, Will R, Kenyon RW, Sobral BW.** 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect. Immun. **79:**4286–4298.
27. **Letunic I, Bork P.** 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23:**127–128.
28. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.** 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25:** 25–29.