Title: Accelerating biological insight for understudied genes

Authors: Kimberly A. Reynolds[*], Eduardo Rosa-Molinar[†], Robert E. Ward[‡], Hongbin Zhang[§], Breeanna R. Urbanowicz[‖], A. Mark Settles[¶,1]

[*]The Green Center for Systems Biology and the Department of Biophysics, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[†]Department of Pharmacology & Toxicology, The University of Kansas, Lawrence, KS 66047, USA

[‡]Department of Biology, Case Western Reserve University, Cleveland, OH 44106, USA

[§]Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843, USA

[‖]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia 30602, USA

[¶]Bioengineering Branch, NASA Ames Research Center, Moffett Field, CA, USA

K.A.R and E.R.M. contributed equally to this work.

[1]Corresponding author; E-mail: andrew.m.settles@nasa.gov; Tel. 352-283-2767; Fax. none

Running title: Biology of understudied genes

Manuscript text (including abstract): 4,489 words

Abstract

The rapid expansion of genome sequence data is increasing the discovery of protein-coding genes across all domains of life. Annotating these genes with reliable functional information is necessary to understand evolution, to define the full biochemical space accessed by nature, and to identify target genes for biotechnology improvements. The vast majority of proteins are annotated based on sequence conservation with no specific biological, biochemical, genetic, or cellular function identified. Recent technical advances throughout the biological sciences enable experimental research on these understudied protein-coding genes in a broader collection of species. However, scientists have incentives and biases to continue focusing on well documented genes within their preferred model organism. This perspective suggests a research model that seeks to break historic silos of research bias by enabling interdisciplinary teams to accelerate biological functional annotation. We propose an initiative to develop coordinated projects of collaborating evolutionary biologists, cell biologists, geneticists, and biochemists that will focus on subsets of target genes in multiple model organisms. Concurrent analysis in multiple organisms takes advantage of evolutionary divergence and selection, which causes individual species to be better suited as experimental models for specific genes. Most importantly, multisystem approaches would encourage transdisciplinary critical thinking and hypothesis testing that is inherently slow in current biological research.

Introduction

In the last 20 years, our capacity to sequence genomes has grown exponentially. NCBI includes well over 200,000 genomes that have discovered more than 150 million non-redundant protein sequences (Li et al. 2021). Despite this rich catalog of information, the biological function of most genes remains unknown. Currently, we consider a protein sequence to be “functionally annotated” when the sequence shows homology to other protein-coding genes that have an assigned function or if the protein contains a conserved domain. These annotations often take the form of Gene Onotology (GO) terms (Ashburner et al. 2000; Gene Ontology Consortium 2021). While the number of protein sequences in UniProtKB has grown exponentially, the number of entries with experimentally supported GO terms has grown linearly (Cozetto and Jones 2016). As a consequence, many genomes remain only partly annotated, even when considering a fairly low bar for defining function. For example, one recent study found that between 52-79% of the average bacterial proteome could be functionally annotated through sequence homology based searches (Lobb et al. 2020).

The predicted proteome of the human genome is arguably the most intensely studied of all organisms, yet 9.6% of the Human Proteome Project’s entries have no direct evidence that the protein is expressed (Adhikari et al. 2020). Intrinsically Disordered Proteins are estimated to comprise 15-40% of the human proteome, but variations in sequence length and composition often preclude alignment, while the lack of a folded domain prohibits the use of structural information to propose annotations based on conservation (Tompa 2012). The genome sequence for the most intensely studied model plant, Arabidopsis thaliana, was released more than 20 years ago, yet nearly 10% of the protein coding genes (2,253 total) lack functional annotation (Cheng et al. 2016). .

Providing accurate functional information for individual proteins is critical to discover new chemical, biochemical, and cellular processes. It is also fundamental to understand genotype to phenotype relationships resulting from variation in protein expression or sequence. Traditionally,

protein function is defined through a one-gene-at-a-time approach requiring years of intensive research to associate a biochemical, cellular, or organismal role. The conventional approach of studying protein function suffers from our own human tendencies to focus attention and effort on a small fraction of the proteome. Well-known proteins are over-studied, while proteins of unknown or poorly defined function are relegated to supplementary tables in -omics publications. This “Matthew effect,” in which the information rich proteins get richer, is illustrated by the distribution of papers that document protein function from NCBI’s gene2pubmed database (Carter et al., 2019). Publications showing functional attributes of genes have a power law distribution with >22% of papers focused on only 1% of human genes (Zwick et al. 2019). The bias in annotation in multiple model species impedes identification of promising drug targets associated with human disease (Haynes et al. 2018).

A related problem is that when biologists do investigate genes of unknown function, consensus may be difficult to achieve for an individual protein’s function. For example, over the last decade multiple groups have studied genes that encode proteins containing the conserved “Domain of Unknown Function 490,” which has been renamed TamB. Despite extensive study in two domains of life, TamB protein function is only partially resolved. The NCBI Conserved Domain Database, annotates the TamB domain as limited to bacteria, yet at least 119 plant species have proteins with the complete domain.

In *E. coli* and other bacteria, tamB and related genes have been associated with outer membrane protein biogenesis and the protein has a partial crystal structure (Selkrig et al., 2012; Selkrig et al. 2015; Josts et al. 2017; Yu et al. 2017; Iqbal et al. 2016; Bialer et al. 2019; Li et al. 2020). TamB proteins have also been studied experimentally in rice, *Arabidopsis thaliana*, and maize. Plant mutants show defects in chloroplasts and other plastid types. In maize, the TamB domain protein appears to have similar function to *E. coli* by promoting outer envelope biogenesis in plastids (Zhang et al 2019). The protein is hypothesized to function in plastid division in rice (Matsushima et al. 2014), and in the chloroplast protein import machinery in

Arabidopsis (Chen et al. 2018). Remote homology predictions of the N-termini of TamB proteins suggest yet another conserved domain that argues for a function in lipid transfer between membranes instead of a direct role in outer membrane protein assembly (Levine, 2019).

Like the parable of the blind men and the elephant, these independent discoveries in multiple model systems led to conflicting interpretations. Nevertheless, all proteins containing a TamB domain in more than 2,500 species are currently counted as "functionally annotated." This example illustrates the challenge of assigning a biochemical and cellular role to just one of the thousands of conserved domains with limited experimental evidence.

TamB proteins are not unique. The vast majority of protein functional annotations are based solely on sequence homology, often derived from superfamilies of conserved domains (Lu et al., 2020). Automated, homology-based predictions of biological function can give a reasonable hypothesis for gene function, but often there are significant annotation errors that are propagated from genome to genome (Promponas et al. 2015). Therefore, it is imperative to develop approaches to determine the biological functions of understudied genes more rapidly.

Recent advances in defining gene function

Biologists take diverse approaches to defining protein-coding gene functions. Historically, scientists considered a gene function to be defined once the DNA corresponding to a trait was cloned into a recombinant DNA vector and proven to control the trait (Cohen et al. 1973). Genetics, biochemistry, molecular biology, or cell biology experiments are used to provide the experimental evidence supporting the role of the cloned DNA in top level Gene Ontology classes of molecular function, biological process, and cellular component.

More recently, improvements in DNA sequencing technologies have transformed quantitative genetics by enabling genome-wide association studies (GWAS) in virtually any organism (Tibbs Cortes et al. 2021). GWAS takes advantage of large-scale genotyping and phenotyping of populations to statistically associate molecular markers with a quantitative trait.

Based on PubMed searches, over 30,000 human GWAS studies have been published in the last ten years and hundreds of studies have been published for each model organism. Each study produces a series of candidate genes that are associated with diverse traits like forage digestibility or sleep disorders. Validating a GWAS association still requires additional experimental evidence, and regrettably associations with understudied genes are typically low priority for researchers (Haynes et al. 2018).

Advances in experimental and computational technologies have greatly increased the capacity for experimental validation of gene-phenotype associations. Here we overview several key disciplines for defining protein function and highlight promising advances and challenges for coordinated transdisciplinary projects to dissect roles of understudied gene families.

Informatics: The rapid increase in genomic “big data” provides rich opportunities for improving prediction of function by statistically-driven evolutionary analyses and machine learning. While a complete discussion of the rich literature on computational function prediction is outside the scope of this review, there are many well established signals of conserved function beyond primary protein sequence. Synteny or conserved chromosomal proximity, phylogenetic profiling for the presence/absence of genes, and protein sequence co-evolution can all aid in predictions of function and functional interactions (Pellegrini et al. 1999; Junier and Rivoire 2016; Rivoire 2013; Kim et al. 2011; Engelhardt et al. 2011; Janga et al. 2005; Huynen et al. 2000; Snel et al. 2002; Szklarczyk et al. 2019). For more than 25 years, Critical Assessment of protein Structure Prediction experiments have gauged the state of the art in protein modeling (Kryshtafovych, 2019). The steady increase in protein structural data has enabled homology-based structural and predictions of inter-residue distances approaches to become increasingly powerful (Yang et al. 2015; Zimmermann et al. 2018). Advances in machine learning and artificial intelligence methods are also dramatically increasing the potential of computational annotation prediction of accurate protein structures (Bileschi et al. 2019; Senior et al., 2020; Callaway et al., 2020). Co-evolutionary analyses suggest the

possibility that cellular systems might be decomposed into small multi-gene functional units (Rivoire et al. 2013; Schober et al. 2019). In this sense, we might recognize or annotate a small group of genes as a physical complex or the unit underlying a particular phenotypic trait.

To take full advantage of these computational approaches, gold-standard training and test set data need to be improved. Expanding the availability of high-quality experimental functional annotations is important to refining and validating the performance of many computational algorithms. For example, the community-driven Critical Assessment of Functional Annotation challenge has been valuable in benchmarking progress and identifying directions requiring further work (Zhou et al. 2019). Increasing the annotations available would further accelerate progress. The highest quality gene annotations are often within a species-centric community database that does not interface well with Genbank, EMBL, and DDBJ. Improving communication between smaller research communities and large clearing house databases would help annotation problems. Automated text processing of peer-reviewed scientific literature to assign GO terms to genes or even to categorize scientific literature is an area of research that has been attempted but has not yet been accepted broadly (Di Len et al. 2015; Van Auken et al. 2014). Further research in natural language processing may be able to reduce the need for manual curation of gene functions.

An additional limitation is that computational annotation methods are typically used by a limited set of expert practitioners. Both the code and analysis results are sometimes inaccessible or restricted in scope to experimental biologists. Like other statistical methods, it is often difficult for experimental biologists to know which tools are the most reliable and appropriate for their application. Therefore, more direct collaboration between experts in computation and experimental biologists is needed. Usage of well-established methods could be further democratized through open-access initiatives that develop consistent standards and interfaces for web-based annotation tools.

Genetics: Mutant alleles allow powerful insights into gene function through observation of the contrasting phenotypes of normal and mutant organisms. Generating knockout collections for individual organisms is a mature technology. Large-scale collections are available for a diversity of model organisms (e.g. Giaever et al. 2002; Bolle et al. 2012; Ramírez-Solis et al. 2012; Varshney et al. 2013; McCarty et al. 2013; Li et al. 2019).

The standard paradigm for microbes and cell cultures has been to measure fitness phenotypes for all knockouts across a gauntlet of controlled environmental conditions. In these types of studies, phenotypes are measured across many growth conditions with the goal of exposing as many slow growth phenotypes as possible (Deutschbauer et al. 2014; Hillenmeyer et al. 2008; Price et al. 2018; Jaffe et al. 2019). However, most genome-wide fitness experiments focus on only one species. Cross species comparisons could provide additional power to detect phenotypes for conserved protein sequences. To this end, recent advances in multiplexed continuous culture devices provide one avenue for broad taxonomic sampling of microbial mutant responses to well-controlled environments (Wong et al. 2018; Toprak et al. 2013).

Single gene and whole genome duplications allow evolution of neo-functionalized genes but also create genetic redundancy. When paralogs retain identical or similar functions, phenotypes may only be observed when two or more genes are simultaneously mutated. Synthetic screens for viability, fitness, or growth have been developed to identify multi-gene knockout phenotypes in microbes, multicellular organisms, and in tumor cells (Klobuchar and Brown 2018; Kuzmin et al. 2020; Zhan and Boutros, 2016). Synthetic screens can reveal phenotypes for genes that are not obvious paralogs allowing high throughput, unbiased discovery of functional redundancy.

Heterologous expression of proteins across species also provides insight on protein function. Since many organisms cannot be grown in a laboratory, it is not possible to study genes with unknown or understudied functions from non-culturable species through direct

genetic manipulation. Expression of understudied and unknown function genes can be used to rescue loss-of-function phenotypes in model organisms (Thiaville et al. 2016; Kim et al. 2010). The throughput of gene complementation studies with heterologous protein expression could be increased by developing libraries of unique genes discovered solely through sequencing diverse and non-culturable species.

Multicellular organisms increase the complexity for genetic manipulation and phenotypic analysis of mutants. The multiple cell types, organs, and complex responses to environment require specialized expertise to adequately interpret phenotypes of homologous genes from divergent species. In addition, there are currently few multicellular organism models that approach genome-wide saturation of gene knockout collections.

The development of CRISPR-based tools has greatly enhanced genetic studies by allowing site-directed mutagenesis (Doudna and Charpentier 2014). In model systems, CRISPR has been adapted to a wide variety of genetic problems, from genome-scale forward genetic screens in human tissue culture to targeted saturation mutagenesis of individual genes in in the crop model rice (Joung et al., 2017; Li et al., 2020). Moreover, CRISPR technologies are applicable broadly to non-model species from marine animals like sea anemones to horticultural crops (Nakanishi and Martindale 2018; Erpen-Dalla Corte et al. 2019). This enables genetic analysis of understudied genes that can be based on levels of genome redundancy as well as developmental and physiological characteristics that make a species a better model for understanding a specific biological process. Genetic redundancy within individual species can require multiple members of gene families to be mutated before any discernible phenotype can be observed, such as the highly redundant families of ethylene and abscisic acid hormone receptors (Hall and Bleecker 2003; Zhao et al. 2018). For gene families, a species with fewer redundant copies would be more accessible for subsequent genetic analysis.

Similarly, developmental differences can allow some species to be better models for the study of conserved molecular processes. For example, maize endosperm is more tolerant of

minor intron splicing defects than human cell cultures or the model plant, Arabidopsis (Gault et al. 2017; Bai et al. 2019). Multispecies comparisons across domains of life would take full advantage of the power of gene editing technologies. Interpreting phenotypes from diverse organisms requires experts who have familiarity with individual or phylogenetically related species. The rapid expansion of technologies enabling genetic studies in almost any organism is both an opportunity to increase interdisciplinary collaborations and a challenge due to the historic isolation of biological research communities. A major barrier for communicating genetic results is developing species agnostic terminology and databases to make gene to phenotype relationships machine readable and more accessible across disciplines.

Biochemistry: Annotation of biochemical function should go beyond vague predictions of potential top-level Enzyme Commission numbers. Even within yeast and human genomes, >30% of proteins are estimated to be lacking a clear biochemical function (Ellens et al. 2017). Specific knowledge of reactions catalyzed, associated catalytic parameters, and substrate specificity or ligand binding affinities are fundamental to identifying new chemistry, defining metabolic pathways, and discovering drug targets. Nonetheless, expressing, purifying and biochemically characterizing individual proteins is labor-intensive and typically low throughput. Combining comparative genomics, structural modeling and high-throughput biochemical assays suggests a path towards broadly defining enzyme function (Bastard et al. 2014; Zhao et al. 2013), but requires additional scaling to keep pace with the expansion in genomic datasets.

Recent advances in microfluidics-based assays, which can enable measurements of catalytic parameters for tens of thousands of enzyme variants, provide one strategy for rapid characterization of homologs or mutants. In these systems, proteins are often transcribed and translated *in vitro*, and measurements of catalytic activity, binding affinity and substrate specificity are collected within isolated microfluidic chambers or droplets (Maxutis et al. 2013; Fordyce et al. 2010). However, these systems are presently limited to relatively well-behaved model enzymes, and typically require known fluorescent substrates. Expanding the utility of

microfluidics approaches would be an enabling technology for annotation of biochemical function and protein engineering.

As throughput is increased and novel enzymatic activities are defined, it becomes imperative that database management expand to allow biochemists to analyze larger data sets. Similar to sequence data, public databases for enzyme properties should be in machine readable format with appropriate meta-data. The BRENDA database (https://www.brenda-enzymes.org/) provides one existing platform for accomplishing this, but the data are not easily searched or downloaded in a readily machine readable format. KEGG provides an organized view of metabolic enzymes and pathway structure (https://www.genome.jp/kegg/), but does not make biochemical parameters for enzyme orthologs readily available.

Cell biology: Cellular imaging and fractionation are critical for determining protein function in the larger context of the cell and organism. The subcellular localization of a protein gives some of the most informative clues about an unknown or understudied protein's molecular function and biological process. Imaging technologies can reveal incredible detail and resolution with the power to test protein-protein interactions. For example, protein-protein interaction imaging using Förster resonance energy transfer (FRET) and fluorescence cross-correlation spectroscopy have been developed into a vast array of applications such as single-molecule FRET; FRET-fluorescence lifetime imaging microscopy; single-molecule protein proximity index; concentric FRET; homogeneous time resolved fluorescence; acceptor photo-bleaching FRET, and correlative acceptor photo-bleaching FRET (Periasamy and Day 1999; Zaccolo 2004; Ciruela 2008; Bucherl et al. 2010; De Los Santos et al. 2015; Shrestha et al. 2015; Geißler and Hildebrandt 2016; Chenab et al. 2019; Tsai et al. 2019). However, most imaging reagents and technologies have been developed for highly abundant, well characterized proteins with stable interaction partners in a limited number of model systems (Daniels et al. 2003; Costes et al. 2004; Bolte et al. 2006; Snapp and Hedge 2006; Comeau et al. 2008). Imaging understudied proteins and intrinsically disordered proteins pose significant challenges due to limitations in

readily available, protein-specific reagents (Perdigão et al. 2015; Lieutaud et al. 2016; Perdigão and Rosa 2019).

There has been massive effort to develop antibody reagents to characterize the human proteome with many antibodies developed commercially (Uhlén and Ponten 2005; Uhlén et al. 2010, 2015). Consequently, the Human Proteome Project tracks more than 4 million antibodies against human proteins (Adhikari et al. 2020). However, few other organisms have garnered as much attention, and there are in fact no organized antibody databases for any plant system. In many organisms, and for understudied proteins in virtually all organisms, antibody reagents are simply not available. In addition, antibodies show significant variation by nature and the quality of an antibody required for different methodologies and workflows can also be variable (Burry 2011; Hewitt et al. 2014; Gautron 2019).

Epitope and fluorescent tags enable studies for proteins without antibody reagents, but most established technologies are labor intensive and poorly scalable. Moreover, efficient methods for tagging proteins and imaging cells with light microscopy are only available for a few model species. For mammalian cells, recent advances in CRISPR-mediated gene editing to insert tags is promising (Thöne et al. 2019). By contrast, genetic lines with fluorescent protein tags are limiting in systems with poor transformation, such as maize (Wu et al. 2013). Better probe design and tagging methods are clearly needed to enable protein assignments to cellular compartments in diverse species.

Enzyme catalyzed proximity labelling is a promising approach to study protein interaction networks at the subcellular level (Branon et al., 2018; Mair et al., 2019; Cho et al., 2020). Proximity labelling uses an engineered biotin ligase fused to a protein of interest, which then labels interacting proteins. Biotinylation allows the purification of labelled proteins with streptavidin for identification with standard proteomic methods (Branon et al., 2018). This technology can be applied to any species that supports expression of recombinant proteins and

would enable comparisons of interacting partners across species to give insight on conserved *in vivo* protein-protein interactions.

A multidisciplinary, multispecies initiative

Annotating protein function is a long-standing problem in biology (Roberts 2004; Gerlt et al. 2011; Earnshaw et al. 2013). As the Arabidopsis genome was nearing completion in 2000, the National Science Foundation initiated the Arabidopsis 2010 program with the goal of discovering the function of every gene in this model plant's genome (Somerville and Dangl 2000). Nevertheless, the scientific community has still not accomplished the goal of understanding all gene functions in *any* genome, yet this basic information is critical to advance all fields of biology.

Given the large scope and complex nature of the problem, we suggest public research initiatives that promote collaborations between experimentalists specializing in different species and disciplines. We envision inter-species projects that would collect data testing the function of understudied proteins at scales greater than one gene at a time, but lower than traditional high-throughput functional genomics studies. This mid-level scale would help customize reductionist experiments to align with understudied protein types. Requiring multispecies studies would engage experts from different scientific communities to evaluate data from diverse fields as it is being generated allowing faster iterations between generating a hypothesis, testing, and revising to the next experiment. This type of structure would also pair scientists who can make genetic progress in one model species with scientists who can make cell biology or biochemistry progress in another species.

The collected experimental information should be compiled in a publicly accessible, searchable database. Making these data available in machine-readable formats with appropriate metadata is important for the training and validation of computational algorithms to propagate conclusions about gene function accurately throughout other genome sequences.

Existing pipelines like the Enzyme Function Initiative provide an exemplar for extending and refining annotation efforts (Zallot et al. 2018; Zallot et al. 2019).

Even at this intermediate scale, deciding which of the vast number of functionally unknown or understudied genes to pursue is challenging. Several types of annotation data that are currently available could be used to narrow the focus of individual projects. For example, evolutionary conservation and genetic redundancy can be used to prioritize subsets of understudied genes. Genes that are found in organisms from multiple domains of life are likely to have more fundamental scientific impact, whereas avoiding genetic redundancy expedites experimental analysis. Moreover. genes that have been identified as essential in more distantly related taxa are more likely to have critical biological functions. We provide a few examples of additional annotation types that could be used to bring realistic scale projects together.

Co-evolution: Understudied protein families are found at different scales of phylogenetic conservation. It is possible to sort conserved proteins by taxa to identify proteins that should function in a cell compartment or metabolic process that is uniquely conserved among the subset of species. As an example, proteins conserved only in bacteria and plants, like the TamB family, generally point to functions conserved in the plant endosymbiotic organelles and in bacteria. A research project to accelerate gene function definition for this pattern would need to have biologists familiar with representative bacterial and plant species. Other co-evolutionary patterns would require different subsets of representative species depending upon unique patterns of taxonomic group co-evolution.

Genetic associations: The massive number of GWAS studies has generated hypotheses for an unknown number of understudied genes that have been relegated to supplementary data in publications. Capturing and analyzing these associations across taxa would provide new targets that show consistent phenotypes in multiple species. Initially data scientists are needed to extract the associations with understudied genes to find gene families that show evidence of

genetic phenotypes in multiple species. Collaborative research teams would focus on analyzing the understudied genes in multiple systems in parallel.

Correlative -omics data: Combining existing co-expression networks, yeast two-hybrid interaction data, and subcellular targeting predictions would identify understudied proteins that show associations with known proteins. Model systems that are best suited to study the implicated processes would identify collaborative research groups.

Broader Impacts and Educational Outcomes:

Many experiments in biology are amenable to science crowd-sourcing and undergraduate laboratory curricula. For example, there are several examples of successful GO term annotation from science community-based efforts as well as undergraduate bioinformatics courses (Antonazzo et al. 2020; Lovering 2017). We envision the possibility of nationwide laboratory courses that assign small groups of students to gene products or protein domains of interest. These proteins can then be characterized using a small set of pre-defined and relatively straightforward assays, including growth rate complementation studies, microscopy-based characterization of localization, assays of cellular phenotypes for knockouts or knockdowns, as well as expression and purification trials. Importantly, these experimental efforts can be combined with informatics and comparative genomics analyses, providing a rich opportunity for interdisciplinary cross-training.

Summary and Potential Impacts

Increasing the throughput and information content of protein annotations is essential to enabling synthetic biology and metabolic engineering, discovering new drug targets, and most fundamentally, understanding the molecular basis of phenotype. Recent advances in high throughput phenotyping, imaging, and biochemistry indicate the potential for high-dimensional annotations of function. Twenty years of large-scale genomics studies has made amazing

progress in understanding gene function, but there is still a significant fraction of the biological universe that is essentially unknown. It is time to go beyond simple classifications or categories for proteins that can be achieved by traditional computational and functional genomics strategies. However, single gene studies in individual model organisms are too slow and narrowly focused to adequately address the explosion of genome sequence. A research initiative that encourages greater integration of experimental research on understudied genes would provide rich annotations of protein localization, physiological roles in multiple species, and quantitative biochemical parameters. Collecting these data into a single well-maintained and well-organized database would greatly improve computational efforts to predict function.

Author Contributions

K.A.R., E.R.M., and A.M.S. developed the concept and wrote the manuscript. R.E.W. and B.R.U. developed the concept and edited the manuscript. All authors approved of the final version of the manuscript.

Funding

This work was funded by the National Science Foundation “Reintegrating Biology Jumpstarts” award 1940791. Research support includes: the Gordon and Betty Moore Foundation Data Driven Discovery Initiative award GBMF4557 to K.A.R.; National Science Foundation award IOS-2111069 to R.E.W.; United States Department of Energy Center for Bioenergy Innovation and grant award DE-SC0015 to B.R.U.; National Institute of Food and Agriculture award 2018-51181-28419 to A.M.S.; Florida Space Research Institute award UF-AWD 98313 to A.M.S.

References

Adhikari S, Nice EC, Deutsch EW, Lane L, Omenn GS, Pennington SR, Paik YK, Overall CM, Corrales FJ, Cristea IM et al. 2020. A high-stringency blueprint of the human proteome. Nat Commun. 11(1):5301.

Antonazzo G, Urbano JM, Marygold SJ, Millburn GH, Brown NH. 2020. Building a pipeline to solicit expert knowledge from the community to aid gene summary curation. Database. 2020:baz152.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet. 25(1):25-29.

Bai F, Corll J, Shodja DN, Davenport R, Feng G, Mudunkothge J, Brigolin CJ, Martin F, Spielbauer G, Tseung CW et al. 2019. Rna binding motif protein 48 is required for u12 splicing and maize endosperm differentiation. Plant Cell. 31(3):715-733.

Bastard K, Smith AA, Vergne-Vaxelaire C, Perret A, Zaparucha A, De Melo-Minardi R, Mariage A, Boutard M, Debard A, Lechaplais C et al. 2014. Revealing the hidden functional diversity of an enzyme family. Nat Chem Biol. 10(1):42-49.

Bialer MG, Ruiz-Ranwez V, Sycz G, Estein SM, Russo DM, Altabe S, Sieira R, Zorreguieta A. 2019. Mapb, the brucella suis tamb homologue, is involved in cell envelope biogenesis, cell division and virulence. Sci Rep. 9(1):2158.

Bileschi ML, Belanger D, Bryant D, Sanderson T, Carter B, Sculley D, DePristo MA, Colwell LJ. 2019. Using deep learning to annotate the protein universe. bioRxiv.626507.

Bolle C, Schneider A, Leister D. 2011. Perspectives on systematic analyses of gene function in arabidopsis thaliana: New tools, topics and trends. Curr Genomics. 12(1):1-14.

Bolte S, Cordelières FP. 2006. A guided tour into subcellular colocalization analysis in light microscopy. J Microsc. 224(Pt 3):213-232.

Branon TC, Bosch JA, Sanchez AD, Udeshi ND, Svinkina T, Carr SA, Feldman JL, Perrimon N, Ting AY. 2018. Efficient proximity labeling in living cells and organisms with turboid. Nat Biotechnol. 36(9):880-887.

Bücherl C, Aker J, de Vries S, Borst JW. 2010. Probing protein-protein interactions with fret-flim. Methods Mol Biol. 655:389-399.

Bunt G, Wouters FS. 2017. Fret from single to multiplexed signaling events. Biophys Rev. 9(2):119-129.

Burry RW. 2011. Controls for immunocytochemistry: An update. J Histochem Cytochem. 59(1):6-12.

Callaway E. 2020. 'it will change everything': Deepmind's ai makes gigantic leap in solving protein structures. Nature. 588(7837):203-204.

Carter AJ, Kraemer O, Zwick M, Mueller-Fahrnow A, Arrowsmith CH, Edwards AM. 2019. Target 2035: Probing the human proteome. Drug Discov Today. 24(11):2111-2115.

Chen YL, Chen LJ, Chu CC, Huang PK, Wen JR, Li HM. 2018. Tic236 links the outer and inner membrane translocons of the chloroplast. Nature. 564(7734):125-129.

Chenab KK, Eivazzadeh-Keihan R, Maleki A, Pashazadeh-Panahi P, Hamblin MR, Mokhtarzadeh A. 2019. Biomedical applications of nanoflares: Targeted intracellular fluorescence probes. Nanomedicine. 17:342-358.

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: A complete reannotation of the arabidopsis thaliana reference genome. Plant J. 89(4):789-804.

Cho KF, Branon TC, Udeshi ND, Myers SA, Carr SA, Ting AY. 2020. Proximity labeling in mammalian cells with turboid and split-turboid. Nat Protoc. 15(12):3971-3999.

Ciruela F. 2008. Fluorescence-based methods in the study of protein-protein interactions in living cells. Curr Opin Biotechnol. 19(4):338-343.

Cohen SN, Chang AC, Boyer HW, Helling RB. 1973. Construction of biologically functional bacterial plasmids in vitro. Proc Natl Acad Sci U S A. 70(11):3240-3244.

Colón-Ramos DA, La Riviere P, Shroff H, Oldenbourg R. 2019. Transforming the development and dissemination of cutting-edge microscopy and computation. Nat Methods. 16(8):667-669.

Comeau JW, Kolin DL, Wiseman PW. 2008. Accurate measurements of protein interactions in cells via improved spatial image cross-correlation spectroscopy. Mol Biosyst. 4(6):672-685.

Costes SV, Daelemans D, Cho EH, Dobbin Z, Pavlakis G, Lockett S. 2004. Automatic and quantitative measurement of protein-protein colocalization in live cells. Biophys J. 86(6):3993-4003.

Cozzetto D, Jones DT. 2017. Computational methods for annotation transfers from sequence. Methods Mol Biol. 1446:55-67.

Daniels G, Jenkins R, Bradshaw D, Andrews D. 2003. Resonance energy transfer: The unified theory revisited. Journal of Chemical Physics. 119(4):2264-2274.

De Los Santos C, Chang CW, Mycek MA, Cardullo RA. 2015. Frap, flim, and fret: Detection and analysis of cellular dynamics on a molecular scale using fluorescence microscopy. Mol Reprod Dev. 82(7-8):587-604.

Deutschbauer A, Price MN, Wetmore KM, Tarjan DR, Xu Z, Shao W, Leon D, Arkin AP, Skerker JM. 2014. Towards an informative mutant phenotype for every bacterial gene. J Bacteriol. 196(20):3643-3655.

Di Lena P, Domeniconi G, Margara L, Moro G. 2015. Gota: Go term annotation of biomedical literature. BMC Bioinformatics. 16:346.

Doudna JA, Charpentier E. 2014. Genome editing. The new frontier of genome engineering with crispr-cas9. Science. 346(6213):1258096.

Earnshaw WC. 2013. Deducing protein function by forensic integrative cell biology. PLoS Biol. 11(12):e1001742.

Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. 2017. Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Res. 45(20):11495-11514.

Engelhardt BE, Jordan MI, Srouji JR, Brenner SE. 2011. Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res. 21(11):1969-1980.

Erpen-Dalla Corte L, M Mahmoud L, S Moraes T, Mou Z, W Grosser J, Dutt M. 2019. Development of improved fruit, vegetable, and ornamental crops using the crispr/cas9 genome editing technique. Plants (Basel). 8(12).

Fordyce PM, Gerber D, Tran D, Zheng J, Li H, DeRisi JL, Quake SR. 2010. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol. 28(9):970-975.

Gault CM, Martin F, Mei W, Bai F, Black JB, Barbazuk WB, Settles AM. 2017. Aberrant splicing in maize. Proc Natl Acad Sci U S A. 114(11):E2195-E2204.

Gautron L. 2019. On the necessity of validating antibodies in the immunohistochemistry literature. Front Neuroanat. 13:46.

Geißler D, Hildebrandt N. 2016. Recent developments in förster resonance energy transfer (fret) diagnostics using quantum dots. Anal Bioanal Chem. 408(17):4475-4483.

Gene Ontology Consortium. 2021. The gene ontology resource: Enriching a gold mine. Nucleic Acids Res. 49(D1):D325-D334.

Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W et al. 2011. The enzyme function initiative. Biochemistry. 50(46):9950-9962.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B et al. 2002. Functional profiling of the saccharomyces cerevisiae genome. Nature. 418(6896):387-391.

Hall AE, Bleecker AB. 2003. Analysis of combinatorial loss-of-function mutants in the arabidopsis ethylene receptors reveals that the ers1 etr1 double mutant has severe developmental defects that are ein2 dependent. Plant Cell. 15(9):2032-2041.

Haynes WA, Tomczak A, Khatri P. 2018. Gene annotation bias impedes biomedical research. Sci Rep. 8(1):1362.

Hewitt SM, Baskin DG, Frevert CW, Stahl WL, Rosa-Molinar E. 2014. Controls for immunohistochemistry: The histochemical society's standards of practice for validation of immunohistochemical assays. J Histochem Cytochem. 62(10):693-697.

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D et al. 2008. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. Science. 320(5874):362-365.

Huynen M, Snel B, Lathe W, Bork P. 2000. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. Genome Res. 10(8):1204-1210.

Iqbal H, Kenedy MR, Lybecker M, Akins DR. 2016. The tamb ortholog of borrelia burgdorferi interacts with the β-barrel assembly machine (bam) complex protein bama. Mol Microbiol. 102(5):757-774.

Jaffe M, Dziulko A, Smith JD, St Onge RP, Levy SF, Sherlock G. 2019. Improved discovery of genetic interactions using crispriseq across multiple environments. Genome Res. 29(4):668-681.

Janga SC, Collado-Vides J, Moreno-Hagelsieb G. 2005. Nebulon: A system for the inference of functional relationships of gene products from the rearrangement of predicted operons. Nucleic Acids Res. 33(8):2521-2530.

Josts I, Stubenrauch CJ, Vadlamani G, Mosbahi K, Walker D, Lithgow T, Grinter R. 2017. The structure of a conserved domain of tamb reveals a hydrophobic β taco fold. Structure. 25(12):1898-1906.e1895.

Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, Sanjana NE, Zhang F. 2017. Genome-scale crispr-cas9 knockout and transcriptional activation screening. Nat Protoc. 12(4):828-863.

Junier I, Rivoire O. 2016. Conserved units of co-expression in bacterial genomes: An evolutionary insight into transcriptional regulation. PLoS One. 11(5):e0155740.

Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. 2010. Three serendipitous pathways in e. Coli can bypass a block in pyridoxal-5'-phosphate synthesis. Mol Syst Biol. 6:436.

Kim PJ, Price ND. 2011. Genetic co-occurrence network across sequenced microbes. PLoS Comput Biol. 7(12):e1002340.

Klobucar K, Brown ED. 2018. Use of genetic and chemical synthetic lethality as probes of complexity in bacterial cell systems. FEMS Microbiol Rev. 42(1).

Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. 2019. Critical assessment of methods of protein structure prediction (casp)-round xiii. Proteins. 87(12):1011-1020.

Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O et al. 2020. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. Science. 368(6498).

Levine TP. 2019. Remote homology searches identify bacterial homologues of eukaryotic lipid transfer proteins, including chorein-n domains in tamb and asma and mdm31p. BMC Mol Cell Biol. 20(1):43.

Li C, Zhang R, Meng X, Chen S, Zong Y, Lu C, Qiu JL, Chen YH, Li J, Gao C. 2020. Targeted, random mutagenesis of plant genes with dual cytosine and adenine base editors. Nat Biotechnol. 38(7):875-882.

Li MF, Jia BB, Sun YY, Sun L. 2020. The translocation and assembly module (tam) of. Front Microbiol. 11:1743.

Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS et al. 2021. Refseq: Expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res. 49(D1):D1020-D1028.

Li X, Patena W, Fauser F, Jinkerson RE, Saroussi S, Meyer MT, Ivanova N, Robertson JM, Yue R, Zhang R et al. 2019. A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis. Nat Genet. 51(4):627-635.

Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S. 2016. How disordered is my protein and what is its disorder for? A guide through the "Dark side" Of the protein universe. Intrinsically Disord Proteins. 4(1):e1259708.

Lobb B, Tremblay BJ, Moreno-Hagelsieb G, Doxey AC. 2020. An assessment of genome annotation coverage across the bacterial tree of life. Microb Genom. 6(3).

Lovering RC. 2017. How does the scientific community contribute to gene ontology? Methods Mol Biol. 1446:85-93.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS et al. 2020. Cdd/sparcle: The conserved domain database in 2020. Nucleic Acids Res. 48(D1):D265-D268.

Mair A, Xu SL, Branon TC, Ting AY, Bergmann DC. 2019. Proximity labeling of protein complexes and cell-type-specific organellar proteomes in. Elife. 8.

Matsushima R, Maekawa M, Kusano M, Kondo H, Fujita N, Kawagoe Y, Sakamoto W. 2014. Amyloplast-localized substandard starch grain4 protein influences the size of starch grains in rice endosperm. Plant Physiol. 164(2):623-636.

Mazutis L, Gilbert J, Ung WL, Weitz DA, Griffiths AD, Heyman JA. 2013. Single-cell analysis and sorting using droplet-based microfluidics. Nat Protoc. 8(5):870-891.

McCarty DR, Suzuki M, Hunter C, Collins J, Avigne WT, Koch KE. 2013. Genetic and molecular analyses of uniformmu transposon insertion lines. Methods Mol Biol. 1057:157-166.

Nakanishi N, Martindale MQ. 2018. Crispr knockouts reveal an endogenous role for ancient neuropeptides in regulating developmental timing in a sea anemone. Elife. 7.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A. 96(8):4285-4288.

Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B et al. 2015. Unexpected features of the dark proteome. Proc Natl Acad Sci U S A. 112(52):15898-15903.

Perdigão N, Rosa A. 2019. Dark proteome database: Studies on dark proteins. High Throughput. 8(2).

Periasamy A, Day RN. 1999. Visualizing protein interactions in living cells using digitized gfp imaging and fret microscopy. Methods Cell Biol. 58:293-314.

Power RM, Huisken J. 2019. Putting advanced microscopy in the hands of biologists. Nat Methods. 16(11):1069-1073.

Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y et al. 2018. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature. 557(7706):503-509.

Promponas VJ, Iliopoulos I, Ouzounis CA. 2015. Annotation inconsistencies beyond sequence similarity-based function prediction - phylogeny and genome structure. Stand Genomic Sci. 10:108.

Ramírez-Solis R, Ryder E, Houghton R, White JK, Bottomley J. 2012. Large-scale mouse knockouts and phenotypes. Wiley Interdiscip Rev Syst Biol Med. 4(6):547-563.

Rivoire O. 2013. Elements of coevolution in biological sequences. Phys Rev Lett. 110(17):178102.

Roberts RJ. 2004. Identifying protein function--a call for community action. PLoS Biol. 2(3):E42.

Schober AF, Mathis AD, Ingle C, Park JO, Chen L, Rabinowitz JD, Junier I, Rivoire O, Reynolds KA. 2019. A two-enzyme adaptive unit within bacterial folate metabolism. Cell Rep. 27(11):3359-3370.e3357.

Selkrig J, Belousoff MJ, Headey SJ, Heinz E, Shiota T, Shen HH, Beckham SA, Bamert RS, Phan MD, Schembri MA et al. 2015. Conserved features in tama enable interaction with tamb to drive the activity of the translocation and assembly module. Sci Rep. 5:12905.

Selkrig J, Mosbahi K, Webb CT, Belousoff MJ, Perry AJ, Wells TJ, Morris F, Leyton DL, Totsika M, Phan MD et al. 2012. Discovery of an archetypal protein transport system in bacterial outer membranes. Nat Struct Mol Biol. 19(5):506-510, S501.

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A et al. 2020. Improved protein structure prediction using potentials from deep learning. Nature. 577(7792):706-710.

Shinogle-Decker H, Martinez-Rivera N, O'Brien J, Powell R, Joshi V, Connell S, Rosa-Molinar E. 2018. Correlative fret: New method improves rigor and reproducibility in determining distances within synaptic nanoscale architecture. Proc. SPIE 10499, Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XXV, 1049920; 23 February 2018.

Shrestha D, Jenei A, Nagy P, Vereb G, Szöllősi J. 2015. Understanding fret as a research tool for cellular studies. Int J Mol Sci. 16(4):6718-6756.

Snapp EL, Hegde RS. 2006. Rational design and evaluation of fret experiments to measure protein proximities in cells. Curr Protoc Cell Biol. Chapter 17:Unit 17.19.

Snel B, Bork P, Huynen MA. 2002. The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci U S A. 99(9):5890-5895.

Somerville C, Dangl. 2000. Genomics. Plant biology in 2010. Science. 290(5499):2077-2078.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P et al. 2019. String v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47(D1):D607-D613.

Thiaville JJ, Flood J, Yurgel S, Prunetti L, Elbadawi-Sidhu M, Hutinet G, Forouhar F, Zhang X, Ganesan V, Reddy P et al. 2016. Members of a novel kinase family (duf1537) can recycle toxic intermediates into an essential metabolite. ACS Chem Biol. 11(8):2304-2311.

Thöne FMB, Kurrle NS, von Melchner H, Schnütgen F. 2019. Crispr/cas9-mediated generic protein tagging in mammalian cells. Methods. 164-165:59-66.

Tibbs Cortes L, Zhang Z, Yu J. 2021. Status and prospects of genome-wide association studies in plants. Plant Genome.e20077.

Tompa P. 2012. Intrinsically disordered proteins: A 10-year recap. Trends Biochem Sci. 37(12):509-516.

Toprak E, Veres A, Yildiz S, Pedraza JM, Chait R, Paulsson J, Kishony R. 2013. Building a morbidostat: An automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. Nat Protoc. 8(3):555-567.

Tsai HY, Kim H, Massey M, Krause KD, Algar WR. 2019. Concentric fret: A review of the emerging concept, theory, and applications. Methods Appl Fluoresc. 7(4):042001.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A et al. 2015. Proteomics. Tissue-based map of the human proteome. Science. 347(6220):1260419.

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S et al. 2010. Towards a knowledge-based human protein atlas. Nat Biotechnol. 28(12):1248-1250.

Uhlen M, Ponten F. 2005. Antibody-based proteomics for human tissue profiling. Mol Cell Proteomics. 4(4):384-393.

Van Auken K, Schaeffer ML, McQuilton P, Laulederkind SJ, Li D, Wang SJ, Hayman GT, Tweedie S, Arighi CN, Done J et al. 2014. Bc4go: A full-text corpus for the biocreative iv go task. Database (Oxford). 2014.

Varshney GK, Burgess SM. 2014. Mutagenesis and phenotyping resources in zebrafish for studying development and human disease. Brief Funct Genomics. 13(2):82-94.

Wong BG, Mancuso CP, Kiriakov S, Bashor CJ, Khalil AS. 2018. Precise, automated control of conditions for high-throughput growth of yeast and bacteria with evolver. Nat Biotechnol. 36(7):614-623.

Wu Q, Luo A, Zadrozny T, Sylvester A, Jackson D. 2013. Fluorescent protein marker lines in maize: Generation and applications. Int J Dev Biol. 57(6-8):535-543.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The i-tasser suite: Protein structure and function prediction. Nat Methods. 12(1):7-8.

Yu J, Li T, Dai S, Weng Y, Li J, Li Q, Xu H, Hua Y, Tian B. 2017. A tamb homolog is involved in maintenance of cell envelope integrity and stress resistance of deinococcus radiodurans. Sci Rep. 7:45929.

Zaccolo M. 2004. Use of chimeric fluorescent proteins and fluorescence resonance energy transfer to monitor cellular responses. Circ Res. 94(7):866-873.

Zallot R, Oberg N, Gerlt JA. 2019. The efi web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry. 58(41):4169-4182.

Zallot R, Oberg NO, Gerlt JA. 2018. 'democratized' genomic enzymology web tools for functional assignment. Curr Opin Chem Biol. 47:77-85.

Zhan T, Boutros M. 2016. Towards a compendium of essential genes - from model organisms to synthetic lethality in cancer cells. Crit Rev Biochem Mol Biol. 51(2):74-85.

Zhang J, Wu S, Boehlein SK, McCarty DR, Song G, Walley JW, Myers A, Settles AM. 2019. Maize. J Cell Biol. 218(8):2638-2658.

Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, Bonanno JB, Hillerich BS, Seidel RD, Babbitt PC et al. 2013. Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature. 502(7473):698-702.

Zhao Y, Zhang Z, Gao J, Wang P, Hu T, Wang Z, Hou YJ, Wan Y, Liu W, Xie S et al. 2018. Arabidopsis duodecuple mutant of pyl aba receptors reveals pyl repression of aba-independent snrk2 activity. Cell Rep. 23(11):3340-3351.e3345.

Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, Lewis KA, Georghiou G, Nguyen HN, Hamid MN et al. 2019. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biology. 20(1):244.

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented mpi bioinformatics toolkit with a new hhpred server at its core. J Mol Biol. 430(15):2237-2243.

Zwick M, Kraemer O, Carter AJ. 2019. Dataset of the frequency patterns of publications annotated to human protein-coding genes, their protein products and genetic relevance. Data Brief. 25:104284.