PNAS

Status of Genome Function Annotation in Model Organisms and Crops

Bo Xue and Seung Y Rhee*

Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA

*Corresponding author

Email: srhee@carnegiescience.edu (SYR); bxue@carnegiescience.edu (BX)

Author Contributions: SYR conceived the project, and BX implemented the project. BX and SYR wrote and edited the manuscript.

Competing Interest Statement: The authors declare no competing interests.

Classification: Biological Science, Systems Biology, Plant Biology

Keywords: genome annotation, gene function, gene ontology, model organisms, bioenergy crops, food crops

This file includes:

Main Text
Figures 1 to 2

Abstract

Since entry into genome-enabled biology 20 years ago, much progress has been made in determining, describing, and disseminating functions of genes and their products. Yet, this information is still difficult to access for many scientists and for most genomes. To provide easy access to the status of genome function annotation for model organisms and bioenergy and food crop species, we created a web application (https://genomeannotation.rheelab.org) to visualize and download genome annotation data for 28 species. The summary graphics and data tables will be updated semi-annually and snapshots will be archived to provide a historical record of the progress of genome function annotation efforts.

Main Text

Background

Rapid advances in DNA sequencing technologies made genome sequences widely available and revealed a plethora of genes encoded within the genomes (1). The timely invention and wide adoption of the Gene Ontology (GO) system transformed how gene and protein functions are described, quantified, and compared across many organisms (2, 3). Despite this tremendous progress in genome biology, it is still not trivial for scientists to get a snapshot of the status of genome function annotation across species.

Status of genome function annotation is not easily accessible for several reasons. First, genome sequences and their annotations are hosted across multiple databases that use different gene/protein/sequence identifier systems. For example, Phytozome (4) uses its own database identifiers for its genes and does not provide cross-database identifier (ID) mapping functionalities. Although some databases include cross database references and provide tools to map IDs, such as UniProt's Retrieve/ID mapping and BioMart's ID conversion (5), these tools are not available for all sequenced genomes. Second, gene function information is not generally annotated using the GO system in the literature and most databases. Third, genome function annotation databases generally only include annotated genes, and it is not trivial to retrieve the number and identity of unannotated genes. The importance of unannotated genes is exemplified by a recent success in identifying the minimal bacterial genome that included 473 essential genes (6). Among these were 149 genes whose molecular functions remain unknown.

To provide scientists and students an easy way to access the status of genome function annotations of model species and bioenergy and food crops, we created a web application that displays these data graphically and tabularly. The website retrieves data from multiple databases, and generates plots that show the percentages of genes with experimental, computational, or no annotations. The snapshots are updated semi-annually, and past snapshots will be archived.

Results

To represent the status of genome function annotation, we selected three groups of organisms: model organisms, bioenergy model and crop species, and most-annotated plant species (**Figure 1**). Model organisms are important experimental tools for investigating biological processes and represent key reference points of biological knowledge for other species (7–9). This panel includes: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Mus musculus*, *Saccharomyces cerevisiae,* and *Schizosaccharomyces pombe* (**Fig. 1A**). We also included *Homo sapiens*, a species for which many model organisms are studied. Next, we selected bioenergy models and crops, which are important in expanding the renewable energy sector needed to combat the climate crisis and steward a more sustainable environment. Biomass is currently the biggest source of renewable energy (10) and is projected to become the biggest source of primary energy by 2050 (11). The bioenergy models and crops we selected include: *Brachypodium distachyon*, *Chlamydomonas reinhardtii*, *Glycine max*, *Miscanthus sinensis*, *Panicum hallii*, *Panicum virgatum*, *Physcomitrium patens*, *Populus trichocarpa*, *Sorghum bicolor*, and *Setaria italica* (**Fig. 1B**). Finally, we included ten additional plant species that have the highest number of GO annotations in UniProt (12), which include: *Oryza sativa Japonica Group* (rice), *Gossypium hirsutum* (cotton), *Spinacia oleracea* (spinach), *Zea mays* (corn), *Medicago truncatula*, *Solanum tuberosum* (potato), *Ricinus communis* (castor bean), *Nicotiana tabacum* (tobacco), *Papaver somniferum* (opium poppy), and *Triticum aestivum* (wheat) (**Fig. 1C**). These include the world's most important cereal crops, such as corn, rice, wheat, and vegetable crops such as potato (13).

There are several ways of accessing the status of genome function annotation for the 28 species. From the front page, visitors can get a quick summary of the state of the genome function annotation as pie charts for the three groups of species (**Figure 1**). These pie charts show the percentage of genes that have: 1) annotations with experimental evidence (green); 2) only the annotations that are computationally generated (blue); or 3) no annotations or annotations as being unknown (gray) (**Figure 1**). Of the eightselected model organisms, *H. sapiens* has the highest percentage of genes with experimental evidence, followed by *S. cerevisiae* and *M. musculus. A. thaliana* has the lowest percentage of "unknown" genes. Among the model organisms, *C. elegans* is the least known species, with the greatest number of genes unannotated or annotated as having unknown function. Most of the plant species have too few GO

annotations based on experimental support to be visible in the pie charts. Visitors can get more detailed information of any of the species by clicking on the species name below the pie charts. Each species page shows additional information about annotation status, including displaying the portion of genes annotated to at least one GO domain (molecular function, cellular component, and biological process (2, 3)), as well as a Venn diagram showing the overlap of genes annotated to more than one GO domain (**Figure 2**). This page also has links to source data and a tabular format of the annotation summary for browsing and downloading.

Discussion

Our website provides a convenient way to obtain the current state of genome function annotation for model organisms and crops for bioenergy, food, and medicine. Our website shows how much is annotated and unannotated in 28 species that represent some of the most intensely studied and arguably most valuable organisms to science and society. These charts also serve as a proxy for illustrating how much is known and unknown. These snapshots will be updated on a semi-annual basis, and comparing the charts across time will reflect how biological knowledge changes over time. These snapshots can be useful in many contexts, including research projects, grant proposals, review articles, annual reports, and outreach materials.

In developing our web application, we encountered a few hurdles. First, there was not a single site where all data were available. To obtain GO annotations from the 28 species, we had to visit several databases. An encouraging finding was that all sites that had GO annotations were using the GO Annotation File (GAF) format. Nevertheless, having a single-entry point where GO annotations of any species can be accessed would be useful. Second, our website includes genes that are unannotated, which are often missing in gene function annotations and enrichment analyses (14). Currently, extracting genes that are not annotated is not trivial and requires many steps that differ across species. Including the unannotated genes in a genome into GAF files would facilitate many downstream applications.

To our surprise, some plant species with well-maintained, species-specific databases seem to have a low number of experimentally supported GO annotated genes in UniProt. Outside of TAIR (15) that provides GO annotations for *A. thaliana*, we were not able to find any database that provides experimental evidence codes to their GO annotations (Supplementary Information). Most plant genome databases stop at computationally generating GO annotations, and some important species do not appear to have dedicated databases. More efforts are needed in experimentally validating functional annotations made from computational approaches, characterizing genes of unknown function, and curating experimentally supported function descriptions in the literature into structured annotations such as GO, which will be crucial for accelerating gene function discovery.

The data summarized on this website can be linked to their sources, which can be used for a variety of investigations. Successful examples include exploring why certain proteins remain unannotated (16), developing pipelines to infer function without relying on sequence similarity (17), and assessing annotation coverage across bacterial proteomes (18). As our society transitions into biology-enabled manufacturing (19), fundamental knowledge of how genes and their products function at various scales will be crucial, as our society transitions into biology-enabled manufacturing and a new bio-economy era.

Acknowledgments

We thank members of the Rhee Lab for discussions and suggestions on the project and Kristen Yawitz for editing the manuscript. This work was supported, in part, by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program grant nos. DE–SC0018277, DE–SC0020366, and DE–SC0021286, and the National Science Foundation grants MCB-2052590, MCB-1916797, and IOS-1546838. This work was done on the ancestral land of the Muwekma Ohlone Tribe, which was and continues to be of great importance to the Ohlone people.

References

1. N. A. O'Leary, *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

2. M. Ashburner, *et al.*, Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

3. N. Acids research, 2021, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

4. D. M. Goodstein, *et al.*, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–86 (2012).

5. J. M. Guberman, *et al.*, BioMart Central Portal: an open database network for the biological community. *Database* **2011**, bar041 (2011).

6. C. A. Hutchison 3rd, *et al.*, Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).

7. R. A. Ankeny, S. Leonelli, "Model Organisms" in *Elements in the Philosophy of Biology*, (Cambridge University Press, 2020) (April 26, 2022).

8. S. Fields, M. Johnston, Cell biology. Whither model organism research? *Science* **307**, 1885–1886 (2005).

9. A. M. Jones, *et al.*, The impact of Arabidopsis on human health: diversifying our portfolio. *Cell* **133**, 939–943 (2008).

10. ,U.S. energy facts explained - consumption and production - U.S. Energy Information Administration (EIA) (June 5, 2022).

11. W. V. Reid, M. K. Ali, C. B. Field, The future of bioenergy. *Glob. Chang. Biol.* **26**, 274–286 (2020).

12. UniProt Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

13. ,FIGURE 21: World production of crops, main commodities. *FAO Statistical Yearbook 2021 Datasets* (2021) https:/doi.org/10.4060/cb4477en-fig21.

14. D. P. Higgins, C. M. Weisman, D. S. Lui, F. A. D'Agostino, A. K. Walker, Defining characteristics and conservation of poorly annotated genes in Caenorhabditis elegans using WormCat 2.0. *Genetics* (2022) https:/doi.org/10.1093/genetics/iyac085.

15. P. Lamesch, *et al.*, The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–10 (2012).

16. V. Wood, *et al.*, Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol.* **9**, 180241 (2019).

17. F. Bossi, *et al.*, Systematic discovery of novel eukaryotic transcriptional regulators using sequence

homology independent prediction. *BMC Genomics* **18**, 480 (2017).

18. B. Lobb, B. J.-M. Tremblay, G. Moreno-Hagelsieb, A. C. Doxey, An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom* **6** (2020).

19. Committee on Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals, Board on Chemical Sciences and Technology, Board on Life Sciences, Division on Earth and Life Studies, National Research Council, *Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals* (National Academies Press (US), 2015).

Figures

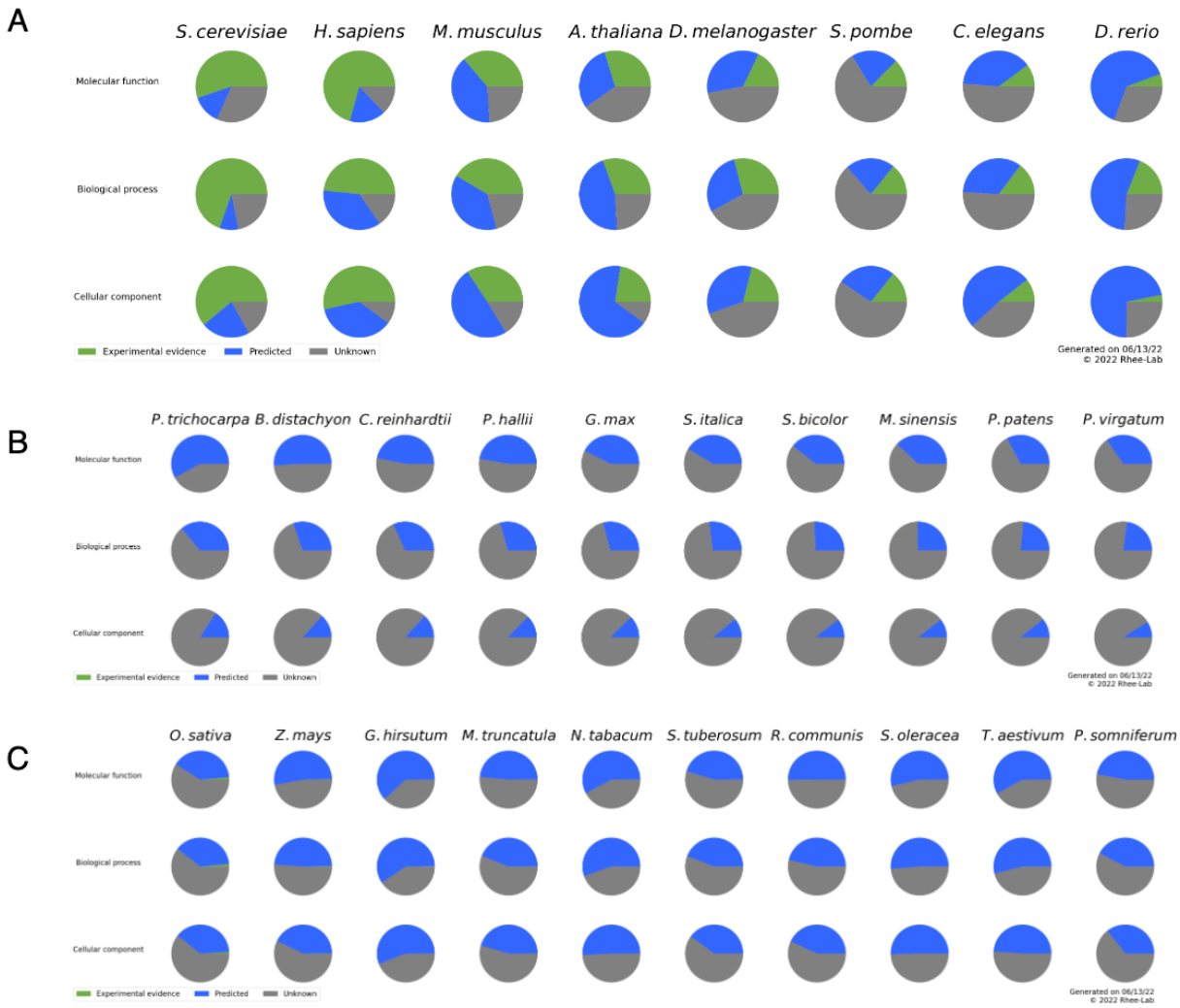


Figure 1 Status of genome function annotations

Each pie chart shows the proportion of genes that are annotated to a domain of Gene Ontology (GO): molecular function, biological process, or cellular component. Green indicates genes that have at least one experimentally validated GO annotation; blue indicates genes that are annotated with computationally predicted GO annotations; and gray indicates genes that do not have any GO annotations or have GO annotations with ND (No biological Data available). The species are ordered by the average percentage

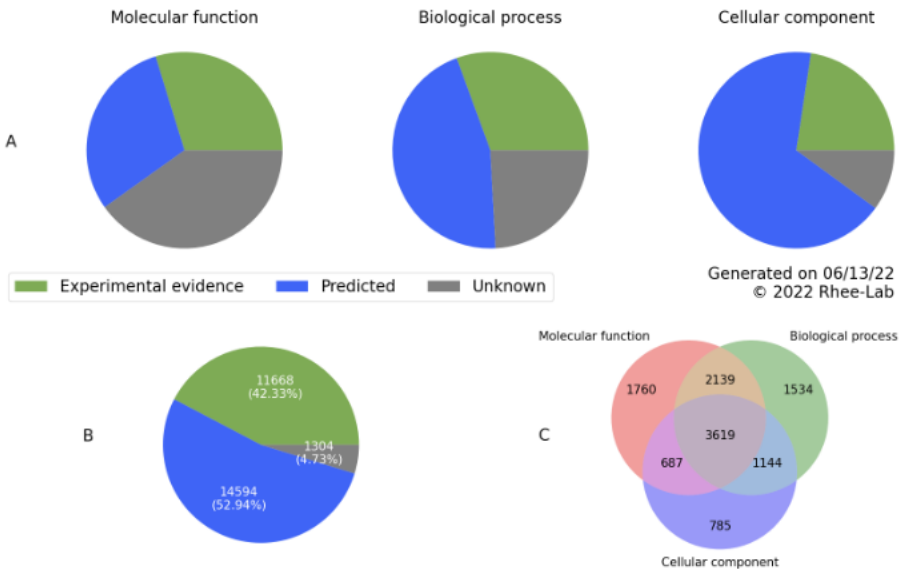
of genes with experimental evidence (or genes with computationally predicted annotations if experimental evidence is not available) of all three domains. A) selected model organisms; B) bioenergy models and crops; C) other plant species with the highest percentage of genes with experimental evidence in UniProt.

Arabidopsis thaliana

GAF Source	Gene Ontology
Sequence Source	The Arabidopsis Information Resource (TAIR)
Download Date	06/13/22

Back to Overview

Status of gene function elucidation and annotation



Copy Link to Image | Download "Experimental Evidence" Gene List | Download "Predicted" Gene List | Download "Unknown" Gene List

Stats of Arabidopsis thaliana (Taxon: 3702)

Toggle Table | Download Table

Num. of Genes	27566
Num. of Genes w/ Experimental evidence	11668
Num. of Genes w/ Computational evidence	14594
Num. of Genes with no GO annotations	1304
Num. of Molecular function (Experimental evidence)	8205 (29.76%)
Num. of Molecular function (Predicted)	8295 (30.09%)
Num. of Molecular function (Unknown)	11066 (40.14%)
Num. of Biological process (Experimental evidence)	8436 (30.60%)
Num. of Biological process (Predicted)	12468 (45.23%)
Num. of Biological process (Unknown)	6662 (24.17%)
Num. of Cellular component (Experimental evidence)	6235 (22.62%)
Num. of Cellular component (Predicted)	18568 (67.36%)
Num. of Cellular component (Unknown)	2763 (10.02%)

Generation Date: 06/13/22

Archived charts

Link to archives.

Figure 2 An example species-specific annotation web page shown for *Arabidopsis thaliana*. It consists of three parts: 1) a table comprising data sources; 2) pie charts showing the proportion of each type of annotations; and 3) a table showing the numbers of genes in each category, which users can toggle to either show or hide.