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

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

*Corresponding author

Keywords: genome annotation, Gene Ontology, model organisms, food crops, bioenergy crops, function annotation, genes of unknown function, Arabidopsis, sorghum, rice

Abstract

Since the 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 by many, especially across 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 27 species. The summary graphics and data tables will be updated semi-annually and snapshots archived to provide a historical record of the progress of genome function annotation efforts.

Background

Rapid advances in DNA sequencing technologies made genome sequences widely available and revealed a plethora of genes encoded within the genomes in the last two decades [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 nontrivial for scientists to get a snapshot of the status of genome function annotation across species.

There are several reasons for the difficulty in obtaining the status of genome function annotation across species. 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 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. 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 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 and Discussion

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*, and *Saccharomyces cerevisiae* (**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 selected ten additional plant species that have the most 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), *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 27 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 (**Figure 1**). Of the 7 selected model organisms, *S. cerevisiae* has the highest percentage of genes with experimental evidence and the least number of genes unannotated or annotated as having unknown function, followed by *H. sapiens* and *A. thaliana.* 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 few GO annotations based on experimental support to be even 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 the 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.

In developing our web application, we came across a few hurdles. First, there was not a single site where all data were available. To obtain GO annotations from the 27 species, we had to visit at least three databases. A positive 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 is 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 are different for each 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 that provides GO annotations for *A. thaliana* [19], we were not able to find any database that provides experimental evidence codes to their GO annotations. Apart from *Nicotiana tabacum* and *Papaver somniferum*, all plants species on our website are included in the most recent version of Phytozome V13, but their GO terms are assigned computationally [4]. The Sol Genomics Network (SGN) (https://solgenomics.net accessed 13 June 2022) [15] hosts genome annotations of Solanaceae species, including *Nicotiana tabacum* and *Solanum tuberosum*. An annotation file for *Nicotiana tabacum* is available [16] but they are assigned with computational support coming from InterProScan [17]. SpudDB [18] (http://spuddb.uga.edu/ accessed 13 June 2022) provides GO annotation for *Solanum tuberosum* but they are generated with InterProScan and by best hit to the Arabidopsis proteome (TAIR10) [19]. MaizeGDB [20] (https://www.maizegdb.org/ accessed 13 June 2022) provides GO annotation for *Zea mays* that are assigned with GO annotation tools including Argot2.5, FANN-GO, and PANNZER [21], which are all computational annotations. SpinachBase (http://www.spinachbase.org/ accessed 13 June 2022) provides a centralized access to *Spinacia oleracea*, and their GO annotations are generated computationally with Blast2GO [22]. *Oryza sativa Japonica Group* GO annotations can be found on Rice Genome Annotation Project [23] and they are assigned with BLASTP searches against Arabidopsis GO-curated proteins [24]. Gramene [31] (https://www.gramene.org/ accessed 13 June 2022) hosts genome data for many species but we could not find GO annotations with evidence codes. We were not able to find species-specific databases that provide GO annotations for *Triticum aestivum*, *Gossypium hirsutu*, *Medicago truncatula*, *Papaver somniferum* or *Ricinus communis*. In summary, 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 both experimentally validating functional annotations made from computational approaches and curating experimentally supported function descriptions in the literature into structured annotations such as GO, which will be crucial for accelerating gene function discovery.

Conclusions

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 the 27 species that represent some of the most intensely studied and arguably the most valuable organisms for science and society. By proxy, these charts illustrate 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. 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 [25], developing pipelines to infer function without relying on sequence similarity [26], and assessing annotation coverage across bacterial proteomes [27]. As our society transitions into biology-enabled manufacturing [32], fundamental knowledge of how genes and their products function at various scales will be crucial in ushering in the era of bio-economy.

Methods

Selecting species and data retrieval

For the seven model organisms, gene function annotations were downloaded as GO Annotation Files (GAF files) from the GO consortium website (http://current.geneontology.org/products/pages/downloads.html accessed 13 June 2022) of the 2022-05-16 release. Genes found in a genome were retrieved from the source indicated on the

GO annotation download page as General Feature Format (GFF) files. A detailed description of the files used to generate charts on our website, including data for the other category of species, can be found in **Table S1**.

Genome annotation and gene list for bioenergy models and crops were downloaded from Phytozome version V13 (https://data.jgi.doe.gov/refine-download/phytozome accessed 13 June 2022). Although some species in this category had GO annotations in the GO consortium database, the sequence identifiers (IDs) for genes could not easily be mapped to Phytozome IDs. To maintain consistency within this category, all annotation files were downloaded from Phytozome. All Phytozome GO annotations are computationally generated [4]. Gene lists were also retrieved from Phytozome V13.

For the last category of plant species, we selected the most annotated plant species from the UniProt GO annotation database [28] GAF files hosted on the GO consortium website (http://current.geneontology.org/products/pages/downloads.html 2022-05-16 release, accessed 13 June 2022). We downloaded these species reference proteomes from the UniProt release 2022_02 and retrieved the number of corresponding genes.

Using the evidence codes provided by GAF files, we generated the numbers of genes annotated with GO supported by experimental evidence. If a gene has at least one GO term annotated using any of the following codes: EXP (Inferred from Experiment), IDA (Inferred from Direct Assay), IPI (Inferred from Physical Interaction), IMP (Inferred from Mutant Phenotype), IGI (Inferred from Genetic Interaction), or IEP (Inferred from Expression Pattern), we categorized the gene as having “Experimental Evidence” for function. Genes that have at least one annotated GO term, but no terms have the evidence codes described above, are categorized as “Predicted”. Since Phytozome has only computationally generated GO annotations, all of their genes are categorized as having their functions “Predicted”. By subtracting the annotated genes from the total number of genes, we retrieved the number of

genes without any GO annotations. These numbers were used to generate pie charts to show the proportions of genes in each category for every species.

All files were processed with scripts written in Python (3.10). All pie charts were generated using Python Matplotlib version 3.5.2 and Venn diagrams were generated using Python matplotlib-venn version 0.11.7. The repository of codes can be found at GitHub (https://github.com/bxuecarnegie/AnnotationStats).

Creating the Website

To create a website for hosting our charts, we used Node.js [29] for our server-side environment, which provides the Application Program Interface (API) for the front end to retrieve the plots generated by Python. The front end of the website uses AngularJS [30].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Data used in this study are all publicly available. GO annotation files were downloaded from (http://current.geneontology.org/annotations/index.html 2022-05-16 release, accessed 13 June 2022) and Phytozome (https://data.jgi.doe.gov/refine-download/phytozome V13, accessed 13 June 2022). Gene data were downloaded from sources indicated on the GO (http://current.geneontology.org/products/pages/downloads.html accessed 13 June 2022), Phytozome, and UniProt (https://www.uniprot.org/ accessed 13 June 2022). Supplemental Table S1 provides detailed information on all species annotation and gene source databases, downloaded versions, and URLs. Graphs and statistics data generated in this study are available at (http://genomeannotation.rheelab.org/ accessed 13 June 2022). Scripts used to process the data and generate the graphs are written in Python 3 and are available at GitHub (https://github.com/bxuecarnegie/AnnotationStats accessed 13 June 2022).

Competing interests

The authors declare no competing interests.

Funding

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-SC0008769, DE-SC0020366 and DE-SC0021286 and the U.S. National Science Foundation grants MCB-1617020 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.

Authors' contributions

SYR conceived the project and BX implemented the project. BX and SYR wrote and edited the manuscript.

Acknowledgements

We would like to thank the members of Rhee lab for the discussions and suggestions on the project.

References

1. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. academic.oup.com; 2016;44:D733–45.

2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. Nature Publishing Group; 2000;25:25–9.

3. Acids research N, 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. Oxford University Press; 2021;49:D325–34.

4. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86.

5. Guberman JM, Ai J, Arnaiz O, Baran J, Blake A, Baldock R, et al. BioMart Central Portal: an open database network for the biological community. Database . academic.oup.com; 2011;2011:bar041.

6. Hutchison CA 3rd, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351:aad6253.

7. Ankeny RA, Leonelli S. Model Organisms. Elements in the Philosophy of Biology. Cambridge University Press; 2020.

8. Fields S, Johnston M. Cell biology. Whither model organism research? Science. 2005;307:1885–6.

9. Jones AM, Chory J, Dangl JL, Estelle M, Jacobsen SE, Meyerowitz EM, et al. The impact of Arabidopsis on human health: diversifying our portfolio. Cell. 2008;133:939–43.

10. U.S. energy facts explained - consumption and production - U.S. Energy Information Administration (EIA) [Internet]. [cited 2022 Jun 5]. Available from: https://www.eia.gov/energyexplained/us-energy-facts/

11. Reid WV, Ali MK, Field CB. The future of bioenergy. Glob Chang Biol. 2020;26:274–86.

12. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. academic.oup.com; 2019;47:D506–15.

13. FIGURE 21: World production of crops, main commodities [Internet]. FAO Statistical Yearbook 2021 Datasets. Food and Agriculture Organization of the United Nations; 2021. Available from: http://www.fao.org/3/cb4477en/StatYearbook2021-fig21.xlsx

14. Higgins DP, Weisman CM, Lui DS, D'Agostino FA, Walker AK. Defining characteristics and conservation of poorly annotated genes in Caenorhabditis elegans using WormCat 2.0. Genetics [Internet]. 2022; Available from: http://dx.doi.org/10.1093/genetics/iyac085

15. Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, et al. The Sol Genomics Network (SGN)--from genotype to phenotype to breeding. Nucleic Acids Res. 2015;43:D1036–41.

16. Edwards KD, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans AD, Bombarely A, et al. A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. BMC Genomics. Springer; 2017;18:448.

17. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.

18. Hirsch CD, Hamilton JP, Childs KL, Cepela J, Crisovan E, Vaillancourt B, et al. Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. Plant Genome. Wiley; 2014;7:lantgenome2013.12.0042.

19. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. academic.oup.com; 2012;40:D1202–10.

20. Woodhouse MR, Cannon EK, Portwood JL 2nd, Harper LC, Gardiner JM, Schaeffer ML, et al. A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biol. 2021;21:385.

21. Wimalanathan K, Friedberg I, Andorf CM, Lawrence-Dill CJ. Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER). Plant Direct. 2018;2:e00052.

22. Collins K, Zhao K, Jiao C, Xu C, Cai X, Wang X, et al. SpinachBase: a central portal for spinach genomics. Database [Internet]. academic.oup.com; 2019;2019. Available from: http://dx.doi.org/10.1093/database/baz072

23. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 2007;35:D883–7.

24. Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, Lin H, et al. The institute for genomic research Osa1 rice genome annotation database. Plant Physiol. 2005;138:18–26.

25. Wood V, Lock A, Harris MA, Rutherford K, Bähler J, Oliver SG. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? Open Biol. The Royal Society; 2019;9:180241.

26. Bossi F, Fan J, Xiao J, Chandra L, Shen M, Dorone Y, et al. Systematic discovery of novel eukaryotic transcriptional regulators using sequence homology independent prediction. BMC Genomics. Springer; 2017;18:480.

27. Lobb B, Tremblay BJ-M, Moreno-Hagelsieb G, Doxey AC. An assessment of genome annotation coverage across the bacterial tree of life. Microb Genom [Internet]. 2020;6. Available from: http://dx.doi.org/10.1099/mgen.0.000341

28. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in UniProt with Gene Ontology. Nucleic Acids Res. 2004;32:D262–6.

29. Tilkov S, Vinoski S. Node.js: Using JavaScript to Build High-Performance Network Programs. IEEE Internet Comput. 2010;14:80–3.

30. Jain, Bhansali, Mehta. AngularJS: A modern MVC framework in JavaScript. Journal of Global Research in Computer [Internet]. jgrcs.info; 2014; Available from: http://www.jgrcs.info/index.php/jgrcs/article/download/952/610

31. Tello-Ruiz MK, Naithani S, Gupta P, Olson A, Wei S, Preece J, et al. Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Res. 2021;49:D1452–63.

32. 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. Washington (DC): 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

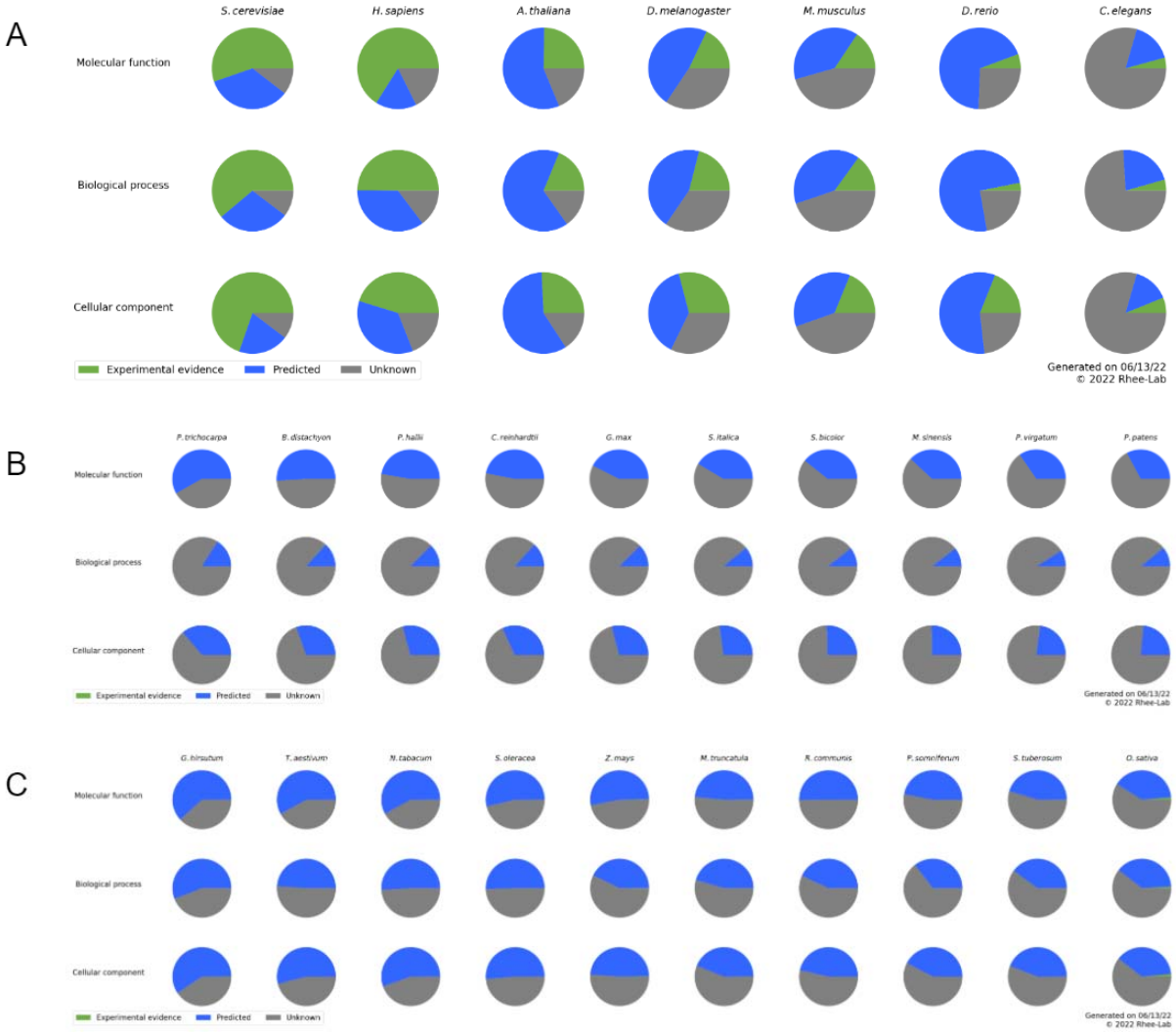
have at least one experimentally validated GO annotation, blue indicates genes that are annotated but none are experimentally annotated, and gray indicates genes that do not have any GO annotations. The species are sorted by the percentage of genes with experimental evidence. A) selected model organisms; B) bioenergy models and crops [1]; C) other plant species with the highest percentage of genes with experimental evidence in UniProt.

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

Supplemental Information

Additional file 1

Table S1 Source data. A table describing the data sources, versions downloaded, and URLs

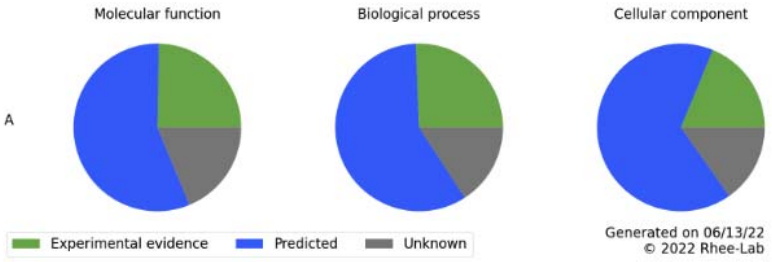


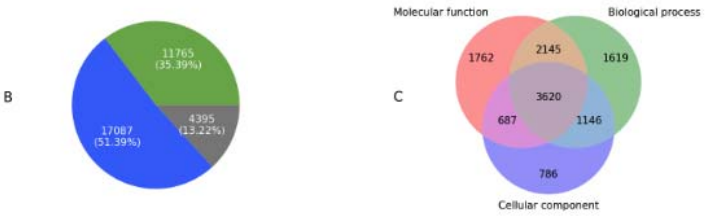
Arabidopsis thaliana

GAF Source	Gene Ontology
Sequence Source	The Arabidopsis Information Resource (TAIR)
Download Date	6/13/22

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	33247
Num. of Genes w/ Experimental evidence	11775
Num. of Genes w/ Computational evidence	17077
Num. of Genes with no GO annotations	4395
Num. of Molecular function (Experimental evidence)	8217 (24.72%)
Num. of Molecular function (Predicted)	18802 (56.55%)
Num. of Molecular function (Unknown)	6228 (18.73%)
Num. of Biological process (Experimental evidence)	8558 (25.74%)
Num. of Biological process (Predicted)	19433 (58.45%)
Num. of Biological process (Unknown)	5256 (15.81%)
Num. of Cellular component (Experimental evidence)	6241 (18.77%)
Num. of Cellular component (Predicted)	21949 (66.02%)
Num. of Cellular component (Unknown)	5057 (15.21%)

Generation Date: 6/13/22

Archived charts

Link to archives.