PhyloGenes: An online phylogenetics and functional genomics resource for plant gene function inference

Peifen Zhang1 | Tanya Z. Berardini1 | Dustin Ebert2 | Qian Li1 | Huaiyu Mi2 | Anushya Muruganujan2 | Trilok Prithvi1 | Leonore Reiser1 | Swapnil Sawant1 | Paul D. Thomas2 | Eva Huala1

1Phoenix Bioinformatics, Fremont, CA, USA
2Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

Correspondence
Eva Huala, Phoenix Bioinformatics, Fremont, California 94538, USA.
Email: huala@phoenixbioinformatics.org

Funding information
This work was supported by the US National Science Foundation (Grant No. DBI-1661543).

Abstract
We aim to enable the accurate and efficient transfer of knowledge about gene function gained from Arabidopsis thaliana and other model organisms to other plant species. This knowledge transfer is frequently challenging in plants due to duplications of individual genes and whole genomes in plant lineages. Such duplications result in complex evolutionary relationships between related genes, which may have similar sequences but highly divergent functions. In such cases, functional inference requires more than a simple sequence similarity calculation. We have developed an online resource, PhyloGenes (phylogenes.org), that displays precomputed phylogenetic trees for plant gene families along with experimentally validated function information for individual genes within the families. A total of 40 plant genomes and 10 non-plant model organisms are represented in over 8,000 gene families. Evolutionary events such as speciation and duplication are clearly labeled on gene trees to distinguish orthologs from paralogs. Nearly 6,000 families have at least one member with an experimentally supported annotation to a Gene Ontology (GO) molecular function or biological process term. By displaying experimentally validated gene functions associated to individual genes within a tree, PhyloGenes enables functional inference for genes of uncharacterized function, based on their evolutionary relationships to experimentally studied genes, in a visually traceable manner. For the many families containing genes that have evolved to perform different functions, PhyloGenes facilitates the use of evolutionary history to determine the most likely function of genes that have not been experimentally characterized. Future work will enrich the resource by incorporating additional gene function datasets such as plant gene expression atlas data.

1 | INTRODUCTION

A major challenge and opportunity for plant biology in the post-genomic era is the transfer of knowledge gleaned from model or reference species to other organisms (Lambing & Heckmann, 2018; Piquerez et al., 2014; Rhee & Mutwil, 2014). In the past two decades, considerable investment has been made in understanding gene function in model species. For Arabidopsis alone, over 54,000 papers have been published since 1965, and the rate at which new papers appear has doubled since Arabidopsis became the first...
plant to have its genome sequenced in 2000 (Arabidopsis Genome Initiative, 2000; Provart et al., 2016). Since that benchmark genome release, genomes have been sequenced for many plants including most species of agricultural importance and many species of special ecological or evolutionary significance (https://plabipd.de/portal/sequenced-plant-genomes). Meanwhile, advances in genomics, bioinformatics and other technologies have vastly expanded our experimental toolkit. We are now poised to gain a new understanding of plant gene function in an evolutionary context, and to reap the benefits of this advance through precision engineering of plant genomes to modify agriculturally important traits such as disease resistance, drought tolerance, and yield.

Although genome sequencing itself has become a routine exercise and structural identification of protein-coding genes has become more reliable, the large-scale, automated assignment of functional roles to genes falls short in accuracy and sensitivity (Friedberg, 2006; Jiang et al., 2016; Radivojac et al., 2013; Schnoes et al., 2009). Simple sequence similarity-based function prediction methods such as BLAST are not sensitive enough to recover similar proteins that share low sequence homology, or accurate enough to distinguish different proteins that share high sequence similarity. The latter situation is amplified for plant genomes where local and whole genome duplications are common, resulting in complex evolutionary relationships between genes (such as orthologs from speciation, paralogs from gene duplication, and homoeologs from successive speciation and allopolyploidization events). These related genes can have similar sequences but divergent functions (Jiang et al., 2016; Karunanithi & Zerbe, 2019; Li et al., 2019; Torrens-Spence et al., 2020; Xu et al., 2007). As a result, analyses relying solely on sequence similarity to guide functional inference may produce false functional predictions. More accurate results can be achieved through integration of multiple lines of information including phylogenetic relationships, sequence alignment, gene expression and protein–protein interaction data as shown by several ensemble learning methods evaluated in the Critical Assessment of protein Function Annotation (CAFA) experiment (Jiang et al., 2016; Radivojac et al., 2013).

The traditional approach for phylogenetic-based functional prediction propagates function from a gene with an experimentally determined function to its orthologs (Eisen, 1998a, 1998b; Tatusov et al., 1997). This approach is based on the ‘ortholog conjecture’, the idea that orthologs are more likely than paralogs to be functionally similar (Nehrt et al., 2011). Other approaches such as SIFTER (Engelhardt et al., 2005) and the GO Phylogenetic Annotation Project (Gaudet et al., 2011) include both orthologs and paralogs for function propagation and are guided by a statistical model or manual expert review. Regardless of method, phylogenetic-based functional prediction requires the simultaneous input and analysis of (a) the phylogenetic trees carrying the evolutionary relationships of member genes, and (b) the characterized functions of family members within the trees. However, phylogenetic tree construction, along with locating, integrating and overlaying the trees with experimentally derived gene function data from which inferences can be made, are time consuming processes. They also require specific expertise, pointing to the need for online resources that make such datasets easily available to the whole research community in an intuitive, graphical and interactive way.

To date, several online resources focused on plant genomes and providing pre-computed gene families with phylogenetic relationships have been developed, including PLAZA (Van Bel et al., 2018), Ensembl Plants (Bolser et al., 2017), and Phyt ozone (Goodstein et al., 2012). While these resources have many strengths, none are comprehensive as each is lacking one or more key features. For example: 1) including the ability to trace functional inferences back to experimental results on important model organisms outside the plant lineage, 2) interactive phylogenetic trees with a full range of customization and data download options, and 3) a comprehensive set of functional data in which experimentally characterized functions are displayed alongside the phylogenetic trees.

To address the need for a web-based resource that effectively integrates and displays phylogenetic trees with comprehensive gene function information, we developed PhyloGenes (www.phylogenesis.org). PhyloGenes displays pre-computed phylogenetic trees adjacent to experimental gene function data to facilitate functional inference for genes of uncharacterized function in plants. Users can customize the display to include only selected species and data, submit a protein sequence from unrepresented genomes to be inserted to a matching tree, and download trees and associated data files. We have followed a user-driven software development methodology, consulting researchers at every stage of the development cycle, to ensure that PhyloGenes addresses the needs of the community in an intuitive fashion.

2 | DATABASE CONTENT

2.1 | Gene families

PhyloGenes gene families and trees are a subset of those available in the PANTHER gene family database (Ml et al., 2019). We chose to use the PANTHER families because they already serve as the basis for the GO Phylogenetic Annotation Project (Gaudet et al., 2011). The GO Phylogenetic Annotation Project is an extensive long-term effort of the Gene Ontology Consortium to manually review all PANTHER gene families containing human genes and functionally annotate uncharacterized members of the families based on their evolutionary relationships to genes with experimentally determined functions. Making use of the same PANTHER families that are used in the GO project allows us to directly incorporate functional inferences for plant lineages within these families into PhyloGenes. In this way we are able to significantly enrich the available functional annotation for plant genes within the 3,776 families also containing human genes. A long-term goal of the PhyloGenes project is to carry out a similar rigorous review and annotation of the plant-specific PANTHER gene families to form a complete and consistent functional annotation dataset that extends across all plant species.

To produce a PANTHER build with a broadly useful set of plant genomes, we first added an additional 30 plant genomes to the set
of 10 already present in PANTHER families. The 40 plant genomes include those with the most abundant experimentally determined function information or the highest agricultural significance, with additional species chosen to maximize taxonomic breadth within the green plant lineage. The decision to limit the total number of plant genomes within both PANTHER and PhyloGenes to 40 allows us to maximize the utility of the families for plant gene function prediction while minimizing difficulties with computing and visualizing very large trees. PANTHER uses the GIGA algorithm to construct phylogenetic trees (Thomas, 2010). The algorithm builds a gene tree using a pairwise sequence distance matrix and a known species tree. It imposes a set of rules in determining a gene tree's topology, including using the species tree to constrain speciation nodes and maximum parsimony to place duplication nodes. The added plant genomes bring the total in the current PANTHER 15 build to 142 genomes spread across all domains of life. All sequences are derived from the UniProt Reference Proteomes set, which designates one representative protein sequence per gene per sequenced genome (UniProt Consortium, 2019).

To generate PhyloGenes trees from the PANTHER build, a subset of the 142 species were filtered to include only the 40 plant genomes, along with 10 non-plant model organisms containing the great majority of all experimentally determined gene functions outside of plants (Table 1, Figure 1a). The inclusion of non-plant model organisms maximizes the power of functional inference for cases in which a gene function was only experimentally characterized in non-plant species (e.g. certain cell cycle genes characterized in yeast). The less informative non-plant species were removed from the trees without altering the tree topology. After this initial tree pruning step, any remaining gene families that lacked plant genes were also removed, reducing the final set to 8,519 gene families in PhyloGenes 2.1. Among these, 1,061 families contain sequences from all major branches of the tree of life, 3,608 families have sequences found only in Eukaryotes, 858 families have sequences found only in green plants, 1,172 families are limited to land plants, and an additional 1,820 families are restricted to more specific lineages such as those only found in Poaceae (Table 1). The taxonomic range of a gene family was determined by the inferred taxon of the root node, the most recent common ancestral taxon from which all genes of the family have evolved. In the event of horizontal transfer in a gene family, the taxonomic range ignores the species whose genes were carried over by the horizontal transfer. The vast majority of gene families (98% of total) contain no more than 1,000 genes (Figure 1b), while 38 very large families each have more than 2,000 family members, for example, families of pentatricopeptide repeat proteins, MADS box proteins, and cytochrome P450s.

### 2.2 Gene function information

Experimentally determined gene function annotations were obtained from the Gene Ontology Consortium (http://geneontology.org). We extracted all annotations to GO molecular function or biological process terms that were supported by experimental evidence codes, including: Inferred from Experiment (EXP), Inferred from Direct Assay (IDA), Inferred from Physical Interaction (IPI), Inferred from Mutant Phenotype (IMP), Inferred from Genetic Interaction (IGI), and Inferred from Expression Pattern (IEP). Import and linking of these annotations to the PhyloGenes families resulted in 5,989 of 8,519 families having at least one gene with an experimentally determined function. Having an experimentally determined function for at least one family member should enable functional inference for their homologs in these 5,989 families.

We further enriched the number of genes with annotations by importing annotations generated by the GO Phylogenetic Annotation Project (aka PAINT annotations) from GO. These are phylogenetic inferred gene functions (evidence code Inferred from Biological aspect of Ancestor, IBA). 541,974 genes from 3,679 families have PAINT IBA annotations.

### 2.3 User interface

PhyloGenes presents gene families on pages containing a tree graph panel on the left and a gene information panel on the right (Figure 2).
Metadata for each gene family, including its name, PANTHER ID, size, a link to its organism list, and taxonomic range, is shown above the tree graph panel. Inside the tree graph panel, users can pan and drag to explore the tree. Nodes that represent different types of evolutionary events, including speciation, duplication, horizontal transfer, and origin of a new subfamily, are indicated with distinct colors and shapes. Clicking individual nodes will collapse or expand sections of the tree at that point. By default, trees are displayed in their compact views, in which only genes with experimentally determined function are displayed and marked with yellow flask icons. All other genes are hidden in collapsed nodes. When a family does not have any gene with an experimentally determined function, the default tree view collapses all child nodes after the first duplication event from the root. The compact view allows faster page loading which is especially useful for displaying very large families, while still highlighting genes with experimentally determined functions. Clicking on the Expand All icon above a tree (Figure 2, arrow a) will expand all collapsed nodes. All leaf nodes (genes) are aligned with the corresponding gene rows in the gene information panel on the right.

The gene information panel displays functional annotations for genes and other data. Experimentally determined or phylogenetically inferred functions of any family member are displayed as GO term names, each in a separate column header. A yellow flask in a gene row under a specific function header indicates direct experimental support for that function in that gene. Inferred functions
from PAINT annotation are indicated by green tree icons. Clicking on either icon displays details of the annotation including the definition of the GO term, the evidence code and the reference used to support the functional annotation. Users can select to show, hide, or reorder the columns (accessed from the cog icon, Figure 2, arrow b). Users can toggle between this view or a multiple sequence alignment (clicking ‘Show MSA’; Figure 2, arrow c) in the gene information panel.

2.4 | Functionalities

2.4.1 | Searching phylogenies

Users can search PhyloGenes using a UniProt ID (e.g. Q38897), gene ID (e.g. AT5G41410), gene symbol (e.g. BEL1), or a keyword that occurs within a gene family name (e.g. homeodomain). The allowable types of gene IDs are described in the User Guide (https://conf.arabidopsis.org/display/PHGSUP/Search). A search that matches a single family will return the page displaying the tree for that family. When more than one family is found (e.g. a search with the keyword ‘MYB’) a list of families is returned, with each family name linking to the gene tree page. Family list results can be further refined using ‘Filter by Taxonomic Range’ or ‘Organisms included’. These filters can be used to further narrow the results, for example to ‘MYB’ families that contain only genes from Poaceae (using the ‘Taxonomic range’ filter) or families that contain at least one maize gene (using the ‘Organisms included’ filter).

2.4.2 | Tree pruning

In some cases a user may wish to hide some species or genes within a large tree in order to make it easier to focus on the most relevant genes and species. The ‘Prune tree by organism’ function enables users to remove genes for a given species from the tree view by deselecting it from the list (Figure 2, arrow d). This function is accessible from the tools icon (Figure 2, arrow e) or by clicking the ‘Organisms’ hyperlink in the family metadata band. The resulting ‘pruned’ tree view hides genes from the deselected species but does not affect the tree topology or the relationships between genes.

2.4.3 | Tree grafting

PhyloGenes allows users to graft one or more sequences onto the PANTHER trees (http://www.phylogenies.org/grafting). This feature was included to allow inclusion of sequences from species not represented in PhyloGenes. The TreeGrafter (Tang et al., 2019) searches for the best matching gene family using HMM scanning. It then finds the appropriate node at which the new sequence should be placed.
within the matched gene tree using the RAxML maximum likelihood method (Stamatakis, 2014) and inserts the new sequence at that node while retaining the original tree topology.

It is important to note that, although the tree construction method (Thomas, 2010) for the precomputed gene trees uses a species tree for gene tree reconciliation, this grafting method does not consider species tree information for a user-submitted sequence (Tang et al., 2019). As a result, the node that includes the grafted sequence sometimes does not resolve the complete evolutionary history of its children. For example, a user’s sequence from species X may be grafted to a duplication node of species Y.

2.4.4 | Download

Gene trees can be downloaded in standard PhyloXML format or as image files (PNG or SVG). The gene information table can be downloaded in CSV (comma-separated values) format for easy further manipulation and analysis.

2.5 | Comparison with other resources that provide gene family data in the context of comparative genomics

Online phylogenetics resources provide platforms for researchers to investigate gene family data without having to install software and carry out their own analysis to cluster gene families and reconstruct gene trees. There are several other databases that provide plant gene family data, each with their own set of features (Bolser et al., 2017; Goodstein et al., 2012; Kriventseva et al., 2019; Mi et al., 2019; Van Bel et al., 2018). Table 2 compares the resources based on the number of plant genomes included, the taxonomic coverage, whether phylogenetic trees are displayed, whether functional information is provided for individual genes and how this information is presented, and whether users are able to customize trees by hiding less relevant species or adding a sequence of interest to a tree in the appropriate position.

These resources can be distinguished by 1) how they represent gene families (tree vs. no tree), 2) how they display functional annotations, 3) inclusion of non-plant species and, 4) the ability to modify trees. Phytozome and OrthoDB provide lists of members of a gene family. Both also provide functional annotations. Phytozome presents gene functional annotations for each family member. The annotations are derived from various methods, but the supporting evidence is not provided. Therefore, it is not obvious to the user, which annotations are experimentally verified and which are computationally predicted, and the basis for the prediction. OrthoDB only displays functional annotations at the family level, also without showing evidence. PhyloGenes, PLAZA, Ensembl Plants and PANTHER all display phylogenetic trees of gene families. PLAZA and Ensembl Plants both display experimentally determined gene functions as GO terms with experimental evidence codes. However, this information is shown on pages separate from the gene trees. PhyloGenes displays experimentally determined gene functions directly adjacent to the tree as GO terms with experimental evidence code. The visual alignment of experimentally determined gene functions with the gene tree allows for easy evaluation of evidence and projection of function to genes lacking direct experimental evidence. While PhyloGenes and Ensembl Plants include non-plant model organisms in gene families, PLAZA does not, which prevents function inferences in cases where experimental evidence is available only for genes outside the plant kingdom. Only PhyloGenes and PLAZA enable users to remove sequences from a gene tree. These

| PhyloGenes | PANTHER | PLAZA | EnsemblPlants | Phytozome | OrthoDB |
|------------|---------|-------|---------------|-----------|---------|
| No. of plant genomes | 40 | 40 | 127 | 79 | 93 | 117 |
| No. of non-plant genomes | 10 | 102 | 0 | 5 | 0 | 7,167 |
| Displays phylogenetic tree | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| User can add sequences to tree | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| User can remove sequences from tree | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| Displays known function of family members next to tree | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| Other gene information displayed next to tree | MSA | MSA | protein domain, gene structure | aligned region | N/A | N/A |

aPhyloGenes and PANTHER can add user sequences (one at a time) by tree grafting without altering the tree’s original topology. PLAZA can add multiple user sequences to a tree by reconstructing the tree.
bPhyloGenes lets user remove sequences of species not of interest without changing the tree’s original topology. PLAZA removes sequences then reconstructs the tree.
cPLAZA and Ensembl Plants both display known gene functions as GO terms with experimental evidence code. However, the information is shown on pages separate from the gene trees.
two resources, along with PANTHER, also enable users to add a new sequence to tree.

2.6 | Using PhyloGenes to infer gene function and/or identify candidate genes

PhyloGenes provides a unique combination of precomputed gene families displayed as phylogenetic trees along with experimentally validated gene functions from model organisms not limited to plants. The platform maximizes the capacity for users to infer the function of uncharacterized genes and rigorously evaluate the evidence supporting such functional inference. It also enables users to discover evolutionary relationships between different gene functions. Below we describe two examples of how to use various features of PhyloGenes to predict the function of uncharacterized genes.

2.6.1 | Distinguishing between two putative functions

The first example uses an uncharacterized grape gene to show how PhyloGenes can be used to better infer function based on available experimental data from model organisms. The grape gene VIT_01s0010g03720 (ID used in PhyloGenes, Ensembl Plants, and UniProt) or GSVIVG01010497001 (ID used in PLAZA, Phytozome) is predicted to be a fatty acid-CoA ligase in Ensembl Plants (release 47) and UniProt (release 2020_02), acting upon medium-chain or long-chain fatty acids. For simplicity we will focus here on the long-chain fatty acid-CoA ligase prediction. PLAZA (release Dicots 4.5) and Phytozome (release 12) have also predicted long-chain fatty acid-CoA ligase activity as well as a 4-coumarate-CoA ligase activity. These are two very different enzyme activities acting on structurally very different substrates. It is not easy to trace the supporting evidence for the predictions nor evaluate the predictions from the above resources. Does the grape gene have both enzyme functions or just one, and if so, which one? How are the two functions evolutionarily related?

Searching in PhyloGenes (release 2.1) returns the PTHR24096 gene family that contains the desired grape gene (http://www.phylogenes.org/tree/PTHR24096, Figure 3). As expected, there is no experimentally characterized function for the grape gene (no flask icon, Figure 3, arrow a). However, its Arabidopsis ortholog 4CLL1 (AT1G62940) is a long-chain fatty acid CoA ligase (Figure 3, arrow b) based on evidence from a direct enzyme assay (Figure 3, arrow c). Using the ortholog conjecture we can infer that our grape gene of unknown function is a long-chain fatty acid-CoA ligase.

FIGURE 3  A use case illustrating how PhyloGenes can be used to predict gene function for a grape gene (gene name in red text on the gene tree). For simplicity, this figure shows a pruned tree view including only genes from grape, Arabidopsis and fruit fly. Genes from Arabidopsis and fruit fly in this family have been experimentally characterized. Red dotted arrows indicate the ancestral node where the annotated function is likely to have arisen.
interest, VIT_01s0010g03720, is a long-chain fatty acid CoA ligase. This inference is also made by the PAINT annotation shown on the screen.

Four other Arabidopsis genes (4CL1, 4CL2, 4CL3, and 4CL4) in the same gene family have been characterized as 4-coumarate CoA ligases (Figure 3, arrow d). These are paralogs of our grape gene of interest (related by duplication [Figure 3, arrow e]). We can also see that there are additional grape genes in this family (e.g. VIT_16s0039g02040, an ortholog of the Arabidopsis 4-coumarate CoA ligases) that appear likely to be 4-coumarate CoA ligases.

In this example, there is an additional evolutionary insight that can be gained from displaying the gene tree along with the characterized gene functions. Using all the genes characterized as long-chain fatty acid CoA ligases, including those from animals, we can trace this function back to the most recent common ancestor with that enzyme function (Figure 3, arrow f). Similarly, we can also locate the most recent common ancestor of the 4-coumarate CoA ligases (Figure 3, arrow g). This gives us the additional insight that the ancestral 4-coumarate CoA ligase gene is a more recent gene that evolved from a long-chain fatty acid CoA ligase after a gene duplication event (Figure 3, arrow e).

2.6.2 | Identification of candidate genes for wet lab analysis

One can also use PhyloGenes to identify candidate genes involved in a particular cellular process, for example, cell wall xylan biosynthesis genes in foxtail millet (Setaria italica). Extensive studies in Arabidopsis identified four homologous pairs of glycosyltransferases involved in the biosynthesis of xylan, with each pair composed of a functionally major and a minor form: irregular xylem9 (IRX9) and IRX9H, IRX10 and IRX10H, IRX14 and IRX14H, and FRA8 (IRX7) and F8H (Wu et al., 2010). Each pair performs a distinct role that is not interchangeable with any other pair. Double mutants of any of the four pairs showed a stunted plant phenotype. Searching PhyloGenes for ‘IRX9’ or ‘IRX14’ returns the gene family PTHR10896 (http://www.phylogen.org/tree/PTHR10896). This family has 329 genes from 44 organisms including foxtail millet. After pruning the gene tree to show only genes from Arabidopsis and foxtail millet, we can easily identify two foxtail millet IRX14/IRX14H orthologs, six IRX9/IRX9H orthologs (Figure 4a). Similarly, searching for IRX7 or IRX10 identified a gene family having one IRX7 ortholog and seven IRX10 orthologs in foxtail millet (http://www.phylogen.org/tree/PTHR1

**FIGURE 4** Screen shots of pruned tree views of gene families showing only foxtail millet and Arabidopsis genes. Foxtail millet orthologs of the Arabidopsis IRX9, IRX9H, and IRX14/IRX14H are indicated within red boxes, whereas paralogs of IRA9H are indicated within a blue box (a). Foxtail millet orthologs of the Arabidopsis IRX7 and IRX10 are indicated within red boxes in (b).
REFERENCES

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408, 796–815.

Bolser, D. M., Staines, D. M., Perry, E., & Kersey, P. J. (2017). Ensembl Arabidopsis Genome Initiative (2000). Analysis of the genome sequence to better facilitate inference of gene function for uncharacterized plant genes. Efforts are underway to extend the types of experimental gene function data to include gene expression atlases and publication lists, among others. We plan to regularly update the gene families and gene trees, substitute higher quality genomes when available and add a small number of additional genomes that fill in important taxonomic gaps, for example, ginkgo to provide coverage of gymnosperms.

ACKNOWLEDGMENTS

We thank the many researchers who participated in our focus groups, user feedback sessions, surveys, and beta testing, and Dr. Monica Munoz-Torres who organized and facilitated the initial planning session.

AUTHOR CONTRIBUTIONS

EH and PT conceived the project. PZ and TZB provided scientific supervision to the development of the PhyloGenes resource; SS and QL wrote code and developed the PhyloGenes database and website; TP provided technical supervision; AM, DE, HM and PT generated and provided the PANTHER data; PZ, TZB, LR and EH wrote the article. EH agrees to serve as the author responsible for contact and ensures communication.

ORCID

Tanya Z. Berardini https://orcid.org/0000-0002-3837-8864

Eva Huala https://orcid.org/0000-0003-4631-7241

Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Briefings in Bioinformatics*, 7, 225–242. https://doi.org/10.1093/bib/bbl004

Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, 12, 449–462. https://doi.org/10.1093/bib/bbr042

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Helfsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, 40, D1178–1186. https://doi.org/10.1093/nar/gkr944

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D’Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., Koo, D. C. E., Penfold-Brown, D., Shasha, D., Youngs, N., Bonneau, R., Lin, A., Sahraeiyan, S. M. E., Martelli, P. L., Profiti, G., ... Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17, 184. https://doi.org/10.1186/s13059-016-1037-6

Karunanithi, P. S., & Zerbe, P. (2019). Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Frontiers in Plant Science*, 10, 1166. https://doi.org/10.3389/fpls.2019.01166

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47, D807–D811. https://doi.org/10.1093/nar/gky1053

Lambing, C., & Heckmann, S. (2018). Tackling plant meiosis: From model research to crop improvement. *Frontiers in Plant Science*, 9, 829. https://doi.org/10.3389/fpls.2018.00829

Li, Y., Zhu, J., Wu, L., Shao, Y., Wu, Y., & Mao, C. (2019). Functional divergence of PIN1 paralogous genes in rice. *Plant and Cell Physiology*, 60, 2720–2732. https://doi.org/10.1093/pcp/pcz159

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47, D419–D426. https://doi.org/10.1093/nar/gky1038

Nehrt, N. L., Clark, W. T., Radivojac, P., & Hahn, M. W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, 7, e1002073. https://doi.org/10.1371/journal.pcbi.1002073

Piquerez, S. J. M., Harvey, S. E., Beynon, J. L., & Ntoukakis, V. (2014). Improving crop disease resistance: Lessons from research on Arabidopsis and tomato. *Frontiers in Plant Science*, 5, 671. https://doi.org/10.3389/fpls.2014.00671

Provart, N. J., Alonso, J., Assmann, S. M., Bergmann, D., Brady, S. M., Brkljacic, J., Browse, J., Chapple, C., Colot, V., Cutler, S., Dangel, J. (2016). 50 Years of Arabidopsis research: Highlights and future directions. *New Phytologist*, 209, 921–944.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Plovesan, D., Casadio, R., Wang, Z., Cheng, J., ... Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10, 221–227. https://doi.org/10.1038/nmeth.2340

Rhee, S. Y., & Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, 19, 212–221. https://doi.org/10.1016/j.tplants.2013.10.006

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5, e1000605. https://doi.org/10.1371/journal.pcbi.1000605
