Received: 16 September 2019 | Revised: 11 November 2019 | Accepted: 13 November 2019

DOI: 10.1002/pld3.191

ORIGINAL RESEARCH

GenFam: A web application and database for gene family-based classification and functional enrichment analysis

Renesh Bedre[1] | Kranthi Mandadi[1,2]

[1]Texas A&M AgriLife Research & Extension Center, Weslaco, TX, USA

[2]Department of Plant Pathology & Microbiology, Texas A&M University, College Station, TX, USA

Correspondence
Kranthi Mandadi, Department of Plant Pathology & Microbiology, Texas A&M AgriLife Research & Extension Center, Texas A&M University System, Weslaco, TX, USA.
Email: kkmandadi@tamu.edu

Abstract

Genome-scale studies using high-throughput sequencing (HTS) technologies generate substantial lists of differentially expressed genes under different experimental conditions. These gene lists need to be further mined to narrow down biologically relevant genes and associated functions in order to guide downstream functional genetic analyses. A popular approach is to determine statistically overrepresented genes in a user-defined list through enrichment analysis tools, which rely on functional annotations of genes based on Gene Ontology (GO) terms. Here, we propose a new computational approach, GenFam, which allows annotation, classification, and enrichment of genes based on their gene family, thus simplifying identification of candidate gene families and associated genes that may be relevant to the query. GenFam and its integrated database comprises of three hundred and eighty-four unique gene families and supports gene family analyses for sixty plant genomes. Four comparative case studies with plant species belonging to different clades and families were performed using GenFam which demonstrated its robustness and comprehensiveness over preexisting functional enrichment tools. To make it readily accessible for plant biologists, GenFam is available as a web-based application where users can input gene IDs and export enrichment results in both tabular and graphical formats. Users can also customize analysis parameters by choosing from the various statistical enrichment tests and multiple testing correction methods. Additionally, the web-based application, source code, and database are freely available to use and download. Website: http://mandadilab.webfactional.com/home/. Source code and database: http://mandadilab.webfactional.com/home/dload/.

KEYWORDS

data integration, database, gene family enrichment analysis, gene ontologies, software, statistics

This manuscript was previously deposited as a preprint at https://www.biorxiv.org/content/10.1101/272187v3

1 | INTRODUCTION

In recent years, genome-wide analyses using high-throughput sequencing (HTS) technologies have become indispensable to life science research. Generating large-scale datasets has become relatively straightforward, as opposed to efficiently interpreting the data to gain intuition into biologically significant mechanisms. Data mining tools that determine, predict, and enrich putative functions among HTS datasets are highly valuable for such genomic analyses (Backes et al., 2007). For instance, RNA-sequencing (RNA-seq) analysis is a high-throughput approach to study transcriptome regulation by determining transcript-level changes in multiple cell- or tissue-types, or among varying experimental conditions (e.g., unstressed vs. stressed). In a typical RNA-seq experiment, the analysis yields hundreds, if not thousands, of genes that are differentially expressed among the experimental conditions. Uncovering enriched biological pathways among these gene lists is a valuable starting step for downstream functional genetic analyses.

The Gene Ontology (GO)-term based enrichment tools (e.g., BinGO (Maere, Heymans, & Kuiper, 2005), Blast2GO (Conesa et al., 2005), AgriGO (Du, Zhou, Ling, Zhang, & Su, 2010), and PlantGSEA (Yi, Du, & Su, 2013)) are widely used by researchers to infer the biological mechanisms of genes identified in HTS experiments (Bedre et al., 2019, 2015; Bedre, Mangu, Srivastava, Sanchez, & Baisakh, 2016; Chen et al., 2013; Li, Dai, Hu, Liu, & Kang, 2017; Mandadi & Scholthof, 2012, 2015). These tools identify overrepresented GO terms associated within a user-defined list of genes by mapping them to the background genome annotations and calculating statistical probability of the enrichment relative to the background database. The enrichment tools can classify genes into GO categories or pathways related to biological process, molecular function, and cellular locations (Du et al., 2010; Goffard & Weiller, 2007). The GO-enrichment and the resultant hierarchy are very useful to understand the complex biological processes that are being enriched. However, information on specific biological attributes of a gene, such as the gene family (a group of homologous genes with common evolutionary origin and biological functions) level information, is hard to glean from GO-enrichment alone (Ashburner et al., 2000; Lee, Katari, & Sachidanandam, 2005). For instance, enrichment of a transcription factor will fetch GO terms for “regulation of transcription (GO:0006355)” or “DNA binding (GO:0003700)” or “response to stress (GO:0006950)” but does not identify which transcription factor family genes (e.g., WRKY and bZIP) being enriched. Having this information allows users to readily interpret large-scale datasets effectively and select favorite gene families for further functional studies. While providing the information for functional studies, gene families also could reveal the accurate gene annotation information that could not be easily determined by BLAST-based tools alone. Further, comparative gene family size analysis can certainly be informative and valuable approach to explore the biologically relevant functions related to genome architecture and adaptation or speciation of various plant species (Guo, 2013).

With the availability of complete genomes and sequence data, identification, and analysis of specific gene families among plant species has become necessary. In this study, we present a unique approach to perform annotation, classification, and enrichment of genes to identify overrepresented gene families (GenFam) in a user-defined query list. We suggest that GenFam is a valuable addition to a plant biologists toolkit to analyze large-scale HTS datasets. By determining overrepresented gene families in a user-defined gene list, rather than GO terms or hierarchy alone, GenFam empowers users to readily interpret information of gene families (e.g., WRKY and bZIP) in their queries, and move forward to selecting favorite overrepresented genes (or families) for downstream studies and interpretation. GenFam is also freely accessible to users on the World Wide Web, as a user-friendly, graphical-user interface.

2 | MATERIALS AND METHODS

2.1 | Background database

GenFam currently supports the analysis of sixty plant genomes. GenFam classifies genes into 384 representative and unique gene families, which to the best of our knowledge the largest collection, based on the well-annotated *Arabidopsis thaliana* (Berardini et al., 2015) and rice (*Oryza sativa*) (Kawahara et al., 2013) genomes, literature search, and Pfam protein families database (El-Gebali et al., 2019). We have identified and used Pfam common conserved domains and domain organization among the homologous gene sequences to assign the gene families. These highly conserved domains define protein functions and classify protein-coding genes into gene families. The conserved signature protein domains have the ability to detect the divergent or distantly related homologs which would be prohibitive with sequence-based similarity analysis tools [e.g., BLAST (Altschul et al., 1997)]. Therefore, domain-based search method would identify more genes belonging to gene families than BLAST-based homology search.

For GenFam implementation, we have leveraged publicly available plant genome resources of Phytozome (v12) and developed a curated database that serves as the background reference. All gene IDs within a user-defined input list are mapped to this reference database to assign genes into families, and subsequently overrepresented gene families in the input list are computed by comparing to the background database. For developing the database, the protein sequences of sixty plant genomes were used to identify conserved protein domains to assign families to known and unclassified or novel genes. The respective protein domains were predicted by HMMER (v3.1b2) using a protein family hidden Markov model (HMM) profiles (Pfam release 32.0) (El-Gebali et al., 2019). We have established rules to classify and assign the genes to gene families based on the presence of signature conserved protein domains and have provided in Table S1. This approach allowed us to maximize classification including orphan genes with missing annotations, genes with incorrect annotations, and novel genes present among the respective genome databases. Lastly, the background databases were curated to remove redundancy and duplication of gene members among families. In summary, we were able to integrate 384 representative gene families and corresponding (on an average ~41%) genes from sixty plant genomes into our database (Table S2). This is a comprehensive collection of gene families spanning sixty plant species, when compared to other existing databases. For instance, the recently published gene family database in poplar (GFDP) has classified 6,551 poplar genes into 145 gene families derived from Arabidopsis genome (Wang et al., 2018). PlantTFDB (v4.0) and PlnTFDB (v3.0) has classified the genes into 58 and 84 transcription factor gene families (Jin et al., 2017; Perez-Rodriguez et al., 2010). Similarly, another database and analysis toolkit, PlantGSEA, supports the gene family analysis for 13 plant species which mostly imports gene families from well-annotated genomes such as rice (118 gene families) and maize (81 gene families) (Yi et al., 2013).

All the gene family data were formatted using the PostgreSQL database to perform classification and enrichment analysis using various statistical enrichment methods. The GenFam database with complete protein domain annotation and gene family classification can be downloaded from the GenFam website (http://mandadilab.webfactional.com/home/dload/). Detailed statistics for the number of genes assigned to each gene family and the total number of background genes are provided in Table S2.

2.2 | Statistical enrichment methods

GenFam performs three main functions: (a) Annotation (b) classification, and (c) enrichment of a user-defined gene list to provide gene family-level attributes. The enrichment is based on the singular enrichment analysis (SEA) method, which computes enrichment of a user-defined list of genes with a precomputed background database (da Huang, Sherman, & Lempicki, 2009). GenFam accepts different types of gene IDs for the analysis. For example, for rice, it accepts gene (e.g., LOC_Os01g06882) and transcript (e.g., LOC_Os01g06882.1) IDs from parent database such as the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/). Additionally, GenFam also accepts Phytozome PAC IDs for a given gene (e.g., 24120792 for LOC_Os01g06882), which provides additional flexibility in performing the analysis. To determine an acceptable ID, the user can run the “check allowed ID type for each species” function on the GenFam analysis page (http://mandadilab.webfactional.com/family/). Once the appropriate gene IDs are provided, GenFam classifies and identifies specific gene families and members that are overrepresented in the input gene list.

Even though there is no defined standard for choosing a reference background, it is ideal to select a background that will increase coverage (or intersection) with an input gene list, as well as that enhances specificity of the enrichment analysis (da Huang et al., 2009). GenFam utilizes the number of total genes categorized/annotated into gene families in each plant species as a reference background, rather than using the whole genome. This feature greatly improves the specificity of the enrichment analysis by implementing statistically stringent criteria. For instance, for case study 1, if enrichment analysis was performed with the whole genome as background, it would result in 35 enriched gene families with much lower P-values, when compared to using the current GenFam background (29 enriched gene families) (Table S3).

GenFam can employ standard statistical tests such as the Fisher exact, chi-square (χ^2), binomial distribution, and hypergeometric tests for enrichment, along with multiple testing corrections to control a false discovery. We recommend using Fisher exact, chi-square (χ^2) and hypergeometric tests for smaller datasets (<1,000) (McDonald, 2009), and binomial distribution for larger datasets (Khatri & Draghici, 2005; Zheng & Wang, 2008). Furthermore, the chi-square (χ^2) test would be appropriate when the user-defined gene list has less overlap with the background database. As a default test, GenFam performs the Fisher exact test, which relies on the proportion of observed data, instead of a value of a test statistic to estimate the probability of genes of interest corresponding to a specific category.

To address the false positives resulting from multiple comparisons especially when the input gene list is large (>1,000), GenFam subsequently employs false discovery correction methods including the Benjamini-Hochberg (Benjamini & Hochberg, 1995), Bonferroni (1936) and Bonferroni-Holm (Holm, 1979). The various statistical tests and false discovery correction methods can be customized by the user as appropriate.

2.3 | Web server implementation

The GenFam web server is implemented using Python 3 (https://www.python.org/), Django 1.11.7 (https://www.djangoproject.com/), and PostgreSQL (https://www.postgresql.org/) database. All the codes for data formatting and statistical analysis are implemented using Python scripting language. Python is a fully fledged programming language which offers well-developed packages for statistical analysis, graphics, and integration with web apps. Therefore, we have chosen Python over other languages such as R for development of GenFam. The high-level Python web framework was constructed using Django. The Python web framework was hosted using WebFaction (https://www.webfaction.com/). The web-based templates were designed using Bootstrap, HTML, and CSS. The GenFam is compatible with all major browsers including Internet Explorer, Microsoft Edge, Google Chrome, Mozilla, and Safari. All the precomputed plant gene family background databases were built using advanced PostgreSQL database. The analyzed data were visualized using the matplotlib (Droettboom et al., 2016) Python plotting library.

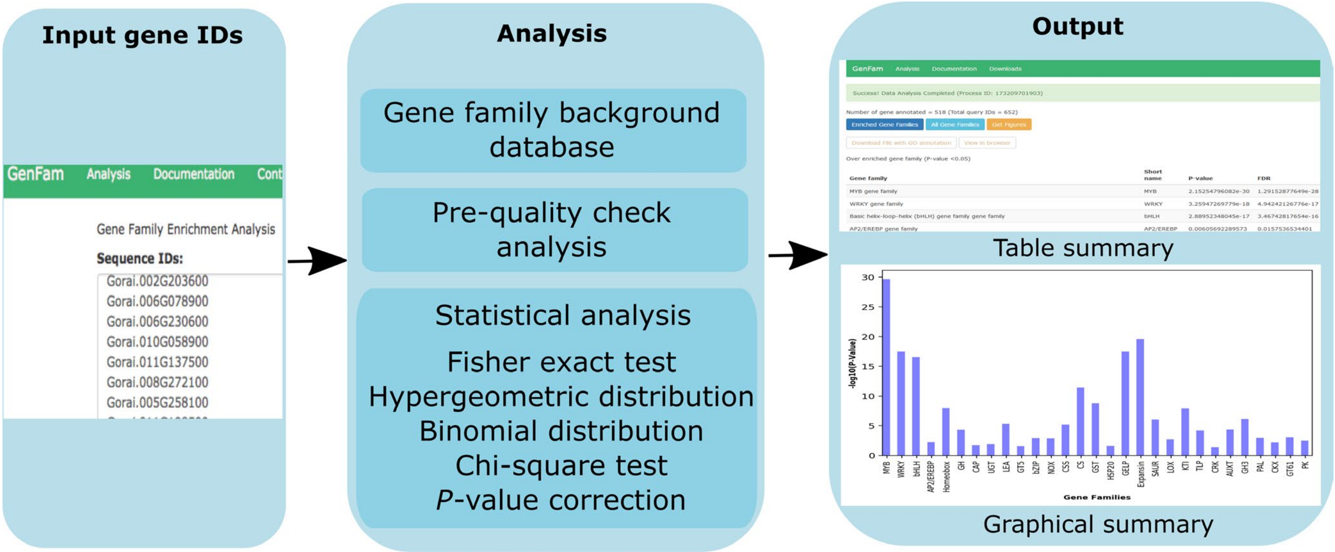
Along with enrichment results for the gene families, GenFam also provides information related to GO terms in biological process, molecular function and cellular component categories associated with the enriched gene families. In addition to GO terms, GenFam also provides the gene family size and gene IDs associated with each gene family. These results can be downloaded as a tabular file ("Enriched Families") or as a graphical figure of the enriched families ("Get Figures"). If users only want to retrieve the classification of genes, GenFam parses another tabular file containing the information of all annotated gene families ("All Families").

3 | RESULTS AND DISCUSSION

A snapshot of the analysis page and workflow is shown in Figure 1. Users have the option to either use the default settings or select desired statistical parameters. The analysis page also guides the users to select gene IDs that are acceptable in GenFam (Figure 1). Users are directed to the results after analysis is completed (Figure 1). The results of GenFam analysis are displayed as summary table (HTML) and graphical chart plotted using the -log$_{10}$(p-value) scores. Higher the -log$_{10}$(p-value) value, greater the confidence in enrichment of the gene family (Figure 2). The enriched and non-enriched gene family results can also be downloaded as tabular files, with further details of associated p-value and FDR statistics, gene family size, gene IDs, and GO terms.

3.1 | Case studies and analysis

To demonstrate the utility of GenFam, we performed four case studies using transcriptome datasets related to plants from different clades and families (cotton, tomato, soybean, and rice) (Bedre et al., 2015; Cui et al., 2018; Dametto et al., 2015; Zeng et al., 2017). We have previously identified 662 differentially expressed genes in cotton (*Gossypium raimondii*, family Malvaceae) infected with *Aspergillus flavus* (Bedre et al., 2015). For the first case study, we used GenFam to determine the enriched gene families among these differentially expressed genes, using the options of Fisher exact test for statistical enrichment, and the Benjamini-Hochberg (Benjamini & Hochberg, 1995) method to control false discovery rate (FDR). Among the 662 genes, 514 genes were annotated and classified into gene families, resulting in ~78% intersection/coverage with the GenFam database. The GenFam enrichment analysis revealed overrepresented gene families such as expansins, kinases, reactive oxygen species (ROS) scavenging enzymes, defense-related genes, heat shock proteins, and transcription factors—genes that we have hypothesized to mediate cell wall modifications, antioxidant activity, and defense signaling in response to *A. flavus* infection (Bedre et al., 2015). Additionally, GenFam also identified new enriched gene families such as bHLH, GH3, glycosyltransferases and thaumatin that were not reported or identified (Figures 1 and 2; Table S3). In the second case study, we analyzed 758 genes which were up-regulated in a cold-tolerant rice (*Oryza sativa*, family Poaceae) (Dametto et al., 2015). Among the 758 genes, 460 genes were annotated and classified into gene families by GenFam, resulting in ~61% intersection/coverage with the GenFam database. GenFam was able to successfully determine enriched gene families related to aquaporins, glutathione S-transferases (GST), transporters, lipid metabolism, transcription factors as well as gene families involved in cell wall-related mechanisms (Table S4)—genes that were hypothesized by Dametto et al. (2015) to play a role in the rice cold stress response. Additionally, GenFam also identified new enriched gene families such as aldehyde dehydrogenase (ADH), kinesins, glycosyltransferases, tubulin, phenylalanine ammonia-lyase (PAL), and thaumatin that were not reported or identified (Table S4). Next, we



Gene Family Enrichment Analysis

Sequence IDs:

Gorai.002G203600
Gorai.006G078900
Gorai.006G230600
Gorai.010G058900
Gorai.011G137500
Gorai.008G272100
Gorai.005G258100

Success! Data Analysis Completed (Process ID: 173209701903)

Number of gene annotated = 518 (Total query IDs = 652)

Over enriched gene family (P-value <0.05)

Gene family	Short name	P-value	FDR
MYB gene family	MYB	2.15254796082e-30	1.29152877649e-28
WRKY gene family	WRKY	3.25947269779e-18	4.94242126776e-17
Basic helix-loop-helix (bHLH) gene family gene family	bHLH	2.88952348045e-17	3.46742817654e-16
AP2/EREBP gene family	AP2/EREBP	0.00605692289573	0.0157536534401

FIGURE 1 GenFam workflow. The list of input gene IDs for respective plant species provided by the user is analyzed for enrichment analysis using various statistical tests. The output of the analysis can be viewed and/or downloaded as a table and/or graphical summary. The results page has multiple options to visualize or download data for both enriched and non-enriched categories (all gene families). The detailed output data from case studies are provided in Tables S3, S4, S5, and S6

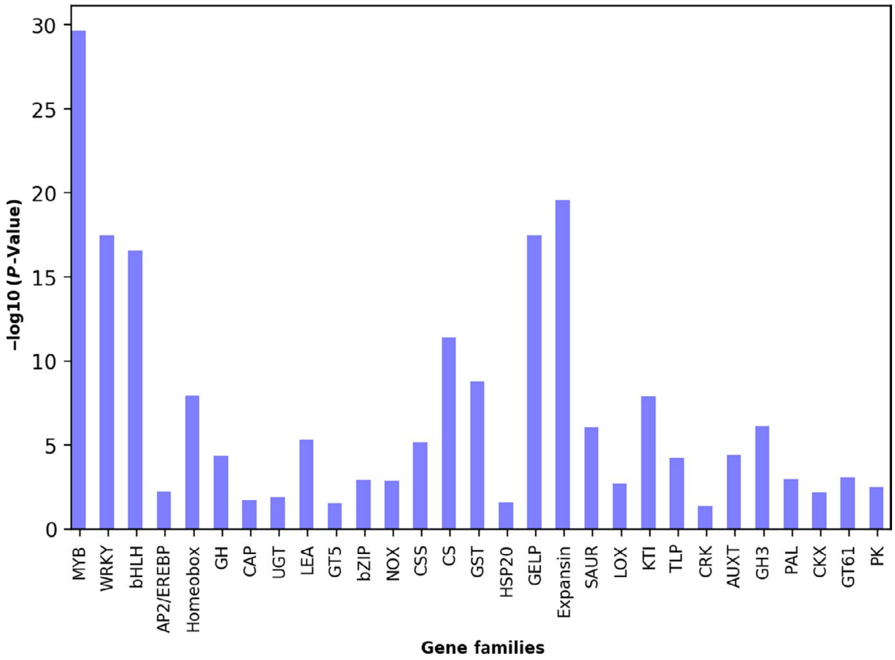


FIGURE 2 Graphical summary of GenFam enrichment analysis of a cotton case study. Results are plotted as bar chart using the $-\log_{10}(p\text{-value})$ scores. Higher the $-\log_{10}(p\text{-value})$ value, greater the confidence in enrichment of the gene family

analyzed the differentially regulated genes from tomato (*Solanum lycopersicum*, family Solanaceae) (Cui et al., 2018) and soybean (*Glycine max*, family Fabaceae) (Zeng et al., 2017) using GenFam (Table S5 and S6). We obtained ~65% and ~59% intersection/coverage with the GenFam database for tomato and soybean, respectively. The GenFam results in both these studies revealed enrichment of several gene families that were overrepresented and reported by Cui et al. (2018) (Cui et al., 2018) and Zeng et al. (2017) (Zeng et al., 2017) (Table S5 and S6). Additionally, GenFam also identified new enriched gene families such as aquaporins, VQ, tify, GST, and PAL in tomato, and BET, dirigent, expansins, asparagine synthase (ASNS), and carbonic anhydrase (CA) in soybean that were not reported or identified (Table S5 and S6). The detailed statistics of enriched gene families for these case studies are provided in Tables S3, S4, S5, and S6.

3.2 | GenFam advantages and comparison with preexisting enrichment tools

To the best of our knowledge, there is only one existing enrichment tool that comes close to the GenFam approach, that is, PlantGSEA (Yi et al., 2013), which also allows users to enrich gene lists using gene family attributes. Hence, we performed a comparative analysis of GenFam and PlantGSEA with a dataset from cotton (662 genes) (Bedre et al., 2015) and employing identical parameters (Fisher's exact test and Benjamini-Hochberg method) for enrichment. GenFam enriched gene families belonging to cell wall modifying genes, ROS scavenging genes, transcription factors, lipid metabolism, and stress-responsive gene families, both new and previously shown to be biologically relevant during *A. flavus* infection of cotton (Bedre et al., 2015), while PlantGSEA missed several of these categories (Table S3 and S7). Upon further examination, we found that several gene family categories such as the ABC transporters, expansins, and glutathione S-transferase were absent in the PlantGSEA *G. raimondii* background database. Moreover, PlantGSEA supports only thirteen plant genomes with several redundant and overlapping genes and gene families, which could impact the accuracy of the enrichment analysis. For instance, in the *A. thaliana* genome there are 37 annotated “C2-C2 Dof” transcription factors. PlantGSEA categorized 36 out of the 37 genes into a “C2-C2 Dof” family, but also into an additional “Dof” family leading to redundant gene family categories. GenFam avoids such discrepancies by curation and filtering redundant categories.

Taken together, we suggest that GenFam is a comprehensive and robust gene family classification and enrichment program over prevailing tools, with several advantages: (a) GenFam is a dedicated and comprehensive platform for gene family-level classification, annotation, and enrichment analysis and supports sixty plant genomes including model and non-model plant species. (b) GenFam background database was constructed from well-annotated gene families of *A. thaliana* and rice genomes, literature search, and as well as a systematic HMM profile search for signature conserved protein domain analysis using the Pfam database. This inclusive strategy enabled us to categorize most of the genes into families, including those which may lack a defined annotation in their corresponding genome database or could be novel genes. As a result, GenFam database is by far the largest collection of gene families (384 families). In contrast, existing databases such as PlantGSEA and GFDP only relies on annotations defined by other databases such as TAIR and MSU annotations and/or other transcription factor databases (Wang et al., 2018; Yi et al., 2013). The lack of additional analysis of protein domains perhaps explains the poor representation of gene families in PlantGSEA and GFDP databases. (c) GenFam background database was curated to remove redundancy and overlapping genes into different gene families that enhances the accuracy of the analysis. (d) In contrast to PlantGSEA, GenFam uses the annotated gene families as reference background instead of the whole genome. This feature ensures decreasing enrichment bias and increasing the accuracy of the analysis (da Huang et al., 2009). (e) GenFam accepts multiple input IDs including, gene IDs, transcript IDs, and PAC IDs; however, PlantGSEA and GFDP are restricted to using only gene IDs. (f) GenFam can be solely used for gene family annotation and classification regardless of enrichment analysis if a user is only interested in annotating genes.

4 | CONCLUSION

Data mining of big datasets (e.g., HTS data) is a very important step and approaches that can systematically mine biologically relevant information from big data are highly desirable. GO term-based enrichment analyses, although very useful to gain insight about the complex biological information, does not reveal specific gene family-level attributes or overrepresented gene families. GenFam can be used as a complementary or alternative approach to GO-based enrichment to interpret biologically relevant information in big datasets by classifying and enriching gene families within a user-defined gene list. This specific information on which gene families are overrepresented allows users to readily identify favorite genes for downstream inquiries. Along with enriching gene families, GenFam can be useful to annotate the large list of genes generated from HTS experiments irrespective of enrichment analysis. In conclusion, we suggest that GenFam would be a valuable and powerful tool for plant biologists utilizing genomics strategies to study plant biology and functional genetics.

ACKNOWLEDGMENTS

We thank Sonia Irigoyen (Texas A&M AgriLife Research) for review and comments during the preparation of this manuscript. The Bioinformatics and database analysis was conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. This work was supported in part by funds from Texas A&M AgriLife Research Insect-vectored Disease Seed Grant and USDA-NIFA-CDRE (2018-70016-28198) to KKM.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

AUTHOR CONTRIBUTIONS

RB conceived the project, developed the database/web server, performed the case studies, and prepared the manuscript. KKM supervised the study, analyzed and interpreted the data. Both authors have read, reviewed, and approved the manuscript.

AVAILABILITY AND REQUIREMENTS

Project name: GenFam

Project home page: http://mandadilab.webfactional.com/home/

Operating system(s): Platform independent

Programming language: Python 3, Django 1.11.7

License: CC BY-NC-ND 4.0

Any restrictions to use by non-academics: License needed.

ORCID

Renesh Bedre https://orcid.org/0000-0001-8874-5100
Kranthi Mandadi https://orcid.org/0000-0003-2986-4016

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/Nar/25.17.3389

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet*, *25*(1), 25–29. https://doi.org/10.1038/75556

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y. A., … Lenhof, H. P. 2007). GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Research*, *35*(Web Server issue), W186–W192. https://doi.org/10.1093/nar/gkm323.

Bedre, R., Irigoyen, S., Schaker, P. D. C., Monteiro-Vitorello, C. B., Da Silva, J. A., & Mandadi, K. K. (2019). Genome-wide alternative splicing landscapes modulated by biotrophic sugarcane smut pathogen. *Scientific Reports*, *9*(1), 8876. https://doi.org/10.1038/s41598-019-45184-1

Bedre, R., Mangu, V. R., Srivastava, S., Sanchez, L. E., & Baisakh, N. (2016). Transcriptome analysis of smooth cordgrass (*Spartina alterniflora* Loisel), a monocot halophyte, reveals candidate genes involved in its adaptation to salinity. *BMC Genomics*, *17*(1), 657. https://doi.org/10.1186/s12864-016-3017-3

Bedre, R., Rajasekaran, K., Mangu, V. R., Timm, L. E. S., Bhatnagar, D., & Baisakh, N. (2015). Genome-wide transcriptome analysis of cotton (*Gossypium hirsutum* L.) identifies candidate gene signatures in response to aflatoxin producing fungus *Aspergillus flavus*. *PLoS ONE*, *10*(9), e0138025. https://doi.org/10.1371/journal.pone.0138025

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, *53*(8), 474–485.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Firenze, Italy: Libreria internazionale Seeber.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., … Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*, 128. https://doi.org/10.1186/1471-2105-14-128

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Cui, J., Xu, P., Meng, J., Li, J., Jiang, N., & Luan, Y. (2018). Transcriptome signatures of tomato leaf induced by Phytophthora infestans and functional identification of transcription factor SpWRKY3. *Theoretical and Applied Genetics*, *131*(4), 787–800. https://doi.org/10.1007/s00122-017-3035-9

da Huang, W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

Dametto, A., Sperotto, R. A., Adamski, J. M., Blasi, É. A. R., Cargnelutti, D., de Oliveira, L. F. V. … Fett, J. P. (2015). Cold tolerance in rice germinating seeds revealed by deep RNAseq analysis of contrasting indica genotypes. *Plant Science*, *238*, 1–12. https://doi.org/10.1016/j.plantsci.2015.05.009

Droettboom, M., Hunter, J., Caswell, T., Firing, E., Nielsen, J., Elson, P., … Giuca, M. (2016). *matplotlib: matplotlib, v1. 5.1*. Retrieved from: https://doi.org/10.5281/zenodo.44579

Du, Z., Zhou, X., Ling, Y., Zhang, Z. H., & Su, Z. (2010). agriGO: A GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, *38*, W64–W70. https://doi.org/10.1093/nar/gkq310

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Goffard, N., & Weiller, G. (2007). PathExpress: A web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Research*, *35*, W176–W181. https://doi.org/10.1093/nar/gkm261

Guo, Y. L. (2013). Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. *The Plant Journal*, *73*(6), 941–951. https://doi.org/10.1111/tpj.12089

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., & Gao, G. E. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, *45*(D1), D1040–D1045. https://doi.org/10.1093/nar/gkw982

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., … Matsumoto, T. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, *6*, 4. https://doi.org/10.1186/1939-8433-6-4

Khatri, P., & Draghici, S. (2005). Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, *21*(18), 3587–3595. https://doi.org/10.1093/bioinformatics/bti565

Lee, J. S., Katari, G., & Sachidanandam, R. (2005). GObar: A gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, *6*, 189. https://doi.org/10.1186/1471-2105-6-189

Li, Y., Dai, C., Hu, C., Liu, Z., & Kang, C. (2017). Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *The Plant Journal*, *90*(1), 164–176. https://doi.org/10.1111/tpj.13462

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, *21*(16), 3448–3449. https://doi.org/10.1093/Bioinformatics/Bti551

Mandadi, K. K., & Scholthof, K. B. (2012). Characterization of a viral synergism in the monocot *Brachypodium distachyon* reveals distinctly altered host molecular processes associated with disease. *Plant Physiology*, *160*(3), 1432–1452. https://doi.org/10.1104/pp.112.204362

Mandadi, K. K., & Scholthof, K.-B.-G. (2015). Genome-wide analysis of alternative splicing landscapes modulated during plant-virus interactions in *Brachypodium distachyon*. *The Plant Cell*, *27*, 71–85. https://doi.org/10.1105/tpc.114.133991

McDonald, J. H. (2009). *Handbook of biological statistics*. Baltimore, MD: Sparky House Publishing.

Perez-Rodriguez, P., Riano-Pachon, D. M., Correa, L. G. G., Rensing, S. A., Kersten, B., & Mueller-Roeber, B. (2010). PlnTFDB: Updated content and new features of the plant transcription factor database. *Nucleic Acids Research*, *38*, D822–D827. https://doi.org/10.1093/nar/gkp805

Wang, H., Yan, H., Liu, H., Liu, R., Chen, J., & Xiang, Y. (2018). GFDP: The gene family database in poplar. *Database*, *2018*, bay107. https://doi.org/10.1093/database/bay107

Yi, X., Du, Z., & Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Research*, *41*(W1), W98–W103. https://doi.org/10.1093/nar/gkt281

Zeng, W., Sun, Z., Cai, Z., Chen, H., Lai, Z., Yang, S., & Tang, X. (2017). Comparative transcriptome analysis of soybean response to bean pyralid larvae. *BMC Genomics*, *18*(1), 871. https://doi.org/10.1186/s12864-017-4256-7

Zheng, Q., & Wang, X. J. (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, *36*(suppl_2), W358–W363. https://doi.org/10.1093/nar/gkn276

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Bedre R, Mandadi K. GenFam: A web application and database for gene family-based classification and functional enrichment analysis. *Plant Direct*. 2019;3:1–7. https://doi.org/10.1002/pld3.191