GenFam: A new web application for gene family-based classification and functional enrichment analysis of plant genomes

Renesh Bedre[1] and Kranthi Mandadi[1,2]*

[1]Texas A&M AgriLife Research & Extension Center, 2415 E. Hwy. 83, Weslaco, Texas 78596

[2]Department of Plant Pathology & Microbiology, Texas A&M University, 2132 TAMU, College Station, Texas 77843

*To whom correspondence should be addressed.

Running Title: Gene family-based enrichment analysis

ABSTRACT

Motivation: Genome-scale studies using next-generation sequencing technologies generate substantial number of differentially-regulated genes. The gene lists need to be further analyzed to identify overrepresented genes and functions in order to guide downstream analyses. Currently available gene enrichment tools rely on functional classifications based on Gene Ontology (GO) terms. A shortcoming of the GO-based classification system is that the GO terms are broad and often redundant, hence necessitating alternate approaches.

Results: We propose a new functional enrichment approach, GenFam, to classify as well as enrich overrepresented gene functions, based on gene family categories. GenFam offers a unique approach to mine valuable, biologically-relevant information, beyond the conventional GO term based enrichment. GenFam is available as a web-based, graphical-user interface, which allows users to readily input gene lists, and export results in both tabular and graphical formats. Additionally, users can customize analysis parameters, by choosing from the different significance tests to conduct advanced statistics. Currently, GenFam supports gene family classification and enrichment analyses for seventy-eight plant genomes and gene identifiers that are available on Phytozome v12.0 database.

Availability and implementation: The GenFam application is open-source and accessible through world-wide web at http://mandadilab.webfactional.com/home/

Contact: kkmandadi@tamu.edu

Supplementary information: Supplementary File 1 and 2

1 INTRODUCTION

In recent years, genome-wide analyses using next-generation sequencing (NGS) 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 NGS datasets are highly valuable for such genomic analyses (Backes *et al.*, 2007). For instance, RNA-sequencing (RNA-seq) analyses 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 genetic analyses.

The Gene Ontology (GO)-term based enrichment tools (e.g., BinGO, Blast2GO, AgriGO) are commonly used by researchers to infer the enriched pathways in NGS experiments (Bedre *et al.*, 2016; Bedre *et al.*, 2015; Chen *et al.*, 2013; Li *et al.*, 2017; Mandadi and Scholthof, 2015; Mandadi and Scholthof, 2012; Schaker *et al.*, 2016). 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 enrichment relative to the background. 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 and Weiller, 2007). However, the GO classifications are often broad and provide limited information on specific biological attributes of the gene (Ashburner *et al.*, 2000). For instance, GO terms in molecular function such as nucleic acid binding (GO:0003676) and DNA binding (GO:0003677) do not provide further information on the class of gene that is being enriched. Further, enriched GO terms can be redundant, that need to be manually filtered before interpretation. Given these shortcomings, new methods to analyze and interpret large-scale datasets to gain further insights into biologically-meaningful information are needed.

In this study, we present a unique approach to perform classification and enrichment analysis of genes, based on gene family (GenFam). The GenFam offers a meaningful way to determine pertinent gene functions by directly classifying and enriching genes, in a user-defined list, based on the encoded-protein and its associated gene family. We present GenFam as a user-friendly, graphical-user interface application that can be launched on the world-wide web.

2 IMPLEMENTATION AND DATA ANALYSIS

2.1 Background database

GenFam classifies and enriches genes into 128 representative and unique gene families, based on the well-annotated reference plant genome, *Arabidopsis thaliana* (Berardini *et al.*, 2015). GenFam currently supports analysis of genes from seventy-eight plant species. The background gene family database for the genomes was manually curated to remove redundancy among the families. Furthermore, we also determined a common protein domain structure for each gene family based on the protein sequences of the family members. The protein domains were predicted using HMMER (v3.1b2) from protein family database (Pfam release 31.0) (Eddy, 2009; Finn *et al.*, 2015). A multi-step annotation approach was used to classify gene sequences to a gene family. First, gene families were assigned based on their sequence similarity to Arabidopsis orthologs. Next, remaining sequences were assigned to a gene family based on their Pfam protein domain signature. All the selected 128 gene families, individual gene sequences, and corresponding gene IDs were formatted using the PostgreSQL database to perform classification and enrichment analysis using various statistical methods.

2.2 Statistical enrichment methods

GenFam provides two main functions: i) classification, and ii) enrichment of user-defined gene lists. The enrichment analysis is based on the singular enrichment analysis methods (Huang da *et al.*, 2009). In a manner similar to GO term enrichment tools (Backes, *et al.*, 2007; Du, *et al.*, 2010; Huang da, *et al.*, 2009), GenFam utilizes the user-defined gene IDs as input to perform statistical enrichment analysis. GenFam accepts different types of gene IDs for the analysis, as defined by the Phytozome database. For example, for rice, it accepts locus (LOC_Os01g06882), transcripts (LOC_Os01g06882.1) and PAC (24120792) IDs. To determine acceptable IDs for all plant species, user can use the “check allowed ID type for each species” function on the GenFam analysis page. Once the appropriate gene IDs are provided, GenFam classifies and identifies specific gene families and members that are overrepresented in the input gene lists. A unique feature of GenFam is that it only utilizes genes categorized to gene family as a reference background, unlike the GO enrichment tools which utilizes the entire genome as a reference background. This feature greatly enhances the sensitivity of the enrichment analysis. GenFam can employ rigorous statistical tests such as the Fisher exact, Chi-Square, Binomial distribution and hypergeometric tests, along with multiple test corrections to control family-wise error rate, in order to report the statistically significant enriched genes.

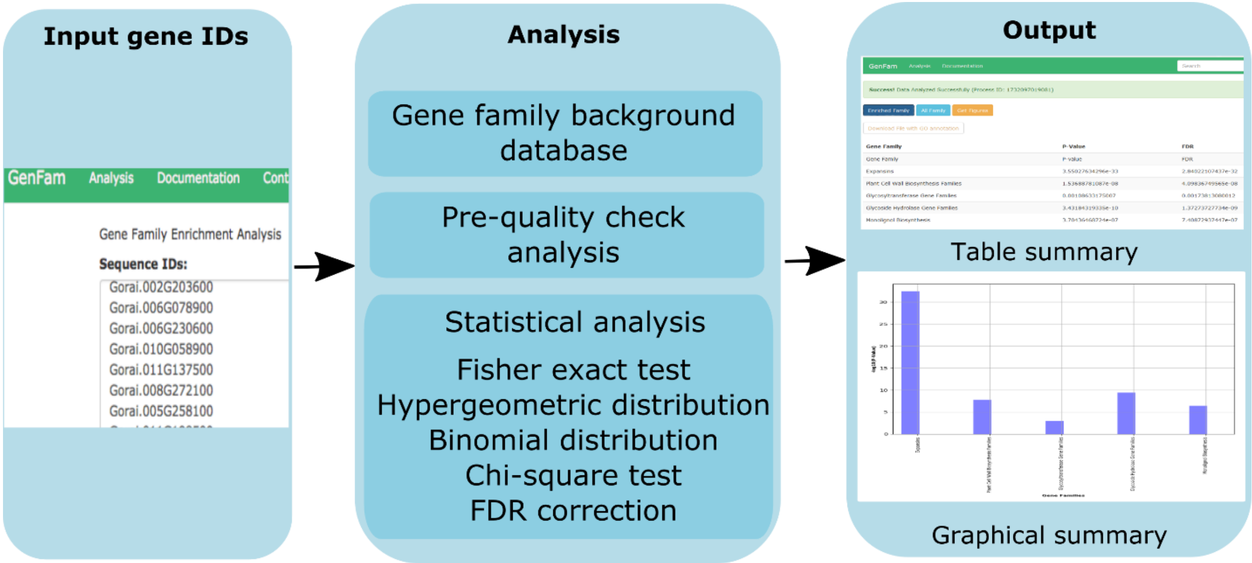
As a default test, GenFam performs the Fisher exact test, which relies on the proportion of observed data, instead of a value of a test statistic to estimate the probability of genes of interest corresponding to a specific category. For instance, suppose we have *n* differentially expressed genes, and among them, *k* falls in a particular gene family category, and there are *m* total genes associated with that gene family in the background reference database among *N* total genes; then Fisher probability that a given gene family is overrepresented in the input of gene list is calculated as,

$$p = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$$

To address the false positives resulting from multiple comparisons especially when the input gene list is large (>1000), GenFam subsequently employs false discovery rate (FDR) correction methods including the Benjamini-Hochberg (Benjamini and Hochberg, 1995), Bonferroni (Bonferroni, 1936) and Bonferroni-Holm (Holm, 1979). The various statistical tests and FDR methods can be customized by the user as appropriate. 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. 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 information of all the annotated gene families (“All Families”).

2.3 Web server implementation

The GenFam web server is implemented using Python3 (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. The high-level Python web framework was constructed using Django. The Django web framework was hosted using WebFaction (https://www.webfaction.com/). The web-based templates were designed using Bootstrap, HTML, and CSS. 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 was visualized using the matplotlib (Droettboom *et al.*, 2016) Python plotting library.



Success! Data Analyzed Successfully (Process ID: 1732097019081)

Gene Family	P-Value	FDR
Gene Family	P-value	FDR
Expansins	3.55027634296e-33	2.84022107437e-32
Plant Cell Wall Biosynthesis Families	1.53688781087e-08	4.09836749565e-08
Glycosyltransferase Gene Families	0.00108633175007	0.00173813080012
Glycoside Hydrolase Gene Families	3.43184319335e-10	1.37273727734e-09
Monolignol Biosynthesis	3.70436468724e-07	7.40872937447e-07

Fig. 1. GenFam workflow. The list of input gene IDs for respective plant species provided by the user are analyzed for enrichment analysis using various statistical tests. The ouput 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 a case study are provided in Supplementary Files 1 and 2.

2.4 Case study and data analysis

To demonstrate the utility of GenFam, we performed two case studies using cotton (a dicot) and rice (a monocot) transcriptome datasets (Bedre, *et al.*, 2015; Dametto *et al.*, 2015). We have previously identified ~662 differentially expressed genes in cotton infected with *Aspergillus flavus* (Bedre, *et al.*, 2015). For the first case study, we used GenFam to determine the enriched gene families among these 662 differentially expressed genes, using the options of Fisher exact

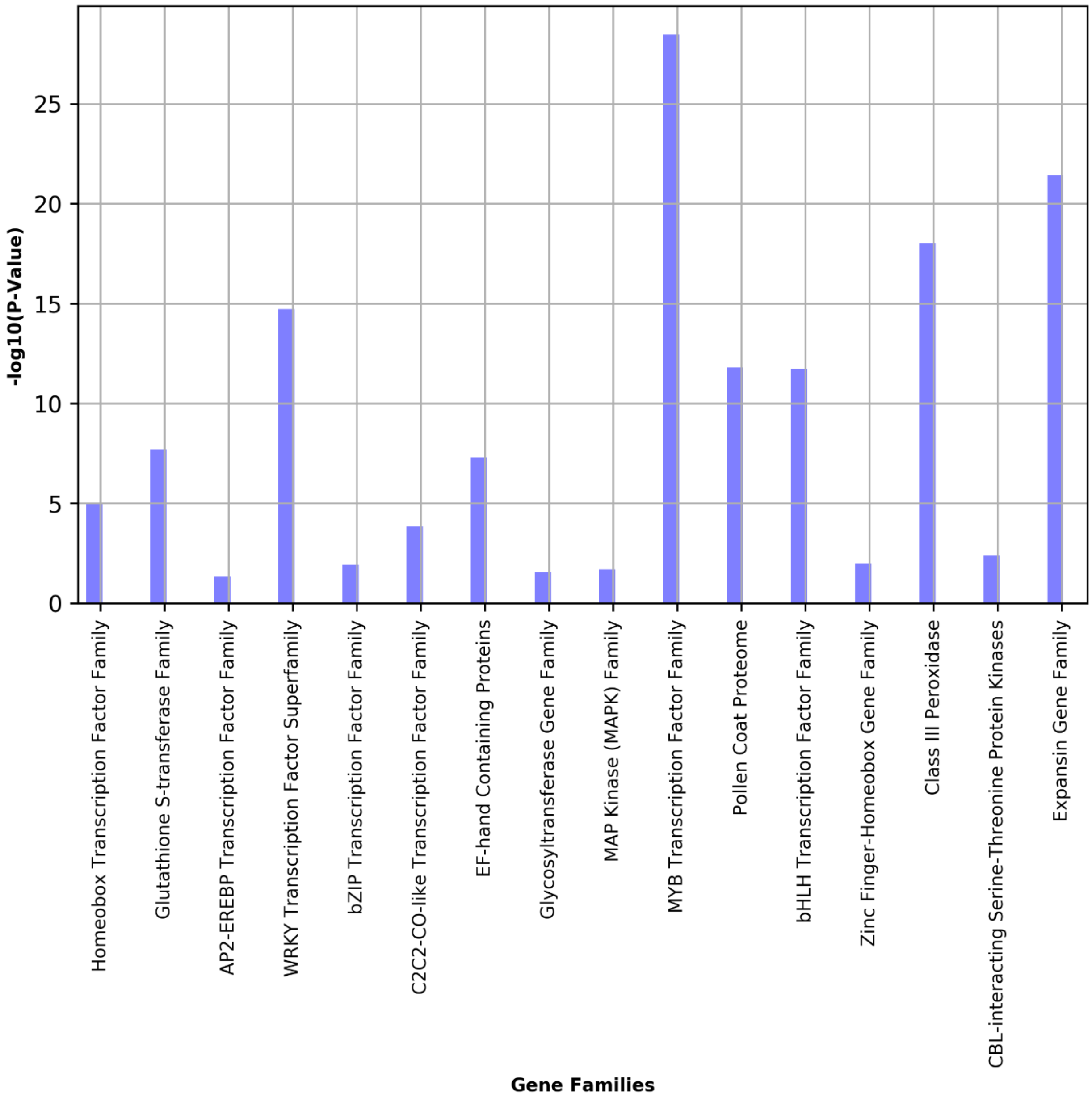


Fig. 2. Graphical summary of GenFam enrichment analysis of a cotton case study. Results are plotted as bar chart using the -$\log_{10}$(P-Value) scores. Higher the -$\log_{10}$(P-Value) value, greater the confidence in enrichment of the gene family.

test for statistical enrichment, and the Benjamini-Hochberg (Benjamini and Hochberg, 1995) method to control FDR. The GenFam classification and enrichment analysis revealed overrepresented gene families such as expansins, kinases, peroxidases, 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) **(Fig. 1** and **2; Supplementary File 1)**. In the second case study, we analyzed ~758 genes which were up-regulated in a cold-tolerant rice genotype (Dametto, *et al.*, 2015). GenFam was able to successfully classify and determine enriched gene families related to aquaporins, peroxidases, glutathione S-transferases, as well as gene families involved in cell wall-related mechanisms **(Supplementary File 2)** —genes that were hypothesized by Dametto *et al*. (2015) to play a role in the rice cold stress response. Together, the information of classified and enriched gene families not only provides understanding of the affected biological processes, but allows the user to readily select favorite gene families for further downstream characterization.

A snapshot of the analysis page and workflow is shown in **Fig. 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 **(Fig. 1)**. Users are directed to the results after analysis is completed **(Fig. 1)**.

2.5 Output summary

The results 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 **(Fig. 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, and GO terms.

3 DISCUSSION

Data mining of big datasets (e.g., NGS data) is a very important step, and approaches that can systematically dissect biologically-relevant information from big data are highly desirable. GO term-based enrichment analyses, although commonly employed, does not provide specific, biologically-relevant, gene family level information. Further, GO classifications can be broad and redundant. We suggest that GenFam is a unique way to extract biologically-relevant, gene family level information among large-scale results. GenFam allows users to readily uncover biologically-relevant functions enriched in large-scale gene datasets by classifying and providing specific information about the enriched gene families— information that could not be inferred by GO enrichment analysis alone. Furthermore, unlike GO enrichment tools, instead of using the whole genome as a background database for enrichment analysis, GenFam uses only genes annotated and classified into a gene family as a reference. This feature ensures decreasing enrichment bias and increasing the accuracy of the analysis (Huang da, *et al.*, 2009). GenFam can be implemented with various statistical enrichment methods such as Fisher exact test, hypergeometric distribution, chi-square test and binomial distribution, thus providing flexibility in the analysis based on the sample size and user preferences. We recommend using Fisher exact test, chi-square test and hypergeometric distribution for smaller datasets (< 1000) (McDonald, 2009), and binomial distribution for larger datasets (Khatri and Draghici, 2005; Zheng and Wang, 2008). To control the false positives, GenFam also supports multiple testing corrections (family-wise error rate) algorithms such as Benjamini-Hochberg (Benjamini and Hochberg, 1995), Bonferroni (Bonferroni, 1936), and Bonferroni-Holm (Holm, 1979).

In conclusion, we suggest that GenFam provides a unique approach to interpret biologically relevant information in big datasets by directly classifying and representing overrepresented genes into gene families. This allows users to readily interpret and identify favorite genes for downstream inquiries.

FUNDING

This work was supported by start-up funds from Texas A&M AgriLife Research to K.K.M.

ACKNOWLEDGEMENTS

We thank Dr. Sonia Irigoyen (Texas A&M AgriLife Research) for comments and feedback during the preparation of this manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

Supplementary File 1: List of the differentially regulated genes and analysis output of the cotton case study.

Supplementary File 2: List of the differentially regulated genes and analysis output of the rice case study.

REFERENCES

Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*,**25**,25-29.

Backes, C. et al. (2007) GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res.*,**35**,W186-192.

Bedre, R. et al. (2016) Transcriptome analysis of smooth cordgrass (*Spartina alterniflora* Loisel), a monocot halophyte, reveals candidate genes involved in its adaptation to salinity. *BMC Genomics*,**17**,657.

Bedre, R. et al. (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**.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)*,**57**,289-300.

Berardini, T.Z. et al. (2015) The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis*,**53**,474-485.

Bonferroni, C.E. Teoria statistica delle classi e calcolo delle probabilita. Libreria internazionale Seeber; 1936.

Chen, E.Y. et al. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*,**14**,128.

Dametto, A. et al. (2015) Cold tolerance in rice germinating seeds revealed by deep RNAseq analysis of contrasting indica genotypes. *Plant Sci.*,**238**,1-12.

Droettboom, M. et al. matplotlib: matplotlib v1. 5.1. In.: doi; 2016.

Du, Z. et al. (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.*,**38**,W64-W70.

Eddy, S.R. A new generation of homology search tools based on probabilistic inference. In, *Genome Informatics 2009: Genome Informatics Series Vol. 23*. World Scientific; 2009. p. 205-211.

Finn, R.D. et al. (2015) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*,**44**,D279-D285.

Goffard, N. and Weiller, G. (2007) PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res.*,**35**,W176-W181.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*,**6**,65-70.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*,**37**,1-13.

Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*,**21**,3587-3595.

Li, Y. et al. (2017) Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *Plant J.*,**90**,164-176.

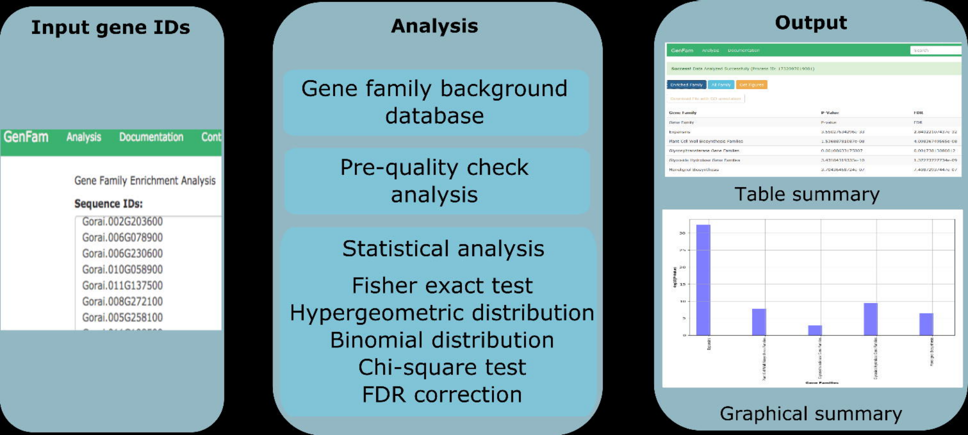
Mandadi, K.K. and Scholthof, K.-B.G. (2015) Genome-wide analysis of alternative splicing landscapes modulated during plant-virus interactions in *Brachypodium distachyon*. *Plant Cell*,**27**,71-85.

Mandadi, K.K. and Scholthof, K.B. (2012) Characterization of a viral synergism in the monocot *Brachypodium distachyon* reveals distinctly altered host molecular processes associated with disease. *Plant Physiol.*,**160**,1432-1452.

McDonald, J.H. Handbook of biological statistics. Sparky House Publishing Baltimore, MD; 2009.

Schaker, P.D. et al. (2016) RNAseq transcriptional profiling following whip development in sugarcane smut disease. *PLoS One*,**11**,e0162237.

Zheng, Q. and Wang, X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*,**36**,W358-363.



Input gene IDs

GenFam Analysis Documentation Cont

Gene Family Enrichment Analysis

Sequence IDs:

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

Analysis

Gene family background database

Pre-quality check analysis

Statistical analysis

Fisher exact test
Hypergeometric distribution
Binomial distribution
Chi-square test
FDR correction

Output

GenFam Analysis Documentation Search

Success! Data Analyzed Successfully (Process ID: 17320970190B1)

Enriched Family | All Family | Get Figures

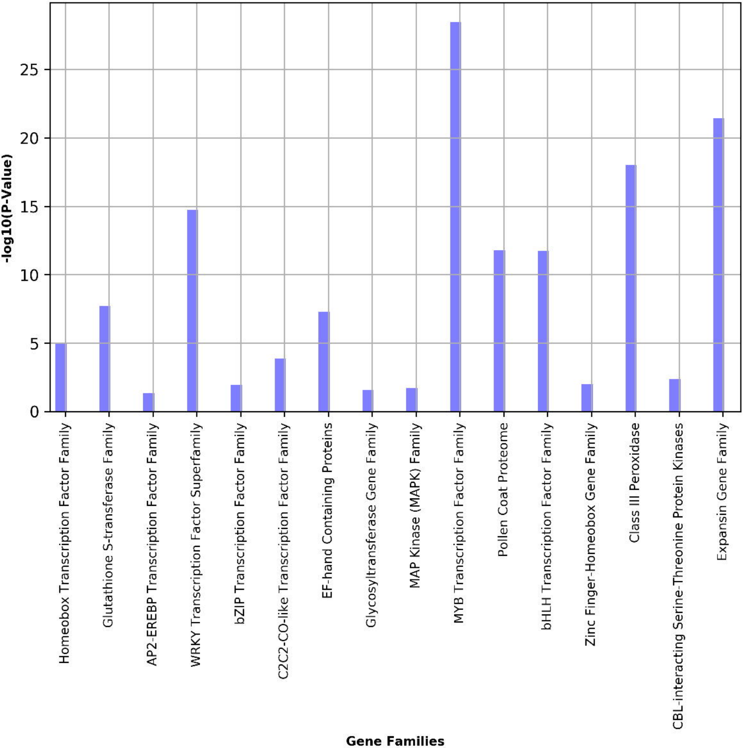
Download File with GO annotation

Gene Family	P-Value	FDR
Gene Family	P-value	FDR
Expansins	3.55027634296e-33	2.84022107437e-32
Plant Cell Wall Biosynthesis Families	1.53688781087e-08	4.09836749565e-08
Glycosyltransferase Gene Families	0.00109633175007	0.00173813080012
Glycoside Hydrolase Gene Families	3.43184319333e-10	1.37273727734e-09
Monolignol Biosynthesis	3.70436468724e-07	7.40872937447e-07

Table summary

-log10(P-value)
30
25
20
15
10
5
0
Expansins
Plant Cell Wall Biosynthesis Families
Glycosyltransferase Gene Families
Glycoside Hydrolase Gene Families
Monolignol Biosynthesis
Gene Families

Graphical summary


-log10(P-Value)
0
5
10
15
20
25
Homeobox Transcription Factor Family
Glutathione S-transferase Family
AP2-EREBP Transcription Factor Family
WRKY Transcription Factor Superfamily
bZIP Transcription Factor Family
C2C2-CO-like Transcription Factor Family
EF-hand Containing Proteins
Glycosyltransferase Gene Family
MAP Kinase (MAPK) Family
MYB Transcription Factor Family
Pollen Coat Proteome
bHLH Transcription Factor Family
Zinc Finger-Homeobox Gene Family
Class III Peroxidase
CBL-interacting Serine-Threonine Protein Kinases
Expansin Gene Family
Gene Families