GenFam: A 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

renesh.bedre@agnet.tamu.edu; kkmandadi@tamu.edu*

*To whom correspondence should be addressed.

Running Title: Gene family-based enrichment analysis

ABSTRACT

Background: Genome-scale studies using next-generation sequencing technologies generate substantial lists of differentially-expressed 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 robust 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.

Conclusions: GenFam implements a unique approach to classify and interpret enriched biological information in terms of gene families from user input of differentially expressed gene datasets**.** The GenFam application is open-source and accessible through world-wide web at http://mandadilab.webfactional.com/home/.

Keywords

Gene families, enrichment analysis, gene ontologies, web based tools, next-generation sequencing, software, databases

BACKGROUND

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 [1]. 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 [2], Blast2GO [3], AgriGO [4], PlantGSEA [5]) are commonly used by researchers to infer the enriched pathways in NGS experiments [6-12]. 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 [4, 13]. However, the GO classifications are often broad and provide limited information on specific biological attributes of the gene [14]. 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. Furthermore, with the availability of complete genomes and sequence data, functional identification, and analysis of specific gene families among plant species has become vital. In this study, we present a unique approach to perform classification and enrichment analysis of genes, based on gene family (GenFam), as a user-friendly, graphical-user interface application that can be launched on the world-wide web. GenFam offers a meaningful and comprehensive 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.

IMPLEMENTATION AND DATA ANALYSIS

Background database

GenFam currently supports analysis of seventy-eight plant genomes. GenFam classifies and enriches genes into 128 representative and unique gene families (superfamilies), based on the well-annotated reference plant genome, *Arabidopsis thaliana* [15]. 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 *A. thaliana* orthologs. As it is well-known that genes within a family possess conserved protein domains, we also used the Pfam common signature domain approach to assign families to 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 31.0) [16, 17]. This combined approach allowed us to maximize classification including orphan genes with missing annotations or novel genes present among the respective genome databases. Lastly, the background databases were manually

curated to remove redundancy and duplication of the gene members among families. All the representative 128 gene families, individual genes of the seventy-eight genomes, and their corresponding gene IDs were formatted using the PostgreSQL database to perform classification and enrichment analysis using various statistical methods.

Statistical enrichment methods

GenFam provides two main functions: i) classification, and ii) enrichment of user-defined gene lists. The enrichment is based on the very efficient singular enrichment analysis (SEA) method, which computes enrichment of a user-defined list of genes with a precomputed background dataset [18]. 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 analysis. To determine an acceptable ID, user can run “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 list. A unique feature of GenFam is that it only utilizes genes categorized to gene family as a reference background, unlike some of the existing GO enrichment tools which utilizes the entire genome as a reference background. This feature greatly enhances the sensitivity of the enrichment analysis [18].

GenFam can employ rigorous statistical tests such as the Fisher exact, Chi-Square, Binomial distribution and Hypergeometric tests for enrichment, along with multiple testing corrections to control family-wise error rate. We recommend using Fisher exact, Chi-square and

Hypergeometric tests for smaller datasets (<1000) [19], and Binomial distribution for larger datasets [20, 21]. Furthermore, Chi-Square test would be appropriate when the user defined gene list has less overlap with background dataset. 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 [22], Bonferroni [23] and Bonferroni-Holm [24]. 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").

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 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 [25] Python plotting library.

RESULTS AND DISCUSSION

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 [7, 26]. We have previously identified ~662 differentially expressed genes in cotton (*G. raimondii*) infected with *Aspergillus flavus* [7]. 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 test for statistical enrichment, and the Benjamini-Hochberg [22] 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 [7] **(Figures 1** and **2; Additional file 1)**. In the second case study, we analyzed ~758 genes which were up-regulated in a cold-tolerant rice genotype [26]. 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 **(Additional 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 **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)**.

Output summary

The results of GenFam 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, and GO terms.

Advantages over existing enrichment tools

To the best of our knowledge, there is only one existing GO enrichment tool, PlantGSEA [5], which also allows users to enrich genes based on gene family. We performed a comparative

enrichment analysis using GenFam and PlantGSEA with a dataset from cotton (662 genes)[7] and employing identical parameters (Fisher exact test and Benjamini-Hochberg method) for enrichment. GenFam accurately enriched important gene families such as expansins, glutathione-S-transferase and peroxidase, previously shown to be biologically-relevant during *A. flavus* infection of cotton [7], while PlantGSEA missed several of these categories **(Table 1, Additional file 3)**. Upon further examination, we found that several gene family categories such as the ABC transporters, expansins, glutathione-S-transferase and peroxidases, were missing in PlantGSEA *G. raimondii* background database. Moreover, PlantGSEA databases had several redundant and overlapping genes and gene families, which could impact the accuracy of the enrichment analysis **(Table 1, Additional file 3)**. For instance, in *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 manual 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:

1) GenFam is a dedicated and comprehensive platform for gene family enrichment and supports seventy-eight plant genomes including model and non-model plant species. PlantGSEA supports only thirteen genomes.

2) GenFam background dataset was constructed from well-annotated gene families of *A. thaliana,* as well as a systematic HMM profile search for signature protein domain analysis using 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. In contrast, PlantGSEA background database only relies on existing

annotations defined by other databases such as TAIR and MSU annotations and other transcription factor databases [5]. The lack of additional analysis of protein domains perhaps explains the poor representation of gene families in PlantGSEA databases.

3) GenFam background dataset was manually curated to remove redundancy and overlapping genes into different gene families, that enhances the accuracy of the analysis. In contrast PlantGSEA appears to be not manually curated, thus leading to redundant classifications and ambiguous analysis.

4) In contrast to PlantGSEA, GenFam uses the annotated gene families as reference background instead of whole genome. This feature ensures decreasing enrichment bias and increasing the accuracy of the analysis [18].

5) GenFam accepts multiple input IDs including, gene IDs, transcript IDs and PAC IDs, however PlantGSEA is restricted to using only gene IDs.

CONCLUSION

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. 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. In conclusion, we suggest that GenFam would be valuable and powerful tool for researchers utilizing genomics-based strategies to gain biological insights.

List of abbreviations: GO: Gene Ontology; NGS: Next-Generation Sequencing; RNA-seq: RNA-sequencing; BiNGO: Biological Networks Gene Ontology; PlantGSEA: Plant Gene Set Enrichment Analysis; DNA: Deoxyriboucleic acid; HMM: Hidden Markov Model; Pfam: Protein Families; SEA: Singular Enrichment Analysis; FDR: False Discovery Rate; HTML: Hyper Text Markup Language; CSS: Cascading Style Sheets; ABC: ATP-binding cassette; Dof: DNA-binding One Zinc Finger; TAIR: The Arabidopsis Information Resource; MSU: Michigan State University; bZIP: Basic Leucine Zipper; MYB: Myeloblastosis; bHLH: basic helix-loop-helix; MAPK: Mitogen-Activated Protein Kinase; AP2-EREBP: APETALA2/Ethylene-Responsive Element Binding Protein; CBL: Calcineurin B-like.

AVAILABILITY AND REQUIREMENTS

Project name: GenFam

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

Operating system(s): Platform independent

Programming language: Python 3, Django

License: CC BY-NC-ND 4.0

Any restrictions to use by non-academics: License needed

DECLARATIONS

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

Author's contributions: RB conceived the project, developed the database/webserver, performed the case studies and prepared the manuscript. KKM supervised the study and reviewed the manuscript for submission. All authors have read and approved the manuscript.

Acknowledgements: We thank Sonia Irigoyen (Texas A&M AgriLife Research) for review and comments during the preparation of this manuscript. All experiments were conducted following the guidelines and appropriate permissions of Institutional Biosafety Committee of Texas A&M University System.

Availability of data and material: All the relevant data is presented within the manuscript. Source code and database used to build GenFam can be downloaded from https://drive.google.com/drive/folders/1nQGa0A6ZGqfwBIUxe4tTVnSkWqXBKtCP?usp=sharin

Competing interests: The authors declare no competing financial interests.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

REFERENCES

1. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail--advanced gene set enrichment analysis. Nucleic Acids Res. 2007;35(Web Server issue):W186-92.
2. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics. 2005;21(16):3448-9.

3. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674-6.

4. Du Z, Zhou X, Ling Y, Zhang ZH, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38:W64-W70.

5. Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res. 2013;41(Web Server issue):W98-103.

6. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.

7. Bedre R, Rajasekaran K, Mangu VR, Timm LES, Bhatnagar D, Baisakh N. Genome-wide transcriptome analysis of cotton (*Gossypium hirsutum* L.) identifies candidate gene signatures in response to aflatoxin producing fungus *Aspergillus flavus*. PLoS One. 2015;10(9).

8. Bedre R, Mangu VR, Srivastava S, Sanchez LE, Baisakh N. Transcriptome analysis of smooth cordgrass (*Spartina alterniflora* Loisel), a monocot halophyte, reveals candidate genes involved in its adaptation to salinity. BMC Genomics. 2016;17(1):657.

9. Mandadi KK, Scholthof K-BG. Genome-wide analysis of alternative splicing landscapes modulated during plant-virus interactions in *Brachypodium distachyon*. Plant Cell. 2015;27:71-85.

10. Mandadi KK, Scholthof KB. Characterization of a viral synergism in the monocot *Brachypodium distachyon* reveals distinctly altered host molecular processes associated with disease. Plant Physiol. 2012;160(3):1432-52.

11. Schaker PD, Palhares AC, Taniguti LM, Peters LP, Creste S, Aitken KS, et al. RNAseq transcriptional profiling following whip development in sugarcane smut disease. PLoS One. 2016;11(9):e0162237.

12. Li Y, Dai C, Hu C, Liu Z, Kang C. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. Plant J. 2017;90(1):164-76.

13. Goffard N, Weiller G. PathExpress: a web-based tool to identify relevant pathways in gene expression data. Nucleic Acids Res. 2007;35:W176-W81.

14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

15. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. Genesis. 2015;53(8):474-85.

16. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2015;44(D1):D279-D85.

17. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Informatics 2009: Genome Informatics Series Vol 23: World Scientific; 2009. p. 205-11.

18. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.

19. McDonald JH. Handbook of biological statistics: Sparky House Publishing Baltimore, MD; 2009.

20. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. 2008;36(Web Server issue):W358-63.

21. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005;21(18):3587-95.

22. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Stat Method). 1995;57(1):289-300.

23. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilita: Libreria internazionale Seeber; 1936.

24. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65-70.

25. Droettboom M, Hunter J, Caswell T, Firing E, Nielsen J, Elson P, et al. matplotlib: matplotlib v1. 5.1. doi; 2016.

26. Dametto A, Sperotto RA, Adamski JM, Blasi EA, Cargnelutti D, de Oliveira LF, et al. Cold tolerance in rice germinating seeds revealed by deep RNAseq analysis of contrasting indica genotypes. Plant Sci. 2015;238:1-12.

FIGURE LEGENDS

Figure 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 Additional files 1 and 2.

Figure 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.

TABLES

Table 1: Comparison of enriched gene family (FDR < 0.05) catgories by GenFam and PlantGSEA using a *G. raimondii* transcriptome dataset [7].

Gene families (enriched gene families from GenFam)	GenFam	PlantGSEA
WRKY Transcription Factor Superfamily	Enriched	Enriched
bZIP Transcription Factor Family	Enriched	Enriched
MYB Transcription Factor Family	Enriched	Enriched
bHLH Transcription Factor Family	Enriched	Enriched
ZF-HD Transcription Factor Family	Enriched	Enriched
MAP Kinase (MAPK) Gene Family	Enriched	Enriched
AP2-EREBP Transcription Factor Family	Enriched	Enriched
*Class III peroxidase Family	Enriched	Not enriched
*Glutathione S-transferase Family	Enriched	Not enriched
Monolignol Biosynthesis Family	Enriched	Not enriched
*Expansin Gene Family	Enriched	Not enriched
*Glycoside Hydrolase Gene Family	Enriched	Not enriched
*Glycosyltransferase Gene Family	Enriched	Not enriched
Homeobox Transcription Factor Family	Enriched	Enriched
Receptor kinase-like protein family	Enriched	Enriched

		(classified into different subfamilies; see additional file 3)
CBL-interacting serione-threonine Protein KinaseFamily	Enriched	Enriched (classifed as *SNF1_RELATED_PROTEIN_KINASE)*
C2C2-CO-like Transcription Factor Family	Enriched	Not enriched
CYTOCHROME_P450_CYP74A5	Enriched (classified into monolignol biosynthesis family)	Enriched
CYTOCHROME_P450_CYP74A	Not enriched	Enriched
TRANSCRIPTION_REGULATOR_ORPHANS	Enriched (classifed into one of the above gene families)	Enriched (classifed as orphan gene family)

*Families unclassified in PlantGSEA *G. raimondii* background database

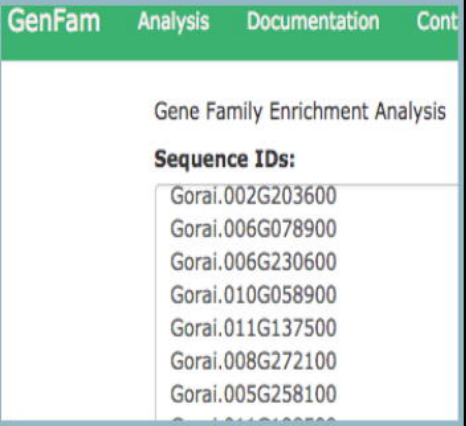
ADDITIONAL MATERIAL

Additional file 1: List of the differentially regulated genes and analysis output of the cotton case study

Additional file 2: List of the differentially regulated genes and analysis output of the rice case study

Additional file 3: PlantGSEA result for gene family enrichment analysis using *G. raimondii* dataset used in GenFam case study [7].

Input gene IDs



Analysis

Gene family background database

Pre-quality check analysis

Statistical analysis

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

Output

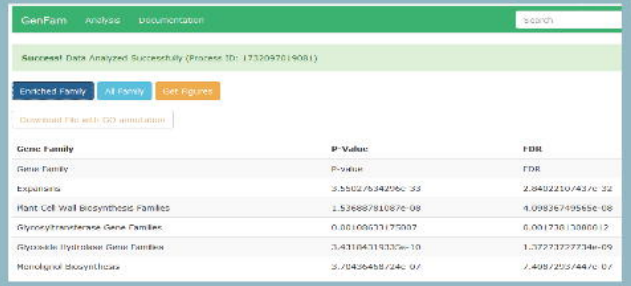
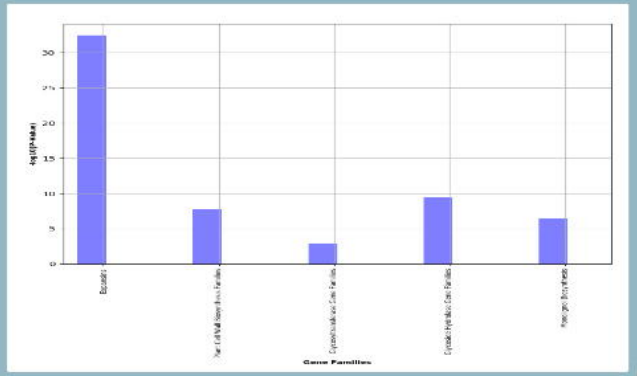
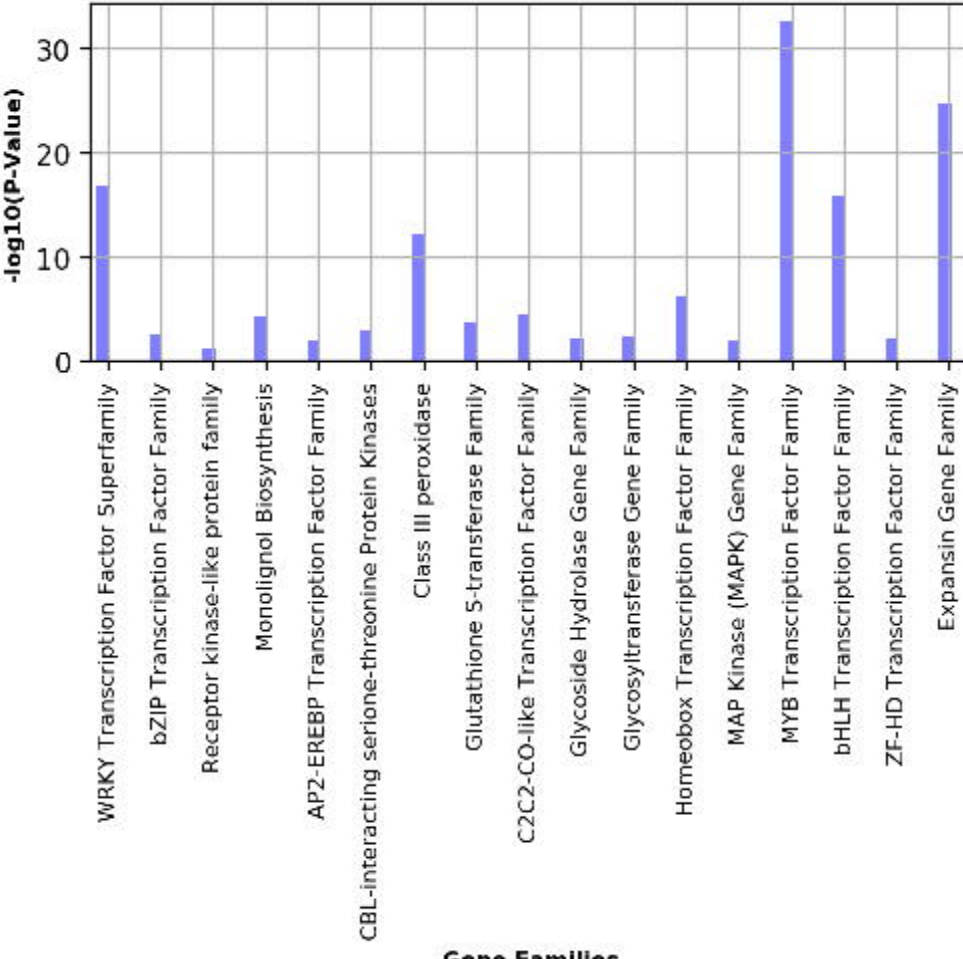


Table summary



Graphical summary


-log10(P-Value)
0
10
20
30
WRKY Transcription Factor Superfamily
bZIP Transcription Factor Family
Receptor kinase-like protein family
Monolignol Biosynthesis
AP2-EREBP Transcription Factor Family
CBL-interacting serione-threonine Protein Kinases
Class III peroxidase
Glutathione S-transferase Family
C2C2-CO-like Transcription Factor Family
Glycoside Hydrolase Gene Family
Glycosyltransferase Gene Family
Homeobox Transcription Factor Family
MAP Kinase (MAPK) Gene Family
MYB Transcription Factor Family
bHLH Transcription Factor Family
ZF-HD Transcription Factor Family
Expansin Gene Family
Gene Families