

1 **GenFam: A web application and database for gene family-based** 2 **classification and functional enrichment analysis**

3
4 Renesh Bedre¹ and Kranthi Mandadi^{1,2*}

5
6 ¹Texas A&M AgriLife Research & Extension Center, Weslaco, TX, USA

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

9
10 *Correspondence:

11 Kranthi Mandadi

12 kkmandadi@tamu.edu

13 14 **ABSTRACT**

15
16 Genome-scale studies using high-throughput sequencing (HTS) technologies generate substantial
17 lists of differentially expressed genes under different experimental conditions. These gene lists
18 need to be further mined to narrow down biologically relevant genes and associated functions in
19 order to guide downstream functional genetic analyses. A popular approach is to determine
20 statistically overrepresented genes in a user-defined list through enrichment analysis tools, which
21 rely on functional annotations of genes based on Gene Ontology (GO) terms. Here, we propose a
22 new approach, GenFam, which allows classification and enrichment of genes based on their gene
23 family, thus simplifying identification of candidate gene families and associated genes that may
24 be relevant to the query. GenFam and its integrated database comprises of three-hundred and
25 eighty-four unique gene families and supports gene family classification and enrichment
26 analyses for sixty plant genomes. Four comparative case studies with plant species belonging to
27 different clades and families were performed using GenFam which demonstrated its robustness
28 and comprehensiveness over preexisting functional enrichment tools. To make it readily
29 accessible for plant biologists, GenFam is available as a web-based application where users can
30 input gene IDs and export enrichment results in both tabular and graphical formats. Users can
31 also customize analysis parameters by choosing from the various statistical enrichment tests and
32 multiple testing correction methods. Additionally, the web-based application, source code and
33 database are freely available to use and download. Website:

34 <http://mandadilab.webfactional.com/home/>. Source code and database:

35 <http://mandadilab.webfactional.com/home/dload/> .

36 37 38 **KEYWORDS**

39
40 Gene family enrichment analysis, gene ontologies, database, software, statistics, data integration
41
42
43
44
45
46

47 INTRODUCTION

48

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

61

62 The Gene Ontology (GO)-term based enrichment tools (e.g., BinGO (Maere et al., 2005),
63 Blast2GO (Conesa et al., 2005), AgriGO (Du et al., 2010), PlantGSEA (Yi et al., 2013)) are
64 widely used by researchers to infer the biological mechanisms of genes identified in HTS
65 experiments (Mandadi and Scholthof, 2012; Chen et al., 2013; Bedre et al., 2015; Mandadi and
66 Scholthof, 2015; Bedre et al., 2016; Li et al., 2017; Bedre et al., 2019). These tools identify
67 overrepresented GO terms associated within a user-defined list of genes by mapping them to the
68 background genome annotations and calculating statistical probability of the enrichment relative
69 to the background. The enrichment tools can classify genes into GO categories or pathways
70 related to biological process, molecular function and cellular locations (Goffard and Weiller,
71 2007; Du et al., 2010). The GO-enrichment and the resultant hierarchy are very useful to
72 understand the complex biological processes that are being enriched. However, information on
73 specific biological attributes of a gene, such as the gene family (a group of homologous genes
74 with common evolutionary origin and biological functions) level information, are hard to glean
75 from GO-enrichment alone (Ashburner et al., 2000; Lee et al., 2005). For instance, enrichment of
76 a transcription factor will fetch GO terms for “regulation of transcription (GO:0006355)” or
77 “DNA binding (GO:0003700)” or “response to stress (GO:0006950)” but does not identify
78 which transcription factor family genes (e.g., WRKY, bZIP) being enriched. Having this
79 information, allows users to readily interpret large-scale datasets effectively and select favorite
80 gene families for further functional studies. While providing the information for functional
81 studies, gene families also could reveal the accurate gene annotation information that could not
82 be easily determined by BLAST-based tools alone. Further, comparative gene family size
83 analysis can certainly be informative and valuable approach to explore the biologically relevant
84 functions related to genome architecture and adaptation or speciation of various plant species
85 (Guo, 2013).

86

87 With the availability of complete genomes and sequence data, identification, and analysis of
88 specific gene families among plant species has become necessary. In this study, we present a
89 unique approach to perform classification and enrichment of genes to identify overrepresented
90 gene families (GenFam) in a user-defined query list. We suggest that GenFam is a valuable
91 addition to a plant biologists toolkit to analyze large-scale HTS datasets. By determining
92 overrepresented gene families in a user-defined gene list, rather than GO terms or hierarchy

93 alone, GenFam empowers users to readily interpret information of gene families (e.g. WRKY,
94 bZIP) in their queries, and move forward to selecting favorite overrepresented genes (or families)
95 for downstream studies and interpretation. GenFam is also freely accessible to users on the
96 world-wide web, as a user-friendly, graphical-user interface.

97

98 **MATERIALS AND METHODS**

99

100 **Background database**

101

102 GenFam currently supports the analysis of sixty plant genomes. GenFam classifies genes into
103 384 representative and unique gene families, which to the best of our knowledge the largest
104 collection, based on the well-annotated *Arabidopsis thaliana* (Berardini et al., 2015) and rice
105 (*Oryza sativa*) (Kawahara et al., 2013) genomes, literature search, and Pfam protein families
106 database (El-Gebali et al., 2019). We have identified and used Pfam common conserved domains
107 and domain organization among the homologous gene sequences to assign the gene families.
108 These highly conserved domains define protein functions and classifies protein-coding genes
109 into gene families. The conserved signature protein domains have the ability to detect the
110 divergent or distantly related homologs which would be prohibitive with sequence based
111 similarity analysis tools [e.g. BLAST (Altschul et al., 1997)]. Therefore, domain-based search
112 method would identify more genes belonging to gene families than BLAST-based homology
113 search.

114

115 To identify and classify gene families in plants, we have leveraged the publicly available
116 genomic resources at Phytozome (v12) database. The protein sequences of sixty plant genomes
117 were used to identify conserved protein domains to assign families to known and unclassified or
118 novel genes. The respective protein domains were predicted by HMMER (v3.1b2) using a
119 protein family hidden Markov model (HMM) profiles (Pfam release 32.0) (El-Gebali et al.,
120 2019). We have established rules to classify and assign the genes to gene families based on the
121 presence of signature conserved protein domains and have provided in **Supplementary Table**
122 **S1**. This approach allowed us to maximize classification including orphan genes with missing
123 annotations, genes with incorrect annotations, and novel genes present among the respective
124 genome databases. Lastly, the background databases were curated to remove redundancy and
125 duplication of gene members among families. In summary, we were able to integrate 384
126 representative gene families and corresponding (on an average ~41%) genes from sixty plant
127 genomes into our database (**Supplementary Table S2**). This is the most comprehensive and
128 largest collection of gene families spanning sixty plant species, when compared to other existing
129 databases. For instance, the recently published gene family database in poplar (GFDP) has
130 classified 6551 poplar genes into 145 gene families derived from Arabidopsis genome (Wang et
131 al., 2018). PlantTFDB (v4.0) and PlnTFDB (v3.0) has classified the genes into 58 and 84
132 transcription factor gene families (Perez-Rodriguez et al., 2010; Jin et al., 2017). Similarly,
133 another database and analysis toolkit, PlantGSEA, supports the gene family analysis for 13 plant
134 species which mostly imports gene families from well-annotated genomes such as rice (118 gene
135 families) and maize (81 gene families) (Yi et al., 2013).

136

137 All the gene family data was formatted using the PostgreSQL database to perform classification
138 and enrichment analysis using various statistical enrichment methods. The GenFam database

139 with complete protein domain annotation and gene family classification can be downloaded from
140 the GenFam website (<http://mandadilab.webfactional.com/home/dload/>). Detailed statistics for
141 the number of genes assigned to each gene family and the total number of background genes are
142 provided in **Supplementary Table S2**.

143

144 **Statistical enrichment methods**

145

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

159

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

170

171 GenFam can employ standard statistical tests such as the Fisher exact, Chi-Square (χ^2), Binomial
172 distribution and Hypergeometric tests for enrichment, along with multiple testing corrections to
173 control a false discovery. We recommend using Fisher exact, Chi-square (χ^2) and
174 Hypergeometric tests for smaller datasets (<1000) (McDonald, 2009), and Binomial distribution
175 for larger datasets (Khatri and Draghici, 2005; Zheng and Wang, 2008). Furthermore, the Chi-
176 Square (χ^2) test would be appropriate when the user defined gene list has less overlap with the
177 background dataset. As a default test, GenFam performs the Fisher exact test, which relies on the
178 proportion of observed data, instead of a value of a test statistic to estimate the probability of
179 genes of interest corresponding to a specific category.

180

181 To address the false positives resulting from multiple comparisons especially when the input
182 gene list is large (>1000), GenFam subsequently employs false discovery correction methods
183 including the Benjamini-Hochberg (Benjamini and Hochberg, 1995), Bonferroni (Bonferroni,

184 1936) and Bonferroni-Holm (Holm, 1979). The various statistical tests and false discovery
185 correction methods can be customized by the user as appropriate.

186

187 **Output summary**

188

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

198

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

207

208 **Web server implementation**

209

210 The GenFam web server is implemented using Python 3 (<https://www.python.org/>), Django
211 1.11.7 (<https://www.djangoproject.com/>) and PostgreSQL (<https://www.postgresql.org/>)
212 database. All the codes for data formatting and statistical analysis are implemented using Python
213 scripting language. Python is a fully-fledged programming language which offers well developed
214 packages for statistical analysis, graphics and integration with web apps. Therefore, we have
215 chosen Python over other languages such as R for development of GenFam. The high-level
216 Python web framework was constructed using Django. The Python web framework was hosted
217 using WebFaction (<https://www.webfaction.com/>). The web-based templates were designed
218 using Bootstrap, HTML, and CSS. The GenFam is compatible with all major browsers including
219 Internet Explorer, Microsoft Edge, Google Chrome, Mozilla and Safari. All the precomputed
220 plant gene family background databases were built using advanced PostgreSQL database. The
221 analyzed data was visualized using the matplotlib (Droettboom et al., 2016) Python plotting
222 library.

223

224 **RESULTS AND DISCUSSION**

225

226 **Case studies and analysis**

227

228 To demonstrate the utility of GenFam, we performed four case studies using transcriptome
229 datasets related to plants from different clades and families (cotton, tomato, soybean and rice)

230 (Bedre et al., 2015; Dametto et al., 2015; Zeng et al., 2017; Cui et al., 2018). We have previously
231 identified 662 differentially expressed genes in cotton (*Gossypium raimondii*, family Malvaceae)
232 infected with *Aspergillus flavus* (Bedre et al., 2015). For the first case study, we used GenFam to
233 determine the enriched gene families among these differentially expressed genes, using the
234 options of Fisher exact test for statistical enrichment, and the Benjamini-Hochberg (Benjamini
235 and Hochberg, 1995) method to control false discovery rate (FDR). Among the 662 genes, 514
236 genes were annotated and classified into gene families, resulting in ~78% intersection/coverage
237 with the GenFam database. The GenFam enrichment analysis revealed overrepresented gene
238 families such as expansins, kinases, reactive oxygen species (ROS) scavenging enzymes, defense
239 related genes, heat shock proteins and transcription factors—genes that we have hypothesized to
240 mediate cell-wall modifications, antioxidant activity and defense signaling in response to *A.*
241 *flavus* infection (Bedre et al., 2015). Additionally, GenFam also identified new enriched gene
242 families such as bHLH, GH3, glycosyltransferases and thaumatin that were not reported or
243 identified (**Figures 1 and 2; Supplementary Table S3**). In the second case study, we analyzed
244 758 genes which were up-regulated in a cold-tolerant rice (*Oryza sativa*, family Poaceae)
245 (Dametto et al., 2015). Among the 758 genes, 460 genes were annotated and classified into gene
246 families by GenFam, resulting in ~61% intersection/coverage with the GenFam database.
247 GenFam was able to successfully determine enriched gene families related to aquaporins,
248 glutathione S-transferases (GST), transporters, lipid metabolism, transcription factors as well as
249 gene families involved in cell wall-related mechanisms (**Supplementary Table S4**)—genes that
250 were hypothesized by Dametto *et al.* (2015) (Dametto et al., 2015) to play a role in the rice cold
251 stress response. Additionally, GenFam also identified new enriched gene families such as
252 aldehyde dehydrogenase (ADH), kinesins, glycosyltransferases, tubulin, phenylalanine ammonia
253 lyase (PAL) and thaumatin that were not reported or identified (**Supplementary Table S4**).
254 Next, we analyzed the differentially regulated genes from tomato (*Solanum lycopersicum*, family
255 Solanaceae) (Cui et al., 2018) and soybean (*Glycine max*, family Fabaceae) (Zeng et al., 2017)
256 using GenFam (**Supplementary Table S5 and S6**). We obtained ~65% and ~59%
257 intersection/coverage with the GenFam database for tomato and soybean respectively. The
258 GenFam results in both these studies revealed enrichment of several gene families that were
259 overrepresented and reported by Cui *et al.* (2018) (Cui et al., 2018) and Zeng *et al.* (2017) (Zeng
260 et al., 2017) (**Supplementary Table S5 and S6**). Additionally, GenFam also identified new
261 enriched gene families such as aquaporins, VQ, tify, GST, and PAL in tomato, and BET,
262 Dirigent, Expansins, Asparagine synthase (ASNS), and Carbonic anhydrase (CA) in soybean that
263 were not reported or identified (**Supplementary Table S5 and S6**). The detailed statistics of
264 enriched gene families for these case studies are provided in **Supplementary Table S3, S4, S5**
265 **and S6**.

266 **GenFam advantages and comparison with preexisting enrichment tools**

268 To the best of our knowledge, there is only one existing enrichment tool that comes close to the
269 GenFam approach, i.e., PlantGSEA (Yi et al., 2013), which also allows users to enrich gene lists
270 using gene family attributes. Hence, we performed a comparative analysis of GenFam and
271 PlantGSEA with a dataset from cotton (662 genes)(Bedre et al., 2015) and employing identical
272 parameters (Fisher exact test and Benjamini-Hochberg method) for enrichment. GenFam
273 enriched gene families belonging to cell-wall modifying genes, ROS scavenging genes,
274 transcription factors, lipid metabolism, and stress responsive gene families, both new and

275 previously shown to be biologically-relevant during *A. flavus* infection of cotton (Bedre et al.,
276 2015), while PlantGSEA missed several of these categories (**Supplementary Table S3 and S7**).
277 Upon further examination, we found that several gene family categories such as the ABC
278 transporters, expansins, and glutathione-S-transferase were absent in the PlantGSEA *G.*
279 *raimondii* background database. Moreover, PlantGSEA supports only thirteen plant genomes
280 with several redundant and overlapping genes and gene families, which could impact the
281 accuracy of the enrichment analysis. For instance, in the *A. thaliana* genome there are 37
282 annotated “C2-C2 Dof” transcription factors. PlantGSEA categorized 36 out of the 37 genes into
283 a “C2-C2 Dof” family, but also into an additional “Dof” family leading to redundant gene family
284 categories. GenFam avoids such discrepancies by curation and filtering redundant categories.

285
286 Taken together, we suggest that GenFam is a comprehensive and robust gene family
287 classification and enrichment program over prevailing tools, with several advantages: i) GenFam
288 is a dedicated and comprehensive platform for gene family-level classification, annotation and
289 enrichment analysis and supports sixty plant genomes including model and non-model plant
290 species. ii) GenFam background dataset was constructed from well-annotated gene families of *A.*
291 *thaliana* and rice genomes, literature search, and as well as a systematic HMM profile search for
292 signature conserved protein domain analysis using the Pfam database. This inclusive strategy
293 enabled us to categorize most of the genes into families, including those which may lack a
294 defined annotation in their corresponding genome database or could be novel genes. As a result,
295 GenFam database is by far the largest collection of gene families (384 families). In contrast,
296 existing databases such as PlantGSEA and GFDP only relies on annotations defined by other
297 databases such as TAIR and MSU annotations and/or other transcription factor databases (Yi et
298 al., 2013; Wang et al., 2018). The lack of additional analysis of protein domains perhaps explains
299 the poor representation of gene families in PlantGSEA and GFDP databases. iii) GenFam
300 background dataset was curated to remove redundancy and overlapping genes into different gene
301 families, that enhances the accuracy of the analysis. iv) In contrast to PlantGSEA, GenFam uses
302 the annotated gene families as reference background instead of the whole genome. This feature
303 ensures decreasing enrichment bias and increasing the accuracy of the analysis (Huang da et al.,
304 2009). v) GenFam accepts multiple input IDs including, gene IDs, transcript IDs and PAC IDs,
305 however PlantGSEA and GFDP are restricted to using only gene IDs. vi) GenFam can be solely
306 used for gene family annotation and classification regardless of enrichment analysis if a user is
307 only interested in annotating genes.

308

309 CONCLUSION

310

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

321 suggest that GenFam would be a valuable and powerful tool for plant biologists utilizing
322 genomics strategies to study plant biology and functional genetics.

323

324 AVAILABILITY AND REQUIREMENTS

325

326 **Project name:** GenFam

327 **Project home page:** <http://mandadilab.webfactional.com/home/>

328 **Operating system(s):** Platform independent

329 **Programming language:** Python 3, Django 1.11.7

330 **License:** CC BY-NC-ND 4.0

331 **Any restrictions to use by non-academics:** License needed

332

333 CONFLICT OF INTERESTS

334

335 The authors declare no competing financial interests.

336

337 AUTHOR CONTRIBUTIONS

338

339 RB conceived the project, developed the database/webserver, performed the case studies and
340 prepared the manuscript. KKM supervised the study, data analysis and interpretation. Both
341 authors have read, reviewed and approved the manuscript.

342

343 ACKNOWLEDGEMENTS

344

345 We thank Sonia Irigoyen (Texas A&M AgriLife Research) for review and comments during the
346 preparation of this manuscript. All experiments were conducted following the guidelines and
347 appropriate permissions of the Institutional Biosafety Committee of Texas A&M University.
348 This work was supported by funds from Texas A&M AgriLife Research Insect-vectorized Disease
349 Seed Grant to KKM.

350

351 REFERENCES

352 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., et al. (1997).

353 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

354 *Nucleic Acids Research* 25(17), 3389-3402. doi: Doi 10.1093/Nar/25.17.3389.

355 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al. (2000). Gene

356 ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1),

357 25-29. doi: 10.1038/75556.

358 Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., et al. (2007).

359 GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res* 35(Web Server issue),

360 W186-192. doi: 10.1093/nar/gkm323.

361 Bedre, R., Irigoyen, S., Schaker, P.D.C., Monteiro-Vitorello, C.B., Da Silva, J.A., and Mandadi,

362 K.K. (2019). Genome-wide alternative splicing landscapes modulated by biotrophic sugarcane

363 smut pathogen. *Sci Rep* 9(1), 8876. doi: 10.1038/s41598-019-45184-1.

- 364 Bedre, R., Mangu, V.R., Srivastava, S., Sanchez, L.E., and Baisakh, N. (2016). Transcriptome
365 analysis of smooth cordgrass (*Spartina alterniflora* Loisel), a monocot halophyte, reveals
366 candidate genes involved in its adaptation to salinity. *BMC Genomics* 17(1), 657. doi:
367 10.1186/s12864-016-3017-3.
- 368 Bedre, R., Rajasekaran, K., Mangu, V.R., Timm, L.E.S., Bhatnagar, D., and Baisakh, N. (2015).
369 Genome-wide transcriptome analysis of cotton (*Gossypium hirsutum* L.) identifies candidate
370 gene signatures in response to aflatoxin producing fungus *Aspergillus flavus*. *Plos One* 10(9),
371 e0138025. doi: ARTN e0138025
372 10.1371/journal.pone.0138025.
- 373 Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and
374 powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-*
375 *Methodological* 57(1), 289-300.
- 376 Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The
377 Arabidopsis information resource: making and mining the “gold standard” annotated reference
378 plant genome. *genesis* 53(8), 474-485.
- 379 Bonferroni, C.E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Libreria
380 internazionale Seeber.
- 381 Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., et al. (2013). Enrichr:
382 interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14,
383 128. doi: 10.1186/1471-2105-14-128.
- 384 Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005).
385 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics
386 research. *Bioinformatics* 21(18), 3674-3676. doi: 10.1093/bioinformatics/bti610.
- 387 Cui, J., Xu, P., Meng, J., Li, J., Jiang, N., and Luan, Y. (2018). Transcriptome signatures of
388 tomato leaf induced by *Phytophthora infestans* and functional identification of transcription
389 factor SpWRKY3. *Theor Appl Genet* 131(4), 787-800. doi: 10.1007/s00122-017-3035-9.
- 390 Dametto, A., Sperotto, R.A., Adamski, J.M., Blasi, E.A., Cargnelutti, D., de Oliveira, L.F., et al.
391 (2015). Cold tolerance in rice germinating seeds revealed by deep RNAseq analysis of
392 contrasting indica genotypes. *Plant Sci* 238, 1-12. doi: 10.1016/j.plantsci.2015.05.009.
- 393 Droettboom, M., Hunter, J., Caswell, T., Firing, E., Nielsen, J., Elson, P., et al. (2016).
394 "matplotlib: matplotlib v1. 5.1". doi).
- 395 Du, Z., Zhou, X., Ling, Y., Zhang, Z.H., and Su, Z. (2010). agriGO: a GO analysis toolkit for the
396 agricultural community. *Nucleic Acids Research* 38, W64-W70. doi: 10.1093/nar/gkq310.
- 397 El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., et al. (2019). The
398 Pfam protein families database in 2019. *Nucleic Acids Res* 47(D1), D427-D432. doi:
399 10.1093/nar/gky995.
- 400 Goffard, N., and Weiller, G. (2007). PathExpress: a web-based tool to identify relevant pathways
401 in gene expression data. *Nucleic Acids Research* 35, W176-W181. doi: 10.1093/nar/gkm261.
- 402 Guo, Y.L. (2013). Gene family evolution in green plants with emphasis on the origination and
403 evolution of Arabidopsis thaliana genes. *Plant J* 73(6), 941-951. doi: 10.1111/tpj.12089.
- 404 Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal*
405 *of statistics* 6(2), 65-70.
- 406 Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools:
407 paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1),
408 1-13. doi: 10.1093/nar/gkn923.

409 Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward
410 a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res*
411 45(D1), D1040-D1045. doi: 10.1093/nar/gkw982.
412 Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., et
413 al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation
414 sequence and optical map data. *Rice* 6. doi: Artn 4
415 10.1186/1939-8433-6-4.
416 Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools,
417 limitations, and open problems. *Bioinformatics* 21(18), 3587-3595. doi:
418 10.1093/bioinformatics/bti565.
419 Lee, J.S., Katari, G., and Sachidanandam, R. (2005). GObar: a gene ontology based analysis and
420 visualization tool for gene sets. *BMC Bioinformatics* 6, 189. doi: 10.1186/1471-2105-6-189.
421 Li, Y., Dai, C., Hu, C., Liu, Z., and Kang, C. (2017). Global identification of alternative splicing
422 via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *Plant J* 90(1),
423 164-176. doi: 10.1111/tpj.13462.
424 Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess
425 overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21(16),
426 3448-3449. doi: Doi 10.1093/Bioinformatics/Bti551.
427 Mandadi, K.K., and Scholthof, K.-B.G. (2015). Genome-wide analysis of alternative splicing
428 landscapes modulated during plant-virus interactions in *Brachypodium distachyon*. *Plant Cell* 27,
429 71-85. doi: 10.1105/tpc.114.133991.
430 Mandadi, K.K., and Scholthof, K.B. (2012). Characterization of a viral synergism in the monocot
431 *Brachypodium distachyon* reveals distinctly altered host molecular processes associated with
432 disease. *Plant Physiol* 160(3), 1432-1452. doi: 10.1104/pp.112.204362.
433 McDonald, J.H. (2009). *Handbook of biological statistics*. Sparky House Publishing Baltimore,
434 MD.
435 Perez-Rodriguez, P., Riano-Pachon, D.M., Correa, L.G.G., Rensing, S.A., Kersten, B., and
436 Mueller-Roeber, B. (2010). PlnTFDB: updated content and new features of the plant
437 transcription factor database. *Nucleic Acids Research* 38, D822-D827. doi: 10.1093/nar/gkp805.
438 Wang, H., Yan, H., Liu, H., Liu, R., Chen, J., and Xiang, Y. (2018). GFDP: the gene family
439 database in poplar. *Database (Oxford)* 2018. doi: 10.1093/database/bay107.
440 Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant
441 community. *Nucleic Acids Res* 41(Web Server issue), W98-103. doi: 10.1093/nar/gkt281.
442 Zeng, W., Sun, Z., Cai, Z., Chen, H., Lai, Z., Yang, S., et al. (2017). Comparative transcriptome
443 analysis of soybean response to bean pyralid larvae. *BMC Genomics* 18(1), 871. doi:
444 10.1186/s12864-017-4256-7.
445 Zheng, Q., and Wang, X.J. (2008). GOEAST: a web-based software toolkit for Gene Ontology
446 enrichment analysis. *Nucleic Acids Res* 36(Web Server issue), W358-363. doi:
447 10.1093/nar/gkn276.

448 449 **FIGURE LEGENDS**

450
451 **Figure 1.** GenFam workflow. The list of input gene IDs for respective plant species provided by
452 the user are analyzed for enrichment analysis using various statistical tests. The output of the
453 analysis can be viewed and/or downloaded as a table and/or graphical summary. The results page
454 has multiple options to visualize or download data for both enriched and non-enriched categories

455 (all gene families). The detailed output data from case studies are provided in **Supplementary**
456 **Tables S3, S4, S5 and S6.**

457

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

461

462 **SUPPLEMENTARY MATERIAL**

463

464 **Supplementary Table S1:** The classification of gene families and assignment of conserved
465 protein domain to each gene family

466 **Supplementary Table S2:** GenFam database statistics for total number of genes classified into
467 gene families and background number of genes in each plant species

468 **Supplementary Table S3:** List of the differentially regulated genes and analysis output of the
469 cotton case study

470 **Supplementary Table S4:** List of the differentially regulated genes and analysis output of the
471 rice case study

472 **Supplementary Table S5:** List of the differentially regulated genes and analysis output of the
473 tomato case study

474 **Supplementary Table S6:** List of the differentially regulated genes and analysis output of the
475 soybean case study

476 **Supplementary Table S7:** PlantGSEA result for gene family enrichment analysis using *G.*
477 *raimondii* dataset used in GenFam case study.

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

P-value correction

Output

GenFam Analysis Documentation Downloads

Success! Data Analysis Completed (Process ID: 173209701903)

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

Enriched Gene Families All Gene Families Get Figures

Download file with GO annotation View in browser

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

Table summary



