5-1-2020

Identification of Plant Gene Families Using Machine Learning of Sequence Similarity, Motif Conservation and Evolutionary Distances

James Stevens
Mississippi State University

Identification of Plant Gene Families Using Machine Learning of Sequence Similarity, Motif Conservation and Evolutionary Distances

By: James Stevens Jr.

A Thesis

Submitted to the Faculty of

Mississippi State University

ACKNOWLEDGEMENTS

Dr. George Popescu has hugely influenced my capacity to perform undergraduate research. He has helped me more than he knows, and for that I am extremely thankful. He has by far aided me most in my academic pursuits by helping me understand Bioinformatics and guiding my research. In addition, I would like to thank my professors who have taught me the foundations on which my understanding of these topics stands. I would also like to give a special thanks to my friends and family who have supported me in my academic career.

TABLE OF CONTENTS

ABBREVIATIONS

MAP3K	Mitogen-activated Protein Kinase Kinase Kinase
RAF	Rapidly Accelerated Fibrosarcoma
HMM	Hidden Markov Model
ILK	Integrin-linked Kinase
IRAK	Interleukin-receptor Associated Kinase
PB1	Phox and Bem1p
ACT	Aspartokinase, chorismate mutase, and TyrA
PAS	Per-Arnt-Sim

LIST OF TABLES

LIST OF FIGURES

Figure App. 1: A tree showing all identified CDKs. Model genes for each type of CDK are marked by a colored line outside of the tree. Novel genes are identified by the coloration of the tree lines.

Figure App. 2: A tree produced with all identified MAPK genes, with groups identified by the colored circle outside of the tree

Figure App. 3: A figure showing all genes identified within the Mitogen-activated protein kinase family. Colors specify individual families.

Figure App. 4: A tree showing genes identified by GeneFamilyRF as LRR-RLKs. Colored strips show the subgroups of previously published genes, with sub-subgroups labelled.

Figure App. 5: Collinearity is represented as lines between chromosomes. Chromosomes are represented by the black curves on the outside of the image with their respective labels. Red lines show collinear relationships involving genes identified as LRR-RLKs, while grey lines represent other collinear connections within the G. hirsutum genome.

ABSTRACT

Gene families are groups of genes of originating from a single ancestral gene, typically sharing similarities as well as conserved domains and structure. We have recently developed GeneFamilyRF, an integrative method that employs ortholog clustering, Hidden Markov Models, and motif identification through presently existing methods to measure factors that indicate familiar relationships among genes. In order to form classifications using these factors, RandomForestClassifier from the Scikit-learn Python package learn creates decision trees using sub-samples of the full dataset with averaging to improve accuracy and assist in prevention of overfitting. This method shows promise in accurately identifying gene family membership. Accurate gene family identification aids in rapid analysis of gene family evolution through the study of the results. To test the program, the WRKY gene family was selected, as it is well-conserved and not well studied in some species. The program identified 99.5% of previously identified genes, in addition to 23 novel genes in these species, 15 of which contained full WRKY DNA-binding domains. Additionally, the RAF gene family has significantly diversified in plants relative to animals, including many genes relating to stress-response and development. To further study family expansion, genes were identified by the seven species within the scope of GeneFamilyRF before being analyzed with phylogeny and motif analysis. This revealed novel motifs within the family, as well as information regarding evolution of particular groups within it.

CHAPTER I

GENEFAMILYRF FUNCTION AND OVERVIEW

INTRODUCTION

What we hoped to do in this study was to develop an improved method of gene family identification using a machine-learning method. The importance of such a task is that gene families are a significant step after genome annotation, as it is a useful grouping for identifying genes which are similar, sometimes in terms of function. Differential expansion within a gene family between species can reveal species-specific adaptations within the family, which can present options for crop improvement.

Random Forest Classifier uses a variety of decision trees, using different metrics (features), to find what the important features are to distinguish the class that is examined. It then takes the average of the trees to discover the metrics to classify the test data. In this study, the test data is all genes which are similar enough to the model genes to fit the similarity threshold which is determined prior to running Random Forest Classifier. The metrics used to determine classification are a variety of motif and sequence similarity data, which are determined by analysis of the amino acid sequences of each gene.

Developing genome functional annotations necessitate means of separating genes into similar groups, with one such annotation being the “gene family”, defined as the set of genes sharing a single common ancestor gene. Current methods to identify gene families include methods based on Hidden Markov Models[1], which have become more popular recently, Markov chain clustering (OrthoMCL), and BLAST[2]–based methods for search of gene similarity. However more bioinformatics tools can be employed to define gene families, including motif search, synteny

calculation tools and even comparative transcription regulation tools. Current tools require significant manual curation and time to filter results following the computational analysis. We have integrated into GeneFamilyRF methods based on hidden Markov models (HMMER package) and Markov chain clustering (OrthoMCL) to analyze sequence similarity and to generate candidates for gene families. Additionally, we implemented motif search and analysis methods from the MEME Suite package - FIMO[3] and MEME[3] – to refine the gene family analysis based on motif conservation. We use a Random Forest classifier to decide on gene family membership based on the evidence collected from all tools for sequence analysis.

HMM is a probabilistic modeling approach used to understand dynamics of discrete systems exhibiting random behavior. In bioinformatics, it is typically used in gene predictions and sequence similarity analysis. For sequence searches, it functions by first building an HMM model for a reference gene. This model can be applied to other genes to produce a score of how similar they are to the modelled genes. HMMER implements the Viterbi method to analyze sequence probabilities in input sequences. MEME uses gapless multiple sequence alignments to search for conserved residues in a set of user-input genes. FIMO scores motif matches to each residue to each position in the motif, and each match within a gene is treated as a singular unit. The RandomForestClassifier[4] is the machine-learning method that uses bootstrapping for supervised classification tasks. In GeneFamilyRF we use RandomForestClassifier to classify genes into families based on inputs calculated from HMMer, OrthoMCL, and MEME/FIMO. We should note that the significance of features used depends upon the family itself: large and diverse families, such as RAFs, will have low or no importance placed upon the E-value, while small and well conserved families will maintain a more balanced feature significance. In this study, we examined two practical applications of the novel GeneFamilyRF method: 1)

identification and classification of genes within the WRKY transcription factors family (including discoveries of WRKYs genes in *Gossypium hirsutum*), and 2) examinations of the evolution of the RAF family across multiple species.

METHODS

The GeneFamilyRF method functions by calling multiple previously designed methods to gather data on genes and their relationships, then feeding the outputs to RandomForestsClassifier, a machine learning function, which is trained on the data for the model genes. Necessary inputs for each family are a list of model genes, either an IUPAC formatted motif(s) or Pfam domain model(s), an optional motif obtained using MEME, and other files or information depending on options and features chosen. Currently, the species analyzed are *Arabidopsis thaliana*[5] (a model species), *Glycine max*[6] (Soybean), *Gossypium raimondii*[7] (a diploid cotton), *Gossypium hirsutum*[8] (tetraploid cotton crop), *Solanum lycopersicum*[9](tomato), *Zea mays*[10] (corn), *and Zostera marina*[11] (a common monocot seagrass). In the first step which is performed prior to actually running GeneFamilyRF, OrthoMCL[12] generates ortholog gene clusters for the seven species (including more species necessitate modifications to the code in addition to generating a new OrthoMCL cluster file which contains genes with cluster annotation.)

First, the configuration file is interpreted and stored. Next, the longest transcriptional variant for each model gene is selected before the amino acid sequence is aligned and used to create an HMM model using HMMbuild from HMMER. HMMsearch is then run on all genes in the database, then put into a ranked list based on score. An E-value threshold is then applied, which is currently at 1×10^{-9} but can be adjusted in the code. HMMbuild uses default settings and HMMsearch uses the specified E-value threshold, along with a simplified output. In order to be used with FIMO, the IUPAC motif is converted into FIMO format with probabilities of 1 for

each residue in the motif. FIMO then uses the genes from the list of genes after the E-value threshold is applied and the produced motif file to score the genes based on the most similar sequence to the motif, with user specified options. If MEME is used instead, the model genes are input using the number of motifs and settings specified in the configuration. The output motif file is then split so that motif-specific options can be used with the dictionary file specified in the configuration if there is some knowledge of settings to improve particular analyses of the individual motifs. When using domains in HMM format in the place of MEME or FIMO motif methods, the program uses HMMsearch in a similar method to before but with a much larger default threshold at 10. A large threshold allows for smaller domains to be detected in some cases, as small, highly variable domains tend to have relatively large E-values. Genes within the ranked HMMsearch output are then subjected to another threshold, which has a cutoff at the last model gene. This newly imposed threshold is then used, along with the clusters which had previously been determined by OrthoMCL, to determine what ratio of each cluster is contained in the list. This ratio, each gene's E-value, the average E-value of the gene's cluster, and the motif/domain q-value/E-value are used to calculate in integrative score for each gene, using the formula: $\log_2$(clust_rat) - $\log_2$(avg_E+1) - $\log_2$(trans_E+1) - qval_scale * $\log_2$(motif 1/ domain 1 score), where clust_rat is the cluster representation, qval_scale is the configuration-specified q-value scale, trans_E is the transcript's E-value, avg_E is the average E-value of the cluster. This score, along with all factors used in calculating the score and each other motif's q-value, are then normalized using numpy and used as features for RandomForestClassifier. RandomForestClassifier then produces a list of genes from all species that it predicts are member of the specified gene family.

A flowchart representing the workflow of GeneFamilyRF is shown in Figure 1.1. OrthoMCL clustering occurs prior to running GeneFamilyRF and is only run once, so the clusters are stored in the Gene IDs in the sequence database. The flowchart assumes that the clusters are already found and attached to the ID.

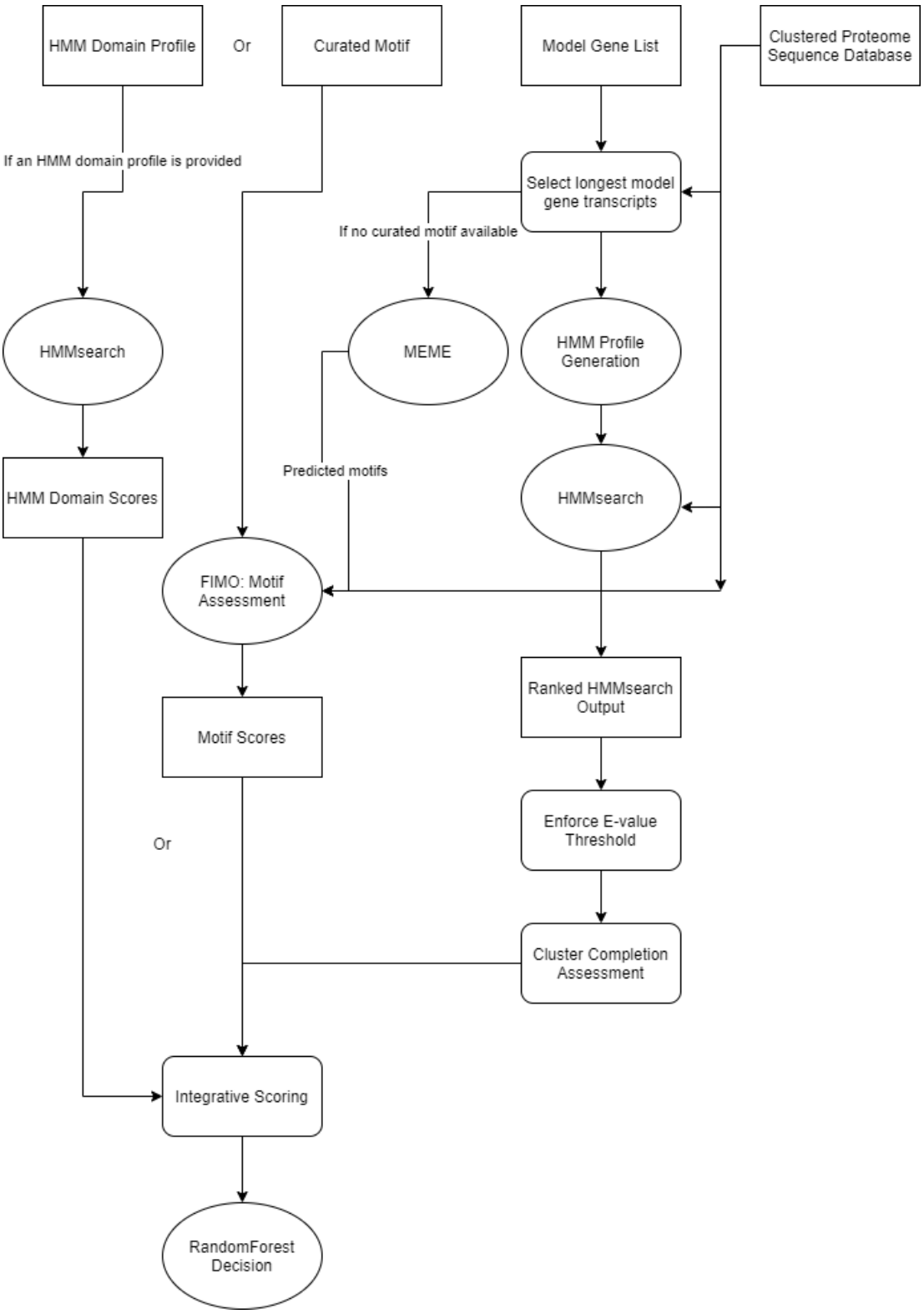

Figure 1.1: Flowchart representing the workflow of GeneFamilyRF

Additional Analyses

We have implemented additional analyses, which can be performed automatically by GeneFamilyRF, which includes alignments using MUSCLE[13] and Phylogeny using MEGA[14]. As such, MUSCLE is run using the default settings with the genes that are assigned to the family as the input. Phylogenetic trees can also be created with MEGA by using the alignment produced by MUSCLE. MEGA must be downloaded by the user and a settings file should be provided if the default settings are not desired. By default, a MEGA options file is provided with the filename gamma_allsites.mao. The settings provided by it are Maximum Likelihood method, no bootstrapping, JTT model, Gamma Distribute with Invariant Sites, 4 discrete gamma categories, Use All Sites, NNI Heuristic Method, and make initial tree automatically using NJ/BioNJ. We implemented MEGA such that it takes the alignment file as input then outputs a .nwk format tree file with the family name included. In order to visualize trees, we import the output .nwk format tree file into iTOL manually. To visualize groups, we use the color strip feature and beginning with the known genes in each family. Comparisons and trees created using GeneFamilyRF that are outside of the scope of this study are included in the Appendix section.

CHAPTER II

WRKY TRANSCRIPTION FACTOR FAMILY

INTRODUCTION

WRKY proteins are a group of Transcription Factors identifiable by the presence of the WRKYGQK and zinc-finger-like motifs within the amino acid sequence with significant roles in the response of plants to pathogens and other stressors, as well as roles in plant development[15]. Currently, there are 3 main groups of WRKY proteins. Group I WRKY proteins contain two WRKY domains, while group II and group III members contain only 1. A primary variation that separates group II and group III groups are the C2-H2 and C2-HC patterns within the zinc-finger-like motifs, respectively. WRKY transcription factors bind to the W-box in DNA, a stretch containing the domain (T)(T)TGAC(C/T). Significant variations in the number of WRKY genes are present between plant species, primarily a result of ploidy. Our examination of previously published literature revealed that WRKY genes have been characterized in several species of plants, including *Arabidopsis thaliana*[16], *Glycine max*[17], *Zea mays*[18], *Gossypium raimondii*[19], and *Solanum lycopersicum*[20]. Previous characterizations of WRKY genes were primarily performed using methods with foundations in BLAST and HMMsearch, along with significant manual curation. As one of the applications of the GeneFamilyRF method, we examined the WRKY gene family composition in seven plants.

METHODS

GeneFamilyRF was used to identify the WRKY family in all species, while the *Arabidopsis thaliana* genes were used as model genes. The first necessary change to the input involves modifying the Arab_kinome_gene_families.txt file to contain only a list of all Arabidopsis

WRKY genes with the family name, WRKY in this case, with a tabbed space after each of them. Next, a configuration file for GeneFamilyRF was produced using the portion following the family parameter to be replaced with "WRKY" and the same change performed on the trial name. The file containing the list of model genes for the family was edited to have the gene ID's from TAIR and the file added to the model_gene_families line of the configuration file. All other portions of the configuration were set to default. The final change was to the motif dictionary file, to which an entry for the WRKY gene family with the motif WRKYGQK was added. As an important family, many species have already had their WRKY family identified.

OrthoMCL is run independent of the GeneFamilyRF method itself. Instead, it is run on all genes for all examined species prior to the usage of the GeneFamilyRF method itself. FIMO is then used to find the most similar stretch to the motif of the searched family. To determine the presence and completeness of the WRKY domains present, an NCBI Conserved Domain Database search[21] was also performed on the newly predicted genes, as well as the ones that were not predicted at all by the methodology. Of those not predicted, excluding GRMZM2G045560 due to reasons described in Results, both contained a WRKY domain, with one (Solyc05g014040.1.1) showing a possible truncation on the N-terminus.

Additional analyses were performed on the previously unidentified genes, including a comparison of FIMO motifs against the WRKYGQK conserved domain. NCBI conserved domain database (CDD) search was also used by inputting the sequences of all newly identified genes with default settings to examine the newly identified and genes that were not identified by the methodology as a WRKY gene. Newly identified genes which did not contain the full DNA-binding motif were excluded as WRKYs. To examine phylogeny, we used the built-in

implementation of GeneFamilyRF to run MEGA, as described in Chapter I. ITOL[22] was used to improve visualization of the relationships between the 3 groups, as well as individual genes.

RESULTS

846 WRKY's in 7 species were identified, including 292 newly identified WRKY genes, with the numbers of genes in all species shown in Table 2.2. 5 of the analyzed species had already been previously characterized: *Arabidopsis thaliana, Glycine max, Gossypium raimondii, Solanum lycopersicum, and Zea mays*. *Gossypium hirsutum* had also been partially characterized by examining orthologs between it and *G. raimondii*, but with a differing gene ID type. *Zostera marina* had had no prior characterizations of the WRKY family, with 44 newly predicted genes. *G. hirsutum* has 226 identified WRKY genes compared to *G. raimondii's* 120, which is approximately a 2:1 ratio as would be expected as a result of being tetraploid to diploid respectively. The feature significances used by RandomForestClassifier is represented in Table 2.1, which correspond to the value each feature had in classifying the family. These significances are relatively uniform in this family.

Table 2.1: A table showing the significance of each feature as determined by RandomForestClassifier. Higher scores represent higher significance,

	Score	q-value	E-value	Avg Cluster E-value	Cluster Representation
RFClassifier Importance	0.132	0.131	0.132	0.114	0.121

Table 2.2: The number of WRKY's identified, as well as those previously published. "Ident." is the number of those previously published that were identified by GeneFamilyRF. "Publ." is the number published, and "New" is the number of novel genes identified.

WRKY				
Species	**Ident.**	**Publ.**	**New**	**Total**
A. thaliana	72	72	0	72
G. max	173	173	9	182
G. hirsutum	N/A	N/A	226	226
G. raimondii	111	111	9	120
Z. mays	118	119	5	123
S.lycopersicum	79	81	0	79
Z. marina	N/A	N/A	44	44
Total				846

Sequence Analysis

Motifs, as determined by FIMO, were also examined in both the newly identified and matched genes from literature. Of the 553 genes in both literature and the prediction, 526 contained the perfectly conserved WRKYGQK motif. Of the newly predicted genes, 15/23 contained the WRKYGQK conserved motif and 19/23 had the WRKYG(Q/K)K motif. Of all 847 genes identified, 777 of them (91.7%) contained the perfectly preserved WRKYGQK motif, while 42 (5%) contained WRKYGKK instead. Of those not predicted by the methodology, one of them (GRMZM2G045560) was not present in current annotations of the Z. mays genome. As such, it is likely an obsolete gene, so it was excluded from the additional analyses.

By only including the variants chosen by the methodology, 13/23 (59.1%) showed a single, complete WRKY domain, 7/23 (31.8%) had exclusively truncated WRKY domains, and 2/23 (9.1%) had multiple WRKY domains (truncated and complete.) Meanwhile, 1 of the 23

identified genes contained no WRKY domain according to the CDD search. All truncated novel genes were truncated on the C-terminal end, as can be see in Table 2.3.

Table 2.3: Novel WRKY genes as identified by GeneFamilyRF. Original Score shows the integrative score from GeneFamilyRF. In addition, the best match for the WRKYGQK motif is shown, followed by whether they contain the DNA-binding motif and the phylogenetic group from the tree.

GeneID	Original Score	Motif 1	Has HX[H/C]?	Phylo Group
AC193630.3_FGP003_Zea	320.883501	WRKYGQK	no	II
Glyma.03G048500.1.p_Glycine	-164.431378	WRYYPLK	no	II
Glyma.05G165800.4.p_Glycine	203.598877	WRKYGKR	yes	II
Glyma.07G161100.1.p_Glycine	320.883501	WRKYGQK	no	I
Glyma.08G078100.1.p_Glycine	320.713576	WRKYGQK	no	II
Glyma.09G127100.1.p_Glycine	-11.309309	WRKYGQK	no	I
Glyma.10G113800.1.p_Glycine	-75.92457	WRKYGKK	no	II
Glyma.10G171100.1.p_Glycine	158.188684	WHQYGLK	yes	II
Glyma.14G100100.1.p_Glycine	256.268239	WRKYGKK	yes	II
Glyma.17G239200.1.p_Glycine	-11.309309	WRKYGQK	no	II
Gorai.003G047800.1_Gossypium	256.268239	WRKYGKK	yes	II
Gorai.003G048100.1_Gossypium	320.883501	WRKYGQK	yes	II
Gorai.004G069500.1_Gossypium	320.883501	WRKYGQK	yes	II
Gorai.006G043200.1_Gossypium	320.883501	WRKYGQK	yes	II
Gorai.007G245200.1_Gossypium	255.531274	WRKYGKK	yes	II
Gorai.007G246500.2_Gossypium	319.883501	WRKYGQK	yes	II
Gorai.008G109600.1_Gossypium	320.342932	WRKYGQK	yes	II
Gorai.008G201000.1_Gossypium	158.088256	WRISEQK	yes	II
Gorai.009G421200.1_Gossypium	319.359939	WRKYGQK	yes	II
GRMZM2G092694_P01_Zea	320.883501	WRKYGQK	no	II
GRMZM2G103742_P01_Zea	320.468463	WRKYGQK	yes	II
GRMZM2G452444_P01_Zea	320.468463	WRKYGQK	yes	II
GRMZM5G849918_P02_Zea	320.468463	WRKYGQK	yes	II

Identified Group III WRKY's always contained a conserved WRKYGQKXIL and Group II tend to have WRKYGQKXXK (or less commonly, WRKYGQKVTR) in addition to the C2HXC and C2HXH motifs present in group.

Gorai.008G200800.2 contains 2 complete WRKY domains, as well as a third truncated one, but was categorized previously as a Group II WRKY. Its ortholog, Gohir.A12G184600.1 also contained the 3 WRKY domains; however, they seemed to have diverged from the base motif with more significant substitutions occurring in the WRKY closest to the C-terminal.

The zinc-finger portion of the protein is encoded by a C-C repeat followed by HXC in group III or HXH in group II WRKY's. The end of the HX(C/H) typically occur between 50 and 60 amino acids after the start of the WRKYGQK motif.

A phylogenetic tree was produced with all WRKY genes identified by GeneFamilyRF, as shown in Figure 2.1.

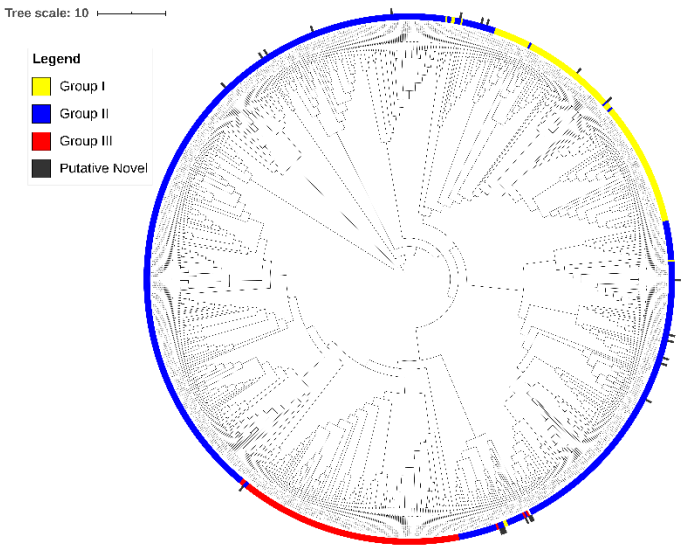

Figure 2.1: Phylogeny of WRKY family reveals mostly monophyletic groupings. Circle colors correspond to groups, while tick marks on the outside show positions of novel genes.

DISCUSSION

Genes not detected by GeneFamilyRF tended to have more poorly conserved motifs and a higher level of substitutions compared to the ones accepted. None of the missed genes contained a WRKYGQK conserved motif with a p-value < 0.0001 according to a FIMO search. Solyc03g082750's most similar stretch to WRKYGQK was WRKR, and Solyc05g014040 had no amino acid sequence resembling the WRKY domain. Solyc03g082750 only had WRKR as a match to the WRKYGQK motif but had the WRKY domain according to the NCBI CDD search. Solyc05g014040 also showed the same type of result in the CDD search but had no conserved stretch representative of a WRKY motif. Prior to the transcriptional variant selection improvement, many of the genes that were newly classified as WRKY's demonstrated some form of truncation when examined with NCBI's CDD search, which may explain their lack of inclusion in previous studies. All of them contained a WRKY domain in some form, helping to validate their inclusion as WRKY genes, as well as most of them containing a sequence with significant similarity to the WRKY motif, as described in Results.

Overall, 553/556 (99.5%) of the previously published WRKY genes were also identified by the method, along with 23 additional genes within the species that previously had their WRKY gene family studied. Among the outlier genes within the phylogenetic tree, some are shown to be very likely misclassified. AT3G01970 was historically placed within group I in previous studies but is shown to be more similar in sequence to group II. This is demonstrated by the presence of a singular group II type WRKY domain, as well as being aligned within group II in the phylogenetic tree. Another such gene is Gorai.008G200800 was previously characterized as group II but shares more similarities to group I in that it has more than 1 WRKY domain. The gene contains 3 WRKY domains featuring zinc-finger binding motifs and is more closely related

to group III based on its position within the tree, which is likely a result of an insertion of another WRKY gene into it. The G. Hirsutum ortholog, Gohir.A12G184600, still appears to contain the remnants of the 3 domains but has diverged significantly in amino acid sequence.

CHAPTER III

RAF FAMILY EVOLUTION

INTRODUCTION

The RAF gene family is a subfamily of MAP3Ks, and contains kinases associated with signal transduction, typically being phosphorylated by membrane-bound proteins then phosphorylating MAP2Ks[23,24]. This family in humans consists of three genes, while this family is significantly expanded within plants, with *Arabidopsis thaliana* containing 48. While RAF lineages within animals remain consistent with very few duplications and variations in function, those within plants feature significant deviations in function. Expansion of RAFs within plants is related to the abundance of duplication events within many plant species, with particularly recent Whole-genome duplication events occurring in *Gossypium hirsutum* and *Glycine max*[25,26]. The most common fate of a gene following duplication is its loss. Many domains have been identified within the RAF family, including EDR1, ACT, Ankyrin Repeat, PAS, and PB1 domains. The EDR1 domain is typically involved in disease resistance and senescence that is ethylene-induced, while also being involved in stress response signaling and programmed cell death regulation.[27] Aspartokinase, chorismate mutase, and TyrA (ACT) domain is a domain typically found within genes that respond to changes in amino acid concentration.[28] Ankyrin repeats and the PAS domain are both common domains, which can be found in a wide variety of genes.[29,30] These genes have diverse functionality. Errors in the ankyrin repeat domain have been demonstrated to induce deleterious phenotypes resulting from structural issues. As such, the folding structure of ankyrin repeats have been observed experimentally. The Phox and Bem1p (PB1) domain found in many signaling and scaffold proteins and is found within genes including MEKK3, a MAP3K, and p62, a scaffold protein, in humans. [31]

METHODS

Initially, gene family membership was determined by using GeneFamilyRF. The motif used as input was GTXX[WY]MAPE[32], while other settings were set to default outside of the option to automatically run Muscle to align all identified genes. Muscle was run by the program using default settings. In order to examine phylogeny and expansion of the RAF gene family, iqtree[33] was used, with automatic model selection and 1000 Ultrafast bootstraps and 1000 SH-like approximate likelihood ratio tests enabled.

Genes were classified primarily based upon their OrthoMCL clusters. Clusters only containing misclassified RLKs were excluded, with RLK classification determined by the presence of a domain that is identified as an IRAK by NCBI's CDD database. Singletons and very small clusters of 2-4 genes belonging to only one species were included in the ortholog group that was most closely related in the tree. Glyma.08G237100 and Glyma.02G215300 were included in the alignments of their cluster despite not containing functional protein kinase domains to better analyze their relationships with their assigned clusters, as they were very likely truncated duplicates of other identified genes. This is the result of these 2 genes featuring a mutation which creates a premature stop codon, such that most of or the entirety of the protein kinase domain is lost. Additionally, they retain, with strong conservation, the domains that are secondary in the closely related genes, indicating either a differentiation of function or that these genes are from very recent duplications.

A CDD search using NCBI's database revealed the domains present within the identified RAF groups, if any were present at all. The CDD search is run using entirely default parameters, with genes of all groups provided through FASTA format files. Following this, Multiple Em for Motif Elicitation (MEME) was run on each group using differential enrichment method, through the

command-line version of MEME Suite 5.05. For this, every group of RAFs outside of the one being examined during the individual run was used as the control in order to identify unique motifs within each family. The number of motifs for MEME to identify within each family was set to 3. Additionally, MEME was run using the same settings on a variety of merged sets of groups for ones without known domains. This grouping was based on monophyletic combinations of groups, such that all groups without known domains are contained within either supergroup. These combinatorial groups will be defined as supergroup A (groups 12, 15, and 25) and supergroup B (groups 6, 11, 12, and 21). Phylogeny of the family was examined using GeneFamilyRF's usage and default settings of MEGA, as specified in Chapter I. Annotation and styling was added by importing the tree into ITOL.

RESULTS

490 genes were identified by GeneFamilyRF, with 17 genes excluded due to their phylogenetic distance, as well as their presence of IRAK domains during CDD search, as can be observed in Figure 3.1 as the genes not assigned to a group. They were, however, included in the tree but not assigned to a RAF group. The number of genes present in each group in each species, along with the identified domain is shown in Tables 3.1.1 and 3.1.2.

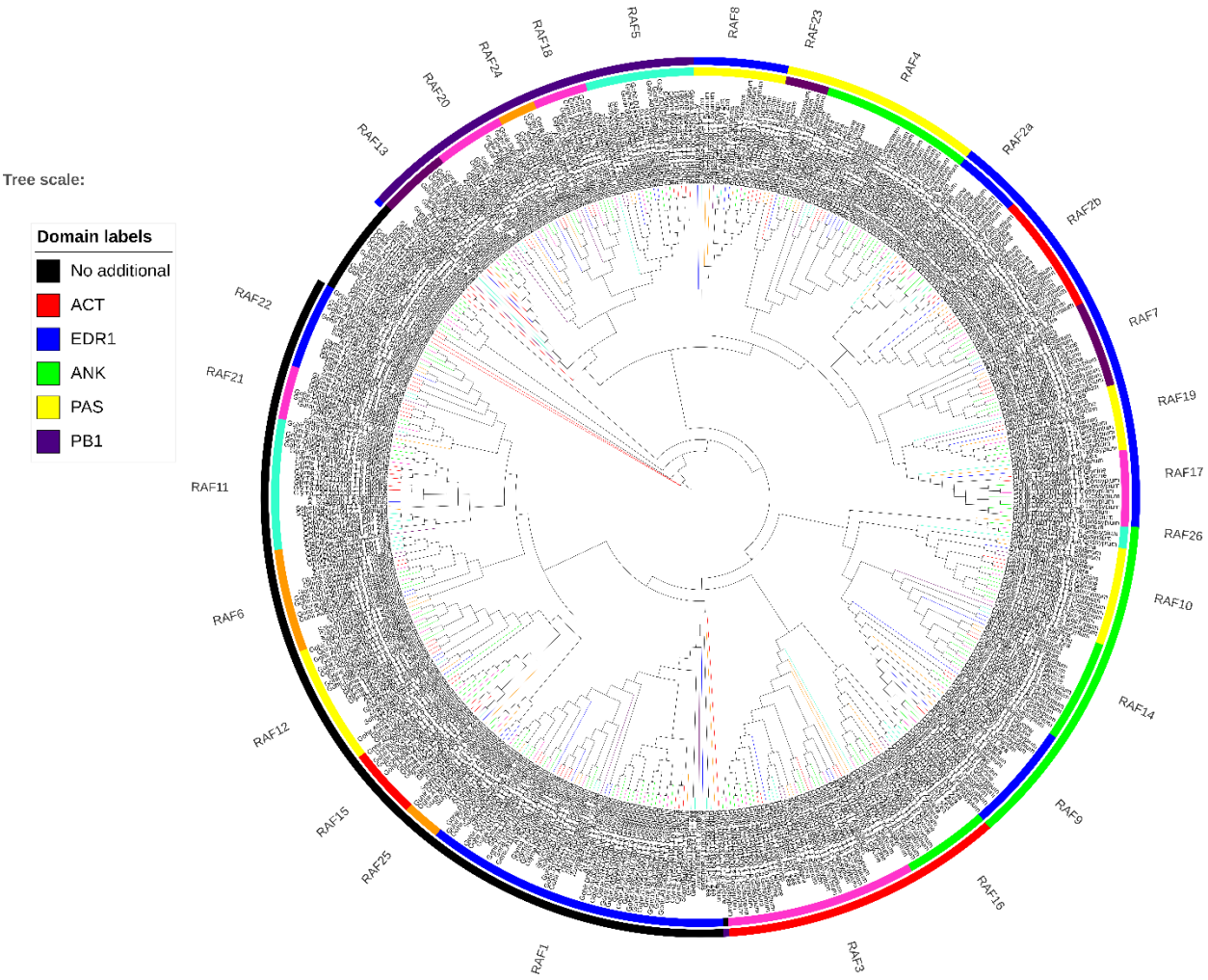

Figure 3.1: Phylogenetic analysis using iqtree reveals mostly monophyletic distribution of cluster-based groups. Outer circle shows domains present, while inner circle differentiates between groups. Colors of groups circles are not assigned to specific groups.

Table 3.1.1: A table showing the number of genes from each species within each group from 1 to 13, as well as the domain present if one was available. Totals for all groups is shown at bottom. Colors correspond to those on the group tree.

Group	At	Gh	Gr	Zmar	Zmaize	Gm	Sl	Total	Domain
1	5	18	10	2	7	9	5	56	
2a	1	3	2	0	1	4	1	12	EDR1
2b	1	7	4	0	1	6	3	22	EDR1
3	3	16	6	2	4	5	4	40	ACT
4	5	9	5	1	3	3	2	28	PAS
5	3	4	2	2	1	7	2	21	PB1
6	2	4	2	2	4	2	3	19	
7	2	4	2	1	1	4	2	16	EDR1
8	3	2	1	1	3	4	3	17	EDR1
9	2	6	3	1	1	4	3	20	ANK
10	1	6	3	2	1	3	2	18	ANK
11	2	4	2	1	5	8	2	24	
12	3	7	3	1	1	6	1	22	
13	1	4	2	1	2	1	1	12	PB1

Table 3.1.2: A table showing the number of genes from each species within each group from 1 to 13, as well as the domain present if one was available. Totals for all groups is shown at bottom. Colors correspond to those on the group tree.

Group	At	Gh	Gr	Zmar	Zmaize	Gm	Sl	Total	Domain
14	3	6	3	1	2	2	2	19	ANK
15	1	4	2	1	1	3	1	13	
16	0	4	2	2	5	2	1	16	ACT
17	1	6	3	0	1	2	1	14	EDR1
18	1	4	2	1	0	2	0	10	PB1
19	1	4	2	1	1	2	1	12	EDR1
20	1	3	1	2	2	2	1	12	PB1
21	1	3	2	2	2	2	1	13	
22	0	2	1	1	1	4	1	10	
23	1	4	2	1	2	4	2	16	PAS
24	1	2	1	0	0	2	1	7	PB1
25	3	2	1	0	0	0	1	7	
26	0	2	1	0	0	0	1	4	ANK
All	48	140	70	29	52	93	48	480	

The grouping method provided in methods resulted in 26 groups of RAF orthologs, with 24 of them being monophyletic within the tree. Two genes, Glyma.08G237100 and Glyma.02G215300, were estimated to be outside of their most similar groups. Glyma.02G215300 features significant similarity to Glyma.14G182700, indicating that it was recently duplicated from this gene. During tree estimation Glyma.02G215300 was placed within group 13, rather than group 2, while Glyma.08G237100 is within group 1 rather than 5.

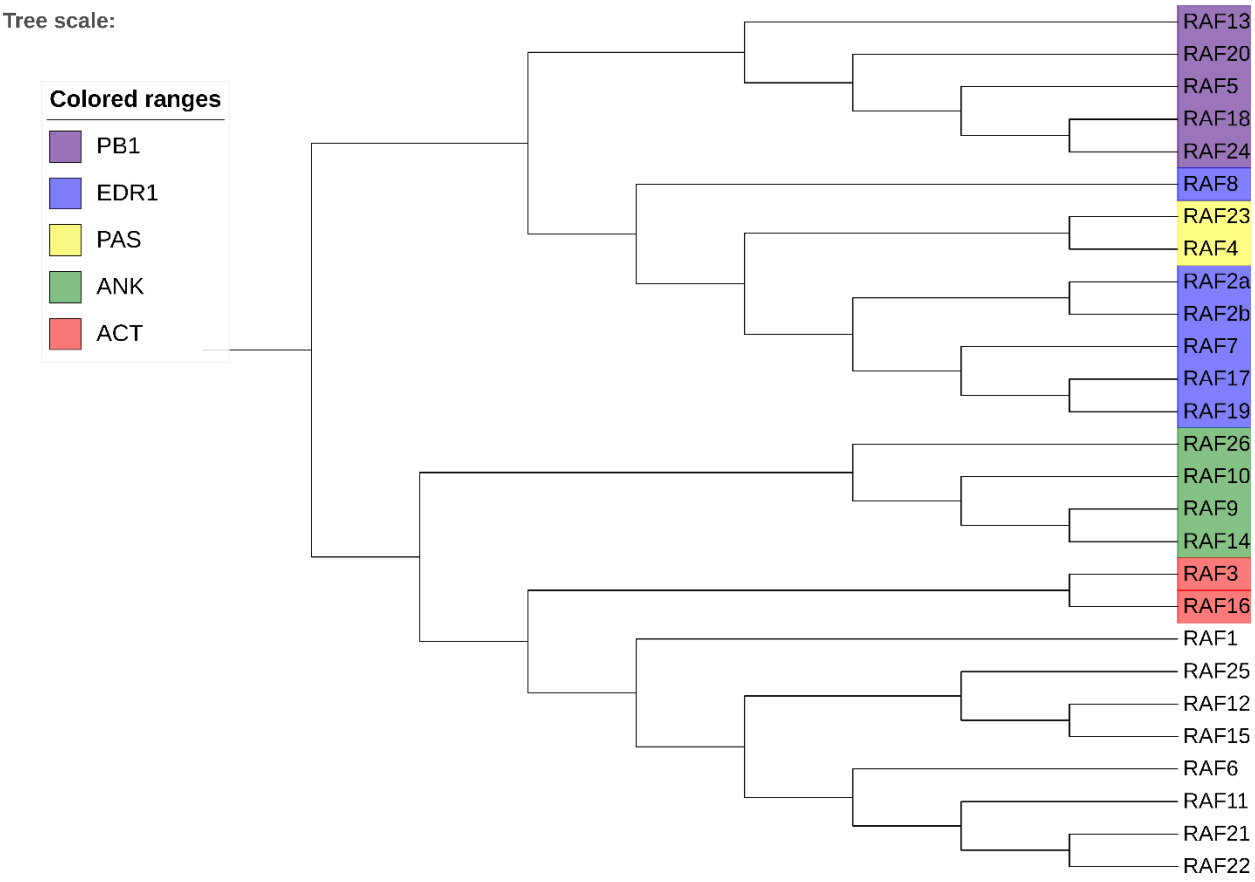

Figure 3.2: Simplified tree showing the groups relative to each other with coloration to indicate their additional domain when present.

Group 16 contained all species examined outside of *Arabidopsis thaliana*. Meanwhile, groups 4 and 23 contain a PAS domain, which is a type of domain commonly found in signaling proteins and functions as a signal sensor domain. The current PAS domain, according to PFAM is a combination of the PAS and PAC motifs.

Running MEME on supergroup A produced an output which revealed large motifs of widths of 25, 39, and 8. This set of groups contains the *A. thaliana* genes ATN1 and PEG7. While many sites within this identified motif were variable, there were some sites with observable conservation. One such site is the IGEG present in sites 20-23 within motif 1. While the IGXG residues are present within other Raf groups, the conserved E is unique to this set of groups.

Another conserved site is positions 10 and 11 in motif 1, which is a D followed by a P residue within all genes in these groups, but this site can also be found in many other of the identified groups so is not unique to this subset. The L residue which occurs immediately before the IDP in sites 9 to 11 is found only within this subset of groups and is only substituted in 4 of the 42 genes within this subset.

Groups 9, 10, 14, and 26 contain Ankyrin Repeat Domain, which is a common protein-protein interaction platform and occur in many proteins of a large variety of functions. Groups 9, 10, and 14 are considered Integrin-linked kinases, with group 9 corresponding to ILKs 4 and 5, group 14 corresponding to ILKs 1 to 3, and group 10 corresponding to ILK 6. All groups of known ILKs are put into a phylogenetic tree in Figure 3.3. Group 26 has not been previously categorized as ILKs and only contains genes from *S. lycopersicum, G. hirsutum, and G. raimondii*. The most closely related *A. thaliana* genes according to BLAST are ILKs 5 and 6, with E-values of 5e-82 and 3e-88, respectively. The query coverages for each were 82% and 79%, while the identity percentages were 40.33% and 45.89%, respectively. When compared against the G. raimondii gene within this group, the coverage is 90% and percent identity is 75.56%.

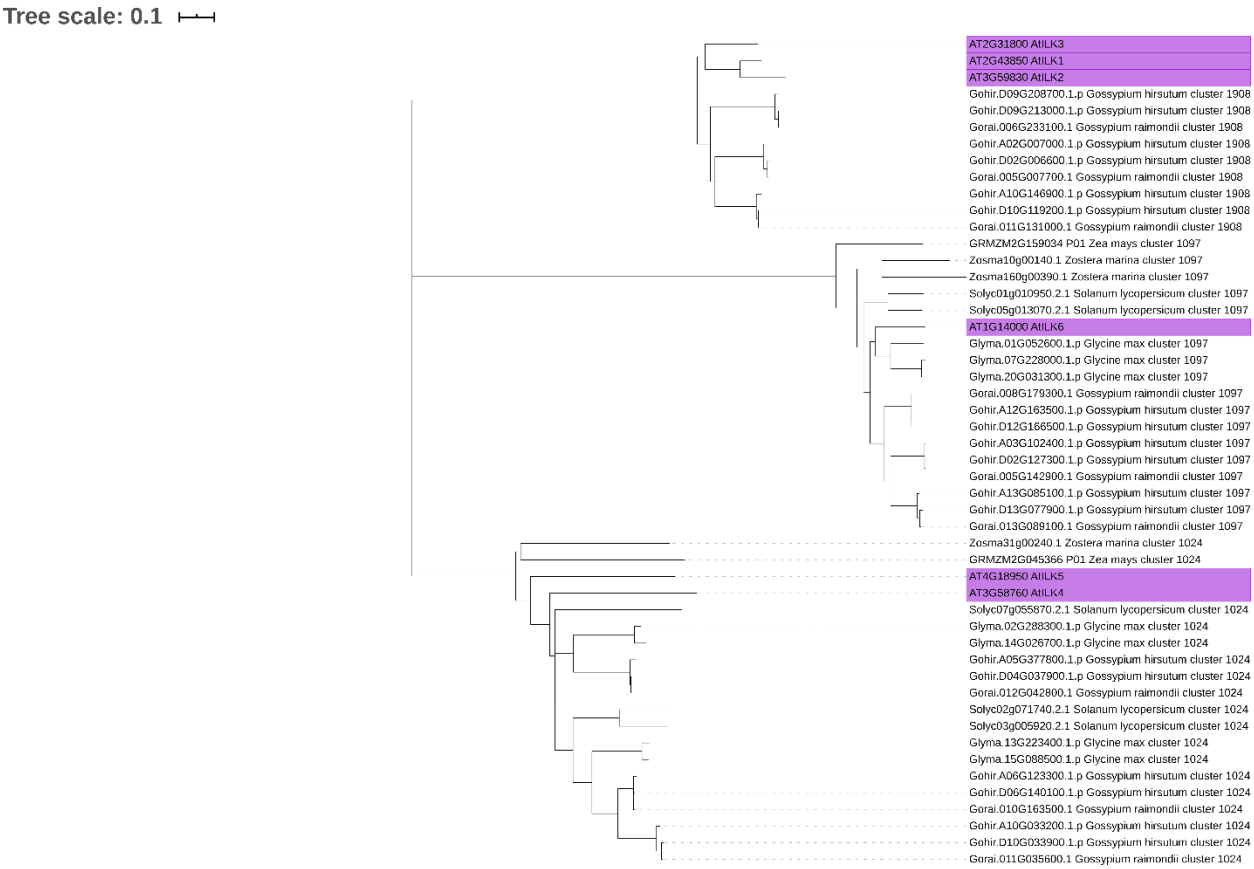

Figure 3.3: A figure showing the ANK-containing groups, except group 26. Known Arabidopsis ILKs are highlighted.

Across almost all RAFs is a conserved glycine 6 sites prior to the APE site, with the only exception being groups 9 and some genes of group 14, both of which are groups containing ILKs. In one group, the glycine residue is substituted with a serine or threonine. The DFG motif, which is C-terminal to the APE site is well-conserved (at most 1 gene featuring a unique substitution) within all groups with the exception of some genes within group 14, which feature a G substitution in the place of the D residue. These genes are the same group as the ones missing the conserved glycine 6 sites prior to the APE site that were contained within group 14.

DISCUSSION

The most expanded domain within RAFs in the examined species is EDR1, with 93 of the 473 (19.7%) examined RAFs containing this domain. Glyma.02G215300's phylogenetic inclusion with group 13 rather than 2 is likely due to the premature truncation, which resulted in errors with phylogeny. Likewise, Glyma.08G237100 is placed within group 1 despite being clustered with and sharing a domain with group 5. These genes maintained strong similarity with their progenitor genes but were truncated and likely became pseudogenes.

Groups sharing a domain clustered monophyletically in all cases except for groups 4 and 23, which have a PAS domain and are positioned among groups with an EDR1 domain. The presence of the PAS domain within RAF groups indicates that pathways have evolved with kinases that respond to amino acid concentration. This is likely due to involvement with amino acid synthesis or anabolism.

The expansion of many other domains within RAFs is further indication that RAFs have expanded significantly in terms of function in addition to number. One such example is the substitution of Aspartic Acid to Glycine. Such a substitution is likely to be involved in a modification of function, as it is perfectly conserved in most other groups. With its position near the binding motif and change from polar to nonpolar, it likely affects binding specificity.

CONCLUSIONS

Overall, the results of the GeneFamilyRF program have been promising regarding rapid gene family identification of novel genomes. As the number of genomes rapidly expands, we find it necessary to design methodology to automatically produce classification of their genes. Our method has some difficulty distinguishing between genes which are of full length and those which were duplicated then truncated by a substitution which results in a non-functional gene. A potential corrective measure would be by taking gene length into account, but this may lead to exclusions of genes without understanding the cause of the shortening. The implementation of shortcuts to phylogenetic analyses also benefits researches, as it allows phylogeny to be produced as soon as the GeneFamilyRF produces its results, while the options also allow disabling of this such that different methods can be used. Improvements and modifications can be tested to producebetter gene family classifications, including either testing other machine learning methods in the place of random forests or using nucleotide sequences in addition to or in place of amino acid sequences. These positive results show that machine-learning can be used in gene family classification, but direct comparisons to other current methods must be made to fully explore efficacy.

The WRKY gene family has significantly expanded in *Gossypium hirsutum*, with roughly double the number of genes when compared to one of its progenitor species, *Gossypium raimondii*. Most of these newly identified genes still maintain strong similarity between duplicate pairs, some with few to no differences. Among the 23 novel genes, 15 contained both the main WRKYGQK, with a few substitutions in some genes, and the C2H2 or C2HC DNA-binding motifs. The remaining 8 were excluded as likely WRKY genes. Of the 8 excluded, 6 belonged to *Glycine max*, while the others were from *Zea mays*. Additionally, 226 genes in *Gossypium hirsutum* and

44 in *Zostera marina* were identified, and phylogenetic analysis showed strong grouping among genes which had their groups previously identified. Novel genes must be further analyzed and can potentially contain insight into stress response or disease resistance.

The RAF gene family is one of the more complex and difficult to study families in plants due to its significant expansion and diversification. The high levels of variability across the family requires lax rules for inclusion, which also results in the inclusion of genes that are unlikely to be legitimate members of the family. The genes included as a result of this belonged to the RLK gene family, which shares many features with RAFs. Following the exclusion, our examinations revealed that many domains were present in the RAF family, including EDR1, PB1, ACT, ANK, and PAS. These featured varying levels of expansion, with the most expanded being EDR1, which is known to be involved with Ethylene-related defense and stress response. In addition, we identified many group-specific motifs which are likely to have functional importance, including a large conserved region present in Supergroup A.

PERSONAL CONTRIBUTIONS

I contributed to the creation and expansion of the GeneFamilyRF program, expanding the motif identification portion by changing the code to allow the integration of multiple motifs, in addition to implementing code that allows the program to process larger gene families. I added code which allows the program to accept HMM domain profiles from PFam or other databases in lieu of an IUPAC motif, as well as code that allows the usage of MEME to identify novel motifs if none is provided. Additionally, I corrected the code for the part of the method that was chosing the transcript variant to analyze (incorrectly in some cases), which resulted in some misclassifications of genes. I performed the classification of WRKY and RAF gene families for the seven plant species analyzed here.

I researched the literature for published gene family classifications and performed the comparisons with my results obtained using GeneFamilyRF, then created tables using these comparisons. In addition, I produced and annotated all phylogenetic trees and figures to visualize the analyzed gene families by using the software tools specified in the thesis.

REFERENCES

[1] Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

[3] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(suppl_2), W202–W208. https://doi.org/10.1093/nar/gkp335

[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825−2830. Retrieved from http://scikit-learn.sourceforge.net.

[5] Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*, *89*(4), 789–804. https://doi.org/10.1111/tpj.13415

[6] Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178–183. https://doi.org/10.1038/nature08670

[7] Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G.,… Yu, S. (2012). The draft genome of a diploid cotton Gossypium raimondii. *Nature Genetics*, *44*(10), 1098–1103. https://doi.org/10.1038/ng.2371

[8] Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., … Chen, Z. J. (2015). Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. *Nature Biotechnology*, *33*(5), 531–537. https://doi.org/10.1038/nbt.3207

[9] Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., … Gianese, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–641. https://doi.org/10.1038/nature11119

[10] Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Wilson, R. K. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, *326*(5956), 1112–1115. https://doi.org/10.1126/science.1178534

[11] Olsen, J. L., Rouzé, P., Verhelst, B., Lin, Y. C., Bayer, T., Collen, J., … Van De Peer, Y. (2016). The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature*, *530*(7590), 331–335. https://doi.org/10.1038/nature16548

[12] Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

[13] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

[14] Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

[15] Agarwal, P., Reddy, M. P., & Chikara, J. (2011). WRKY: its structure, evolutionary relationship, DNA-binding selectivity, role in stress tolerance and development of plants. *Molecular Biology Reports*, *38*(6), 3883–3896. https://doi.org/10.1007/s11033-010-0504-5

[16] Eulgem, T., Rushton, P. J., Robatzek, S., & Somssich, I. E. (2000, May 1). The WRKY superfamily of plant transcription factors. *Trends in Plant Science*, Vol. 5, pp. 199–206. https://doi.org/10.1016/S1360-1385(00)01600-9

[17] Yang, Y., Zhou, Y., Chi, Y., Fan, B., & Chen, Z. (2017). Characterization of Soybean WRKY Gene Family and Identification of Soybean WRKY Genes that Promote Resistance to Soybean Cyst Nematode. *Scientific Reports*, *7*(1), 17804. https://doi.org/10.1038/s41598-017-18235-8

[18] Wei, K.-F., Chen, J., Chen, Y.-F., Wu, L.-J., & Xie, D.-X. (2012). Molecular phylogenetic and expression analysis of the complete WRKY transcription factor family in maize. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*, *19*(2), 153–164. https://doi.org/10.1093/dnares/dsr048

[19] Ding, M., Chen, J., Jiang, Y., Lin, L., Cao, Y., Wang, M., … Ye, W. (2015). Genome-wide investigation and transcriptome analysis of the WRKY gene family in Gossypium. *Molecular Genetics and Genomics*, *290*(1), 151–171. https://doi.org/10.1007/s00438-014-0904-7

[20] Huang, S., Gao, Y., Liu, J., Peng, X., Niu, X., Fei, Z., … Liu, Y. (2012). Genome-wide analysis of WRKY transcription factors in Solanum lycopersicum. *Molecular Genetics and Genomics*, *287*(6), 495–513. https://doi.org/10.1007/s00438-012-0696-6

[21] Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., … Bryant, S. H. (2016). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, *45*(D1), D200–D203. https://doi.org/10.1093/nar/gkw1129

[22] Interactive Tree of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. (n.d.). *Nucleic Acids Res.*, *44*, W242–W245.

[23] Jonak, C., Ökrész, L., Bögre, L., & Hirt, H. (2002, October 1). Complexity, cross talk and integration of plant MAP kinase signalling. *Current Opinion in Plant Biology*, Vol. 5, pp. 415–424. https://doi.org/10.1016/S1369-5266(02)00285-6

[24] Hanks, S. K., & Hunter, T. (1995). The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification 1 . *The FASEB Journal*, *9*(8), 576–596. https://doi.org/10.1096/fasebj.9.8.7768349

[25] Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., … Schmutz, J. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*, *492*(7429), 423–427. https://doi.org/10.1038/nature11798

[26] N Panchy, M. L.-S. S.-H. S. (2016). Evolution of gene duplication in Plants1[OPEN]. *Plant Physiol*, *171*, 2294–2316.

[27] Zhao, C., Nie, H., Shen, Q., Zhang, S., Lukowitz, W., & Tang, D. (2014). EDR1 Physically Interacts with MKK4/MKK5 and Negatively Regulates a MAP Kinase Cascade to Modulate Plant Innate Immunity. *PLoS Genetics*, *10*(5), e1004389. https://doi.org/10.1371/journal.pgen.1004389

[28] Chipman, D. M., & Shaanan, B. (2001, December 1). The ACT domain family. *Current Opinion in Structural Biology*, Vol. 11, pp. 694–700. https://doi.org/10.1016/S0959-440X(01)00272-X

[29] Li, J., Mahajan, A., & Tsai, M. D. (2006, December 26). Ankyrin repeat: A unique motif mediating protein-protein interactions. *Biochemistry*, Vol. 45, pp. 15168–15178. https://doi.org/10.1021/bi062188q

[30] Henry, J. T., & Crosson, S. (2011). Ligand-binding PAS domains in a genomic, cellular, and structural context. *Annual Review of Microbiology*, *65*, 261–286. https://doi.org/10.1146/annurev-micro-121809-151631

[31] Lamark, T., Perander, M., Outzen, H., Kristiansen, K., Øvervatn, A., Michaelsen, E., … Johansen, T. (2003). Interaction Codes within the Family of Mammalian Phox and Bem1p Domain-containing Proteins. *Journal of Biological Chemistry*, *278*(36), 34568–34581. https://doi.org/10.1074/jbc.M303221200

[32] Bokros, N., Popescu, S. C., & Popescu, G. V. (2019). Multispecies genome-wide analysis defines the MAP3K gene family in Gossypium hirsutum and reveals conserved family expansions. *BMC Bioinformatics*, *20*(S2), 99. https://doi.org/10.1186/s12859-019-2624-9

[33] IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. (n.d.). *Mol. Biol. Evol.*, *32*, 268–274.

APPENDIX

The Arabidopsis kinome is currently the only to be fully characterized, featuring over 1000 genes. The families considered in the kinome analysis are Receptor-like Kinases (RLK), Cyclin-Dependent Kinases (CDKs), all families in the Mitogen-activated protein kinase cascade (MAPK, MAP2K, MAP3K, MAP4K), SnRKs, AGCs, NEKs, AURORAs, and SHAGGY-LIKE families. The most challenging kinase families in our experiments were RLKs and CDKs. We had to implement additional enhancements that have enabled the analysis of the Leucine-Rich Repeat Receptor-Like Kinase (LRR-RLK) and Cyclin-dependent Kinase (CDK) families. LRR-RLKs feature 2 motifs, the Leucine-Rich Repeat and the kinase domain, requiring multiple motif analysis, while CDKs require MEME with differential enrichment to find a motif not shared with other kinases.

1. Analysis of the Cyclin Family in Arabidopsis and Comparative Analysis of Cyclin-Like Proteins in seven plant species

Cyclin-Dependent Kinases (CDKs) are proteins that interact with Cyclins to regulate transcription and processes related to the cell cycle. CDKs form complexes with Cyclins when they are phosphorylated by other Kinases.

Methods/Results

Using differential enrichment option with MEME, the model MAPK Arabidopsis genes as a control group, and model CDK genes from TAIR as the inputs to scan for motifs within resulted in an output of CDK genes from FamSync which included all model genes. The reasoning for using differential enrichment was to potentially identify the amino acid sequence involved in the binding of the Cyclin to search for with FIMO.

The number of genes varied significantly between species, with the most identified in Glycine max at 23 CDKs, even more than in Gossypium hirsutum despite being diploid and nearly double that of Arabidopsis thaliana. The Glycine max genes within each family of CDKs were more similar than those of other species, with an exception of a single gene (Glyma.07G021100) in CDKB indicating that CDKs likely underwent multiple duplications within Glycine max. The lowest numbers of CDKs per species were in Gossypium raimondii, Zea mays, and Zostera marina at 9 identified genes each. All identified CDKs are arranged in App. Table 1 and assigned to CDK type.

App. Table 1: The number of CDKs identified by type and species.

Cyclin-Dependent Kinases								
	CDK Type	A	B	C	D	E	F	Species Total
Species								
A. thaliana		1	4	2	3	1	1	12
G. max		4	5	3	4	3	4	23
G. raimondii		1	1	1	2	2	2	9
G. hirsutum		3	4	2	3	4	3	19
S. lycopersicum		3	2	2	1	1	1	10
Z. mays		2	2	1	2	1	1	9
Z. marina		2	2	2	1	1	1	9
Type Total		16	20	13	16	13	13	91

This output was used to generate a tree with muscle and MEGA through the command-line version of MEGA. The method used to generate the tree was maximum likelihood. The tree was then uploaded to iTOL and all Arabidopsis genes colored based on their CDK group. The tree including branch lengths is shown in Figure App. 2. This tree demonstrates strong grouping within each family and simple grouping that allows the identification of the specific families for each of the identified genes.

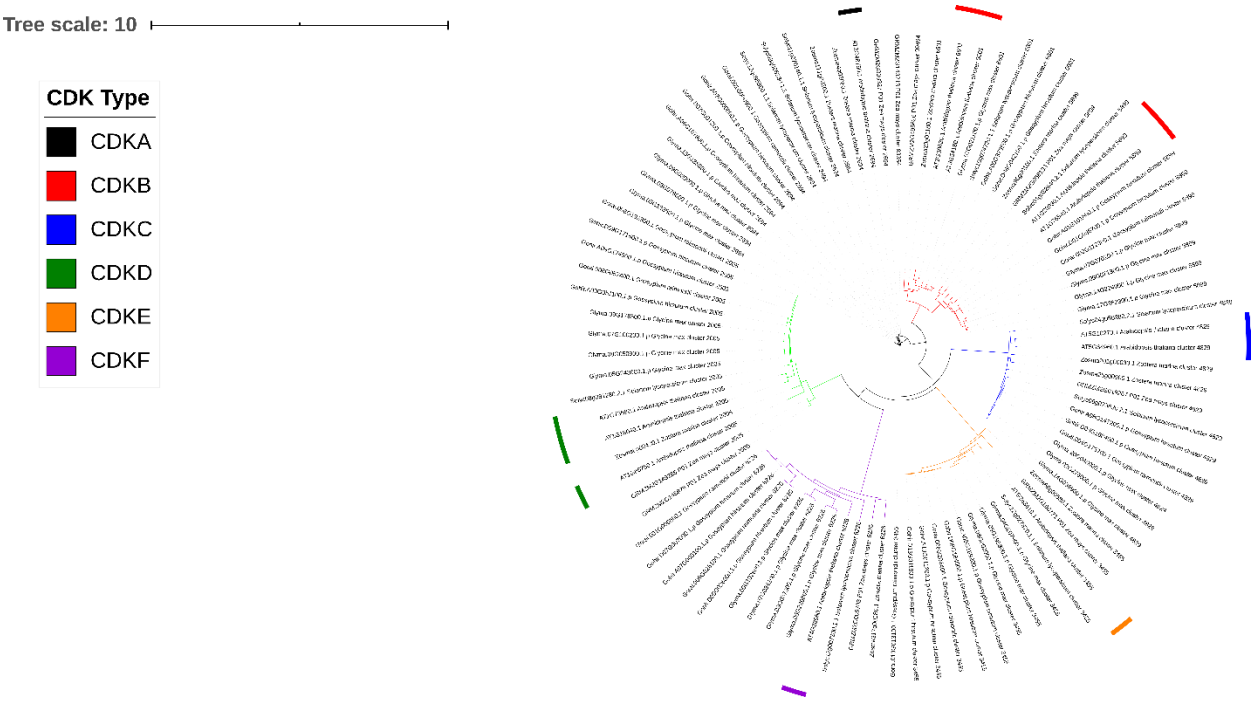

Figure App. 1: A tree showing all identified CDKs. Model genes for each type of CDK are marked by a colored line outside of the tree. Novel genes are identified by the coloration of the tree lines.

2. Analysis of the MAPK Family in seven plant species

Protein modifications play a significant role in the regulation of cellular processes. One of the most common post-translation modifications is phosphorylation. Phosphorylation is typically started by receptors then carried and amplified by Kinase proteins to transmit information from external stimuli, such as pathogens and temperature changes. The primary cytosolic kinases families involved in the MAP (mitogen-activated protein) kinase pathway are MAPK, MAP2K, MAP3K, and in some cases MAP4K, with the chain of phosphorylation occurring sequentially in the order of MAP4K -> MAP3K -> MAP2K -> MAPK. Mitogens are substances (typically proteins) that trigger cell division.

METHODS/RESULTS

MAPK genes were identified using GeneFamilyRF, an integrative method using motif identification, HMMsearch, and ortholog clustering to predict gene families.

MAPK Comparisons to https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4363184/.

A paper on MAPK gene identification had previously been published, identifying MAPK's in 40 species, and the comparisons can be seen in App. Table 2. Of the ones GeneFamilyRF currently identifies, it contained 5 of the 7 species identified by GeneFamilyRF. 2 species that were identified by GeneFamilyRF that the paper had not identified were Gossypium hirsutum and Zostera marina. Within their proteomes, 54 and 17 genes were identified. The gene ID system used was from a previous version of Phytozome's Glycine max genome. These could be converted easily into the current format by searching the genes in Phytozome's Phytomine and using the updated ID's returned.

App. Table 2: A comparison of literature MAPKs to ones identified by GeneFamilyRF. Ident. is the number identified in the previous study examined, Publ. is the number previously published, and New is the number of putative novel MAPK genes.

	MAPK			
	Ident	Publ.	New	Total
A. thaliana	20	20	0	20
G. raimondii	28	28	0	28
G. hirsutum	N/A	N/A	55	55
Z. mays	19	19	1	20
G. max	31	31	1	32
S. lycopersicum	17	17	0	17
Z. marina	N/A	N/A	14	14
** G. max genes in the paper used a different type of Gene ID and count of both still given				

All 115 of the previously identified genes were identified by GeneFamilyRF, along with 2 additional genes that had not been published, GRMZM2G063144 and Glyma.07G255400. Motif analysis of all identified genes was performed. Of the genes identified, 166 of the 186 genes identified featured a motif in the form of T[ED]YVxTRWYRAPE. 17 of the 20 identified genes with variations contained only 1 amino acid substitution within the motif. The most common variation in the motif, occurring in 11 of the 19 genes featuring changes in the motif, was a substitution of N-terminal APE, resulting in SPE or PPE. Both of the identified genes that had not been previously published both featured a MAPK motif of TDYVATRWYRAPE, a match of the recognized motif T[ED]YVxTRWYRAPE.

Both newly identified genes were queried in NCBI's Conserved Domain Database

In order to verify that none of the genes were misclassified membrane proteins, TMHMM was used to scan for membrane-bound stretches. Only 5 genes were identified as having at least 1 likely transmembrane amino acid. The one with the greatest number was Glyma.17G018800, which had 14, still fewer than the number for a complete transmembrane sequence. This gene featured a complete MAPK motif but is likely to be membrane-bound. The number of amino acids in predicted transmembrane helices of the rest of the genes was at most 2.6.

A tree of MAPK genes identified by GeneFamilyRF was made. Using this tree, MAPK genes featuring the TDY activation loop were marked with a black strip. A single member of each group identified in the comparison paper was colored with a strip, then the main branch containing each was also marked with the same color. The genes showing unique variation in the activation loop were labelled with stars. All genes with these variations are within groups A or B, with 3 of the 4 in group B. We performed similar calculation of gene families for MAP2Ks,

MAP3Ks (partially reported here in chapter 3) and MAP4Ks. The results can be seen in App. Figure

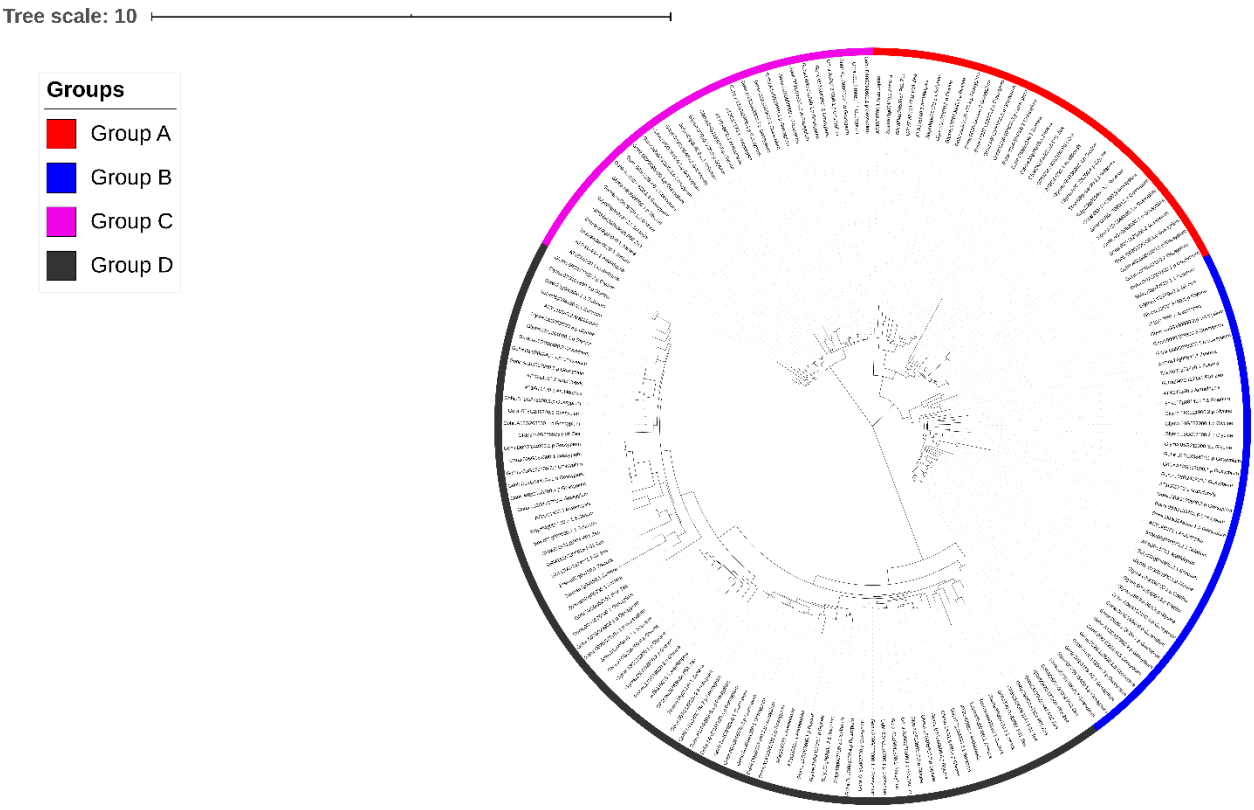

Figure App. 2: A tree produced with all identified MAPK genes, with groups identified by the colored circle outside of the tree

A phylogenetic tree including the entire MAPK signaling cascade (MAPK, MAP2Ks, MAP3Ks, and MAP4Ks) genes identified by GeneFamilyRF is shown in the Figure App. 3 below. The number of identified genes and comparisons to previous literature can be found in App. Table 3.

App. Table 3: A table showing the number of identified genes in each Mitogen-activated Protein Kinase cascade family. MAP4Ks had no genome-wide studies outside of A. thaliana, which had all members identified. The missing MAP3K was one that had previously been predicted to be misclassified.

	MAPK						MAP2K						MAP3K						MAP4K
	Ident	Publ.	New	Total	SN	SP	Ident.	Publ.	New	Total	SN	SP	Ident.	Publ.	New	Total	SN	SP	Ident.
A. thaliana	20	20	0	20	1	1	10	10	0	10	1	1	92	93	0	92	0.9892	0.9892	10
G. raimondii	28	28	0	28	1	1	11	11	0	11	1	1	107	110	3	110	0.9727	0.9727	15
G. hirsutum	N/A	N/A	54	54	N/A	N/A	N/A	N/A	20	20	N/A	N/A	208	210	9	217	0.9905	0.9585	28
Z. mays	18	19	0	18	0.9474	1	9	6	3	9	1.5	1	78	81	0	78	0.963	1	9
G. max	31	31	1	32	1	0.9688	10	10	2	12	1	0.8333	172	172	5	172	1	1	15
S. lycopersicum	17	17	0	17	1	1	5	5	0	5	1	1	87	93	2	89	0.9355	0.9775	8
Z. marina	N/A	N/A	14	14	N/A	N/A	N/A	N/A	0	5	N/A	N/A	51	51	1	64	1	0.7969	8

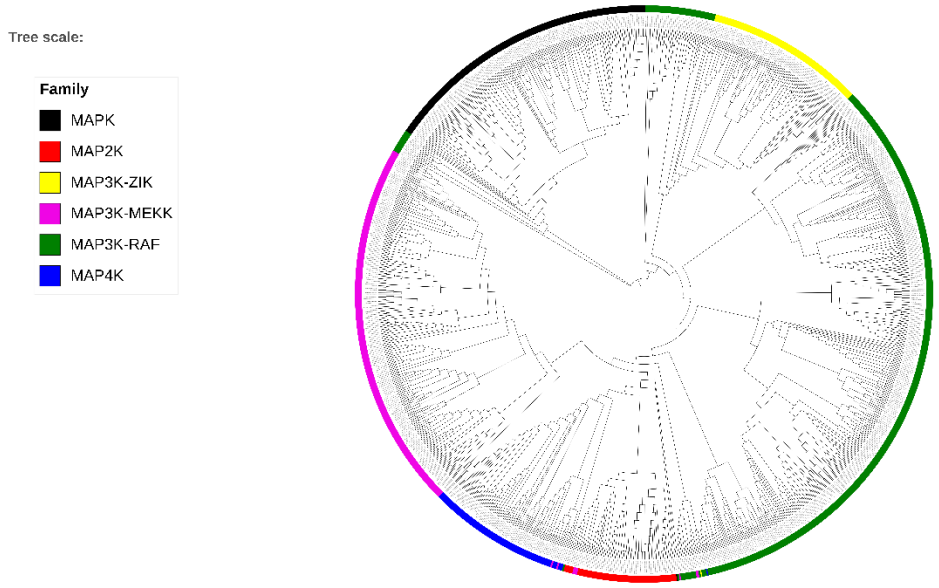

Figure App. 3: A figure showing all genes identified within the Mitogen-activated protein kinase family. Colors specify individual families.

3. Analysis of the LRR-RLK Family in seven plant species

Introduction

Receptor-like kinases (RLK's) are kinases that interact with external stimuli to trigger phosphorylation cascades. RLK's are the most abundant kinase family in plants, with 610 representatives in Arabidopsis alone [1]. These can be further divided into 44 subfamilies based on the presence of additional domains such as Leucine-Rich Repeats. Most RLK's are transmembrane in nature, but a group of RLK's known as Receptor-like Cytosolic Kinases (RLCK's) primarily in the cytoplasm, with little to no extracellular domains. The intracellular portions of all RLK's contain the protein kinase domain, which is necessary for phosphorylation and kinase activity.

LRR-RLK's are distinguished from other RLK's due to their abundance of Leucine-Rich Repeats. Likewise, many of the other RLK subfamilies are characterized by their additional

motifs. Leucine Rich Repeats typically consist of the motif: LxLxxNxL and can occur a variable number of times.

Methods and Results

In the analysis of LRR-RLK's relating to the 2017 paper on ortholog identification, GeneFamilyRF predicted an Arabidopsis LRR-RLK that was not predicted in the paper, while another was predicted by the paper, but not by GeneFamilyRF. The one predicted by GeneFamilyRF but not the paper was AT1G29730 and the other that had been predicted in the paper but not GeneFamilyRF was AT3G46350. AT3G46350 when ran through NCBI's CDD search showed that it had an LRR-RLK domain with the Leucine-rich repeats near the C-terminal end of the gene, as well as a large Malectin-like domain, meaning that it was still very likely to be an LRR-RLK. AT1G29730 showed a conventional LRR-RLK domain layout, with Leucine-rich Repeats near the N-terminus and a complete protein kinase domain. This gene has also been identified as an LRR-RLK in previous papers as well [1,2]. Using a q-value threshold for the Leucine-rich repeat motif of .23 resulted in many sequences not demonstrating it in CDD search meaning that a good q-value threshold is necessary for proper identification of motifs. AT3G46350 was not present in the FIMO output indicating that it may have been removed due to either being below the threshold or the maximum number of transcripts stored forced it to be removed to save memory. Increasing the allowed maximum number of transcripts stored from 100000 to 120000 did not show any improvements. Incremental increases to the amount lead to setting the amount to 8000000, an 80-fold increase. While this prevented entries from being removed, the other Arabidopsis gene was still excluded from FIMO's output. After further testing, this was discovered to be a result of exclusion due to the relatively high p-value for the

first motif. The effective fix for this was starting at a very relaxed threshold and tightening it to the point where all of the Arabidopsis training genes were within the output.

When using the 2 motifs, FIMO seemed to be unable to pick all LRR motifs without also choosing some motifs that are not LRRs. This was determined to be the result of The way that it was worked around was using MEME to find the best Leucine Rich Repeat and Kinase domain motifs to use by scanning for 2 motifs on Arabidopsis LRR-RLKs then feeding the output motif into FIMO through GeneFamilyRF rather than an IUPAC format previously determined motif. This removed most erroneously identified genes, such as Lectin-RLKs and increased the number of genes identified that were previously published as LRR-RLKs. The change resulted in going from 198 genes previously published but not predicted by GeneFamilyRF and 201 "new" genes to 74 and 57 respectively.

GeneFamilyRF Results

In addition to all of the previously published Arabidopsis LRR-RLK genes, 2 more were identified. These 2 genes showed LRRs and protein kinase domains in NCBI's CDD search and were previously identified by other publications to be LRR-RLKs.

Gossypium hirsutum and Zostera marina, which had not previously had LRR-RLK genes identified, contained 634 and 164 genes identified respectively.

In Gossypium raimondii, 368 of the 385 (95.6%) previously published genes were identified by GeneFamilyRF. Of Solanum lycopersicum's 218 LRR-RLK genes, 210 were identified, which is 96.3%. 470 of 484 (97.1%) genes in Glycine max were identified by GeneFamilyRF.

In Zea mays, only 175 of 210 (83.3%) genes were identified. One of these, GRMZM2G316474, was not in the database file for GeneFamilyRF or in Phytozome. Of the remainder of mays genes, 9 did not display Leucine-Rich Repeats in NCBI's CDD search, either due to a lack of

them or the LRRs having low E-values. GRMZM2G404647, which was also one of the 9 without an LRR, and GRMZM2G144923 listed in Phytozome as cytosolic kinases. Numbers and comparisons of genes can be seen in App. Table 4.

App. Table 4: A comparison of literature LRR-RLK to the ones identified by GeneFamilyRF. Ident. is the number identified in the previous study examined, Publ. is the number previously published, and New is the number of putative novel MAPK genes.

LRR-RLK				
Species	Publ.	Ident.	New	Spec. Total
A. thaliana	222	222	2	224
G. raimondii	385	368	11	379
G. hirsutum	N/A	N/A	634	634
Z. mays	210	175	6	181
S. lycopersicum	218	210	7	217
G. max	484	470	31	501
Z. marina	N/A	N/A	164	164
Category Total	1519	1445	855	2300

CDD Search

A query for NCBI's CDD database was made using default settings with the new and missing genes as inputs. This revealed that 9 of the missing genes were missing Leucine-Rich repeats, all of which were Z. mays genes. In addition to this, only 7 of the newly identified genes were missing Leucine-Rich Repeats as well, possibly as a result of partial LRR degradation due to changes in amino acid sequence or by CDD search not detecting them. 5 of the 7 were from Glycine max, while the other 2 belonged to Z. mays and G. raimondii. All genes in both categories contained what was identified as at least a segment of the kinase domain, which is not especially scrutinized as CDD search also showed only segments in some of the characterized Arabidopsis genes.

Transmembrane Domain Analysis

Leucine-Rich Repeat Receptor-like Kinases are transmembrane proteins, featuring an external portion of the protein which contains the Leucine-Rich repeats used to detect external stimuli. As a result of this, all LRR-RLKs are expected to contain transmembrane helices, which can be predicted using various software. For this analysis, TMHMM was chosen to identify transmembrane helices in the newly identified and previously published but not identified genes. This revealed that 11 of the previously published but not identified genes did not contain a predicted TM helix, along with 2 that potentially had no helix due to the number of predicted transmembrane amino acids was slightly below the number typically representative of transmembrane helices. Interestingly, 6 of the 11 not predicted to contain transmembrane helices were from the species Glycine max, possibly due to differences in the more recent Glycine max gene annotation. Another revelation from the TMHMM results was that 14 of the newly identified genes also did not have likely transmembrane helices. This was most prevalent in Zea mays genes, which represented 10 of the 14. The remainder were 2 in Glycine max, 1 in Solanum lycopersicum, and 1 in Gossypium raimondii.

A tree has been made with all LRR-RLKs identified by GeneFamilyRF, as can be observed in Figure App. 4. Currently, no bootstrapping has been performed on the dataset. A tree was made using muscle through MEGA to obtain alignments, which were then used to produce a tree with MEGA using maximum likelihood method. The tree was then uploaded to iTOL and the historical subgroups labelled, but not the sub-subgroups. Overall, the subgroups seemed to group with their own with a few subgroup portions lying within a different subgroup. Additionally, there is a small branch of the tree near the bottom that is a conglomeration of 7 different subgroups but containing only 16 genes. All sub-subgroups except for XII were separated and

were grouped near other subgroups. Three small clusters of SGIII were spread within and between other subgroups despite having no sub-subgroups listed. A majority of subgroup IV was within a branch of subgroup VI with a small cluster in a branch of SGXI. XIIIb was also found in a large branch of XI. Outside of these exceptions, there were very few genes outside of the primary branch of their subgroup, as can be observed in Figure App. 4.

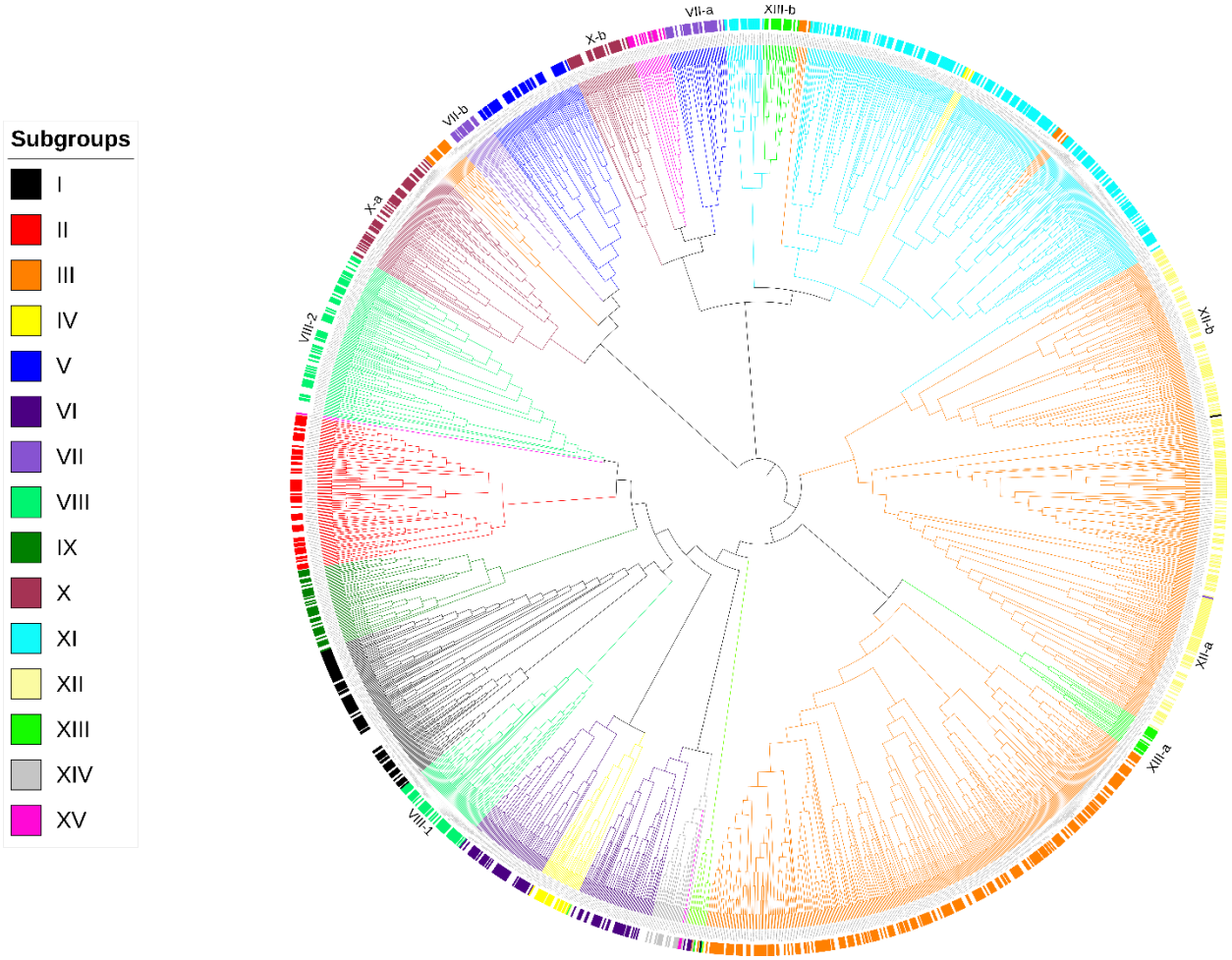

Figure App. 4: A tree showing genes identified by GeneFamilyRF as LRR-RLKs. Colored strips show the subgroups of previously published genes, with sub-subgroups labelled.

Gene Duplication Analysis

The information obtained with MCScanX was the duplication type involved in the genesis of the genes in the family, as well as the genes related to LRR-RLKs through duplication events. Of particular interest was Gossypium hirsutum due to its allotetraploidy from its progenitor species, with a Circos-produced collinearity shown in Figure App. 5. A more specific analysis of genes on the primary chromosomes was performed. This resulted in MCScanX identifying 530 of the 628 (634 including other chromosomes) genes identified as LRR-RLKs as being involved in Whole-Genome Duplication (WGD) or segmental duplication. MCScanX was also used to calculate Ka (nonsynonymous mutations) and Ks (synonymous mutations) then using these to calculate the Ka/Ks ratio, which typically represents selection pressure on the gene, with ratios > 1 showing positive selection towards changes in the amino acid sequence and < 1 showing selection of mutations which conserve amino acid sequence. Then, the collinearity data from MCScanX was used to create a collinearity chart in Circos v69, as seen below. The collinearity chart visually shows how much of the collinear relationships are between the A and D genomes, as is expected for genes in Gossypium hirsutum.

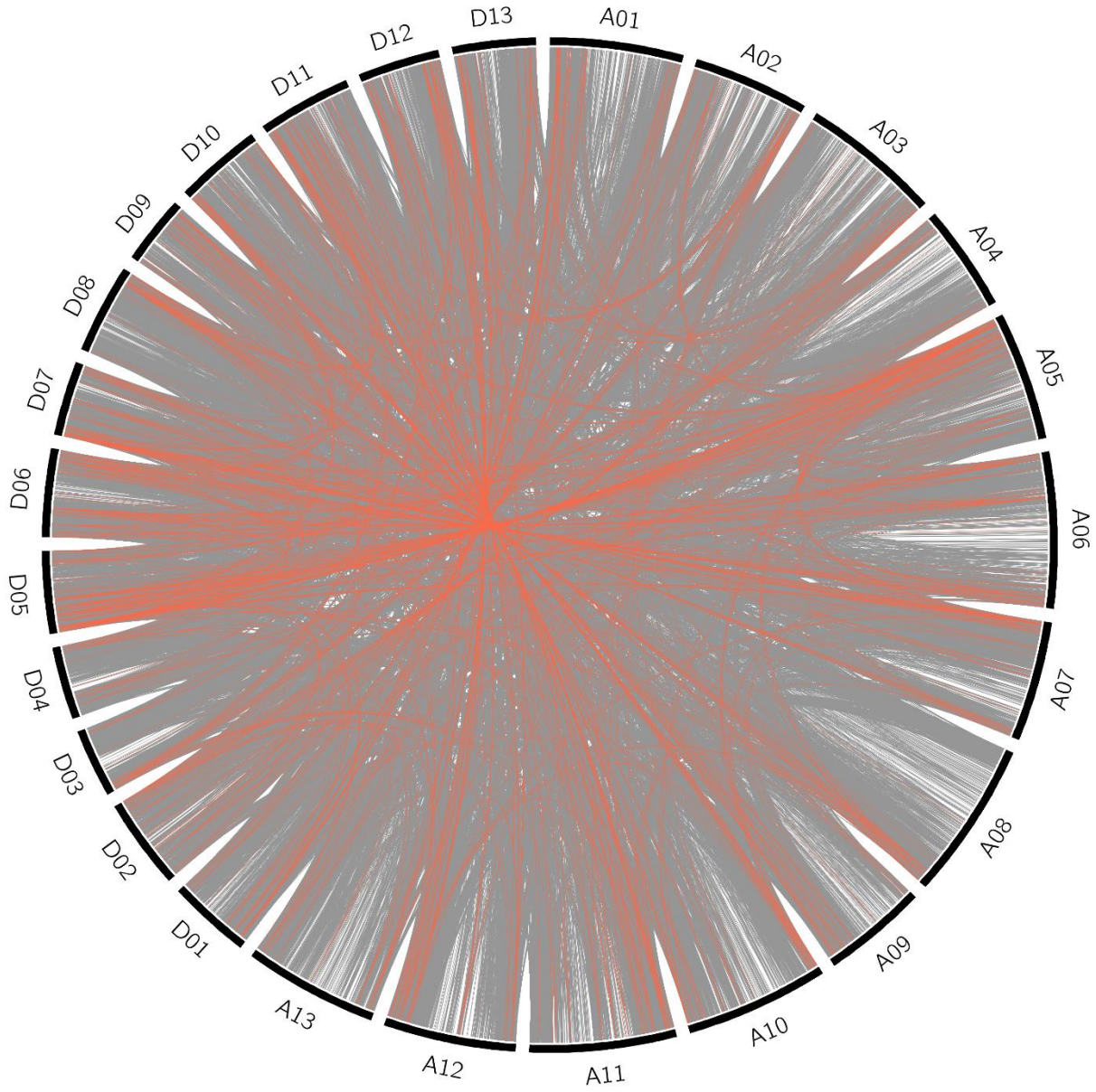

Figure App. 5: Collinearity is represented as lines between chromosomes. Chromosomes are represented by the black curves on the outside of the image with their respective labels. Red lines show collinear relationships involving genes identified as LRR-RLKs, while grey lines represent other collinear connections within the G. hirsutum genome.

Expression Analysis

Additionally, an analysis of the expression of the identified Gossypium hirsutum genes using cottonFGD (cottonfgd.org) was performed. All time periods for each stressor were averaged and compared to the control for fold-change. This revealed that 63 identified genes were downregulated by at least 50% under cold-stress and 69 genes altogether were downregulated by at least 50% during at least 1 stressor, which means that 92.8% of all >50% downregulated genes are downregulated during cold stress. In addition, 244 of 299 (81.6%) of the genes that were downregulated by at least 20% were downregulated under cold stress.

17 of 29 genes upregulated by at least 50% did so under PEG (drought simulation) treatment, and 11 of 29 did so during salt treatment. 108 genes showed at least 20% upregulation in at least one treatment, of which only 23 were upregulated during cold treatment.

The most significantly upregulated gene during any stress treatment was Gohir.A05G251900, which featured a 5.4-fold change during salt-stress, resulting in a change in FPKM from 4.46 to 28.66. This gene was also upregulated by at least 50% in every stress category except cold stress where it was downregulated by ~25%.

Previous Publications

A paper identifying LRR-RLKs in Gossypium Hirsutum has been found [2]. This paper uses BLASTP as the primary method to identify similar sequences to LRR-RLKs in Arabidopsis, then using hmmscan and CDD search to verify the ones that contain both a kinase domain and at least one LRR. The hirsutum genome searched within for LRR-RLKs was obtained from the CottonGen database, which uses a different ID system and likely a different annotation of the genome. The number of Gossypium hirsutum genes identified by the paper was 543, with a focus on the possible orthologs of Arabidopsis Stress Induced Factor (SIF) genes.

References

1. **Shiu, S.-H.**, and Bleecker, A. B. (2001). Plant receptor-like kinase gene family: diversity, function, and signaling. *Science STKE* 2001, RE22.

2. Yuan N, Rai KM, Balasubramanian VK, Upadhyay SK, Luo H, Mendu V. Genome-wide identification and characterization of LRR-RLKs reveal functional conservation of the SIF subfamily in cotton (Gossypium hirsutum). *BMC Plant Biol*. 2018;18(1):185. doi:10.1186/s12870-018-1395-1