Peer-reviewed version available at Genes 2019, 10; doi:10.3390/genes10120963

Phylostratigraphic analysis shows the earliest origination of the stress associated genes in *A.thaliana*

Zakhar Mustafin[1], Dmitrii Konstantinov[1,2], Vladimir Zamyatin[1,2], Aleksej Doroshkov[1,2], Sergey Lashin[1,2*], Dmitry Afonnikov[1,2,*]

mustafinzs@bionet.nsc.ru

konstantinov@bionet.nsc.ru

zamyatin@bionet.nsc.ru

ad@bionet.nsc.ru

lashin@bionet.nsc.ru

ada@bionet.nsc.ru

Affiliations:

1. The Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (IC&G SB RAS);

2. Novosibirsk State University (NSU);

* Correspondence: lashin@bionet.nsc.ru; Tel.: +7-383 363-49-63; ada@bionet.nsc.ru; Tel+7-383 363-49-63

Abstract: Phylostratigraphic analysis is a way to look anew on phylogenetic data in the evolutionary aspect. It allows counting the evolutionary age based on the analysis of genes, their orthologs and finding the last common ancestor. We performed phylostratigraphic analysis of *Arabidopsis thaliana* genes associated with several types of abiotic stresses (heat, cold, water-related, light, osmotic, salt, and oxidative) determined by the Gene Ontology annotation. Comparison of the distributions of ages of genes associated with stresses of different type has shown the heat stress to involve older genes while the light stress – younger genes. At the same time, all types of stress are characterized by a significantly higher proportion of old genes (common to all eukaryotes) compared to the whole set of *A.thaliana genes*. This can be explained by the involvement of basic molecular processes in plant cells into the stress response. Reconstruction and graphical analysis of the gene network of the heat stress educed several clusters associated with different response functions. Some of these clusters contain only ancient genes. The results obtained show that the phylostratigraphic analysis reveals the fundamental features of the organization of gene networks and their evolution.

Keywords: gene network, network analysis, transcription regulation network, Cytoscape, gene family evolution, divergence, A. thaliana, abiotic stress.

1. Introduction

The *de novo* origin of genes [1] is one of the key evolutionary processes. It causes the changes in repertoire of genes in plants (and other organisms) as well as duplication and specialization [2]. The origin of genes may be associated with the processes of genetic recombination, the activity of viruses and transposons, horizontal transfer of genetic material. Further consolidation of novel genes in the genome and their evolution is caused by the emergence of new adaptive functions in response to environmental changes [3].

The evolutionary history of a gene including the determination of the moment of its origination may be back-traced using phylostratigraphic analysis [4]. It is based on the reconstruction of a species tree for the analyzed organism. At first, taxa of interest to the researcher are distinguished, as a rule, reflecting all the most important events in the pedigree of the organism, relying on the reliability of phylogenetic relations and the data available to researchers. Every two boundary taxon of the built tree is determined by one phylostratum. For each gene under consideration, the analysis of its orthologs determines a taxon that will be basal, including the earliest ancestor of the gene and all its orthologs-descendants.

Phylostratigraphic analysis of genes is of great interest in relation to identification of important stages of genome evolution, where rapid growth of new genes took place [5] identification of Lineage specific genes. On the other hand, a close relationship between the age of genes and the level of their expression in the process of embryogenesis was shown for both animals [6] and plants [7]. Interestingly, an hourglass-like pattern of gene expression by age was also identified in the response of tobacco plants to biotic stress [8].

One of the interesting tasks is to find functional features of genes that differ in age. In particular, a number of data suggest that genes associated with fundamental processes in cells usually are older than

other genes. For instance, the study [5] reported that human genes referred to such phylostrata as Cellular organism and Eukaryota, are generally associated with basal cellular functions (metabolic processes, transcription regulation), while the genes originated in the later stages of evolution are associated with the genes of the immune response and reproduction. In plants, older genes are also associated primarily with fundamental cellular processes (photosynthesis, RNA transcription and processing, primary metabolism), and younger genes are associated with secondary metabolism, hormonal regulation, and transcription regulation [9]. Expression of the youngest genes of *Arabidopsis thaliana* showed a bias to mature pollen, and was enriched in a gene co-expression module that correlates with mature pollen [10].

Since the molecular basis of the phenotype of the organism are gene networks, the study of the relationship of the age of genes with their functions and interactions in gene networks is of particular interest. Large-scale analysis of co-expression networks and gene ages in *A. thaliana*, *Oryza sativa* (rice) and *Physcomitrella patens* (moss) demonstrated, that genes from the same evolutionary period tend to be connected, whereas old and young genes tend to be disconnected and the modules of the same age emerged at a specific time in plant evolution [9].

Previously, we developed a Cytoscape application Orthoscape for analysis and visualization of the ages of genes in the context of the structure of their gene networks [11]. In the present study, the Orthoscape was used to analyze genes associated with seven types of abiotic stress in *A. thaliana*. Stress response in plants is crucial for their adaptation to environment and evolution. Genetic systems for responding to abiotic stresses have been studied in model plants such as rice and Arabidopsis [12,13]. These systems consist of coordinately functioning genes, they have a level of evolutionarily plasticity, and their composition may significantly change in the process of evolution due to the large role of segmental and full-genome duplications in plants [14]. Using the Orthoscape application, we carry out phylostratigraphic analysis of genes of plant stress, including the assessment of the distributions of these genes according to the evolutionary age as well as the reconstruction and graphical visualization of gene networks by the example of the network of the heat stress response. Our results have shown that the genes associated with different types of stress differ in age. However, in general, the response to stress involves a significant proportion of old genes. Graphical analysis of the reconstructed gene network of the heat stress has demonstrated its modular organization where some modules are represented by age-homogeneous genes. The study demonstrates that the use of phylostratigraphic analysis allows to obtain new interesting data on the evolution of genes of stress response in plants.

2. Materials and Methods

2.1 Gene sets preparation

We analyzed *Arabidopsis thaliana* sets of genes associated with seven types of abiotic stress: heat stress, cold stress, water-related stress, light stress, osmotic stress, salt stress, and oxidative stress. TAIR database v. 20170930 [15] annotation has been used. Gene sets have been formed on the basis of Gene Ontology (GO) terms [16], represented in the TAIR annotation (totally 49963 terms). When selecting an annotation, only the terms of the following confidence levels were used: inferred from direct assay (IDA), inferred from mutant phenotype (IMP), inferred from genetic interaction (IGI), inferred from physical interaction (IPI).

In the first step, extended lists of GO terms associated with each type of stress were formed. To do this, we selected all the terms that contained the keyword "stress" in either title or description, as well as all their child terms. After the formation of the initial list, its refinement was carried out, the terms GO not associated with this type of stress were removed. As a result, we have selected 161 terms that characterize particular types of stress. Subsequent analysis showed that the lists of terms associated with the keyword 'water' and 'drought' were substantially overlapped: 25 terms were associated with the keyword 'water' and 10 with 'drought', 6 terms were common. Therefore, these two lists in our analysis were combined under the name "water-related stress".

2.2 Network reconstruction for gene sets

To reconstruct gene network for the set of genes, interactions with the level of confidence above 0.7 were searched using the STRING database. It should be noted that the search in the STRING database can change the composition of genes in the reconstructed gene network both by excluding genes from the input list for which no interactions were detected, and by adding new genes (in this paper, we allowed to add no more than five additional genes to the existing list). Generated STRING tables were then loaded into Cytoscape [17] for visual network reconstruction and analysis via Orthoscape application [11].

2.3 PAI/DI calculation and network visualization.

We used the Orthoscape application [11] for analysis for gene sets of plant stress response and visualization of their reconstructed networks. Orthoscape loads lists of genes and their network relationships either from KEGG database or user-defined file. For each gene in the network, the Orthoscape calculates two evolutionary indices. First, the phylostratigraphic age index (PAI), order number of a phylostratum, indicating the evolutionary age of a gene based on the finding of the most basal taxon, common for the gene and every of its orthologs [4]. The lower PAI is, the lower is the phylostratum number and the earlier is the gene appeared in the course of the organismal evolution [4]. The Orthoscape uses the KEGG Organisms database [18] to get taxonomic trees. It performs a search of orthologous genes, populates the tree of species these genes belong to and then analyses the resulting tree [11]. In the KEGG database, the taxonomic tree for *A.thaliana* contains the following 18 taxa (corresponding PAI values are shown in parentheses): Cellular Organisms (0), Eukaryota (1), Viridiplantae (2), Streptophyta (3), Embryophyta (4), Tracheophyta (5), Spermatophyta (6), Magnoliophyta (7), Eudicotyledons (8), Gunneridae (9), Pentapetalae (10), Rosids (11), Malvids (12), Brassicales (13), Brassicaceae (14), Camelineae (15), Arabidopsis (16), *A.thaliana* (17). However, as a result of the analysis it was found that some of the taxonomic groups

(Streptophyta, Spermatophyta, Gunneridae) were not basal for any ortholog group of the studied *A.thaliana* genes. Therefore, these taxa were excluded from further analysis. It should also be noted that the list of the genes described in KEGG contains 32690 elements (http://rest.kegg.jp/list/ath). However, only those genes have been selected for the analysis, for which at least one annotation term was found in the Gene Ontology database. This list included 25843 genes, and below it is assumed as the background *A. thaliana* genes list. As a result, such a reduction allows us to take into account the fact that younger genes are less annotated in the GO.

Second evolutionary index is the divergence index (DI) of a gene indicating the influence of natural selection on gene evolution. It is based on the estimation of the Ka/Ks ratio [19] between the gene from the analyzed organism and the most similar ortholog from its closest relative organism, *Arabidopsis lyrata*. The larger the DI value is the higher is the pressure of the Darwinian selection on its sequence. Low values of the DI indicate stabilizing selection acting on a gene.

The Orthoscape report the following results: a graphical representation of a gene network graph in which each node of the network corresponding to a gene is colored according to PAI or DI; the value of PAI values for genes in the gene network, and the result of Ka/Ks ratio evaluation for genes. Orthoscape also provide its output in HTML format along with the generated R scripts that can be used for drawing violin plot for all the distributions obtained. HTML reports contain also the data of specific PAI calculated using weights according the node connectivity.

3. Results

3.1 GO terms and genes associated with abiotic stress

The list of GO terms associated with stress and the list of *A.thaliana* genes, annotations of which contain these terms are presented in Supplementary file 1. The number of GO terms and genes associated with each type of stress are presented in the table 1. The least specific GO terms (4) were found for cold stress, the most terms (48) were found to be related to the light stress. For other stresses we have found from 14 to 28 associations with GO terms.

A list of genes associated with different types of stress is given in the Supplementary file 2. For each of the types of stress we have identified no less than 100 genes (minimum, 102, genes for heat stress; maximum, 232, genes for salt stress). Interestingly, there was no significant linear correlation between the number of GO terms and the number of genes associated with these terms (Pearson correlation coefficient between these values was found to be 0.09).

Identification of genes from the resulting list in the KEGG database allowed us to find almost all genes corresponding the TAIR annotations: for most lists of genes, only 1-6 genes were not detected; only for the list of the light stress 13 genes were excluded.

Table 1. Number of GO terms and genes that have identified associations with studied stresses.

Stress type	Number of GO terms	Number of genes	KEGG number of genes
Salt	17	232	231

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2018 doi:10.20944/preprints201811.0352.v1

Peer-reviewed version available at Genes 2019, 10; doi:10.3390/genes10120963

Heat	14	102	102
Light	48	155	142
Water-related	27	210	205
Cold	4	146	140
Osmotic	23	115	113
Oxidative	28	153	151

There are genes common for different gene sets. For example, 13 genes of salt stress (5.6% of the total number, 232) included in the heat stress dataset (TAIR annotation). The number of genes common between pairs of stress datasets provided in table 2 along with the numbers of unique genes for each stress. It is apparent from the table that the gene set for osmotic stress shares largest fraction of genes with other datasets (40% with salt, 25% with water-related, 15% with cold and 10% with oxidative stresses). From the other hand, large fraction of gene sets have a number of genes in common with salt stress dataset (5 out of 6 types have more than 10% of common genes with this type of stress). However, the majority of comparisons yield less than 10% of common genes (28 out of 42). The fraction of unique genes for datasets is lower than 50% for only one type of stress, osmotic (30%), for three datasets it is greater than 70%, for other three datasets it is greater than 50% (table 2).

Therefore, we will analyze the seven types of gene sets separately, however bearing in mind that some pairs of gene sets may overlap quite remarkably.

Table 2. The number of common genes between pairs if stress gene sets. Each cell in the table represent the fraction (and number, in parentheses) of genes from the set of the row in common with the set of the column. The cells with fraction of genes larger than 0.1 shown in bold. The last column represent the number of unique genes for the stress in the row.

	Salt	Heat	Light	Water-related	Cold	Osmotic	Oxidative	Unique genes
Salt	232	13 (0.06)	7 (0.03)	**41 (0.18)**	18 (0.08)	**47 (0.20)**	18 (0.08)	126 (0.54)
Heat	**13 (0.13)**	102	8 (0.08)	**11 (0.11)**	8 (0.08)	8 (0.08)	6 (0.06)	72 (0.71)
Light	7 (0.05)	8 (0.05)	155	12 (0.08)	9 (0.06)	3 (0.02)	6 (0.04)	120 (0.77)
Water-related	**41 (0.20)**	11 (0.05)	12 (0.06)	210	18 (0.09)	**29 (0.14)**	11 (0.05)	124 (0.59)
Cold	**18 (0.12)**	8 (0.05)	9 (0.06)	**18 (0.12)**	146	**17 (0.12)**	6 (0.04)	93 (0.64)
Osmotic	**47 (0.41)**	8 (0.07)	3 (0.03)	**29 (0.25)**	**17 (0.15)**	115	12 (0.10)	35 (0.30)
Oxidative	**18 (0.12)**	6 (0.04)	6 (0.04)	11 (0.07)	6 (0.04)	12 (0.08)	153	118 (0.77)

3.2 Analysis of PAI and DI indices

We have calculated PAI indices for each gene from the stress datasets and for all genes in *A.thaliana* genome. To estimate the difference between PAI distributions for genes from the stress datasets and the background set of *A.thaliana* genes (see Methods) we have calculated difference between frequencies of occurrence of PAI values in the stress dataset and all *A.thaliana* genes. The diagrams for these differences are shown in figure 1 for each taxonomic level (at the similarity threshold 0.7).

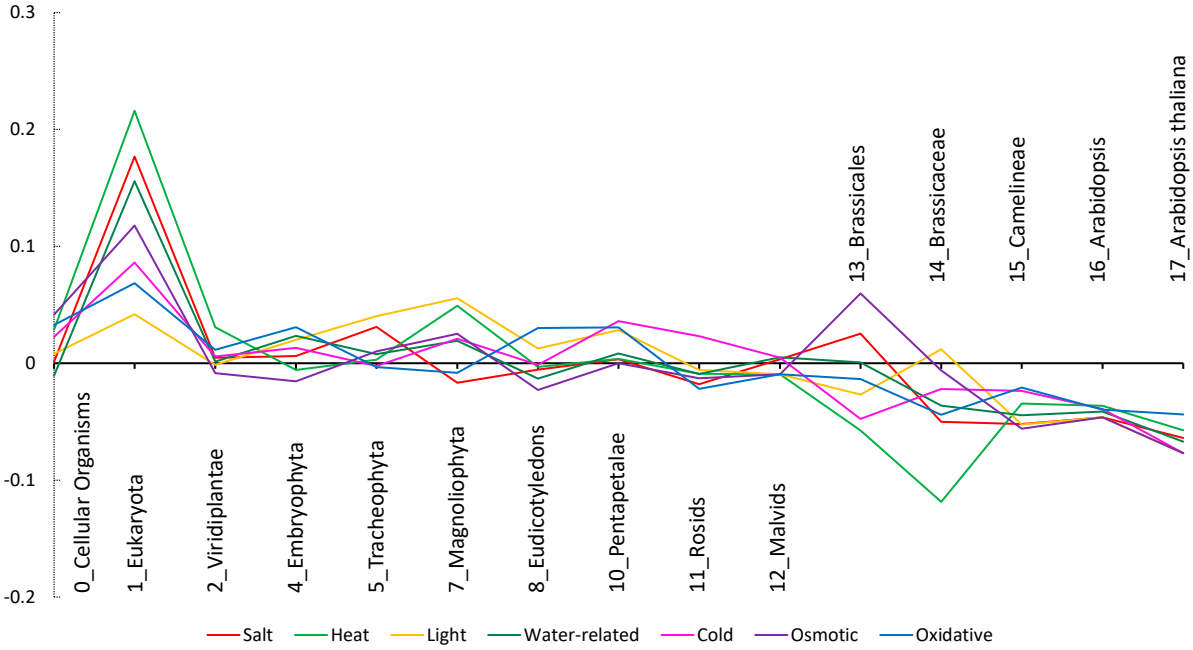

Figure 1. Difference between the frequencies of occurrence of PAI values in stress dataset and all *A.thaliana* genes. X axis represent the taxon from Cellular organisms (PAI=0) to *A.thaliana* (PAI=17). Plots for different sets of genes shown in different colors.

This diagram clearly demonstrates that PAI for genes from the analyzed stress datasets have higher fraction of genes with lower PAI values in comparison with all genes distribution. For instance, large excess of genes from stress the datasets is observed for Eukaryota taxon (all difference values in stress datasets are positive). The positive values for this plot is also the characteristics of the stress datasets at the Cellular organism, Viridiplantae, Embryophyta and other taxonomic levels, which are lower than Rosids. For large PAI values (> 10) most of the difference values are below zero. Interestingly, the PAI values of the heat stress genes demonstrate the most pronounced shift towards smaller value: largest fraction of genes with PAI=1 (Eukaryota) and smallest fraction of genes with PAI = 12, 13, 14 (Malvids, Brassicales, Brassicacea).

We performed comparison of the PAI distributions for stress datasets and all *A.thaliana* genes distributions using chi-square test. It should be noted that the data on the number of stress response genes for different taxonomic levels showed a lot of zero values. For instance, the genes involved into the osmotic stress do not have any orthologs at Viridiplantae, Embryophyta, Eudicotyledons, Malvids, Arabidopsis and *A. thaliana* taxa, while for Tracheophyta, Pentapetalae, Rosids, Camelinae taxa number of orthologous

genes is less than 5. Moreover, the expected number of orthologous genes for this type of stress estimated based on their proportion among all *A. thaliana* genes for Cellular Organisms, Viridiplantae, Embryophyta, Tracheophyta, Eudicotyledons, Pentapetalae, Malvids taxa was also found to be less than 5. These circumstances did not allow the Chi-square test to be used to compare distributions directly. Therefore, when comparing the distributions by the number of genes for different PAI, we combined the taxa into 4 large groups, in which the number of genes for all types of stress in the PAI distribution is not less than 5: (Cellular Organisms, Eukaryota), (Viridiplantae, Embryophyta, Tracheophyta, Magnoliophyta), (Eudicotyledons, Pentapetalae, Rosids, Malvids, Brassicales), (Brassicaceae, Camelineae, Arabidopsis, *A. thaliana*). The results of the analysis of differences between PAI distributions for all stress-type gene lists are shown in table 3. The analysis demonstrates, that PAI distribution for all stress datasets differs significantly from PAI distribution for all genes in *A.thaliana* (at significance level $p<0.05$).

Table 3. Results of the chi-square test of comparison between PAI distribution in various stress gene sets and all *A.thaliana* genes.

Stress type	Chi-square	Significance p-value
Salt	66.50	$2.39 \cdot 10^{-14}$
Heat	55.40	$5.63 \cdot 10^{-12}$
Light	23.20	$3.66 \cdot 10^{-5}$
Water-related	45.13	$8.69 \cdot 10^{-10}$
Cold	19.30	0.000243
Osmotic	25.37	$1.29 \cdot 10^{-5}$
Oxidative	17.55	0.000544

The average values of PAI for various type of stress-associated gene sets at different values of the sequence similarity thresholds are shown in table 4. This table demonstrates that the heat stress genes have the smallest average PAI values for all similarity thresholds. This suggests that the genes associated with the heat stress diverged mostly at the early stages of the plant evolution. This is in agreement with the Figure 1: the heat stress gene set has largest fraction of genes at low PAI (1, Eukaryota; 2, Viridiplantae) and lowest at high PAI (13, Brassicales; 14, Brassicaceae). The second type of stress that in general have low PAI values for genes is the salt stress.

Table 4. PAI and DI values for the set of genes associated with different types of stress. PAI were calculated using various sequence identity threshold.

	PAI					DI
Sequence identity	0.5	0.6	0.7	0.8	0.9	
Salt	2.61	5.05	7.46	9.45	12.20	0.18*

Heat	2.48*	4.20*	6.28*	8.97*	12.15*	0.20
Light	2.91	5.93**	8.26**	10.74**	13.32	0.19
Water-related	2.53	5.25	7.75	9.67	12.66	0.19
Cold	2.86	5.68	8.09	10.61	13.58**	0.22**
Osmotic	2.88	5.61	7.98	10.24	12.50	0.18*
Oxidative	3.12**	5.50	8.21	10.56	13.50	0.22**

231 * Lowest value among all stress datasets for specific identity threshold.
232 ** Highest value among all stress datasets for specific identity threshold.
233

234 The opposite case is the gene set related to light stress. This sets of genes has the largest average PAI
235 values at 3 of 5 thresholds used (table 4). This is again in agreement with data shown in Fig. 1. Light stress
236 set has the smallest fraction of genes with PAI=1 (Eukaryota) and the largest fraction of genes at PAI=5 and
237 7 (Tracheophyta, Magnoliophyta). Oxidative stress has also relatively small number of genes with PAI=1,
238 and larger number of genes with PAI=8 and 10 (Eudicotiledons, Pentapetalae). This suggests that these two
239 gene sets contain more 'younger' genes than others.

240 The DI average values (table 4, last column) are within the range 0.18-0.22. The greatest values of DI
241 (0.22) have been found in genes related to the oxidative and cold stresses.

242 In the previous section, it was shown that stress gene lists contain both genes unique for each type of
243 stress and genes common to two or even more lists. When analyzing the genes presented in KEGG, it turned
244 out that of all the genes we analyzed, 654 (78%) are represented only in one of the stress gene lists (unique),
245 and 185 (22%) in two and more lists (shared). We compared the distributions of PAI values for the four
246 generalized value intervals (see above for a description of the distribution comparison) to avoid a small
247 number of genes (less than 5) for these intervals. The results are presented in figure 2. The figure shows
248 that for genes that are associated with only one type of stress, the distribution contains more genes with
249 higher PAI values (intervals [8,13] and [14,17]) compared to genes that are associated with two or more
250 types of stress response. The Chi-square test showed a significant (at $p<0.05$) difference between the two
251 distributions ($p=0.029$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2018 doi:10.20944/preprints201811.0352.v1

Peer-reviewed version available at Genes 2019, 10; doi:10.3390/genes10120963

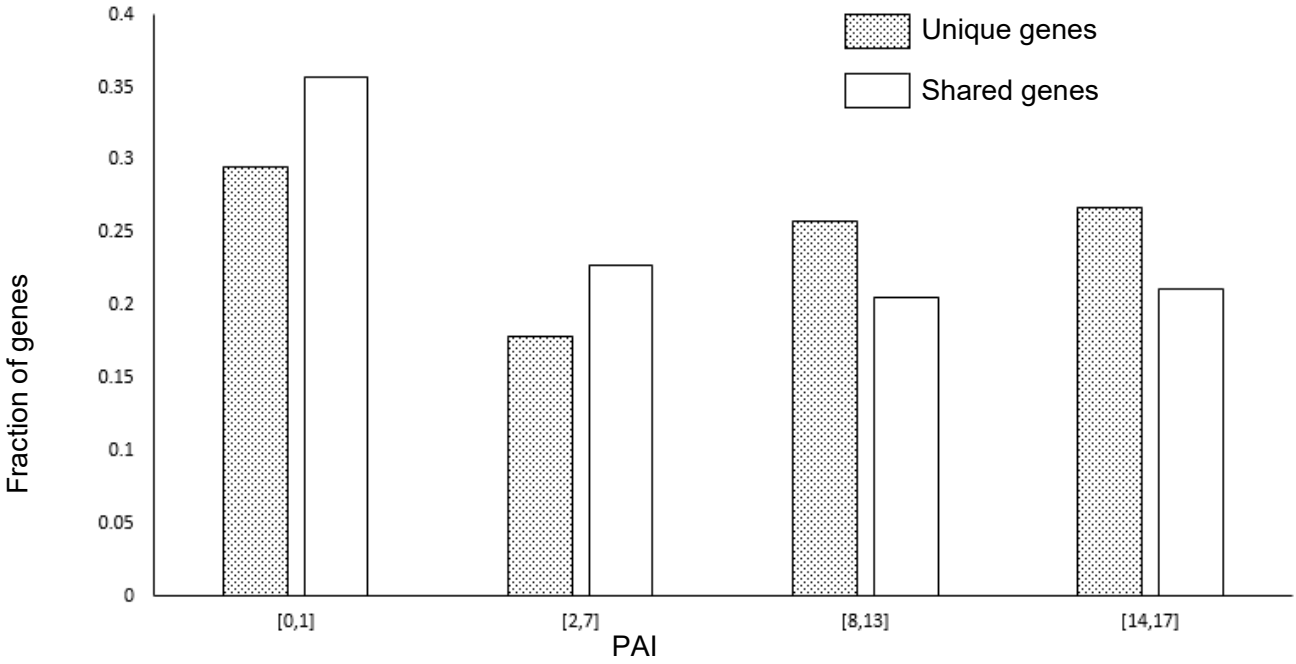

Figure 2. Frequency of occurrence of stress-associated genes (y-axis) for PAI values combined into larger intervals (x-axis). Dotted bars represent the histogram for genes that occur only in one type of stress lists (unique); the histogram of genes common to 2 or more types of stress (shared) is shown by white bars. PAI intervals correspond to groups of taxa: [0,1] Cellular Organisms, Eukaryota, [2,7] Viridiplantae, Embryophyta, Tracheophyta, Magnoliophyta, [8,13] Eudicotyledons, Pentapetalae, Rosids, Malvids, Brassicales, [14,17] Brassicaceae, Camelineae, Arabidopsis, *A. thaliana*.

3.3 Analysis of the gene networks for heat stress gene set

We reconstructed gene network for genes represented in heat stress list. The network was visualized using the Orthoscape application (figure 3). In this network, five clusters have been identified. Cluster 1 comprises 23 genes. 13 genes are coding for heat shock proteins performing chaperone functions (gene ID is shown after the slash): BOB1/AT5G53400, HSBP/AT4G15802, BAG7/AT5G62390, Fes1A/AT3G09350, HSP21/AT4G27670, HSF3/AT5G16820, BIP2/AT5G42020, AR192/AT4G26780, HSC70-1/AT5G02500, HSP101/AT1G74310, BIP3/AT1G09080, ATERDJ3A/AT3G08970, HSP81-3/AT5G56010. 2 genes related to thioredoxin (GRXS17/AT4G04950, TDX/AT3G17880) and one gene, PP7/AT5G63870 is a housekeeping gene. Functions of other 8 genes in this cluster are less clear. We may speculate that this cluster of genes is responsible for protective functions related to heat shock proteins. It contains the majority of genes with low PAI and they have dense network of interactions. Most of these genes are unique for heat stress gene set.

Cluster 2 contains 20 genes. It includes transcription factors of WRKY (WRKY25/ AT2G30250, WRKY33/AT2G38470) and C2H2 (RHL41/AT5G59820) types, receptors for ethylene (ETR1/AT1G66340, XRN4/AT1G54490, EBP/AT3G16770), salicylic acid (NPR1/AT1G64280), abscisic acid (ABI1/ AT4G26080), hormone biosynthesis (ABA1/AT5G67030, ABA3/AT1G16540) are two enzymes controlling the first and the last steps of abscisic acid biosynthesis, respectively) and chromatin modifying protein ATCHR12/ AT3G06010. It is likely that these genes are responsible for regulatory functions in heat stress response.

Cluster 2 contains more genes with higher PAI values and more genes shared with other stress types. The interaction network for this cluster is sparser in comparison with Cluster 1.

Clusters 1 and 2 connected via hub gene, TE1 (ERECTA/AT2G26330), a receptor protein kinase, which is a pleiotropic regulator of developmental and physiological processes, as well as it is a modulator of responses to environmental stimuli [20], including heat stress [21].

Cluster 3 contains 6 genes. SUMO1/AT4G26840 [22] and SIZ1/ AT5G60410 [23] are related to ubiquitination. Two genes involved in the repair of strand breaks, and the excision repair in response to ultraviolet radiation: UVH6/ AT1G03190 [24] and UVH3/AT3G28030 [25]. Two remaining genes from this cluster involved in the mitochondrial genome stability, MSH1/AT3G24320 (a plant-specific protein involved in organellar genome stability in mitochondria and plastids [26] and RECA3/AT3G10140 [27]. It is likely that this cluster is related to the functions of responding to damage of proteins and DNA in a cell as a result of heat stress. Like the other regulatory cluster, 2, cluster 3 has sparse interactions and large fraction of genes with medium/high PAI.

The fourth cluster contains 13 genes, 10 of which are the only genes added to the initial heat stress gene set by STRING. Genes from this cluster have no connections to other clusters via STRING interactions. They tightly interconnected within the cluster. All of genes included in this cluster are proteasomal genes. It is likely that the function of this cluster is related to the degradation of proteins unfolded due to the heat stress.

We combined the rest four genes outside clusters 1-4 into the fifth cluster. It contain pair of genes associated with the biosynthesis of ascorbic acid: CYT1/AT2G39770, GDP-D-mannose pyrophosphorylase VTC1 [28] и VTC2/AT4G26850 [29]. Two other genes, a DEAD box RNA helicase LOS4/AT3G53110 [30] and exprotin XPO1A/AT5G17020 [31] are associated with the export of mRNA. These four genes have low PAI values (less than 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2018 doi:10.20944/preprints201811.0352.v1

Peer-reviewed version available at Genes 2019, 10; doi:10.3390/genes10120963

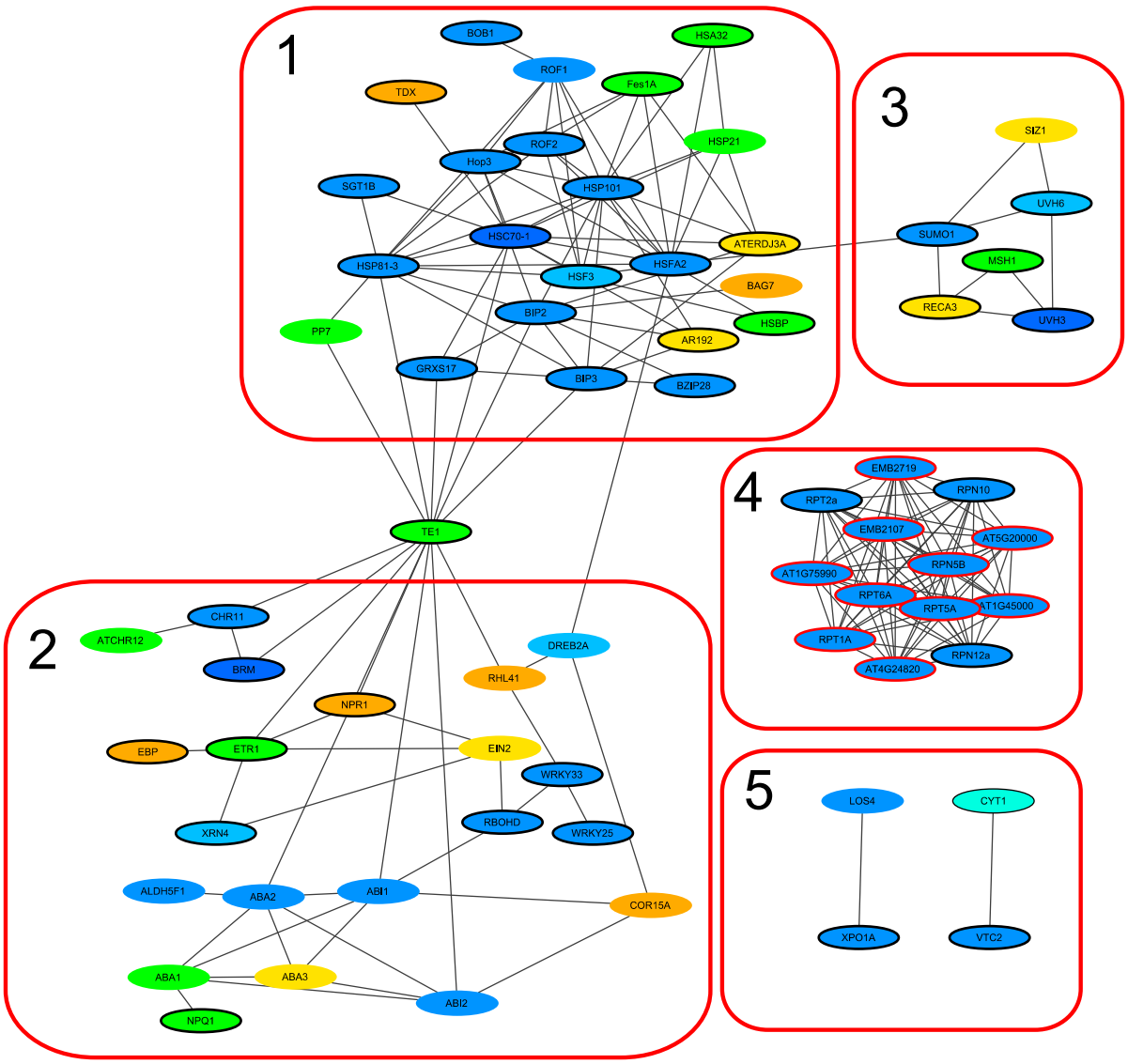

Figure 3. Gene network reconstructed for heat associated gene set using STRING tool. Node color correspond to the PAI index of the gene from 0 (dark blue) to 17 (red). Nodes that added to the gene set by STRING procedure of network reconstruction outlined by red color. Nodes corresponding to genes unique to heat gene set outlined by black color. Five clusters of genes shown by red rectangles and numbered.

Thus, the presented visual analysis of the gene network of plant response to the heat stress allowed to identify gene clusters associated with a number of key functions, as well as to visualize the gene network graph in accordance with the ages of genes.

4. Discussion

4.1 Features of gene annotation for different types of stress

The response of plants to stress of any nature affects a large number of molecular processes [32]. For example, the heat stress leads to the triggering of such processes in plant cells as change in membrane

fluidity, increase of the reactive oxygen species (ROS), change in the transport of Ca+ ions and restructuring of the cytoskeleton, the denaturation of proteins and RNA, changing the structure of chromatin and the expression of miRNAs [33]. The drought stress activates specific signaling pathways and transcription factors, detoxification enzymes, enzymes of the biosynthesis of osmolytes, system of transporters and water channels, response to protein denaturation [34]. The heat stress activates heat shock proteins, sumoylation systems, chromatin remodeling, dehydration control [35]. In response to the salt stress, genes of photosynthesis and carbon production, cell wall components, water channels, ion transport, ROS protection system, detoxification system, signaling pathways and specific transcription factors are involved [36]. It should be noted that the system of response to the osmotic and the oxidative stress themselves are involved in responses to other types of abiotic stress [37]. Thus, the systems of response to abiotic stresses in plants are closely interconnected. Our analysis of annotations of the stress genes in *A. thaliana* has indeed shown that the involvement of some genes in several stress responses is one of the features of stress genes. This was most noticeable for such stress as osmotic. More than 60% of the genes involved in responding to this stress are also involved in responding to other stress (table 2). A significant number of genes common to some lists were also identified for salt (almost 50%) and water-related stress (40%). At the same time, for some types of stress, more than 70% of genes were unique (heat, light, oxidative, see table 2). Overlapping lists of genes for stress pairs in most cases, however, was not more than 20%. Such an overlap generally looks natural; for example, the systems of salt, osmotic, water-related and cold stresses contain a significant proportion of common genes, since they are all closely related to the water and ion regime of cells. The presence of common and unique genes can be explained by the multilevel structure of molecular systems of response to stress [32]: as a rule, these systems include stress sensors, signal transmission systems (including hormonal response), triggering transcription of stress response genes, molecular response to the occurrence of stress conditions to minimize its consequences. Systems of the first and second level, as well as partly the regulation of genes, are mainly specific for each type of abiotic stress. At the same time, the molecular response to cell stress for different types of stress has many common features: control of reactive oxygen species (ROS), change of ion transport, cell detoxification, control of protein denaturation. The presence of specific and generic stress response genes seems to be related to the proportion of genes from lists that are relevant to either specific response levels or non-specific levels and how this is reflected in the GO annotation.

4.1 Age of genes involved in stress response

The Orthoscape application allowed us to estimate the age of the genes involved in the stress response and compare the distribution of these ages to the General distribution of all genes of the *A.thaliana*. As a result of this comparison, it was found that stress gene systems contain a greater number of genes, the origin of which is associated with the oldest taxa of living organisms, mainly with the levels of Cellular organisms and Eukaryotes. The explanation for this may lie in the fact that the stress response involves genes significant part of which is associated with the very basal functions of cells, functions that had already formed in unicellular organisms. These groups of genes include, in particular, chaperones (heat-shock proteins, Hsps) responsible for protein folding, assembly, translocation and degradation [38]. For example, some of them were identified as members of the heat stress response network (Fig. 3). One of these proteins, HSC70-1/AT5G02500, belongs to the HSP70 family, which includes the most conserved proteins present in all kingdoms of life [39].

Some transcription factors involved in the stress response are of ancient origin. For example, we have identified several transcription factors of the WRKY family in the gene network of heat stress response (Figure 3). For the domains of these transcription factors, homologues were found presented in such non-plant eukaryotes as the unicellular protist *Giardia lamblia*, one of the most primitive organisms that represent the earliest branching among extant eukaryotes, and the slime mold *Dictyostelium discoideum*, which belongs to the Mycetozoa, a lineage more closely related to animals and fungi than to green plants [40].

Signaling proteins-receptors, such as receptor-like kinases (RLKs) may sense change in the fluidity of cellular phospholipid membranes induced by heat and cold stresses [41]. It was shown that these proteins belong to the group monophyletic with respect to kinase domains when compared with the other eukaryotic kinase families [42]. Such ancient genes are more often involved in the overall functional core of the response to abiotic stress, which is demonstrated by the results of comparing the common and unique genes associated with stress (figure 2).

These results are in good agreement with data from Ruprecht et al [9], who showed that general biological processes, such as photosynthesis, glycolysis, DNA synthesis and others were already present in the ancestors of green plants. In the study above, based on the analysis of rice and moss genes it was shown that the 'stress' term is significantly enriched for ancient phylostrata like 'green plants' and 'vascular plants'. The same analysis for *A.thaliana* genes showed the 'stress' enrichment for 'vascular plants' phylostratum. Thus, it can be concluded that in accordance with our results, stress genes in plants have a rather ancient origin.

Previously in [43] they have identified Lineage - specific genes (LGSs) in *A. thaliana* that are restricted to the Brassicaceae family. The authors showed that new genes are more likely to exhibit differential expression in the conditions of plant response to stress (compared to other genes). These results, however, do not contradict ours. Although we have shown that stress response gene networks include a significant portion of ancient genes, some young genes are also involved in these networks. These young genes may be involved into regulatory modules of gene networks (Fig. 3), as well as in the system of sensitivity to stressors and therefore primarily respond to changes in its expression in response to external factors.

We have shown that there are differences between the ages of genes involved in different types of stress. Thus, the genes of the response to the heat stress contain the largest proportion of ancient representatives and the lowest values of PAI values. This is most likely due to the involvement in this response of such ancient families as chaperones and proteasomal proteins, which represent a significant proportion of all proteins in the set (Fig. 3).

The response to the light stress is characterized by the presence of younger genes, the average PAI values for them is the highest (Table 4). For this type of stress, the high value of the proportion genes belonging to such phylostrata as Trachaeophyta (vascular plants), Magnoliophyta (flowering plants) and Brassicaceae is observed (Fig. 1). As for the first two phylostrata, we can suggest that the high value of the proportion of the light stress genes for them may be due to the fact that notably vascular plants are characterized by the formation of leaves in the process of the plant evolution [44]. The leaves are considered as one of the innovations of vascular plants [45,46]. On the other hand, the vascular plants, mainly flowering,are characterized by a wide variety of leaf shapes [47]. It should be noted that one of the most important factors that affect the shape of the sheet is the light [48]. It is possible that the processes of

formation of the leaf, the plant organ, which is most closely related to the absorption of the light, is associated with the appearance of a noticeable part of the genes of the response to the light stress.

4.3 Gene networks and phylostratigraphic indices

The results of our analysis on the example of the gene network of the heat stress show several functional blocks. Gene age analysis demonstrated results similar to those obtained by Ruprecht et al [9]. Among the selected clusters in this network, three (1,4,5) represented the vast majority of genes of ancient phylostrata (Figure 3).

Thus, the results presented in this paper show that the analysis of phylostratigraphic indices at the level of gene networks, their visualization, can provide useful information about the relationship between the structural and functional features of gene networks and the evolution of genes that form them.

Supplementary Materials: The following are available online, Supplementary_file_1.xlsx: List of GO terms characterizing seven types of plant stress response. Supplementary_file_2.xlsx: List of A. thaliana genes associated with seven abiotic stresses.

Author Contributions: Conceptualization, Sergey Lashin and Dmitry Afonnikov; Data curation, Zakhar Mustafin, Dmitry Konstantinov, Vladimir Zamyatin and Dmitry Afonnikov; Funding acquisition, Dmitry Afonnikov; Methodology, Zakhar Mustafin, Dmitry Konstantinov and Sergey Lashin; Software, Zakhar Mustafin and Vladimir Zamyatin; Supervision, Sergey Lashin and Dmitry Afonnikov; Validation, Zakhar Mustafin, Alexey Doroshkov and Sergey Lashin; Writing – original draft, Zakhar Mustafin, Dmitry Konstantinov, Vladimir Zamyatin, Alexey Doroshkov, Sergey Lashin and Dmitry Afonnikov; Writing – review & editing, Sergey Lashin and Dmitry Afonnikov.

Funding: This research was funded by Russian Scientific Foundation, grant number 18-14-00293.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tautz, D.; Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **2011**, *12*, 692–702, doi:10.1038/nrg3053.
2. Flagel, L.E.; Wendel, J.F. Gene duplication and evolutionary novelty in plants. *New Phytol.* **2009**, *183*, 557–564, doi:10.1111/j.1469-8137.2009.02923.x.
3. Oh, D.H.; Dassanayake, M.; Bohnert, H.J.; Cheeseman, J.M. Life at the extreme: Lessons from the genome. *Genome Biol.* **2012**, *13*, 1–9, doi:10.1186/gb-2012-13-3-241.
4. Domazet-Lošo, T.; Brajković, J.; Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **2007**, *23*, 533–539, doi:10.1016/j.tig.2007.08.014.
5. Domazet-Loso, T.; Tautz, D. An Ancient Evolutionary Origin of Genes Associated with Human Genetic Diseases. *Mol. Biol. Evol.* **2008**, *25*, 2699–2707, doi:10.1093/molbev/msn214.
6. Domazet-Lošo, T.; Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **2010**, *468*, 815–818, doi:10.1038/nature09632.
7. Quint, M.; Drost, H.-G.; Gabel, A.; Ullrich, K.K.; Bönn, M.; Grosse, I. A transcriptomic hourglass in plant embryogenesis. *Nature* **2012**, *490*, 98–101, doi:10.1038/nature11394.
8. Durrant, M.; Boyer, J.; Zhou, W.; Baldwin, I.T.; Xu, S. Evidence of an evolutionary hourglass pattern in herbivory-induced transcriptomic responses. *New Phytol.* **2017**, *215*, 1264–1273,

437 doi:10.1111/nph.14644.
438 9. Ruprecht, C.; Proost, S.; Hernandez-Coronado, M.; Ortiz-Ramirez, C.; Lang, D.; Rensing, S.A.;
439 Becker, J.D.; Vandepoele, K.; Mutwil, M. Phylogenomic analysis of gene co-expression networks
440 reveals the evolution of functional modules. *Plant J.* **2017**, *90*, 447–465, doi:10.1111/tpj.13502.
441 10. Cui, X.; Lv, Y.; Chen, M.; Nikoloski, Z.; Twell, D.; Zhang, D. Young Genes out of the Male: An
442 Insight from Evolutionary Age Analysis of the Pollen Transcriptome. *Mol. Plant* **2015**, *8*, 935–945,
443 doi:10.1016/j.molp.2014.12.008.
444 11. Mustafin, Z.S.; Lashin, S.A.; Matushkin, Y.G.; Gunbin, K.V.; Afonnikov, D.A. Orthoscape: a
445 cytoscape application for grouping and visualization KEGG based gene networks by taxonomy
446 and homology principles. *BMC Bioinformatics* **2017**, *18*, 1–9, doi:10.1186/s12859-016-1427-5.
447 12. Golldack, D.; Lüking, I.; Yang, O. Plant tolerance to drought and salinity: stress regulating
448 transcription factors and their functional significance in the cellular transcriptional network. *Plant*
449 *Cell Rep.* **2011**, *30*, 1383–1391, doi:10.1007/s00299-011-1068-0.
450 13. Deinlein, U.; Stephan, A.B.; Horie, T.; Luo, W.; Xu, G.; Schroeder, J.I. Plant salt-tolerance
451 mechanisms. *Trends Plant Sci.* **2014**, *19*, 371–379, doi:10.1016/j.tplants.2014.02.001.
452 14. Panchy, N.; Lehti-Shiu, M.D.; Shiu, S.-H. Evolution of gene duplication in plants. *Plant Physiol.*
453 **2016**, pp.00523.2016, doi:10.1104/pp.16.00523.
454 15. Lamesch, P.; Berardini, T.Z.; Li, D.; Swarbreck, D.; Wilks, C.; Sasidharan, R.; Muller, R.; Dreher, K.;
455 Alexander, D.L.; Garcia-Hernandez, M.; Karthikeyan, A.S.; Lee, C.H.; Nelson, W.D.; Ploetz, L.;
456 Singh, S.; Wensel, A.; Huala, E. The Arabidopsis Information Resource (TAIR): improved gene
457 annotation and new tools. *Nucleic Acids Res.* **2012**, *40*, D1202–D1210, doi:10.1093/nar/gkr1090.
458 16. Gene Ontology Consortium Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **2015**,
459 *43*, D1049–D1056, doi:10.1093/nar/gku1179.
460 17. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski,
461 B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction
462 networks. *Genome Res.* **2003**, *13*, 2498–504, doi:10.1101/gr.1239303.
463 18. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New perspectives on
464 genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2017**, *45*, D353–D361,
465 doi:10.1093/nar/gkw1092.
466 19. Zhang, Z.; Li, J.; Zhao, X.-Q.; Wang, J.; Wong, G.K.-S.; Yu, J. KaKs_Calculator: Calculating Ka and
467 Ks Through Model Selection and Model Averaging. *Genomics. Proteomics Bioinformatics* **2006**, *4*,
468 259–263, doi:10.1016/S1672-0229(07)60007-2.
469 20. van Zanten, M.; Snoek, L.B.; Proveniers, M.C.G.; Peeters, A.J.M. The many functions of ERECTA.
470 *Trends Plant Sci.* **2009**, *14*, 214–218, doi:10.1016/J.TPLANTS.2009.01.010.
471 21. Shen, H.; Zhong, X.; Zhao, F.; Wang, Y.; Yan, B.; Li, Q.; Chen, G.; Mao, B.; Wang, J.; Li, Y.; Xiao, G.;
472 He, Y.; Xiao, H.; Li, J.; He, Z. Overexpression of receptor-like kinase ERECTA improves
473 thermotolerance in rice and tomato. *Nat. Biotechnol.* **2015**, *33*, 996–1003, doi:10.1038/nbt.3321.
474 22. Kurepa, J.; Walker, J.M.; Smalle, J.; Gosink, M.M.; Davis, S.J.; Durham, T.L.; Sung, D.-Y.; Vierstra,
475 R.D. The small ubiquitin-like modifier (SUMO) protein modification system in Arabidopsis.
476 Accumulation of SUMO1 and -2 conjugates is increased by stress. *J. Biol. Chem.* **2003**, *278*, 6862–72,
477 doi:10.1074/jbc.M209694200.
478 23. Datta, M.; Kaushik, S.; Jyoti, A.; Mathur, N.; Kothari, S.L.; Jain, A. SIZ1-mediated SUMOylation
479 during phosphate homeostasis in plants: Looking beyond the tip of the iceberg. *Semin. Cell Dev.*
480 *Biol.* **2018**, *74*, 123–132, doi:10.1016/J.SEMCDB.2017.09.016.
481 24. Bilichak, A.; Yao, Y.; Titov, V.; Golubov, A.; Kovalchuk, I. Genome stability in the uvh6 mutant of
482 Arabidopsis thaliana. *Plant Cell Rep.* **2014**, *33*, 979–991, doi:10.1007/s00299-014-1580-0.
483 25. Liu, Z.; Hall, J.D.; Mount, D.W. Arabidopsis UVH3 gene is a homolog of the Saccharomyces
484 cerevisiae RAD2 and human XPG DNA repair genes. *Plant J.* **2001**, *26*, 329–338, doi:10.1046/j.1365-

313X.2001.01031.x.

26. Virdi, K.S.; Wamboldt, Y.; Kundariya, H.; Laurie, J.D.; Keren, I.; Kumar, K.R.S.; Block, A.; Basset, G.; Luebker, S.; Elowsky, C.; Day, P.M.; Roose, J.L.; Bricker, T.M.; Elthon, T.; Mackenzie, S.A. MSH1 Is a Plant Organellar DNA Binding and Thylakoid Protein under Precise Spatial Regulation to Alter Development. *Mol. Plant* **2016**, *9*, 245–260, doi:10.1016/J.MOLP.2015.10.011.
27. Vikas Shedge, Jaime Davila, Maria P. Arrieta-Montiel, Saleem Mohammed, S.A.M. Extensive Rearrangement of the Arabidopsis Mitochondrial Genome Elicits Cellular Conditions for Thermotolerance. *PLANT Physiol.* **2010**, *152(4)*, 1960–1970, doi:https://doi.org/10.1104/pp.109.152827.
28. Zhao, S.; Liu, L.; IUCr Expression and crystallographic studies of the *Arabidopsis thaliana* GDP- D -mannose pyrophosphorylase VTC1. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2016**, *72*, 795–798, doi:10.1107/S2053230X16013406.
29. William A. Laing, Marcela Martínez-Sánchez, Michele A. Wright, Sean M. Bulley, Di Brewster, Andrew P. Dare, Maysoon Rassam, Daisy Wang, Roy Storey, Richard C. Macknight, R.P.H. An Upstream Open Reading Frame Is Essential for Feedback Regulation of Ascorbate Biosynthesis in Arabidopsis. *Plant Cell* **2015**, *tpc-114*, doi:https://doi.org/10.1105/tpc.114.133777.
30. Zhizhong Gong, Chun-Hai Dong, Hojoung Lee, Jianhua Zhu, Liming Xiong, Deming Gong, Becky Stevenson, J.-K.Z. A DEAD Box RNA Helicase Is Essential for mRNA Export and Important for Development and Stress Responses in Arabidopsis. *Plant Cell* **2005**, *17(1)*, 256–267, doi:https://doi.org/10.1105/tpc.104.027557.
31. Wu, S.-J.; Wang, L.-C.; Yeh, C.-H.; Lu, C.-A.; Wu, S.-J. Isolation and characterization of the Arabidopsis heat-intolerant 2 (hit2) mutant reveal the essential role of the nuclear export receptor EXPORTIN1A (XPO1A) in plant heat tolerance. *New Phytol.* **2010**, *186*, 833–842, doi:10.1111/j.1469-8137.2010.03225.x.
32. Cramer, G.R.; Urano, K.; Delrot, S.; Pezzotti, M.; Shinozaki, K. Effects of abiotic stress on plants: a systems biology perspective. *BMC Plant Biol.* **2011**, *11*, 163, doi:10.1186/1471-2229-11-163.
33. Wahid, A.; Gelani, S.; Ashraf, M.; Foolad, M.R. Heat tolerance in plants: An overview. *Environ. Exp. Bot.* **2007**, *61*, 199–223, doi:10.1016/J.ENVEXPBOT.2007.05.011.
34. Shinozaki, K.; Yamaguchi-Shinozaki, K. Gene networks involved in drought stress response and tolerance. *J. Exp. Bot.* **2006**, *58*, 221–227, doi:10.1093/jxb/erl164.
35. Ohama, N.; Sato, H.; Shinozaki, K.; Yamaguchi-Shinozaki, K. Transcriptional Regulatory Network of Plant Heat Stress Response. *Trends Plant Sci.* **2017**, *22*, 53–65, doi:10.1016/j.tplants.2016.08.015.
36. Parida, A.K.; Das, A.B. Salt tolerance and salinity effects on plants: a review. *Ecotoxicol. Environ. Saf.* **2005**, *60*, 324–349, doi:10.1016/J.ECOENV.2004.06.010.
37. Hazman, M.; Hause, B.; Eiche, E.; Riemann, M.; Nick, P. Different forms of osmotic stress evoke qualitatively different responses in rice. *J. Plant Physiol.* **2016**, *202*, 45–56, doi:10.1016/J.JPLPH.2016.05.027.
38. Wang, W.; Vinocur, B.; Shoseyov, O.; Altman, A. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* **2004**, *9*, 244–252, doi:10.1016/J.TPLANTS.2004.03.006.
39. Gupta, R.; Golding, G.B. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. *J. Mol. Evol.* **1993**, *37*, 573–582, doi:10.1007/BF00182743.
40. Zhang, Y.; Wang, L. The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evol. Biol.* **2005**, *5*, 1, doi:10.1186/1471-2148-5-1.
41. Zhu, J.-K. Abiotic Stress Signaling and Responses in Plants. *Cell* **2016**, *167*, 313–324, doi:10.1016/J.CELL.2016.08.029.
42. Bleecker, S.-H.S. and A.B. Receptor-like kinases from Arabidopsis form a monophyletic gene

family related to animal receptor kinases. *Proc. Natl. Acad. Sci.* **2001**, *98(19)*, 10763–10768, doi:https://doi.org/10.1073/pnas.181141598.

43. Donoghue, M.T.A.; Keshavaiah, C.; Swamidatta, S.H.; Spillane, C. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. **2011**, 1–23.
44. Tomescu, A.M.F. Megaphylls, microphylls and the evolution of leaf development. *Trends Plant Sci.* **2009**, *14*, 5–12, doi:10.1016/J.TPLANTS.2008.10.008.
45. Harrison, C.J. Development and genetics in the evolution of land plant body plans. *Phil. Trans. R. Soc. B* **2016**, *372(1713)*, doi:DOI: 10.1098/rstb.2015.0490.
46. C. Jill Harrison, J.L.M. The origin and early evolution of vascular plant shoots and leaves. *Phil. Trans. R. Soc. B* **2017**, *373(1739)*, doi:10.1098/rstb.2016.0496.
47. Dkhar, J.; Pareek, A. What determines a leaf's shape? *Evodevo* **2014**, *5*, 47, doi:10.1186/2041-9139-5-47.
48. Tsukaya, H. Leaf shape: genetic controls and environmental factors. *Int. J. Dev. Biol.* **2005**, *49*, 547–55, doi:10.1387/ijdb.041921ht.