Zakayo Kazibwe

Gene Family Evolution of RNA Helicases in Tobbaco, Rice and Arabidopsis

Introduction and Project Motivation

RNA helicases (here after referred to as RHs) are proteins involved in various aspects of RNA metabolism in both plants and animals. They are responsible for the unwinding of double stranded RNA molecules in an energy dependent manner, mainly through hydrolysis of NTP's (Umate *et al*. 2010). RHs are involved in functions such as ribosome biogenesis, mRNA splicing, and editing and transport of RNA in and out of the cell organelles. The latter function is the interest of my research. Various reports have shown the involvement of RHs in the growth, development, and stress response in plants. To date, RHs have been classified into three subfamilies, these are DEAD-box, DEAH-box and DExD/H-box helicases. More than 100 Arabidopsis thaliana RHSs (here after referred to as AtRHs) have been reported in literature, and it is found that their number doubles that found in other species like yeast and mammals (Umate et al. 2010). Phylogenetic Analysis of the housekeeping At*RHs* suggested a scenario for the evolution of duplicated genes, leading to both highly and poorly transcribed genes in the same terminal branch of the phylogenetic tree (Xu et al. 2013). It appears that the general evolutionary drive of the *AtRHs* family, after duplication of a highly transcribed ancestral *AtRHs,* was towards an iteration of the transcriptional activity of the divergent duplicates through successive events of suppression of the TATA-box and/or the 5 ′UTR intron (Tuteja 2010). That said more genomic data containing RHs have been generated, and it is worth looking into this data to reconstruct phylogeny and better understand the evolutionary changes in this gene family. This work is aimed at identifying the relevant homologs of DEAD-box/DEAH-box/DExD subfamilies in Arabidopsis Rice and Tobacco. We intend to replicate the phylogenetic analyses of the DEAD-box/DEAH-box which have been done previously, using neighbor joining approach. We propose to use other methods like maximum likelihood to reconstruct these trees and update them, at least in part using the new available data. Lastly, we plan to predict the evolutionary rates of DEAD-box/DEAH-box/DExD/H-box proteins in Arabidopsis Rice and Tobacco.

Methods

Sequences and Multiple Sequence Alignment

Relevant proteins ID's were collected from literature, most of which are published in Xu *et al*. (2013) and a few most recent from elsewhere. For Arabidopsis, protein and DNA sequences were retrieved from TAIR (https://www.arabidopsis.org/), while protein sequences for Rice and Tobacco were obtained from Uniport (https://www.uniprot.org). The protein coding DNA Sequences for Tobacco and Rice were extracted from (https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi). In Maximum Likelihood and Distance **A**nalysis, a total of 30 RHs proteins sequences, comprising of 15 from Arabidopsis, 7 from Tobacco and 8 from Rice which are either DEAD-box/DEAH or DExD/H-box genes were used. The retrieved sequences were aligned using MAFFT and they were saved in fasta, nexus or phylip formats depending on the methods of analysis. A different set of sequences was used for hypothesis Testing using codeml as described below.

Phylogenetic analyses

Maximum likelihood

The phylogenetic trees were reconstructed using the maximum likelihood method implemented in the PhyML program (v3.1/3.0 aLRT) and RAxML7 v.8.0.22 with the GTRGAMMA model of nucleotide substitutions. A 50% majority rule consensus tree was constructed using the bayesian inference method implemented in MrBayes program (v3.2.6).

Distance analyses (PHYLIP)

Pairwise distances among the taxa as inputs for phylogenetic reconstruction were calculated using Neighbor Joining approach. Distance matrices were estimated using models of amino acid substitution available in protdist for DNA sequences. For amino acids, these models are Dayhoff PAM matrix, the JTT (Jones-Taylor-Thornton) model, the PMB (Probability Matrix from Blocks) model, Kimura's distance, and Categories distances). For each of these models, NJ trees were constructed using an appropriate outgroup. For each analysis, bootstrapping with 1,000 replicates was also performed and the trees were viewed or printed in Figtree or MEGA-X software.

Hypothesis Testing and Detecting Selection with codeml

To determine the rates of evolution in the RHS genes and ascertain the underlying selection pressure, I used codeml, a PAML (Phylogenetic Analysis by Maximum Likelihood) package, by setting seqtype to 1, using DNA Sequences. For the meaningful interpretation of the results, I pulled 15 genes from each species, and I used 5 DNA sequences from each subfamily. Therefore, this analysis, used a different dataset compared to the previous analyses and ML analysis was done by codon substitution models as described in (Goldman and Yang 1994). To this end, each species and subfamily was analyzed independently and compared the omega (ω) ratios between the Arabidopsis, Rice and Tobacco genes and between the DEAD-box/DEAH/ DExD/H-box genes. The Omega (ω) ratio (ω = dN/dS) is a measure of natural selection acting on the protein and it is very informative in understanding natural selection acting on genomes of species. The dS represent synonymous substitutions and dN is nonsynonymous substitutions in a gene. I propose that the rates of change in the RHs family proteins differs between species and it is more pronounced within than between subfamilies.

Results and Discussion

RNA helicases form Two Distinct Paralogous Groups

From the phylogenetic analysis of protein coding sequences of RHs genes, it can be observed that all the 30 RHs genes are homologous to each other and are clustered into several subgroups, forming distinct clades, with Tobacco genes forming their separate clade (Fig.1a, and 1b). Our analysis by MrBayes program showed that RHs family proteins could be further classified into more than 10 subgroups (Figures 1 a &b). In previous reports, two Arabidopsis DEAD-box RNA helicases, RHS10ARAB and RHS13ARAB, were shown to be involved in responses to multiple, abiotic stresses (Gong et al. 2002). Their investigations indicated that DEAD-box RNA helicases may play an important role in building resistance to abiotic stress during plant growth and development. Figure 3 shows that, RHS17ORYSJ and Tobacco Q75WU9TOB have high homology to RHS10ARAB. In addition, the phylogenetic tree analyses showed that all the Tobacco RHS are homologous, with very limited differences, as evidenced by the branch length and bootstrap values and therefore all of them could be involved in stress responses (figure 1 a).

When the protein groups from the three species were combined and their amino acid sequences analyzed by neighbor joining, the resulting tree shows that the proteins are paralogs and they separate into three major groups which later duplicated independently (Fig 2). This may imply that they arose as a result of gene duplication in the MRCA (Most Recent Common ancestor). However, within each protein subgroup, all proteins are homologous to each other. The neighbor joining trees established by Boudet *et al.* (2001) indicated that a first event of duplication generated the ancestors of two groups of more recently duplicated RNA helicase genes. This was confirmed by Mingam et al. (2004) who reported that the two subgroups exhibit a completely different intronic structure in the translated regions. In my analysis, there is another distinct group of helicases that has gone through little or no duplication, composed of Arabidopsis and Rice RHS10 and it is more evident in NJ tree (figure 2).

RNA Helicase Genes are Under Different Selection Pressures

Our current analysis shows that, the omega (ω = dN/dS) between Rice, Arabidopsis and Tobacco, as well as between DEAD-box, DEAH-box and DExD/H-box genes is different (Table 1). Hypothesis testing using codeml revealed that DEAH-box genes are under positive selection pressure (ω <1) at least in Rice and Arabidopsis whereas DExD/H-box and DEAH-box genes are under negative selection (ω >1, Table 1). Negative selection, also called purifying selection, means that selection is purging changes that cause deleterious impacts on the fitness of the host giving them less chances of survival. On the other hand, in positive selection, variants increase in frequency until they are fixed in the relevant population (https://www.researchgate.net/). The reason behind negative selection pressure among DExD/H-box/DEAD-box helicases is unknown. Possible explanation may stem from the fact that these genes are too long, and they are more prone to mutations (Pentzold et al. 2018, Lynch et al. 2016). For example, the average length of the DExD/H-box RHS is two times the length of DEAH-box genes. On the other hand, however, mutation in any of the SKI complex member, a subgroup of DExD/H-box results into similar phenotypes and constitutive autophagy in Arabidopsis thaliana, meaning that they are highly duplicated and essential for plant metabolism (Zhao and Kunst, 2016).

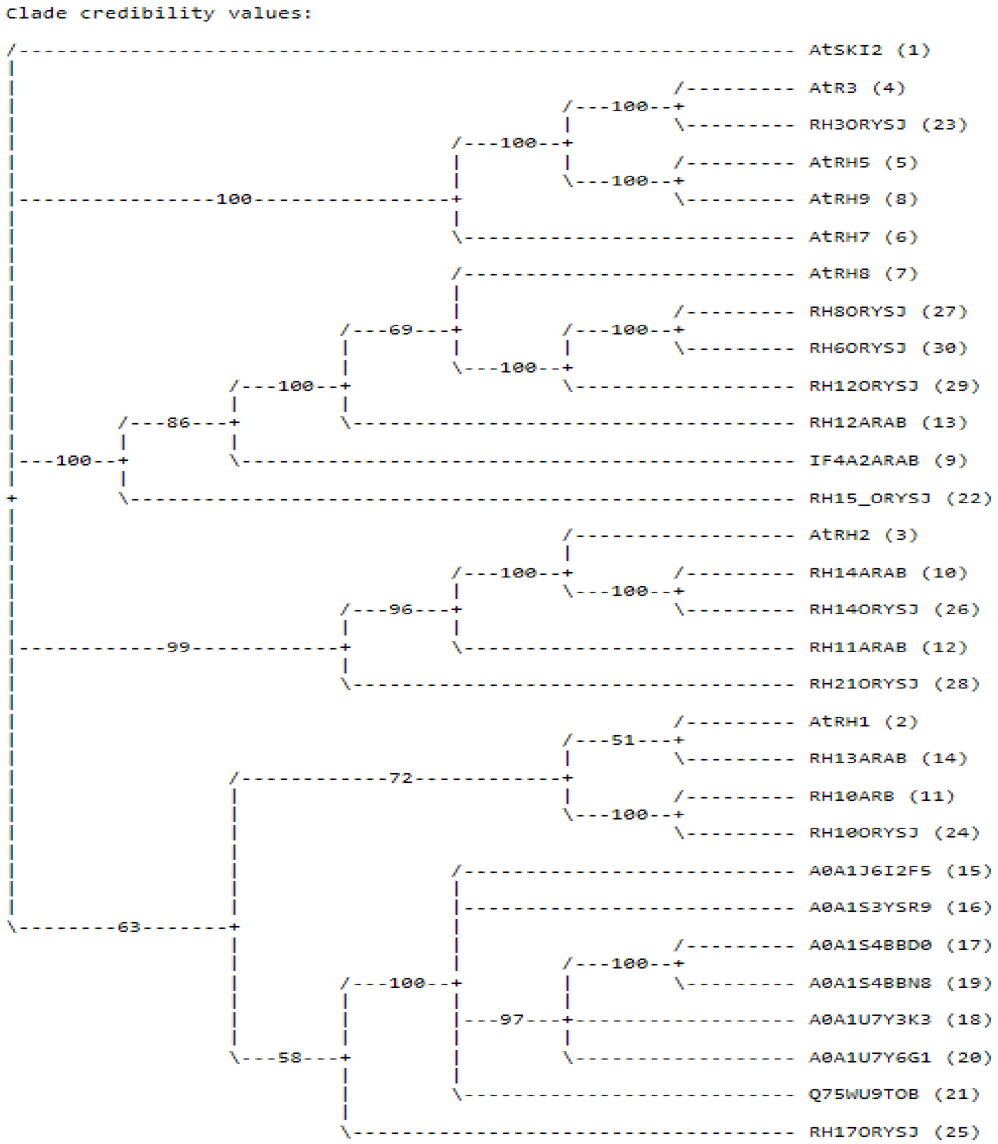
It should be noted that purifying selection is more effective in large populations than in small populations yielding smaller dN/dS ratio (Gillespie.1994). In Arabidopsis, the DExD/H-box evolved with more than 70 protein coding genes compared to DEAD-box and DEAH-box genes, 50 and 40 respectively (Xu *et al*. 2013). This is in line with the smaller dN/dS ratio observed for DExD/H-box genes . Therefore, the DExD/H-box genes could be very essential for plant growth and development.

Conclusion

RNA helicases are found in various organisms, ranging from prokaryotes to mammals, and have become a focus of interest in recent years due to their participation in diverse cellular processes Linder and Owttrim (2009). The phylogenetic tree analyses identified the relevant homologs of DEAD-box/ DExD/H-box/DEAH-box RNA helicase proteins in each of the three plant species. In addition, we have produced different tree topologies by likelihood methods, as opposed to those available in literature produced by only NJ methods. We also showed that the RHs genes are under different selection pressure, with DEAD-box/DExD/H-box under negative selection, while DEAH-box RHs are under positive selection. Even though the list of RHs considered in this paper is not exhaustive, it gives a bird's eye view on the evolutionary patterns of one of the most important plant protein families.

RNA Helicase Subfamily	**Arabidopsis**			**Rice**			**Tobacco**		
	dN	dS	**dN/dS**	dN	dS	**dN/dS**	dN	dS	**dN/dS**
DEAD-box	0.65	0.761	**0.854**	0.591	0.81	**0.73**	0.1196	0.52	**0.23**
DEAH-box	0.788	0.63	**1.25**	0.397	0.21	**1.89**	0.427	0.44	**0.97**
DExD/H-box	0.351	0.45	**0.78**	0.125	0.89	**0.14**	0.401	0.59	**0.68**

Table 1: Omega Values of the RNA Helicases in different Species. *Synonymous (dS) and nonsynonymous (dN) substitution rates were obtained by codelm implemented in PAML package. Omega (***dN/dS***) values were estimated for protein-coding DNA sequences of Arabidopsis, Rice and Tobacco and for the three subfamilies independently.*

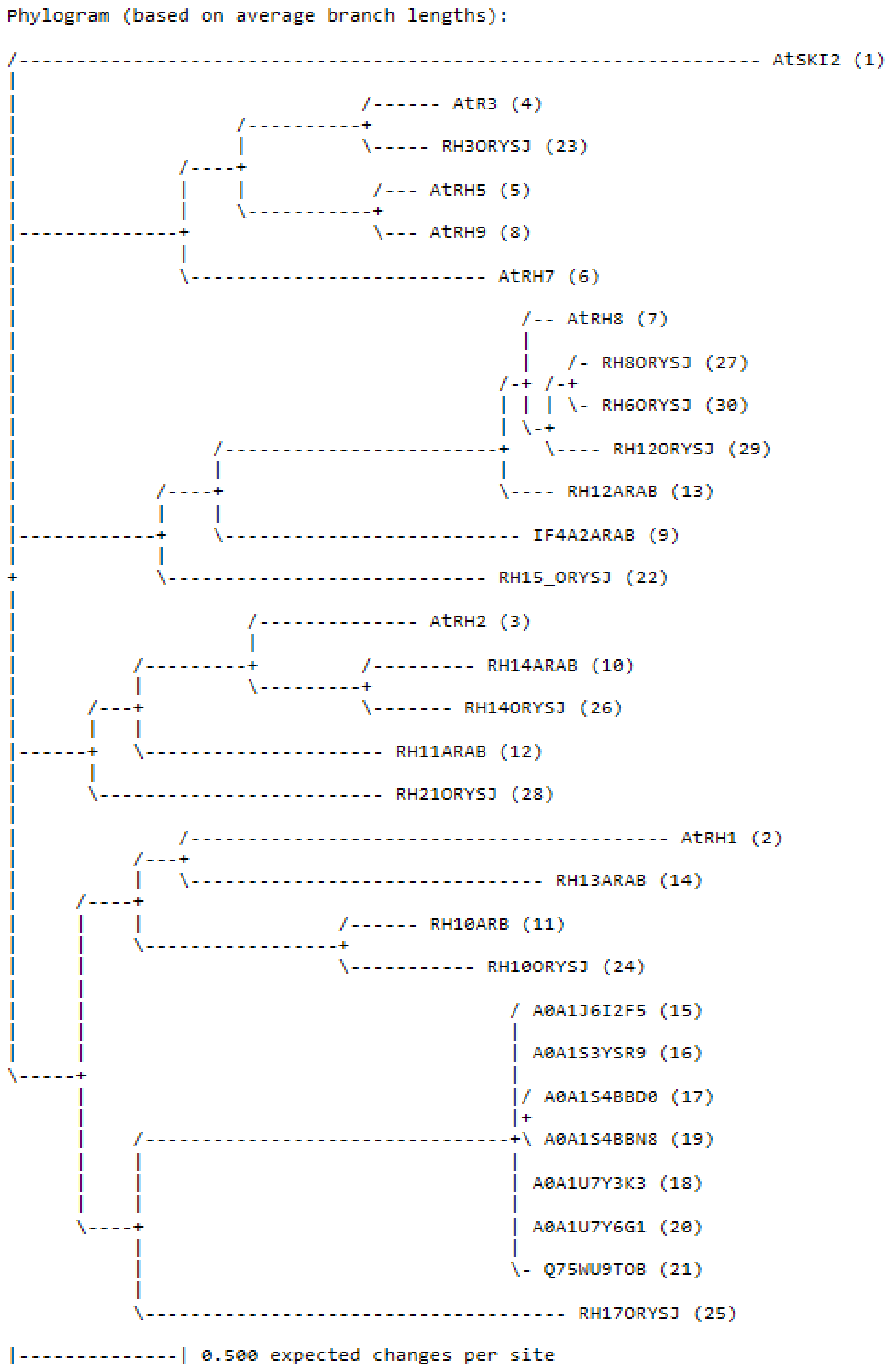


```
Clade credibility values:

/-------------------------------------------------------------- AtSKI2 (1)
|
|                                                     /-------- AtR3 (4)
|                                            /---100--+
|                                            |        \-------- RH3ORYSJ (23)
|                                   /---100--+
|                                   |        |        /-------- AtRH5 (5)
|                                   |        \---100--+
|----------------100----------------+                 \-------- AtRH9 (8)
|                                   |
|                                   \-------------------------- AtRH7 (6)
|
|                                   /-------------------------- AtRH8 (7)
|                                   |
|                                   |                 /-------- RH8ORYSJ (27)
|                          /---69---+        /---100--+
|                          |        |        |        \-------- RH6ORYSJ (30)
|                          |        \---100--+
|                 /---100--+                 \----------------- RH12ORYSJ (29)
|                 |        |
|        /---86---+        \----------------------------------- RH12ARAB (13)
|        |        |
|---100--+        \-------------------------------------------- IF4A2ARAB (9)
|        |
+        \----------------------------------------------------- RH15_ORYSJ (22)
|
|                                            /----------------- AtRH2 (3)
|                                            |
|                                   /---100--+        /-------- RH14ARAB (10)
|                                   |        \---100--+
|                          /---96---+                 \-------- RH14ORYSJ (26)
|                          |        |
|------------99------------+        \-------------------------- RH11ARAB (12)
|                          |
|                          \----------------------------------- RH21ORYSJ (28)
|
|                                                     /-------- AtRH1 (2)
|                                            /---51---+
|                                            |        \-------- RH13ARAB (14)
|                 /------------72------------+
|                 |                          |        /-------- RH10ARB (11)
|                 |                          \---100--+
|                 |                                   \-------- RH10ORYSJ (24)
|                 |
|                 |                 /-------------------------- A0A1J6I2F5 (15)
|                 |                 |
|                 |                 |-------------------------- A0A1S3YSR9 (16)
\--------63-------+                 |
                  |                 |                 /-------- A0A1S4BBD0 (17)
                  |                 |        /---100--+
                  |        /---100--+        |        \-------- A0A1S4BBN8 (19)
                  |        |        |        |
                  |        |        |---97---+----------------- A0A1U7Y3K3 (18)
                  |        |        |        |
                  \---58---+        |        \----------------- A0A1U7Y6G1 (20)
                           |        |
                           |        \-------------------------- Q75WU9TOB (21)
                           |
                           \----------------------------------- RH17ORYSJ (25)
```

Figure 1 (a)and 1 (b). The phylogenetic trees were reconstructed using the bayesian inference method implemented in MrBayes program (v3.2.6). The number of substitution types was fixed to 1. The blosum62 model was used for substitution, while rates variation across sites was fixed to "invgamma". Four Markov Chain Monte Carlo (MCMC) chains were run for 10000 generations, sampling every 10 generations, with the first 250 sampled trees discarded as "burn-in". Finally, a 50% majority rule consensus tree was constructed. Trees were visualized and printed in Figtree program. Numbers on the branches in figure 1 (a) refer to the bootstrap values after 1000 replicates.

Zakayo Kazibwe



```
Phylogram (based on average branch lengths):

/----------------------------------------------------------------- AtSKI2 (1)
|
|                              /------ AtR3 (4)
|                   /----------+
|                   |          \----- RH3ORYSJ (23)
|              /----+
|              |    |           /--- AtRH5 (5)
|              |    \-----------+
|--------------+                \--- AtRH9 (8)
|              |
|              \-------------------------- AtRH7 (6)
|
|                                            /-- AtRH8 (7)
|                                            |
|                                            |   /- RH8ORYSJ (27)
|                                          /-+ /-+
|                                          | | | \- RH6ORYSJ (30)
|                                          | \-+
|                 /------------------------+   \---- RH12ORYSJ (29)
|                 |                        |
|            /----+                        \---- RH12ARAB (13)
|            |    |
|------------+    \-------------------------- IF4A2ARAB (9)
|            |
+            \---------------------------- RH15_ORYSJ (22)
|
|                    /-------------- AtRH2 (3)
|                    |
|          /---------+         /--------- RH14ARAB (10)
|          |         \---------+
|      /---+                   \------- RH14ORYSJ (26)
|      |   |
|------+   \--------------------- RH11ARAB (12)
|      |
|      \------------------------- RH21ORYSJ (28)
|
|              /------------------------------------------ AtRH1 (2)
|          /---+
|          |   \------------------------------- RH13ARAB (14)
|     /----+
|     |    |                 /------ RH10ARB (11)
|     |    \-----------------+
|     |                      \----------- RH10ORYSJ (24)
|     |
|     |                                     / A0A1J6I2F5 (15)
|     |                                     |
|     |                                     | A0A1S3YSR9 (16)
\-----+                                     |
      |                                     |/ A0A1S4BBD0 (17)
      |                                     |+
      |    /--------------------------------+\ A0A1S4BBN8 (19)
      |    |                                |
      |    |                                | A0A1U7Y3K3 (18)
      |    |                                |
      \----+                                | A0A1U7Y6G1 (20)
           |                                |
           |                                \- Q75WU9TOB (21)
           |
           \------------------------------------- RH17ORYSJ (25)

|--------------| 0.500 expected changes per site
```

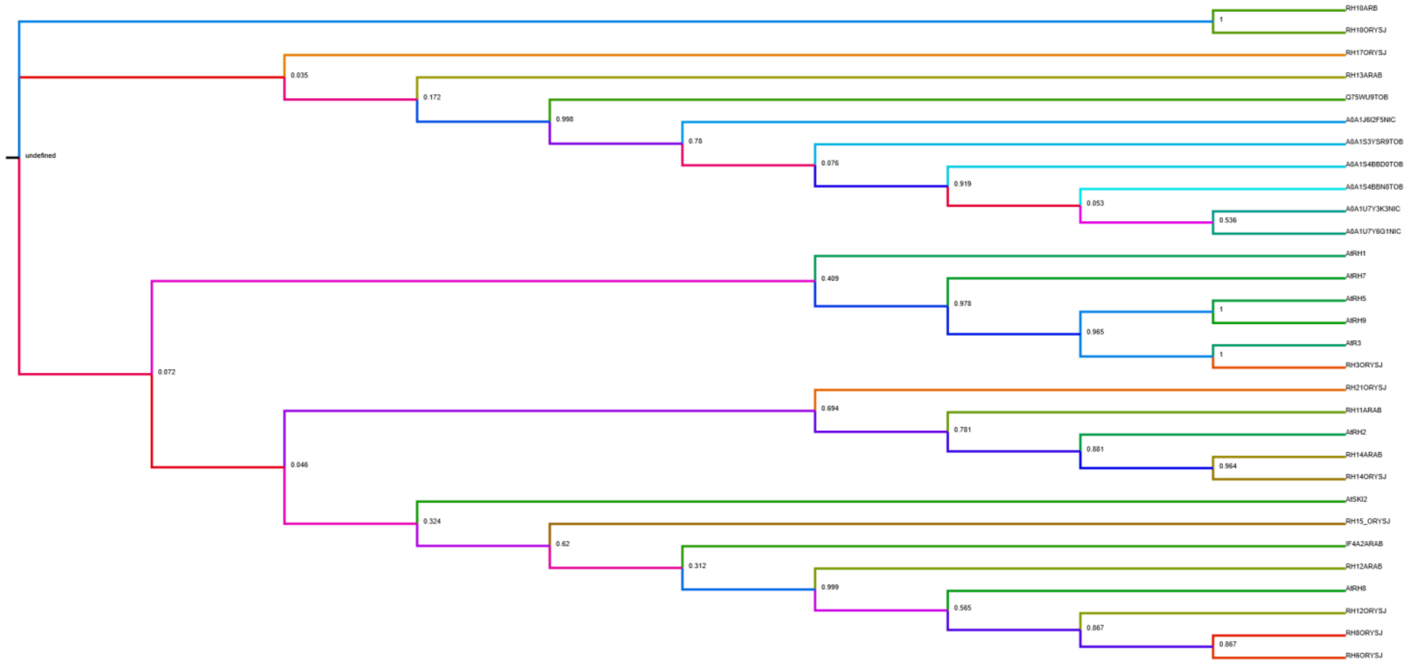


Figure 2. Phylogenetic analysis of RHs proteins using Distance analysis. *The phylogenetic tree was reconstructed using the neighbor joining method implemented in the BioNJ program. Distances were calculated using ProtDist. The DAY subsitution model was selected for the analysis. The tree was visualized and edited in Mega-x software. Numbers on the branches refer to the support values. The tree was visualized by Figtree and rooted at the MRCA (Most Recent Common Ancestor) of the three protein groups.*

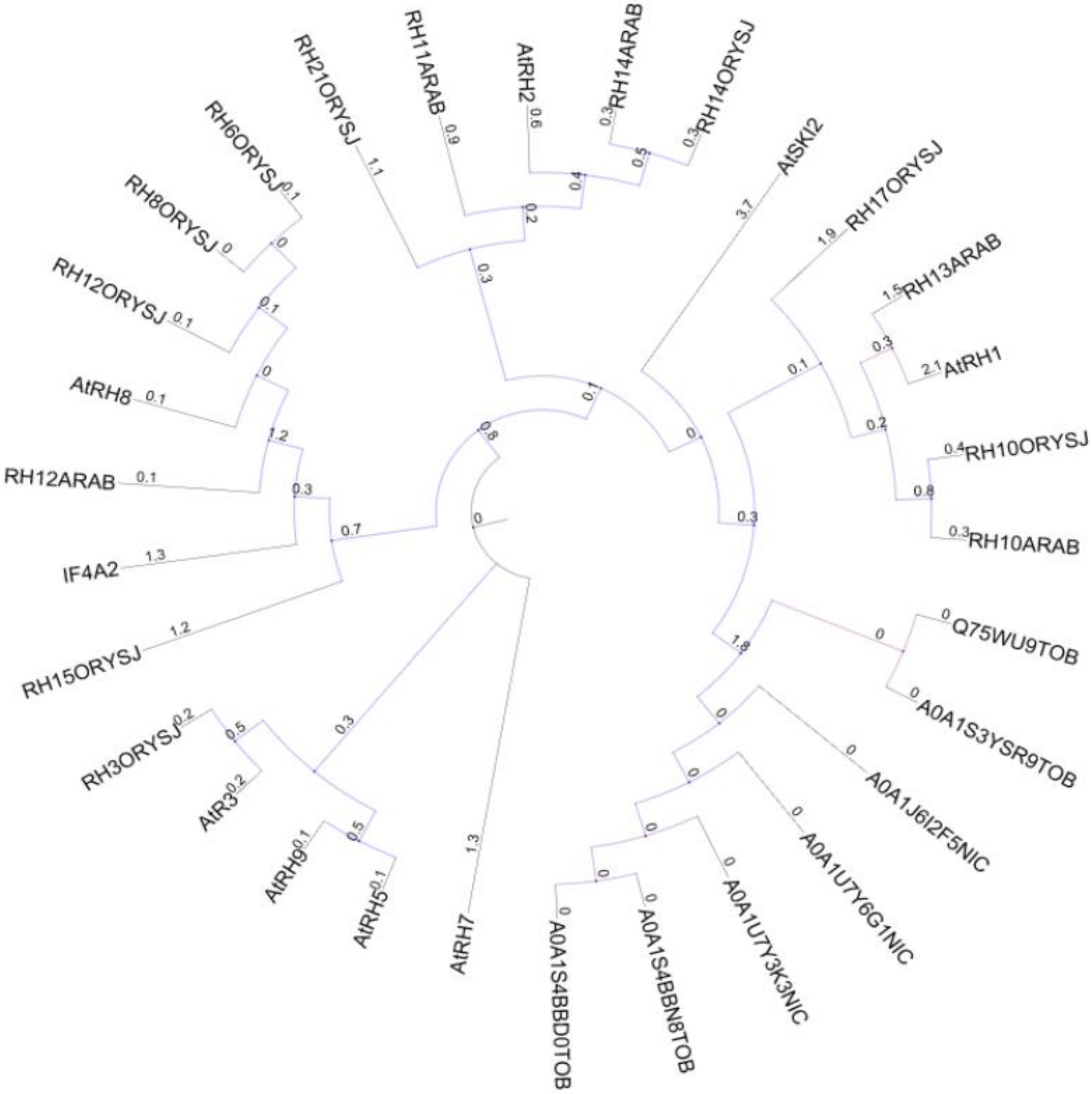


Figure 4. ***Phylogenetic analysis of RHs proteins using Distance analysis.*** *The phylogenetic tree was reconstructed using the maximum likelihood method implemented in the PhyML program (v3.1/3.0 aLRT). RaxML produced the same tree. The JTT substitution model was selected assuming an estimated proportion of invariant sites (of 0.024) and 4 gamma-distributed rate categories to account for rate heterogeneity across sites. The gamma shape parameter was estimated directly from the data (gamma=1.277). Reliability for internal branch was assessed using the aLRT test (SH-Like). The tree was visualized in ITOL server (https://itol.embl.de/tree). The values on branches indicate average branch length.*

References

1. Xu R, Zhang S, Huang J, Zheng C (2013) Genome-Wide Comparative In Silico Analysis of the RNA Helicase Gene Family in Zea mays and Glycine max: A Comparison with Arabidopsis and Oryza sativa. PLoS ONE 8(11): e78982.
2. Umate P, Tuteja R, Tueja N (2010) Genome-wide analysis of helicase gene family from rice and Arabidopsis: a comparison with yeast and human. Plant Mol Biol 73 (4–5): 449–465
3. **N Goldman Z Yang.** A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, Volume 11, Issue 5, 1994, 725–736, **https://doi.org/10.1093/oxfordjournals.molbev.a040153**
4. Gong ZZ, Lee H, Xiong LM, Jagendorf A, Stevenson B, et al. (2002) RNA helicase-like protein as an early regulator of transcription factors for plant chilling and freezing tolerance. Proc Natl Acad Sci USA 99 (17): 11507–11512.
5. Linder P, Owttrim GW (2009) Plant RNA helicases: linking aberrant and silencing RNA. Trends Plant Sci 14 (6): 344–352.
6. Pentzold C, Shah SA, Hansen NR, Le Tallec B, Seguin-Orlando A, Debatisse M, Lisby M, Oestergaard VH. FANCD2 binding identifies conserved fragile sites at large transcribed genes in avian cells. Nucleic Acids Res. 2018 Feb 16;46(3):1280-1294.
7. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016 Oct 14;17(11):704-714.
8. http://envgen.nox.ac.uk/bioinformatics/docs/codeml.html
9. http://abacus.gene.ucl.ac.uk/software/pamlFAQs.pdf
10. **https://www.researchgate.net/**
11. *https://itol.embl.de/tree*