Comprehensive Classification of the RNaseH-like domain-containing Proteins in Plants

Shuai Li
Tsinghua University School of Life Sciences https://orcid.org/0000-0002-2913-1229
Kunpeng Liu
Tsinghua University School of Life Sciences
Qianwen Sun (sunqianwen@mail.tsinghua.edu.cn)
https://orcid.org/0000-0003-0111-5400

Research article

Keywords: R-loop, RNaseH-like protein, transposon, phylogenetic tree

Posted Date: October 1st, 2019

DOI: https://doi.org/10.21203/rs.2.15377/v1

Abstract

Background

R-loop is a nucleic acid structure containing an RNA-DNA hybrid and a displaced single-stranded DNA. Recently, accumulated evidences showed that R-loops occur frequently in various organisms' genomes and have crucial physiological functions, including replication, transcription and DNA repair. RNaseH-like superfamily (RNHLS) domain-containing proteins, such as RNase H enzymes, RNase H1 and RNase H2, are important in restricting R-loop levels. However, little is known about other RNHLS proteins and their relationship and function on modulating R-loops, especially in plants.

Results

In this study we characterized 6193 RNHLS proteins from 13 representative plant species, and clustered them into 27 clusters, among which reverse transcriptases and exonucleases are the two largest groups. Moreover, we found there are 691 RNHLS proteins in Arabidopsis, and the catalytic alpha helix and beta sheet domain are conserved among them. Interestingly, each of the Arabidopsis RNHLS proteins is composed of RNHLS domains and other distinctive protein domains. we further found that RNHLS genes are highly expressed in diverse meristems and metabolic tissues. Various RNHLS genes expressed in different tissues grouped together indicate that they work both isolatedly and cooperatively.

Conclusions

In summary, we systematically analyzed RNHL proteins in plant and found that there are mainly 27 subclusters of them. Most of these proteins are implicated in DNA replication, RNA transcription and nucleic acid degradation. We characterized and classified the RNHLS proteins in plant, which affords new insights on investigation of novel regulatory mechanisms and functions in R-loop biology.

Introduction

R-loop is a nucleotide structure consists of RNA-DNA hybrid with the displaced single-stranded DNA [1]. Formation of R-loops *in vivo* was found in process of DNA replication, retrovirus or single stranded RNA virus infection or retrotransposon life cycle [2]. For a long time, R-loops have been considered as transcription by-products since the nascent transcripts could bind back to their DNA templates [3]. However, recent studies in different organisms revealed that R-loops are widespread on genomes in yeast, plants and mammals [4–10]. These evidences together revealed that R-loops should be accurately modulated in different cellular processes.

Ribonucleases H (RNase H), a type of endonuclease, specifically degrades RNA moiety of RNA-DNA hybrid therefore can destroy and remove R-loops [11–13]. RNA molecule might be the origin of life and RNase H mainly disrupts the direct interaction of RNA and DNA [14]. Therefore, it is reasonable that RNase H is one of the most ancient protein and that many enzymes involved in RNA and DNA related

processes employs RNase H domain during evolution [15, 16]. The RNase H-like superfamily (RNHLS) proteins are groups of proteins with the similar 3D structure of RNase H domain, including the type 1 and type 2 RNase Hs, retroviral integrases, DNA transposases, Holliday junction resolvases, Piwi/Argonaute nucleases, exonucleases and key spliceosome component Pre-mRNA-processing-splicing factor 8 (Prp8) and other RNA nucleases [17–23]. Hence, RNHLS proteins play important roles in DNA replication and repair, RNA transcription and interference, homologous and non-homologous recombination, transposon and retrovirus transposition, and single-stranded RNA virus infection [15]. Previous study using bioinformatics approaches to search database for all RNHLS proteins revealed that the RNHLS proteins could be grouped into 152 families and divided mainly into exonucleases and endonucleases [15]. Plants are sessile and need to overcome different biotic or abiotic stress over their life cycle. Recently, it is been reported that the regulation of R-loop levels could impact antisense transcription at *FLC* locus in Arabidopsis[24]. Moreover, the misexpression of topoisomerase 1, a key R-loop regulator, could compromise the expression of auxin-related genes in rice[25]. In addition, there are three RNase H1s with different organelle localization and the chloroplast-localized RNase H1 is critical for maintaining chloroplast genome integrity in Arabidopsis[26]. Furthermore, R-loops can perform their function as chromatin features and read by proteins such as ALBA proteins in Arabidopsis and GADD45A proteins in mammals[27, 28]. Therefore, expression of *RNHLS* genes should be fine-tuned to maintain development and overcome these stresses. However, little is known about RNHLS proteins and their relationship in plants especially on regulation of R-loops.

In this study, to systemically analyze the relationship and diversity of RNHLS proteins in plant, we using bioinformatics tools to search the database and collect the RNHLS proteins of 13 representative plant species, including *Solanum lycopersicum (*tomato), *Solanum tuberosum (*potato), *Populus trichocarpa* (western balsam poplar), *Glycine max* (soybean), *Citrus sinensis* (sweet orange), *Arabidopsis thaliana* (arabidopsis), *Oryza sativa* (rice), *Physcomitrella patens* (moss), *Marchantia polymorpha* (liverwort), *Klebsormidium flaccidum (*filamentous charophyte green algae), *Chlamydomonas eustigma* (unicellular flagellate green algae), *Cyanidioschyzon merolae* (red alga),and *Chondrus crispus* (carrageen Irish moss). We characterized 6193 RNHLS proteins in these representative plants. They are clustered into 27 clusters, among them reverse transcriptases and exonucleases are two largest groups. In Arabidopsis, there are 691 RNHLS proteins, consisting of RNHLS domain with various types of protein domains in these proteins, while the catalytic alpha helix and beta sheet domain is conserved among them. In addition, *RNHLS* genes are actively expressed in meristems and other metabolic tissues. Various *RNHLS* genes expressed in different tissues are grouped together indicating they work both individually and cooperatively. In summary, we characterized and classified the RNHLS proteins in plants, which will extend our understanding of R-loop regulation *in vivo.*

Materials And Methods

Co-occurrence of RNHLS domain with other protein domains

The co-occurrence matrix of RNHLS domain (3.30.420.10) with other protein domains was downloaded from Gene3D [29]. The heatmap was processed by online software Morpheus (https://software.broadinstitute.org/morpheus/) using the default settings.

Protein sequence search

To identify the RNaseH-like superfamily proteins in Arabidopsis comprehensively, we used the proteins sequences from RNHLS (IPR012337) of InterPro database (http://www.ebi.ac.uk/interpro/) filtered with different species. In addition, we performed sequence BLAST with AtRNH1A (At3g01410) as a query in the Phytozome and NCBI database [30, 31]. We carried out the analysis after combining the sequence from the three methods and removed the duplicates.

Protein sequence cluster

The identified proteins were first analyzed with CD-HIT and the threshold was set at 60% with word size of 3 [32]. The clustered proteins were used the BLAST to compute the similarity of RNHLS proteins with an average pairwise E-value at least of 1×10^{-18} and redundancy proteins were removed [31]. The network was visualized using Cytoscape3.7.0 and laid out using the organic layout algorithm and annotated manually [33].

Protein sequence alignment

To analysis the protein sequences systemically, we used Clustal Omega online service with the default parameters except for all the iterations set as five [34]. The alignment results were visualized with UGENE and MEGA7 [35]. The structure-based sequence alignment was performed to find the conserved secondary structure using the PROMALS3D and JPred4 and manually adjusted with UGENE to identify the homologous positions with shift insertions and deletions from conserved secondary structure elements [36, 37].

Phylogenetic tree construction

The phylogenetic tree of the subgroups and Arabidopsis RNHLS proteins were generated using Clustal Omega at the same parameters for protein sequence alignment. For the total RNHLS proteins phylogenetic tree, we employed the MAFFT software for multiple sequence alignment and constructed phylogenetic tree using FastTree [38, 39]. We further modified the phylogenetic tree with online software Interactive tree of life (iTOL) [40].

Protein domain diagram

The information of protein domains was from Pfam (http://pfam.xfam.org/), InterPro (http://www.ebi.ac.uk/interpro/) and UniProt (https://www.uniprot.org/). The diagram were depicted with IBS [41].

RNA-seq data of Arabidopsis various tissues

The RNA-seq data of Arabidopsis *RNHLS* genes used here were downloaded from TraVA (Transcriptome Variation Analysis, http://travadb.org/), except for which were not available [42]. The heatmap was visualized using Morpheus online software with default settings. GO analysis of highly expressed genes in meristems was performed on the online database agriGO v2.0 [43].

Results

The species selected to cover the characters of RNHLS proteins in plants

To acquire representative plant species for analysis the RNHLS proteins, we employed the Gene3D to investigate co-occurring domain families, with value for RNHLS proteins with other protein domain families in 33 plant species (Fig. 1A and Fig. S1) [29]. The co-occurrence of RNHLS domain with DNA polymerase domain, nuclease and nucleotide triphosphate hydrolase domain are in most of the plant species (Fig. S1). Based on these results, we selected 10 representative plant species and added three other species including *C. sinensis, M. polymorpha,*and *K. flaccidum* to further represent the evolutionary tree of plants for further investigation (Fig. 1B) [44].

A global view of RNHLS protein clusters

To understand the function and distribution of RNHLS proteins in the 13 plant species, we acquired the RNHLS proteins from the selected plants using the protein database InterPro [45]. A total of 6193 proteins were collected through this method. To elucidate the relationship among these proteins, we conducted a clustering analysis which relied on the sequence similarity of proteins. First, we chose the representative proteins using CD-HIT with a threshold of 50 percent and achieved 2066 representative RNHLS proteins. Then, we compared the similarity of the these RNHLS proteins with total RNHLS (6193) using BLAST with an average pairwise E-value at least of 1×10^{-18}. The network showed that there are 27 clusters (Fig. 2). Most part of them are Ty1/Copia or Ty3/Gypsy reverse transcriptases and other transposases or integrases. The second largest part of them are different types of nucleic acid endonucleases and exonucleases (Fig. 2). The third part of them are Argonaute proteins and DNA polymerases. The

separation of these clusters was consistent with their biological function and confirmed that our SSN map can show the relative information of the investigated RNHLS proteins. To further investigation the relationships of all RNHLS proteins and within the 27 subclusters, we constructed phylogenetic trees of total RNHLS proteins and each of the 27 subclusters. Consistent with the SSN network, the RNHLS proteins are diversified and most of them are transposon polyprotein (Fig. S3 and Fig. S4).

To further investigate the composition of these clusters at species level, we indicated the species information with different colors at the SSN map (Fig. S2). The results showed that most of the proteins comes from *O. sativa* are clustered in the Ty1/Copia and Ty3/Gypsy polyprotein clusters. In the Ty1/Copia polyprotein group, proteins from *A. thaliana* and *O. sativa* are clustered in different subclusters (Fig. S2). Moreover, proteins from *C. sinensis, C. crispus, S. lycopersicum* and *C. eustigma* also form subclusters in this cluster. Interestingly, transposases from *C. crispus* and *M. polymorpha* are clustered into two relative isolated groups (Fig. S2). Consistent with this phenomenon, most of transposon-related polyprotein clusters prefer to cluster together in a species dependent manner. Otherwise, the other clusters such as AGO, PRP8, DNA polymerase and RNH2 are distributed more evenly compared with transposon-related polyprotein clusters (Fig. S2). Additionally, the AGO cluster is separated into two sub-clusters consistent with their biological function (Fig. 2). Together, these results indicate that RNHLS proteins are very diversified and the evolution of transposon within each species are exclusive.

Phylogenetic tree of the Arabidopsis RNHLS proteins

To better understand the RNHLS proteins in plants, we used the RNHLS proteins from Arabidopsis as candidates to conduct alignment and construct a phylogenetic tree (Fig. 3). Consistent with the protein clusters, the retrotransposon reverse transcriptases occupy a big part of the clades (Fig. 3). There are 5 clades of Ty1/copia reverse transcriptases, which consistent with the cluster results that there are more than eight sub-clusters (Fig. 3). Meanwhile, the HAT transposon proteins are also clustered together in the tree (Fig. 3). Furthermore, the non-LTR retrotransposon reverse transcriptases comprise another three clades of the tree. The CAF1 proteins are separated into two clades compared with MEN or SDN proteins which shows only one clade on the tree (Fig. 3). Additionally, the exonucleases are clustered together. Interestingly, we found that the AGO family proteins are separated into four clades of them with AGO1/5/10, AGO2/3, AGO4/6/8/9 and AGO7 and this result is identical to previous reports [46]. Various of DNA polymerases with RNHLS domain are clustered into one clade (Fig. 3). Moreover, we found that one type of RNHLS proteins come from Ring-type E3 ubiquitin transferases. The PRP8 family proteins are clustered in one clade (Fig. 3). Together these results give us a comprehensive view of RNHLS proteins and the relationship among them in Arabidopsis.

The domain construction of different types of RNHLS proteins

To better characterize the properties of the RNHLS proteins in plant, we chose the representative Arabidopsis RNHLS proteins in each of the clades to draw a protein domain diagram to indicate their domain composition (Fig. 4). The RNH1 contains three important domains with HBD, Dis and RNH domain. HBD is important for RNA:DNA hybrid binding while RNH domain is for catalytic activity [47]. The Dis domain provides the ability for RNH to catalyze a variety of hybrids without unloading (Fig. 4) [47]. Based on this, most RNHLS proteins combine RNH domain with other domains to exercise their destroying RNA:DNA functions during different biological processes. To analysis the relationship of RNHLS domain among different RNHLS proteins, we performed a multiple sequence alignment with the truncated sequences of each representative Arabidopsis RNHLS proteins (Fig. 5). The results indicate that the conserved alpha helix and beta sheet domain present in these RNHLS proteins which consistent with their annotated function in InterPro (Fig. 4 and Fig. 5).

RNHLS proteins are abundant in plant meristems

To predict the biological function of RNHLS proteins in plants, we first analyzed their expression pattern in Arabidopsis and rice using RNA-seq data from database (Fig. 6 and Fig. S5). In Arabidopsis and rice, various *RNHLS* genes expressed in different tissues are grouped together. In Arabidopsis, we found these genes are mainly expressed at meristem tissues, such as shoots and root meristems and inflorescences (Fig. 6A). Some of the genes are mainly expressed in anthers and pistils (Fig. 6A). In addition, in the young and mature anthers different types of *RNHLS* genes show clearly different expression patterns (Fig. 6A). Meanwhile, the dry seeds accumulate another group expression of *RNHLS* genes (Fig. 6A). However, some genes are expressed at senescent tissues due to programmed cell death accompanied with nucleic acid degradation (Fig. 6A). We perform a GO analyzer of the meristem expressed genes and it shows that these proteins are involved in nucleic acid metabolism (Fig. 6B). In rice, the result shows that post-emergence inflorescences and pre-emergence inflorescences are grouped with totally different expression of *RNHLS* genes consistent with Arabidopsis (Fig. S5). In the embryo–25 dap (days after plant), there is another group of *RNHLS* expressed (Fig. S5). Moreover, in seedling four-leaf stage, half of the investigated *RNHLS* genes are expressed in this tissue (Fig. S5). Taken together, the expression of *RNHLS* genes in different groups are throughout Arabidopsis and rice development and consistent with their biological function. These results indicate the importance of *RNHLS* genes in plants.

Discussion

In this study, we have characterized 6193 RNHLS proteins in 13 plants. Among them, 2066 representative RNHLS proteins are clustered into 27 clusters, including reverse transcriptases, transposases, integrases, nucleic acid endonucleases, exonucleases, key spliceosome component Prp8, as well as other RNA

nucleases. Moreover, RNHLS domain is combined with various types of protein domains in different RNHLS proteins. And the catalytic alpha helix and beta sheet domain are conserved among them in Arabidopsis. In addition, *RNHLS* genes are actively expressed in meristems, flower tissues and other metabolic tissues, and various *RNHLS* genes expressed in different tissues are grouped together.

The selected 13 plant species represent the evolutionarily process of plants (Fig. 1B) [44]. The distribution of RNHLS proteins in all species of plants suggests they are a type of the evolutionarily oldest proteins [16]. Among 27 clustered RNHLS proteins, DNA polymerase, GTF2B, CAF1, PRP8, RRP6L, AGO and EXO worked in basic nucleic acid metabolic processes are conserved in plant kingdom [15]. WRN is able to digest the RNA moiety of the RNA:DNA hybrid [48]. Recently, it was reported that WRN is involved in R-loop-related genome stability [49]. Moreover, SDN proteins are involved in degrading small RNAs and could trimming dsRNA therefore raising the possibility of trimming RNA strand in the RNA:DNA hybrids because of high structure similarity of dsRNA and RNA-DNA hybrid [50]. The family numbers of TE-related RNHLS proteins are larger, especially for Ty1/Copia and Ty3/Gypsy reverse transcriptase family (Fig.2 and Fig. S4). Some of the TE-related RNHLS proteins are not present in all selected plants, such as RICESLEEPER is only present in rice (Fig. S2), Ty1/Copia, Ty3/Gypsy and Non-LTR reverse transcriptase are mainly enriched in rice, Arabidopsis and other higher plants (Fig. 2 and Fig. S2). This is consistent with the size and composition of their genome [44]. As retrotransposons are abundant in higher plants, the TE-related RNHLS proteins may play a major role in driving genetic diversity and evolution of these organisms [13, 51].

RNHLS protein is a large group of evolutionarily related, but strongly diverged protein family [15]. The first identified RNHLS protein is RNase H from *Escherichia coli* [52]. There are two types of RNase H enzymes, which differ in structure and substrate specificity. RNase H1 is monomeric, whereas RNase H2 is monomeric in bacteria but composed of three subunits in eukaryotes: RNH2A (the catalytic subunit), RNH2B and RNH2C. Both types of RNase H enzyme can remove RNA-DNA hybrids in addition to having different specialized roles. The phylogenetic analysis of RNHLS proteins in Arabidopsis shows that some proteins with similar function were clustered in different group according to their sequence, such as AGO proteins. While some proteins from different protein families clustered together, such as some Ty1/Copia and non-LTR reverse transcriptase proteins. All these indicate that both amino acid construction and the motif composition of RNHLS proteins are diversity. Domain diagram analysis of representative proteins shows that all the RNHLS proteins have RNH domains except AGOs, which use a PIWI domain instead (Fig.4). The Piwi domain have been reported adopts an RNase H-like fold and enables some, but not all, AGO proteins to cleave target RNAs complementary to the bound sRNAs [19]. A catalytic triad (Asp-Asp-His/Asp, DDH/D) is generally thought to be responsible for slicer activity is present in both RNH domains and PIWI domain [53]. Further study need to test whether AGOs have ability to cleave the RNA strand in RNA-DNA hybrid or worked as a sensor for RNA-DNA hybrid. In addition, RNHLS proteins are abundant in meristems, flower tissues and other metabolic tissues, and various *RNHLS* genes expressed in different tissues grouped together indicate they worked division and cooperation (Fig. 6). As RNHLS proteins work on nucleic acid metabolism especially on R-loop homeostasis, this study characterized the RNHLS

Proteins in plants and provide new direction on R-loop regulation and function identification study in the future.

Conclusions

In this study, we employed the bioinformatics approaches to analyze the RNHLS proteins in representative 13 plant species. A SSN network showed that these proteins can be separated into 27 clusters and consistent with their biological functions. Moreover, the *RNHLS* genes expression pattern during arabidopsis and rice development further revealed that these proteins are mainly expressed in meristems and metabolic tissues. These evidences shed light on our understanding of R-loop regulation in plants.

Abbreviation

AGO, Argonaute; CAF1, putative CCR4-associated factor 1 homolog; Dis, disorder protein domain; DUF1744, domain of unknown function 1744 domain; Exo, exonuclease; Gag-pol, putative gag-pol polyprotein; GTF2B, general transcription factor 2-related zinc finger protein; HBD, hybrid binding domain; non-LTR, putative non-LTR retroelement reverse transcriptase; Prp8, pre-mRNA-processing-splicing factor 8; RBRE3UT, RBR-type E3 ubiquitin transferase; RNase H, ribonucleases H; RNH2, ribonuclease H2; RNHLS, RNase H-like superfamily; RRP6L, RRP6-like protein; SDN, small RNA degrading nuclease; SSN, sequence similarity network; WRN, werner syndrome ATP-dependent helicase; ZBED1, zinc finger BED domain-containing protein DAYSLEEPER (Transposase-like protein DAYSLEEPER).

Declarations

Author contributions: Q. S. conceptualized the study. S. L. analyzed the sequences and prepared the figures, and prepared the manuscript with K. L. and Q. S.

Acknowledgements

This work was funded by grants from the Ministry of Science and Technology of China (grant no. 2016YFA0500800 to Q. S.); and the National Natural Science Foundation of China (grant nos. 91740105 and 31822028 to Q. S.). The Sun Lab was supported by Tsinghua-Peking Joint Center for Life Sciences, Tsinghua University Initiative Scientific Research Program, and the 1000 Young Talent Program of China. K. L. was supported by postdoctoral fellowship from Tsinghua-Peking Joint Center for Life Sciences.

Competing interests: The authors declare no competing interests.

References

1.Thomas M, White RL, Davis RW. Hybridization of RNA to double-stranded DNA: formation of R-loops. Proc Natl Acad Sci U S A. 1976;73(7):2294–2298.

2.Santos-Pereira JM, Aguilera A. R loops: new modulators of genome dynamics and function. Nat Rev Genet. 2015: 16(10):583–597.

3.Huertas P, Aguilera A. Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. Mol Cell. 2003; 12(3):711–721.

4.Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell. 2012; 45(6):814–825.

5.Chan YA, Aristizabal MJ, Lu PY, Luo Z, Hamza A, Kobor MS, Stirling PC, Hieter P. Genome-wide profiling of yeast DNA:RNA hybrid prone sites with DRIP-chip. PLoS Genet. 2014; 10(4):e1004288.

6.Wahba L, Costantino L, Tan FJ, Zimmer A, Koshland D. S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. Genes Dev. 2016; 30(11):1327–1338.

7.Xu W, Xu H, Li K, Fan Y, Liu Y, Yang X, Sun Q. The R-loop is a common chromatin feature of the Arabidopsis genome. Nat Plants. 2017; 3(9):704–714.

8.Dumelie JG, Jaffrey SR. Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. Elife. 2017, 6.

9.Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, Xu X, Chedin F. Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. Mol Cell. 2016;63(1):167–178.

10.Fang Y, Chen L, Lin K, Feng Y, Zhang P, Pan X, Sanders J, Wu Y, Wang XE, Su Z et al. Characterization of functional relationships of R-loops with gene transcription and epigenetic modifications in rice. Genome Res. 2019.

11.Stein H, Hausen P. Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase. Science. 1969;166(3903):393–395.

12.Hausen P, Stein H. Ribonuclease H. An enzyme degrading the RNA moiety of DNA-RNA hybrids. Eur J Biochem. 1970;14(2):278–283.

13.Smyshlyaev G, Voigt F, Blinov A, Barabas O, Novikova O. Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. Proc Natl Acad Sci U S A 2013;110(50):20140–20145.

14.Joyce GF. The antiquity of RNA-based evolution. Nature. 2002;418(6894):214–221.

15.Majorek KA, Dunin-Horkawicz S, Steczkiewicz K, Muszewska A, Nowotny M, Ginalski K, Bujnicki JM. The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic Acids Res. 2014;42(7):4160–4179.

16.Ma BG, Chen L, Ji HF, Chen ZH, Yang FR, Wang L, Qu G, Jiang YY, Ji C, Zhang HY. Characters of very ancient proteins. Biochem Biophy Res Co. 2008;366(3):607–611.

17.Pena V, Rozov A, Fabrizio P, Luhrmann R, Cwahl M. Structure and function of an RNase H domain at the heart of the spliceosome. EMBO J. 2008;27.

18.Zuo Y, Deutscher MP. Exoribonuclease superfamilies: structural analysis and phylogenetic distribution. Nucleic Acids Res.2001;29(5):1017–1026.

19.Song JJ, Smith SK, Hannon GJ, Joshua-Tor L. Crystal structure of Argonaute and its implications for RISC slicer activity. Science 2004;305(5689):1434–1437.

20.Rice PA, Baker TA. Comparative architecture of transposase and integrase complexes. Nat Struct Biol. 2001;8(4):302–307.

21.Parker JS, Roe SM, Barford D. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. EMBO J 2004;23(24):4727–4737.

22.Ariyoshi M, Vassylyev DG, Iwasaki H, Nakamura H, Shinagawa H, Morikawa K. Atomic structure of the RuvC resolvase: a holliday junction-specific endonuclease from E. coli. Cell. 1994;78(6):1063–1072.

23.Nowotny M, Gaidamakov SA, Crouch RJ, Yang W. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. Cell. 2005;121(7):1005–1016.

24.Sun Q, Csorba T, Skourti-Stathaki K, Proudfoot NJ, Dean C. R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. Science. 2013;340(6132):619–621.

25.Shafiq S, Chen C, Yang J, Cheng L, Ma F, Widemann E, Sun Q. DNA Topoisomerase 1 Prevents R-loop Accumulation to Modulate Auxin-Regulated Root Development in Rice. Mol Plant 2017;10(6):821–833.

26.Yang Z, Hou Q, Cheng L, Xu W, Hong Y, Li S, Sun Q. RNase H1 Cooperates with DNA Gyrases to Restrict R-Loops and Maintain Genome Integrity in Arabidopsis Chloroplasts. Plant Cell 2017; 29(10):2478–2497.

27.Arab K, Karaulanov E, Musheev M, Trnka P, Schafer A, Grummt I, Niehrs C. GADD45A binds R-loops and recruits TET1 to CpG island promoters. Nat Genet 2019;51(2):217–223.

28.Yuan W, Zhou J, Tong J, Zhuo W, Wang L, Li Y, Sun Q, Qian W. ALBA protein complex reads genic R-loops to maintain genome stability in Arabidopsis. Sci Adv 2019;5(5):eaav9040.

29.Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C, Lees J. Gene3D: Extensive prediction of globular domains in proteins. Nucleic Acids Res. 2018;46(D1):D435-D439.

30.Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–1186.

31.Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–410.

32.Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–3152.

33.Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–2504.

34.Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(1):539.

35.Okonechnikov K, Golosova O, Fursov M, team U. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166–1167.

36.Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36(7):2295–2300.

37.Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015;43(W1):W389–394.

38.Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–3066.

39.Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641–1650.

40.Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;l44(W1):W242–245.

41.Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, Lahrmann U, Zhao Q, Zheng Y, Zhao Y et al. IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics. 2015;31(20):3359–3361.

42.Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J. 2016;88(6):1058–1070.

43.Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z.agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017;45(W1):W122-W129.

44.Chang C, Bowman JL, Meyerowitz EM.Field Guide to Plant Model Systems. Cell. 2016;167(2):325–339.

45.Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:D211–215.

46.Fang X, Qi Y. RNAi in Plants: An Argonaute-Centered View. Plant Cell. 2016;28(2):272–285.

47.Cerritelli SM, Crouch RJ. Ribonuclease H: the enzymes in eukaryotes. FEBS J. 2009;276(6):1494–1505.

48.Suzuki N, Shiratori M, Goto M, Furuichi Y. Werner syndrome helicase contains a 5'—>3' exonuclease activity that digests DNA and RNA strands in DNA/DNA and RNA/DNA duplexes dependent on unwinding. Nucleic Acids Res. 1999;27(11):2361–2368.

49.Marabitti V, Lillo G, Malacaria E, Palermo V, Sanchez M, Pichierri P, Franchitto A. ATM pathway activation limits R-loop-associated genomic instability in Werner syndrome cells. Nucleic Acids Res. 2019.

50.Chen J, Liu L, You C, Gu J, Ruan W, Zhang L, Gan J, Cao C, Huang Y, Chen X et al. Structural and biochemical insights into small RNA 3' end trimming by Arabidopsis SDN1. Nat Commun. 2018;9(1):3585.

51.Moelling K, Broecker F, Russo G, Sunagawa S. RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense. Front Microbio. 2017;8:1745.

52.Leis JP, Berkower I, Hurwitz J. Mechanism of action of ribonuclease H isolated from avian myeloblastosis virus and Escherichia coli. Proc Natl Acad Sci U S A. 1973;70(2):466–470.

53.Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ. Argonaute2 is the catalytic engine of mammalian RNAi. Science. 2004;305(5689):1437–1441.

Supplementary Material Legends

Supplementary material 1: Supplementary figure 1 to supplementary figure 5. (PDF 2560 kb) Supplementary material 2: A representative SSN of the RNHLS proteins. (CYS 4687 kb) Supplementary material 3: A representative SSN of the RNHLS proteins colored with species information. (CYS 4657 kb) Supplementary material 4: Supplementary table 1. Gene3D domain family matrix of RNHLS domain that co-occur with other protein domains. (XLSX 52.4 kb) Supplementary material 5: Supplementary table 2. RNHLS protein information used in this study. (XLS 668 kb) Supplementary material 6: Supplementary table 3. tissue-specific expression data of RNHLS genes of Arabidopsis and rice. (XLSX 187 kb)

Figures

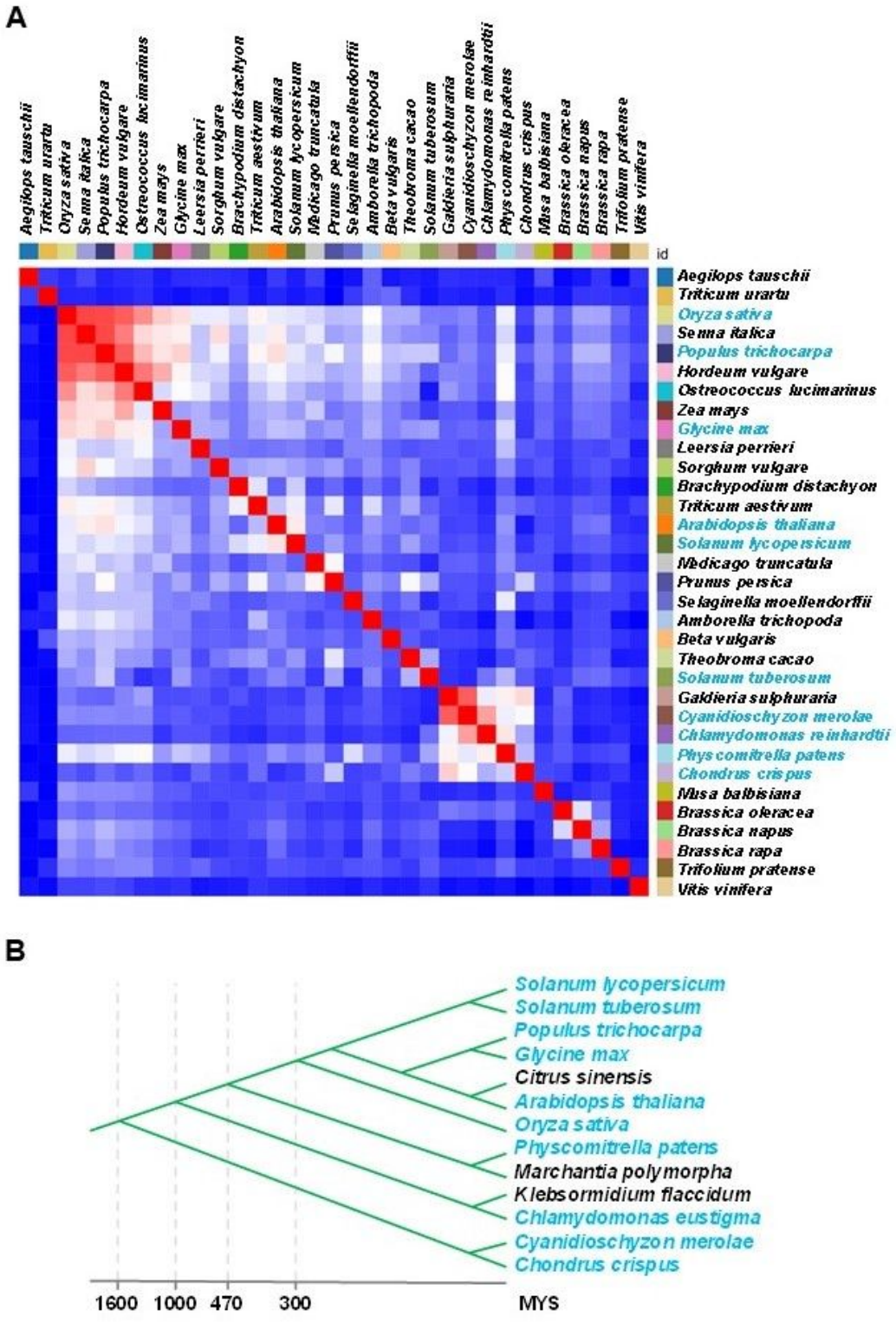

Figure 1

Selection of candidate species for investigation of plant RNHLS proteins. (A) The similarity of Gene3D domain family co-occurrence of RNHLS domain with other protein domains in all available plant species. The species to be investigated in this study were marked in blue font. The original matrix of other protein domains that co-occurred with RNHLS domain were downloaded from Gene3D (http://gene3d.biochem.ucl.ac.uk). The similarity heatmap were performed in Morpheus

(https://software.broadinstitute.org/morpheus/) using the default setting. (B) The cladogram of species for investigation of RNHLS proteins. The blue font is marked the selected species from the similarity matrix while the black font is marked the species which are supplemented for better covering the plant evolutionary.

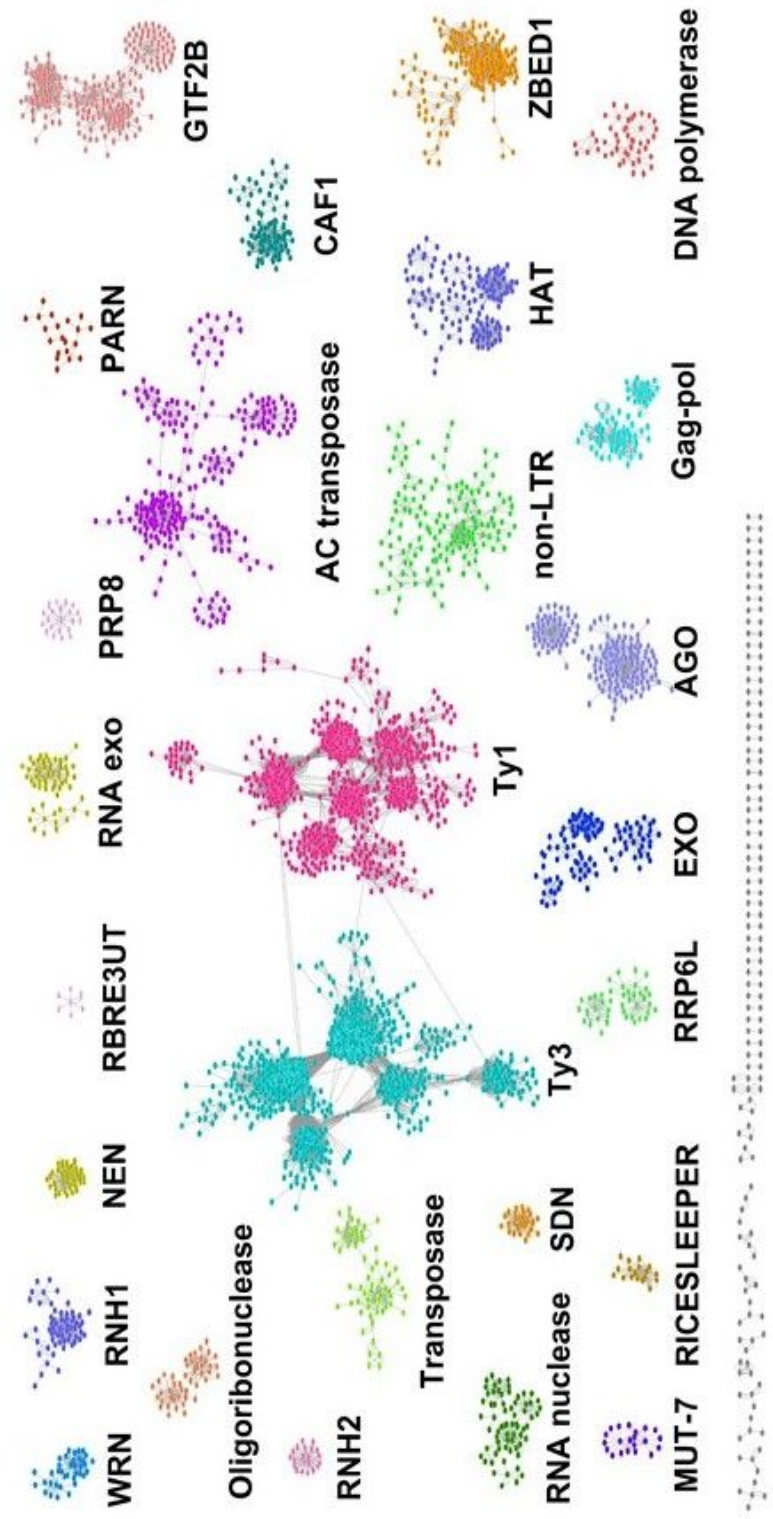

Figure 2

A representative SSN of the RNHLS proteins. A total of 4,755 protein sequences were depicted by 2,066 nodes (circles), which represent proteins sharing >50% sequence identity. Edges between nodes indicate an average pairwise BLAST E-value of at least 1 × 10-18. Node coloring represents subgroup classification. Names indicate subgroups that contain at least one protein with literature-documented functional information except for nodes marked with gray. The network was visualized by Cytoscape3.7.0 using the organic layout algorithm. WRN, Werner syndrome ATP-dependent helicase; SDN, small RNA degrading nuclease; RNH2, ribonuclease H2; AGO, Argonaute; RNH1, ribonuclease H1; RRP6L, RRP6-like protein; RBRE3UT, RBR-type E3 ubiquitin transferase; Exo, exonuclease; Ty1, Ty1/copia-element polyprotein; Ty3, Gypsy/Ty3 element polyprotein; non-LTR, putative non-LTR retroelement reverse transcriptase; CAF1, putative CCR4-associated factor 1 homolog; HAT, HAT dimerisation domain-containing protein; GTF2B, general transcription factor 2-related zinc finger protein; ZBED1, zinc finger BED domain-containing protein DAYSLEEPER (Transposase-like protein DAYSLEEPER); Gag-pol, putative gag-pol polyprotein; PRP8, Pre-mRNA-processing-splicing factor 8.

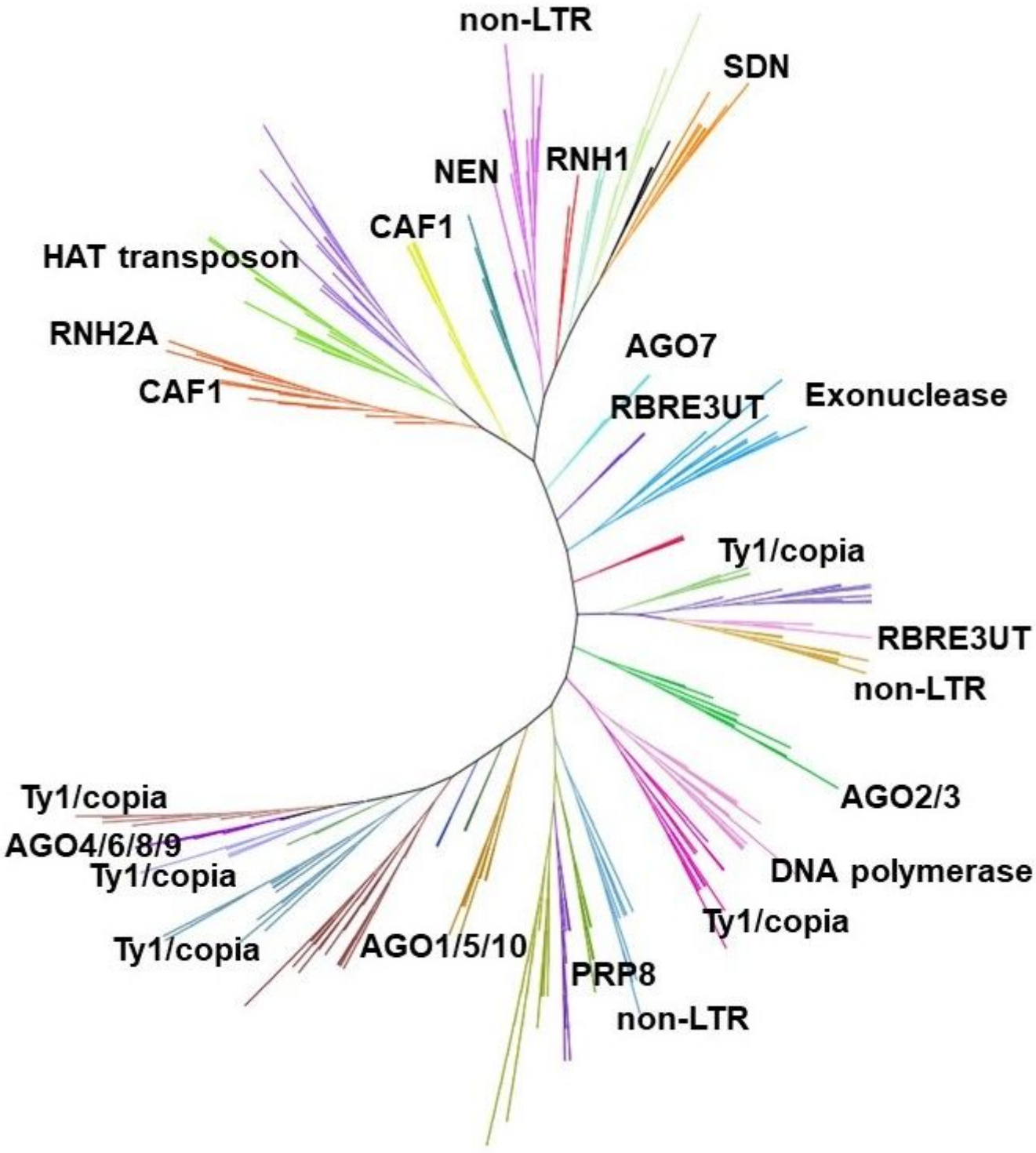

Figure 3

Phylogenetic tree of the Arabidopsis RNHLS proteins. A total of 691 Arabidopsis RNHLS protein sequences were aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). The phylogenetic tree was constructed using the results from this alignment and modified using iTOL (https://itol.embl.de/). Different types of nucleases and polymerases are clearly separated to different clades. The distribution of transposon-related proteins is more diversified than AGOs, SDNs and MENs.

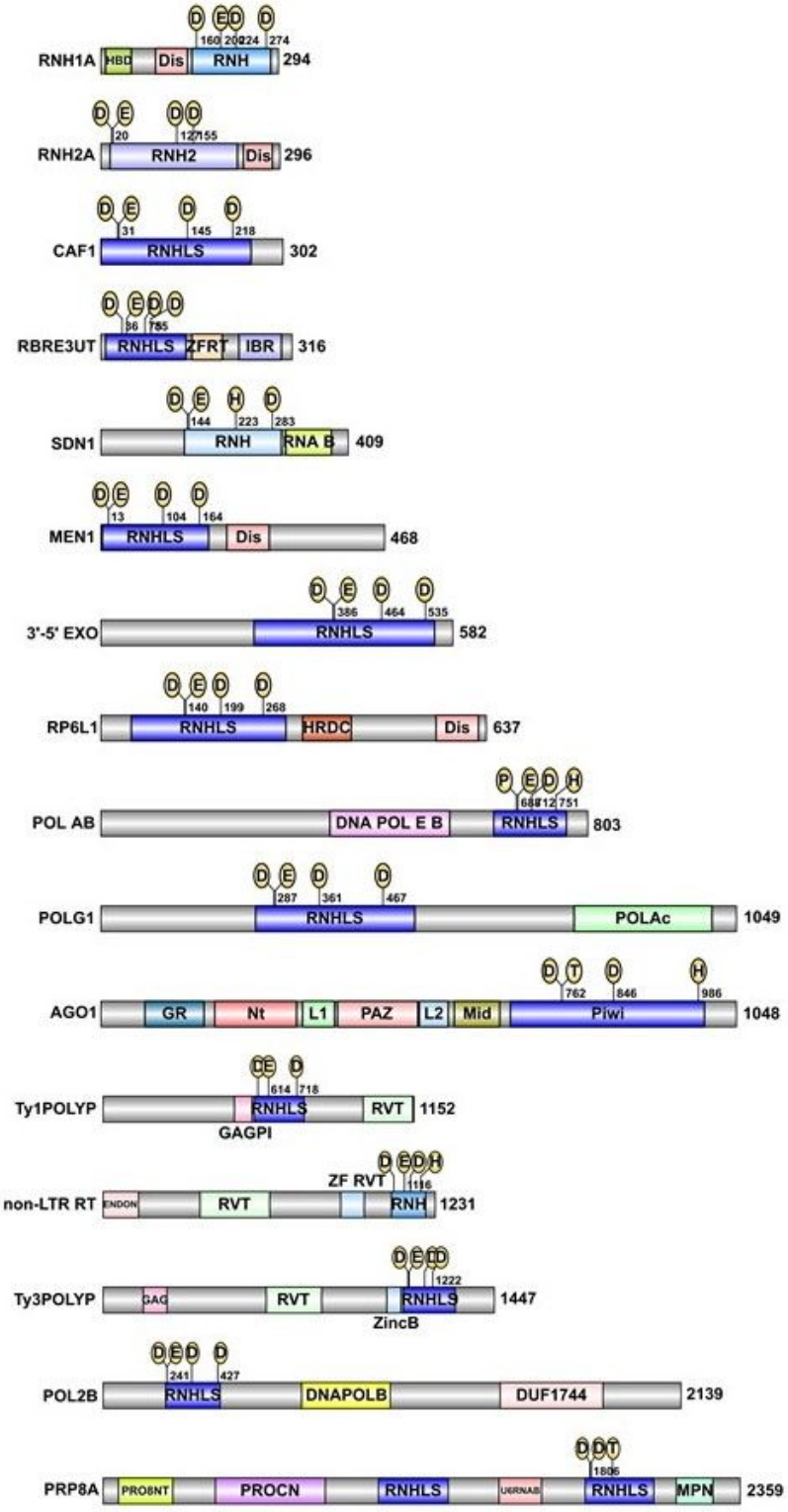

Figure 4

Protein domain diagram of Arabidopsis representative RNHLS proteins. The domain annotations of these proteins were combined from Pfam (http://pfam.xfam.org/), InterPro (http://www.ebi.ac.uk/interpro/) and UniProt (https://www.uniprot.org/). HBD, hybrid binding domain; Dis, disorder protein domain; RNH, ribonuclease H domain; RNHS, Ribonuclease H-like superfamily domain; ZFRT, zinc finger ring-type domain; IBR, In Between Ring fingers domain; RNA B, RNA binding domain; HRDC, helicase and RNase D

C-terminal; DNA POL E/B domain, DNA polymerase alpha/epsilon subunit B domain; POLAc, DNA polymerase A domain; GR, glycine-rich domain; Nt, N-terminal domain; L1, linker 1 domain; AGO1 Argonaute 1, consists of GR, glycine-rich domain, Nt, N-terminal domain, L1, linker 1 domain; PAZ, PAZ domain; L2, linker 2 domain; Mid, Mid domain; Piwi, Mid domain; Ty1POLYP, Ty1/copia-element polyprotein; GAGPI, GAG-pre-integrase domain; RVT, reverse transcriptase, RNA-dependent DNA polymerase domain; ENDON, Endonuclease superfamily domain; ZF RVT, reverse transcriptase zinc-binding domain; Ty3POLYP, Ty3/Gypsy-element polyprotein; GAG, retrotransposon gag domain; ZincB, integrase zinc-binding domain; POL2B, DNA polymerase epsilon catalytic subunit; DNAPOLB, DNA-directed DNA polymerase family B multifunctional domain; DUF1744, domain of unknown function 1744 domain; PRP8, Pre-mRNA-processing-splicing factor 8; PRO8NT, N terminus of PRP8 domain; PROCN, central domain of PRP8 domain; U6RNAB, U6-snRNA-binding of PRP8 domain; MPN (Mpr1, Pad1 N-terminal) domain. D represents for aspartic acid. E represents for glutamic acid. H represents for histidine. T represents for threonine. The numbers besides the protein indicate the amino acid number of that protein.

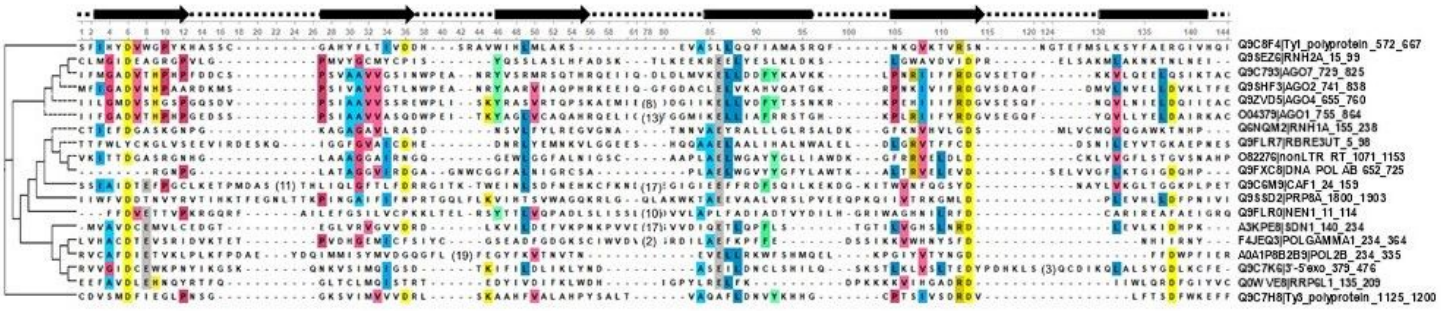

Figure 5

Multiple sequence alignment of Arabidopsis representative RNHLS proteins. The RNHLS domains of each protein were selected for this alignment. Each of RNHLS domains were predicted the secondary structure using Jpred4 (http://www.compbio.dundee.ac.uk/jpred/index.html). The alignment was visualized using UGENE and adjusted manually.

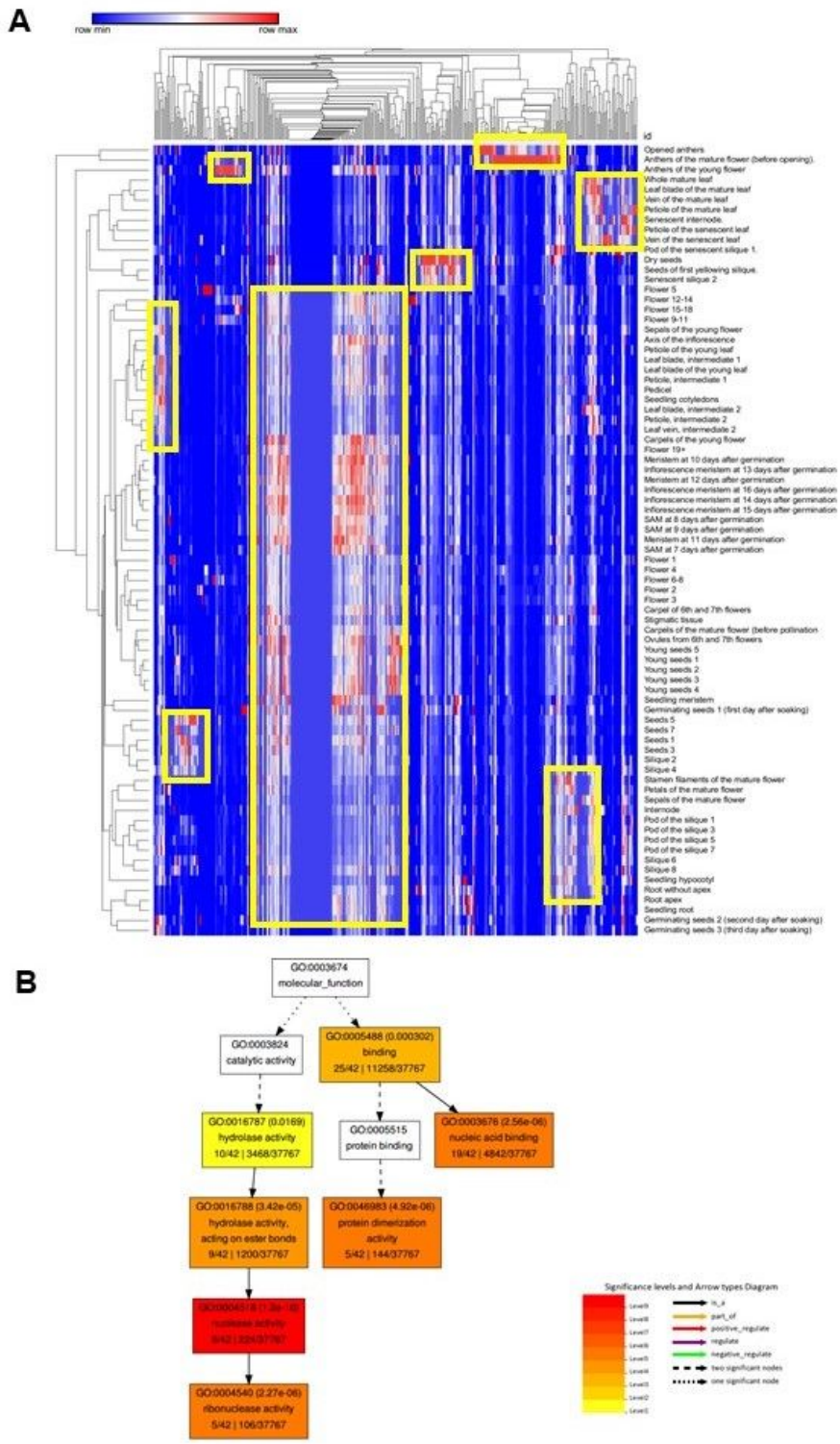

Figure 6

Tissue-specific expression of Arabidopsis RNHLS genes. (A) The transcriptome of RNHLS genes were downloaded from TRAVA (http://travadb.org/browse/) except for some of the genes could not be retrieved from that database. These genes are mainly expressed at the meristems such as flowers inflorescences and root meristems. The yellow rectangles indicate the clusters of genes highly expressed

in these tissues. (B) GO annotation of the genes expressed in Arabidopsis meristems. These genes are mainly involved in nucleic acid metabolism.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RNHLfiguresupplementary.pptx
- Supplementarytable1.xlsx
- Supplementarytable3.xlsx
- Supplementarytable2.xls