Title

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

Shuai Li, Kunpeng Liu and Qianwen Sun[#]

Tsinghua-Peking Joint Center for Life Sciences and Center for Plant Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

[#] Corresponding author Qianwen Sun: sunqianwen@mail.tsinghua.edu.cn.

Main conclusion

This study showed that RNaseH-like domain containing proteins can form 27 subclusters and mainly regulate nucleic acid metabolism in meristem tissue, revealing these proteins is critical in R-loop regulation.

Abstract

R-loop is a nucleic acid structure containing an RNA-DNA hybrid and a displaced single-stranded DNA. Recently, accumulated evidences show 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. In this study we characterized 6193 RNHLS proteins from 13 representative plant species, and clustered them into

27 clusters, among which reverse transcriptase and exonuclease are the two largest groups. Further 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 combined with 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. In summary, we systematically analyzed the RNHL protein in plant and found that there are mainly 27 sub-clusters of them. Most of these proteins are implicated in DNA replication, RNA transcription and nucleic acid degradation. we characterize and classify the RNHLS proteins in plants, which affords new insights on investigation of novel regulatory mechanisms and functions in R-loop biology.

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

Introduction

R-loop is a nucleotide structure consists of RNA-DNA hybrid with the displaced single-stranded DNA (Thomas et al. 1976). Formation of R-loops *in vivo* was found in process of DNA replication, retrovirus or single stranded RNA virus infection or retrotransposon life cycle (Santos-Pereira and Aguilera 2015). For a long time, R-loops had been considered as transcription by-product since the nascent transcript could bind back to its template DNA (Huertas and Aguilera 2003). However, recent studies in different organisms showed that R-loops are widespread on genomes in yeast, plant and mammals (Ginno et al. 2012; Chan et al. 2014; Wahba et al. 2016; Xu et al. 2017; Dumelie and Jaffrey 2017; Sanz et al. 2016). These evidences together revealed that R-loop levels 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 could destroy and remove R-loops (Stein and Hausen 1969; Hausen and Stein 1970; Smyshlyaev et al. 2013). RNA molecule might be the origin of life and RNase H mainly disrupts the direct interaction of RNA and DNA (Joyce 2002). 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 (Majorek et al. 2014; Ma et al. 2008). The RNase H-like superfamily (RNHLS) proteins are group 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 (V. et al. 2008; Zuo and Deutscher 2001; Song et al. 2004; Rice and Baker 2001; Parker et al. 2004; Ariyoshi et al. 1994; Nowotny et al. 2005). 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 (Majorek et al. 2014). 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 (Majorek et al. 2014). Plants are sessile and need to overcome different biotic or abiotic stress over their life cycle. Therefore, the 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 plant 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 clustered into 27 clusters, among them reverse transcriptase and exonuclease are the 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 proteins with other protein domains

The co-occurrence matrix of RNHLS domain (3.30.420.10) with other protein domains was downloaded from Gene3D (Lewis et al. 2018). The heatmap was processed by online software Morpheus (https://software.broadinstitute.org/morpheus/) using the default settings.

Protein Sequence search

To identify the RNase H superfamily proteins in Arabidopsis comprehensively, we use the proteins sequences from RNHLS (IPR012337) of InterPro database (http://www.ebi.ac.uk/interpro/) filtered with different species. In addition, we perform sequence BLAST with AtRNH1A (At3g01410) as a query in the Phytozome and NCBI database (Goodstein et al. 2012; Altschul et al. 1990). We carry out the analysis after combining the sequence from these three and remove the duplication from those results.

Protein sequence cluster

The identified proteins are first analyzed with CD-HIT and the threshold was set at 60% with word size of 3 (Fu et al. 2012). 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 remove of redundancy protein (Altschul et al. 1990). The network was visualized using Cytoscape3.7.0 and layout using the organic layout algorithm and annotated manually (Shannon et al. 2003).

Protein sequence alignment

To analysis the protein sequences systemically, we use Clustal Omega online service with the default parameters except for all the iterations set as five (Sievers et al. 2011). The alignment results were visualized with UGENE and MEGA7 (Okonechnikov et al. 2012). 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 (Pei et al. 2008; Drozdetskiy et al. 2015).

Phylogenetic tree construction

The phylogenetic tree of the subgroups and arabidospsis 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 (Katoh et al. 2002; Price et al. 2009). We further modify the phylogenetic tree with online software Interactive tree of life (iTOL) (Letunic and Bork 2016).

Protein domain diagram

The information of protein domain is from Pfam (http://pfam.xfam.org/), InterPro (http://www.ebi.ac.uk/interpro/) and UniProt (https://www.uniprot.org/). The diagram depicted with IBS (Liu et al. 2015).

RNA-seq data of Arabidopsis various tissues

The RNA-seq data of Arabidopsis *RNHLS* genes used here are downloaded from TraVA (Transcriptome Variation Analysis, http://travadb.org/), except for which are not available (Klepikova et al. 2016). The heatmap was visualized using Morpheus online software with default settings. GO analysis of the meristem highly expressed genes was performed on the online database agriGO v2.0 (Tian et al. 2017).

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) (Lewis et al. 2018). The co-occurrence of RNHLS domain with DNA polymerase, nuclease and nucleotide triphosphate hydrolase domain families 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) (Chang et al. 2016).

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 (Hunter et al. 2009). 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 similarities of proteins. First, we chose the representative proteins using CD-HIT with a threshold of 50 percent and got 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 of them are Ty1/Copia or Ty3/Gypsy reverse transcriptase and other transposase or integrases. The second largest part is different types of nucleic acid endonuclease and exonuclease (Fig. 2). The third part is Argonaute proteins and DNA polymerases. The separation of these clusters is constant with their biological function and confirm 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 sub-clusters (Fig. S2). Moreover, *C. sinensis*, *C. crispus*, *S. lycopersicum* and *C. eustigma* also form sub-clusters in this cluster. Interestingly, transposases from *C. crispus* and *M. polymorpha* cluster into two relative isolated groups (Fig. S2). Consistent with this phenomenon, most of transposon-related polyprotein clusters prefer to clustering 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 are separated into two sub-clusters constant with their biological function (Fig. 2). Together, these information indicates 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 plant, we use the RNHLS proteins from Arabidopsis as candidates to conducted alignment and constructed a phylogenetic tree (Fig. 3). Constant with the protein clusters, the retrotransposon reverse transcriptase is a big part of the clades (Fig. 3). There are 5 clades of Ty1/copia reverse transcriptase, which consistent with the cluster result 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 transcriptase comprises another three clades on the tree. The CAF1 proteins are separated into two clades compare with MEN or SDN proteins which shows only one clade on the tree (Fig. 3). Additionally, the exonucleases cluster together. Interestingly, we found that the AGO family proteins are separated to four clades of them with AGO1/5/10, AGO2/3, AGO4/6/8/9 and AGO7 and this result is identity to previous reports (Fang and Qi 2016). Various of DNA polymerases with RNHLS domain 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 choose 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 (Cerritelli and Crouch 2009). The Dis domain provides the ability for RNH to catalyze a variety of hybrids without unloading (Fig. 4) (Cerritelli and Crouch 2009). Based on this, most RNHLS proteins combine RNH domain with other domains to exercise their destroying RNA:DNA functions during different biological process. 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 plant, we first analysis their expression pattern in Arabidopsis and rice using RNA-seq data from database (Fig. 6 and Fig. S5). In in Arabidopsis and rice, various *RNHLS* genes expressed in different tissues are grouped together. In Arabidopsis, we find these genes are mainly expressed at meristem tissue, such as shoot and root meristem and inflorescence (Fig. 6A). Some of the genes are mainly expressed in anther and pistil (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 tissue due to programmed cell death are 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 inflorescence and pre-emergence inflorescence are grouped with totally different expression of *RNHLS* genes constant with Arabidopsis (Fig. S5). In the embryo-25 dap (days after plant), there are another group of *RNHLS* expressed (Fig. S5). Moreover, in seedling four-leaf stage, half of the investigated *RNHLS* expressed in this tissue (Fig. S5). Taken together, the expression of *RNHLS* genes in different groups are throughout Arabidopsis and rice development and constant 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 transcriptase,

transposases, integrase, nucleic acid endonuclease, exonuclease, key spliceosome components Prp8, as well as other RNA nucleases. Moreover, RNHLS domain is combined with various types of protein domains. 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) (Chang et al. 2016). The distribution of RNHLS proteins in all species of plants suggest they are one type of the evolutionarily oldest proteins (Ma et al. 2008). Among 27 clustered RNHLS proteins, DNA polymerase, GTF2B, CAF1, PRP8, RRP6L, AGO and EXO worked in basic nucleic acid metabolic process are conserved in plant kingdom (Majorek et al. 2014). WRN is able to digest the RNA moiety of the RNA:DNA hybrid (Suzuki et al. 1999). Recently, it is reported that WRN is involved in R-loop-related genome stability (Marabitti et al. 2019). 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 (Chen et al. 2018). 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 plant genome (Chang et al. 2016). 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 (Smyshlyaev et al. 2013; Moelling et al. 2017).

RNHLS protein is a large group of evolutionarily related, but strongly diverged protein family (Majorek et al. 2014). The first identified RNHLS protein is RNase H from *Escherichia coli* (Leis et al. 1973). 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 have the ability to remove RNA-DNA hybrids in addition to having different specialized roles. The phylogenetic analysis of RNHLS proteins in Arabidopsis show 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 show 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 (Song et al. 2004). 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 (Liu et al. 2004). 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 meristem, flower tissues and other metabolic tissues, and various *RNHLS* genes expressed in different tissues are grouped together indicating 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.

Conclusion

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 function. 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 evidence shed light on our understanding of R-loop regulation in plants.

Supplementary material

Supplementary material 1: Supplementary figure 1 to supplementary figure 5. (PDF 922 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 558 kb)

Supplementary material 6: Supplementary table 3. tissue-specific expression data of *RNHLS* genes of Arabidopsis and rice. (XLSX 187 kb)

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).

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

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215 (3):403-410. doi:10.1016/S0022-2836(05)80360-2

Ariyoshi M, Vassylyev DG, Iwasaki H, Nakamura H, Shinagawa H, Morikawa K (1994) Atomic structure of the RuvC resolvase: a holliday junction-specific endonuclease from *E. coli*. Cell 78 (6):1063-1072. doi:10.1016/0092-8674(94)90280-1

Cerritelli SM, Crouch RJ (2009) Ribonuclease H: the enzymes in eukaryotes. FEBS J 276 (6):1494-1505. doi:10.1111/j.1742-4658.2009.06908.x

Chan YA, Aristizabal MJ, Lu PY, Luo Z, Hamza A, Kobor MS, Stirling PC, Hieter P (2014) Genome-wide profiling of yeast DNA:RNA hybrid prone sites with DRIP-chip. PLoS Genet 10 (4):e1004288. doi:10.1371/journal.pgen.1004288

Chang C, Bowman JL, Meyerowitz EM (2016) Field Guide to Plant Model Systems. Cell 167 (2):325-339. doi:10.1016/j.cell.2016.08.031

Chen J, Liu L, You C, Gu J, Ruan W, Zhang L, Gan J, Cao C, Huang Y, Chen X, Ma J (2018) Structural and biochemical insights into small RNA 3' end trimming by Arabidopsis SDN1. Nat Commun 9 (1):3585. doi:10.1038/s41467-018-05942-7

Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43 (W1):W389-394. doi:10.1093/nar/gkv332

Dumelie JG, Jaffrey SR (2017) Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. Elife 6. doi:10.7554/eLife.28306

Fang X, Qi Y (2016) RNAi in Plants: An Argonaute-Centered View. Plant Cell 28 (2):272-285. doi:10.1105/tpc.15.00920

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28 (23):3150-3152. doi:10.1093/bioinformatics/bts565

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 45 (6):814-825. doi:10.1016/j.molcel.2012.01.017

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res (40) :D1178-1186. doi:10.1093/nar/gkr944

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

Huertas P, Aguilera A (2003) Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. Mol Cell 12 (3):711-721. doi:http://dx.doi.org/10.1016/j.molcel.2003.08.010

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res (37) :D211-215. doi:10.1093/nar/gkn785

Joyce GF (2002) The antiquity of RNA-based evolution. Nature 418 (6894):214-221. doi:10.1038/418214a

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30 (14):3059-3066. doi:DOI 10.1093/nar/gkf436

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J 88 (6):1058-1070. doi:10.1111/tpj.13312

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

Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44 (W1):W242-245. doi:10.1093/nar/gkw290

Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C, Lees J (2018) Gene3D: Extensive prediction of globular domains in proteins. Nucleic Acids Res 46 (D1):D435-D439. doi:10.1093/nar/gkx1069

Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ (2004) Argonaute2 is the catalytic engine of mammalian RNAi. Science 305 (5689):1437-1441. doi:10.1126/science.1102513

Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, Lahrmann U, Zhao Q, Zheng Y, Zhao Y, Xue Y, Ren J (2015) IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics 31 (20):3359-3361. doi:10.1093/bioinformatics/btv362

Ma BG, Chen L, Ji HF, Chen ZH, Yang FR, Wang L, Qu G, Jiang YY, Ji C, Zhang HY (2008) Characters of very ancient proteins. Biochem Biophy Res Co 366 (3):607-611. doi:10.1016/j.bbrc.2007.12.014

Majorek KA, Dunin-Horkawicz S, Steczkiewicz K, Muszewska A, Nowotny M, Ginalski K, Bujnicki JM (2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic Acids Res 42 (7):4160-4179. doi:10.1093/nar/gkt1414

Marabitti V, Lillo G, Malacaria E, Palermo V, Sanchez M, Pichierri P, Franchitto A (2019) ATM pathway activation limits R-loop-associated genomic instability in Werner syndrome cells. Nucleic Acids Res. doi:10.1093/nar/gkz025

Moelling K, Broecker F, Russo G, Sunagawa S (2017) RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense. Front Microbio 8:1745. doi:10.3389/fmicb.2017.01745

Nowotny M, Gaidamakov SA, Crouch RJ, Yang W (2005) Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. Cell 121 (7):1005-1016. doi:10.1016/j.cell.2005.04.024

Okonechnikov K, Golosova O, Fursov M, team U (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28 (8):1166-1167. doi:10.1093/bioinformatics/bts091

Parker JS, Roe SM, Barford D (2004) Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. EMBO J 23 (24):4727-4737. doi:10.1038/sj.emboj.7600488

Pei J, Kim BH, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36 (7):2295-2300. doi:10.1093/nar/gkn072

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26 (7):1641-1650. doi:10.1093/molbev/msp077

Rice PA, Baker TA (2001) Comparative architecture of transposase and integrase complexes. Nat Struct Biol 8 (4):302-307. doi:10.1038/86166

Santos-Pereira JM, Aguilera A (2015) R loops: new modulators of genome dynamics and function. Nat Rev Genet 16 (10):583-597. doi:10.1038/nrg3961

Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, Xu X, Chedin F (2016) Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. Mol Cell 63 (1):167-178. doi:10.1016/j.molcel.2016.05.032

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13 (11):2498-2504. doi:10.1101/gr.1239303

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7 (1):539. doi:10.1038/msb.2011.75

Smyshlyaev G, Voigt F, Blinov A, Barabas O, Novikova O (2013) Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. Proc Natl Acad Sci U S A 110 (50):20140-20145. doi:10.1073/pnas.1310958110

Song JJ, Smith SK, Hannon GJ, Joshua-Tor L (2004) Crystal structure of Argonaute and its implications for RISC slicer activity. Science 305 (5689):1434-1437. doi:10.1126/science.1102514

Stein H, Hausen P (1969) Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase. Science 166 (3903):393-395. doi:10.1126/science.166.3903.393

Suzuki N, Shiratori M, Goto M, Furuichi Y (1999) 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 27 (11):2361-2368

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

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z (2017) agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res 45 (W1):W122-W129. doi:10.1093/nar/gkx382

Pena V, Rozov A, Fabrizio P, Luhrmann R, Cwahl M (2008) Structure and function of an RNase H domain at the heart of the spliceosome. EMBO J 27:2929–40.

Wahba L, Costantino L, Tan FJ, Zimmer A, Koshland D (2016) S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. Genes Dev 30 (11):1327-1338. doi:10.1101/gad.280834.116

Xu W, Xu H, Li K, Fan Y, Liu Y, Yang X, Sun Q (2017) The R-loop is a common chromatin feature of the Arabidopsis genome. Nat plants 3 (9):704-714. doi:10.1038/s41477-017-0004-x

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

Figure legends

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 are marked in blue font. The original matrix of other protein domains that co-occurred with RNHLS domain are download from Gene3D (http://gene3d.biochem.ucl.ac.uk). The similarity heatmap are 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 are 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 is 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 are aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). The phylogenetic tree is constructed using the results from this alignment and modified using iTOL (https://itol.embl.de/). Different types of nuclease and polymerase are clearly separated to different clades. The distribution of transposon-related protein is more diversified than AGOs, SDNs and MEN.

Figure 4. Protein domain diagram of Arabidopsis representative RNHLS proteins.

The domain annotation of these proteins is 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 domain of each protein are selected for this alignment. Each of the RNHLS domain are

predicted the secondary structure using Jpred4 (http://www.compbio.dundee.ac.uk/jpred/index.html). The alignment is visualized using UGENE and adjusted manually.

Figure 6. Tissue-specific expression of Arabidopsis *RNHLS* genes.

(A) The RNA expression levels of RNHLS proteins are downloaded from TRAVA (http://travadb.org/browse/) except for some of the genes cannot retrieved from that database. These genes are mainly expressed at the meristem such as flower inflorescence and root meristem. The yellow rectangles indicate the clusters of genes highly expressed in these tissues. (B) GO annotation of the genes expressed in Arabidopsis meristem. These genes are mainly involved in nucleic acid metabolism.

Figure 1

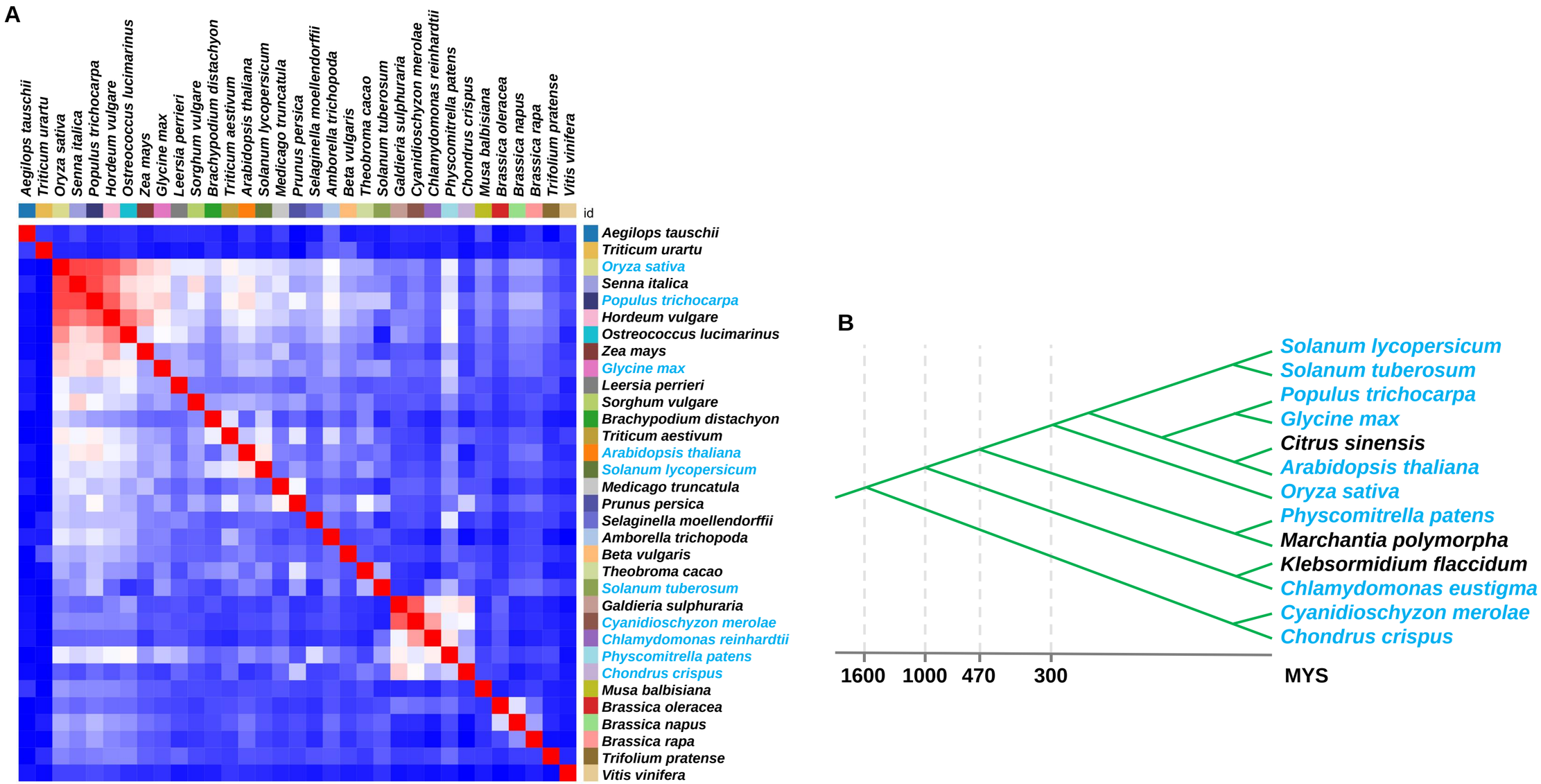

A
Aegilops tauschii
Triticum urartu
Oryza sativa
Senna italica
Populus trichocarpa
Hordeum vulgare
Ostreococcus lucimarinus
Zea mays
Glycine max
Leersia perrieri
Sorghum vulgare
Brachypodium distachyon
Triticum aestivum
Arabidopsis thaliana
Solanum lycopersicum
Medicago truncatula
Prunus persica
Selaginella moellendorffii
Amborella trichopoda
Beta vulgaris
Theobroma cacao
Solanum tuberosum
Galdieria sulphuraria
Cyanidioschyzon merolae
Chlamydomonas reinhardtii
Physcomitrella patens
Chondrus crispus
Musa balbisiana
Brassica oleracea
Brassica napus
Brassica rapa
Trifolium pratense
Vitis vinifera
id
B
Solanum lycopersicum
Solanum tuberosum
Populus trichocarpa
Glycine max
Citrus sinensis
Arabidopsis thaliana
Oryza sativa
Physcomitrella patens
Marchantia polymorpha
Klebsormidium flaccidum
Chlamydomonas eustigma
Cyanidioschyzon merolae
Chondrus crispus
1600
1000
470
300
MYS

Figure 2

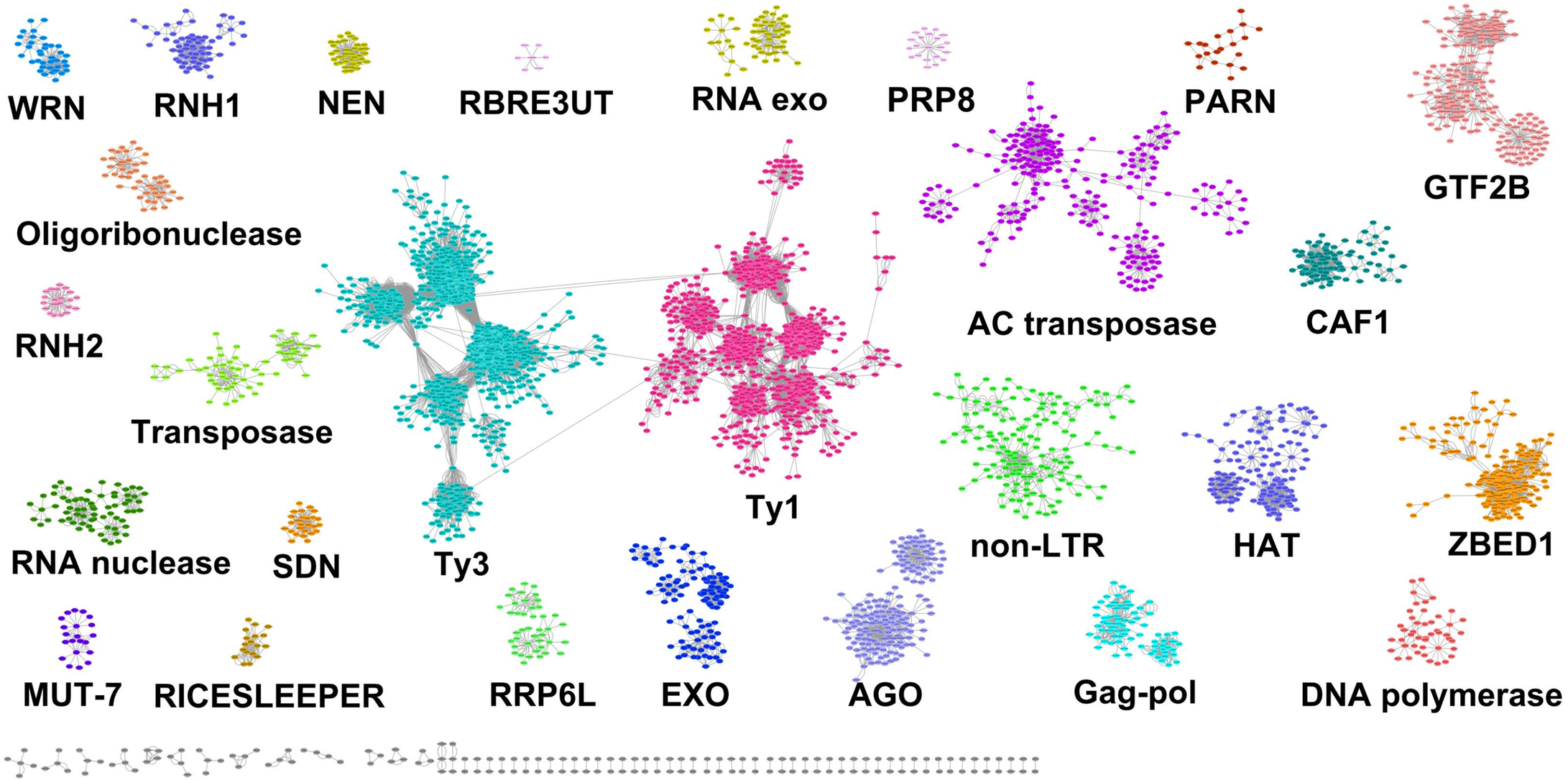

WRN
RNH1
NEN
RBRE3UT
RNA exo
PRP8
PARN
GTF2B
Oligoribonuclease
RNH2
Transposase
Ty3
Ty1
AC transposase
CAF1
RNA nuclease
SDN
non-LTR
HAT
ZBED1
MUT-7
RICESLEEPER
RRP6L
EXO
AGO
Gag-pol
DNA polymerase

Figure 3

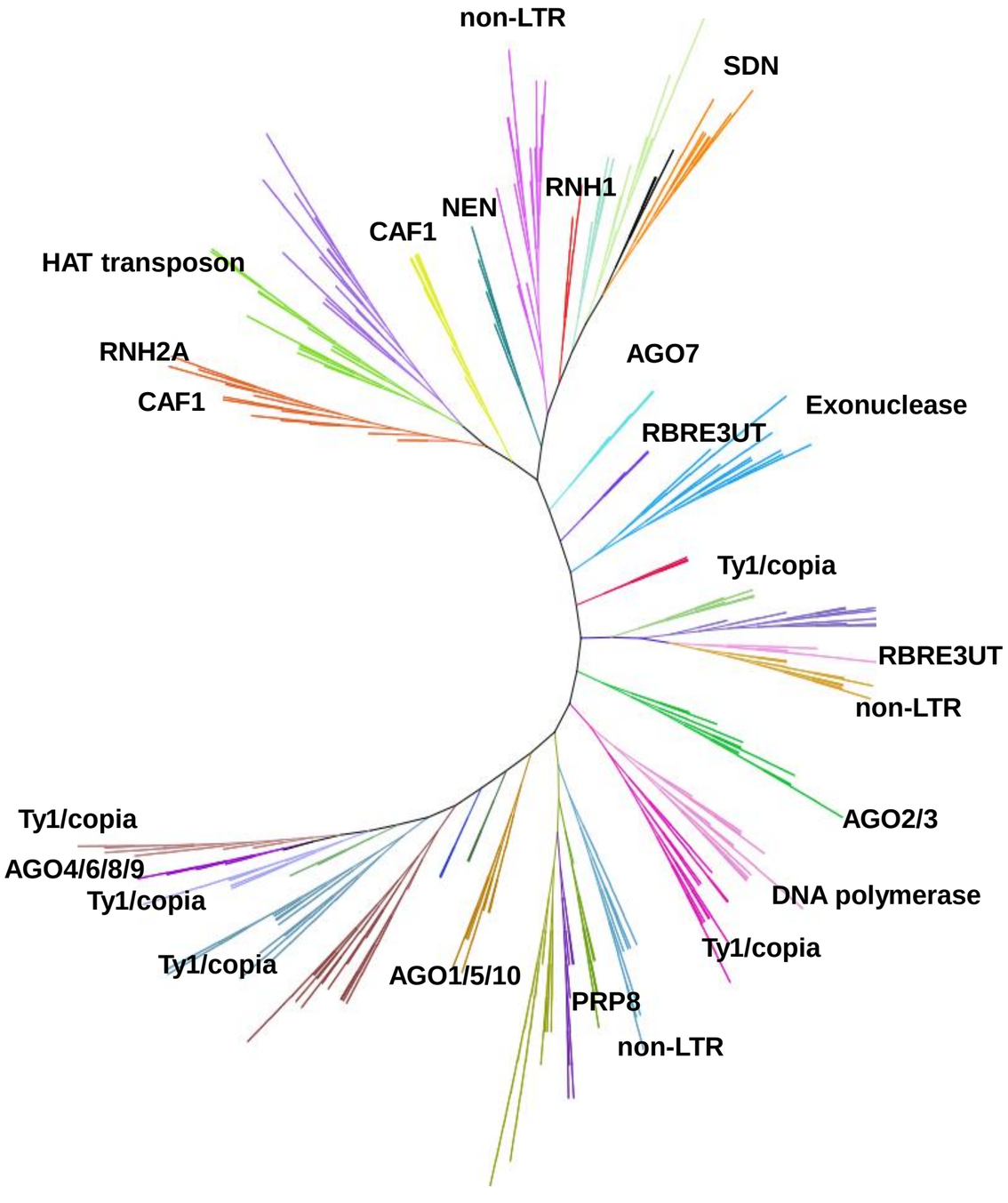

non-LTR
SDN
RNH1
NEN
CAF1
HAT transposon
RNH2A
CAF1
AGO7
Exonuclease
RBRE3UT
Ty1/copia
RBRE3UT
non-LTR
AGO2/3
DNA polymerase
Ty1/copia
Ty1/copia
AGO4/6/8/9
Ty1/copia
Ty1/copia
AGO1/5/10
PRP8
non-LTR

Figure 4

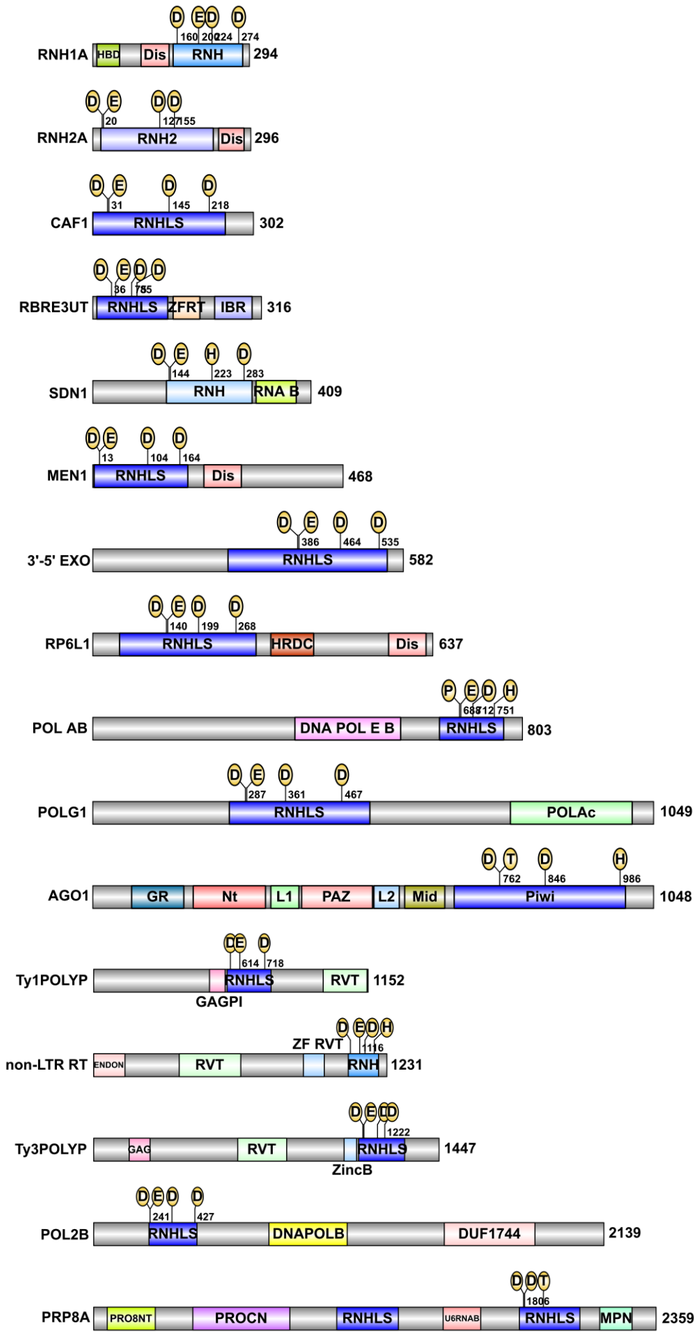

RNH1A
HBD
Dis
RNH
294
RNH2A
RNH2
Dis
296
CAF1
RNHLS
302
RBRE3UT
RNHLS
ZFRT
IBR
316
SDN1
RNH
RNA B
409
MEN1
RNHLS
Dis
468
3'-5' EXO
RNHLS
582
RP6L1
RNHLS
HRDC
Dis
637
POL AB
DNA POL E B
RNHLS
803
POLG1
RNHLS
POLAc
1049
AGO1
GR
Nt
L1
PAZ
L2
Mid
Piwi
1048
Ty1POLYP
RNHLS
RVT
GAGPI
1152
non-LTR RT
ENDON
RVT
ZF RVT
RNH
1231
Ty3POLYP
GAG
RVT
RNHLS
ZincB
1447
POL2B
RNHLS
DNAPOLB
DUF1744
2139
PRP8A
PRO8NT
PROCN
RNHLS
U6RNAB
RNHLS
MPN
2359

Figure 5

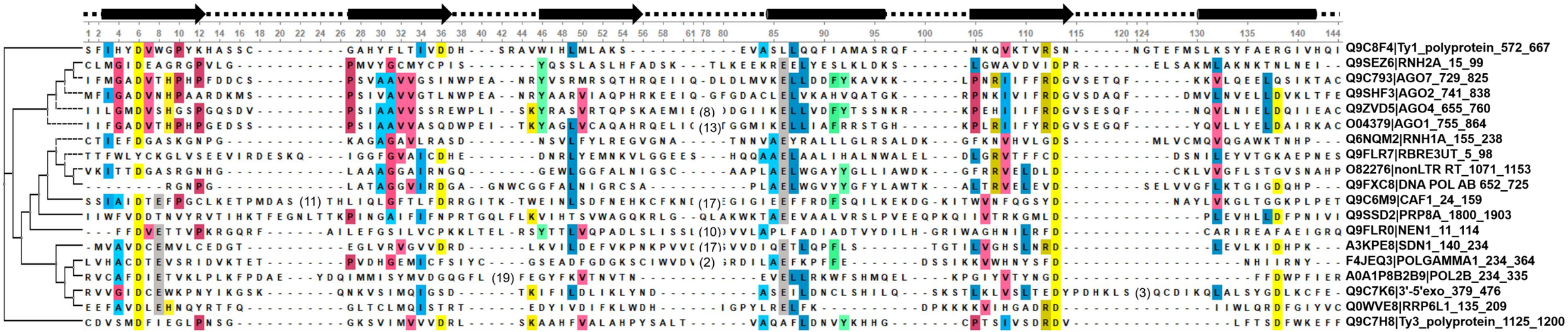
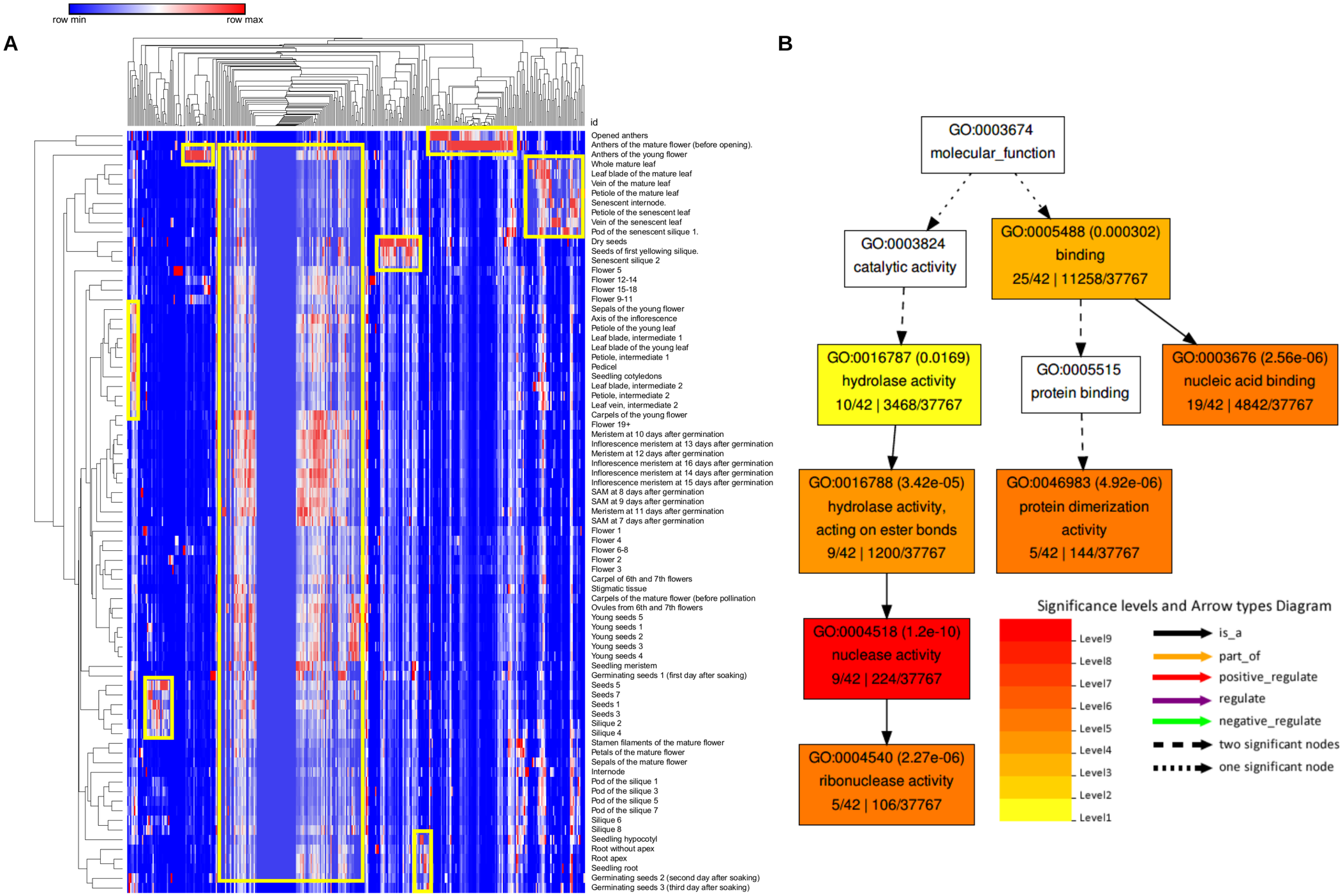

Q9C8F4|Ty1_polyprotein_572_667
Q9SEZ6|RNH2A_15_99
Q9C793|AGO7_729_825
Q9SHF3|AGO2_741_838
Q9ZVD5|AGO4_655_760
O04379|AGO1_755_864
Q6NQM2|RNH1A_155_238
Q9FLR7|RBRE3UT_5_98
O82276|nonLTR RT_1071_1153
Q9FXC8|DNA POL AB 652_725
Q9C6M9|CAF1_24_159
Q9SSD2|PRP8A_1800_1903
Q9FLR0|NEN1_11_114
A3KPE8|SDN1_140_234
F4JEQ3|POLGAMMA1_234_364
A0A1P8B2B9|POL2B_234_335
Q9C7K6|3'-5'exo_379_476
Q0WVE8|RRP6L1_135_209
Q9C7H8|Ty3_polyprotein_1125_1200

Figure 6


row min
row max
A
id
Opened anthers
Anthers of the mature flower (before opening).
Anthers of the young flower
Whole mature leaf
Leaf blade of the mature leaf
Vein of the mature leaf
Petiole of the mature leaf
Senescent internode.
Petiole of the senescent leaf
Vein of the senescent leaf
Pod of the senescent silique 1.
Dry seeds
Seeds of first yellowing silique.
Senescent silique 2
Flower 5
Flower 12-14
Flower 15-18
Flower 9-11
Sepals of the young flower
Axis of the inflorescence
Petiole of the young leaf
Leaf blade, intermediate 1
Leaf blade of the young leaf
Petiole, intermediate 1
Pedicel
Seedling cotyledons
Leaf blade, intermediate 2
Petiole, intermediate 2
Leaf vein, intermediate 2
Carpels of the young flower
Flower 19+
Meristem at 10 days after germination
Inflorescence meristem at 13 days after germination
Meristem at 12 days after germination
Inflorescence meristem at 16 days after germination
Inflorescence meristem at 14 days after germination
Inflorescence meristem at 15 days after germination
SAM at 8 days after germination
SAM at 9 days after germination
Meristem at 11 days after germination
SAM at 7 days after germination
Flower 1
Flower 4
Flower 6-8
Flower 2
Flower 3
Carpel of 6th and 7th flowers
Stigmatic tissue
Carpels of the mature flower (before pollination
Ovules from 6th and 7th flowers
Young seeds 5
Young seeds 1
Young seeds 2
Young seeds 3
Young seeds 4
Seedling meristem
Germinating seeds 1 (first day after soaking)
Seeds 5
Seeds 7
Seeds 1
Seeds 3
Silique 2
Silique 4
Stamen filaments of the mature flower
Petals of the mature flower
Sepals of the mature flower
Internode
Pod of the silique 1
Pod of the silique 3
Pod of the silique 5
Pod of the silique 7
Silique 6
Silique 8
Seedling hypocotyl
Root without apex
Root apex
Seedling root
Germinating seeds 2 (second day after soaking)
Germinating seeds 3 (third day after soaking)
B
GO:0003674
molecular_function
GO:0003824
catalytic activity
GO:0005488 (0.000302)
binding
25/42 | 11258/37767
GO:0016787 (0.0169)
hydrolase activity
10/42 | 3468/37767
GO:0005515
protein binding
GO:0003676 (2.56e-06)
nucleic acid binding
19/42 | 4842/37767
GO:0016788 (3.42e-05)
hydrolase activity,
acting on ester bonds
9/42 | 1200/37767
GO:0046983 (4.92e-06)
protein dimerization
activity
5/42 | 144/37767
GO:0004518 (1.2e-10)
nuclease activity
9/42 | 224/37767
GO:0004540 (2.27e-06)
ribonuclease activity
5/42 | 106/37767
Significance levels and Arrow types Diagram
Level9
Level8
Level7
Level6
Level5
Level4
Level3
Level2
Level1
is_a
part_of
positive_regulate
regulate
negative_regulate
two significant nodes
one significant node