Study of small non-coding RNAs in plants by developing novel pipelines

Dissertation

zur Erlangung des

Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät III

Agrar- und Ernährungswissenschaften,

Geowissenschaften und Informatik

der Martin–Luther–Universität Halle–Wittenberg,

vorgelegt von

Frau Deblina Patra Bhattacharya

Geb. am 21. August 1983 in Krishnanagar Nadia

Gutachter:
Prof. Dr. Ivo Große
Prof. Dr. Peter F. Stadler
Prof. Dr. Ivo Hofacker

Datum der Verteidigung: 27.04.2017

Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

Signature: _____

Date: _____

Nowadays, high throughput sequencing technologies have become essential in studies on genomics, epigenomics, and transcriptomics since they are capable of sequencing multiple DNA molecules in parallel, enabling hundreds of millions of DNA molecules to be sequenced at a time. This is a great advantage which allows HTS to be used to create large data sets, generating more comprehensive insights into the cellular genomic and transcriptomic signatures of various diseases and developmental stages.

Small non-coding RNAs make up much of the RNA content of a cell and have the potential to regulate gene expression on many different levels. And it is now possible for the sake of high-throughput sequencing techniques to assay an organism's entire repertoire of small non-coding RNAs (ncRNAs) in an efficient and cost-effective manner.

Due to the moderate size of small RNA-seq datasets, it is convenient and feasible to provide free web servers to the research community that provide many basic features of a small RNA-seq analyses, including quality control, read normalization, ncRNA quantification, and the prediction of putative novel ncRNAs. We introduced such web server `plantDARIO` in order to provide comprehensive analysis for plant small non-coding RNAs (sncRNAs) which includes major modifications to cope with plant-specific sncRNA processing.

During analysis of small non-coding RNAs, small nucleolar RNAs (snoRNAs) are found to be the most ancient as well as conserved families amongst non-protein-coding RNAs. SnoRNAs are ubiquitous in Archaea and Eukarya but are absent in bacteria. Their main function is to target chemical modifications of ribosomal RNAs. They fall into two classes, box C/D snoRNAs and box H/ACA snoRNAs, which are clearly distinguished by conserved sequence motifs and the type of chemical modification that they govern. And like other small non-coding RNAs, in animals, snoRNAs and their evolution have been studied in much detail.

However, very little attention is paid to the plant snoRNAs. In order to chart the phylogenetic distribution of individual snoRNA families in plants, a sophisticated approach for identifying homologs of known plant snoRNAs across the plant kingdom is applied and we identified 296 families of snoRNAs in 24 species and traced their evolution throughout the plant kingdom.

Many of the plant snoRNA families comprise paralogs. The sequence conservation of snoRNAs is sufficient to establish homologies between phyla. The degree of this conservation tapers off, however, between land plants and algae. It is also found

that targets are well-conserved for most snoRNA families and plant snoRNAs are frequently organized in highly conserved spatial clusters.

Since the snoRNAs are evolutionary ancient as well as conserved, it is speculated that novel snoRNAs if predicted and provided they are not false predictions, then they are also conserved in more than one species. In this context we applied `plantDARIO` server to find novel snoRNAs from publicly available small RNA-seq dataset and studied their phylogenetic distribution using the same sophisitcated approach for identifying homologs of known plant snoRNAs across the plant kingdom.

We intended to find how the novel predicted are distributed amongst the 24 species in the plant kingdom. We find 11 novel snoRNA families classified into 9 box C/D snoRNA families and 2 box H/ACA families along with their targets.

Words would never be sufficient to express my gratitude to the people I name here. However, I make a humble effort to bring those names together and also re-iterate that without them this work would not be successfully done.

I thank my advisors Peter Stadler and Ivo Grosse for being an amazing positive influence on my ambitions. I thank Peter for having the patience to listen through my numerous project related problems and offering instant tips to resolve them and Ivo for being such a supportive advisor. I am grateful for all the motivation and support. This work was funded by Deutsche Forschungsgemeinschaft grant no. GR 3526/2 and JU 205/19, under the auspices of the Priority Program 1530 "Flowering Time Control – from Natural Variation to Crop improvement". I really learned a lot from the Seminars and Symposiums conducted by Priority Program 1530 "Flowering Time Control – from Natural Variation to Crop improvement".

I thank all my colleagues in Leipzig and Halle for all the support and help. I am thankful to David and Mario for the initial help and support to introduce me to the world of "small non-coding RNAs". Thanks to Jana, Steffi and Sebastian for the support in finishing my work. A big thanks to Maribel, Bia, Haleh, Heni, Rozin for making life in Leipzig so much fun and easier. Thanks to Lydia, Petra for always solving my German-document related problems and Ulf for the fun with the free German lessons. Big thanks to Ioana, Claus, Hajk, Alex, Martin and Ralf for all the discussions and help.

I thank Maa, Baba, Bhai for motivating me and remaining by my side always. Thanks to Subhamoy for being my inspiration and Mummy, Baba, Didibhai, Pramitda, Bittu, Arka for the emotional support. Thanks to the almighty GOD for listening to my prayers and helping to fulfil them.

Contents

List of Figures

List of Tables

Abbreviations

ncRNAs	**N**on **C**oding **RNA**s
snc-RNA	**S**mall **N**on **C**oding **RNA**
miRNAs	**M**icro **RNA**
tRNA	**T**ransfer **RNA**
mRNA	**M**essenger **RNA**
rRNA	**R**ibosomal **RNA**
dsRNA	**d**ouble **s**tranded **RNA**
siRNA	**S**mall **I**nterfering **RNA**
phasiRNA	**P**hased **S**mall **I**nterfering **RNA**
easiRNA	**E**pigenetically **A**ctivated **S**mall **I**nterfering **RNA**
piRNA	**P**iwi **I**nteracting **RNA**
rasiRNA	**R**epeat **A**ssociated **S**mall **I**nterfering **RNA**
tasiRNA	**T**rans **A**cting **S**mall **I**nterfering **RNA**
natsiRNA	**N**natural **A**n-**T**isense **S**mall **I**nterfering **RNA**
hcsiRNA	**H**etero **C**hromatic **S**mall **I**nterfering **RNA**
21URNA	**21**21-mer with 5 **U**ridine **RNA**
qiRNA	**Q**DE2 **I**nteracting small **RNA**
DCL	**D**icer **L**ike
snRNA	**S**mall **N**uclear **RNA**
snoRNA	**S**mall **NO**ucleolar **RNA**
RDR	**RNA D**ependent **RNA**-polyemerase
AGO	**Argonaute**
HTTP	**HT**ypertext **T**ransfer **P**rotocol
HTML	**HT**ypertext **M**arkup **L**anguage

RdDM	**R**NA **d**irected **D**NA **M**ethylation
RdDM	**R**NA **d**irected **D**NA **M**ethylation
HYL1	**HY**PONASTIC **L**eaves **1**
PTGS	**P**ost **T**ranscriptional **G**ene **S**ilencing
RISC	**RNA I**nduced **S**ilencing **C**omplex
LSU	**L**arge **S**ub **U**nit
SSU	**S**mall **S**ub **U**nit
TEs	**T**ransposable **E**lements
TUT	**T**erminal **U**ridylyl **T**ransfer
K-turn	**K**ink-turn
NCBI	**N**ational **C**entre for **B**iotechnology **I**nformation
Rfam	**R**NA **fam**ily database
ePoPE	**e**fficient **P**rediction **o**f **P**aralog **E**volution

Dedicated to Maa and Baba

Chapter 1

Introduction

Since the time of late 1800s, RNA is known, but its importance in cell functioning has long been not discovered. During the period of 1950s, when the DNA molecular structure was established, and from that time RNA is proposed to be an intermediate molecule in the information flux between DNA and proteins. And then later, experimental demonstration revealed that during gene expression, DNA is copied in a molecule of messenger RNA (mRNA) that is then translated into proteins with the help of other RNA molecules like transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) [1]. The thought that RNAs are much more than molecules involved in storage and transfer of information emerged with the discovery of ribozymes, RNA molecules that have, like proteins, active roles as catalysts of chemical reactions in cells.

The two ribozymes identified first have RNAs as substrates and are the Tetrahymena intron of the 26S rRNA that is a self-sufficient catalytic unit capable of autoexcision and autocyclization [2], and the ribonucleoprotein, RNase P, an enzyme containing an RNA subunit essential for the catalysis required for the synthesis of tRNAs [3]. These discoveries clearly encouraged a variety of studies to search for potential new roles of RNA molecules in vivo, and led to the re-evaluation of RNAs as crucial molecules in the evolution of life. In view of the ability of RNAs to catalyze biological reactions, it is conceivable that the first organisms could rely only on RNA molecules and that only later an evolution of a more complex system based on proteins is established. This hypothesis gave support to the model of a primordial "RNA World" [4, 5].

1.1 Emergence of Non-coding RNAs

Studies and research hinted towards the existence of non-coding RNAs much longer before the non-coding RNA revolution. Back in the early period, evidences were derived from the the labs of Phil Sharp [6] and of Louise Chow, Tom Broker, and Rich Roberts [7] from studies of adenoviral early mRNAs from the labs of Phil Sharp and of Louise Chow, Tom Broker, and Rich Roberts which showed that final mRNAs can be formed from the transcript stretches arising from distinct and distant portions of the viral genome (exons) when pieced together. This led to the question whether any jumping RNA polymerases are responsible for such results but then appar- ently the hnRNAs are found to be full-length transcripts with the excision and discard of intron sequences which also explains that newly synthesized RNA documented for mammalian cell nuclei have been hugely wasted.

But all these findings directed to the single question that what kind of cellular machinery could be responsible behind all these actions? Finally, the evidence of non-coding RNAs (ncRNAs) answered of all these questions. Earlier studies had uncovered the presence of small (100–300 nt) highly abundant U-rich RNAs in the nuclei of vertebrate cells [8, 9] and led to the discovery of non-coding RNAs.

Soon after the discovery of ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) in 1950s, the central roles of these principal RNA participants in gene expression and protein synthesis is firmly established [2].

Non-coding RNAs (ncRNAs) are functional RNA molecules that are transcribed from DNA but are not translated into proteins (see in 1.1)(adapted from [10]) In general ncRNAs function to regulate gene expression at the transcriptional and post-transcriptional level, getting involved in the chain process of central dogma from transcription to splicing to translation and contributing to genome organization and stability.

Non-coding RNAs also play role in RNA editing events e.g. nucleotide exchanges or very small (1–3 nt) insertion/deletions within an RNA transcript. It is found that mRNAs editing in the mitochondria of kinetoplastid protozoa can result in alteration of as many as 50% of the coding nucleotides. Short (40–80 nt) guide (g)RNAs actively participate in base pairing with the editing sites to direct the action of endonucleases and U-specific exonucleases or TUTases (terminal uridylyl transfer- ases), executing the deletion or insertion of U residues [11].

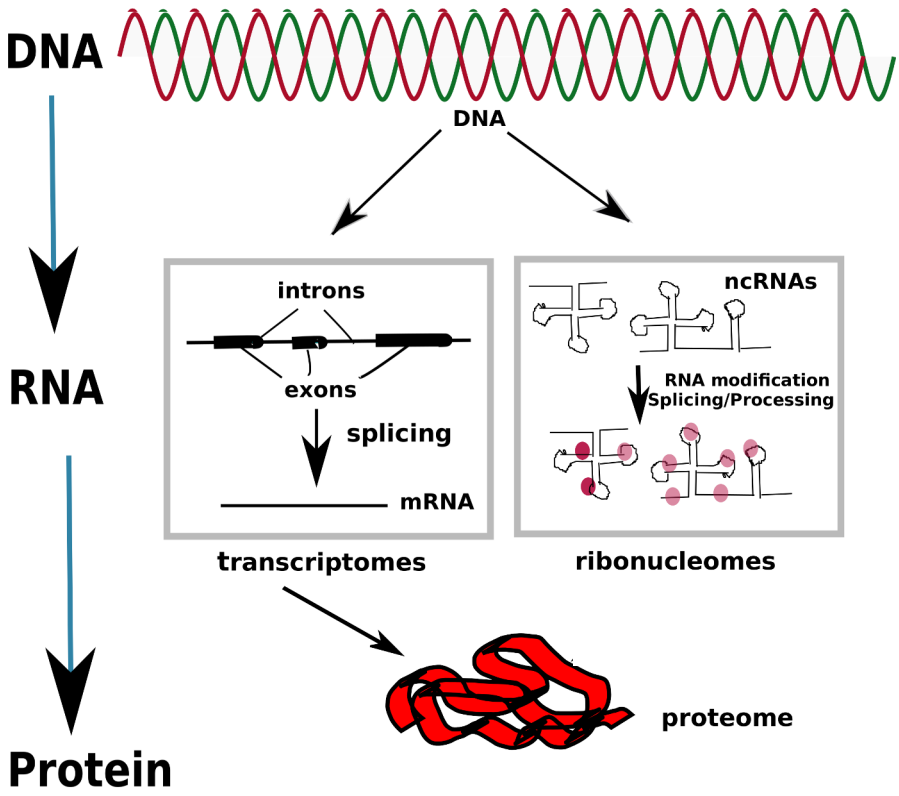


FIGURE 1.1: From genes to non-coding RNAs

The ncRNAs can be divided into many classes depending on their size and function. The short length small non-coding RNAs belong to one of the classes of the ncRNAs.

1.2 Small non-coding RNAs

The discovery of small ncRNAs in recent years has shaken the world of RNAs, bringing to the limelight very tiny yet powerful regula- tors of gene expression. The characteristic features of small ncRNAs are their short length (20–30 nt) and their function in regulating gene expression. In addition to these defining features, different classes of small ncRNAs guide diverse as well as complex schemes of gene regulation [12]. Amongst the small non-coding RNAs siRNAs, snRNAs, rRNAs, tRNAs, miRNAs and snoRNAs are the small RNAs which are commonly known and are the focus of the study (especially the last two).

Since the literature keeps on growing in case of small ncRNAs, various newer acronyms, such as piRNA (Piwi-interacting RNA), rasiRNA (repeat-associated

siRNA), tasiRNA (trans-acting siRNA), natsiRNA (natural an-tisense transcript siRNA), hcsiRNA (heterochromatic siRNA), scnRNA (small scan RNA), 21U RNA (21-mer with 5 uridine), and qiRNA (QDE2-interacting small RNA) have also found their place in the literature and many more yet to be added in the near future.

1.3 Plant small non-coding RNAs

The universe of plant sncRNAs (small non-coding RNAs) is much more complex and diverse than its counterpart in animals. Longer approximately or perfectly doublestranded RNA (dsRNA) precursors are cut by Dicerlike (DCL) proteins into small RNA duplexes [13].

So, these small RNA duplexes are produced intitially and then, later, one strand from the initial duplex becomes associated with an Argonaute(AGO) protein. Interestingly, the small RNAs are diversified based on the duplication of genes encoding DCLs and RNA-dependent RNA polymerase (RDRs) [14, 15].

The AGOs are diverse in types as well, which result in the development of distinct gene-silencing processes depending on differential AGO affinities to small RNAs [16]. The small RNAs bound to AGOs hybridize the target RNAs and upon successful paring, the AGO protein directly catalyzes the repressive activities on the target. The repression of the AGO-organized target can take place at the levels of repressive chromatin modifications, decreased RNA stability, and lowered translational efficiency. Hence, the AGObound smallRNA ensemble in any specific plant is considered a reservoir of negative regulators of specific sequences [13].

The small RNA duplexes can be loaded onto different classes of Argonaute (AGO) proteins present in complexes of different functions that mediate the interaction of the incorporated smRNAs with their targets. For e.g. AGO1 acts mainly in microRNA (miRNA) pathways for post-transcriptional gene silencing (PTGS) [17]. In case of miRNA duplexes, while the guide strands are incorporated into AGO1 of the RNA-induced silencing complex (RISC), the passenger strands called miRNA star (miRNA*) are mostly degraded [18]. Small RNAs loaded onto other Argonaute-containing complexes have different functions, e.g. heterochromatin maintenance.

The role of small RNAs not only in plant development but in reproduction and genome reprogramming marked the snoRNAs to be very significant in plants. Even the phenotypic plasticity in plants is contributed by the large variety of small-RNA pathways in plants and it is now known and accepted that these pathways have evolved as cellular defence mechanism against RNA viruses and transposable elements [19]. Later, these pathways are adapted to regulate the expression of endogenous genes.

In contrast to animals, all the plant small RNAs are modified at the 3'-end by 2' O methylation, since 2' O methylation seems essential conferring stability and pro tection from 3' -uridylation and degradation. MiRNAs especially in plants are generally involved in post-transcriptional gene silencing (PTGS) by transcript cleavage or translational repression and might trigger secondary siRNA produc tion from RNA polymerase II (Pol II)derived cleaved transcripts. It has become quite evident that many small RNAs are involved in PTGS. However, the majority of siRNAs in plants are associated with RNA-directed DNA methylation (RdDM) and transcriptional gene silencing (TGS)[13].

1.3.1 Classification of plant small non-coding RNAs

The diversity of plant DCL/AGO small RNAs based primarily on their distinct modes of biogenesis can be described by hierarchical classification (see in 1.2, adapted from [13]). As described, the double-stranded RNA (dsRNA) precursors are cut by Dicer-like (DCL) proteins into small RNA duplexes and further processed into different types of small RNAs [13]. The precursors of siRNAs consist of dsRNA molecules (see [20] for a recent review) whereas less heavily structured single-stranded RNAs serve as the precursors of microRNAs [21].

It is a fact that most small RNA classes generally have significant role in defence responses as well as in epigenetic regulation. Although relative importance and overlap varies from plant to plant species, however the consistency persists amongst all plants [13].

As the details of the precursors of small RNAs are considered, an elementary division seems to emerge between small RNAs derived from doublestranded precursors, generally formed by the intermolecular hybridization of two complementary RNA strands and the small RNAs derived from single stranded precursors that acquire

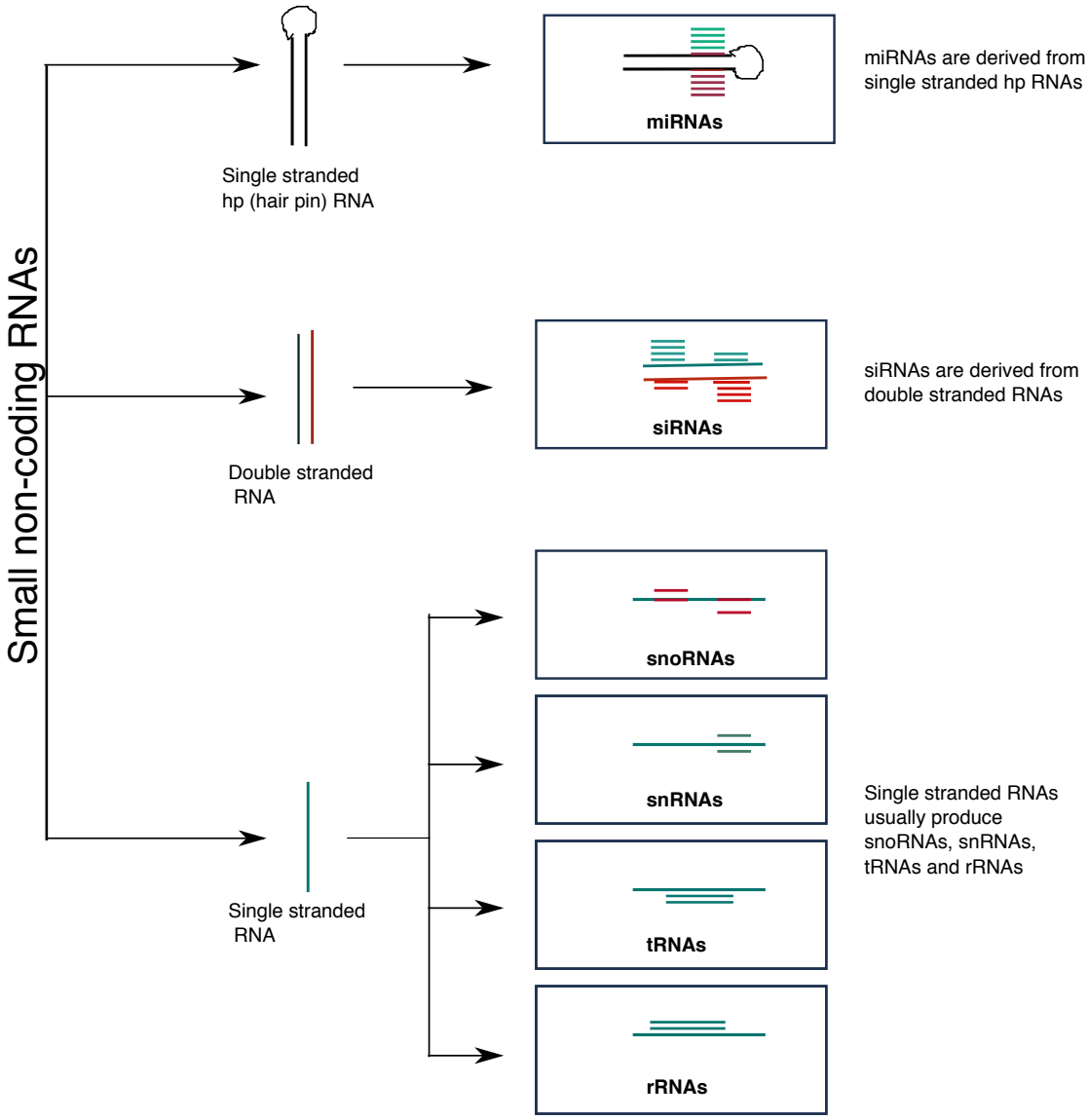


FIGURE 1.2: miRNAs, siRNAs, snoRNAs, snRNAs, rRNAs and tRNAs

an intramolecular, "hairpin" structure (fig:heirarchy) which is self complementary [13].

Small RNAs derived from double-stranded RNA (dsRNA) precursors are referred as small interfering RNAs (siRNAs), whereas the "hairpin" singe stranded structures give rise to microRNAs (miRNAs) and the small derived from the single-stranded precursors are again catergorized into: small nucleoar RNAs (snoRNAs), small nuclear RNAs (snRNAs), transfer RNAs (tRNAs) and rRNAs (ribosomal RNAs).

The precursors of siRNAs consist of dsRNA molecules (see [20] for a recent review)

rather than more or less heavily structured single-stranded RNAs that serve as the precursors of microRNAs [21]. The small RNA duplexes can be loaded onto different classes of Argonaute (AGO) proteins present in complexes of different functions that mediate the interaction of the incorporated smRNAs with their targets. For e.g. AGO1 acts mainly in microRNA (miRNA) pathways for post-transcriptional gene silencing (PTGS) [17]. In case of miRNA duplexes, while the guide strands are incorporated into AGO1 of the RNA-induced silencing complex (RISC), the passenger strands called miRNA star (miRNA*) are mostly degraded [18]. Small RNAs loaded onto other Argonaute-containing complexes have different functions as for example heterochromatin maintenance.

In plants, even more extensive groups of sncRNAs have been described, comprising in addition a variety of distinct types of small interfering RNAs (siRNAs) such as trans-acting siRNAs (ta-siRNAs), natural antisense siRNAs (nat-siRNAs), and double-strand break interacting RNAs (diRNAs) [22–25]. Heterochromatic (hc-)siRNAs are the most abundant class of small RNAs in many plants. The transcripts yielding hc-siRNAs are transcribed by the plant-specific RNA polymerase IV and enter the RNA-directed DNA methylation (RdDM) pathway, comprising first the synthesis of dsRNA by RDR2 and subsequent cleavage by DCL3. The resulting 24 nt long hc-siRNAs are then bound to AGO4 [26]. In contrast to miRNAs whose genomic loci are conserved between species, hc-siRNAs genomic loci are not, because they overlap with transposable elements (TEs), which are known to rapidly change their position and copy number in the genomes during plant evolution [13].

Analyses on small RNAs are discussed in more details in chapter 2, which includes transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs(snRNAs), small interfering RNAs (siRNAs), especially more about microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs). And plant snoRNAs are vividly studied, described and analyzed in chapter 3.

1.3.2 Transfer RNAs (tRNAs)

Transfer RNA (tRNA) plays a very important role in translation of genetic information into proteins, and understanding its molecular evolution is important if we are to understand the genetic code. Small RNAs of about 76-90 nucleotides termed transfer RNAs (tRNAs), act as adaptor molecules that physically link

the nucleotide sequence containing genetic information to an amino acid during protein synthesis [27]. In the reaction catalyzed by aminoacyl-tRNA synthetase (aaRS), only one type of amino acid is attached to each type of tRNA [28].

The tRNAs are stucturally different in variable regions and based on that tRNAs can be classified into two groups. Short variable region of 4–5 nucleotides known as class I tRNAs, whereas class II tRNAs have a long variable-arm (V-arm) structure containing ten or more nucleotides [29], and are therefore also called "V-arm-containing tRNAs." In class II category, tRNA(Leu), tRNA(Ser), and bacterial and organellar tRNA(Tyr) are classified in class II, and all other tRNA isotypes are classified in class I. The V-arm structure generally play an important role as a recognition site for the aaRSs during the aminoacylation of class II tRNAs [30–33]. The V-arms of tRNALeu and tRNASer are found to be conserved in all organisms, however the V-arms of archaeal and eukaryotic tRNA(Tyr) seem to be lost soon after the separation of these domains from the domain Bacteria [34].

Nuclear or cytosolic compartments are the storage for most of the proteins required by the plant chroloplasts or mitochondria, including the aminoacyl-transfer RNA synthetases. However, the plant choloroplasts retain all tRNA genes and plant mitochondria retain most tRNA genes needed for translation. The tRNAs encoded by chloroplast and mitochondria mostly resemble their prokaryotic counterparts and show very little homology to cytosolic species. Therefore, a plant cell contains a variety of different tRNAs following classical structural rules [35].

The nuclear tRNA genes in plants as well as in animals and yeast found to exist as multi gene families that are either scattered throughout the genome or found to be clustered at single chromosomal sites. The first situation is illustrated in tobacco nuclear tRNATyr genes where the minimal number of individual tRNA(Tyr) gene loci appears to be about 14 [36].

The presence of "cytosol-like" tRNAs in plant mitochondria served as the initial evidence for tRNA import from cytosol. The genetic origin of the mitochondrial tRNA population is diverse in case of plants, which is a striking feature in plants. This underlines the complexity of both genetic information transfers between the plant cell compartments and mitochondrial gene divergence during evolution [35].

1.3.3 Ribosomal RNAs (rRNAs)

The riosomes are generally smaller in prokaryotes when compared to the eukaryotes, with a sedimentation coefficient of 70 Svedberg units (abbreviated as 70S), while eukaryote ribosomes have a sedimentation coefficient of 80 Svedberg units (80S). Like the other higher eukaryotes, the nuclear RNA genes (rDNA) in higher plants are arranged in long tandem repeating units.

The importance of ribosomal RNAs (rRNAS) is found to be very evident in evolutionary biology, since the ribosomal RNA is considered the most conserved (least variable) gene in all cells, the role of rRNAs in evolutionary biology cannot be neglected [37]. Therefore in order to identify an organism's taxonomic group, calculate related groups, and estimate rates of species divergence, the genes that encode the rRNA (rDNA) are sequenced. As a consequence, thousands of rRNA sequences are known and stored in specialized databases such as RDP-II [38] and the European SSU database [39].

The mature rRNAs are produced by processing of the pre-RNA, which sometimes requires a number of snoRNAs and nucleolar proteins. The ribosomal RNAs generally complex with proteins to form large subunit (LSU) and small subunit (SSU), and between the small and large subunits there is mRNA, and the ribosome catalyzes the formation of a peptide bond between the two amino acids that are contained in the rRNA. A ribosome also has three binding sites called A, P, and E.

While the ribosomal subunits are quite similar between prokaryotes and eukaryotes, the 70S ribosomes contain proportionally more RNA than protein, while the 80S ribosomes are composed of less RNA than protein [37]. In comparing the two subunits themselves, the proportions of rRNA and protein are approximately equal. The spacer has been most frequently designated the "non-transcribed spacer". Plants generally have more rRNA genes than the other groups of organisms [40].

The 70S ribosomes have three different types of rRNA: 23S rRNA, 16S rRNA, and 5S rRNA. There are four different types of rRNA in 80s ribosomes: 28s rRNA (but 25-26S rRNA in plants, fungi, and protozoans), 18S rRNA, 5S rRNA, and 5.8S rRNA. The 3' end of the 16S rRNA (in a ribosome) binds to a sequence on the 5' end of mRNA called the Shine-Dalgarno sequence. The 18S rRNA in most

eukaryotes is in the small ribosomal subunit, and the large subunit contains three rRNA species (the 5S, 5.8S and 28S rRNAs) [41].

1.3.4 microRNAs (miRNAs)

A group of sRNAs originating from endogenous loci and regulating other target RNAs are the micro RNAs (miRNAs), which are a well-studied subset of hpRNAs defined by the highly precise excision of one or sometimes a few functional products [42].

Generally, plant miRNAs are 21–22 nt long mediating gene silencing at post-transcriptional level. The major functions of miRNAs are entailing endonucleolytic cleavage (slicing) and transla- tional repression of a target mRNA, and the miRNAs often have a defined set of mRNA tar- gets. It is found that individual miRNA families can be conserved over long evolutionary distances [43].

Although many animal miRNAs are derived from introns or untranslated regions of coding messages or primary transcripts containing tandem precursors [44, 45] most plant miRNA- encoding loci comprise independent, non-protein- coding transcription units [46]. However, there are some known examples of transcripts harboring tandem precursors in plants [47, 48] and precursors located in mRNA untranslated regions.

Most plant miRNAs are generated from their own transcriptional units. The miRNA primary transcripts contain an in- ternal stem–loop secondary structure (miRNA pre- cursor) with the miRNA located in one of the arms.

Recognized by the miRNA processing machinery, structural determinants in the miRNA precursors and produces staggered cleavages in the dsRNA, sepa- rated 21 nt of each other. These cuts release the miRNA together with the opposing fragment of the precursor that is interacting with it, called miRNA* as in 1.3 (adapted from [42]).

The miRNA processing in Arabidopsis is dependent on DICER-LIKE1 (DCL1) machinery, which is the RNAse type III that produces all cuts in the miRNA precursors [49, 50]. DCL1 acts together with the dsRNA-binding protein HYPONASTIC LEAVES1 (HYL1) [51, 52] and the C2H2 zinc-finger protein SERRATE (SE) [53, 54]

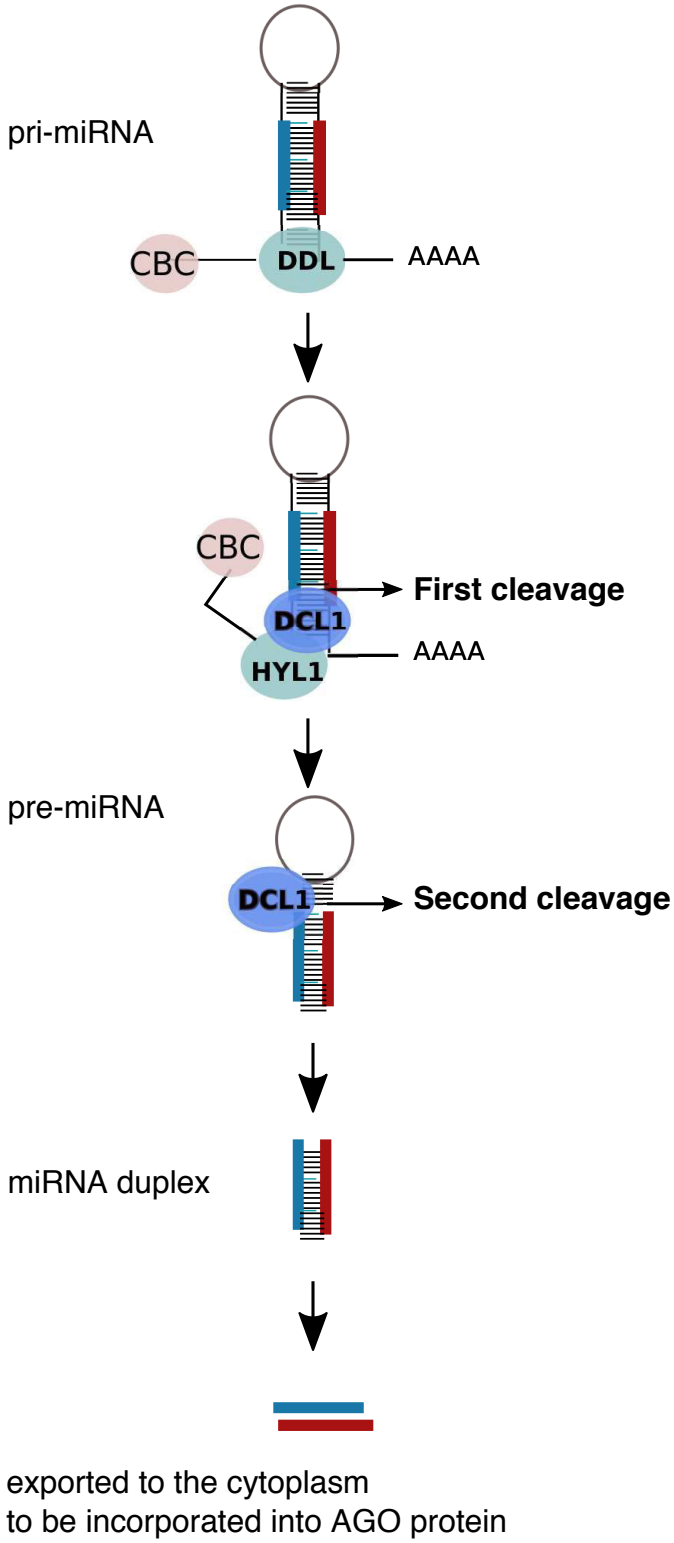


FIGURE 1.3: Biogenesis of plant microRNA

It is presumed that DAWDLE, a DCL1 interacting protein stabilizes miRNA primary transcripts until they are processed by DCL1 [55]. There is a homolog of the animal Exportin5 named HASTY which contributes to the levels of certain miRNAs [56].

In case of animals, where HASTY as Exportin5 transports animal pre-miRNAs to the cytoplasm, whereas the molecular role of HASTY is not clear in plants since all molecular processing steps occur within plant nucleus. However, HASTY might be associated with other cargo, such as the miRNA/miRNA*. After the processing of the pre- cursors, the miRNA/miRNA* duplex is released [57].

When compared to plants, miRNA biogenesis is com- partmentalized in animals. The primary transcripts are first trimmed in the nucleus to separate the stem-loop precursor from the rest of the transcript and the process is generally achieved by a Microprocessor complex. This complex is generally formed by an RNase III-like enzyme termed Drosha and the dsRNA-binding protein DiGeorge syndrome critical region gene 8 (DGCR8; Pasha in Drosophila melanogaster and Caenorhabditis elegans) [51, 58].

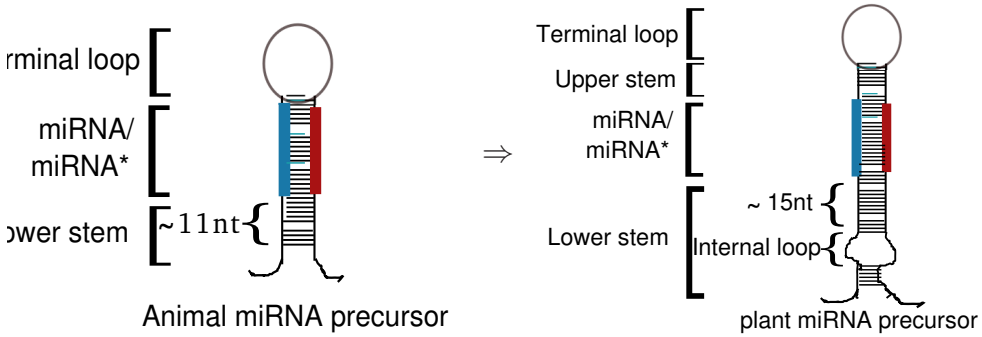
In general, the processing of the fold-back precursors by the RNAse III complexes causes the release of miRNA/miRNA* duplexes (2.3), where miRNA is incorporated into an AGO complex, whereas the miRNA* is generally degraded.

The main important point is the precision in the position of the cuts along the precursor as they determine the sequence of the miRNA and therefore its target specificity. The selection of the position for the first cleavage reaction is of special importance because the second cut is usually performed by measuring a fixed distance from the end of the precursor [42].

In some cases, the biogenesis of the miRNAs could proceed by other pathways e.g. cases have been found where recently evolved miRNAs have been shown to depend on DCL4 rather than DCL1 in *Arabidopsis thaliana*[59].

Some examples are also found where the activity of Drosha can be bypassed by the splicing machinery, generat- ing mirtrons [60, 61], which have been also described in plants [62, 63]. However, the biogenesis of most animal and plant miRNAs is generally canalized through Drosha and DCL1, respectively [42].

Plant microRNAs are much longer compared to animal miRNAs [43] (see in 2.3), adapted from [42]



(A) Animal microRNA precursor (B) Plant microRNA precursor

FIGURE 1.4: Difference of structural determinants in animal and plant miRNA precursors

The typical miRNA precursor in animals generally comprises a stem of approximately three helical turns (miRNA/miRNA* duplex plus a lower stem), a terminal loop and long ssRNA flanking sequence (1.4a), whereas the plant miRNA precursor seems to be different.

A plant miRNA precursor could be divided into four parts comprising a lower stem, the miRNA/ miRNA*, an upper stem and the terminal loop. Many plant miRNA precursors have a lower stem of 15 nt below the miRNA/miRNA* that is followed by a large bulge [42](1.4b).

1.3.5 Small nuclear RNAs (snRNAs)

Previous studies had uncovered the presence of small (100–300 nt) highly abundant U-rich RNAs in the nuclei of vertebrate cells [8, 9] and named as small nuclear RNAs (snRNAs), which mainly form the core spilceosome, sometimes also responsible for pre-processing of mRNA. These snRNAs are discovered to associate tightly with a set of proteins that are targets of autoantibodies (anti-Sm) found in patients with lupus, forming so-called Sm snRNPs [64].

The Sm proteins are now known to be related to Hfq, which binds multiple ncRNAs in bacteria [65]. Failure to assemble snRNAs with due to a deficit in the cellular assembly factor SMN (survival of motor neurons) leads to a devastating disease atleast in case of animals [66].

Hence the spliceosomal RNAs U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac grouped together with SRP RNA and RNase MRP RNA are together grouped into the class “snRNAs” and discussed more in chapter 2.

1.3.6 Small interfering RNAs (siRNAs)

smallRNA derived from DCL-catalyzed processing of dsRNA i.e. small-RNA targeting of an initial primary transcript leads to recruitment of a RNA dependent RNA polymerase (RDR) leading to synthesis of the complementary RNA strand and ultimately processing of the resulting dsRNA into secondary siRNAs as mentioned previously [13].

Small-RNA targeting of an initial primary transcript leads to recruitment of an RDR, synthesis of the complementary RNA strand, and processing of the resulting dsRNA into secondary siRNAs. The three major clades of eukaryotic RDRs are RDRα, RDRβ and RDRγ; the RDRα clade is present in the fungal, plant and animal kingdoms, whereas RDRβ has been found in only animals and fungi and RDRγ in only plants and fungi [14, 67]. Generally, RDR genes are found in RNA viruses, plants, fungi, protists and some animals, but found to be absent in flies, mice and humans [14].

The endogenous siRNAs in plants are primarily processed by DCL2, DCL3 and DCL4, and have been categorized into secondary siRNAs. Secondary siRNAs include different subclasses, such as trans-acting siRNAs (tasiRNAs), phased siRNAs (phasiRNAs), epigenetically activated siRNAs (easiRNAs) and natsiRNAs.

Amongst these, the most abundant small RNAs are the 24nucleotide heterochromatic siRNAs (hetsiRNAs), which mediate transcriptional silencing of transposons and pericentromeric repeats through RNA-directed DNA methylation (RdDM) [68, 69]. Heterochromatic siRNAs are generally very consistent with requirements for specific mem- bers of the RDR, DCL, and AGO gene families. They depend specifically on RDR2 and DCL3 for their biogenesis [70, 71] and on members of the AGO4 clade of AGOs (AGO4, -6, and -9 in Arabidopsis) for their function. Whereas in case of most heterochromatic siRNAs, they depend on an alternative DNA-dependent RNA polymerase, Pol IV, for their biogenesis [72].

1.3.7 Small nucleolar RNAs (snoRNAs)

Amongst the ever-increasing number of families of small RNAs are small nucleolar RNAs (snoRNAs) which represent an abundant class of 50-300 nucleotide trans-acting RNAs in all eukaryotes [73] generally involved in RNA metabolism and gene

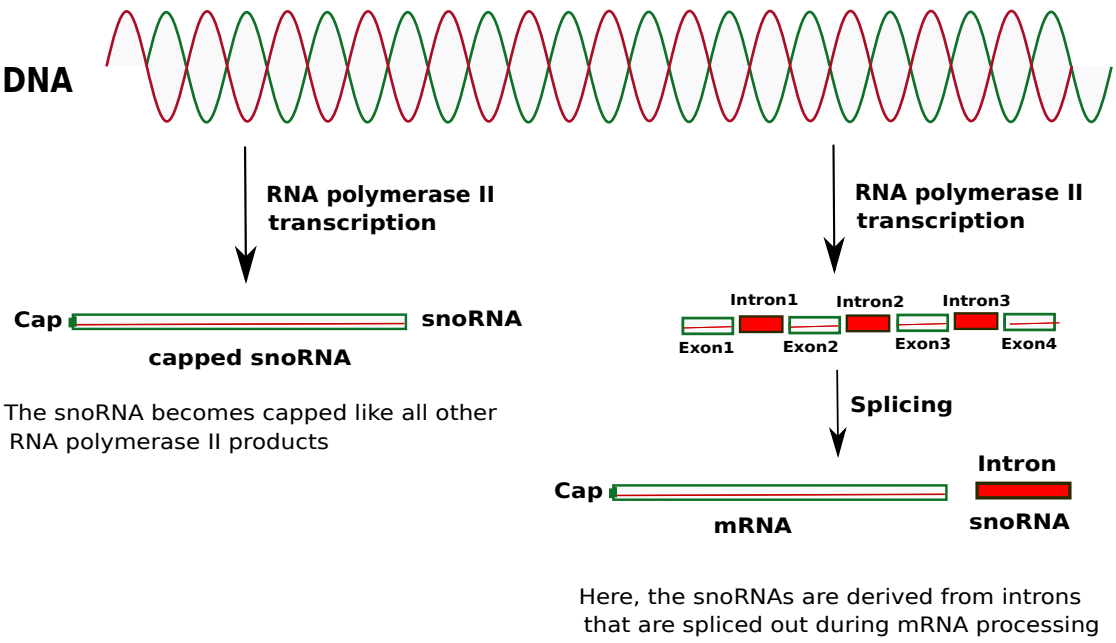


FIGURE 1.5: Synthesis of snoRNA

expression in eukaryotes.

snoRNAs are found to be synthesised by two major mechanisms: i.e. the snoRNAs can either be directly transcribed by RNA polymerase II or produced through splicing. In the first case the snoRNA becomes capped like all other RNA polymerase II products whereas in the latter, the snoRNAs are derived from introns that are spliced out during mRNA processing (1.5 adpated from [1])

Although the snoRNAs are absent in bacteria, but are present in archae, where they are named as sRNAs (for small RNAs) stating the ancient origin of snoRNAs [74]. snoRNAs are composed of two subclasses, C/D snoRNAs (1.6, adapted from [75]) and H/ACA snoRNAs (1.7, adapted from [75]), both of which have been shown to function as guides in site-specific RNA modification [76, 77]. Mature snoRNAs are produced by processing of a pre-snoRNA that can be polycistronic, intronic or monocistronic [78].

A key factor in the processing of polycistronic pre-snoRNA in yeast, is Rnt1p, an RNase III endonuclease which cleaves the pre-snoRNA and liberates the individual snoRNA with 3' and 5' extensions. These extensions are eliminated by exonuclease activities [79], whereas the mature snoRNA ends are protected by assembly of snoRNP core proteins [80].

[1] https://www.nobelprize.org/educational/medicine/dna/a/translation/snorna_bio.html

In vertebrates, mature snoRNAs are mainly produced from introns of precursors that can be both protein-coding mRNAs or non-coding "host genes." In contrast, only a few snoRNAs are intronic in budding yeast and plants [78, 81]. Moreover, the loss of introns through widespread degeneration of splicing signals has lead to snoRNA host genes that carry snoRNAs as exons in yeast. [82].

Plant pre-snoRNAs have been detected in Cajal bodies, supporting a role in processing [38]. In addition, there is the unexplained observation of the accumulation of plant snoRNAs in the nucleolar cavity (a central, transcriptionally inactive region of plant nucleoli).

There is a tendency for polycistronic snoRNA precursors in general. In plants, however, polycistronic precursors are the standard [83–85]. A curious exception are the tRNA(Gly)-snoRNA and tRNA(Met)-snoRNA cotranscripts in dicots and monocots, respectively [86].

In general, individual snoRNAs are excised from the precursor by RNase III endonucleases and are then trimmed by exonucleases [79, 87]. The mature snoRNA ends are protected by assembly of snoRNP core proteins [80] from further degradation.

1.3.7.1 Box C/D snoRNAs

The box C/D snoRNAs share the conserved sequence motif C (`RUGAUGA`) close to the 5'-end and D (`CUGA`) near the 3'-end, which are tethered by a terminal stem-loop (1.6). In addition, internal C' and D' box can be found in many of the box C/D snoRNA. They have the same consensus sequence as C and D box, but show more variance.

The spacing between the box C/D and D'/C' motifs is highly conserved in archaeal box C/D snoRNAs. The spatial positioning of the two constituent RNPs within the sRNP complex is critical for nucleotide modification activity. The terminal box C/D core motif is comprised of boxes C and D, whereas internally located D' and C' sequences usually fold to form the D'/C' motif. Both motifs often form a kink-turn (K-turn) motif, which is typified by two tandem-sheared G:A pairs hydrogen-bonding across the asymmetric bulge [88–90].

In yeast and vertebrates, box-C/D snoRNAs associate with Snu13p, Nop56p, Nop58p and Nop1p (fibrillarin in animals and plants) [91, 92]. The core snoRNP

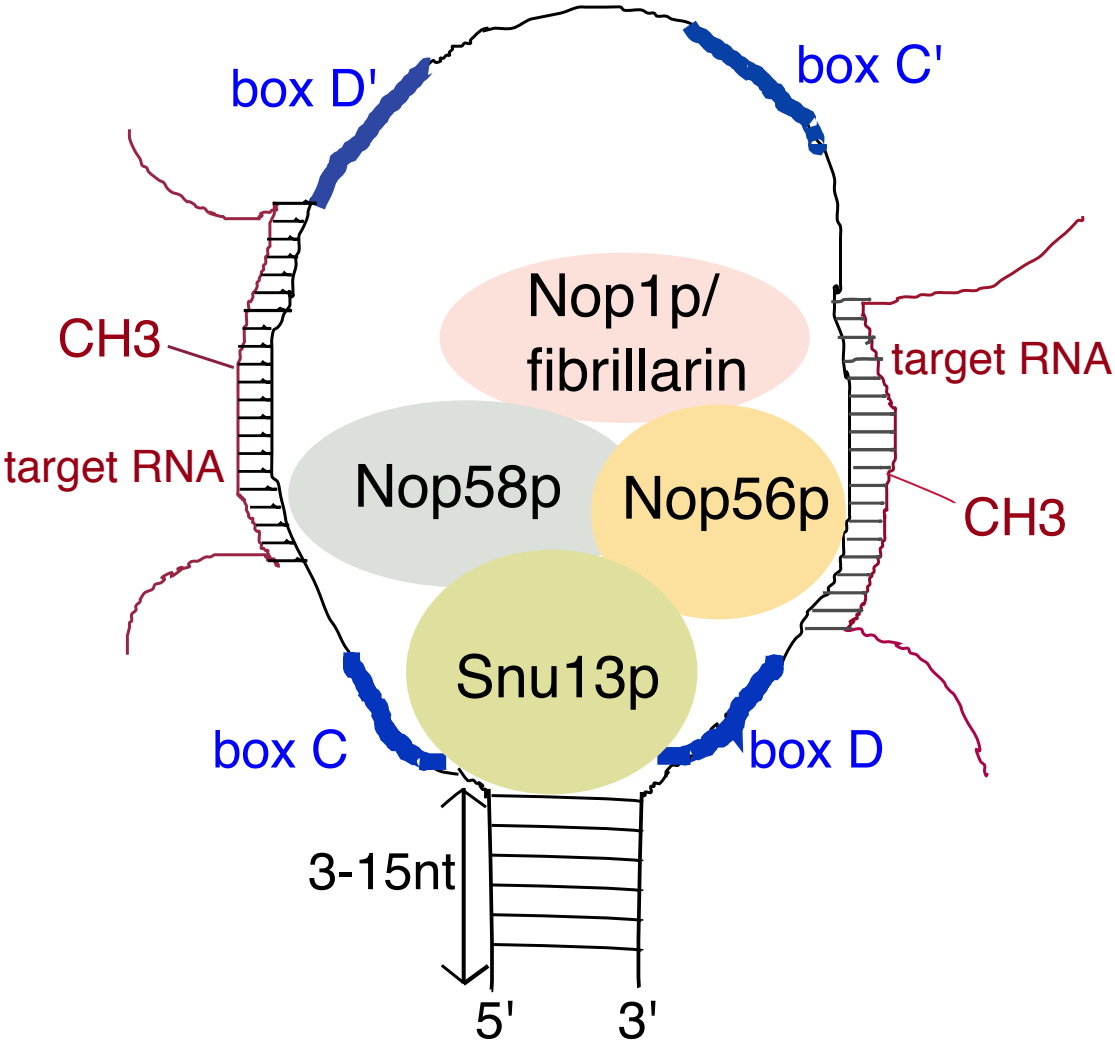


FIGURE 1.6: Box C/D snoRNA

proteins in case of plants are highly conserved and have all been found in the nucleolar proteome from Arabidopsis [93].

Although the plant box-C/D snoRNAs are found similar to their metazoan and yeast counter- parts in size and structure but, in alignments, only the box-C and -D elements, and the RNA-complementary regions are found higly conserved [83].

Generally, the assembly of C/D snoRNPs is initiated by the binding of 15.5K protein to the kink–turn fold of C/D boxes. This 15.5K is the only core snoRNP protein directly interacting with the snoRNA which is followed by the recruitment of the three other core C/D snoRNP proteins. This is a process that involve ther factors [88, 94].

Experimental validation identified NUFIP (nuclear Fragile X mental retardation protein- interacting protein; Rsa1 in yeast) as a central protein which directs this process. It is found that NUFIP interacts directly with the 15.5K protein serving as a bridge to recruit the other core proteins [95]. NUFIP is also found to be implicated in the assembly of U4 snRNP that contains the 15.5K protein and other nuclear RNP-containing proteins from the L7Ae family, such as NHP2 from H/ACA snoRNPs [95].

1.3.7.2 Box H/ACA snoRNAs

The box H/ACA snoRNAs are distinguished by the presence of an `ACA` triplet at their 3'-end and a characteristic hairpin-hinge-hairpin-tail secondary structure with the H box (`ANANNA`) located in the hinge region [96, 97] (1.7).

The H/ACA snoRNAs associate with dyskerin/NAP57 (Cbf5p in yeast), which is the pseudouridine synthase, NHP2, NOP10 and GAR1 [74, 77, 83]. Yeast and vertebrate box-H/ACA snoRNAs associate with Cbf5p (NAP57 in vertebrates), Gar1p, Nhp2p and Nop10p [81, 92]. In these complexes, Nop1p/fibrillarin is the rRNA methylase and Cbf5/Nap57 is the rRNA pseudouridine synthase [98].

In the case of H/ACA the snoRNP assembly is directed by a different set of proteins, including NAF1 (nuclear assembly factor 1), a key assembly factor both in vertebrates and yeast (Table 1). NAF1 interacts with dyskerin/NAP57 (Cbf5p in case of yeast) and subsequently recruits the other H/ACA core proteins. Most importantly, NAF1 binds to the CTD (C-terminal domain) of Pol II, and the assembly of H/ACA snoRNPs is tightly coupled to transcription. NAF1 binds to one of the core proteins, NAP57, and then shuttles between nucleus and cytoplasm. Both proteins are equally essential for stable H/ACA RNA accumulation. It is found that NAF1 and GAR1 bind NAP57 competitively, suggesting a sequential interaction [99]. It is repoted that the assembly factor Naf1p and the core components Cbf5p and Nhp2p are generally recruited during early period of transcription on H/ACA snoRNA genes. It is known that cotranscriptional recruitment of Naf1p and Cbf5p is Ctk1p dependent and that Ctk1p and Cbf5p are required for preventing the readthrough into the snoRNA downstream genes and also known that proper cotranscriptional snoRNP assembly controls 3'-end formation of snoRNAs [99–101].

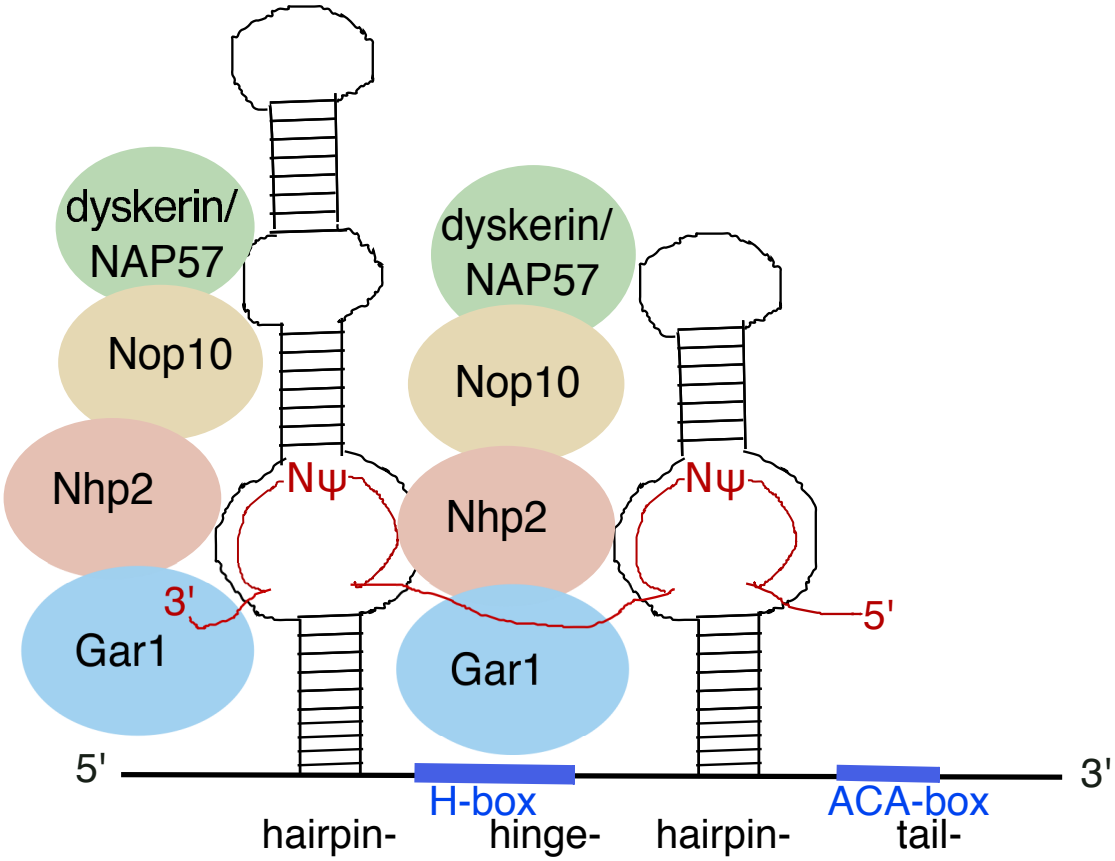


FIGURE 1.7: Box H/ACA snoRNA

1.3.7.3 Functional roles of snoRNAs

Beyond their function as guides for chemical modifications, a few snoRNAs are required for the cleavage of the ribosomal RNA precursors [102], among them in particular the U3 and the U14 snoRNAs. As it is known that the abundant U3 and U14 snoRNAs are required for 18S rRNA production like the unique RNase MRP snoRNA (which is involved in 25S –28S rRNA production), are conserved in all eukaryotes including plants.

In contrast to the modification guides, these snoRNAs are essential for cell survival in human and yeast. They are also ubiquitously present throughout eukaryotes [91, 103, 104]. Some snoRNAs are involved in regulating gene expression, e.g. by modulating mRNA splicing or editing [74, 77].

More recently, snoRNAs have also been identified as a source of miRNA-like small RNAs that function in mRNA silencing found in diverse organisms from archaea

to humans [105, 106]. SnoRNAs have even been found to be important players in cancer, suggesting that they fullfil multiple additional function in cellular regulation [107, 108].

Prior to the divergence of Archaea from eukaryotes, the common ancestors of snoRNAs already contained multiple snoRNA genes [109, 110]. By studying the evolution of snoRNAs in Archaea, yeast and vertebrates, it seems to have occurred through a repeated series of duplications, mutations and selections for their ability to associate into stable snoRNPs and to influence ribosome assembly and function [92, 93].

Owing to the prevalence of polyploidy in plants, there is a high degree of gene duplication and potential gene redundancy in plant snoRNAs, providing more opportunity to accumu- late mutations. Thus, plant snoRNA genes provide a useful model for observing mechanisms playing important role in gene evolution [83].

1.3.7.4 Plant Phylogeny

For the global biodiversity, climatic change is believed to be one of the major factors which is responsible [111, 112]. Aging towards life, world's climatic fluctuations have most likely caused major extinctions [113] leading to the development of new ecosystems and new biotic interactions are promoted leading to the evolution of novel adaptive traits. And diversification of dynamic events can be studied through phylogenetic trees and their detail analysis.

Diverse life forms are exhibited by the green plants or termed Viridiplantae, which are actually a clade of perhaps 500,000 species [114–116] also including some of the smallest and largest eukaryotes [117]. According to fossil evidence the clade is at least 750 million years old, whereas the molecular data estimated divergence time suggest that it may be more than one billion years old [118]. But reconstructing the phylogenetic relationships across green plants is really challenging not only because of the age of the clade but also the extinction of major lineages and extreme molecular rate and compositional heterogeneity [119, 120]. Two well supported subclades of Viridiplantae, Chlorophyta and Streptophyta [121] have been recovered from most phylogenetic analyses. Chlorophyta contain comprises of "green algae," and Streptophyta contain the land plants (Embryophyta), as well as

several other lineages also considered "green algae". Land plants include the seed plants or the flowering plants (gymnosperms and angiosperms; Spermatophyta), which consist of 270,000 to 450,000 species [114].

Generally the broad analyses of green plant relationships based on nuclear gene sequence data have been largely dependent on 18S/26S rDNA sequences [122, 123]. Recent reserach and analyses have also employed numerous nuclear genes which are involved. Studies have often used mitochondrial gene sequence as well as oher other data [124], but the green plant phylogeny research is largely dependent on the employed chloroplast genes (e.g.,[124, 125]). It has now also came to the picture that sequence data too from the plastid genome have been an important and valuable resource for transformed plant systematics. It has greatly contributed to the current view of plant relationships.

Hence it's a fact that detail study of phylogenetic tree from the aspect of non-coding RNAs will also lead to more interesting facts and informations. This phylogenetic tree (1.8)(created through taxonomy browser, consisting of some clades from the phylogenetic tree mentioned in [126]) seems to cover the almost all the major clades from algae to flowering plants.

In chapter 3, this phylogentic tree (1.8) is used to study the phylogenetic distribution of plant snoRNAs leading to interesting evolutionary facts and informations of plant snoRNA families.

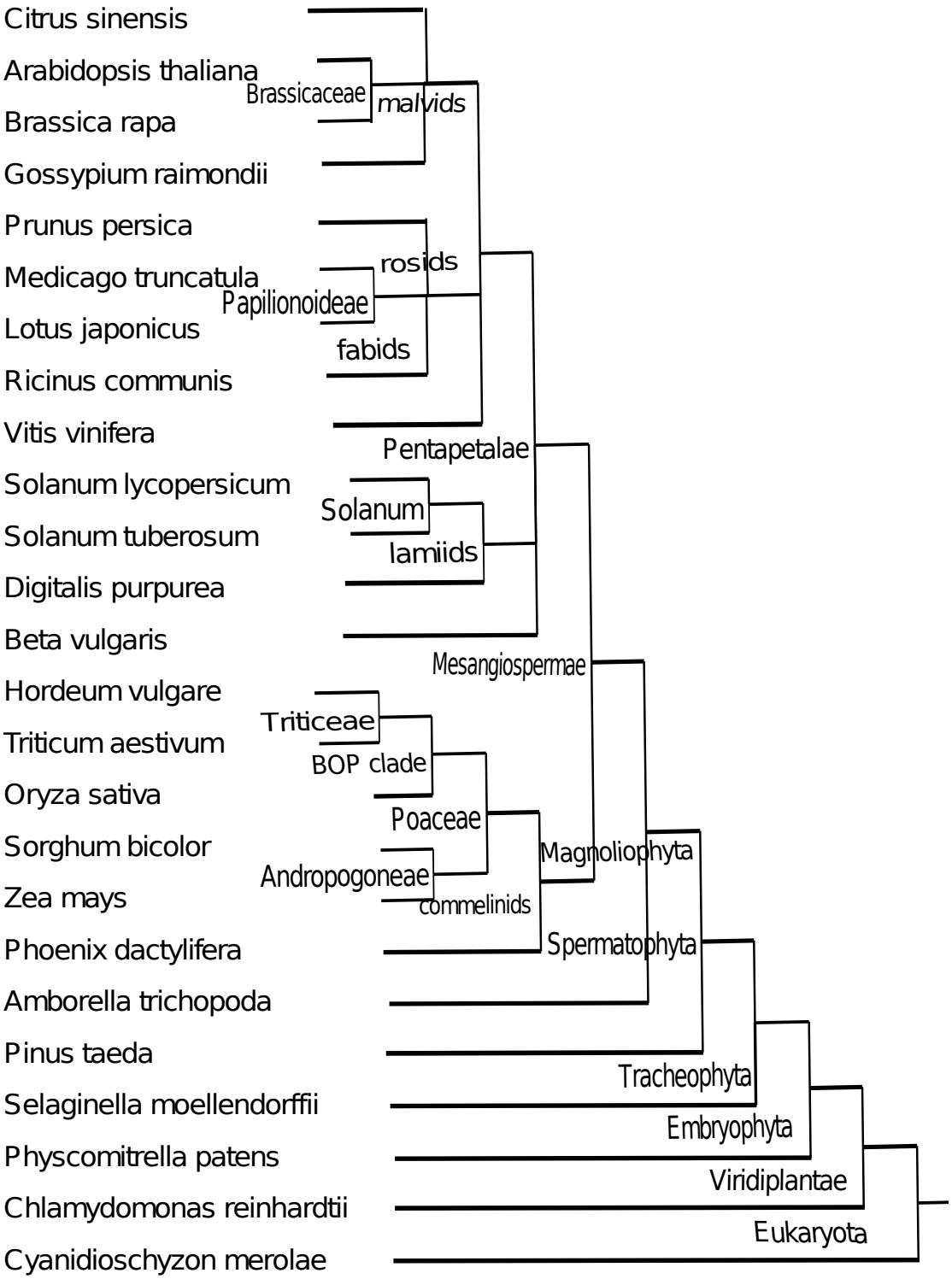


FIGURE 1.8: Phylogenetic tree covering major clades

Chapter 2

plantDARIO web-server for analyzing small non-coding RNAs

In the previous chapter, we learned about the plant small non-coding RNAs and their classification. In this chapter we are going to learn about the analysis of small RNA-seq data, including quality control, read normalization, ncRNA quantification, and the prediction of putative novel ncRNAs with the implementation of plantDARIO web-server.

2.1 Background of creating plantDARIO web-server

As discussed in the previous chapter, the universe of plant sncRNAs is more complex and diverse than its counterpart in animals and it has been found that plant sncRNAs from seedlings upto inflorescences have been shown to have a broad range of biological functions in the model plant *Arabidopsis thaliana* [127].

As revealed in case of animals after detailed analyses of small RNA-seq samples, which were primarily produced with the aim of measuring miRNA expression [128, 129], are actually derived from virtually all of the housekeeping ncRNAs including tRNAs [130, 131], snoRNAs [132, 133], and snRNAs [134, 135], as well as from many previously undescribed genomic loci including promoters and transcriptional termini of most protein-coding genes [136].

As discussed in case of plants, even more extensive groups of sncRNAs have been described, comprising in addition a variety of distinct types of small interfering

RNAs (siRNAs) such as trans-acting siRNAs (tasiRNAs), natural antisense siRNAs (natsiRNAs), and double-strand break interacting RNAs (diRNAs) [22–25]. Heterochromatic (hc-)siRNAs are the most abundant class of small RNAs in many plants.

A large array of computational tools has been developed and published for the analysis of the RNA-seq data, mainly focusing on the prediction and quantification of sncRNA genes for e.g `ShortStack` [137], `mirDeep` [138], and others. There are also tools like `PsRobot` [139] that combine plant small RNA annotation and target analysis, while `psRNATarget` [140] and `SoMART` [141] are mostly concerned with target prediction. While `miRanalyzer` and `omiRas` are the only webtools installed and run locally, requiring more than basic computer skills. In case of `CPSS` and `PsRobot`, the data needs to be formated to fasta format manually. The other sncRNA prediction tools need to be downloaded, installed and run locally, requiring more than basic computer skills.

The main drawback of all these tools are the integrated adapter clipping and read mapping steps. Given the differences in the performance of read mappers, in particular regarding sequences mapping multiple times and the handling of mismatches arising from polymorphisms [142] or editing [143]. These can be actually problematic since different library preparations and sequencing runs which result in sequencing data that should be handled independently, therefore it is desirable that the researcher should use the tools of his/her choice. Furthermore, the sheer size of the raw sequencing data (several gigabyte) compared to their mapping coordinates (some megabyte) and abundances suggests the conclusion, that for a web-tool mapping coordinates are the upload format of choice.

In 2011, `DARIO` a webserver was introduced for the analysis of small RNA-seq data in animals was introduced [144] which is designed to perform quality control of input samples along with expression analyses of annotated and user-defined sncRNAs, as well as a prediction of new non-coding RNAs. The main feature is that it provides exploratory analyses for mapped, but also unannotated reads. Keeping in mind all the features of this web-service, we have modified this versatile web server into a version which is specifically tailored to plants. And accordingly implemented the needed modiciations in the workflow.

Since plant pre-miRNAs are much more heterogeneous than their animal counterparts and have a different distribution of genomics contexts in which they reside

[44–46], hence they are more difficult to annotate [145]. In contrast to most animals, plant genomes (with the exception of *Arabidopsis thaliana*) are poorly annotated for ncRNAs and thus a careful and manual annotation of their sncRNAs was essential.

A classification of different sncRNAs solely based in their read patterns, as it has been used in `DARIO` [144], was not possible in plants. Hence, `plantDARIO` uses third-party tools that also consider sequence and structure information for their predictions. Furthermore, due to a lack of one genome browser covering all plants, it was necessary to adapt and utilize different ones, allowing the researcher to take a look on the read distribution of the known and newly predicted sncRNAs.

2.2 Material and Methods

The current version of `plantDARIO` handles data for *A. thaliana* (TAIR9 and TAIR10)[1], *B. vulgaris* (RefBeet-1.1)[2] [146], and *S. lycopersicum* (SL2.40)[3] [147].

2.2.1 Concept of Web-server and its implementation

Storing, processing, and delivering the web pages to the client are the main features of a web-server. By the utilization of Hypertext Transfer Protocol (HTTP), communication between client and server takes place. The HTML (HyperText Markup Language) documents are generally delivered as pages including images, style sheets and scripts along with text content. Sometimes multiple web servers are used for a high traffic website [148].

Communication is initiated by a web browser or web crawler or an user agent by making a request for a specific resource using HTTP and the server responds with the content of that resource or an error message if unable to do so. Generally there is a secondary server storage in the form of a real file or folder which is the resource. In this case also, we have implemented such server storage for all the informations (files and folders) to create the webserver `plantDARIO`, which can be accessed at `http://plantdario.bioinf.uni-leipzig.de/`.

[1] `ftp://ftp.arabidopsis.org`
[2] `http://bvseq.molgen.mpg.de`
[3] `http://solgenomics.net/organism/Solanum_lycopersicum/genome`

2.2.2 The Workflow pipeline

The input by the user to the plantDARIO web-service is a list of sequencing read positions which are mapped to one of the supported reference genomes. Data generated from any sequencing platform and mapped with the read alignment tool of user's choice can be used. However, only data originating from experiments prepared with the small RNA-seq protocol and thus predominantly covering read lengths of about 21–26 nt can be analyzed. Mapped reads can be uploaded in either BAM or bed format.

We provide the PERL script map2bed.pl for converting mapped reads to bed format and for merging reads to tags, i.e., unique reads, that are represented as coordinate pairs rather than sequences. This reduces the volume of data to be transferred over the internet to a manageable amount: 1 GB of SAM formatted mapper output is converted to about 15 MB of compressed bed file that can be uploaded to plantDARIO. Additionally, user-defined annotations can easily be added to the annotation information stored in plantDARIO's internal database by uploading a list of loci, again in BED format.

Figure 2.1 summarizes plantDARIO's workflow. The usage of plantDARIO is detailed on the separate help page ([4]).

Instead of featuring a big extensive pipeline in the workflow, we have collated several analytical works as one step in the workflow. The first component of the pipeline performs a global statistical analysis of the input and provides the aggregate data for several quality control tools. The second component is concerned with the quantitative expression analysis at known and user-defined loci. The third component supports the discovery of novel microRNAs, snoRNAs, and tRNA-like loci. Output is displayed as HTML web pages and provided as machine-readable text files for download. A single job typically takes between 1 and 2 hours.

2.2.3 Quality control of the input data

A wide variety of errors and biases have been described in high-throuput sequencing data, which may originate from sample handling, library preparation, or the sequencing itself. It is thus necessary to assess the quality and integrity of the

[4] http://plantdario.bioinf.uni-leipzig.de/help.py

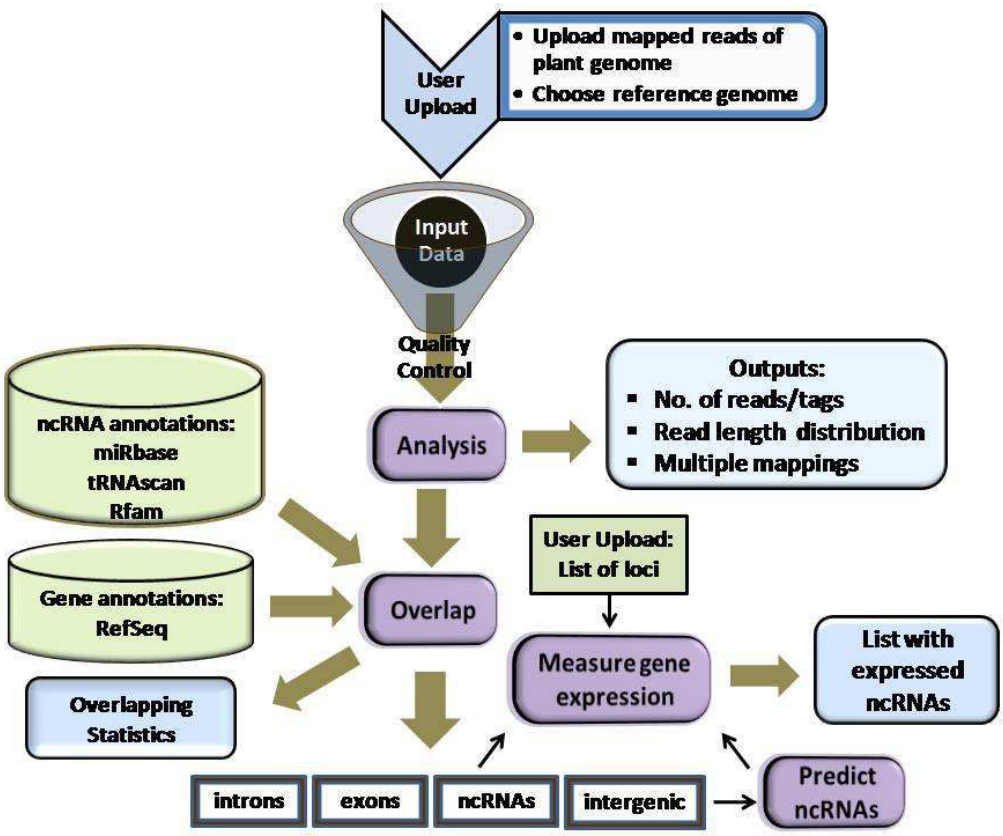


FIGURE 2.1: Workflow design of `plantDARIO`. Several analyses are integrated into one step e.g. quantification, normalization processes are merged into the step 'Measure gene expression'.

experimental data before they are analyzed for biological content [149–151]. Important measures include the number of mappable reads and the number of tags (distinct read sequences), the distribution of read length, and the sequence composition of mapped reads.

A set of plots provides a convenient overview of the dataset (Figure 2.2). `plantDARIO` also computes a summary of the distribution of reads among annotation items such as introns and exons and the major classes of annotated non-coding RNAs such as miRNA, snRNA, rRNA, tRNA, tasiRNA, and snoRNAs.

2.2.4 RNA Quantification

Mapped loci are overlapped with annotated ncRNAs. To this end, `plantDARIO` includes an internal database of ncRNAs comprising microRNAs from `miRBase` [152],

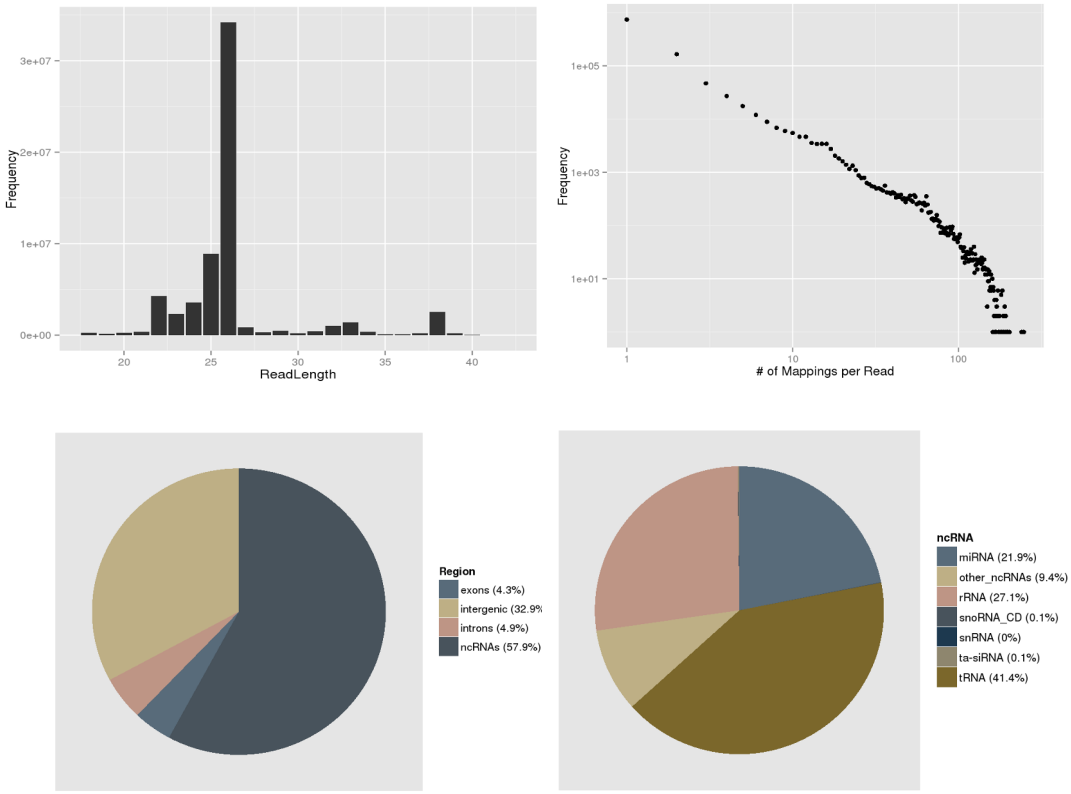


FIGURE 2.2: Initial quality control. `plantDARIO` provides overviews of the read length distribution, the distribution of read-length multiplicities, the distribution of genomic locations, and known annotations (separated into known ncRNAs, exons, introns, and intergenic regions). Here, an overview of data-set SRR952330 from *A. thaliana* is shown as an example.

tRNA annotations from `tRNAscan-SE` [153], tasiRNA annotations from TAIR [5]and tasiRNAdb [6] [154], plant specific literature data [146, 155, 156], as well as dedicated homology-based annotations for each individual genome. This internal annotation can be complemented by user-defined loci, which are then fully included in all downstream analyses. To handle multiple mappings, the number of reads for each sequence tag is divided by the number of its mapping loci, and this normalized expression value is assigned to each mapping locus.

The webserver generates a list of expressed ncRNAs, itemized by ncRNA classes. For each of them, a normalized expression value based on RPM (Reads per million) and the number of mapped reads (both in raw form and normalized for multiple mapping) is displayed.

[5] ftp://ftp.arabidopsis.org
[6] http://bioinfo.jit.edu.cn/tasiRNADatabase/

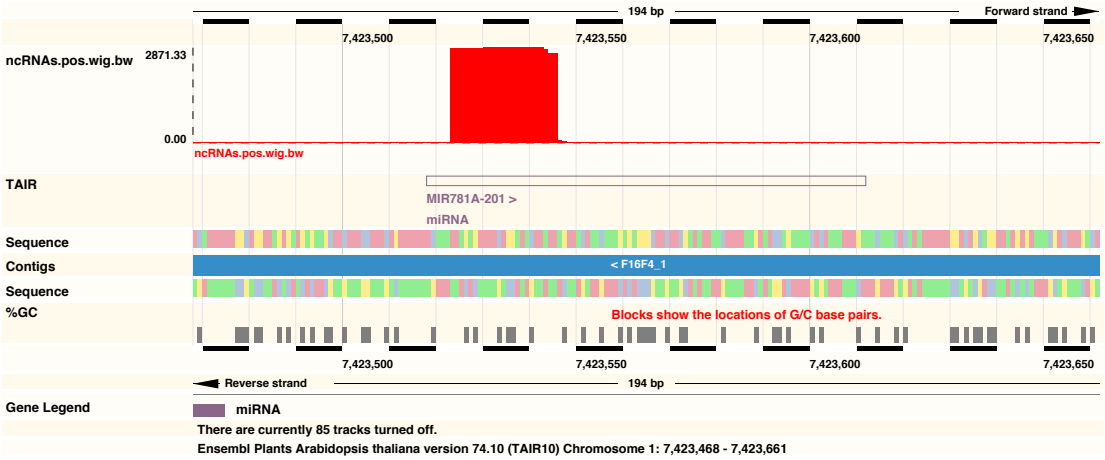


FIGURE 2.3: A link to the Ensemble genome browser (http://plants.ensembl.org) allows the instantaneous inspection of ncRNAs with help of ncRNA annotation tracks and conservation. The example shows the MIR781A-2.1 locus.

In addition a link to a genome browser is generated that allows for the user to conveniently inspect the expression pattern at each individual locus (Figure 2.3). This can he helpful e.g. to distinguish between *bona fide* microRNAs from other RNA classes in case of misannotations [157], to inspect microRNA genes for the presence of offset RNAs [158, 159], or to look for short reads generated from the antisense locus [160].

2.2.5 Analysis of Unannotated Loci

Mapped tags are merged to blocks and are aggregated to regions of blocks using `blockbuster` [159] with default parameters. Contrary to animals, the processing patterns of microRNAs are not very consistent in plants (Figure 2.4) so that patterns of mapped reads alone do not allow a sufficiently accurate classification. The same is true for snoRNAs.

Hence the prediction of microRNAs and snoRNAs is assisted by the integration `novomir` [161] and `snoReport` [162] in `plantDARIO`. The tools are integrated as algorithms or scripts locally and interfaced the the output internally to `plantDARIO`. Both tools implement RNA folding and machine learning approaches to classify intervals of genomic sequences. We use `blockbuster` to identify accumulations of reads and then run the two tools on these loci.

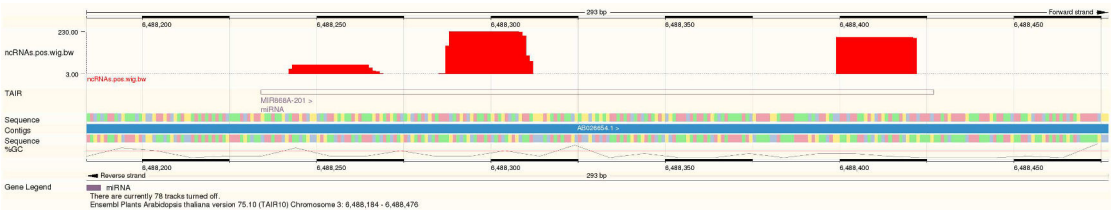


FIGURE 2.4: Usual read patterns of plant microRNAs. The example shows the MIR868A-201 locus.

2.2.6 ncRNA Annotation in *Solanum lycopersicum*

Non-coding RNAs have not been comprehensively annotated in many published genomes. This is also the case for *S. lycopersicum*, whereas most relevant annotation data were already available for the arabidopsis and sugar beet genomes. Here is the annotation track focussing on microRNAs, snoRNAs, and tRNAs for the tomato genome roughly following the workflow employed for the annotation of the *B. vulgaris* genome [146]:

1. For microRNAs, plant microRNA pre-cursors were downloaded from `miRBase` and mapped against the genome using `blast`, employing a minimum alignment length of 60 nt and a sequence similarity of 80% as filter criteria. Overlapping matches were combined.

2. For snoRNAs, all plant snoRNAs were downloaded from the `Rfam` database and mapped against the genome with `blast`, employing a minimum alignment length of 70 nt and a sequence similarity of 80% as filter criteria. Overlapping matches were combined.

3. For tRNAs, `tRNAscan` [153] was run against the whole genome of *S. lycopersicum*.

The annotations can be downloaded from [7].

[7] http://plantdario.bioinf.uni-leipzig.de/annotations/

2.2.7 snRNA annotation in *Solanum lycopersicum* and *Arabidopsis thaliana*

For the *B. vulgaris* genome, snRNAs are already annotated and available along with other non-coding genes from the *B. vulgaris* resource [146]. For *A. thaliana* and *S. lycopersicum*, snRNA covariance models were downloaded from Rfam ([8]), and `infernal` [163] was run against the respective genomes. For the purpose of providing a brief summary statistics, the spliceosomal RNAs U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac are grouped together with SRP RNA and RNase MRP RNA in the class "snRNAs". They can be downloaded from the annotation URL given above.

2.2.8 Genomes and Visualization

`plantDARIO` references to the `Ensembl` genome browser [164] to visualize the read coverage at annotated loci and predictions as custom tracks for *A. thaliana*. This allows an interpretation of the user data in the context of information provided by the Gramene database [165], a resource for plant comparative genomics. For sugarbeet and tomato, we rely on the genome browser from the *B. vulgaris* resource [146] and sol genomics network (SGN) [147], respectively, for visualization.

2.2.9 Implementation Details

The technical details of `plantDARIO` parallel those of `DARIO` [144]. Web pages are created by `python` scripts making use of the `Mako` template engine. Graphics are created using `R` and the graphics package `ggplot2` [166]. A queuing system is used to distribute analysis jobs.

2.3 Results and Discussion

`plantDARIO` implements basic workflows for the analysis of RNA-seq data. It allows the user to obtain a comprehensive overview starting after read mapping. To demonstrate the versatility of `plantDARIO` we re-analyzed publicly available small

[8] ftp://ftp.ebi.ac.uk/pub/databases/Rfam/

TABLE 2.1: Known and novel sncRNAs in four test datasets. For both microRNAs and snoRNAs, the number of expressed annotated sncRNA loci ("known") and the number of novel candidates ("new") is reported.

		miRNAs		snoRNAs	
Data	Species	known	new	known	new
SRR167709	*A. th.*	276	121	78	348
SRR167710	*A. th.*	236	139	71	268
SRR786984	*S. ly.*	268	65	121	202
SRR868805	*B. vu.*	197	41	60	22

RNA-seq datasets from *Arabidopsis* SRR952330, [SRR167709 and SRR167710; 167], sugarbeet (SRR868805) [146], and tomato (SRR786984) [168]. We used `segemehl` [169] with default parameters to map the sequencing data to the respective reference genomes. Unlike many other mapping tools, `segemehl` has full support for multiple-mapping reads which is very important for small RNA-seq [170].

2.3.1 Novel miRNAs and snoRNAs

In addition to more than 200 known microRNAs, we observed more than 100 expressed putative novel microRNAs in each of the datasets. An example of a newly predicted microRNA is shown in Figure 2.5. It represents a perfect plant microRNA pattern as expected for sncRNAs processed by a plant dicer-like enzyme [50], resulting in one functional arm (proper read block in the figure) in this case. The irregular patterns found as little bumps in the structure are bulge loops or internal loops present in the pre-miRNA structure, which are usual, i.e., which are a thermodynamic feature of the RNA. Furthermore, the read pattern matches a stem-loop when traced back to a likely pre-microRNA, as shown in Figure 2.5.

For snoRNAs, we observed an even larger number of candidates. An example is detailed in Figure 2.6. The structure pattern shows a candidate snoRNA with typical C box and D box sequence patterns close to the ends. The middle region, presumably a loop, contains box C' and D' regions frequently found in box C/D snoRNAs.

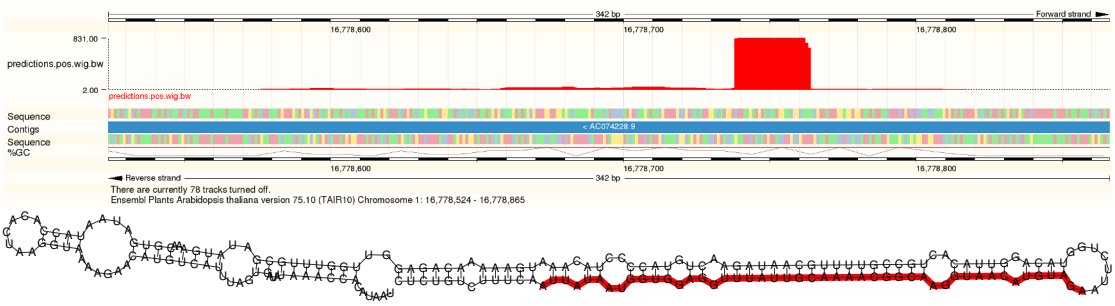


FIGURE 2.5: A novel microRNA discovered by `plantDARIO`. Top: Visualization of the expression profile. Bottom: Secondary structure of the predicted microRNA precursor.

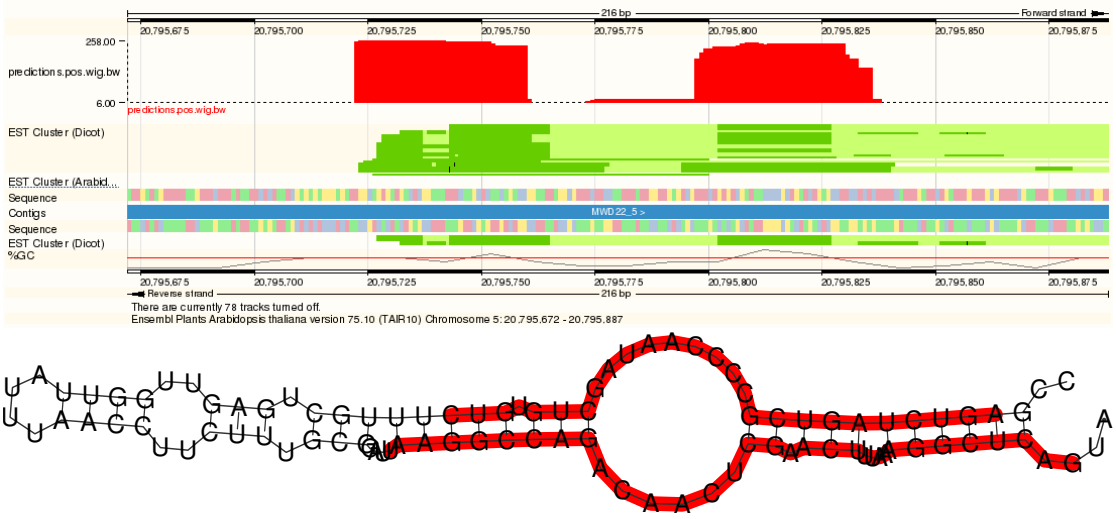


FIGURE 2.6: A novel CD box snoRNA discovered by `plantDARIO`. Top: Visualization of the expression profile. Bottom: predicted secondary structure; the orgin of the observed short reads is marked in red.

2.3.2 Differential expression

In order to demonstrate that the output of `plantDARIO` is easy to use for downstream analyses, we compared small RNA expression for microRNA and snoRNA in the two *A. thaliana* datasets SRR167709 and SSR167710 [167] representing populations of small RNAs from *Arabidopsis* immature flowers of WT and drb2 mutants, respectively. The original study aimed at the antagonistic impact of double-stranded RNA binding proteins DRB2 and DRB4 on polymerase dependent siRNA levels. Figure 2.7 shows that, overall, the microRNA expression levels correlate positively between the two data-sets for both previously annotated and newly predicted microRNAs.

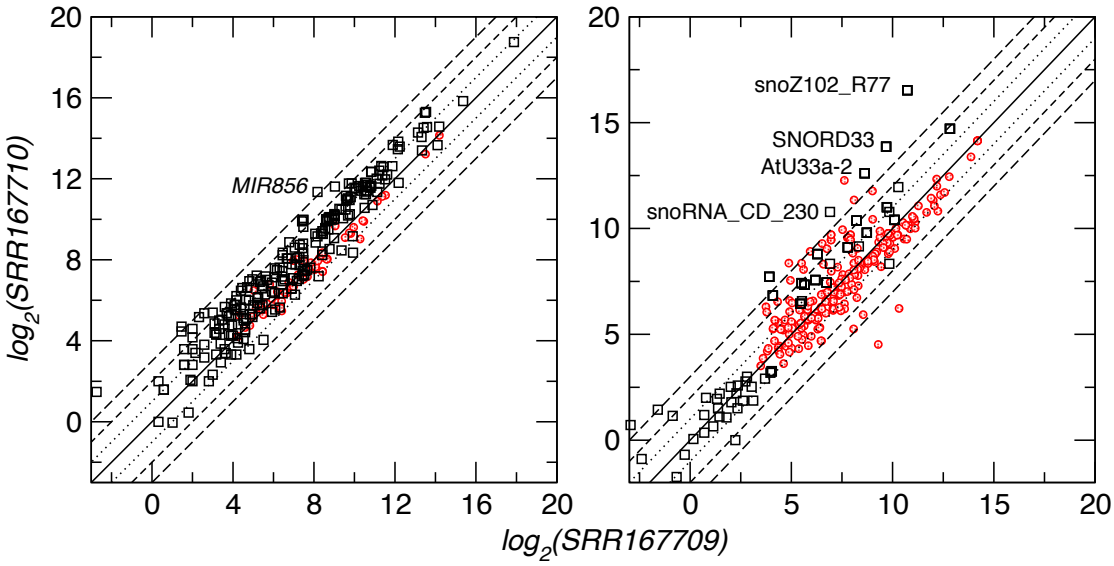


FIGURE 2.7: Differential expression of microRNAs (left panel) and snoRNA-derived small RNAs (right panel) for two *A. thaliana* datasets. Diagonal lines indicate differences between 2^3 and 2^{-3} fold. Black symbols indicate annotated microRNA and snoRNA loci, red dots refer to novel predictions. A few loci with extreme expression differences are labeled.

One of the microRNAs with extreme (> 8fold) change in expression level is ath-MIR856. This microRNA, which is predominantly expressed in the floral organ [63], belongs to a set of microRNAs that are evolutionary transient within the genus *Arabidopsis* [171, 172] and shows an exceptional evolutionary behavior with relatively low levels of polymorphism but the highest level of divergence [173].

Surprisingly, we observe a much larger variability for the processing products of snoRNAs. The extreme case, snoZ102_R77, is a box C/D snoRNA belonging to the SNORD44 clan. Box C/D snoRNA_CD_230 (*Arabidopsis* chr1 6697176 6697261) is related to snoR16 and snoR72 families according to a search in `Rfam`. All these snoRNAs have a primary function in ribosomal RNA processing [83]. Interestingly, the examples with extreme differential expression belong to the box C/D class of snoRNAs that is not processed by Dicer but utilizes another, hitherto unknown, processing pathway at least in mammals [174].

2.4 Concluding Remarks

High-throughput sequencing has become the method of choice for the analysis of transcriptome data. For the special case of small RNA-seq data, webservices

provide a convenient means of conducting standard analyses. In this way the user can avoid the need to install, maintain, and update an array of individual tools. plantDARIO is such a service that, in contrast to comprehensive analysis environments like GALAXY [175], provides a ready-to-use analysis workflow for small RNA-seq data.

Together with precompiled sncRNA annotations this allows to inspect analysis results quickly after uploading the user data. In summary, plantDARIO provides the user with a valuable combination of annotation-based, standardized quantitative analysis and a simple facility for guided discoveries of novel small RNA loci.

The webservice also provides the results in a bed format, which can easily be used for downstream analysis tasks such as the assessment of differential expression. Using publicly available small RNA-seq data for *A. thaliana* we noticed extreme differences in the levels of small RNAs processed from box C/D snoRNAs.

Some of these sncRNAs are known to have a regulatory role in animals, so it might be of possible interest to further characterize small RNA processing from "house-keeping ncRNAs" in plants, and plantDARIO might be a convenient and versatile tool for this purpose.

Chapter 3

Phylogenetic distribution of plant snoRNAs

Until this chapter, we already read about the plant small non-coding RNAs and their analysis. In this chapter we are going to analyze and evaluate plant snoRNAs with phylogenetic tree studying their distribution.

3.1 Background of analyzing plant snoRNAs and their phylogenetic distribution

It is already known that small nucleolar RNAs function as guides in site-specific RNA modification [76, 77]. They fall into two distinct classes: box H/ACA snoRNAs responsible for targeting pseudouridylation sites and box C/D snoRNAs directing 2'-O-methylation of ribonucleotides and both are part of well-defined ribonucleo-particles the snoRNPs [73].

Based on sequence similarity, snoRNAs fall into many well-defined families of homologous genes. As a consequence of the frequent segmental, chromosomal, and whole genome duplications in plant genome evolution, most plant snoRNA families have multiple paralogous members both in spatial clusters and spread throughout the genome [106]. Despite their ancient ancestry as a class [176], not very much is known about the evolution of the individual snoRNA families.

Several studies showed that many snoRNA families are conserved at phylum or even kingdom level in animals [177], plants [83], and fungi [75]. The genome-wide analysis of chicken snoRNAs provided direct evidence for extensive recombination and separation of guiding function [178]. Similarly, multicellular fungi exhibit a more complex pattern of methylation guided by box C/D snoRNAs than unicellular yeasts [179]. Nevertheless, conserved snoRNA targets typically have conserved modification sites, although there is some redundancy and an appreciable level of turnover throughout the animal kingdom [177].

Again matching the situation in microRNAs [180], there is good evidence for clade specific *de-novo* innovation of novel snoRNA families found in fungi as well as in humans[76, 181]. The long and the short of it is that so far there is no clear picture if and how the evolution of plant snoRNAs differs from the situation in fungi although a lot of data are available, dispersed throughout the literature. However, no clue is found how snoRNAs emerges in case of plants, and we aim to find out if clade specific innovation of snoRNA families occurs in plants too.

Although there is good evidence for the conservation of many of the chemical modification sites on rRNAs and snRNAs between eukaryotic kingdoms [182], it has remained an open question to what extent individual snoRNA families are homologous at such large phylogenetic distances.

This is difficult to address since snoRNA sequences evolve quite rapidly apart from the conserved boxes and the antisense region. To tackle this question, it is necessary to first understand the evolutionary patterns of snoRNAs within each kingdom in detail. Secondly, snoRNA families that originated in the eukaryotic ancestor need to be distinguished from those that originated more recently. This may provide an answer of how the phylogenetic distances affected the homology of snoRNAs.

In this contribution we reconstruct the evolutionary history of snoRNAs in the plant kingdom. Henceforth, the paper is structured in studying the phylogenetic distribution of the snoRNA families with the identification of additional homologs and several interesting patterns of conserved snoRNA families and spatial clusters along with systematic tracing of the evolution of each individual snoRNA family back to its last common ancestor.

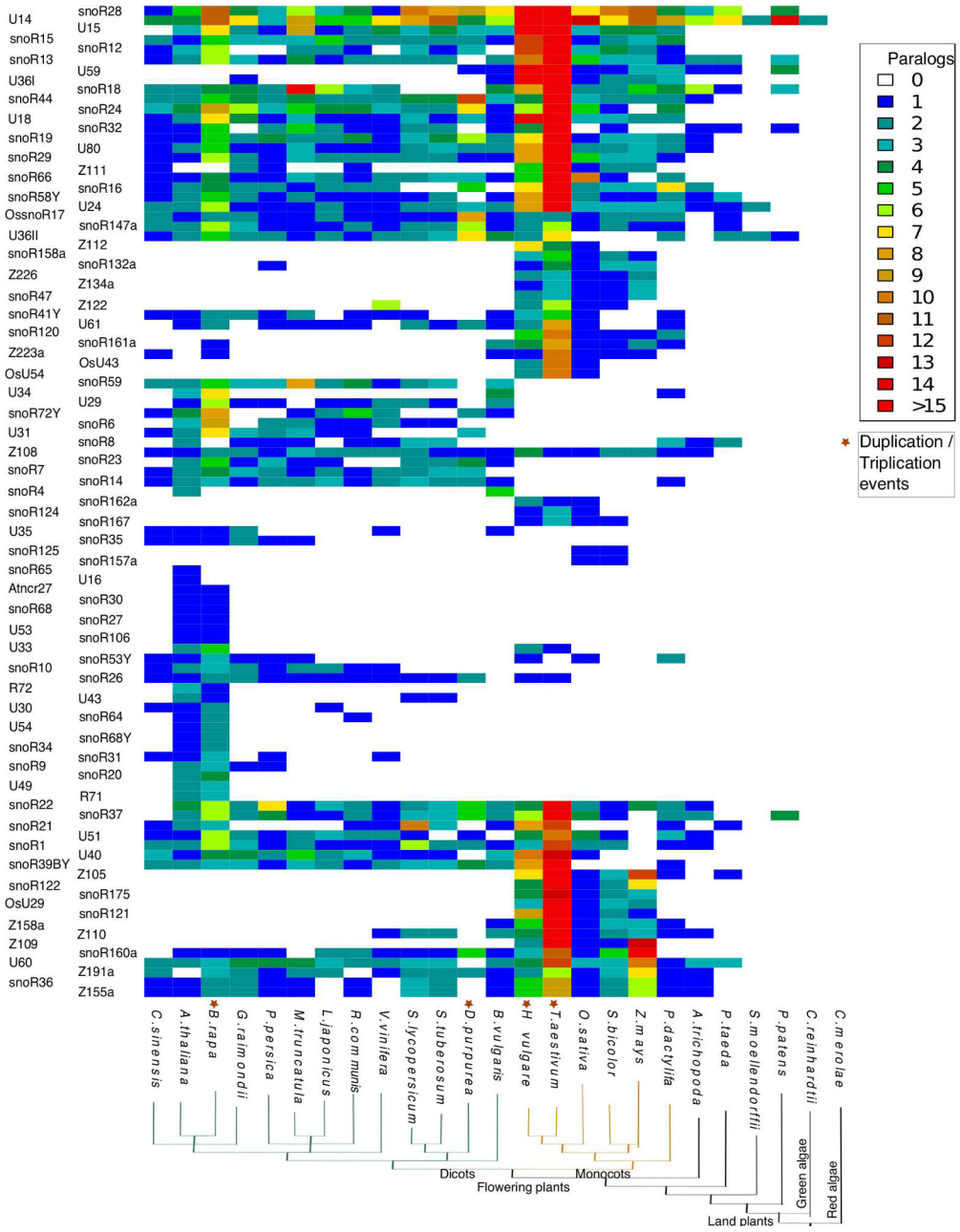


FIGURE 3.1: The heatmap (built in R with heatmap.2 version) shows the box C/D snoRNA families and their distribution amongst the plant species. The colour code reflects the number of box C/D paralogs found within each species. The phylogenetic tree was constructed from recent literature and NCBI Taxonomy information.

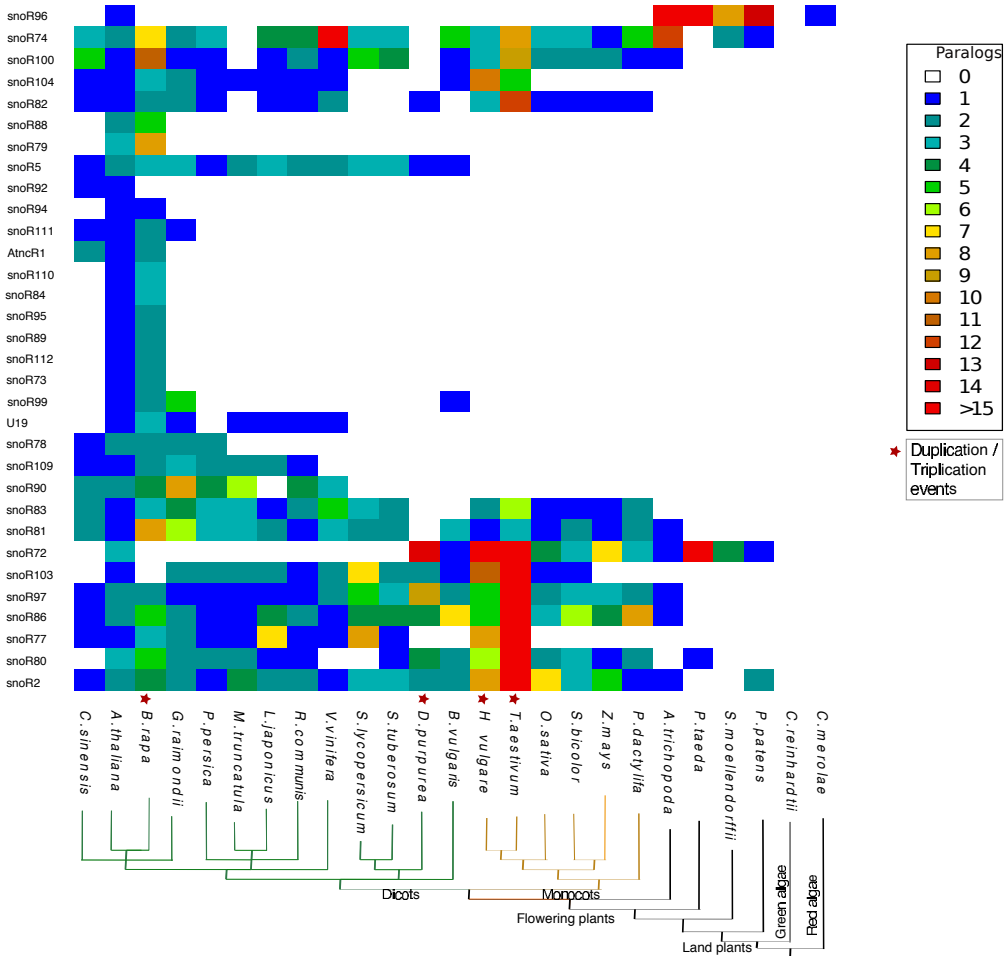


FIGURE 3.2: The heatmap (built in R with heatmap.2 version) shows the box H/ACA snoRNA families and their distribution amongst the plant species. The colour code reflects the number of box C/D paralogs found within each species. The phylogenetic tree was constructed from recent literature and NCBI Taxonomy information.

3.2 Material and Methods

Data sources

We selected 24 plant species with completely sequenced genomes covering the plant kingdom, see Figs. 3.1 and 3.2. Among crown group (living representatives of the collection together with their ancestors back to their most recent common ancestor as well as all of that ancestor's descendants) eudicots, we preferentially included species for which snoRNAs had been described in the literature.

We collected all available plant snoRNA sequences from the SnoRNA orthologous gene database (SNOPY [183]) and the plant snoRNA database [184]. In addition we extracted snoRNA sequences from the literature [84, 155, 156, 185–188].

We considered only the rRNAs/snRNAs as potential targets. Ribosomal RNA sequences of the 24 plant and red algae species are downloaded from the SILVA database [189]. The snRNAs comprising of U1, U2, U4, U4atac, U5, U6, U6atac, U11, and U12 are imported from datasets of the plantDARIO webserver [190].

3.2.1 Curation of initial snoRNA data

From the initial set of collected snoRNAs, the box motifs are annotated and categorized into box C/D and box H/ACA snoRNAs. The characteristic boxes (C, D', C', D, H, ACA) are annotated manually using the sequence patterns as constraints given in [191].

3.2.1.1 SnoRNA box motifs

Previous analyses from the Bachellerie laboratory showed conserved spacing between the box C/D core motif and the internal D'/C' motif of the archaeal box C/D snoRNAs[109]. Although alteration of D and D' spacer distances does not affect box C/D and D'/C' RNP assembly, the spacer distances severely affect box C/D and D'/C' RNP-guided methylation of target RNAs [191].

Hence, box motives are annotated based on both known pattern of conserved nucleotides and likely spacer distances, usually 12nt, between the box C/D and D'/C' motifs. Only snoRNAs with boxes that could be annotated with high certainty are selected for the initial query set. The sequences are then grouped into gene families based on known orthology and sequence similarity.

3.2.2 Homology search

In the next step all snoRNA families were mapped to all plant genomes. The list of all genomes with accession numbers is provided in Appendix A. The snoStrip pipeline [75] was used to search each of the 24 plant genomes for homologs of each of the query families. In a nutshell, snoStrip is an automatic annotation pipeline

that is developed specifically for comparative genomics of snoRNAs. It first uses both a `blast` search with relaxed parameters and `infernal` [192] to retrieve initial candidates.

The expected boxes and the anti-sense elements were annotated based on sequence alignments, and candidates were filtered for the presence of the boxes. Then secondary structure features are validated. In the final step a family-wide alignment of all retained candidate sequences was calculated. The alignments produced by `snoStrip` are manually inspected.

Data were then aggregated to heatmaps showing the number of family members in each species. SnoRNA clusters were identified by proximities of genomic coordinates.

The history of gains and losses in each snoRNA family was reconstructed using a Dollo parsimony approach implemented in the `ePoPe` programm [193].

Since the nomenclature of plant snoRNAs only partially respects known or detectable sequence homology we used a unique internal family identifier throughout this study. These identifiers are re-translated to a consolidated family nomenclature that is based, in this order, on the nomenclature for *Arabidopsis*, *Oryza*, and *Chlamydomonas*. A complete table of family names and their species-specific synonyms is provided in Appendix B.

3.3 Results and Discussion

From the intial set of collected and curated snoRNA families, snoRNAs are mapped to all the plant genomes, family-wide alignments of all retained candidate sequences were calculated and finally the history of gains and losses in each snoRNA family was reconstructed. The initial query set of 554 snoRNA genes was comprised of a collation of all available (plant) snoRNA databases. These sequences were assigned to 222 box C/D and 74 box H/ACA snoRNA families after manual curation and annotation of the box C/D and box H/ACA snoRNAs. We identified a total of 5116 additional homologs in the 24 plant species under consideration.

3.3.1 Heatmaps of snoRNA families

The phylogenetic distribution of the snoRNA families is shown in Figure 3.1 and Figure 3.2 in form of heatmaps color-coding the number of family members.

3.3.2 Patterns in heatmaps of snoRNA families

Several patterns are apparent. With the exception of the highly conserved U14 family and the snoR96 family that shows a much more scattered distribution, snoRNAs from land plants do not have identifyable homologs in green algae. Seven families of box C/D snoRNAs (snoR28, U14, snoR13, snoR18, snoR32, U36II, and snoR37) as well as four H/ACA snoRNA families (snoR2, snoR72, snoR96, and snoR74) are present throughout land plants. The largest fraction of identified snoRNAs (76 box C/D and 20 box H/ACA families) are common to the flowering plants including both monocots and dicots.

It is possible that many of these families are in fact evolutionarily older and that the apparent restriction to land plants or flowering plants is a consequence of the limited sensitivity of state-of-the-art homology search methods. The consensus box motifs within some snoRNA families are very well conserved across the plant kingdom, see Figure 3.3 for an example.

A very interesting pattern is the large block of box C/D snoRNAs (20 families) that is only present in monocots. A similar pattern is not visible for box H/ACA snoRNAs. There is also no such pattern of dicot-specific box C/D snoRNAs or dicot-specific box H/ACA snoRNAs. Hence, it is very unlikely that the monocot specific families of box C/D snoRNAs are just an artefact caused by limitations in the homology search method. So they should be interpreted as true monocot innovations.

Finally, focussing on column-wise patterns we observe a systematically elevated number of snoRNA paralogs in some species. The most prominent examples are *Brassica rapa* and *Digitalis purpurea* among dicots, as well as *Triticum aestivum* and *Hordeum vulgare* among monocots. By comparison with the Plant Genome Duplication Database [194] this observation is readily explained by phylogenetically recent genome duplication or triplication events.

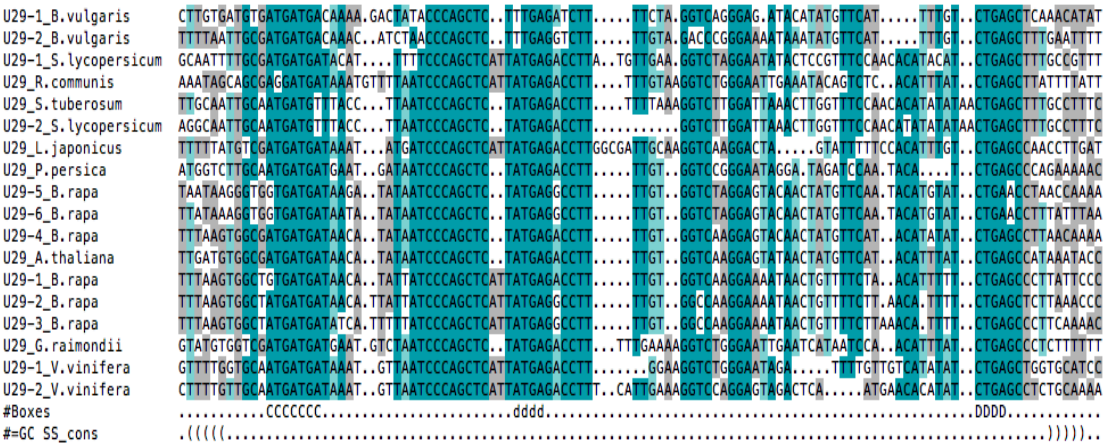


FIGURE 3.3: Conservation pattern of snoRNA U29. In the #Boxes line nt marked with C, D, and d belong to the box C, box D, and box D', respectivley. The consensus secondary structure in dot-bracket notation provides the typical terminal stem with the unpaired nucleotides inbetween. The region upstream of the box D' is highly conserved. It is the putative antisense element for guiding a modification. The region upstream of the box D is less conserved than box D'.

3.3.3 Exceptional snoRNA families

On the other hand, there are many families with a very narrow phylogenetic distribution: 27 families are found only in *Arabidopsis*, e.g. snoR107, 28 families appear to be specific to *Oryza*, e.g. snoR146a, and 131 families appear only in *Chlamydomonas*, e.g. CrACA02. Either these sequences have evolved extremely rapidly, essentially at neutral rates, or they are true species or genus-specific innovations.

3.3.4 snoRNA clusters

SnoRNAs that are encoded or positioned closely together in the same chromosomal region are considered as "snoRNA clusters". In order to study the long-term integrity of those clusters we investigated representative examples: the 68 rice snoRNA clusters described in [84]. Multiple snoRNA clusters have also been identified and studied in some detail in *A. thaliana* [155]. In this case, we find 10 snoRNA clusters that are conserved in rice and at least in some of the selected 24 plant species considered here, 5 of which have also been described in *A. thaliana* [155].

The 10 genomic clusters involve 22 distinct snoRNA families. A subset of the clusters comprises highly conserved snoRNAs, whereas most of the rice clusters are

not conserved in other species. Several snoRNA families have members in distinct clusters. Figure 3.4 summarizes the evolutionary history of "U15a-U15b-snoR7b-snoR18b cluster" termed "cluster 5" in rice [84], which consists of U15a, U15b, snoR7b, and snoR18b, respectively. While two members of the U15 family (U15A and U15B) and snoR18b date back to the magnoliophyte ancestor (*P.dactylifera*), snoR7b is a more recent addition, incorporated in the dicot ancestor. Its homolog in *A. thaliana* was discussed in [155] as the "U15a-U15b-snoR7.1 cluster".

The U36Ia-U36IIa-U36IIb cluster named as "cluster 1" in rice is only present in the flowering plants (3.5).

In the snoR12-U24 cluster ("cluster 19") (3.6), which was termed "U12.2-U24.2 cluster" in *A.thaliana*[155], U24 was present already in the ancestor of viridiplantae. In contrast, snoR12 comparatively has its origin in mesangiospermae or the flowering plants as seen in 3.6.

In cluster snoR22a-snoR23-snoR22b ("cluster 20") (3.7), the *A. thaliana* "U32.2-U27.2-U80.2 cluster" [155], snoR22b dating back to the magniliophyte ancestor whereas, snoR22a appears in the monocots and also in few recent dicot plants. However, snoR23 is the prominent addition in the dicot plants.

In cluster U27-U80b ("cluster 43") (3.8), amongst U27 and U80b, U27 is the recent snoRNA appearing in the mesangiospermae family, while U80b can be traced back to magniliophyta. It is also found in *A. thaliana* [155] as the "U32.2-U27.2-U80.2 cluster".

In the cluster U61-snoR14 ("cluster 49") (3.9) corresponding to the "U61-U14.1-U56" cluster" in *A.thaliana* [155], both U61 and snoR14 appear in the measangiospermae family, however, snoR14 is more consistently conserved in the mesangiospermae plant species.

Cluster snoR44-snoR17-snoR147a ("cluster 53") (3.10) consists of snoR44, snoR17, and snoR147. snoR147 is the ancestral snoRNA dating back to spermatophyte ancestor, followed by snoR44 dating back to the magniliophyte ancestor, whereas snoR17 appear to be recent emergence in the mesangiospermae or flowering plants.

snoR167-snoR47 cluster ("cluster 56")(3.11) comprising snoR167 and snoR47, both of them appear only in the monocots without any innovation in the recent species.

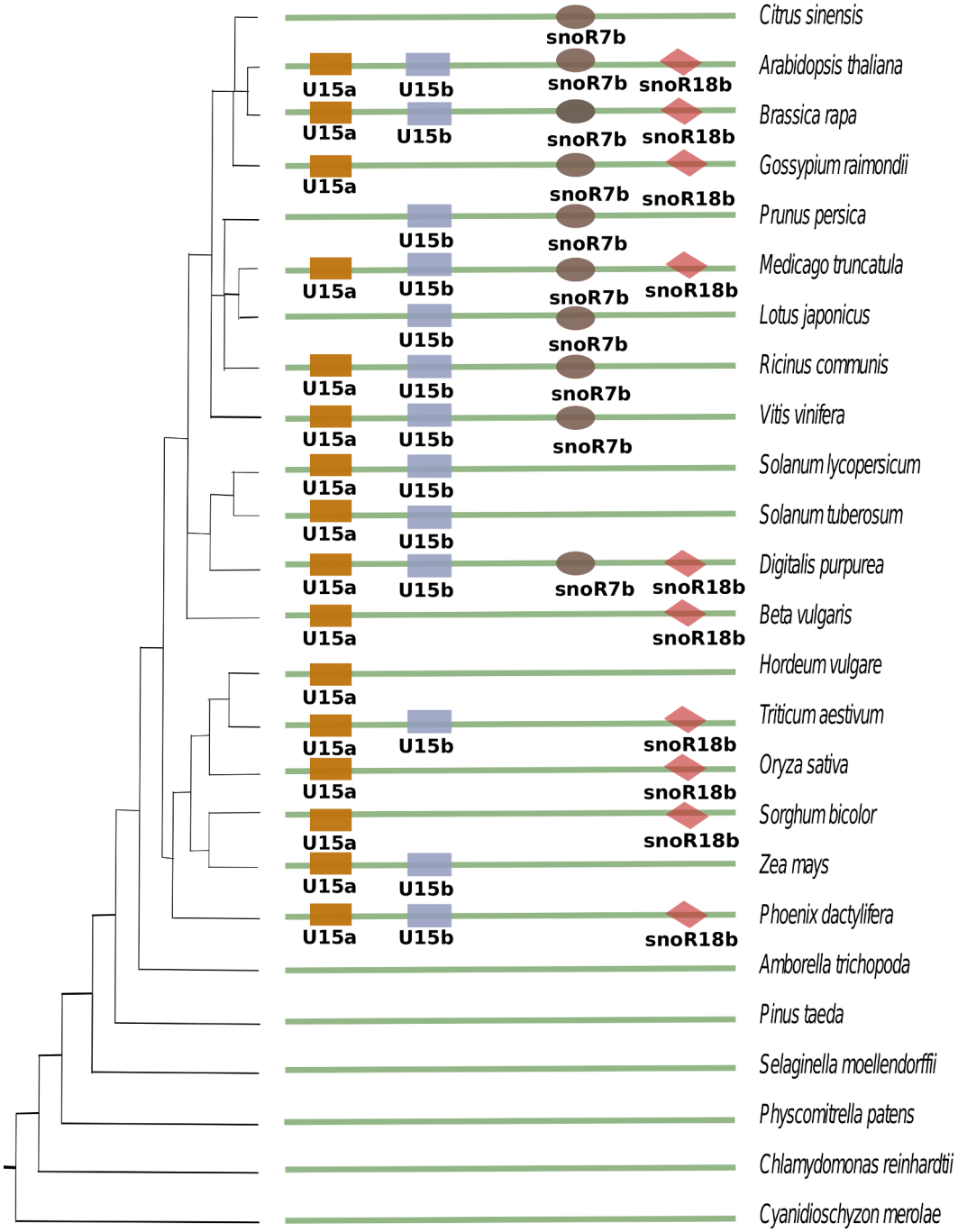


FIGURE 3.4: Evolutionary observation of snoRNA "U15a-U15b-snoR7b-snoR18b cluster", where we find two members of the U15 family (U15A and U15B) and snoR18b date back to the magnoliophyte ancestor (*P.dactylifera*), whereas snoR7b seems to be a recent innovation [84].

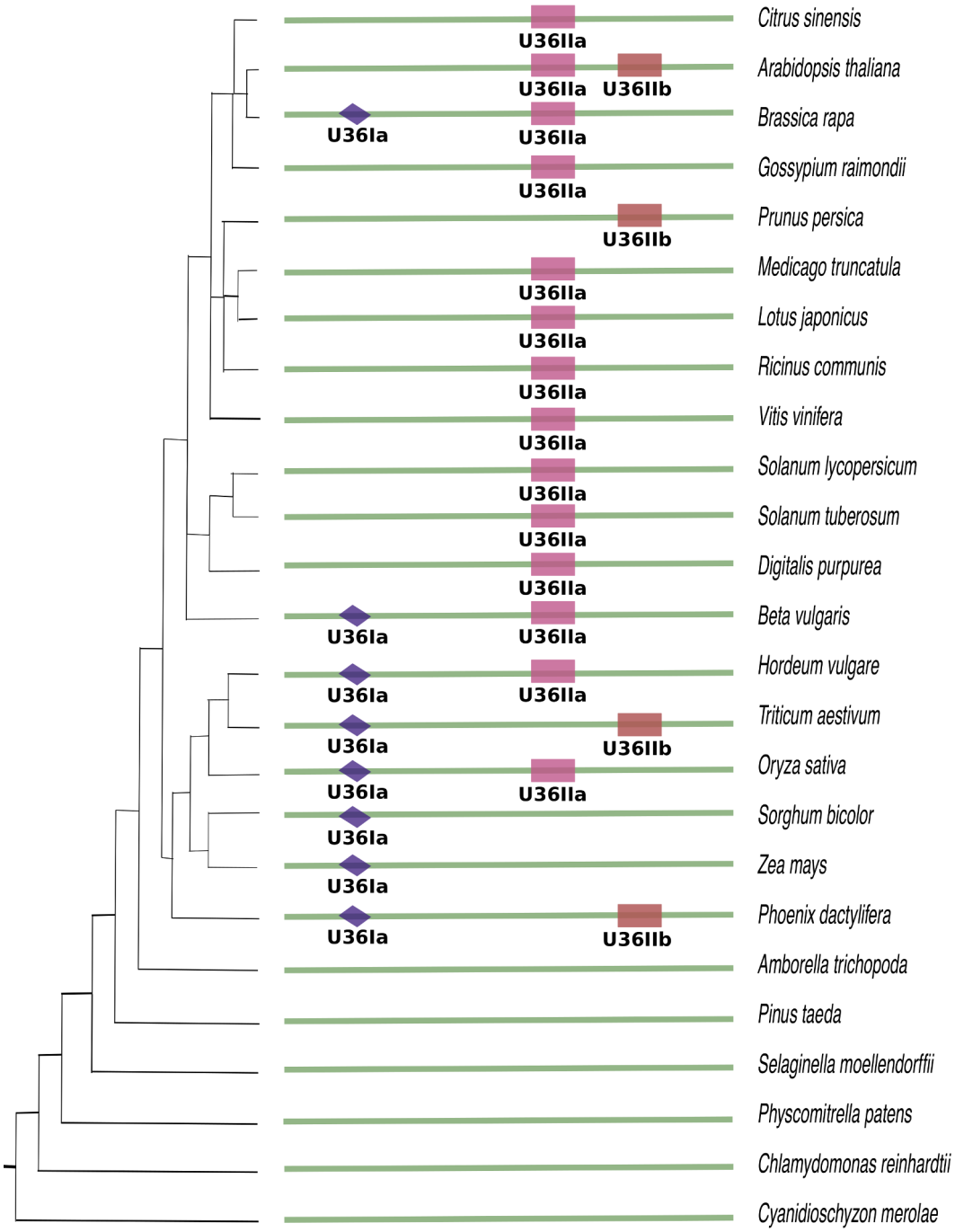


FIGURE 3.5: Evolutionary observation of snoRNA "U36Ia-U36IIa-U36IIb cluster"

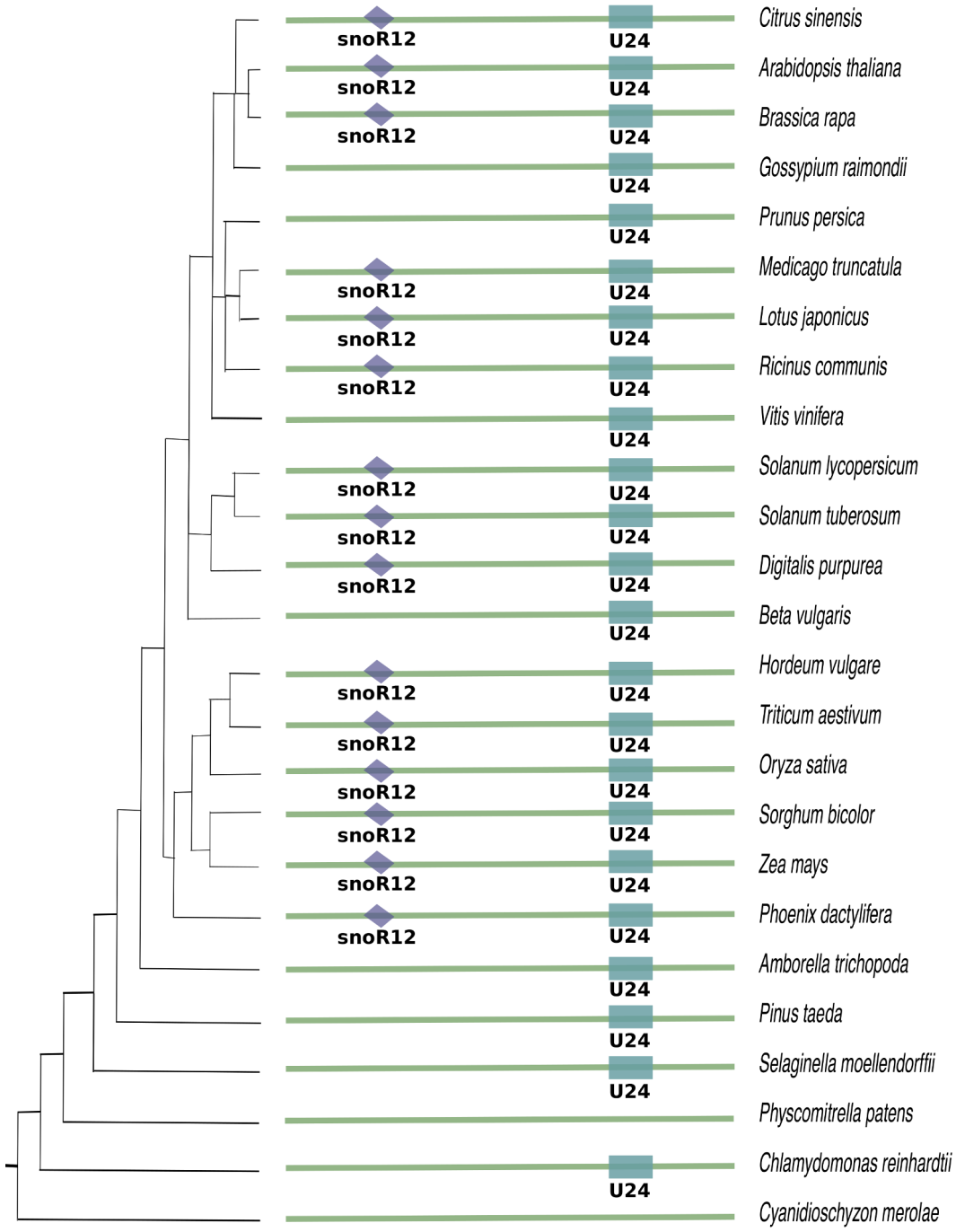


FIGURE 3.6: Evolutionary observation of snoRNA "snoR12-U24 cluster"

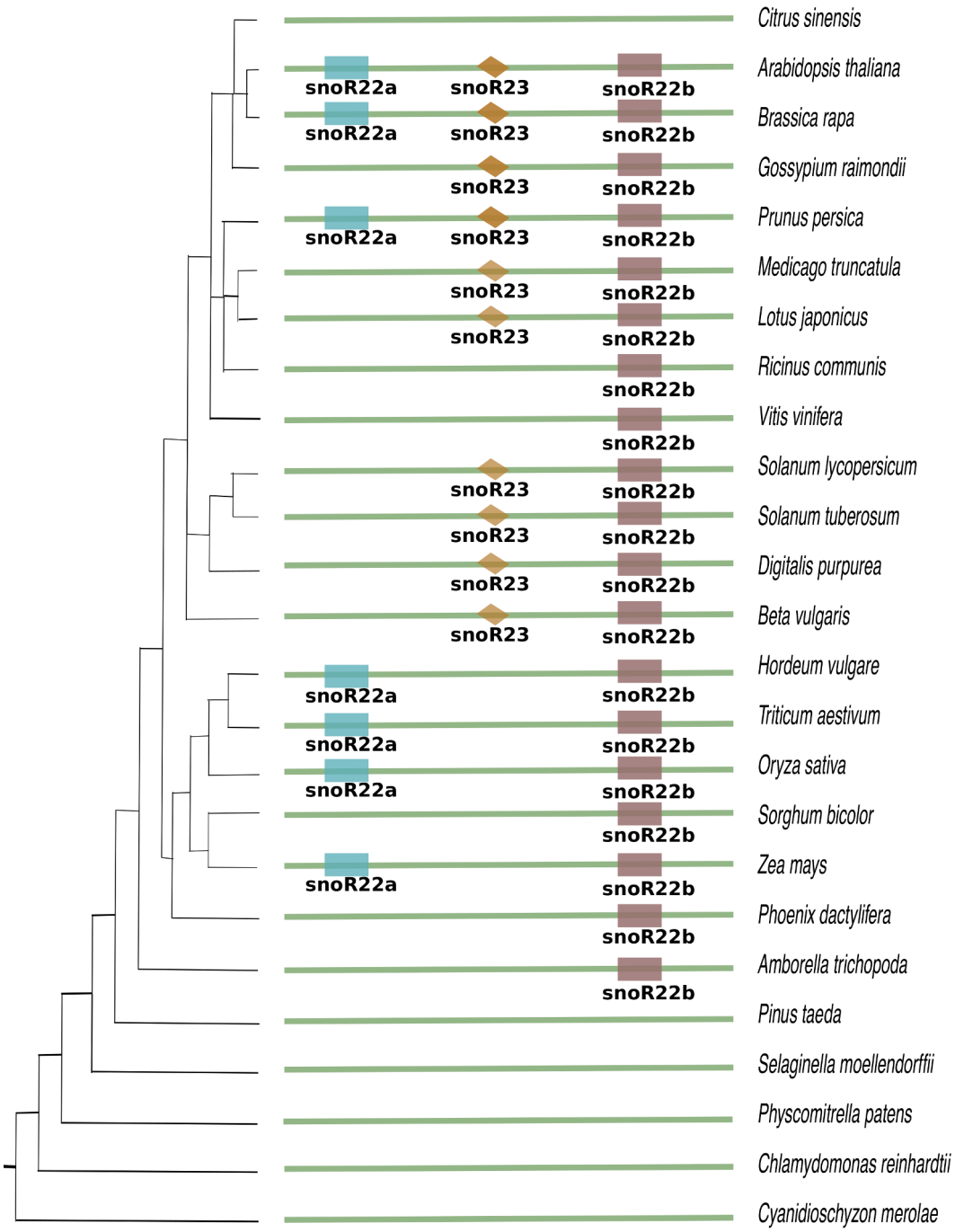


FIGURE 3.7: Evolutionary observation of snoRNA "snoR22a-snoR23-snoR22b cluster"

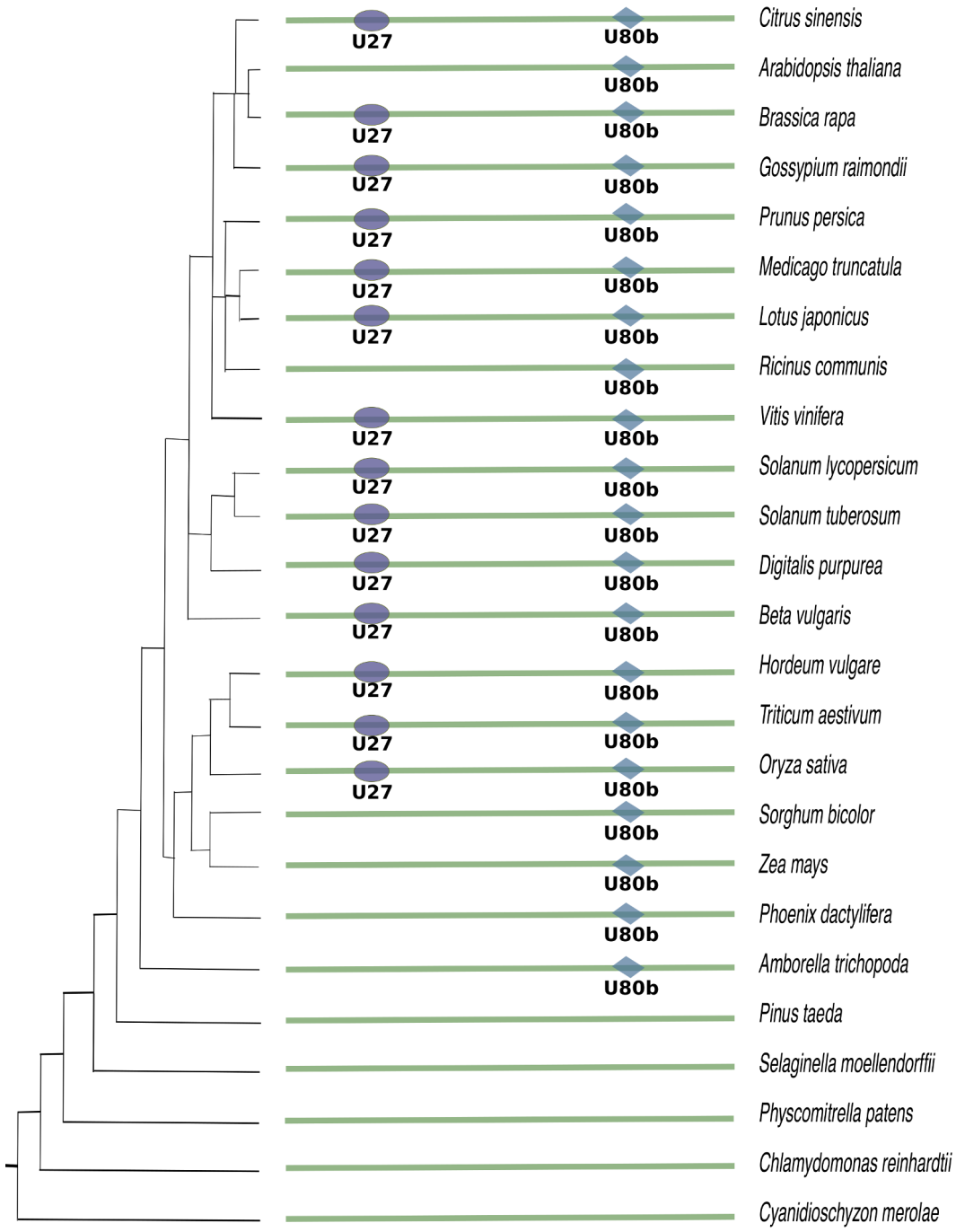


FIGURE 3.8: Evolutionary observation of snoRNA "U27-U80b cluster"

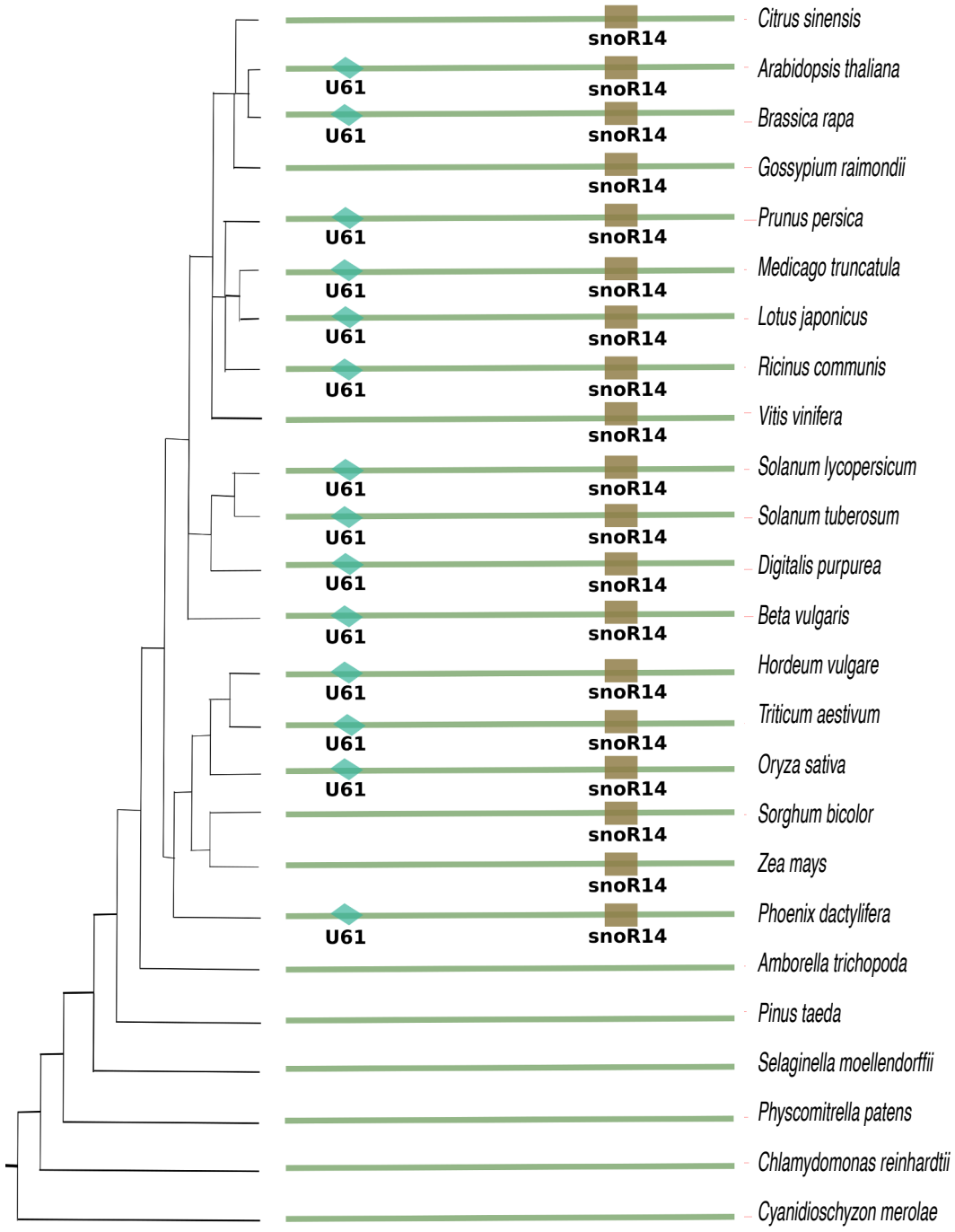


FIGURE 3.9: Evolutionary observation of snoRNA "U61-snoR14 cluster"

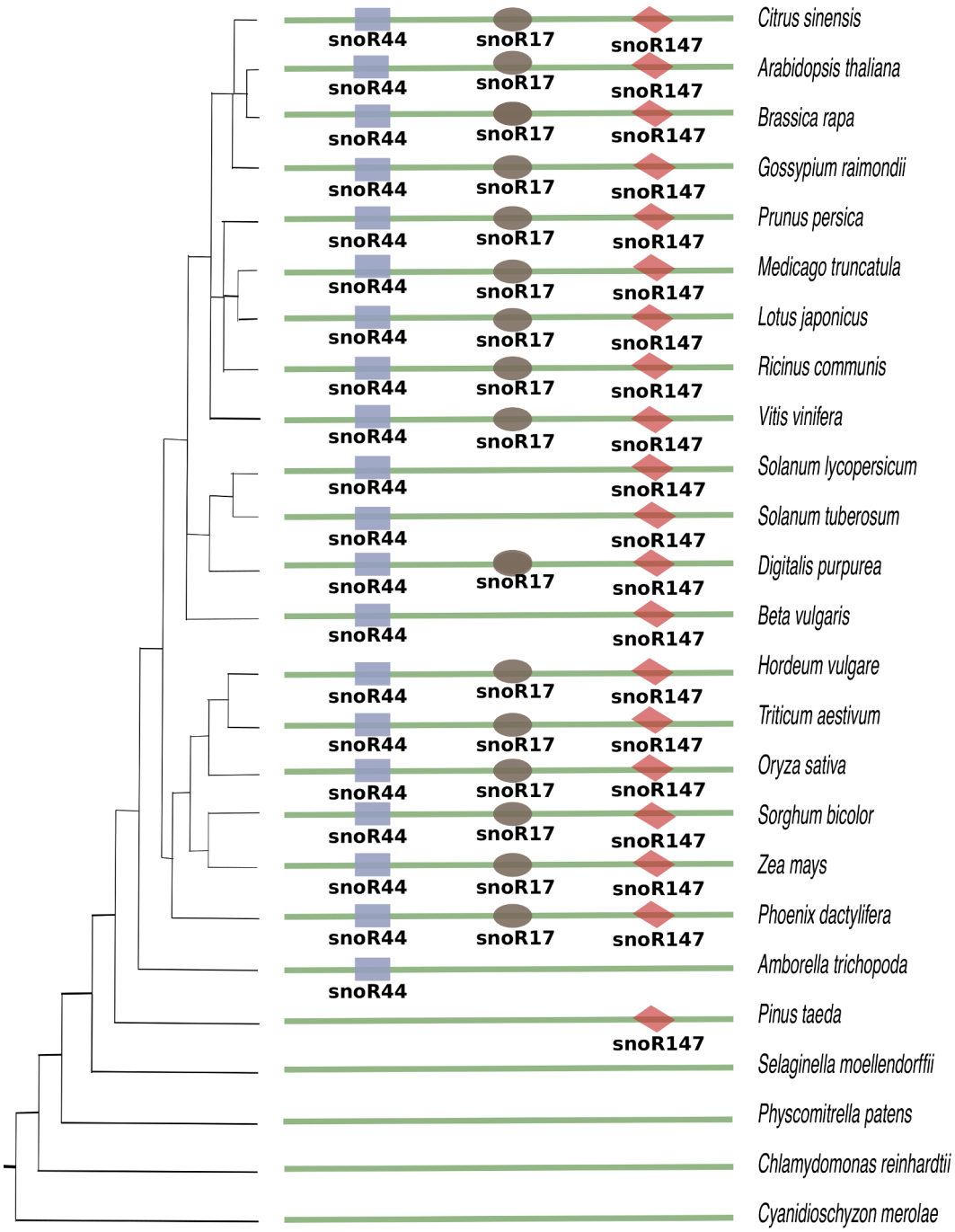


FIGURE 3.10: Evolutionary observation of snoRNA "snoR44-snoR17-snoR147a cluster"

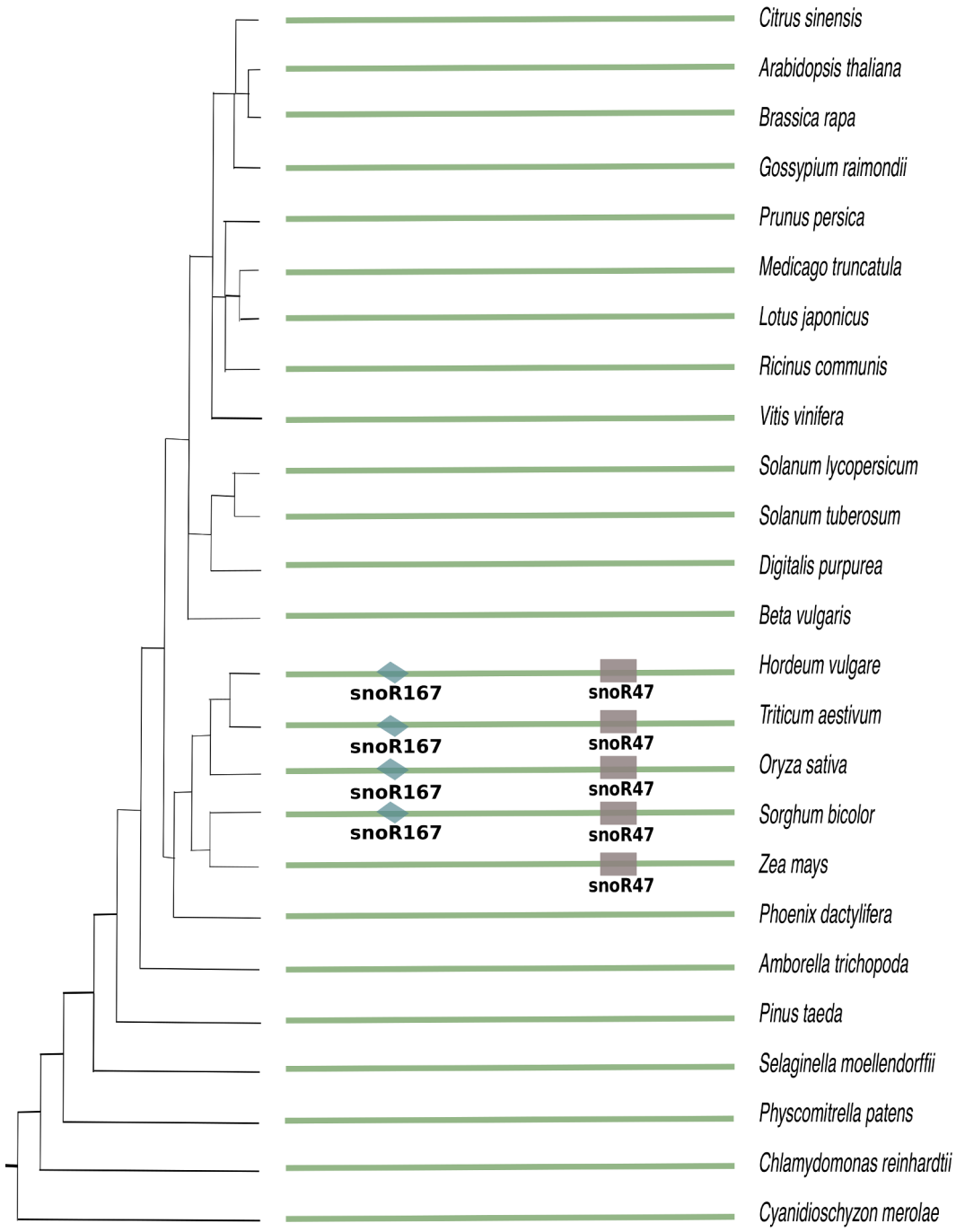


FIGURE 3.11: Evolutionary observation of snoRNA "snoR167-snoR47 cluster"

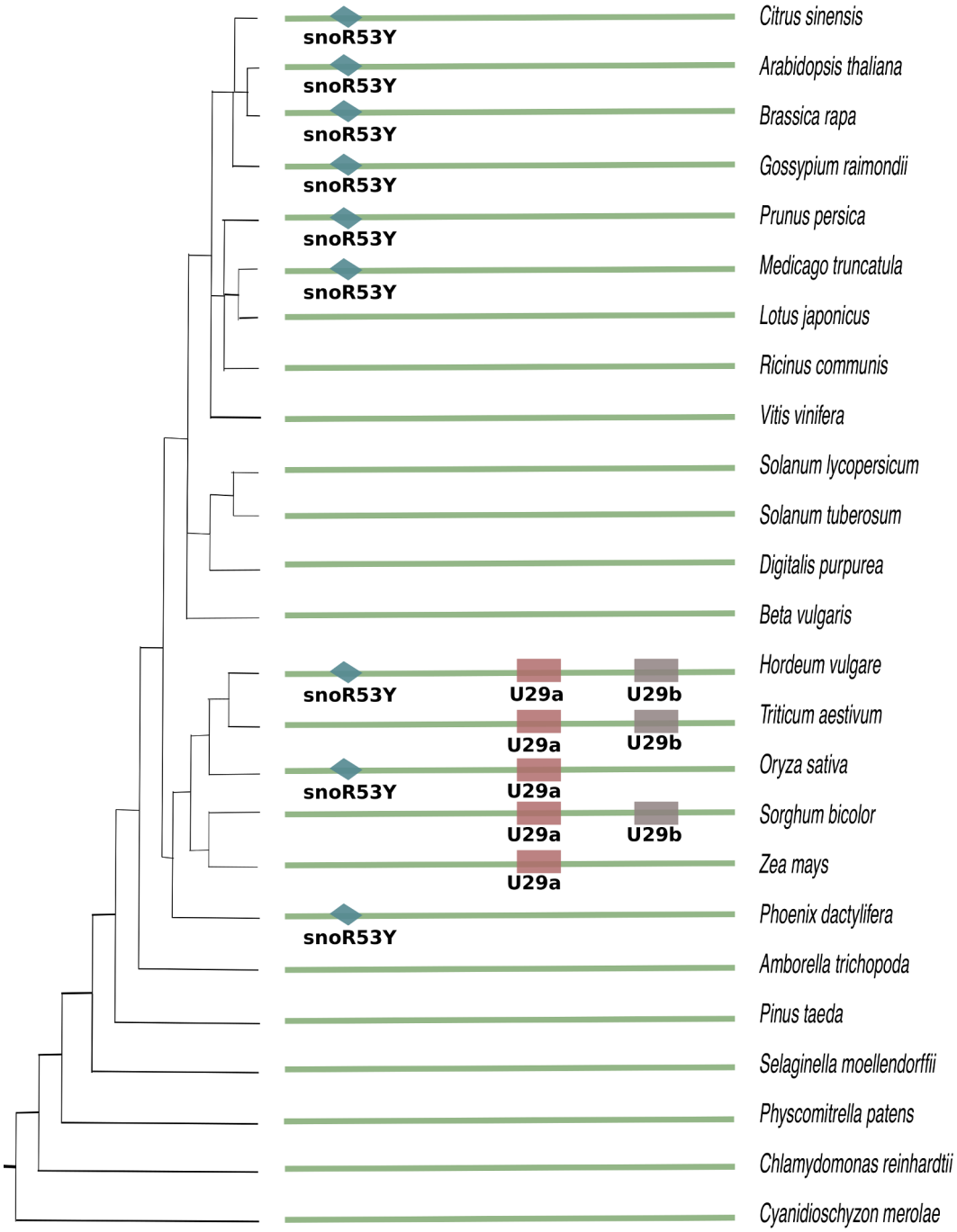


FIGURE 3.12: Evolutionary observation of snoRNA "snoR53Y-U29a-U29b cluster"

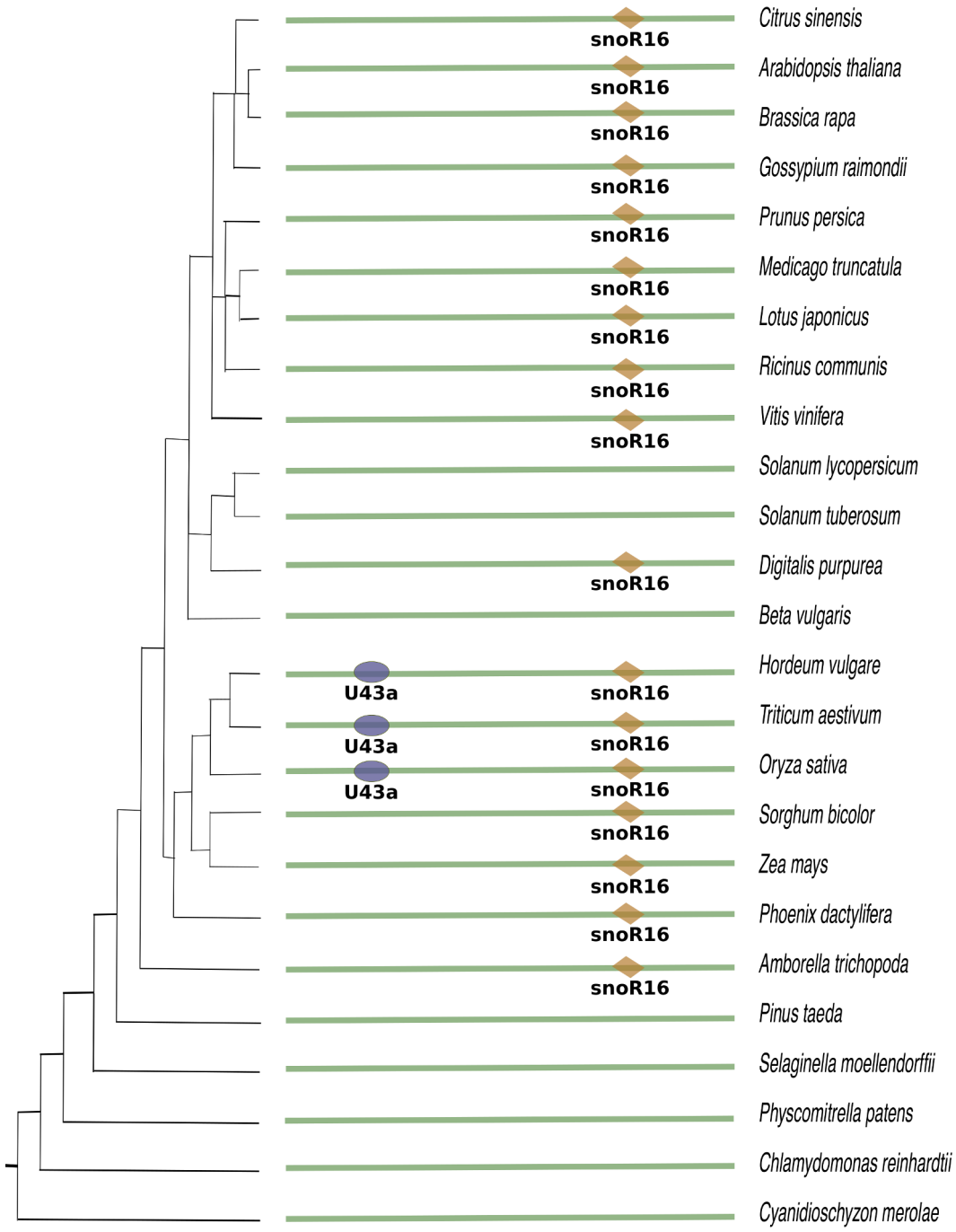


FIGURE 3.13: Evolutionary observation of snoRNA "U43a-snoR16 cluster"

In cluster snoR53Y-U29a-U29b cluster ("cluster 58") (3.12), although snoR53Y emerges in the mesangiospermae family but is not consistently conserved throughout but also re-appears in recent dicots, whereas both U29a and U29b are restricted to monocots.

Cluster U43a-snoR16 ("cluster 66") (3.13) comprising U43a and snoR16, snoR16 seems to date back to magniliophyte ancestor whereas U43a although is a recent addition but restricted to subfamily BOP Clade. This cluster is also already mentioned in *A. thaliana* [155] as "snoR16.1-U43.1 cluster". The conservation of many snoRNA clusters independently strongly supports the results of the homology-based family assignments.

3.3.5 snoRNA targets

Systematic prediction of snoRNA targets in rRNAs and snRNAs showed that known and many predicted targets are usually conserved when the snoRNA is conserved. As an example, Figure 3.14 shows the predicted targets for snoR28 in the ribosomal RNA 18S. While we were able to identify putative targets for most snoRNA families, several orphan snoRNAs (where no target RNAs are found) remain: snoR8, snoR9, snoR106, snoR107, snoR109, snoR112, CrCD72, CrCD74, CrACA54, and CrACA55. Orphan snoRNAs for which we could not find any rRNA or snRNA target may have a different function, e.g. they may target other RNAs such as mRNAs, or they may act as precursor molecules for the production of small regulatory RNAs [85].

Targets are also found to be conserved to a great extent. The target prediction employed by the `snoStrip` pipeline [75] suggests that 12 of the target sites in rRNAs are conserved throughout the plant kingdom. These 12 targest sites, from the aspect of snoRNA families (snoR1, snoR12, snoR14, snoR15, snoR22, snoR24, snoR28, snoR32, snoR37, snoR44, snoR59, U15) are higly conserved as well.

3.3.6 Phylogenetic tree and the evolution of snoRNA families

To draw a comprehensive picture of the snoRNA evolution in the 24 plant species we used the compational approach `ePoPE` [193]. It implements a parsimony-based

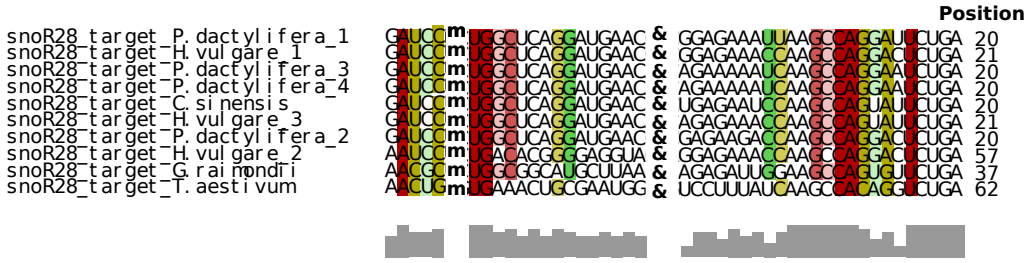


FIGURE 3.14: Conservation of the interaction between the region upstream of D-box of snoRNA family snoR28 (right side) and the region around the 2'-O-methylated cytosine in 18S rRNA (left side). Target RNA segment and ASE are separated by &. The methylated residue is marked with M. The position of the predicted modification in the 18S rRNA sequence within each species is given at the end of each row. Red and green columns highlight conservation of the RNA-RNA interaction. Completely conserved base pairs are shown in red. Green columns mark base pairs with compensatory mutations. Lighter colors indicate loss of base pairs in individual species. The gray bars at the bottom correspond to the degree of sequence conservation. The last three snoR28 paralogs are more divergent and presumably address different targets.

presence/absence analysis of genes within a gene family. Given the phylogenetic tree of our plants of interest and the built alignments this program systematically traced each individual snoRNA family back to its last common ancestor. The `ePoPE` program also returns a most parsimonious solution for the history of gains and losses of genes along the phylogenetic tree. A summary of this study over *all* plant snoRNA families is given in Figures 3.15 (box C/D snoRNAs) and 3.16 (box H/ACA snoRNAs). which includes the annotation of the last common ancestor of this snoRNA family, the predicted number of snoRNA genes that emerged and diverged at each branch and the number of genes that is observed in the species (at the leafs).

3.4 Concluding Remarks

Many snoRNA families are deeply conserved in the plant kingdom. Surprisingly, only a few families can unambiguously be traced back to the ancestor of land plants. Some families are innovations that emerged later during plant evolution. We hypothesize that at least 8 snoRNA families are recent innovations, i.e. snoR59, U29, snoR72Y, snoR6, U31, snoR8, snoR23, and snoR7. This hypothesis is supported by a large group of monocot-specific snoRNAs.

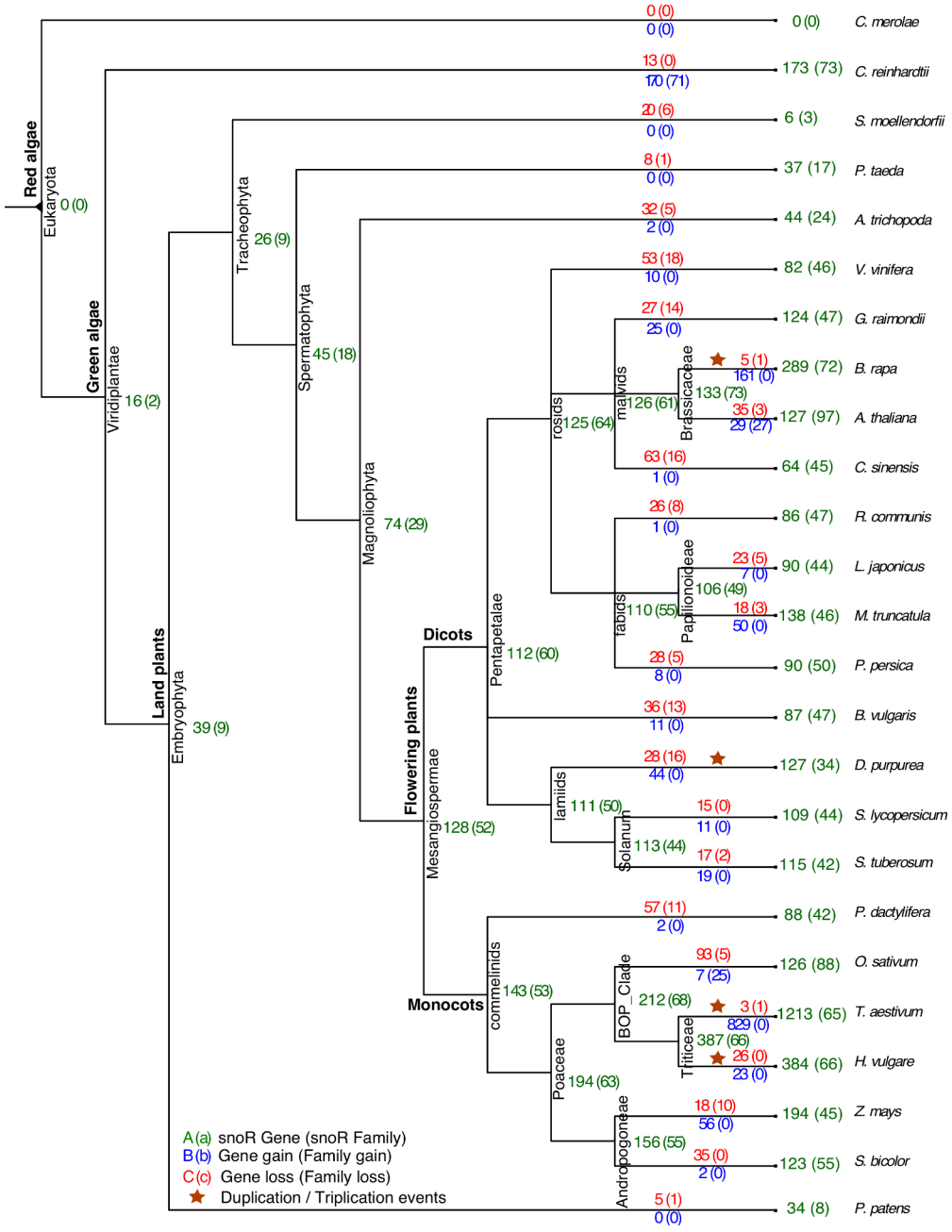


FIGURE 3.15: Phylogenetic tree of C/D snoRNAs of 24 plant species and red alga (*C. merolae*). The phylogenetic tree was constructed from recent literature and NCBI Taxonomy information. The species are assigned to the leaves. The numbers summarize the results of all `ePoPE` runs - that trace each snoRNA family back to its LCA and annotates the inner nodes of the tree with a putative number of observed paralogs. Green numbers refer to the predicted number of observed genes (families) at each node. Red numbers refer to the number of lost genes (families) while blue numbers to the number of gained genes (families). Prominent duplication and triplication events in certain plant species are also depicted in the figure.

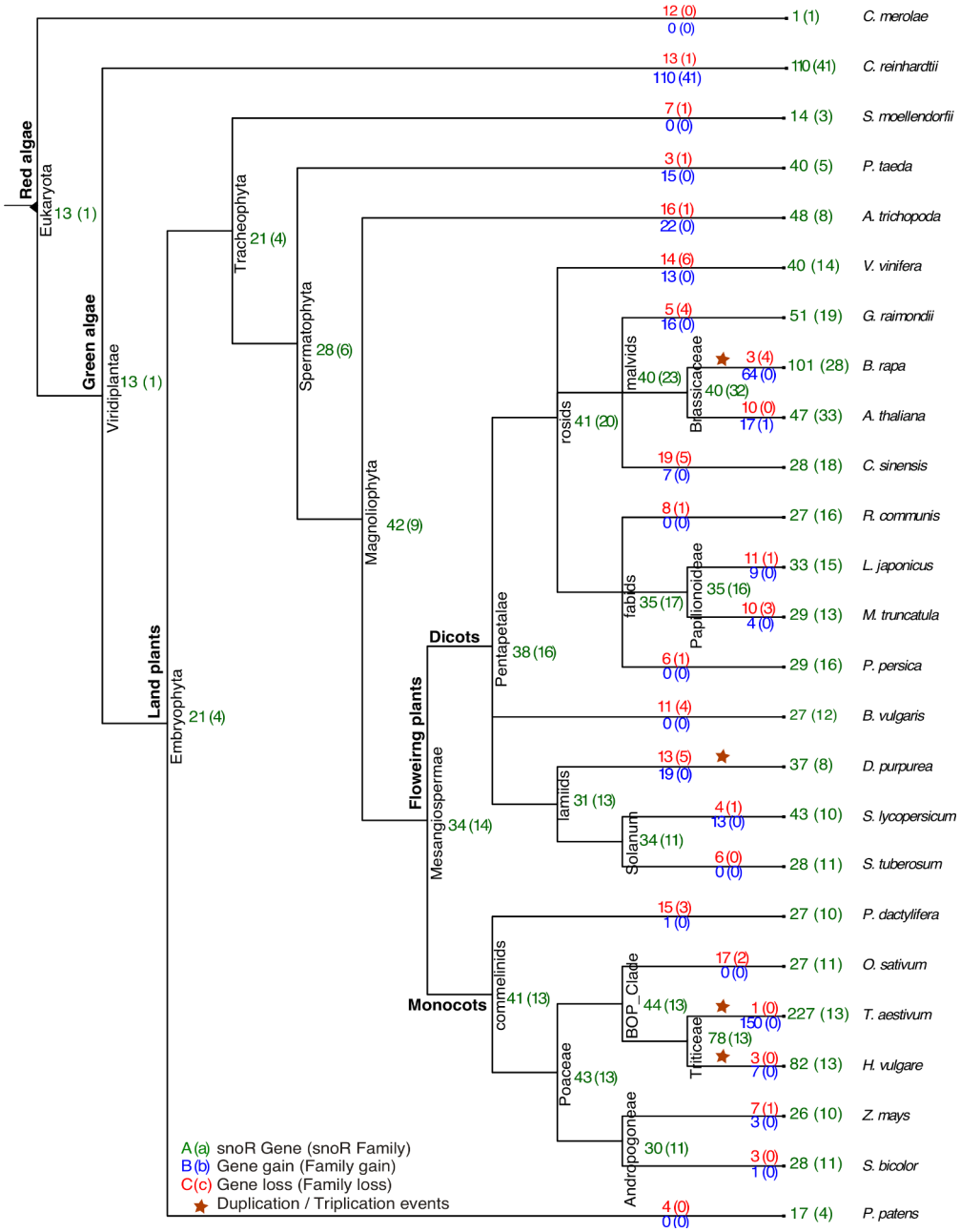


FIGURE 3.16: Phylogenetic tree of H/ACA snoRNAs of 24 plant species and red alga (*C. merolae*). The species are assigned to the leaves. The numbers summarize the results of all `ePoPE` runs - that trace each snoRNA family back to its LCA and annotates the inner nodes of the tree. Green numbers refer to the predicted number of observed genes (families) at each node. Red numbers refer to the number of lost genes (families) while blue numbers to the number of gained genes (families). Prominent duplication and triplication events in certain plant sepecies are depicted in the figure.

The strong conservation of some chemical modification sites in ribosomal RNAs, however, supports the idea that there is a core of snoRNA genes that are ubiquituously present in Eukarya and possibly even in Archaea. Several interesting patterns on snoRNA evolution in plants can be observed. Many snoRNA families have well-identifiable paralogs. Furthermore, distinction between evolutionarily old families and a collection of evolutionarily young innovations is observed see Figs. 3.1 and 3.2.

Chapter 4

Prediction of novel snoRNAs in plants

Since in the last chapter (chapter 3), we found how the annotated snoRNAs are phylogenetically distributed and their evolutionary significance, here we intended to find novel snoRNAs in plants and whether they are conserved to an extent like the annotated snoRNAs.

4.1 Background

Small nucleoar RNAs (snoRNAs), although absent in bacteria but present in archae group demonstrates an ancient origin and the phylogenetic distribution of plant snoRNAs (in chapter 3) proved the conservation of snoRNAs to an extent along the selected plant species which includes green algae, land plants as well as flowering plants. Therefore, since the snoRNAs are evolutionary ancient and categorized into two box C/D and box H/ACA snoRNAs, hence it is speculated that novel predicted snoRNAs should also be conserved in more than one species unless they are really species specific innovation or false prediction. In this context, we thought to implement the application of `plantDARIO` web server, described in chapter 2, which is mainly used for the analysis of small RNAs in plants along with the initial quality control and prediction of novel microRNAs and snoRNAs.

Therefore we intented to combine the studies as well as pipelines from chapter 2 and chapter 3 related to `plantDARIO` and `Phylogenetic distribution of`

`plant snoRNAs` respectively to predict the novel snoRNAs and to find how many of them are conserved. The plan is to combine the workflow pipelines from chapter 2 and chapter 3 in order to find whether novel snoRNAs predicted from any small RNA-seq study can lead to other paralogs in other species and the targets as well.

4.2 Material and Methods

4.2.1 Small RNA-seq dataset

The small RNA-seq dataset from tomato, SRR786984 [168] is downloaded from short read archive. The dataset is then analyzed, sorted then `segemehl` [169] is used with default parameters to map the sequencing data to the respective reference genome (*S.lycopersicum* genome). We use `segemehl`, since it has full support for multiple-mapping reads which is very important for small RNA-seq data. The actual study of SRR786984 small RNA-seq dataset aimed to examine small RNAs from *B. cinerea*-treated tomato leaf and fruit tissue over a time course.

As it is already known, `plantDARIO` implements basic workflows for the analysis of RNA-seq data and allows the user to obtain a comprehensive overview starting after read mapping, we provided the already mapped RNA-seq data as input to `plantDARIO`.

4.2.2 `plantDARIO` analysis

The plantDARIO intially assess the quality and integrity of the data before they are analyzed further; the measures include the number of mappable reads and the number of tags (distinct read sequences), the distribution of read length, and the sequence composition of mapped reads (4.1). Generally a wide variety of errors and biases have been described in high-throughput sequencing data, which may originate from sample handling, library preparation, or the sequencing itself. It is thus very necessary to assess the quality and integrity of the experimental data before they are analyzed for biological content [149–151].

An overview of the dataset is obtained since `plantDARIO` computes a summary of the distribution of reads among annotation items such as introns and exons;

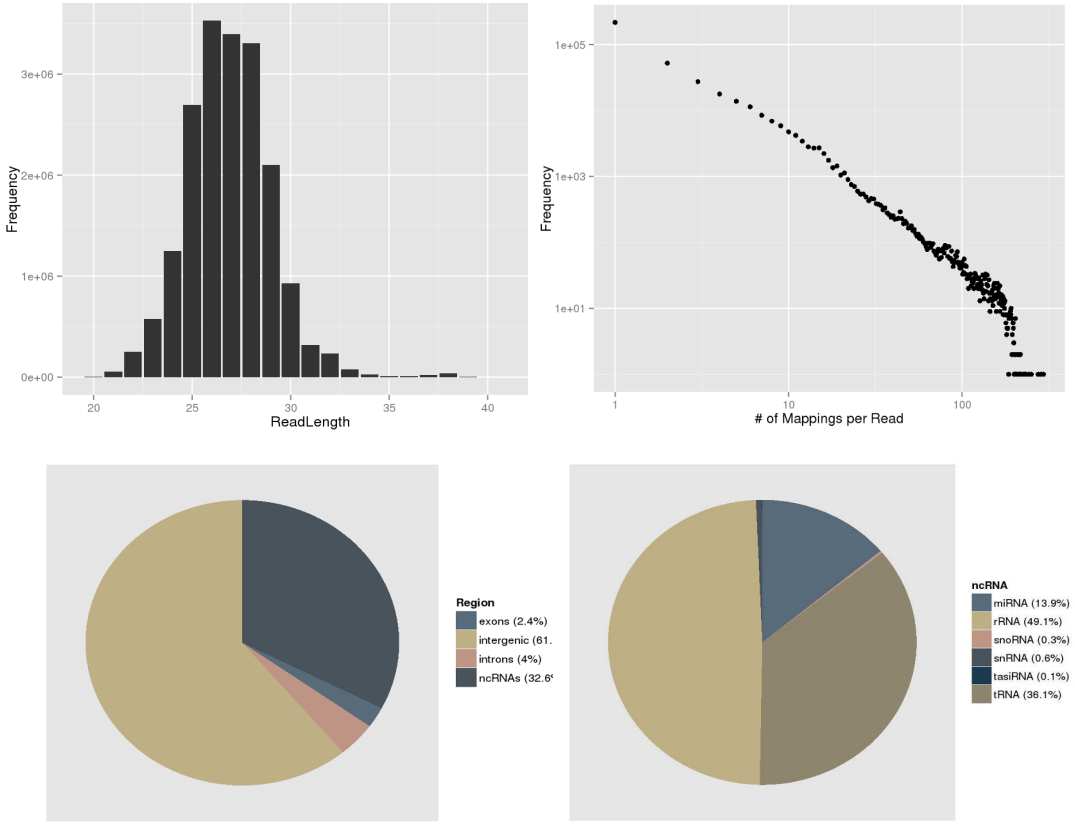


FIGURE 4.1: Initial quality control. `plantDARIO` provides overviews of the read length distribution, the distribution of read-length multiplicities, the distribution of genomic locations, and known annotations (separated into known ncRNAs, exons, introns, and intergenic regions). This is the overview of the dataset SRR786984 from *S. lycopersicum*

the major classes of annotated non-coding RNAs such as miRNA, snRNA, rRNA, tRNA, ta-siRNA, snoRNAs and predict novel miRNAs and snoRNAs. Since already discussed, that contrary to animals, the processing patterns of microRNAs as well as snoRNAs are not very consistent in plants so that patterns of mapped reads alone do not allow a sufficiently accurate classification. Therefore the prediction of microRNAs and snoRNAs is assisted by the integration of `novomir` [161] and `snoReport` [162] in `plantDARIO` as algorithms or scripts locally and interfaced the the output internally to `plantDARIO`. From the output results, predicted novel snoRNAs are derived for further study and analysis.

4.2.3 Curation of the derived predicted snoRNA data

The predicted novel snoRNAs derived are then categorized into box C/D and box H/ACA snoRNAs respectively, and the characteristic boxes (C, D', C', D, H, ACA) are annotated manually using the sequence patterns as constraints given in [191] in the similar way as in chapter 3.

Since from the previous analyses from the Bachellerie laboratory, it is already known that there is conserved spacing between the box C/D core motif and the internal D'/C' motif of the archaeal box C/D snoRNAs[109]. Hence, the box motifs are annotated based on both known pattern of conserved nucleotides and likely spacer distances, usually 12nt, between the box C/D and D'/C' motifs. Only snoRNAs with boxes that could be annotated with high certainty are selected for the initial query set. The sequences are then grouped into gene families based on known orthology and sequence similarity as followed in chapter 3.

4.2.4 Homology search for predicted snoRNA data

In the next all predicted snoRNA families were mapped to all plant genomes similarly as the annotated snoRNAs are mapped in chapter 3. The list of all genomes with accession numbers is provided in Appendix A already mentioned in chapter. The `snoStrip` pipeline [75] was used to search each of the 24 plant genomes for homologs of each of the query families. It is an automatic annotation pipeline that is developed specifically for comparative genomics of snoRNAs which first uses both a `blast` search with relaxed parameters and `infernal` [192] to retrieve initial candidates.

And then like in chapter 3, the expected boxes and the anti-sense elements were annotated based on sequence alignments, and candidates were filtered for the presence of the boxes. And a family-wide alignment of all retained candidate sequences was calculated and the alignments produced by `snoStrip` are manually inspected.

In the final step, data were aggregated to heatmaps showing the number of family members in each species. SnoRNA clusters were identified by proximities of genomic coordinates.

4.2.5 Targets of the predicted snoRNAs

Likewise in chapter 3, we considered the rRNAs/snRNAs as potential targets. Ribosomal RNA sequences of the 24 plant and red algae species are downloaded from the `SILVA` database [189]. The snRNAs comprising of U1, U2, U4, U4atac, U5, U6, U6atac, U11, and U12 are imported from datasets of the `plantDARIO` web server [190].

4.3 Results and Discussion

4.3.1 Heatmaps of predicted snoRNA families

The obtained heatmap represents the number of novel predicted snoRNA families and how they are distributed amongst the selected plants, 4.2. The heatmap shows 11 novel snoRNA families classified into 9 box C/D snoRNA families and 2 box H/ACA families.

The 9 box C/D snoRNA families are found to be present only within "Solanaceae" family, evident in *S.tuberosum* and *S.lycopersicum* only. Perhaps these box C/D snoRNAs are only "Solanaceae" family clade specific innovations and we suspect that these snoRNAs are also found in other "Solanaceae" family members: like eggplant (*S. melongena*), the pepper (*Capsicum annuum*), tobacco (*N.tabacum*, belladonna (*A.belladonna*) and others. The box H/ACA snoRNA families are found to be present comparatively in more species. The H/ACA box snoRNA, sly HACA01 is found to be present in the red alga, green alga, land plants and the monocots. However it is totally absent in the dicots. Whereas, sly HACA02 is absent in red alga, green alga, land plants or monocots but mostly present in the dicots. The targets are found to be the rRNAs.

4.3.2 Expression of predicted novel snoRNA candidates

Most of the predicted novel snoRNAs rom SRR786984 [168] are found to be expressed in the *S.lycopersicum* and other species where they are found. The expression of the novel snoRNAs are observed with visualization browser, see example 4.3.

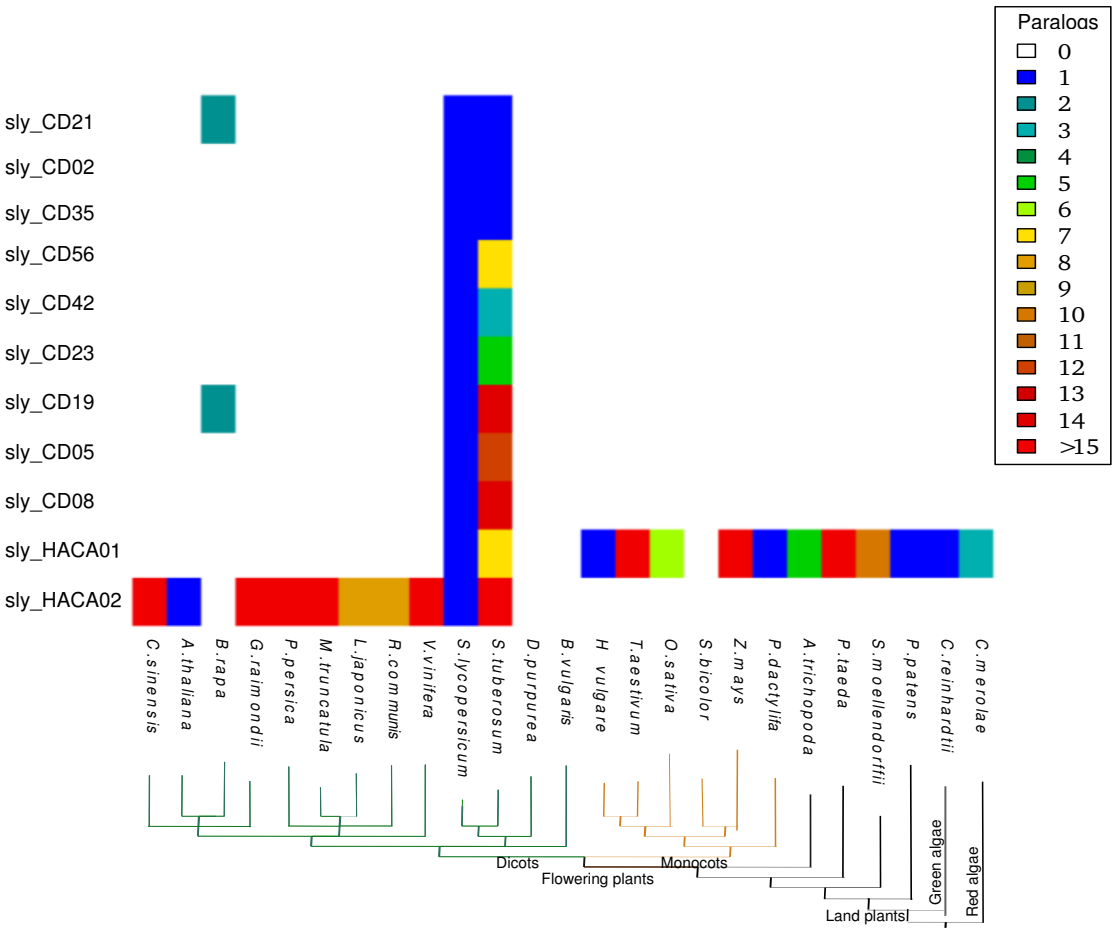


FIGURE 4.2: Heatmap of predicted novel snoRNAs (built in R heatmap.2 version)

Figure 4.3 shows the expression of box H/ACA snoRNA in *S.lycopersicum* named HACA 01 as found in the snoRNA heatmap. The box motifs are quite evident and clear according to the figure which adds an extra support to the fact that the predicted snoRNAs are presumably true and not false positives.

4.4 Concluding Remarks

It can hypothesized from the heatmap that some snoRNAs are family specific innovation (mostly predicted box C/D snoRNAs) which are limited only to Solanaceae family. Nevertheless, the predicted box H/ACA snoRNAs are found to be extensively present and conserved to an extent. The box H/ACA snoRNAs show very interesting patterns. Box H/ACA 01 snoRNA is found to be dating back to algae species and missing in the advanced recent species. Whereas the case is vice versa

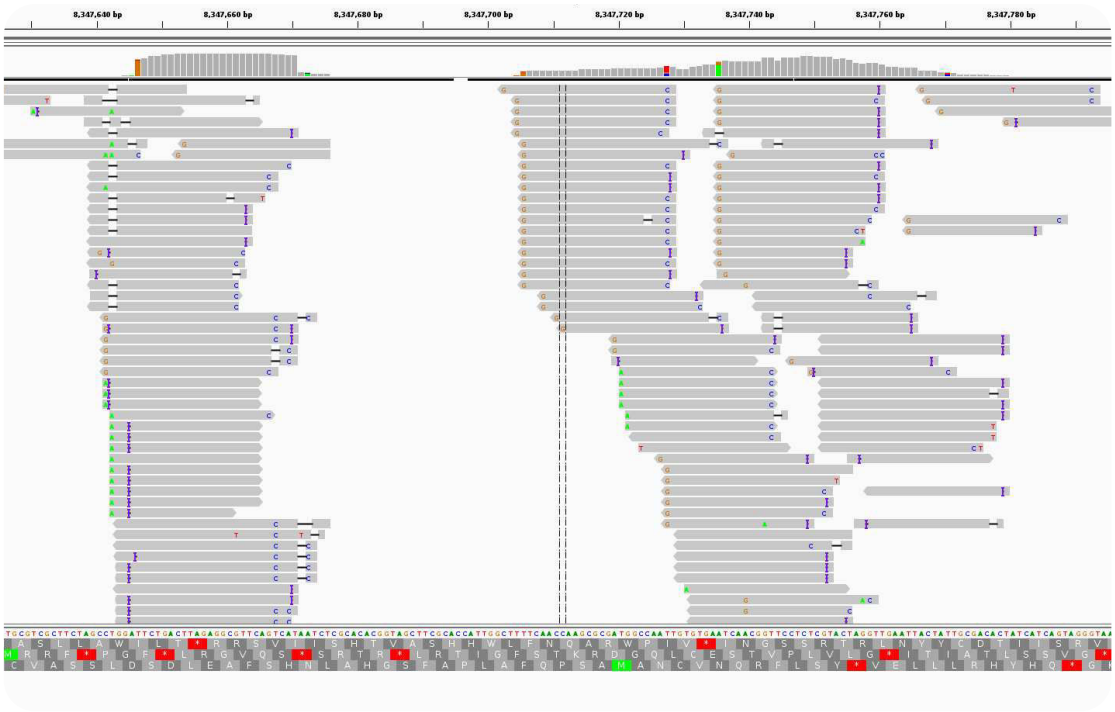


FIGURE 4.3: Visualization of HACA 01 snoRNA gene expression (viewed in IGV genome browser)

in case of box H/ACA 02. It is found only in the advanced recent species and date back only to the Solanaceae family.

Hence, merging the pipelines of `plantDARIO` and phylogenetic distribution in case of other small RNA-seq datasets can identify more new snoRNAs leading to many interesting facts and new inputs important in plants.

Chapter 5

Conclusions

Since, high-throughput sequencing now has become one of the choicest methods for the analysis of transcriptome data and for special case of small RNA-seq data, web servers provide convenient means of conducting standard analyses. In fact, web servers help to avoid the the need to install, maintain, and update an array of individual tools. From chapter 2, we find that `plantDARIO` is such a web server which provides a ready-to-use analysis workflow for small RNA-seq data. In short, `plantDARIO` provides the user with a valuable combination of annotation-based, standardized quantitative analysis and a simple facility for guided discoveries of novel small RNA loci as well as to carry out other analyses i.e. the web server also provides the results in a bed format, which can easily be used for downstream analysis tasks such as the assessment of differential expression, for example 2.7, discussed in chapter 2.

Hence, the best part is that we can even use publicly available small RNA-seq data in order to do any comparative analysis. Using publicly available dataset e.g. SRR167709 and SSR167710 datasets from *A. thaliana*, analyzing them, we noticed extreme differences in the levels of small RNAs processed from box C/D snoRNAs. Some of these snoRNAs are known to have a regulatory role in animals, so it might be of possible interest to further characterize small RNA processing from "house-keeping ncRNAs" in plants, and `plantDARIO` seems to be a convenient and versatile tool which helps to serve this cause.

From the analysis of small non-coding RNAs, we observed small nucleolar RNAs (snoRNAs) are found to be the most ancient as well as conserved families amongst non-protein-coding RNAs. Hence we thought to use phylogenetic tree analysis

in order find the distribution of snoRNAs in the plants belonging to different hierarchical families and clades as discussed in chapter 3. The phylogenetic tree comprises major clades from algae to the flowering plants, which is used for the phylogenetic study of the distribution of the plant snoRNAs.

Detail study of the phylogenetic distribution in chapter 3 led to many interesting facts and informations. Many snoRNA families are found to be deeply conserved in the plant kingdom. However, surprisingly, only a few families can unambiguously be traced back to the ancestor of land plants and some families are found to be innovations that emerged later during plant evolution. We hypothesize from the study that at least 8 snoRNA families are recent innovations, i.e. snoR59, U29, snoR72Y, snoR6, U31, snoR8, snoR23, and snoR7, which is supported by a large group of monocot-specific snoRNAs amongst the flowering plants.

The strong conservation of some chemical modification sites in ribosomal RNAs, however, supports the idea that there is a core of snoRNA genes that are ubiquitously present in Eukarya and possibly even in Archaea. The small size, the relative fast rate of evolution, and limitations of available homology search techniques, however, make it hard to directly test this hypothesis. Surprisingly, homology search methods fail, with very few exceptions, to identify homologs of landplant snoRNAs in green algae. We suspect, however, that this rather a limitation of the state of the art in homology search.

Despite the limitations, several interesting patterns on snoRNA evolution in plants are observed. Many snoRNA families also have well-identifiable paralogs. Furthermore, distinction between evolutionarily old families and a collection of evolutionarily young innovations is observed see Figs. 3.1 and 3.2 in chapter 3. The rapidly increasing collection of completely sequenced rosids, for example, can serve as an excellent starting point for a systematic study of snoRNA turnover.

The main challenge in the part was the nomenclature of plant snoRNAs, since it is often species specific and it respects only partially known orthology relationships at the level of individual snoRNAs families. In particular, this is the case where data go beyond the plant snoRNA database [184]. In some cases, naming convention for different species are even contradictory. This poses a serious obstacle for large-scale comparative studies and causes the danger of mis-interpreting the results of comparative surveys. Hence in chapter 3, in this contribution, we used the *Arabidopsis* or *Oryza* names for snoRNA families wherever possible based on the

assumption that these are most widely used. A comprehensive table of synonyms is provided in Appendix B.

The nomenclature of such type in plant snoRNAs is a great asset in order to facilitate comparative studies. This nomenclature is (a) designed to be applicable to all (land) plant species, (b) strives to honor homologies, and (c) distinguishes box H/ACA and box C/D snoRNAs.

Therefore in chapter 3, we studied as well as analyzed all the available annotated plant snoRNAs and provide a comprehensive, well curated collection of homologous snoRNAs in 24 plant species evenly covering the plant kingdom. For each individual snoRNA family we prepared and analyzed multiple sequence alignments in the `Rfam`-compatible STOCKHOLM[1] format. These alignments would only help to count the number of species aligned but also to see how the box motifs are conserved in the species.

Moreover, apart from the aligned sequences these files contain the predicted conserved secondary structure and the positions of the characteristic box motifs of snoRNAs. In addition, all data regarding target prediction, snoRNA distribution and evolution provided valuable resourceful information regarding snoRNAs and their evolution in the plant kingdom.

In chapter 4, we intended to find and see whether novel snoRNAs when annotated are conserved or how they are distributed in the plant kingdom. Hence, we used plantDARIO to predict novel snoRNAs from small RNA-seq dataset of tomato SRR786984 [168] and studied their phylogenetic distribution. We find that some snoRNAs are family specific innovations (most predicted box C/D snoRNAs). Whereas the predicted box H/ACA snoRNAs are found to be extensively present and conserved to an extent in the plant kingdom. From the genomic coordinates, we intended to find the neighbouring genes and their nature in order to predict the characteristics of the novel genes.

Hence to conclude overall, the study covered the types of small non-coding RNAs, their annotations, differentiation and analyses with `plantDARIO` including further analysis. The study also covered the annotated plant snoRNAs and their evolutionary significance through phylogenetic distribution along with annotation and phylogenetic distribution of novel predicted snoRNAs.

[1] https://en.wikipedia.org/wiki/Stockholm_format

Bibliography

[1] A Q Gomes, S Nolasco, and H Soares. Non-Coding RNAs: Multi-Tasking molecules in the cell. *Int J Mol Sci*, 14:16010–16039, 2013.

[2] T R Cech, A J Zaug, and P J Grabowski. Tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27: 487–496, 1981.

[3] C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman. The rna moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35: 849–857, 1983.

[4] J P Dworkin, A Lazcano, and S L Miller. The roads to and from the RNA world. *J. Theor. Boil*, 222:127–134, 2003.

[5] Lehman N. RNA in evolution. *Wiley Interdiscip. Rev*, 1:202–213, 2010.

[6] S M Berget, C Moore, and P A Sharp. Spliced segments at the 50 terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*, 74:3171–3175, 1977.

[7] L T Chow, R E Gelinas, T R Broker, and R J Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12:1–8, 1977.

[8] H Busch, R Reddy, L Rothblum, and Y C Choi. SnRNAs, SnRNPs, and RNA processing. *Annu. Rev. Biochem*, 51:617–654, 1982.

[9] R A Weinberg and S Penman. Small molecular weight monodisperse nuclear RNA. *J. Mol. Biol*, 38:289–304, 1968.

[10] L J Collins, B Schönfeld, and X S Chen. *The epigenetics of non-coding RNA. In T. Tollefsbol (Ed.).* Handbook of epigenetics: the new molecular and medical genetics, London: Academic, 2011.

[11] S Hajduk and T Ochsenreiter. RNA editing in kinetoplastids. *RNA Biol*, 7:229–236, 2010.

[12] S Choudhuri. Small noncoding rnas: biogenesis, function, and emerging significance in toxicology. *J Biochem Mol Toxicol*, 24:195–216, 2009.

[13] M J Axtell. Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol*, 64:137–159, 2013.

[14] M R Willmann, M W Endres, R T Cook, and B D Gregory. *The functions of RNAdependent RNA polymerases in Arabidopsis*, volume 9. Arabidopsis Book, 2011.

[15] K Mukherjee, H Campos, and B Kolaczkowski. Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. *Mol. Biol. Evol*, 30:627–641, 2013.

[16] B Czech and G J Hannon. Small RNA sorting: matchmaking for Argonautes. *Nat. Rev. Genet*, 12:19–31, 2010.

[17] H Wang, X Zhang, J Liu, T Kiba, J Woo, T Ojo, M Hafner, T Tuschl, NH Chua, and XJ Wang. Deep sequencing of small RNAs specifically associated with arabidopsis AGO1 and AGO4 uncovers new AGO functions. *Plant J*, 67:292–304, 2011.

[18] X Wang, JD Laurie, T Liu, J Wentz, and XS Liu. Computational dissection of arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites. *Genomics*, 97:235–243, 2011.

[19] F Borges and R A Martienssen. The expanding world of small rnas in plants. *Nat Rev Mol Cell Biol*, 16:727–741, 2015.

[20] NG Bologna and O Voinnet. The diversity, biogenesis, and activities of endogenous silencing small RNAs in arabidopsis. *Annu Rev Plant Biol*, 65: 473–503, 2014.

[21] YX Liu, M Wang, and XJ Wang. Endogenous small RNA clusters in plants. *Genomics Proteomics Bioinformatics*, 12:64–71, 2014.

[22] V Ramachandran and X Chen. Small RNA metabolism in Arabidopsis. *Trends Plant Sci*, 13:368–74, 2008.

[23] A C Mallory and H Vaucheret. Functions of microRNAs and related small RNAs in plants. *Nat Genet*, 38:S31–S36, 2006.

[24] W Wei, Z Ba, M Gao, Y Wu, Y Ma, S Amiard, C I White, J M Rendtlew Danielsen, Y G Yang, and Y Qi. A role for small RNAs in DNA double-strand break repair. *Cell*, 149:101–112, 2012.

[25] M Yoshikawa. Biogenesis of trans-acting siRNAs, endogenous secondary siRNAs in plants. *Genes Genet Syst.*, 88:77–84, 2013.

[26] Marjori A Matzke and Rebecca A Mosher. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature Reviews Genetics*, 6:394–408, 2014.

[27] K Hamashima, M Tomita, and A Kanai. Expansion of noncanonical v-arm-containing trnas in eukaryotes. *Mol Biol Evo*, 33:530–540, 2015.

[28] W H McClain. Rules that govern trna identity in protein synthesis. *J Mol Biol*, 234:257–280, 1993.

[29] M Sprinzl, C Horn, M Brown, A Ioudovitch, and S Steinberg. Compilation of trna sequences and sequences of trna genes. *Nucleic Acids Res*, 26: 148–153, 1998.

[30] V Biou, A Yaremchuk, M Tukalo, and S Cusack. The 2.9 a crystal structure of t. thermophilus seryl-trna synthetase complexed with trna(ser). *Science*, 263:1404–1410, 1994.

[31] K Breitschopf, T Achsel, K Busch, and H J Gross. Identity elements of human tRNA(leu): structural requirements for converting human tRNA(ser) into a leucine acceptor in vitro. *Nucleic Acids Res*, 23: 3633–3637, 1995.

[32] A Soma, K Uchiyama, T Sakamoto, M Maeda, and H Himeno. Unique recognition style of tRNA(leu) by haloferax volcanii leucyl-trna synthetase. *J Mol Biol*, 293:1029–1038, 1999.

[33] A Yaremchuk, I Kriklivyi, M Tukalo, and S Cusack. Class i tyrosyl-tRNA synthetase has a class ii mode of cognate trna recognition. *EMBO J*, 21: 3829–3840, 2002.

[34] T Achsel and H J Gross. Identity determinants of human tRNAser: sequence elements necessary for serylation and maturation of tRNA with a long extra arm. *EMBO J*, 12:3333–3338, 1993.

[35] L Maréchal-Drouard, J H Weil, and A Dietrich. Transfer RNAs and transfer RNA genes in plants. *EMBO J*, 21:3829–3840, 2002.

[36] D Beier, N Stange, H J Gross, and H Beier. Nuclear tRNA(Tyr) genes are highly amplified at a single chromosomal site in the genome of Arabidopsis thaliana. *Mol.Gen.Genet*, 225:72–80, 1991.

[37] S Smit, J Widmann, and R Knight. Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Res*, 35:3339–3354, 2007.

[38] J R Cole, B Chai, T L Marsh, R J Farris, Q Wang, S A Kulam, S Chandra, D M McGarrell, T M Schmidt, G M Garrity, and J M Tiedje. The ribosomal database project (RDP-II): Previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*, 31:442–443, 2003.

[39] J Wuyts, Y Van de Peer, T Winkelmans, and R De Wachter. The European database on small subunit ribosomal RNA. *Nucleic Acids Res*, 30:183–185, 2002.

[40] R Appels and J Dvoŕáfik. The wheat ribosomal DNA spacer region: Its structure and variation in populations and among species. *Theor Appl Genet*, 63:337–348, 1982.

[41] J Wuyts, Y Van de Peer, T Winkelmans, and R De Wachter. The european database on small subunit ribosomal rna. *Nucleic Acids Res*, 30: 183–185, 2008.

[42] N G Bologna, A L Schapire, and J F Palatnik. Processing of plant microrna precursors. *Brief Funct Genomics*, 12:37–45, 2012.

[43] Y Kurihara and Y Watanabe. Processing of microrna precursors. *Methods Mol Biol*, 592:231–241, 2010.

[44] R W Carthew and E J Sontheimer. Origins and mechanisms of miRNAs and siRNAs. *Cell*, 136:642–655, 2009.

[45] V N Kim, J Han, and M C Siomi. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol*, 10:126–139, 2009.

[46] E Allen, Z Xie, A M Gustafson, G H Sung, J W Spatafora, and J C Carrington. Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nat Genet*, 36:1282–1290, 2004.

[47] G Chuck, A M Cigan, K Saeteurn, and S Hake. The heterochronic maize mutant corngrass1 results from overexpression of a tandem microRNA. *Nat Genet*, 39:544–549, 2007.

[48] A Boualem, P Laporte, M Jovanovic, C Laffont, J Plet, J P Combier, A Niebel, M Crespi, and F Frugier. MicroRNA166 controls root and nodule development in medicago truncatula. *PlantJ*, 54:876–887, 2008.

[49] W Park, J Li, R Song, J Messing, and X Chen. CARPEL FACTORY, a dicer homolog, and HEN1, a novel protein, act in microrna metabolism in Arabidopsis thaliana. *Curr Biol*, 12:1484–1495, 2002.

[50] Y Kurihara and Y Watanabe. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A*, 101:12753–12758, 2004.

[51] M H Han, S Goud, L Song, and N Fedoroff. The Arabidopsis double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation. *Proc Natl Acad Sci U S A*, 101: 1093–1098, 2004.

[52] F Vazquez, T Blevins, J Ailhas, T Boller, and F Jr. Meins. Evolution of arabidopsis MIR genes generates novel microRNA classes. *Nucleic Acids Res*, 36:6429–6438, 2004.

[53] D Lobbes, G Rallapalli, D D Schmidt, C Martin, and J Clarke. SERRATE: a new player on the plant microRNA scene. *EMBO Rep*, 7: 1052–1058, 2006.

[54] L Yang, Z Liu, F Lu, A Dong, and H Huang. SERRATE is a novel nuclear regulator in primary microRNA processing in arabidopsis. *Plant J*, 47: 841–850, 2006.

[55] B Yu, L Bi, B Zheng, L Ji, D Chevalier, M Agarwal, V Ramachandran, W Li, T Lagrange, J C Walker, and X Chen. The FHA domain proteins DAWDLE in arabidopsis and SNIP1 in humans act in small RNA biogenesis. *Proc Natl Acad Sci U S A*, 105:10073–10078, 2008.

[56] M Y Park, G Wu, A Gonzalez-Sulser, H Vaucheret, and R S Poethig. Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A*, 102:3691–3696, 2005.

[57] S Boutet, F Vazquez, J Liu, C Béclin, M Fagard, A Gratias, J B Morel, P Crété, X Chen, and H Vaucheret. Arabidopsis HEN1: a genetic link between endogenous miRNA controlling development and siRNA controlling transgene silencing and virus resistance. *Curr Bio*, 13:843–848, 2003.

[58] B D Gregory, R C O'Malley, R Lister, M A Urich, J Tonti-Filippini, H Chen, A H Millar, and J R Ecker. A link between RNA metabolism and silencing affecting Arabidopsis development. *Dev Cell*, 14:854–866, 2008.

[59] R Rajagopalan, H Vaucheret, J Trejo, and D P Bartel. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*, 20:3407–3425, 2006.

[60] K Okamura, J W Hagen, H Duan, D M Tyler, and E C Lai. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell*, 130:89–100, 2007.

[61] J G Ruby, C H Jan, and D P Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448:83–86, 2007.

[62] Q H Zhu, A Spriggs, L Matthew, L Fan, G Kennedy, F Gubler, and C Helliwell. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res*, 18:1456–1465, 2008.

[63] Yijun Meng, Chaogang Shao, Xiaoxia Ma, Huizhong Wang, and Ming Chen. Expression-based functional investigation of the organ-specific microRNAs in Arabidopsis. *PLoS One*, 11:e50870, 2012.

[64] M R Lerner and J A Steitz. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc Natl Acad Sci U S A*, 76:5495–5499, 1979.

[65] J Vogel and B F Luisi. Hfq and its constellation of RNA. *Nat Rev Microbiol*, 9:578–589, 2011.

[66] B Wirth, L Brichta, and E Hahnen. Spinal muscular atrophy and therapeutic prospects. *Prog Mol Subcell Biol*, 44:109–132, 2006.

[67] J Zong, X Yao, J Yin, D Zhang, and H Ma. Evolution of the RNA-dependent RNA polymerase (RdRP)genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene*, 447:29–39, 2009.

[68] R K Slotkin, M Vaughn, F Borges, M Tanurdzic, and J D Becker. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, 136:461–472, 2007.

[69] M A Matzke and R A Mosher. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet*, 15:570, 2014.

[70] K D Kasschau, N Fahlgren, E J Chapman, C M Sullivan, J S Cumbie, S A Givan, and J C Carrington. Genome-wide profiling and analysis of arabidopsis siRNAs. *PLoS Biol*, 5:e57, 2007.

[71] C Lu, K Kulkarni, F F Souret, R MuthuValliappan, S S Tej, R S Poethig, I R Henderson, S E Jacobsen, W Wang, P J Green, and B C Meyers. Microrna and other small rnas enriched in the arabidopsis rna-dependent rna polymerase-2 mutant. *Genome Res*, 16:1276–1288, 2006.

[72] R A Mosher and C W Melnyk. siRNAs and DNA methylation: seedy epigenetics. *Trends Plant Sci*, 15:204–210, 2010.

[73] J Rodor, I Letelier, L Holuigue, and M Echeverria. Nucleolar RNPs: from genes to functional snoRNAs in plants. *Biochem Soc Trans*, 38:672–676, 2010.

[74] J P Bachellerie, J Cavaillé, and A Hüttenhofer. The expanding snoRNA world. *Biochimie*, 84:775–790, 2002.

[75] S Bartschat, S Kehr, H Tafer, P F Stadler, and Hertel J. snoStrip: a snoRNA annotation pipeline. *Bioinformatics*, 30:115–116, 2014.

[76] G Dieci, M Preti, and B Montanini. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, 94:83–88, 2009.

[77] A G Matera, R M Terns, and M P Terns. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Genomics*, 8:209–220, 2007.

[78] T Kiss and W Filipowicz. Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns. *Genes Dev*, 9:1411–1424, 1995.

[79] C Allmang, J Kufel, G Chanfreau, P Mitchell, E Petfalski, and D Tollervey. Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *EMBO J*, 18:5399–5410, 1999.

[80] E Caffarelli, L Maggi, A Fatica, E De Gregorio, P Frangapane, and I Bozzoni. Processing of the intron-encoded U16 and U18 snoRNAs: the conserved c and d boxes control both the processing reaction and the stability of the mature snoRNA. *EMBO J*, 15:1121–1131, 1996.

[81] W Filipowicz and V Pogacić. Biogenesis of small nucleolar ribonucleoproteins. *Curr Opin Cell Biol*, 14:319–327, 2002. doi:10.1016/S0955-0674(02)00334-4.

[82] Q M Mitrovich, B B Tuch, F M De La Vega, C Guthrie, and A D Johnson. Evolution of yeast noncoding RNAs reveals an alternative mechanism for widespread intron loss. *Science*, 330:838–841, 2010.

[83] J W Brown, M Echeverria, and L H Qu. Plant snoRNAs: functional evolution and new modes of gene expression. *Trends Plant Sci*, 8:42–49, 2003.

[84] C L Chen, D Liang, H Zhou, M Zhuo, Y Q Chen, and L H Qu. The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from Oryza sativa. *Nucleic Acids Res*, 31:2601–2613, 2003.

[85] SH Kim, M Spensley, S K Choi, C P Calixto, A F Pendle, O Koroleva, P J Shaw, and J W Brown. Plant U13 orthologues and orphan snoRNAs identified by RNomics of RNA from Arabidopsis nucleoli. *Nucleic Acids Res*, 38:3054–3067, 2010.

[86] M Michaud, V Cognat, A M Duchêne, and L Maréchal-Drouard. A global picture of tRNA genes in plant genomes. *Plant J.*, 66:80–93, 2011.

[87] D J Leader, G P Clark, J Watters, A F Beven, P J Shaw, and J W Brown. Splicing-independent processing of plant box C/D and box H/ACA small nucleolar RNAs. *Plant Mol. Biol*, 39:1091–1100, 1999.

[88] N Watkins, V Segault, B Charpentier, S Nottrott, P Fabrizio, A Bachi, M Wilm, M Roshbash, C Branlant, and R Luhrrmann. A common core RNP structure shared between the small nucleolar box C/D RNPs and the spliceosomal U4 snRNP. *Cell*, 103:457–466, 2000.

[89] D Klein, T Schmeing, P Moore, and T Steitz. The kink-turn: A new RNA secondary structure motif. *EMBO J*, 20:4214–4221, 2001.

[90] J Kuhn, E Tran, and E S Maxwell. Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. *Nucleic Acids Res*, 30:931–941, 2002.

[91] J Venema and D Tollervey. Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu. Rev. Genet*, 33:261–311, 1999.

[92] T Kiss. Small nucleolar RNAs: an abundant group of non-coding RNAs with diverse cellular functions. *Cell*, 109:145–148, 2002.

[93] T Kiss. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J*, 20:3617–3622, 2001.

[94] K S McKeegan, C M Debieux, S Boulon, E Bertrand, and N J Watkins. A dynamic scaffold of pre-snoRNP factors facilitates human box C/D snoRNP assembly. *Mol Cell Biol*, 27:6782–6793, 2007.

[95] S Boulon, N Marmier-Gourrier, B Pradet-Balade, L Wurth, C Verheggen, B Jády B E, Rothé, C Pescia, M C Robert, T Kiss, B Bardoni, A Krol, C Branlant, C Allmang, E Bertrand, and B Charpentier. The Hsp90 chaperone controls the biogenesis of L7Ae RNPs through conserved machinery. *J Cell Biol*, 180:579–595, 2008.

[96] C Torchet, G Badis, F Devaux, G Costanzo, M Werner, and A Jacquier. The complete set of h/aca snornas that guide rrna pseudouridylations in saccharomyces cerevisiae. *RNA*, 11:928–938, 2005.

[97] A G Balakin, L Smith, and M J Fournier. The rna world of the nucleolus: two major families of small rnas defined by different box elements with related functions. *Cell*, 86:823–834, 1996.

[98] H Wang, D Boisvert, K K Kim, R Kim, and S H Kim. Crystal structure of a fibrillarin homologue from Methanococcus jannaschii, a hyperthermophile, at 1.6 A resolution. *EMBO J*, 19:317–323, 2000.

[99] X Darzacq, N Kittur, S Roy, Y Shav-Tal, R H Singer, and U T Meier. Stepwise RNP assembly at the site of H/ACA RNA transcription in human cells. *J Cell Bio*, 173:207–218, 2006.

[100] P K Yang, C Hoareau, C Froment, B Monsarrat, Y Henry, and G Chanfreau. Cotranscriptional recruitment of the pseudouridylsynthetase Cbf5p and of the RNA binding protein Naf1p during H/ACA snoRNP assembly. *Mol Cell Biol*, 25:3295–3304, 2005.

[101] M Ballarino, M Morlando, F Pagano, A Fatica, and I Bozzoni. The cotranscriptional assembly of snoRNPs controls the biosynthesis of h/aca snornas in Saccharomyces cerevisiae. *Mol Cell Biol*, 25:5396–5403, 2005.

[102] J Venema, H R Vos, A W Faber, W J van Venrooij, and H A Raué. Yeast Rrp9p is an evolutionarily conserved U3 snoRNP protein essential for early pre-rRNA processing cleavages and requires box C for its association. *RNA*, 6:1660–1671, 2000.

[103] D L J Lafontaine and D Tollervey. The function and synthesis of ribosomes. *Nat. Rev. Mol. Cell Biol*, 2:514–520, 2001.

[104] Manja Marz and Peter F. Stadler. Comparative analysis of eukaryotic U3 snoRNAs. *RNA Biol.*, 6:503–507, 2009.

[105] Michelle Scott and Motoharu Ono. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie*, 93:1987–1992, 2011.

[106] T T Liu, D Zhu, W Chen, W Deng, H He, G He, B Bai, Y Qi, R Chen, and X W Deng. A global identification and analysis of small nucleolar RNAs and possible intermediate-sized non-coding RNAs in Oryza sativa. *Mol Plant*, 6:830–486, 2013.

[107] Eva K Herter, Maria Stauch, Maria Gallant, Elmar Wolf, Thomas Raabe, and Peter Gallant. snoRNAs are a novel class of biologically relevant Myc targets. *BMC Biology*, 13:25, 2015.

[108] F Dupuis-Sandoval, M Poirier, and Scott M S. The emerging landscape of small nucleolar rnas in cell biology. *Wiley Interdiscip Rev RNA*, 6:381–397, 2015.

[109] C Gaspin, J Cavaillé, G Erauso, and J P Bachellerie. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. *J Mol Biol*, 297:895–906, 2000.

[110] A D Omer, T M Lowe, A G Russell, H Ebhardt, S R Eddy, and P P Dennis. Homologs of small nucleolar RNAs in Archaea. *Science*, 288: 517–522, 2000.

[111] C Parmesan and G Yohe. A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, 421:37–42, 2003.

[112] T L Root, J T Price, K R Hall, H Schneider, C Rosenzweig, and J A Pounds. Fingerprints of global warming on wild animals and plants. *Nature*, 421:57–60, 2003.

[113] P B Wignall. *Causes of mass extinctions. Extinctions in the history of life.* Cambridge University Press, 2004.

[114] R Govaerts. How many species of seed plants are there? - a response. *Taxon*, 52:583–584, 2003.

[115] R Govaerts. How many species of seed plants are there? *Taxon*, 50: 1085–1090, 2001.

[116] M D Guiry. How many species of algae are there? *J Phycol*, 48:1057–1063, 2012.

[117] C Courties, A Vaquer, M Troussellier, J Lautier, M J Chretiennot-Dinet, J Neveux, C Machado, and H Claustre. Smallest eukaryotic organism. *Nature*, 370:255, 1994.

[118] H S Yoon, J D Hackett, C Ciniglia, G Pinto, and D Bhattacharya. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol*, 21:809–818, 2004.

[119] C J Rothfels, A Larsson, L Y Kuo, P Korall, W L Chiou, and K M Pryer. Overcoming deep roots, fast rates, and short internodes to resolve the ancient rapid radiation of eupolypod ii ferns. *Syst Biol*, 61:490–509, 2012.

[120] P S Soltis, D E Soltis, V Savolainen, P R Crane, and T G Barraclough. Rate heterogeneity among lineages of tracheophytes: integration of

molecular and fossil data and evidence for molecular living fossils. *Proc Natl Acad Sci USA*, 99:4430–4435, 2002.

[121] K G Karol, R M McCourt, and C F Cimino, M Tand Delwiche. The closest living relatives of land plants. *Science*, 294:2351–2353, 2001.

[122] D L Nickrent, C L Parkinson, J D Palmer, and R J Duff. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol Biol Evol*, 17:1885–1895, 2000.

[123] K S Renzaglia, S Schuette, R J Duff, R Ligrone, A J Shaw, B D Mishler, and J G Duckett. Bryophyte phylogeny: advancing the molecular and morphological frontiers. *Bryologist*, 110:179–213, 2007.

[124] Y L Qiu, L Li, B Wang, Z Chen, V Knoop, M Groth-Malonek, O Dombrovska, J Lee, L Kent, and J Rest. The deepest divergences in land plants inferred from phylogenomic evidence. *Proc Natl Acad Sci USA*, 103:15511–15516, 2006.

[125] S Magallon and M J Sanderson. Relationships among seed plants inferred from highly conserved genes: Sorting conflicting phylogenetic signals among ancient lineages. *Amer J Bot*, 89:1991–2006, 2002.

[126] B R Ruhfel, M A Gitzendanner, P S Soltis, D E Soltis, and J G Burleigh. From algae to angiosperms-inferring the phylogeny of green plants (viridiplantae) from 360 plastid genomes. *BMC Evol Biol*, 14:23, 2014.

[127] C Lu, S S Tej, S Luo, CD Haudenschild, B C Meyers, and P J Green. Elucidation of the small RNA component of the transcriptome. *Science*, 309:1567–1569, 2005.

[128] M Hafner, P Landgraf, J Ludwig, A Rice, T Ojo, C Lin, D Holoch, C Lim, and T Tuschl. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, 44:3–12, 2008.

[129] C J Creighton, J G Reid, and P H Gunaratne. Expression profiling of microRNAs by deep sequencing. *Brief Bioinform*, 10:490–497, 2009.

[130] Y S Lee, Y Shibata, A Malhotra, and A Dutta. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.*, 23:2639–2649, 2009.

[131] A Sobala and G Hutvagner. Transfer RNA-derived fragments: origins, processing, and functions. *Wiley Interdiscip Rev RNA*, 2:853–862, 2011.

[132] C Ender, A Krek, M R Friedl??nder, M Beitzinger, L Weinmann, W Chen, S Pfeffer, N Rajewsky, and G Meister. A human snoRNA with microRNA-like functions. *Mol Cell*, 32:519–28, 2008.

[133] M Falaleeva and S Stamm. Processing of snoRNAs as a new source of regulatory non-coding RNAs: snoRNA fragments form a new class of functional RNAs. *Bioessays*, 35:46–54, 2013.

[134] David Langenberger, Clara Bermudez-Santana, Peter F. Stadler, and Steve Hoffmann. Identification and classification of small RNAs in transcriptome sequence data. *Pac. Symp. Biocomput.*, 15:80–87, 2010.

[135] Z Li, C Ender, G Meister, P S Moore, Y Chang, and B John. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Res.*, 40:6787–6799, 2012.

[136] P Kapranov, J Cheng, S. Dike, D Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermüller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, G. Madhavan, A Piccolboni, V Sementchenko, H. Tammana, and T. R. Gingeras. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488, 2007.

[137] M J Axtell. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, 19:740–751, 2013.

[138] M R Friedländer, W Chen, C Adamidi, J Maaskola, R Einspanier, S Knespel, and N Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol*, 26:407–415, 2008.

[139] HJ Wu, YK Ma, T Chen, M Wang, and XJ Wang. PsRobot: a web-based plant small RNA meta-analysis toolbox. *Nucleic Acids Res*, 2012.

[140] Xinbin Dai and Patrick X Zhao. psRNATarget: A plant small RNA target analysis server. *Nucleic Acids Res*, 39:W155–W159, 2011.

[141] F Li, R Orban, and B Baker. Somart: a webserver for plant mirna, tasirna and target gene analysis. *Plant J*, 70:891–901, 2012.

[142] Minja Zorc, Dasa Jevsinek Skok, Irena Godnic, George Adrian Calin, Simon Horvat, Zhihua Jiang, Peter Dovc, and Tanja Kunej. Catalog of microRNA seed polymorphisms in vertebrates. *PLoS One*, 7:e30737, 2012.

[143] Shahar Alon, Eyal Mor, Francois Vigneault, George M. Church, Franco Locatelli, Federica Galeano, Angela Gallo, Noam Shomron, and Eli Eisenberg. Systematic identification of edited microRNAs in the human brain. *Genome Res*, 22:1533–1540, 2012.

[144] Mario Fasold, David Langenberger, Hans Binder, Peter F. Stadler, and Steve Hoffmann. DARIO: A ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*, 39: W112–W117, 2011.

[145] C Coruh, S Shahid, and MJ Axtell. Seeing the forest for the trees: annotating small RNA producing genes in plants. *Curr Opin Plant Biol*, 18:87–95, 2014.

[146] JC Dohm, A E Minoche, D Holtgräwe, S Capella-Gutiérrez, F Zakrzewski, H Tafer, O Rupp, TR Sörensen, R Stracke, R Reinhardt, A Goesmann, T Kraft, B Schulz, PF Stadler, T Schmidt, T Gabaldón, H Lehrach, B Weisshaar, and H Himmelbauer. The genome of the recently domesticated crop plant sugar beet Beta vulgaris. *Nature*, 7484:546–549, 2014.

[147] Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485:635–641, 2012.

[148] D E Comer and D L Stevens. Vol iii: Client-server programming and applications. *Department of Computer Sciences, Purdue University, West Lafayette, IN 479*, 13:ISBN 0–13–474222–2, 1993.

[149] J C Dohm, C Lottaz, T Borodina, and H Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 36:e105, 2008.

[150] S E Linsen, E deWit, G Janssens, S Heater, L Chapman, R K Parkin, B Fritz, S K Wyman, E deBruijn, E E Voest, S Kuersten, M Tewari, and Edwin Cuppen. Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods*, 6:474–476, 2009.

[151] K D Hansen, S E Brenner, and S Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*, 38: e131, 2010.

[152] A Kozomara and S Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39:D152–D157, 2011.

[153] TM Lowe and SR Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res*, 25:955–964, 1997.

[154] C Zhang, G Li, S Zhu, S Zhang, and J Fang. tasiRNAdb: a database of ta-siRNA regulatory pathways. *Bioinformatics*, 30:1045–1046, 2014.

[155] JW Brown, GP Clark, DJ Leader, CG Simpson, and T Lowe. Multiple snoRNA gene clusters from Arabidopsis. *RNA.*, 12:1817–1832, 2001.

[156] F Barneche, C Gaspin, R Guyot, and M Echeverria. Identification of 66 box c/d snornas in arabidopsis thaliana: extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2'-o-methylation sites. *J Mol Biol*, 1:57–73, 2001.

[157] David Langenberger, Sebastian Bartschat, Jana Hertel, Steve Hoffmann, Hakim Tafer, and Peter F. Stadler. *MicroRNA or Not MicroRNA?*, volume 6832 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2011.

[158] W Shi, D Hendrix, M Levine, and B Haley. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat Struct Mol Biol*, 16:183–189, 2009.

[159] David Langenberger, Clara Bermudez-Santana, Jana Hertel, Steve Hoffmann, Steve Khaitovich, and Peter F. Stadler. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, 25: 2298–2301, 2009.

[160] A Stark, N Bushati, C H Jan, P Kheradpour, E Hodges, J Brennecke, D P Bartel, S M Cohen, and M Kellis. A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. *Genes Dev*, 22:8–13, 2008.

[161] J H Teune and G Steger. NOVOMIR: *De Novo* prediction of microRNA-coding regions in a single plant-genome. *J Nucleic Acids.*, 10: 495904, 2010.

[162] J Hertel, IL Hofacker, and PF Stadler. `SnoReport`: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24: 158–164, 2008.

[163] EP Nawrocki. Annotating functional RNAs in genomes using infernal. *Methods Mol Biol*, 1097:163–197, 2014.

[164] T Hubbard, D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, J Cuff, V Curwen, T Down, R Durbin, E Eyras, J Gilbert, M Hammond, L Huminiecki, A Kasprzyk, H Lehvaslaiho, P Lijnzaad, C Melsopp, E Mongin, R Pettett, M Pocock, S Potter, A Rust, E Schmidt, S Searle, G Slater, J Smith, W Spooner, A Stabenau, J Stalker, E Stupka, A Ureta-Vidal, I Vastrik, and M Clamp. The Ensembl genome database project. *Nucleic Acids Res*, 30:38–41, 2002.

[165] K Youens-Clark, E Buckler, T Casstevens, C Chen, G Declerck, P Derwent, P Dharmawardhana, P Jaiswal, P Kersey, A S Karthikeyan, J Lu, S R McCouch, L Ren, W Spooner, J C Stein, J Thomason, S Wei, and D Ware. Gramene database in 2010: updates and extensions. *Nucleic Acids Res*, 39:1085–1094, 2010.

[166] H Wickham. *`ggplot2`: Elegant Graphics for Data Analysis*. Springer, New York, 2009.

[167] T Pélissier, M Clavel, C Chaparro, M N Pouch-Pélissier, H Vaucheret, and J M Deragon. Double-stranded RNA binding proteins DRB2 and DRB4 have an antagonistic impact on polymerase IV-dependent siRNA levels in Arabidopsis. *RNA*, 17:1502–1510, 2011.

[168] A Weiberg, M Wang, F M Lin, H Zhao, Z Zhang, I Kaloshian, H D Huang, and H Jin. Fungal small rnas suppress plant immunity by hijacking host rna interference pathways. *Science*, 342:118–123, 2013.

[169] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comp. Biol.*, 5:e1000502, 2009.

[170] C Otto, PF Stadler, and S Hoffmann. Lacking alignments? the next-generation sequencing mapper segemehl revisited. *Bioinformatics*, 2014.

[171] Z Ma, C Coruh, and M J Axtell. Arabidopsis lyrata small RNAs: transient MIRNA and small interfering RNA loci within Arabidopsis genus. *Plant Cell*, 22:1090–1103, 2010.

[172] Chaogang Shao, Xiaoxia Ma, Ming Chen, and Yijun Meng. Characterization of expression patterns of small RNAs among various organs in Arabidopsis and rice based on 454 platform generated high throughput sequencing data. *Plant Omics*, 3:298–304, 2012.

[173] J de Meaux, J Y Hu, U Tartler, and U Goebel. Structurally different alleles of the ath-MIR824 microRNA precursor are maintained at high frequency in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 26:8994–8999, 2008.

[174] David Langenberger, M. Volkan Çakir, Steve Hoffmann, and Peter F. Stadler. Dicer-processed small RNAs: Rules and exceptions. *J. Exp. Zool: Mol. Dev. Evol.*, 320:35–46, 2012.

[175] J Goecks, A Nekrutenko, J Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11: R86, 2010.

[176] M P Hoeppner and A M Poole. Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC Evol Biol*, 12:183, 2012.

[177] Stephanie Kehr, Sebastian Bartschat, Hakim Tafer, Peter F. Stadler, and Jana Hertel. Matching of soulmates: Coevolution of snoRNAs and their targets. *Mol Biol Evol*, 31:455–467, 2014.

[178] P Shao, J H Yang, H Zhou, D G Guan, and L H Qu. Genome-wide analysis of chicken snornas provides unique implications for the evolution of vertebrate snoRNAs. *BMC Genomics*, 10:86, 2009.

[179] N Liu, Z D Xiao, C H Yu, P Shao, Y T Liang, D G Guan, J H Yang, C L Chen, L H Qu, and H Zhou. SnoRNAs from the filamentous fungus

Neurospora crassa: structural, functional and evolutionary insights. *BMC Genomics*, 10:515, 2009.

[180] Jana Hertel, Manuela Lindemeyer, Kristin Missal, Claudia Fried, Andrea Tanzer, Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and The Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. *BMC Genomics*, 7:15, 2006.

[181] Hadi Jorjani, Stephanie Kehr, Dominik J. Jedlinski, Rafal Gumienny, Jana Hertel, Peter F. Stadler, Mihaela Zavolan, and Andreas R. Gruber. An updated human snoRNAome. *Nucl Acids Res*, 44:5068–5082, 2016. doi: 10.1093/nar/gkw386.

[182] L Lestrade and M J Weber. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, 34: D158–162, 2006.

[183] M Yoshihama, A Nakao, and N Kenmochi. snOPY: a small nucleolar RNA orthological gene database. *BMC Research Notes*, 6:426, 2013.

[184] J W Brown, M Echeverria, L H Qu, T M Lowe, J P Bachellerie, A Hüttenhofer, J P Kastenmayer, P J Green, P Shaw, and Marshall D F. Plant snoRNA database. *Nucleic Acids Res*, 31:432–435, 2003.

[185] F Barneche, F Steinmetz, and M Echeverria. Fibrillarin genes encode both a conserved nucleolar protein and a novel small nucleolar RNA involved in ribosomal RNA methylation in Arabidopsis thaliana. *J Biol Chem*, 275: 27212–27220, 2000.

[186] L H Qu, Q Meng, H Zhou, and Y Q Chen. Identification of 10 novel snoRNA gene clusters from Arabidopsis thaliana. *Nucleic Acids Res*, 29: 1623–1630, 2001.

[187] C L Chen, C J Chen, O Vallon, Z P Huang, H Zhou, and L H Qu. Genomewide analysis of box C/D and box H/ACA snoRNAs in Chlamydomonas reinhardtii reveals an extensive organization into intronic gene clusters. *Genetics*, 179:21–30, 2008.

[188] G Qu, K Kruszka, P Plewka, Chiou T J Yang S Y, A Jarmolowski, Z Szweykowska-Kulinska, M Echeverria, and W M Karlowski. Promoter-based identification of novel non-coding rnas reveals the presence

of dicistronic snorna-mirna genes in Arabidopsis thaliana. *BMC Genomics*, 16:1009, 2015. doi: 10.1186/s12864-015-2221-x.

[189] C Quast, E Pruesse, P Yilmaz, J Gerken, T Schweer, P Yarza, J Peplies, and F O Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, 41: D590–D596, 2013.

[190] Deblina Patra, Mario Fasold, David Langenberger, Gerhard Steger, Ivo Grosse, and Peter F. Stadler. plantdario: web based quantitative and qualitative analysis of small rna-seq data in plants. *Front Plant Sci*, 5:708, 2014. doi: 10.3389/fpls.2014.00708.

[191] E Tran, X Zhang, L Lackey, and E S Maxwell. Conserved spacing between the box C/D and C'/D' RNPs of the archaeal box C/D sRNP complex is required for efficient 2'-O-methylation of target RNAs. *RNA*, 11:285–293, 2005.

[192] E P Nawrocki, D L Kolbe, and S R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25:1335–1337, 2009.

[193] Jana Hertel and Peter F. Stadler. The expansion of animal microrna families revisited. *Life (Basel)*, 5:905–920, 2015. doi: 10.3390/life5010905.

[194] T H Lee, H Tang, X Wang, and A H Paterson. PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Res*, 41:D1152–D1158, 2013.

Appendix A: Reference genomes of selected plant species

The list of the genomes discussed in chapter 3 as well as in chapter 4, used for of the different selected species are downloaded from different sources and they are cited with the sources and their accession numbers in the following page:

Species	Genome Assembly	Genome Source
Citrus sinensis	$\text{Csinensis}_1 54$	phytozome database
Arabidopsis thaliana	TAIR10	ensembl database
Brassica rapa	IVFCAASv1	ensembl database
Gossypium raimondii	JGI v221	JGI database
Prunus persica	$\text{Ppersica}_1 39$	phytozome database
Medicago trunculata	$\text{MedtrA17}_3.5_G CA_0 00219495.1$	ftp.jcvi.org database
Lotus japonicus	Lj2.5	ftp.kazusa.or.jp database
Ricinus communis	$\text{Rcommunis}_1 19$	phytozome database
Vitis vinifera	$\text{V.vinifera}_I GGP_1 2x$	ensembl database
Solanum lycopersicum	SL2.40	cornell database
Solanum tuberosum	$\text{SolTub}_3.0$	ensembl database
Digitalis purpurea	$\text{dpa}_a ssembly_{v1} 0072011$	plantbiology database
Beta vulgaris	RefBeet-1.2	bvseq.molgen.mpg.de database
Hordeum vulgare	ASM32608v1	ensembl database
Triticum aestivum	IWGSP1	ensembl database
Oryza sativa	IRGSP-1.0	ensembl database
Sorghum bicolor	Sorbi1	ensembl database
Zea mays	AGPv3	ensembl database
Phoenix dactylifera	PdactyKAssembly1.0	cornell database
Amborella trichopoda	AMTR1.0	ensembl database
Pinus taeda	ptaeda.v1.0	dendrome database
Selaginella moellendorffii	v1.0	ensembl database
Physcomitrella patens	ASM242v1	ensembl database
Chlamydomonas reinhardtii	v3.0	JGI database
Cyanidioschyzon merolae	ASM9120v1	ensembl database

Appendix B: Nomenclature of snoRNA families

As discussed in chapter 3, the nomenclature of plant snoRNAs only partially respects known or detectable sequence homology and so we used a unique internal family identifier throughout the study. We provide here the complete table of snoRNA unique internal family identifiers, types and their species-specific synonyms in the corresponding sources. The nomenclature data table along with comparisons from the different sources is attached in the following page:

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR1 C/D	AtsnoR1	—	SnoR1a/snoR1	—	—
snoR10 C/D	AtsnoR10	snoR10	—	—	—
snoR11 C/D	AtsnoR11	snoR11	—	—	—
snoR12 C/D	AtsnoR12	snoR12	Z131a/snoR12	—	—
snoR13 C/D	AtsnoR13	snoR13	Z199a/snoR13	—	—
snoR14 C/D	AtsnoR14	snoR14	snoR14	—	—
snoR15 C/D	AtsnoR15	snoR15	Z101/snoR15	—	—
snoR16 C/D	AtsnoR16	snoR16	snoR16	—	—
snoR17 C/D	AtsnoR17	snoR17		—	—
snoR18 C/D	AtsnoR18	snoR18	Z102/snoR18	—	—
snoR20 C/D	AtsnoR20	snoR20	Z160a/snoR20	—	—
snoR21 C/D	AtsnoR21	snoR21	Z221/snoR21	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR22 C/D	AtsnoR22	snoR22	Z151a/snoR22	—	—
snoR23 C/D	AtsnoR23	snoR23	Z152/snoR23	—	—
snoR24 C/D	AtsnoR24	snoR24	SnoR24a/snoR24	—	—
snoR25 C/D	AtsnoR25	snoR25	—	—	—
snoR29 C/D	AtsnoR29	snoR29	Z107/snoR29	—	—
snoR31 C/D	AtsnoR31	snoR31	—	—	—
snoR32 C/D	AtsnoR32	snoR32	—	—	—
snoR37 C/D	AtsnoR37	snoR37	Z157a/snoR37	—	—
snoR4 C/D	AtsnoR4	snoR4	—	—	—
snoR44 C/D	AtsnoR44	snoR44	snoR44	—	—
snoR59 C/D	AtsnoR59	snoR59	—	—	—
snoR64 C/D	AtsnoR64	snoR64	—	—	—
snoR7 C/D	AtsnoR7	snoR7	—	—	—
snoR8 C/D	AtsnoR8	snoR8	—	—	—
snoR9 C/D	AtsnoR9	snoR9	—	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
U27 C/D	AtU27	U27	—	—	—
U29 C/D	AtU29	U29	—	—	—
U30 C/D	AtU30	U30	—	—	—
U31 C/D	AtU31	U31	—	—	—
U33 C/D	AtU33	U33	Z195a/U33	—	—
U34 C/D	AtU34	U34	Z181a/U34	—	—
U35 C/D	AtU35	U35	Z228a/U35	—	—
U36II C/D	AtU36	U36	U36A/U36II	—	—
U37 C/D	AtU37	U37	—	—	—
U38 C/D	AtU38	U38	—	—	—
U43 C/D	AtU43	U43	—	—	—
U49 C/D	AtU49	U49	—	—	—
U51 C/D	AtU51	U51	Z196a/U51	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
U53 C/D	AtU53	U53	—	—	—
U55 C/D	AtU55	U55	—	—	—
U56 C/D	AtU56	U56	—	—	—
U61 C/D	AtU61	U61	U61/U61	—	—
U80 C/D	AtU80	U80	Z193a/U80	—	—
snoR38Y C/D	AtsnoR38Y	snoR38Y	—	—	—
snoR53Y C/D	AtsnoR53Y	snoR53Y	snoR53Y	—	—
AtsnoR68Y C/D	AtsnoR68Y	snoR68Y	—	—	—
snoR69Y C/D	AtsnoR69Y	snoR69Y	—	—	—
snoR39BY C/D	—	snoR39BY	Z125a/snoR39BY	—	—
snoR102 C/D	—	snoR102	—	—	—
snoR106 C/D	—	snoR106	—	—	—
snoR107 C/D	—	snoR107	—	—	—
snoR108 C/D	—	snoR108	—	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR19 C/D	—	snoR19	SnoR19a/snoR19	—	—
snoR26 C/D	—	snoR26	—	—	—
snoR27 C/D	—	snoR27	—	—	—
snoR28 C/D	—	snoR28	Z103a/snoR28	—	—
snoR33 C/D	—	snoR33	—	—	—
snoR34 C/D	—	snoR34	—	—	—
snoR35 C/D	—	snoR35	—	—	—
snoR36 C/D	—	snoR36	—	—	—
snoR41Y C/D	—	snoR41Y	Z154a/snoR41Y	—	—
snoR58Y C/D	—	snoR58Y	Z200a/snoR58Y	—	—
snoR6 C/D	—	snoR6	—	—	—
snoR65 C/D	—	snoR65	—	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR66 C/D	—	snoR66	Z269a/snoR66	—	—
snoR68 C/D	—	snoR68	—	—	—
snoR72Y C/D	—	snoR72Y	—	—	—
snoR77Y C/D	—	snoR77Y	—	—	—
U14 C/D	—	U14	Z114a/U14	—	—
U15 C/D	—	U15	Z104a/U15	—	—
U16 C/D	—	U16	—	—	—
U18 C/D	—	U18	Z106/U18	—	—
U24 C/D	—	U24	Z132a/U24	—	CrCD59/U24
U54 C/D	—	U54	—	—	—
snoR60 C/D	—	U60	snoR60	—	—
snoR30 C/D	—	snoR30	—	—	—
U40 C/D	—	U40	Z153a/U40	—	—
R72 C/D	—	R72	—	—	—
R71 C/D	—	R71	—	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
Z155a C/D	—	—	snoR36/Z155a	—	—
U36I C/D	—	—	Z100a/U36I	—	—
U59 C/D	—	—	Z159a/U59	—	—
Z158a C/D	—	—	snoR38Ya/Z158a	—	—
Z226 C/D	—	—	snoR68Y/Z226	—	—
Z122 C/D	—	—	snoR72Y/Z122	—	—
Z111 C/D	—	—	snoR77Ya/Z111	—	—
Z105 C/D	—	—	SnoR7a/Z105	—	—
Z110 C/D	—	—	SnoR10a/Z110	—	—
OssnoR17 C/D	—	—	snoR17	—	—
Z108 C/D	—	—	SnoR30/Z108	—	—
Z109 C/D	—	—	SnoR31/Z109	—	—
OssnoR32 C/D	—	—	snoR32	—	—
Z155a C/D	—	—	snoR36/Z155a	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR47 C/D	—	—	Z271/snoR47	—	—
Z134a C/D	—	—	SnoR64/Z134a	—	—
snoR120 C/D	—	—	Z192/sno120	—	—
snoR121 C/D	—	—	Z118a/snoR121	—	—
snoR122 C/D	—	—	Z119a/snoR122	—	—
snoR123 C/D	—	—	Z121/snoR123	—	—
snoR124 C/D	—	—	Z123/snoR124	—	—
snoR125 C/D	—	—	Z124/snoR125	—	—
snoR126a C/D	—	—	Z278a/snoR126a	—	—
snoR132a C/D	—	—	Z162a/snoR132a	—	—
snoR135a C/D	—	—	Z165a/snoR135a	—	—
snoR136a C/D	—	—	Z166/snoR136a	—	—
snoR137a C/D	—	—	Z168a/snoR137a	—	—
snoR138 C/D	—	—	Z170/snoR138	—	—
snoR139 C/D	—	—	Z171/snoR139	—	—
snoR140a C/D	—	—	Z172a/snoR140a	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
sno141a C/D	—	—	Z173a/snoR141a	—	—
snoR144 C/D	—	—	Z250/snoR144	—	—
snoR145a C/D	—	—	Z177a/snoR145a	—	—
snoR146a C/D	—	—	Z178a/snoR146a	—	—
snoR147a C/D	—	—	Z267a/snoR147	—	—
snoR148 C/D	—	—	Z252/snoR148	—	—
snoR149a C/D	—	—	Z182a/snoR149a	—	—
snoR150a C/D	—	—	Z183a/snoR150a	—	—
snoR153 C/D	—	—	Z187/snoR153	—	—
snoR154a C/D	—	—	Z188a/snoR154a	—	—
snoR156 C/D	—	—	Z190/snoR156	—	—
snoR157a C/D	—	—	Z194a/snoR157a	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR158a C/D	—	—	Z268a/snoR158a	—	—
snoR159a C/D	—	—	Z198a/snoR159a	—	—
snoR160a C/D	—	—	Z270a/snoR160a	—	—
snoR161a C/D	—	—	Z279a/snoR161a	—	—
snoR162a C/D	—	—	Z203/snoR162a	—	—
snoR165 C/D	—	—	Z225/snoR165	—	—
snoR167 C/D	—	—	Z229/snoR167	—	—
snoR169 C/D	—	—	Z240/snoR169	—	—
snoR170 C/D	—	—	Z241/snoR170	—	—
snoR172 C/D	—	—	Z243/snoR172	—	—
snoR175 C/D	—	—	Z274/snoR175	—	—
snoR176 C/D	—	—	Z275/snoR176	—	—
Z191a C/D	—	—	U27/Z191a	—	—
OsU29 C/D	—	—	U29a	—	—
Z223a C/D	—	—	U38/Z223a	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
OsU43 C/D	—	—	U43	—	—
Z112 C/D	—	—	U49/Z112	—	—
OsU54 C/D	—	—	U54a	—	—
AtncR4 C/D	—	—	—	ncR4	—
AtncR6 C/D	—	—	—	ncR6	—
AtncR7 C/D	—	—	—	ncR7	—
AtncR10 C/D	—	—	—	ncR10	—
AtncR12 C/D	—	—	—	ncR12	—
AtncR16 C/D	—	—	—	ncR16	—
AtncR17 C/D	—	—	—	ncR17	—
AtncR27 C/D	—	—	—	ncR27	—
AtncR28 C/D	—	—	—	ncR28	—
CrCD01 C/D	—	—	—	—	snoR1/CrCD01

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
CrCD02 C/D	—	—	—	—	CrCD02
CrCD03 C/D	—	—	—	—	snoR68/CrCD03
CrCD04 C/D	—	—	—	—	CrCD04
CrCD05 C/D	—	—	—	—	CrCD05
CrCD06 C/D	—	—	—	—	snoR69/CrCD06
CrCD07 C/D	—	—	—	—	snoR24/CrCD07
CrCD09 C/D	—	—	—	—	U51/CrCD09
CrCD10 C/D	—	—	—	—	snoR41Y/CrCD10
CrCD11 C/D	—	—	—	—	snoR120/snoR162/CrCD11
CrCD14 C/D	—	—	—	—	CrCD14
CrCD15 C/D	—	—	—	—	CrCD15
CrCD16 C/D	—	—	—	—	snoR13/CrCD16
CrCD17 C/D	—	—	—	—	U34/CrCD17
CrCD18 C/D	—	—	—	—	Z270/CrCD18

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
CrCD19 C/D	—	—	—	—	U54/CrCD19
CrCD20 C/D	—	—	—	—	U59/CrCD20
CrCD21 C/D	—	—	—	—	U49/CrCD21
CrCD22 C/D	—	—	—	—	snoR60/CrCD22
CrCD23 C/D	—	—	—	—	snoR130/CrCD23
CrCD24 C/D	—	—	—	—	CrCD24
CrCD25 C/D	—	—	—	—	J27/U36/CrCD25
CrCD26 C/D	—	—	—	—	U14/CrCD26
CrCD27 C/D	—	—	—	—	U35/CrCD27
CrCD28 C/D	—	—	—	—	U35/CrCD28
CrCD29 C/D	—	—	—	—	U29/CrCD29
CrCD30 C/D	—	—	—	—	U18/CrCD30
CrCD31 C/D	—	—	—	—	snoR37/CrCD31
CrCD32 C/D	—	—	—	—	snoR41YII/CrCD32

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
CrCD33 C/D	—	—	—	—	snoR66/CrCD33
CrCD34 C/D	—	—	—	—	snoR31/snoR9/CrCD34
CrCD35 C/D	—	—	—	—	snoR10/CrCD35
CrCD36 C/D	—	—	—	—	U59/CrCD36
CrCD37 C/D	—	—	—	—	U30/CrCD37
CrCD38 C/D	—	—	—	—	Z267/CrCD38
CrCD39 C/D	—	—	—	—	CrCD39
CrCD40 C/D	—	—	—	—	snoR44/CrCD40
CrCD41 C/D	—	—	—	—	CrCD41
CrCD43 C/D	—	—	—	—	snoR14/U61/CrCD43
CrCD44 C/D	—	—	—	—	snoR77Y/CrCD44
CrCD45 C/D	—	—	—	—	snoR19/crCD45
CrCD46 C/D	—	—	—	—	snoR15/CrCD46
CrCD47 C/D	—	—	—	—	CrCD47
CrCD48 C/D	—	—	—	—	U36/CrCD48
CrCD49 C/D	—	—	—	—	snoR68/CrCD49
CrCD50 C/D	—	—	—	—	U15/CrCD50
CrCD51 C/D	—	—	—	—	snoR72Y/CrCD51

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
CrCD52 C/D	—	—	—	—	CrCD52
CrCD53 C/D	—	—	—	—	U80/CrCD53
CrCD54 C/D	—	—	—	—	U80/CrCD54
CrCD55 C/D	—	—	—	—	U27/CrCD55
CrCD56 C/D	—	—	—	—	snoR53Y/CrCD56
CrCD57 C/D	—	—	—	—	U38/CrCD57
CrCD58 C/D	—	—	—	—	CrCD58
CrCD60 C/D	—	—	—	—	CrCD60
CrCD61 C/D	—	—	—	—	snoR19/CrCD61
CrCD62 C/D	—	—	—	—	snoR39BY/CrCD62
CrCD63 C/D	—	—	—	—	U43/CrCD63
CrCD64 C/D	—	—	—	—	U40/CrCD64
CrCD65 C/D	—	—	—	—	snoR133/CrCD65

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
CrCD66 C/D	—	—	—	—	CrCD66
CrCD67 C/D	—	—	—	—	snoR7/CrCD67
CrCD68 C/D	—	—	—	—	snoR38Y/snoR18/CrCD68
CrCD69 C/D	—	—	—	—	snoR29/CrCD69
CrCD70 C/D	—	—	—	—	snoR12/CrCD70
CrCD71 C/D	—	—	—	—	CrCD71
CrCD72 C/D	—	—	—	—	CrCD72
CrCD73 C/D	—	—	—	—	snoR32/CrCD73
CrCD74 C/D	—	—	—	—	CrCD74
CrCD75	—	—	—	—	CrCD75
C/D CrCD76 C/D	—	—	—	—	SnoR22/CrCD76
CrCD77 C/D	—	—	—	—	CrCD77
snoR2 H/ACA	—	SnoR2a/U65	Z113/snoR2	—	—
snoR5 H/ACA	—	snoR5a	SnoR5/snoR5a	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR100 H/ACA	—	snoR100	—	—	—
snoR103 H/ACA	—	snoR103	—	—	—
snoR104 H/ACA	—	snoR104	—	—	—
snoR109 H/ACA	—	snoR109	—	—	—
snoR110 H/ACA	—	snoR110	—	—	—
snoR111 H/ACA	—	snoR111	—	—	—
snoR112 H/ACA	—	snoR112	—	—	—
snoR72 H/ACA	—	snoR72	—	—	—
snoR73 H/ACA	—	snoR73	—	—	—
snoR74 H/ACA	—	snoR74	—	—	—
snoR77 H/ACA	—	snoR77	—	—	—
snoR78 H/ACA	—	snoR78	—	—	—
snoR79 H/ACA	—	snoR79	—	—	—
snoR80 H/ACA	—	snoR80	—	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
snoR81 H/ACA	—	snoR81	—	—	—
snoR82 H/ACA	—	snoR82	—	—	—
snoR83 H/ACA	—	snoR83	—	—	—
snoR84 H/ACA	—	snoR84	—	—	—
snoR86 H/ACA	—	snoR86	—	—	—
snoR88 H/ACA	—	snoR88	—	—	—
snoR89 H/ACA	—	snoR89	—	—	—
snoR90 H/ACA	—	snoR90	—	—	—
snoR92 H/ACA	—	snoR92	—	—	—
snoR93 H/ACA	—	snoR93	—	—	—
snoR94 H/ACA	—	snoR94	—	—	—
snoR95 H/ACA	—	snoR95	—	—	—
snoR96 H/ACA	—	snoR96	—	—	—
snoR97 H/ACA	—	snoR97	—	—	—
snoR99 H/ACA	—	snoR99	—	—	—
U19 H/ACA	—	U19	—	—	—

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
AtncR1 H/ACA	—	—	—	ncR1	—
CrACA01 H/ACA	—	—	—	—	snoR83/snoR82/CrACA01
CrACA02 H/ACA	—	—	—	—	CrACA02
CrACA03 H/ACA	—	—	—	—	CrACA03
CrACA04 H/ACA	—	—	—	—	CrACA04
CrACA05 H/ACA	—	—	—	—	snoR99/CrACA05
CrACA06 H/ACA	—	—	—	—	Osaca052/CrACA06
CrACA07 H/ACA	—	—	—	—	CrACA07
CrACA08 H/ACA	—	—	—	—	CrACA08
CrACA09 H/ACA	—	—	—	—	Osaca003/CrACA09
CrACA10 H/ACA	—	—	—	—	CrACA10
CrACA13 H/ACA	—	—	—	—	SnoR2/CrACA13
CrACA16 H/ACA	—	—	—	—	snoR100/CrCD16
CrACA18 H/ACA	—	—	—	—	CrACA18
CrACA19 H/ACA	—	—	—	—	CrACA19
CrACA21 H/ACA	—	—	—	—	Osaca019/CrACA21
CrACA22 H/ACA	—	—	—	—	CrACA22
CrACA23 H/ACA	—	—	—	—	CrACA23
CrACA24 H/ACA	—	—	—	—	Osaca003/CrACA24

snoRNAs Type	Plant-snoRNAdb	snOPYdb	O.sativa-paper	A.thaliana-paper	C.reinhardtii-paper
CrACA26 H/ACA	—	—	—	—	snoR91/CrACA26
CrACA28 H/ACA	—	—	—	—	CrACA28
CrACA29 H/ACA	—	—	—	—	Osaca019/Osaca003/CrACA29
CrACA30 H/ACA	—	—	—	—	CrACA30
CrACA31 H/ACA	—	—	—	—	snoR87/CrACA31
CrACA32 H/ACA	—	—	—	—	snoR96/CrACA32
CrACA33 H/ACA	—	—	—	—	snoR91/CrACA33
CrACA35 H/ACA	—	—	—	—	CrACA35
CrACA36 H/ACA	—	—	—	—	snoR78/CrACA36
CrACA37 H/ACA	—	—	—	—	snoR86/CrACA37
CrACA38 H/ACA	—	—	—	—	Osaca053/CrACA38
CrACA39 H/ACA	—	—	—	—	Osaca025/CrACA39
CrACA40 H/ACA	—	—	—	—	SnoR82/snoR77/CrACA40
CrACA41 H/ACA	—	—	—	—	CrACA40
CrACA42 H/ACA	—	—	—	—	CrACA42
CrACA43 H/ACA	—	—	—	—	Osaca069/CrACA43
CrACA44 H/ACA	—	—	—	—	CrACA44
CrACA45 H/ACA	—	—	—	—	CrACA45
CrACA46 H/ACA	—	—	—	—	snoR83/CrACA46
CrACA48 H/ACA	—	—	—	—	CrACA48
CrACA50 H/ACA	—	—	—	—	U19/CrACA50
CrACA51 H/ACA	—	—	—	—	CrACA51
CrACA54 H/ACA	—	—	—	—	CrACA54

Deblina Patra Bhattacharya; Mob. +49 (0)17661361750; E-mail: deblina@bioinf.uni-leipzig.de;

Curriculum Vitae

Personal Information

Surname	Patra Bhattacharya (née Patra)
First Name	Deblina
Birth date	21.08.1983
Birthplace	Krishnanagar, India
Nationality	Indian
Marital status	Married

Research Experience

Feb 2013 – April 2017	Scientific Researcher at Universität Leipzig & Martin-Luther-Universität Halle-Wittenberg Project title: "*Annotation and analysis of plant small non-coding RNAs, Studying evolutionary significance of plant snoRNAs*" Supervisors: Prof. Dr. Peter F. Stadler & Prof. Dr. Ivo Große
Aug 2011 – Oct 2012	Scientific Researcher at Freiburg Institute for Advanced Studies, Albert-Ludwigs-Universität Freiburg Project title: "*In-silico data analyses platforms for large-scale proteomics experiments and studying underlying principles of mass spectrometry*" Advisor: Prof. Dr. Jörn Dengjel
Jan 2012 - Jan 2013	Collaborative Project at Universitätsklinikum Freiburg Project title: "*Template based homology modelling and docking to identify phagosomal immunological receptor of ssRNA*" Advisor: Dr. rer. nat. Sachin D. Deshmukh
Jul 2010 – Jul 2011	Diamond Jubilee Research Intern at Council of Scientific Industrial Research (CSIR), New Delhi, India Project title: "*Creating and integrating workflows and webservices through Galaxy workflow and Taverna in Open Source Drug Discovery Program*" Advisor: Dr. Anshu Bhardwaj
Dec 2008 – Apr 2010	Project Assistant Level II at Indian Institute of Chemical Biology, Kolkata, India Project title: "*Studying GSTM1 null genotype and its compensation by other GSTM family members*" Advisors: Dr. A.K. Giri and Dr. Nanda Ghoshal

Education

2013 - 2017	Dr. rer. nat., Martin-Luther-Universität Halle-Wittenberg, Germany, Dissertation Thesis Title: "*Study of small non-coding RNAs in plants by developing novel pipelines*"
2006 - 2008	Masters in Bioinformatics (MSc.), Utkal University, Orissa (India) Ranked 1/200 students, MSc.Thesis Title: "*Bioinformatics approaches for studying GSTM polymorphisms*", Achieved 1st division
2003 - 2006	Bachelors in Botany (BSc.), University of Calcutta, Kolkata (India), Achieved 1st division

Deblina Patra Bhattacharya; Mob. +49 (0)17661361750; E-mail: deblina@bioinf.uni-leipzig.de;

■ Publications

Patra Bhattacharya, D.; Hertel, J.; Bartschat, S.; Kehr, S.; Grosse, I.; Stadler, P.F.
Phylogenetic distribution of plant snoRNAs. BMC Genomics, 17:969, 2016.

Patra, D.; Fasold, M.; Langenberger, D.; Grosse, I.; Stadler, P. F.
plantDARIO: web based quantitative and qualitative analysis of small RNA-seq data in plants. Front Plant Sci, 5:708, 2014.

Bhattacharjee, P.; Paul, S.; Banerjee, M.; Patra, D.; Banerjee, P.; Ghoshal, N.; Bandyopadhyay, A.; Giri, A.K.
Functional compensation of glutathione S-transferase M1 (GSTM1) null by another GST superfamily member, GSTM2. Sci Rep., 3:2704, 2013.

Bhardwaj, A.; Scaria, V.; Patra, D.
Open Source Drug Discovery : A Global Collaborative Drug Discovery Model for Tuberculosis. Science and Culture, vol. 77, nos. 12, 22-26, 2011.

■ Professional skills and Interests

- Large-scale analysis of high-throughput next generation sequence data and RNA-seq data analysis
- Annotation and integration of plant small non-coding RNA genes, tool development for variant analysis of small RNA-seq data
- Interests: Designing pipelines and tools for high-throughput sequence data for large-scale data analysis

■ Technical Skills

O/S: Windows, Linux (Fedora, Redhat), Ubuntu, Mac

Programing languages: C, C++, Perl, Python, Shell, R, PHP, HTML, SQL

Specialized Software: MATLAB

Software and Applications : MS Office, LibreOffice and LaTeX

■ Extra-curricular Activities

Activities / Hobbies	Dramatics and Dance (trained in Indian classical dance)
	Hiking and Music listening (Indian and Western)

■ Language Skills

English	TOEFL, 2009 score: 108/120
German	A2 (University of Leipzig)
Hindi & Bengali	Native speaker