Università degli Studi di Ferrara

Ph.D. Course in

Evolutionary Biology and Ecology

In cooperation with

Università degli studi di Parma, Università degli studi di Firenze

CYCLE XXXI

COORDINATOR Prof. Guido Barbujani

# Utility of RNA-Seq and bioinformatics for gene discovery and evolutionary analyses in the Arundinoideae (Poaceae)

Scientific/Disciplinary Sector (SDS) BIO/18

| **Candidate** | **Supervisor** | **Co-supervisor** |
|---|---|---|
| Dott. Jike Wuhe | Prof. Bertorelle Giorgio | Dr. Varotto Claudio |

Years 2015/2018

**Abbreviations**

| | |
|---|---|
| Aco | *Arundo collina* |
| Ado | *Arundo donax* |
| Adof | *Arundo donaciformis* |
| Afo | *Arundo formosana* |
| Ama | *Arundo macrophylla* |
| Ami | *Arundo micrantha* |
| Apl | *Arundo plinii* |
| Hma | *Hakonechloa macra* |
| Mca | *Molinia caerulea* |
| Pau | *Phragmites australis* |
| At | *Arabidopsis thaliana* |
| Os | *Oryza sativa* |
| ATR | *Amborella trichopoda* |
| AFLP | Amplified Fragment Length Polymorphism |
| HGT | horizontal gene transfer |
| SAGE | Serial Analysis Of Gene Expression |
| pri-miRNA | primary miRNA |
| pre-miRNA | precursor miRNA |
| miRNA | MicroRNA |
| MFE | minimal folding free energy |
| MFEI | minimal folding energy index |
| RISC | RNA-induced silencing complex |
| PTGS | Post-transcriptional gene silencing |
| GO | Gene Ontology |
| FDR | False discovery rate |
| BUSTED | Branch-site Unrestricted Statistical Test for Episodic Diversification |
| aBSRE | adaptive branch-site random effects likelihood |
| FUBAR | Fast Unconstrained Bayesian AppRoximation |
| MEME | Mixed Effects Model of Evolution |

**Abstract**

Perennial grass species have been used as renewable resource to produce biofuel, and especially the Arundinoideae, one subfamily of Poaceae perennial grasses has attracted great research attention. *Arundo donax* L. (also known as the giant reed) is a perennial $C_3$ grass with fast growth. It is considered as the next generation most promising bioenergy and phytoremediation crop in the Mediterranea area. Despite its importance and value, many fundamental aspects of its biology and the precise phylogenetic relationships with respect to other species of the *Arundo* genus still remain to be fully elucidated.

In this PhD thesis I first applied, based on the reference transcriptome of *Arundo donax*, a computational step-by-step workflow for identifying a total of 141 miRNAs belonging to 14 families and a total of 462 high-confidence predicted targets in *A. donax*. Among the different miRNA families identified, MIR444 family was commonly expressed in four tissues (bud, culm, root and leaf) in *A. donax*, indicating it may be related to the high resistance to viruses of these species.

Secondly, a total of 235 miRNAs belonging to 37 miRNA families and a total of 175 high-confidence putative targets were identified by using computational approaches in *de novo* assembly of several *Arundo* species leaf transcriptomes. Conserved miRNAs tend to regulate homologous targets at conserved target sites in different species.

The in depth analysis of the leaf transcriptomes for all taxa of the *Arundo* genus and closely related outgroups yielded a total of 1,016,877 unigenes with average length ranging from 741 to 1065 bp. Phylogenomic reconstruction based on 150 one-to-one orthologous groups (OGs) showed that *A. formosana* was sister to the other members of the *Arundo* genus. The probabilistic models suggested that the ancestral haploid chromosome number of *Arundo* was thirty-six and revealed that demi-duplication was responsible for the evolutionary increase in chromosome numbers throughout the *Arundo* genus radiation. In addition, evolutionary analyses identified some genes under positive selection, suggesting their potential for future gene functional validation and improvement of the biomass species *A. donax*.

Lastly, *in silico* identification and comparative analysis of lignin and cellulose biosynthesis gene families across the Arundinoideae (Poaceae) was carried out. A total of 741 CesA/Csl protein sequences and 1118 lignin biosynthesis proteins were identified from the *de novo* assemblies of Arundinoideae leaf transcriptomes. Phylogenetic analysis of CesA/Csl proteins showed that CesA/Csl genes classified into 8 clades. CSLA and CSLC subfamily is an independent lineage to other CesA/Csl genes family, indicating that they probably originated from a separate duplication event. Lignin biosynthetic gene were highly divergent between eudicots and monocots, indicating that these genes might have experienced expansion after species differentiation. Further, these gene families divided into different groups based on reference species, namely rice, *Arabidopsis* and *Amborella*, indicating that diverse functions might exist in these gene family. The cellulose and lignin biosynthesis genes identified in this study will be helpful for

establishing mutagenesis-based reverse genetics and functional genomics approaches in *A.donax*.

In summary, leveraging on the availability of the leaf transcriptomes for all taxa of the *Arundo* genus and closely related outgroups, and reference transcriptomes of *Arundo donax*, gene discovery and evolutionary analyses were carried out in this study. These results pave the road to further elucidate the biology and evolution of *Arundo donax* and other *Arundo* species. The dissection of the patterns of evolution in *Arundo* genus will support ongoing efforts to establish reverse genetics and functional genomics approaches in *A. donax*, thus contributing to provide promising candidate genes for the improvement of this biomass species.

**Keywords:** *Arundo donax*; Arundinoideae; leaf RNA-Seq; comparative transcriptomics; phylogenomics; Chromosome number evolution; microRNA and their targets; lignin biosynthetic genes; cellulose and hemicellulose biosynthetic genes; Positive selection.

**Riassunto**

Le Arundinoideae, una sottofamiglia di erbe perenni di Poaceae, ha attirato grande attenzione come risorsa rinnovabile per la produzione di biocarburanti. *Arundo donax* L. (canna gigante) è un erba perenne di tipo $C_3$ a crescita rapida. È considerata la specie più promettente per produzione di bioenergia e fitodepurazione nella zona mediterranea. Nonostante la sua importanza e valore, molti aspetti fondamentali della sua biologia e le precise relazioni filogenetiche rispetto ad altre specie del genere *Arundo* restano ancora da chiarire.

In questa tesi di dottorato ho dapprima applicato una pipeline computazionale per identificare un totale di 141 miRNA appartenenti a 14 famiglie e un totale di 462 geni target in *A. donax*. Tra le diverse famiglie di miRNA identificate, la famiglia MIR444 è comunemente espressa in quattro tessuti (gemma, fusto, radice e foglia) in *A. donax*, indicando che potrebbe essere correlata all'elevata resistenza ai virus di queste specie.

In secondo luogo, un totale di 235 miRNA appartenenti a 37 famiglie di miRNA e un totale di 175 target putativi sono stati identificati utilizzando approcci computazionali tramite l'assemblaggio *de novo* di diversi trascrittomi fogliari di specie del genere *Arundo*. I miRNA conservati tendono a regolare obiettivi omologhi presso siti bersaglio conservati in diverse specie.

L'analisi dei transcrittomi di foglia per tutti i taxa del genere *Arundo* e outgroup strettamente correlati ha prodotto un totale di 1.016.877 unigenes con una lunghezza media di 741-1065 bp. La ricostruzione filogenomica basata su 150 gruppi ortologhi uno-a-uno (OG) ha dimostrato che *A. formosana* è la specie sorella degli altri membri del genere *Arundo*. I modelli probabilistici suggeriscono che il numero aploide di cromosomi ancestrale di *Arundo* è di 36 e rivela che la semi-duplicazione è stata responsabile dell'aumento evolutivo dei numeri cromosomici in tutto il genere *Arundo*. Inoltre, le analisi evolutive hanno identificato alcuni geni soggetti a selezione positiva, suggerendo il loro potenziale per il miglioramento della specie da biomassa *A. donax*.

Infine, è stata effettuata l'identificazione *in silico* e l'analisi comparativa delle famiglie di geni di biosintesi della lignina e della cellulosa nelle Arundinoideae (Poaceae). Un totale di 741 sequenze di proteine CesA / Csl e 1118 proteine di biosintesi della lignina sono state identificate dagli assemblaggi *de novo* dei trascrittomi di foglie di Arundinoideae. L'analisi filogenetica delle proteine CesA / Csl ha dimostrato che i geni CesA / Csl sono classificati in 8 cladi. Le sottofamiglie CSLA e CSLC sono una linea evolutiva indipendente rispetto ad altre famiglie di geni CesA / Csl, indicando che probabilmente hanno avuto origine da un evento di duplicazione separato. I geni biosintetici della lignina sono altamente divergenti tra eudicotiledoni e monocotiledoni, indicando che questi geni potrebbero essere andati incontro ad espansione dopo la differenziazione delle specieInoltre, queste famiglie di geni si dividono in diversi gruppi basati sulle specie di riferimento, indicando che potrebbero esistere diverse funzioni in questa famiglia di geni. I geni identificati in questo studio saranno utili per stabilire approcci di genomica funzionale e di genetica inversa basati sulla mutagenesi in *A.donax*.

In sintesi, facendo leva sulla disponibilità dei trascrittomi fogliari per tutti i taxa del genere *Arundo*, in questo studio sono stati condotti sia la scoperta genica che l'analisi evolutiva in Arundinoidee. Questi risultati aprono la strada per chiarire ulteriormente la biologia e l'evoluzione di *Arundo donax* e di altre specie di *Arundo*. La dissezione dei modelli di evoluzione nel genere *Arundo* sosterrà gli sforzi in corso per metter a punto gli approcci di genetica inversa e genomica funzionale in *A. donax*, contribuendo così a fornire promettenti geni candidati per il miglioramento di questa specie da biomassa.

**Parole chiave:** *Arundo donax*; Arundinoideae; RNA-Seq di foglia; trascrittomica comparativa; filogenomica; evoluzione del numero di cromosomi; microRNA e geni bersaglio; geni per la biosintesi della lignina; geni per la biosintesi della cellulosa ed emicellulosa; selezione positiva.

**Table of Contents**

Introduction of the Thesis....1
Overview of the Arundinoideae subfamily....2
Transcriptome and RNA-Seq Method....4
Phylogenomics and Polyploid speciation....6
MicroRNAs and Target mRNAs....7
Lignocellulose biosynthetic gene families....8
Bioinformatics....9
Goals of the Thesis....10
CHAPTER 1 *In silico* identification and characterization of a diverse subset of conserved microRNAs in bioenergy crop *Arundo donax* L....11
1.1 Abstract....12
1.2 Introduction....12
1.3 Materials and methods....14
1.3.1 Prediction of potential *Arundo donax* miRNAs....14
1.3.2 Position-specific base composition of mature microRNAs....15
1.3.3 Structural and phylogenetic reconstruction of different microRNA families....15
1.3.4 Prediction and functional annotation of putative *Arundo donax* miRNA targets 15
1.3.5 Comparative genomic analyses of miRNA targets in *Arundo donax* and other plants....15
1.4 Results....16
1.4.1 Identification of putative miRNAs in *Arundo donax* and their characteristics....16
1.4.2 Analysis of position-specific nucleotide preference in *Arundo donax* mature miRNAs....19
1.4.3 Variability of stem-loop structures in *Arundo donax* pre-miRNAs....20
1.4.4 Target prediction of *Arundo donax* miRNAs....22
1.4.5 Functional annotation of predicted targets....23
1.4.6 Conservation of miRNA targets among *Arundo donax* and other plant species..24
1.5 Discussion....25
1.5.1 Identification and characterization of conserved miRNAs in *Arundo donax*....25
1.5.2 Functional annotation of putative targets....26
1.6 Conclusion....27
CHAPTER 2 Computational predictions and comparative analyses of conserved microRNAs from *Arundo* leaf transcriptomes....28
2.1 Abstract....29
2.2 Introduction....29
2.3 Materials and methods....30

2.3.1 Plant materials, transcriptome dataset and reference miRNA .....30
2.3.2 In silico prediction of potential miRNAs.....31
2.3.3 Phylogenetic analysis of the putative miRNAs.....31
2.3.4 Prediction of miRNA targets and Functional annotation.....31
2.3.5 Comparative analyses of miRNA targets in *Arundo* species .....32
2.4 Results.....32
2.4.1 Identification of putative miRNAs in different *Arundo* species and their characteristics.....32
2.4.2 Phylogenetic analysis .....35
2.4.3 Target prediction of miRNAs and Functional annotation.....36
2.4.4 Conservation of miRNA targets among *Arundo* genus .....37
2.5 Discussion .....38
2.5.1 Identification and comparative analyses of conserved miRNAs in *Arundo* genus.....38
2.5.2 Functional annotation of putative miRNA targets.....39
2.6 Conclusion.....40
CHAPTER 3 Phylogeny and adaptive trait evolution in the *Arundo* genus (Poaceae) .....41
3.1 Abstract .....42
3.2 Introduction .....42
3.3 Materials and Methods.....44
3.3.1 Transcriptome De novo Assembly .....44
3.3.2 Gene functional annotation .....44
3.3.3 Orthologous Groups Identification and supermatrix construction .....44
3.3.4 Phylogenomic reconstruction of *Arundo* species .....45
3.3.5 Inference of chromosome-number change.....45
3.3.6 Analysis of molecular evolution.....46
3.4 Results.....46
3.4.1 De novo assembly of Illumina reads .....46
3.4.2 Functional annotation .....47
3.4.3 Phylogenomic reconstruction of *Arundo* species .....50
3.4.4 Chromosome evolution.....51
3.4.5 Evolutionary analysis .....53
3.5 Discussion .....58
3.5.1 De novo assembly of transcriptome data and function annotation .....58
3.5.2 Phylogenomic reconstruction of *Arundo* species .....59
3.5.3 Chromosome evolution.....60
3.5.4 Evolutionary analysis .....61

3.6 Conclusion....62
CHAPTER 4 *In silico* identification and comparative analysis of lignin and cellulose biosynthesis gene families across the Arundinoideae (Poaceae)....63
4.1 Abstract....64
4.2 Introduction....64
4.3 Materials and Methods....65
4.3.1 Computational identification of cellulose and lignin biosynthesis gene families 65
4.3.2 Sequence alignment and phylogenetic analysis....66
4.4 Results....66
4.4.1 Identification of CesA/Csl gene superfamily and lignin biosynthetic gene families....66
4.4.2 Phylogenetic analysis of CesA/Csl proteins....69
4.4.3 Phylogenetic analysis of lignin biosynthesis proteins....70
4.5 Discussion....72
4.5.1 Identification and phylogenetic analysis of CesA/Csl gene families....73
4.5.2 Identification and phylogenetic analysis of lignin biosynthesis gene families....73
4.6 Conclusion....74
Conclusion of the Thesis....75
References....79
Supplementary Figures and Tables....103
Acknowledgment....142

# Introduction of the Thesis

## Overview of the Arundinoideae subfamily

As the reduction of fossil energy reserves and global warming accelerate their pace, the world urgently require the development of environmental-friendly renewable sources of energy. Biomass can mitigate the dependence from petrol and coal by providing a nearly carbon-neutral source of energy. Plant biomass absorbs solar radiation energy from sunlight and fixes it together with atmospheric carbon dioxide into organic compounds by photosynthesis, making it available for production of biofuel energy (Donald, 2004). Biomass derived from many different food crops, likes rice, wheat and corn, leads to the the production of so-called first-generation biofuels. However, due to increase of both population and energy demand, competition for soils between first-generation biofuels and food crops can potentially lead to shortages of food supplies. Thus, second generation biofuels, based on plant species like grasses with lower requirements for the high quality arable land required by food crops, were developed (Naik et al., 2010). Biomass plants contain large amounts of polysaccharides in their cell walls, so these polysaccharides can be used as major renewable resources for biofuel production by converting them first into sugar and then bioethanol by enzymatic and microbial action (Pauly and Keegstra, 2008).

Perennial rhizomatous crops as promising sources of biomass, which can reduce the competition for land between food/feed and bioenergy crops, thus much attention has been recently devoted to these plant species (McKendry, 2002). These perennial grasses have been already used as source of bioenergy with many advantages in both Europe and USA. In particular, some species from the Arundinoideae, a subfamily of perennial grass from Poaceae, show good promise as bioenergy crops in the Mediterranean area (Lewandowski, 2003). Previous studies indicated that the Arundinoideae subfamily contain roughly 40 species belonging to 16 genera, and 2 tribes, Molinieae and Arundineae (Soreng et al., 2015; Soreng et al., 2017). Molinieae include 24 species from 13 genera, while Arundineae include 16 species from only 3 genera：*Amphipogon*, *Arundo*, and *Monachather* (Soreng et al., 2015). The popular ornamental genera *Hakonechloa*, *Molinia* and *Phragmites* are classified into the Molinieae tribe. *Hakonechloa macra*, also named "Japanese forest grass", which is a tough and ornamental grass, grows in slightly wet conditions, can resistant to disease and pests, but shading density and division size are important factors limiting the growth and production of this speices (Harvey and Brand, 2002). *Molinia caerulea* is a species of flowering and stress-tolerant plants with fast growth and high utilization of nutrition (G.W. Heil and M. Bruggink, 1987). It adapts to many types of soil, such as the Upper Teesdale area (Pigott, 1956). Previous genetic study showed that this native species exists as a tetraploid with chromosome number 2n = 36 (Taylor et al., 2001)**.** *Phragmites australis*, also named common reed, is a large perennial grass with fast growth and growing in some extreme conditions, even in high heavy metal concentration and saline conditions (Lissner and Schierup, 1997; Bonanno and Giudice, 2010; Bragato et al., 2006). *Phragmites australis* is considered as one among the invasive species via rapid rhizome growth, spreading and reproduction through rhizome fragments and seeds by water flow or human intervention (Lambert et al., 2010), so it has been suggested for

production of bioenergy (Patuzzi et al., 2003).

Arundineae is another the tribe of the Arundinoideae. *Arundo* is one of the most renown genera in the Arundineae tribe because of its large and gorgeous perennial species with ornamental value. Recent revisions indicated that up to five taxa may be included in the *Arundo* genus (Hardion et al., 2012). Most of these plants are occurring in some lowlands and sites affected by human disturbance of the landscape. Specifically, *A. collina*, formerly called *A. hellenica*, is distributed across several Mediterranean countries, and thanks to its drought-resistance it is often used for protecting bare hillside from erosion. However, due to fierce competition, there are some unresolved problems in the natural or artificial regeneration of this species (Danin et al., 2002; Danin, 2004). *Arundo plinii*, with a height up to roughly 2 meters, is commonly distributed in Italy almost over 1000 kilometers from Bologna to Sicily and it is present also in Malta, Croatia and Greece. It grows from seed and it is dispersed by wind propagation. Amplified Fragment Length Polymorphism (AFLP) and chloroplast DNA data showed that *Arundo plinii* is characterized by a large decrease of genetic diversity from the southern part of Italy (Sicily and Calabria), where populations with different ploidy exist (2n = 12X, with 72 chromosomes; 2n = 18X, with 108 chromosomes), to central and northern Italy, where only one ploidy level exists (12X uniform haplotype). However, there is haplotype diversity in both the south and north of the Balkans (Hardion et al., 2014). Based on detailed morphometric studies supported by molecular data, it has been recently proposed that *A. collina* is a synonym of *A. plinii*, which should possibly be used as species name also for all populations formerly attributed to *A. collina* (Hardion et al., 2014).

*Arundo formosana*, is one of the local Taiwan grasses growing along the steep slopes of the island. It grows faster in rainfall environments and it is dispersed through wind via seeds. *Arundo formosana* is considered as a very important grass which can be used for protecting hillslopes from soil erosion in the local area (Lin et al., 2006). *Arundo micrantha*, formerly called *Arundo mediterranea*, is distributed in the Mediterranean region and north Africa along rivers and streams. This species of grasses is endangered by the competition of another invasive species from the Arundineae, *Arundo donax* (Hardion et al., 2012; Mascia et al., 2013). *Arundo donaciformis* is a polyploid species (2n=108) with asexual reproduction, it occurs mainly in southern France and northwest Italy and plays an important role in preventing soil erosion through its powerful rhizomes (Hardion et al., 2012; Hardion et al., 2015).

Previous studies proposed four species as the most promising bioenergy crops among rhizomatous grasses. Specifically, *Arundo donax* and *Phalaris arundinacea* are $C_3$ species, while *Miscanthus* and *switchgrass* are $C_4$ grasses (Lewandowski, 2003). As promising next-generation biomass crop, *Arundo donax*, also called “giant cane” or “giant reed”, has been studied in detail. It is a perennial rhizomatous $C_3$ grass and an infertile polyploid plant. The cause of its sterility is still under discussion, because of the difficulty in precisely counting its chromosome number, which is predicted most probably ranging from 108 to 110. However, it seems now sure that the high polyploidy

characterizing this species is at the base of its sterility (Hardion et al., 2015). Two hypotheses about the origin of *Arundo donax* have been put forth in a recent study: one hypothesis is that octadecaploid *Arundo donax* originated from *Arundo plinii* via auto-polyploidization. The second hypothesis is allo-polyploidization, according to which the fertile *Arundo plinii* intercrossed with *Phragmites australis* to produce *Arundo donax* (Bucci et al., 2013). Genetic studies indicated that *Arundo donax* geographically originated in Eastern Asia, from where is spreaded all around the world, possibly by human intervention for different purposes. *A. donax* requires little management input and lacks natural competition, which, in addition to a great adaptability, makes it at the same time a badly invasive grass and a valuable bioenergy crop (Pilu et al., 2012). Previous studies showed that *A. donax* has also high potential for phytoremediation of heavy metal contamination of soil and especially of water (Bruno et al., 2015). These strong features make it a very productive biomass species, as in the right environment *Arundo donax* fields become productive from the second year and produce high dry biomass yield up to roughly 40 tons/hectare/year. Thus, it is considered as one of the most promising species for the biomass production in the Mediterranea (Angelini et al., 2009).

**Transcriptome and RNA-Seq Method**

Recently, the transcriptomics technologies are developing at a vary high pace, paralleling the rise of next-generation sequencing, and RNA-seq is a popular method for the massive study of gene expression (McGettigan, 2013). Transcriptomes include all transcripts of an RNA sample, irrespective on whether it is generated from a specific cell or populations of cells from different tissues. Comparative trancriptome analysis is important to understand the functional details of genomes and development processes, as transcription is the essential step for gene expression where genetic information from DNA is copies into RNA by the RNA polymerase. Transcriptome sequencing is an efficient and cost-effective way to produce large amounts of RNA transcript sequences used for reconstructing phylogeny and for gene discovery in eukaryotes, such as non-coding RNAs and mRNAs, particularly in some non-model organisms (Lemmon and R. Lemmon, 2013). In recent studies, more and more researchers used transcriptome data for resolving the evolutionary relationships among different plants lineages, such as reconstructing the phylogeny and define the origin of the land plants and its sister lineages (Timme et al., 2012; Wickett et al., 2014), or resolving relationships among species of the grape plant family (Wen et al., 2013). Besides evolutionary relationships reconstruction and gene discovery, comparative transcriptomics has also been used in many other systematic biology studies (Figure 1). For example, comparative transcriptomics provide a new insight into our understanding of horizontal gene transfer (HGT), which is an important way of microorganism evolution (Zhang et al., 2014).

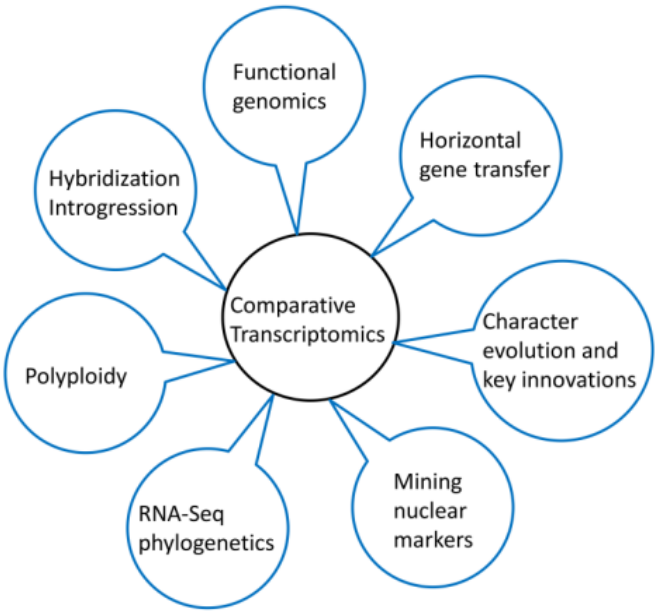



**Figure 1.** The application of comparative transcriptomics in plant systematics (Wen et al., 2015).

Many high-throughput technologies have been developed for generating transcriptome data, including sequence-based or hybridization-based approaches, such as SAGE (Serial Analysis Of Gene Expression) (Velculescu, 2000) and DNA microarrays (Schena et al., 1995). However, several limitations exist in both of these methods for high-throughput sequencing. Previous studies showed that the hybridization-based method has poor sensitivity and it is difficult to compare expression levels from different experiments, requiring complicated normalization methods. On the other hand, sequence-based methods need a reference genome, only some parts of the transcripts are analyzed and it is difficult to distinguish isoforms from each other, limiting the annotation for the transcriptomes (Wang et al., 2009). RNA-Seq is a novel high-throughput RNA sequencing that overcomes such limits and has great advantages over other sequencing technologies (Nagalakshmi et al., 2010; Hrdlickova et al., 2017). RNA sequencing (RNA-Seq or whole-transcriptome shotgun sequencing) is in fact able to reveal at the same time both the sequence and the expression levels of cellular RNAs. RNA-Seq is the principal high-throughput method used for transcriptome sequencing being highly accurate in quantifying expression level and highly reproducible. Additonally reads can be mapped either to a reference genome or transcripts, but most significantly they can also be assembled *de novo* without any reference genomic sequence in non-model organisms, which is very important in analyzing systematics aspects of biology (Wang et al., 2009). With the decreasing of sequencing costs and its advantages, de novo RNA-seq have been applied in some special fields for important fruits or plants, such as for intance investigating metabolic pathways in blackberry, pea and *Panax japonicus* (Schlüter et al., 2016; Garcia-Seco et al., 2015; Rai et al., 2016; Alves-Carvalho et al., 2015).

**Phylogenomics and Polyploid speciation**

Next-generation sequencing technologies (NGS) are generating huge amounts of whole-genome sequences or transcriptome data from different conditions and thousands of closely related species, samples or individuals. Phylogenomics is the study of the evolutionary relationships by comparative genome or transcriptome sequence analysis. This is important for understanding the diversity of biological characteristics among species, for example, the origin and evolution of species, gene functional annotation, reconstructing the tree of life, or evolutionary analysis (Chan and Ragan, 2013). Some important species and model organism genomes have been sequenced, and these sequences can be obtained from public databases (such as *Arabidopsis thaliana*, the fly *Drosophila melanogaster*) (Cao et al., 2011; Adams et al., 2000), but it is common to use transcriptome data in phylogenomic analysis when genome data are unavailable (Oliveira et al., 2012). Orthologous genes are inherited from a common recent ancestor in different species and these genes tend to retain the same function. Orthologs identified from transcriptome datasets have been utilized mainly to reconstruct the phylogenetic relationships among related taxa (Kocot et al., 2011). RNA-Seq phylogenetics and orthologous genes are also used for inferring the evolutionary rate of genes from large transcriptome sequences to explore natural selection footprints (Yang, 2005). However, there are some limitations, as confounding effect such as gene deletion, gene fusion, and gene recombination, can constitute a significant challenge for both phylogenetic analysis and natural selection detection. Thus, development of novel and more accurate algorithms is of paramount imporance to improve computational efficiency and overcome these challenges for phylogenomic reconstruction (Chan and Ragan, 2013).

Polyploidy is the conditions by which all the cells of a given individual are containing two or more sets of homologous chromosomes. Polyploidy is occasionally occurring in animals, while it is a relatively common condition in plants, which is recognized as a major driver of speciation and genetic diversity in the plant kingdom (Soltis et al., 1992). Polyploids can be distinguished in allopolyploids when the multiple sets of homologous chromosomes come from hybridization of evolutionary divergent taxa or autopolyploids when they come from a single parental species. In plants, allopolyploids are more common than autopolyploids (Soltis et al. 1992). Especially autopolyploids with high ploidy levels are often sterile, because homologous chromosomes are sufficiently similar to form multivalent during meiosis, which generate stochastic segregation of the chromatids and production of unbalanced, sterile gametes during meiosis. Previous studies carried on the potential influences of polyploidy on the reproduction of *Arundo* in the Mediterranean showed that infertility of the *Arundo* genus (Poaceae) is usually due to its high polyploidy (octadecaploidy; Hardion et al., 2015). Polyploid species shown generally great ability to adapt to different environments, even in harsh conditions (Ramsey and Schemske, 2002). It is widely demonstrated that autopolyploidy and allopolyploidy happened both in domesticated and wild plant species, and it is considered that the formation rate of autopolyploid is much faster than the formation rate of allopolyploid (Ramsey and Schemske, 1998) by integrating molecular genetic and phylogenetic methods (Wendel, 2000). At the same time, transcriptomics is a

significant way for studying the evolutionary history of allopolyploid genomes, and such next-generation sequencing datasets have been used, for instance, for studying evolutionary processes in the natural allopolyploid *Tragopogon* (Buggs et al., 2012).

**MicroRNAs and Target mRNAs**

MicroRNAs (miRNAs) are a class of small non-coding RNA molecules (usually about 22 nucleotides long) found in both plants and animals. They function in post-transcriptional gene regulation by targeting mRNAs with high sequence complementarity to them and preventing/reducing their expression (Bartel, 2004). In plants, miRNA genes are transcribed by RNA polymerase II into primary miRNAs (pri-miRNA) containing usually one or sometimes more hairpin-loop secondary structures (Lee et al., 2004). In the nucleus, pri-miRNAs are processed by the Dicer-like 1 (DCL1) enzyme into shorter stem-loop hairpins, called precursor miRNAs (pre-miRNAs). After transportation into the cytoplasm by exportin 5, the pre-miRNAs are further cleaved by DCL1 to produce a duplex formed by the mature miRNA and the star miRNA (miRNA*), the nearly perfect reverse complementary RNA derived from the pre-miRNA stem (Lee, 2002; Kurihara and Watanabe, 2004). In plants, miRNAs are involved in various biological processes, such as development, response to environmental stress and gene expression regulation, being most of their targets transcriptional factors (Zhang et al., 2006; Mallory and Vaucheret, 2006), miRNAs identified in the model species *Arabidopsis thaliana* suggested that thay play a fundamental role during plant development (Park et al., 2002). A total of 5071 miRNA loci have been identified in 58 different species and these datasets are available in the public database miRBase, which is a good resource available for identification of the miRNAs in other new plants by comparative genomics methods (Griffiths-Jones et al., 2008).

miRNAs are conserved both in plants and animals, and their function in gene regulation (Axtell and Bartel, 2005; Tanzer et al., 2004) can take place through two different mechanisms, mRNA nucleolytic degradation (cleavage) and translational repression (C. Vella and J. Slack, 2005). Despite their overall functional similarity, some major differences exist between miRNAs from plants and animals, which can provide a better understanding of miRNA biological functions in plants. Plant miRNAs are near-perfect complementary to their target genes, and this important character makes computational identification of plant miRNA targets relatively straightfoward. By contrast, animal miRNAs are less complementary to their targets, so it is difficult to identify the target genes by computational approaches alone, so development a more accurate algorithms for computational identification of miRNA targets with high specificity is still an ongoing process for animals. In plants, miRNAs are generated from stem-loop regions in primary transcripts through a Dicer-like protein, and thay usually induce target genes suppression by cleavage of the complementary target mRNAs (Jones-Rhoades et al., 2006). miRNA families are many but usually small in the animal genomes, in contrast to plant miRNA gene families, which are fewer and larger. Plant miRNA gene family members are highly similar, suggesting that there have been recent

expansion events through segmental duplication and gene duplication (Li and Mao, 2007) and that the purifying selecetion acting on their sequences is possibly higher than in animals. To understand the biological function of any miRNA, a fundamental step is the identification its mRNA target(s). Computational methods have been used successful in plants, based on the nearly perfect complementary of plant miRNAs to their targets. In general, target identification suggested that miRNAs in plant play roles as small interfering RNAs and are involved in the cleavage of mRNA (Rhoades et al., 2002).

In recent years, more and more genome and transcriptome sequences have been produced from many plants with multiple sequencing approaches. It therefore takes considerable efforts to identify miRNAs and their target genes from the increasing plant genomes and transcriptome databases. At the same time, this large amount of sequencing data allows not only to elucidate the miRNA functional mechanisms but also the evolutionary history of the corresponding miRNA genes. Identification of the miRNAs in plants can provide a chance to better understanding conservation of miRNAs at different evolutionary distances. There are in fact many miRNAs which are species-specific, indicating the fast origination and divergence of miRNA genes (Egan et al., 2012; Cui et al., 2017).

**Lignocellulose biosynthetic gene families**

Lignocellulose is a complex constituted by three main components, namely cellulose, hemicellulose and lignin. Cellulose and hemicellulose are a class of heteroglycans used for bioconversion into biofuels, while lignin is a class of aromatic polymer for structural support and resistance to pathogens in plant cell walls, but it needs to be removed from lignocellulose biomass for improving production of biofuel (Mussatto and Teixeira, 2010). There are some important genes involved in lignocellulosic biosynthesis pathways namely CesA (Cellulose-related cellulose synthases), Csl (hemicellulose-related cellulose synthase-like), CAD (cinnamyl alcohol dehydrogenase), CCoAOMT (caffeoyl-CoA O-methyltransferase), 4CL (4-coumarate: CoA ligase), CCR (cinnamoyl-CoA reductase), PAL (phenylalanine ammonia-lyase), C4H (cinnamate 4-hydroxylase), HCT (hydroxycinnamoyl-CoA shikimate/Quinate hydroxycinnamoyl transferase), COMT (caffeic acid O-methyl transferase), C3H (p-coumarate 3-hydroxylase) and F5H (ferulate 5-hydroxylase). These cellulose and lignin biosynthetic gene families are together responsible for the construction of the most important components of plant cell wall, and thus are also intimately involved in plant development and growth (Hamann et al., 2004; Suzuki et al., 2006; Liu et al., 2018). Specifically, the CesA gene family encodes cellulose synthase, which functions in primary and secondary plant cell wall formation, whereas the Csl gene family encode cellulose synthases-like enzymes. Usually Csl genes are expressed in some specific cell types, such as the CSLC genes, which encode β-1,4 glucan synthase, are involved in xyloglucan biosynthesis and function in structural support (Holland et al., 2000; Cocuron et al., 2007). Among the lignin biosynthetic genes, the 4CL protein functions in reducing lignin content, and it is involved in the development of rice and regulation of rice blast resistance (Liu et al., 2017). CCoAOMT is involved in caffeoyl CoA methylation and

hydroxycinnamates 5-methoxylation, which can also be used for the reduction of lignin content (Zhong et al., 2000). More and more cellulose and lignin biosynthetic gene families have been identified and their functional annotations were reported in plants, providing useful information for improvement of bioenergy crops.

**Bioinformatics**

With the development of the next generation sequencing technologies, progressively more massive biological data became available, such as the giant panda and turkey genomes (Li et al., 2010; Kerstens et al., 2009). In parallel, bioinformatics has played an important role in sequence assembly, the management of huge data, gene discovery, analysis and interpretation of these biological data (Oliver et al., 2015). These applications can be roughly divided into three steps, namely (1) next generation sequencing data generation and processing of the raw reads, (2) alignment and *de novo* assembly of raw reads, and (3) interpretation of biological data (Moorthie et al., 2013).

The high-quality raw reads generated from RNA-Seq technologies cannot be used as such, as they are usually too short to be of practical utility directly. Thus, these reads need to be either mapped to the reference genomes or de novo assembled into long contigs. There are several algorithms used for aligning short reads to the genome, such as SOAP (FM-index algorithm), Bowtie (FM-index algorithm), BWA (FM-index algorithm) and Novoalign (hash table algorithm) (Yu et al., 2012). Meanwhile, several programs have been developed for multiple sequences alignment, likes MUSCLE, MAFFT and T-Coffee (Pervez et al., 2014). Multiple sequences alignment is a very important step for interpreting biological data, such as functional annotation and phylogenetic analysis, but there is an unavoidable limitation constituted by the differences of the alignments generated by the different alignment approaches (Essoussi et al., 2008). Based on different data characteristics and purposes, choosing the proper alignment algorithms and programs for sequences alignment is an essential step for downstream analyses. For example, MUSCLE is used for large data alignments, as it has no limitation for the number of aligned sequences, it is fast and accurate. However, the input sequences format has special requirements in this program (Sedaghatinia et al., 2009). For analyzing biological data from raw reads generated by NGS approach, many algorithms have been developed for de novo assembly of short genome reads assembly, like Velvet and SOAPdenovo (Zerbino and Birney, 2008; Li et al., 2010). Some programs like Trinity and SOAPdenovo-Trans are utilized for de novo transcriptome assembly (Grabherr et al., 2011; Xie et al., 2014). Bioinformatics has become a fundamental and practical approach for interpreting biological data and mining biological information in the high-throughput sequencing era (Kanehisa and Bork, 2003). However, even if these programs and algorithms have many applications in sequences alignment, *de novo* assembly and interpretating biological data, there is still the onging need to develop new algorithms and programs for improving accuracy and reducing computational cost for analyzing and mining the massive biological data generated by NGS.

**Goals of the Thesis**

Leveraging on the availability of the leaf transcriptomes for all taxa of the *Arundo* genus and closely related outgroups (*Hakonechloa macra*, *Molinia caerulea* and *Phragmites australis*), and availability of the reference transcriptomes of *Arundo donax* (from leaf, root, bud and culm), the proposed project aims at:

(1) The *in silico* identification and characterization of a diverse subset of conserved microRNAs in bioenergy crop *Arundo donax* L based on the reference transcriptomes.

(2) The computational prediction and comparative analysis of conserved microRNAs from *de novo* assemly of leaf transcriptomes of taxa from the *Arundo* genus.

(3) The transcriptome-based phylogenomic reconstruction of the relationships among *Arundo* species and the identification of the origin of the biomass species *Arundo donax*. In particular, the contrasting hypotheses of auto-polyploidization *vs.* allo-polyploidization have been assessed in light of the process of chromosomal evolution in the *Arundo* genus.

(4) The dissection of the selective constraints controlling gene evolution in Arundinoideae, through identification of candidate genes for positive selection. In particular genes involved in traits that are known to affect *Arundo donax* growth and productivity (lignocellulosic biomass content and saccharification efficiency) have been identified.

The goal and expected impact of the thesis is to pave the road to further elucidate the fundamental aspects of the biology and evolution of *Arundo donax* and other *Arundo* species. The dissection of the patterns of evolution in the *Arundo* genus will support ongoing efforts to establish reverse genetics and functional genomics approaches in *Arundo donax*, thus contributing to provide promising candidate genes for the improvement of this biomass species.

# CHAPTER 1

*In silico* identification and characterization of a diverse subset of conserved microRNAs in bioenergy crop *Arundo donax* L.

## 1.1 Abstract

MicroRNAs (miRNAs) are small non-coding RNA molecules involved in the post-transcriptional regulation of gene expression. *Arundo donax* L. is a perennial $C_3$ grass considered one of the most promising bioenergy crops. Despite its relevance, many fundamental aspects of its biology still remain to be elucidated. In the present study, the first in silico mining and tissue-specific characterization of microRNAs and their putative targets in *Arundo donax* was performed. This study identified a total of 141 miRNAs belonging to 14 families along with the corresponding primary miRNAs, precursor miRNAs and a total of 462 high-confidence predicted targets. Gene Ontology functional annotation showed that miRNA targets are constituted mainly by transcription factors involved in important biological processes. Folding variability of pre-miRNAs loops and phylogenetic analysis indicate variable selective pressure acting on the different miRNA families. The set of miRNAs identified in this study will pave the road to further miRNA research in *Arundo donax* and contribute towards a better understanding of miRNA-mediated gene regulatory processes in other bioenergy crops.

## 1.2 Introduction

MicroRNAs (miRNAs) are endogenous small non-coding RNA molecules, containing approximately 22 nucleotides (nt), playing important roles in the regulation of gene expression at the post-transcriptional level (Bartel, 2004). In plants, miRNA genes are transcribed by RNA polymerase II into primary miRNAs (pri-miRNA) (Lee et al., 2004). In the nucleus, pri-miRNAs are processed by the Dicer-like 1 (DCL1) enzyme into shorter stem-loop hairpins, called precursor miRNAs (pre-miRNAs). After transportation into the cytoplasm by exportin 5, the pre-miRNAs are further cleaved by DCL1 to produce a duplex formed by the mature miRNA and its star miRNA (miRNA*), the nearly perfect reverse complementary RNA derived from the pre-miRNA stem (Kurihara and Watanabe, 2004). Subsequently, the single strand of mature duplex corresponding to the mature miRNA is assembled with an Argonaute (AGO) RNA binding protein to form the RNA-induced silencing complex (RISC), which faciliate the interaction of mature miRNAs with their target mRNAs (Davis and Hata, 2009; Baumberger and Baulcombe, 2005). RISC-associated plant miRNAs recognize their target mRNA sequences by their nearly perfect or perfect complementarity, allowing them to identify with extremely high specificity only a small fraction of all transcribed mRNAs. This very high specificity is the key to enable microRNAs to regulate the expression of their targets. For the majority of plant miRNAs, target gene expression regulation is achieved by transcript cleavage, usually occurring between the $10^{th}$ and $11^{th}$ nucleotide at the 5' end of the miRNA (German et al., 2008). However, translational inhibition can be an additional/alternative mechanism used by some microRNAs to downregulate target expression (Kidner and Martienssen, 2005; Zhang et al., 2006). In addition, the ability of a single miRNA to be potentially involved in the regulation of multiple target genes or of multiple miRNAs (Dehury et al., 2013) makes microRNAs very flexible regulators in a wide variety of metabolic and biological process during all major growth and developmental processes of plants (Singh et al., 2016).

The majority of miRNAs are highly conserved in plants, and modern high-throughput sequencing technologies hold great promise to produce large sets of genomic or transcriptomic data in different tissues and at different developmental processes, which provides useful sequence resources to predict and analyze miRNAs in non-model plant species with non-sequenced genomes. In plants, computational approaches have been successfully applied and demonstrated effectively to attain a comprehensive prediction of potential miRNAs, such as in *Cassava*, strawberry, *Arabidopsis thaliana* and many others (Lindow and Krogh, 2005; Patanun et al., 2013; Dong et al., 2012; Wang et al., 2004; Archak and Nagaraju, 2007). Besides, the existence of curated online databases like miRBase (Griffiths-Jones et al., 2008), collecting and organizing in a reference repository all miRNAs predicted by computational approaches, enormuously simplifies microRNA identification in novel transcriptomes. Both homology searches based on miRNA conservation among different plant species and the secondary hairpin loop structures of the pre-miRNA sequences along with the high negative minimal folding energy (MFE) are reliable criteria for the computational identification of miRNAs. This is why, even without experimental validation, miRNAs can reliably be distinguished from other types of small RNAs, thus reducing the number of false positive among predicted miRNAs (Yin et al., 2008; Bonnet et al., 2004). Especially in plant species, the feature of nearly perfect or perfect complementarity of miRNAs to their target mRNA sequences also allows the reliable computational prediction of miRNA target genes. This, in turn, is very important for *in silico* prediction of miRNA functions, which allowed the genome-wide identification of microRNA genes (Rhoades et al., 2002; Schwab et al., 2005; Devi et al., 2016).

*Arundo donax L.*, also called "giant reed", is a perennial $C_3$ fast growing grass (Rossa et al., 1998). Genetic studies indicate that *Arundo donax* originated in Eastern Asia, from where it spread, possibly by human intervention, to the Middle East and the Mediterranean. More recently it was introduced in Africa and even Australia. In the large majority of its distribution area *Arundo donax* is reported to be a sterile species, and it has been suggested that it may be a hybrid with uneven ploidy or possibly a (pseudo-) triploid species (Pilu et al., 2012; Bucci et al., 2013). Despite its sterility, the vigorous growth and lack of natural antagonists allowed *Arundo donax* to become one of the most invasive riparian species in Southern USA (especially California). This robustness makes it a very productive biomass species, as in optimal conditions *Arundo donax* fields become productive already after the second year and can provide dry biomass yields up to 40 tons per hectare for the next ten years. These yields are higher than other perennial rhizomatous grasses, thus constituting one of the most promising species for the production of biomass in the Mediterranea area (Angelini et al., 2009). In addition, *Arundo donax* requires little management input, as it is resistant to most pests and pathogens. It can grow without significant P or N fertilization, and is highly tolerant to heavy metals and saline soil (Calheiros et al., 2012; Raspolli Galletti et al., 2013; Papazoglou et al., 2005). *Arundo donax* has also been utilized as a raw material for bioethanol production with dilute oxalic acid pre-treatment, which is important to overcome recalcitrance of lignocellulose for ethanol production (Scordia et al., 2011).

Recently, several transcriptomic studies have started to elucidate the content of expressed genes in *Arundo donax* (e.g. Sablok et al., 2014; Barrero et al., 2015; Fu et al., 2016), providing the opportunity for the in-depth mining of its gene space.

In the present study, the first computational identification and characterization of miRNAs for the biomass and bioenergy crop *Arundo donax* using tissue-specific transcriptomic data was performed. This could provide novel insights into the mechanism of *Arundo donax* development, metabolism and biology. The analyses also predicted the putative target transcripts of miRNAs, providing through network analysis an in-depth dissection of gene ontologies and functional annotations for both putative miRNAs and target genes. These findings advance the understanding of miRNAs in *Arundo donax*, and have the potential to be further utilized for controlling secondary metabolism for improving the production of biomass and fermentation efficiency.

## 1.3 Materials and methods

### 1.3.1 Prediction of potential *Arundo donax* miRNAs

All previously known 1616 miRNA precursor sequences from 12 monocotyledon species were downloaded from the miRBase database (Release 21.0; http://www.mirbase.org/) (Kozomara et al., 2014). These precursor miRNAs were used as query sequences for BLASTN searches against the reference bud, culm, leaf and root transcriptomes of *Arundo donax* (Sablok et al., 2014) using default parameters and an E-value cut-off of 10. Only the best hit for each query sequence was retained and after elimination of redundant hits, these candidate primary miRNA sequences were scanned for hairpin-like secondary structures using the miRNA identification pipeline of the C-mii software (Supplementary Figure 1.1; Numnark et al., 2012). For prediction, only the miRNAs of *Oryza sativa* were used as reference, as the annotation of microRNA genes in this species is by far the most complete and reliable. To reduce type I (false positive) errors at the possible expense of a somehow inflated number of false negatives, we applied a stringent filtering of the primary microRNAs (pri-miRNAs) identified by C-mii (Supplementary Figure 1.1). Only candidate sequences fitting the following criteria were considered as putative miRNAs in *Arundo donax*: (1) The length of predicted mature miRNAs should be in the range of 19–25 nucleotides; (2) A maximum of two mismatches compared with known rice mature miRNAs should be allowed for predicted mature miRNAs; (3) The mature miRNA should be localized in only one arm within the predicted stem–loop structure; (4) No more than five mismatches should be allowed between miRNA sequence and guide miRNA sequence in the stem–loop structure; (5) miRNAs should have high A + U content (30-70%); and (6) minimal folding free energy (MFE) and minimal free energy index (MFEI) value of the secondary structure should be highly negative, with a cut-off value of -0.85 kcal/mol (Prakash et al., 2015; Singh et al., 2016; Xu et al., 2008). *Arundo donax* putative microRNAs were renamed according to the closest homologous locus in rice, identified as the best hit in BLASTN searches against all rice pre-miRNAs. Clustering of *Arundo donax* pri-miRNAs was finally carried out to identify tentative genetic loci, as no reference genome sequence is available for this species. Sequences with less than 2 mismatches in

BLASTN searches over the whole alignment length were considered alleles of the same locus and renamed accordingly.

### 1.3.2 Position-specific base composition of mature microRNAs

Nucleotide composition and their dominance at particular positions in mature *Arundo donax* miRNAs and reference *O.sativa* were analyzed by using BioEdit (Hall et al., 2011). Base composition frequency were calculated for each position of *A. donax* and *O.sativa* mature miRNAs, the average percentage of A, C, G and U bases was then calculated across all families. The position-specific nucleotide frequency of predicted mature miRNAs was summarized in graphical form.

### 1.3.3 Structural and phylogenetic reconstruction of different microRNA families

The precursor sequences of the identified *Arundo donax* miRNAs were further analyzed to investigate stem-loop structure variabilities of pre-miRNAs. MUSCLE was used to align sequences which were subsequently used for phylogenetic analysis in MEGA 7.0 by employing the Neighbor joining method with 1000 bootstrap replicates (Kumar et al., 2016).

### 1.3.4 Prediction and functional annotation of putative *Arundo donax* miRNA targets

Putative microRNA targets were identified with two different programs: psRNATarget (Dai et al., 2011) and TargetFinder (Bo et al., 2005). The parameters set for prediction by the psRNATarget server (http://plantgrn.noble.org/psRNATarget/home) were: maximum expectation of the score between small RNAs and their target transcripts: 3.0; complementarity scoring length (hspsize): 20; maximum allowed unpaired energy (UPE): 25; flanking length for analysis of target accessibility: 17 nt upstream and 13 nt downstream of the target site; central mismatch range leading to translation inhibition: 10-11 nt. Prediction of candidate targets with a stand-alone version of the TargetFinder program was carried out with default parameters. Sequences with a score of less than 4 were regarded as predicted miRNA target genes. The transcripts identified by both programs and removed non-coding transcripts by sequence similarity search against known bological protein database were considered as putative microRNA targets and used for subsequent analyses. To better understand the function of *Arundo donax* miRNAs and their regulating targets, Gene Ontology (GO) annotation of the predicted *Arundo donax* miRNA targets were predicted by using the annotation web tool FunctionAnnotator (http://fa.cgu.edu.tw/index.php) (Chen et al., 2012). Further functional annotation of the predicted targets was carried out performing BLASTX searches against the *Arabidopsis thaliana* (https://www.arabidopsis.org/) and *Setaria italica* (http://www.uniprot.org/) protein databases using default parameters and an E-value cut-off of 1e-5. The biological networks formed by the putative miRNAs and their targets were visualized by Cytoscape version 3.5 (Shannon et al., 2003).

### 1.3.5 Comparative genomic analyses of miRNA targets in *Arundo donax* and other plants

All predicted targets of the 11 conserved miRNA families from *Oryza sativa*, *Zea*

*mays*, *Arabidopsis thaliana* and *Vitis vinifera* were downloaded from the PNRD database (http://structuralbiology.cau.edu.cn/PNRD ) (Yi et al., 2015) and used for TBLASTN and BLASTX searches against the putative *Arundo donax* targets with an E-value cut-off of 1e−5. Hits with a Score value greater than 50 and with sequence coverage to the query greater than 50% (R. Pearson, 2013) were retained as conserved homologs, while the others were considered novel targets.

## 1.4 Results

### 1.4.1 Identification of putative miRNAs in *Arundo donax* and their characteristics

Through blast searches of the reference transcriptome of *Arundo donax* (1,195,562 transcript sequences) and microRNA prediction with the C-mii program we identified a total of three hundred and ten miRNA candidates. For reducing the false positives and improving the accuracy of the prediction, the study retained for subsequent analyses only the predicted pre-miRNA with highly negative values of *MFEI* (<= -0.85 kcal/mol). In this way, identified a total of 141 high-confidence putative miRNAs belonging to 14 different families (Supplementary Table 1.1), corresponding to the most common miRNA families in *O. sativa*. *Arundo donax* putative miRNAs varied from 20 to 22 nucleotides in length, with the majority of them being 21 nt in length (85.82%), followed by 20 nt (7.80%), and 22 nt (6.38%), respectively. The lengths of precursor miRNAs varied from 60 to 193 nt with an average value of 99 nt , in line with what has been found in other plant species. The *Ado-MIR444d-1c_b, Ado-MIR444d-2c_b, Ado-MIR444d-2c_c, Ado-MIR444d-2c_l, Ado-MIR444d-3c_c, Ado-MIR444d-4c_l and Ado-MIR444d-5c_r* exhibited the shortest precursor length of 60 nt, whereas *Ado-MIR169n-2_r* showed the longest precursor length of 193 nt (Supplementary Figure 1.2 and Supplementary Table 1.1). Among the 14 miRNA families, 10 (MIR166, MIR396, MIR529, MIR827, MIR160, MIR319, MIR1430, MIR167, MIR171 and MIR172) contained one to nine members, while the remaining four families (MIR156, MIR169, MIR393 and MIR444) were found to have more than ten members. The MIR444 family was the largest family with 55 members (Supplementary Table 1.1 and Figure 1.1). The study used the 94 miRNA loci of *O. sativa* corresponding to the 14 families from miRBase as reference to reliably identify tentative miRNA loci in *A. donax*. Based on the number of paralogs present in rice, A total of 69 loci were identified in *Arundo donax*. The MIR169 family had the highest number of loci both in *O. sativa* and *Arundo donax*. 18.75% of *Arundo donax* MIR169 loci corresponded to single loci in *O. sativa*. The highest number of *Arundo donax* loci per rice gene were seven, in line with the polyploidy of the giant reed (Table 1.1). 26.24% of the miRNAs generated from unique loci and primary transcripts, e.g., locus_17 and locus_18. The miRNA identified from these loci did not show marked tissue-specific preferences, with 29 loci expressed in roots, followed by 20, 17 and 15 in culms, buds and leaves, respectively. However, 13.04% of loci generated multiple primary transcripts and miRNAs. For instance, there were three transcripts corresponding to locus_63. The miRNA identified from this locus belong to the MIR444 family, and these miRNAs were expressed in buds, culms and leaves; another example was the two transcripts from locus_69, which belonged to MIR827 family and they were expressed in

buds and roots (Supplementary Table 1.1).

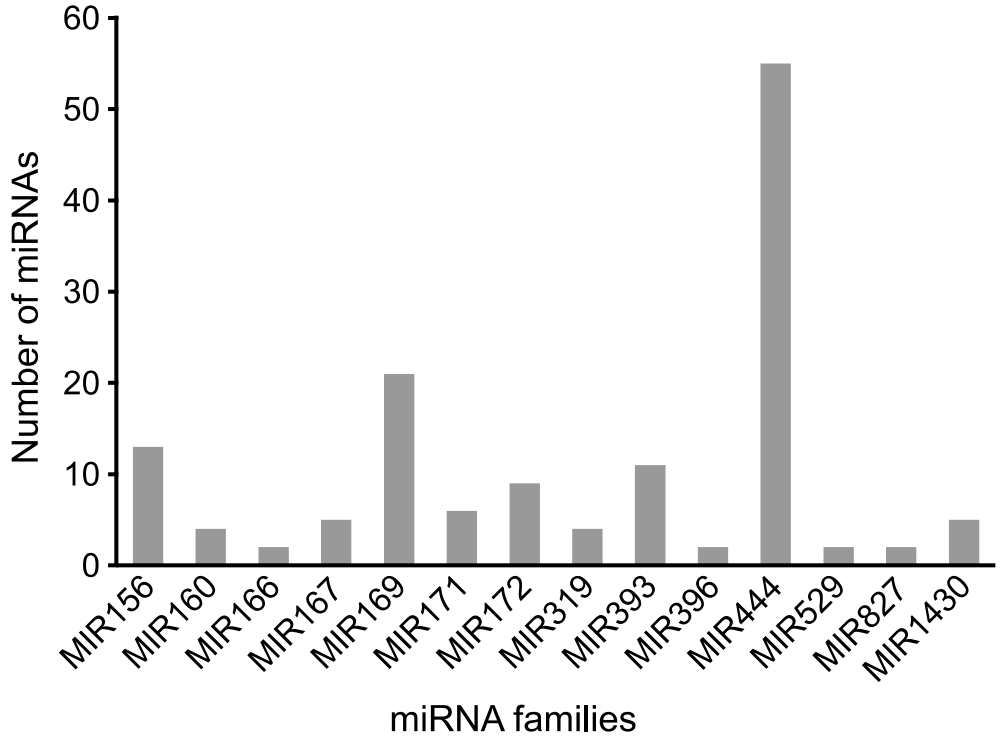



**Figure 1.1.** Number of miRNAs predicted for each of 14 families identified in *Arundo donax*.

**Table 1.1.** Rice homologs of *Arundo donax* microRNA loci.

| miRNA family | loci in *Oryza sativa* | loci in *Arundo donax* | loci in common[#] |
|---|---|---|---|
| MIR156 | 12 | 8 | a(1), b(1), c(1), g(3), j(1), k(1) |
| MIR160 | 6 | 4 | b(3), c(1) |
| MIR166 | 13 | 2 | a(2) |
| MIR167 | 10 | 5 | d(2), g(3) |
| MIR169 | 18 | 16 | a(1), c(3), i(1), n(3), p(1), q(7) |
| MIR171 | 9 | 5 | a(1), c(2), f(1), i(1) |
| MIR172 | 4 | 3 | b(1), d(2) |
| MIR319 | 2 | 2 | a(2) |
| MIR393 | 2 | 4 | a(1), b(3) |
| MIR396 | 8 | 2 | a(1), b(1) |
| MIR444 | 6 | 12 | a(2), c(4), d(5), e(1) |
| MIR529 | 2 | 1 | a(1) |
| MIR827 | 1 | 1 | *(1) |
| MIR1430 | 1 | 4 | *(4) |

#: letters correspond to the names of single loci in *O. sativa*, numbers in brackets correspond to the number of inferred loci in *A. donax*; *: single locus in *O. sativa*.

By analyzing more in-depth the tissue-specific co-expression pattern, the study found that most of the identified miRNA families were preferentially expressed in root (31.9%; Figure 1.2), followed by culm (26.2%), bud (22.7%) and leaf (19.1%). Among the conserved miRNAs, about one third of the families were expressed in all four tissues studied, namely MIR444 (the largest family distributed on a per-tissue basis), MIR169, MIR167, MIR393 and MIR172. Only in a minority of the sampled families were expressed in only one tissue, namely MIR166, which showed specific expression in the root, and MIR396 and MIR529, which were specifically expressed only in the culm (Supplementary Table 1.2). Overall, a relatively limited differential expression was observed for all the predicted miRNA in the four tissues.

MFE is an important parameter for determining the reliability of secondary structures of pre-miRNA, as the stability of the stem-loop structures of the precursor miRNAs is more stable when MFE has highly negative values. In the present study, the range of MFE (-kcal/mol) calculated was -26.4 to -81.8 (kcal/mol) with an average value of -48.67 (kcal/mol). MFEI is the minimal folding energy index, which can be used to distinguish pre-miRNA from other coding or non-coding RNA and RNA fragments. MFEI values ranged from -0.85 (kcal/mol; the maximal cut-off used for prediction) to -1.402 (kcal/mol) with an average of -1.03 (kcal/mol). These values were significantly lower than other reported small RNAs such as tRNAs (-0.64 kcal/mol), rRNAs (-0.59 kcal/mol) and mRNAs (0.62–0.66 kcal/mol), indicating that the identified *Arundo donax* miRNAs were putative miRNAs with high confidence (Supplementary Table 1.1; Adai et al., 2005; Zhang et al., 2006; Bonnet et al., 2004).

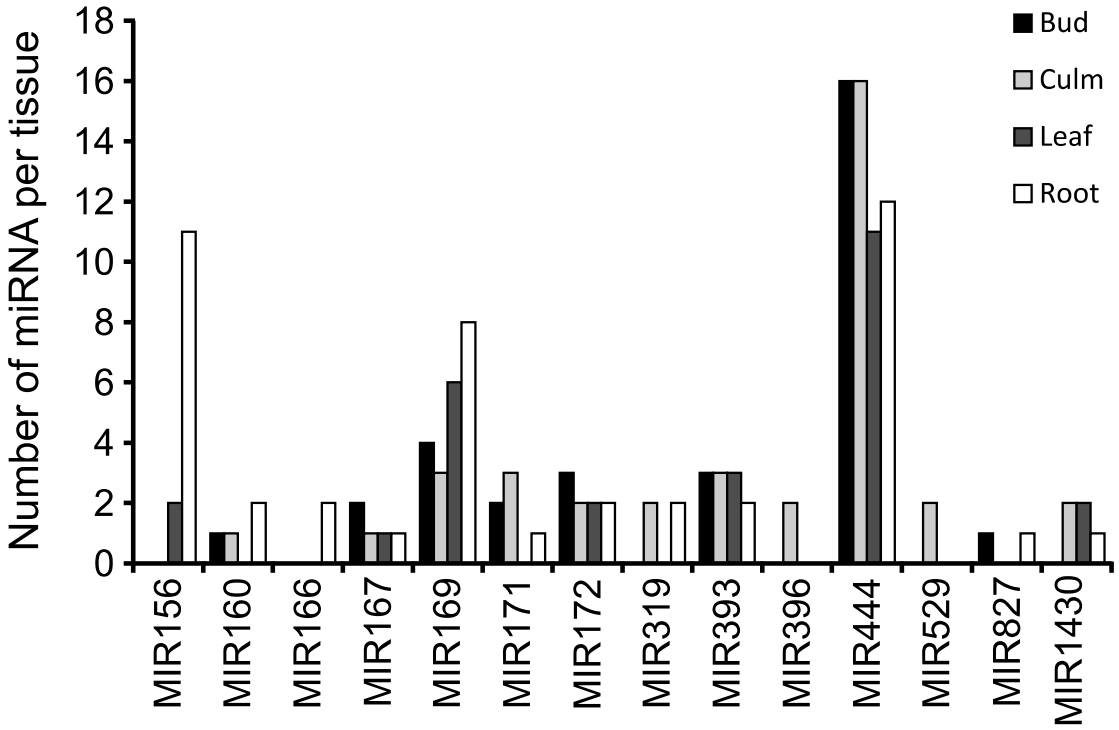



**Figure 1.2.** Number of mature miRNAs present in different tissues (bud, culm, leaf and root) of *Arundo donax*.

### 1.4.2 Analysis of position-specific nucleotide preference in *Arundo donax* mature miRNAs

The overall percentage of each base of mature *Arundo donax* miRNAs was found to be 28.46% for uracil, 24.54% for cytosine, 25.62% for guanine and 21.39% for adenine (Supplementary Table 1.3). These values are in line with base compositions in *Oryza sativa* mature miRNA. In general, a slightly lower GC content was apparent in *Arundo donax* (50.15%) than in *O. sativa* (50.69%) (Figure 1.3A; Supplementary Table 1.3).

In the 5'-end of *Arundo donax* miRNA, uracil was found in 85.11% of the sequences, while in *O. sativa* it was present in 62.71% of the cases. Also other positions showed different base preferences as compared to rice. Cytosine was found to be a dominating base at position 19 (50.35%) in *A. donax*, while in rice it was present in only 40.68% of the cases. Adenine (50.35%) was abundant at the 10th nt position of A. donax mature miRNAs, but less abundant in the same position in rice (35.59%; Figure 1.3B ; Supplementary Table 1.3).

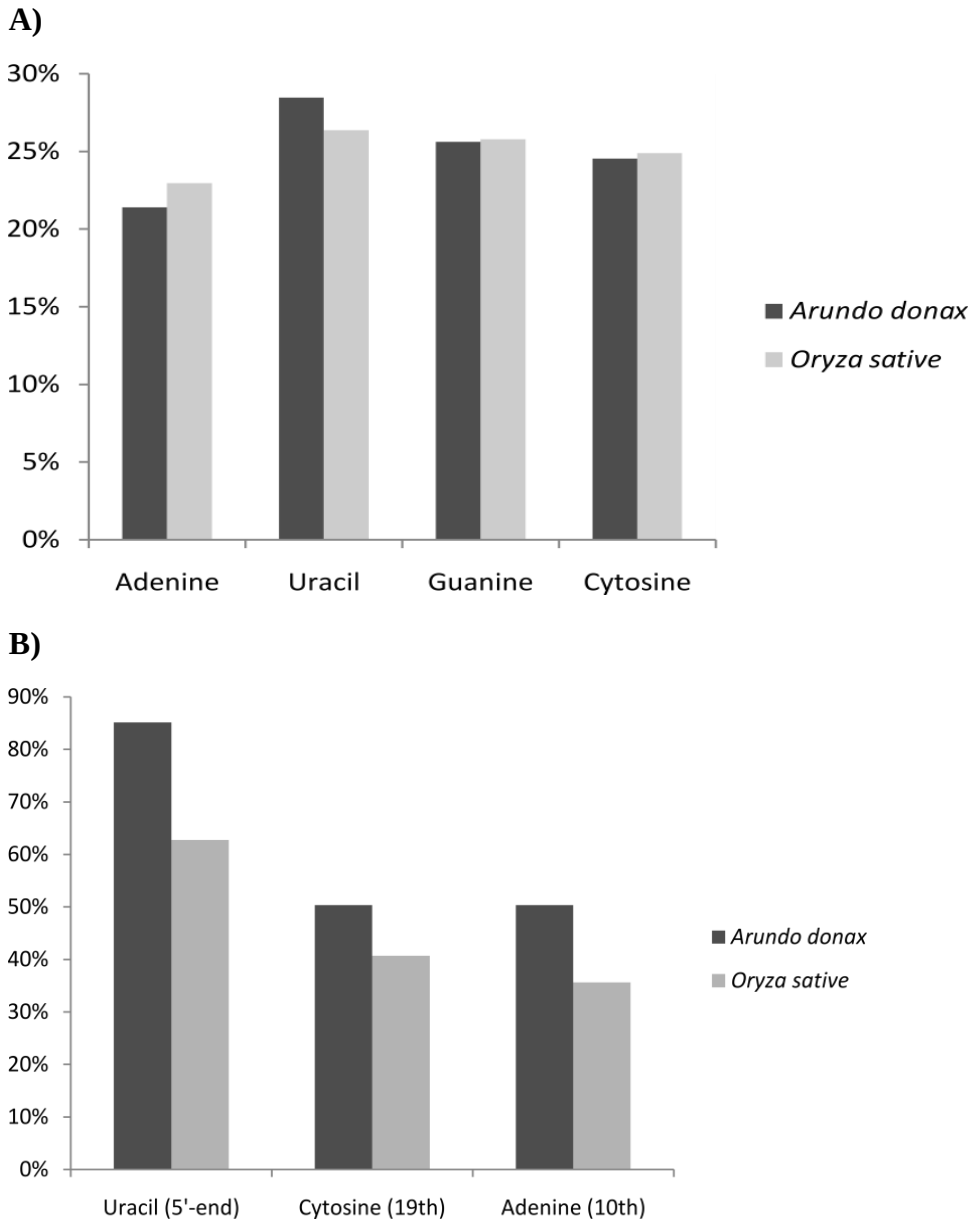



**Figure 1.3.** Composition of nucleotides in mature miRNAs from *A. donax* and *O. sativa*. A) Overall nucleotide compositions (%) compared among mature miRNAs of *A. donax* and *O. sativa*. B) Selected position-specific nucleotide compositions of mature microRNAs in *A. donax* and *O. sativa*.

### 1.4.3 Variability of stem-loop structures in *Arundo donax* pre-miRNAs

To understand folding variability of stem-loop structures and phylogenetic relationship among *Arundo donax* pre-miRNAs, the stem-loop structures of each family member and phylogenetic trees of corresponding family members were constructed using MIR169c, MIR172d and MIR444c multicopy loci as examples. As shown in Figure 1.4, the stem structures were conserved among MIR169c (Figure 1.4A) and MIR172d (Figure 1.4B) paralogs, while divergent in MIR444c (Figure 1.4C). On the contrary, the loop strctures were conserved among MIR444c paralogs, while divergent in MIR169c and MIR172d subfamilies. These results were consistent with the phylogenetic relationships

of MIR169c (Figure 1.4A), MIR172d (Figure 1.4B) and MIR444c (Figure 1.4C) loci. In general, the more similar the miRNA structures were, the more likely they formed supported clades in the respective phylogenetic trees.

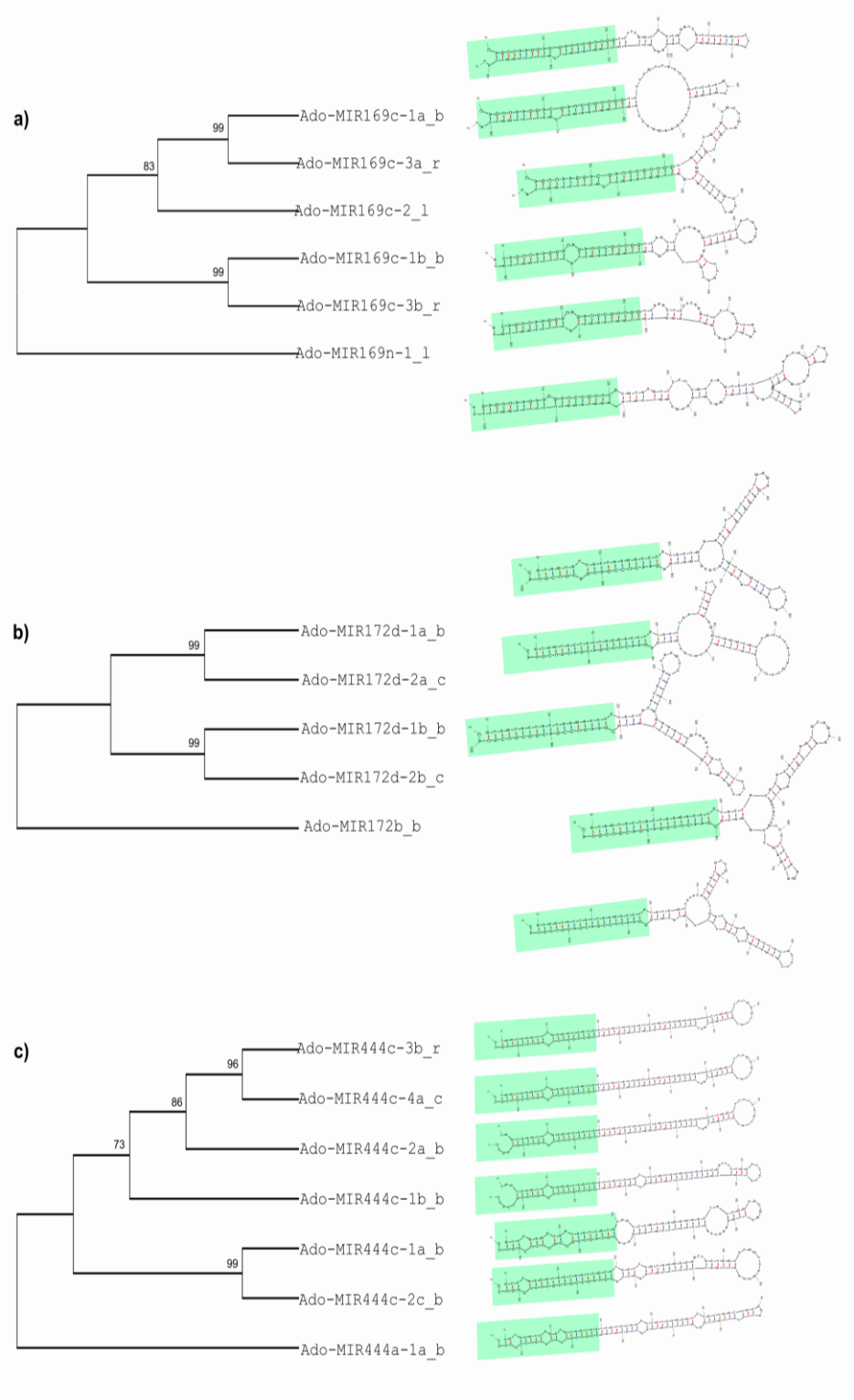



**Figure 1.4.** Stem-loop structures and phylogenetic relationships of *A. donax* pre-miRNA MIR169c (a) , MIR172d (b) and MIR444c subfamilies (c). The single outgroups are at the base of the each cladogram. Uppercase letters identify mature

microRNAs. Regions highlighted in light green are the stem structure of each pre-miRNA. Pre-miRNA structures were drawn with the C-mii software.

### 1.4.4 Target prediction of *Arundo donax* miRNAs

Despite the high sequence complementarity of plant microRNAs and their targets, different algorithms display marked differences in the total number of predicted targets from the same set of microRNAs/mRNA transcripts. For this reason we combined the predictions carried out with two popular programs, Targetfinder and psRNATarget, to attain a more reliable set of true positive targets (Srivastava et al., 2014). By considering only the common hits obtained with these two programs (Supplementary Figure 1.3), and removed non-coding transcripts by similarity search against known bological protein database, the study predicted a total of 107 mature miRNAs out of 141 mature miRNAs to regulate with high reliability 462 non-redundant target transcripts in *Arundo donax*, and the miRNA distributed in families of MIR156, MIR160, MIR166, MIR169, MIR171, MIR172, MIR319, MIR393, MIR444, MIR529 and MIR827 were observed to regulate more than one gene (Figure 1.5). A total of 102 out of 107 *Arundo donax* miRNAs were found to target more than one type of transcript, whereas Ado-MIR160b-2_r, Ado-MIR160b-3_r, Ado-MIR319a-1a_c, Ado-MIR393b-1a_b and Ado-MIR393b-2c_c only targeted single genes (Supplementary Figure 1.4). miRNAs, therefore, tend to regulate multiple distinct genes, and the targets were belonged to several gene families involved in different biological process, cellular component and molecular function (Figure 1.6; Supplementary Figure 1.4). These results suggested that miRNAs combinatorially control multiple processes by regulating different target genes during plant growth and development , as suggested for other species (Yuan et al., 2009; Dehury et al., 2013).

The type of post-translational regulation was analyzed for the predicted targets. Out of 462 targets, 449 were predicted to undergo a cleavage type of inhibition. In particular, MIR156, MIR160, MIR166, MIR169, MIR171, MIR172, MIR319, MIR393, MIR444, MIR529 and MIR827 showed cleavage type of inhibition for all predicted target transcripts. Besides the 449 target transcripts regulated by transcriptional cleavage, MIR444 was the only one predicted to regulate also a total of 13 transcripts by translational suppression. These results indicate that miRNA-mediated Post-transcriptional gene silencing (PTGS) in *Arundo donax* took place mainly through cleavage of the target transcripts, while, in line with what previously observed in plants (Sunkar et al., 2005), translational regulation contributes marginally to overall microRNA regulatory activity.

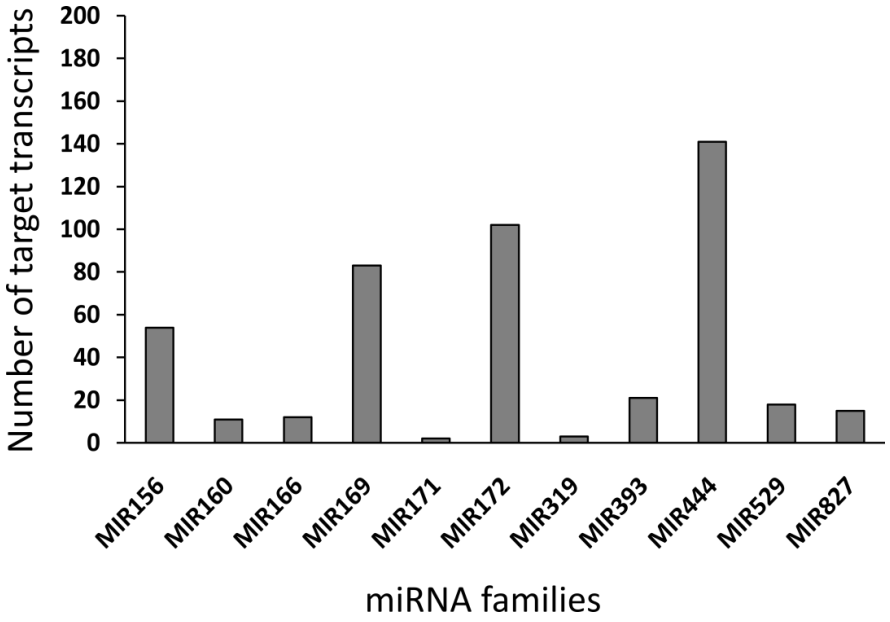



**Figure 1.5.** Number of predicted targets for potential *A. donax* miRNA families.

### 1.4.5 Functional annotation of predicted targets

For functional annotation, gene ontology (GO) analysis was carried out for the predicted *Arundo donax* targets, which indicated their involvement in regulating diverse physiological processes (Figure 1.6). In line with the major role of microRNAs in the regulation of transcriptional cascades by targeting transcription factors (Nazarov et al., 2013), the main molecular functions associated to the predicted *Arundo donax* target genes were binding and more specifically nucleic acid binding transcription factor activity, while only a minority of the functions referred to catalytic activity. For the cellular component category, the majority of the target genes were associated with cell part, organelle and macromolecular complex. Under the biological process category, the majority of targets were associated with biological regulation, developmental process and reproductive process. The functional annotation of the targets also was performed by sequence similarity searches against the *Arabidopsis thaliana* and *Setaria italica* proteins using the BLASTx algorithm, a total of 425 (91.99%) target genes functional hits correspond to *Arabidopsis thaliana*, and a total of 455 (98.48%) target genes functional hits correspond to *Setaria italica*, two popular species with fully sequenced genomes.

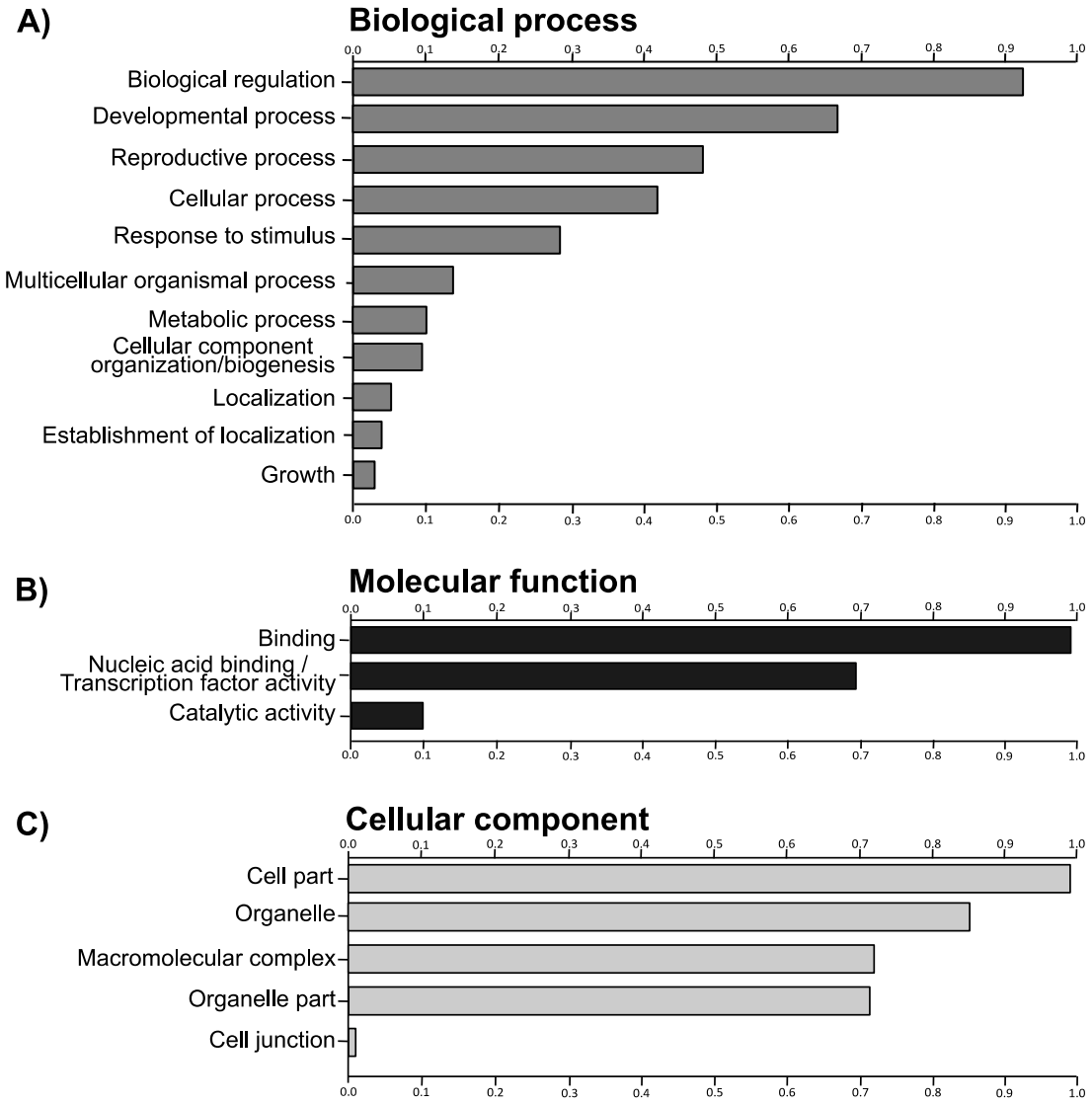



**Figure 1.6.** Functional annotation of predicted targets. (A) biological process, (B) molecular function and (C) cellular component.

### 1.4.6 Conservation of miRNA targets among *Arundo donax* and other plant species

In order to investigate the conservation of the putative miRNA targets in *Arundo donax* compared to other known species namely *Oryza sativa*, *Zea mays*, *Arabidopsis thaliana* and *Vitis vinifera*, BLAST sequence similarity search approach was utilized to identify homolog miRNA targets. There were 390 homolog targets for 10 conserved miRNA families were identified in *Arundo donax* and 72 targets for 8 *Arundo donax* conserved miRNA families were also identified as novel targets (Table 1.2). 84.42% of *Arundo donax* conserved miRNA targets were found to have homologs in other plants, showing that the majority of conserved miRNA targeted conserved target genes in *Arundo donax*, and there were also some conserved miRNAs targeted novel targets.

**Table 1.2.** Numbers of homolog miRNA targets and novel targets in conserved miRNA families in plant genomes.

| miRNA families | *A. donax* | *O. sativa* | *Z. mays* | *A. thaliana* | *V. vinifera* | Novel targets |
|---|---|---|---|---|---|---|
| MIR156 | 50 | 48 | 45 | 40 | 47 | 4 |
| MIR160 | 11 | 10 | 9 | 9 | 10 | 0 |
| MIR166 | 11 | 11 | 11 | 11 | 11 | 1 |
| MIR169 | 77 | 75 | 77 | 68 | 58 | 6 |
| MIR171 | 2 | 2 | 2 | 2 | 2 | 0 |
| MIR172 | 74 | 74 | 74 | 51 | 60 | 28 |
| MIR319 | 0 | 0 | 0 | 0 | 0 | 3 |
| MIR393 | 19 | 19 | 19 | 19 | 19 | 2 |
| MIR444 | 120 | 118 | 110 | - | - | 21 |
| MIR529 | 11 | 10 | 1 | - | - | 7 |
| MIR827 | 15 | 15 | 15 | 0 | - | 0 |
| Total | 390 | 382 | 363 | 200 | 207 | 72 |

**Note: "-" represent the miRNA targets absence.**

## 1.5 Discussion

### 1.5.1 Identification and characterization of conserved miRNAs in *Arundo donax*

Identification of expressed microRNA loci from *Arundo donax*, an important biomass species, may have relevant applications to improve the quality/amount of its biomass, thanks to establishment of promising protocols for regeneration and genetic transformation (Takahashi et al. 2010; Dhir et al. 2010). In addition, it can significantly contribute to elucidate the evolutionary trajectories followed by such important regulators of gene expression in polyploid species (Hardion et al. 2014). Transcriptome sequencing is one of the most efficient and cost-effective approaches available for gene discovery, and can provide massive and valuable information for the identification of low abundance or tissue-specific miRNA and their targets (Xu et al., 2015; Ling et al., 2017; Chen et al., 2015; Fahlgren et al., 2007), phylogenetic inferences and characterization of polyploid speciation (Wang et al., 2017; Buggs et al., 2012). In addition, the choice to leverage on the reference transcriptome of *Arundo donax* for microRNA identification has been dictated by the lack of a fully sequenced genome for this species. As such, this study did not expect to be able to obtain a full representation of all microRNA loci existing in *Arundo donax*. In the study, we decided to carry out a stringent identification of the putative microRNA pri-miRNA by relying on a series of filtering steps. Especially the use of a threshold of -0.85 kal/mol for MFEI provided a set of well resolved miRNA candidates with respect to tRNAs (average MFEI = -0.64), rRNAs (MFEI = -0.59) and mRNAs (MFEI = -0.65) (Zhang et al., 2006; Patanun et al., 2013). The length distributions of predicted pre- and mature-microRNAs (from 60 to193 nt and from 20 to 22 nt, respectively) are in line with those observed in other species (Patanun et al., 2013; Wang et al., 2004; Zhang et al., 2008; Xuan et al., 2011), thus supporting the reliability of the identification. Also the analysis of position-specific nucleotide preferences confirms

similarity of *Arundo donax* mature miRNAs to those from other species. In particular, dominance of uracil at the first position of the 5' terminus may play an important role in miRNA biogenesis or RISC formation, while preference for cytosine at position 19 seems to be relevant for targeting RISC or Dicer-mediated cleavage to specific sites in pre-miRNAs (Zhang et al., 2008; Prakash et al., 2016; Zhang et al., 2006). Comparison of the number of microRNA loci present in the rice genome with the number of tentative loci identified in *Arundo donax* indicates that the major and evolutionarily most conserved families have been identified in our screening (Zhang et al., 2006). The relatively higher numbers of loci identified for families MIR444 and MIR1430, and the involvement of MIR444 in mediating perception of viral infections is well established (see below), the function of MIR1430 remains to be fully elucidated. Previous studies indicate that in rice MIR1430 post-transcriptionally regulate a gene encoding a member of the nuclear factor Y-A (NF-YA) gene family (LOC_Os12g42400; Sun et al. 2015; Thirumurugan et al. 2008). The transcript of LOC_Os12g42400 is expressed specifically in callus, flower and panicles (Liu et al. 2017) and it is co-regulated by MIR169 family members (Sun et al. 2015). In *Arabidopsis* miR169 takes part in stress-induced early flowering (Xu et al. 2014). MIR1430 family expansion may predate *Arundo donax* loss of fertility associated to polyploidization (Hardion et al. 2015), when transition to the reproductive phase likely still constituted a relevant stress-escape strategy for this species. Another possibility is that the likely lineage-specific expansion of MIR1430 gene family may be related to suppression of fungal resistance, as differential repression of NF-YA genes by miR169 has been found to negatively regulate rice immunity against the blast fungus *M. oryzae* (Li et al. 2017).

### 1.5.2 Functional annotation of putative targets

In line with the known tendence of microRNAs to preferentially target transcription factors (Cui et al., 2007), we found that the major molecular function classes targeted in *Arundo donax* were “binding” and “nucleic acid binding transcripton factor activity”. Also the most common biological process GO classes indicate that the set of microRNAs identified in *Arundo donax* play an important role in the regulation of development, reproduction, metabolism as well as stress response. In particular, network analysis of *Arundo donax* miRNAs and their target genes highlighted the large number of targets regulated by MIR444, MIR172, MIR169 and MIR156. In *Oryza sativa*, RNA-dependent RNA ploymerase1 (RDR1) is a central component in the antiviral RNA-silencing pathway, and MIR444 plays an important role in transducing the antiviral signal from virus infection to RDR1 expression (Wang et al., 2016). The capacity of *Arundo donax* to asyntomatically stand viral infections may, thus, in part depend on the possible expansion of the MIR444 familiy in this species (Ingwell et al. 2014). Previous studies showed that MIR156 regulates developmental timing by repressing the expression of functionally distinct SPL transcription factors, while MIR172 regulates flowering time and flower formation by regulating the expression of AP2-like transcription factors (Wu et al., 2009; Zhu et al., 2011). MIR172 could, therefore, be a useful tool to modulate flowering time in *Arundo donax*, analogously to what observed in Sorghum (Calviño et al., 2011). The MIR169/NF-YA (Nuclear factor Y, subunit A) is a well established regulatory module

functioning in developmental processes and responding to environmental stresses (Sorin et al., 2014; Li et al., 2010). Also MIR166 may possibly respond to environmental stress and it controls root architecture (Boualem et al., 2008), a trait that may be relevant for improving *Arundo donax* tolerance to drought (Fu et al. 2016).

## 1.6 Conclusion

The growing interest towards improvement of biofuel/bioenergy crops has stimulated in recent years the search for novel approaches to improve their productivity. As specific miRNAs regulate several bioenergy traits, genetic transformation of bioenergy crops like switchgrass and poplar with selected miRNAs has been already demonstrated a viable option to improve plant biomass, decreasing the lignin content, modulating stress responses and flowering time (Fu et al. 2012; Rubinelli et al. 2013). This is, however, still just a small fraction of the 30 microRNA families of potential interest for bioenergy crop improvement (Trumbo et al. 2015). Dissecting the genetic architecture of miRNA loci in the crop of interest is the first fundamental step for any subsequent attempt to improve biofuel feedstock species (Trumbo et al. 2015). The putative microRNA loci identified in the present study from the transcriptome of *Arundo donax* provide novel opportunities for the genetic improvement of biomass yield and quality in this emerging biomass species. They also shed new light into the complex dynamics of microRNA evolution in this highly polyploid species, providing evidence for lineage-specific amplification of microRNA families involved in important aspects of plant fitness and productivity, like viral resistance and regulation of flowering time.

# CHAPTER 2

Computational predictions and comparative analyses of conserved microRNAs from *Arundo* leaf transcriptomes

## 2.1 Abstract

MicroRNAs (miRNAs) are a kind of small non-coding RNA molecules with the length of 19-25 nucleotides regulating gene expression at the post-transcriptional level in plants. In this study, a total of 235 miRNAs belonging to 37 miRNA families and a total of 175 high-confidence putative targets were identified by using computational approach in *Arundo* leaf transcriptomes. The minimum value of precursor miRNA minimal folding energy index (MFEI) was presented to be -0.662 kcal/mol. Gene Ontology functional annotation showed that most of the miRNA targets are involved in multiple biological processes. Among the different miRNA families identified, MIR444, MIR167, MIR159 and MIR162 are universally expressed among *Arundo* species. Phylogenetic analysis based on the highly conserved miRNA159 family indicated that different miRNAs evolved with different rates in *Arundo* genus, confirmed that miRNAs are evolutionarily conserved in plants, and suggested that conserved miRNA regulated homologous targets at the conserved target sites. This is the first *in silico* mining and comparative analyses of miRNAs and their putative targets from *de novo* assemly of leaf transcriptomes of taxa from the *Arundo* genus. These findings increased our understanding of the miRNA regulation in *Arundo* species. The functional annotation of predicted miRNA targets may provide useful information for further functional analyses and experimental validation.

## 2.2 Introduction

MicroRNAs (miRNAs) are small non-coding RNA molecules including almost 22 nucleotides in length, they are generated from ~ 70 nucleotide precursors miRNA with stem-loop hairpin structures (Lee et al., 2002). MicroRNAs are found in plants and animals, and these miRNAs function in post-transcriptional level and gene regulation by targeting mRNAs (Bartel 2004). At present, there are lots of plant miRNAs having been identified and their functions also have been confirmed in plant growth and development via different experimental designs and computational approaches (Zhang et al., 2006). In plants, computational approaches on comparative studies have been successfully applied and demonstrated effectively to attain a comprehensive prediction and characterization of potential miRNAs (Lindow and Krogh, 2005; Patanun et al., 2013; Dong et al., 2012; Wang et al., 2004; Archak and Nagaraju, 2007). In addition, there are useful and relatively complete online miRNA databases like miRBase (Griffiths-Jones et al., 2008), collecting and depositing all putative miRNAs predicted via in silico approaches, greatly helpful for microRNA identification in novel transcriptomes, and the feature of nearly perfect or perfect complementarity of miRNAs to their target mRNA sequences, which allowed the genome-wide identification of microRNA genes (Rhoades et al., 2002; Schwab et al., 2005; Devi et al., 2016).

*Arundo* genus belongs to the Arundineae tribe of Arundinoideae, in poaceae family and the different species in *Arundo* genus are normally tall, hard stalk, propagated by rhizomatous and wetland adaptable. Recent revisions of the genus suggested that up to five taxa are included in *Arundo* genus (Hardion et al., 2012), they are occurring in some lowlands and the landscape of human disturbance. *Arundo formosana* is a local Taiwan grass with fast growth in rainfall environment (Lin et al., 2006). *Arundo micrantha* is

distributed in the Mediterranean region and even in north Africa,threatened in freshwater habitat due to the invasive species *Arundo donax* (Hardion et al., 2012; Mascia et al., 2013). *Arundo donaciformis* occurs mainly in the southern France and northwest Italy and functions in preventing soil erosion from powerful rhizomes (Hardion et al., 2012; Hardion et al., 2015). *Arundo collina* is a drought-resistance plant for protecting bare hillside erosion, However, due to fierce competition, there are some problems unresolved in the natural regeneration and artificial regeneration of this species (Danin et al., 2002; Danin, 2004). *Arundo plinii*, roughly growing up to 2 meters, is commonly distributed in Malta and Italy, Croatia and Greece, It grows and disperses through wind via seeds (Hardion et al., 2014). *Arundo donax* is a perennial rhizome $C_3$ grass and an infertile polyploidy plant, there is no appropriate explanation for the cause of its sterility, because the chromosome number is not determined and is predicted most probably ranging from 108 to 110 (Bucci et al., 2013). It requires little management input and lacks of natural competition, and great adaptability and resistance to most pests and pathogens, non-native and threaten biodiversity, so it is regarded as an invasive grass (Pilu et al., 2012). This robustness also makes it a very productive biomass species, as in optimal conditions *Arundo donax* fields become productive already after the second year and can provide dry biomass yields up to 40 tons per hectare for the next ten years (Angelini et al., 2009). In this study, the computational identification and comparative analyses of conserved miRNAs and their targets were carried out for *Arundo* genus using *de novo* assembly of leaf transciptomes. These findings may provide basis information for further functional analyses and experimental validation in the future studies.

## 2.3 Materials and methods

### 2.3.1 Plant materials, transcriptome dataset and reference miRNA

Seven *Arundo* species were used in this study: *Arundo collina*, *Arundo donaciformis*, *Arundo donax*, *Arundo formosana*, *Arundo macrophylla*, *Arundo micrantha* and *Arundo plinii*. The plant material was collected from 3 individuals from each species, grown in a commen garden experiment at the Edmund Mach Foundation. Leaf tissues were sampled from the first and second fully developed leaves of each individual, in order to standardize as much as possible transcript sampling, and Paired-end RNA-Seq libraries were prepared using the TruSeq RNA Sample Prep V2 kit (Illumina, San Diego, CA), pooled in equimolar ratio and sequenced on an Illumina HiSeq 2000 sequencing platform. These *Arundo* leaf transcriptome data were used for the miRNA and their targets prediction. A total of 740, 618 assembled *Arundo* transcriptome unigenes were utilized for computational prediction of miRNA (Detailed information of assemble refer to CHAPTER 3 of this thesis). All previously known 1616 miRNA precursor sequences from 12 monocotyledon species were downloaded from the miRBase database (Release 21.0; http://www.mirbase.org/) (Kozomara et al., 2014). These precursor miRNAs were used as query sequences for BLASTN searches against the leaf transcriptomes of *Arundo* genus using default parameters and an E-value cut-off of 10.

### 2.3.2 *In silico* prediction of potential miRNAs

Those sequences which the best hit for each query sequence were retained and after elimination of redundant hits, these candidate primary miRNA sequences were scanned for hairpin-like secondary structures using C-mii software (Numnark et al., 2012). And the miRNAs of *Oryza sativa* were used as reference, due to the microRNA genes in this species is the most complete and reliable. The candidate sequences were considered as putative miRNAs in *Arundo* genus by the criteria set as CHAPTER 1 of this thesis except the following criterion for relaxation: (1) A maximum of four mismatches allowed between the known rice miRNAs and predicted mature miRNAs; and (2) minimal folding free energy (MFE) and minimal free energy index (MFEI) value of the secondary structure should be highly negative, with a cut-off value of −0.66 kcal/mol (Prakash et al., 2015; Singh et al., 2016; Xu et al., 2008). Putative microRNAs in *Arundo* genus were renamed according to the closest homologous locus in rice, identified as the best hit in BLASTN searches against all rice pre-miRNAs.

### 2.3.3 Phylogenetic analysis of the putative miRNAs

The precursor sequences of the identified *Arundo* miRNAs were further analyzed to investigate the evolutionary relationships among *Arundo* miRNA and *Oryza sativa* miRNA (http://www.mirbase.org/). The sequences were aligned by MUSCLE and the program Gblock v0.91 (Castresana, 2000) was used to select conserved blocks with less strict parameters: maximum number of contiguous nonconserved positions = 8, minimum length of a block = 5, and allowed gap positions: with half. The best-fitting substitution model of nucleic acid evolution inferred by SMS server web (Smart Model Selection, web server: http://www.atgc-montpellier.fr/sms/) (Lefort et al., 2017), Maximum-likelihood (ML) tree was reconducted by PhyML v3.0 (Guindon et al., 2010), Branch Support was calculated using aLRT (approximate Likelihood-Ratio Test) (Anisimova and Gascuel, 2006).

### 2.3.4 Prediction of miRNA targets and Functional annotation

In order to investigate the function of putative miRNA, their targets were identified via two different programs: psRNATarget and TargetFinder (Detailed information of the parameter sets refer to CHAPTER 1 of this thesis) (Dai et al., 2011; Bo et al., 2005). The candidate targets identified by both programs after removing non-coding genes by sequence similarity search against known biological protein database were considered as putative microRNA targets and used for subsequent analyses. Gene Ontology (GO) annotation of the predicted miRNA targets was carried out by using the annotation web tool FunctionAnnotator (http://fa.cgu.edu.tw/index.php) (Chen et al., 2012). Further functional annotation of the predicted targets was carried out performing BLASTX searches against the *Arabidopsis thaliana* (https://www.arabidopsis.org/) protein databases using default parameters and an E-value cut-off of 1e-5, query coverage larger than 65. (Shannon et al., 2003).

### 2.3.5 Comparative analyses of miRNA targets in *Arundo* species

In order to investigate the conservation of the putative miRNA and their target sites across *Arundo* genus, in the study, multiple alignment software MUSCLE implemented in MEGA 7.0 was utilized to align sequences of conserved miRNA family and their homolog targets generated by local blast search against *Arundo* genus with the corresponding reference miRNA and targets from *O. sativa* (Kumar et al., 2016).

## 2.4 Results

### 2.4.1 Identification of putative miRNAs in different *Arundo* species and their characteristics

The Pipeline utilized for prediction of miRNA and targets from *Arundo* leaf transcriptome is presented in Figure 2.1. Through blast homology searches of the leaf transcriptome in *Arundo* species (740,618 transcript sequences) and microRNA prediction with the C-mii program this study identified a total of 387 redundant miRNA candidates. For reducing the false positives and redundancy, the pre-miRNAs were predicted with manual inspection applying the values of MFEI (<= -0.66 kcal/mol) for distinguishing pre-miRNA from other coding or non-coding RNA and RNA fragments. In the present study, the precursor miRNA MFEI values were ranged from -0.662 (kcal/mol) to -1.617 (kcal/mol), these values were lower than other reported small RNAs such as tRNAs (-0.64 kcal/mol), rRNAs (-0.59 kcal/mol) and mRNAs (0.65 kcal/mol), indicating that the identified *Arundo* miRNAs were most likely putative miRNAs (Supplementary Table 2.1; Adai et al., 2005; Zhang et al., 2006; Bonnet et al., 2004; Patanun et al., 2013). Finally, a total of 235 putative miRNAs belonging to 37 different families were identified in seven *Arundo* species leaves: 43 from *Arundo collina*, 27 from *Arundo donaciformis*, 38 from *Arundo donax*, 15 from *Arundo formosana*, 36 from *Arundo macrophylla*, 23 from *Arundo micrantha* and 53 from *Arundo plinii* (Figure 2.2). In this study, MIR444, MIR167, MIR159 and MIR162 families were found across seven *Arundo* species with 80 putative mature miRNAs identified. Besides, comparative analyses of miRNA families indicated that there were preferential expression of some miRNA families in leaves of different *Arundo* species. For example, MIR1862 was identified in *Arundo donaciformis* leaf transcriptome, MIR171 and MIR2275 found in *Arundo plinii*, MIR5337 and MIR5831 found in *Arundo macrophylla*, MIR530 and MIR1879 in *Arundo donax*, MIR1432 and MIR164 in *Arundo collina*, MIR11340, MIR1866 and MIR5824 in *Arundo micrantha* (Table 2.1).

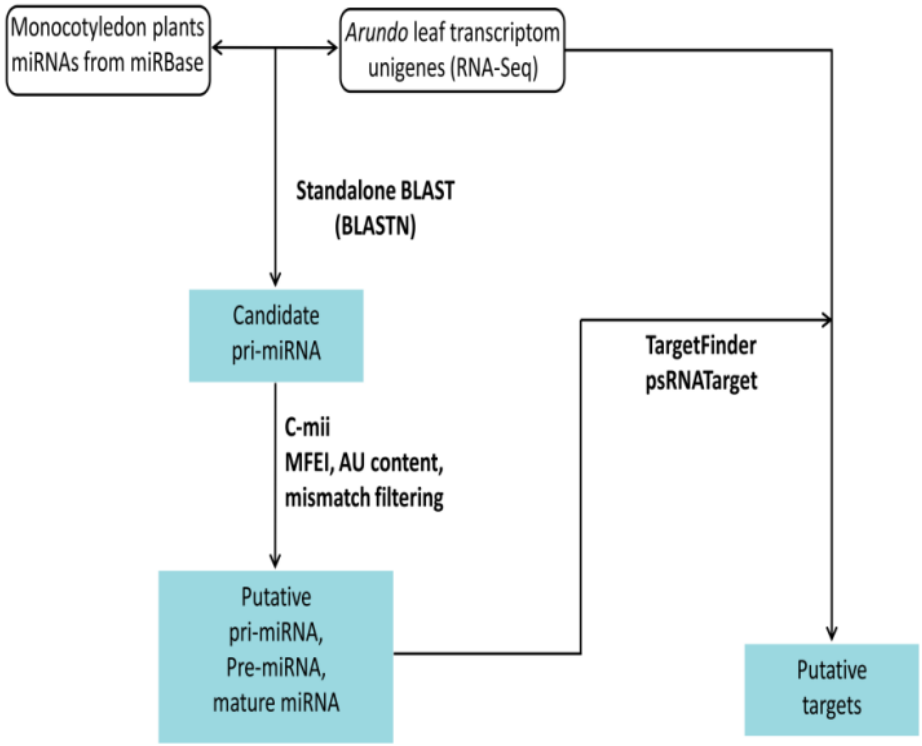



**Figure 2.1.** Schematic workflow of the analyses for the identification of microRNAs and their targets in *Arundo* genus.

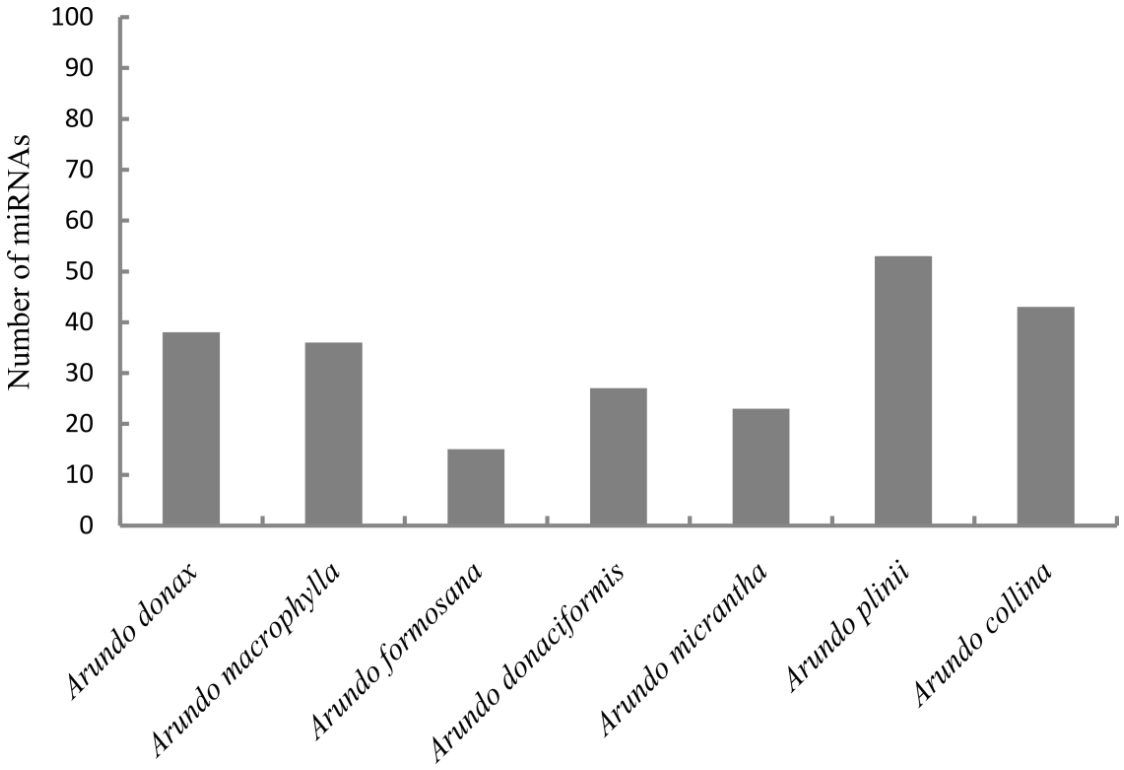



**Figure 2. 2.** Number of putative mature miRNA in *Arundo* genus.

A total of putative miRNAs in *Arundo* species varied from 20 to 24 nucleotides in length with the majority of them being 21 nt in length (83.40%), followed by 20 nt (9.79%), 22 nt (4.26%) and 24 (2.55%), respectively. The lengths of precursor miRNAs varied from 40 to 317 nt with an average value of 134 nt shown in Figure 2.3, in line with what has been found in other plant species. The Ami-MIR818a-16 exhibited the shortest precursor length of 40 nt, whereas Adof-MIR1862e-3 showed the longest precursor length of 317 nt among *Arundo* species.

**Table 2.1.** MiRNA families identified in *Arundo* genus.

| **miRNA families** | **Ado** | **Ama** | **Afo** | **Adof** | **Ami** | **Apl** | **Aco** |
|---|---|---|---|---|---|---|---|
| MIR156 | 0 | 3 | 2 | 2 | 1 | 0 | 2 |
| MIR159 | 2 | 3 | 2 | 2 | 2 | 3 | 4 |
| MIR160 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| MIR162 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| MIR164 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| MIR166 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| MIR167 | 3 | 4 | 3 | 1 | 1 | 2 | 2 |
| MIR168 | 0 | 0 | 1 | 1 | 1 | 0 | 2 |
| MIR169 | 8 | 0 | 1 | 4 | 1 | 3 | 3 |
| MIR171 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| MIR172 | 2 | 2 | 0 | 0 | 2 | 5 | 2 |
| MIR393 | 2 | 1 | 1 | 0 | 0 | 0 | 0 |
| MIR396 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| MIR399 | 0 | 1 | 0 | 0 | 0 | 4 | 1 |
| MIR408 | 0 | 1 | 0 | 1 | 1 | 1 | 2 |
| MIR437 | 2 | 4 | 0 | 3 | 0 | 1 | 1 |
| MIR444 | 12 | 4 | 3 | 6 | 3 | 5 | 6 |
| MIR528 | 0 | 1 | 1 | 1 | 1 | 1 | 2 |
| MIR530 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| MIR812 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| MIR818 | 1 | 2 | 0 | 0 | 2 | 8 | 3 |
| MIR827 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| MIR1430 | 0 | 0 | 0 | 2 | 0 | 1 | 1 |
| MIR1432 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| MIR1435 | 0 | 2 | 0 | 0 | 0 | 1 | 0 |
| MIR1439 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| MIR1862 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| MIR1866 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| MIR1879 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| MIR2275 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| MIR2905 | 0 | 2 | 0 | 0 | 0 | 0 | 1 |
| MIR5143 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| MIR5161 | 1 | 0 | 0 | 1 | 2 | 4 | 5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| MIR5337 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| MIR5824 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| MIR5831 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| MIR11340 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

**Note:** The number of the miRNAs distributed in each miRNAs families are available in the table.

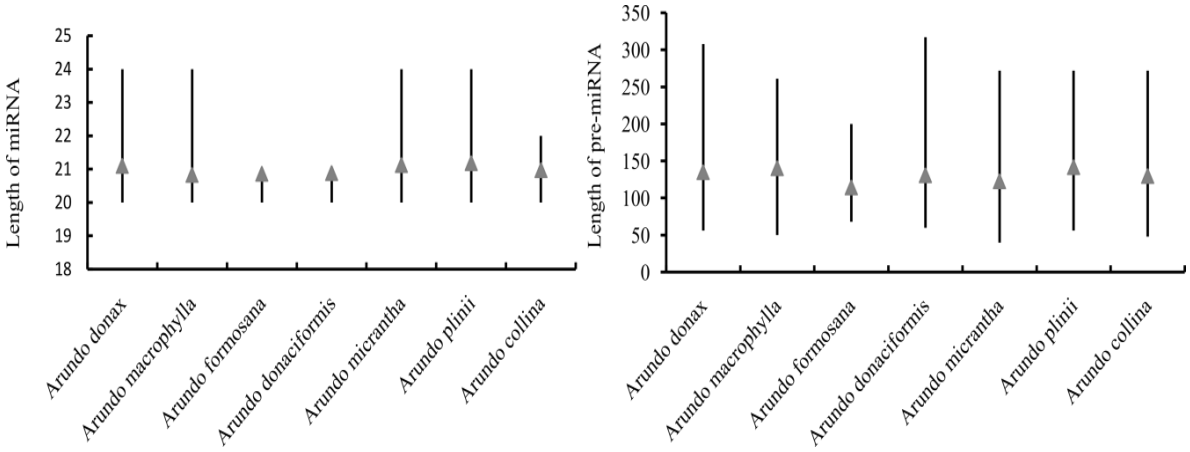



**Figure 2. 3.** Length distributions of mature miRNA and pre-miRNA in *Arundo* genus.

### 2.4.2 Phylogenetic analysis

To better understand phylogenetic relationships among *Arundo* species pre-miRNAs and precursors miRNA in *Oryza sativa* (Osa), the phylogenetic trees of corresponding family members were constructed. The phylogenetic tree reconstruction of the conserved MIR159 family sequences showed that the pre-miRNA in MIR159 family from seven *Arundo* species and *Oryza sativa* were divided into three groups (MIR159a, MIR159b and MIR159d), indicating that different miRNAs evolved with different rates and further confirming that miRNAs were evolutionarily conserved in different plant species (Figure 2.4).

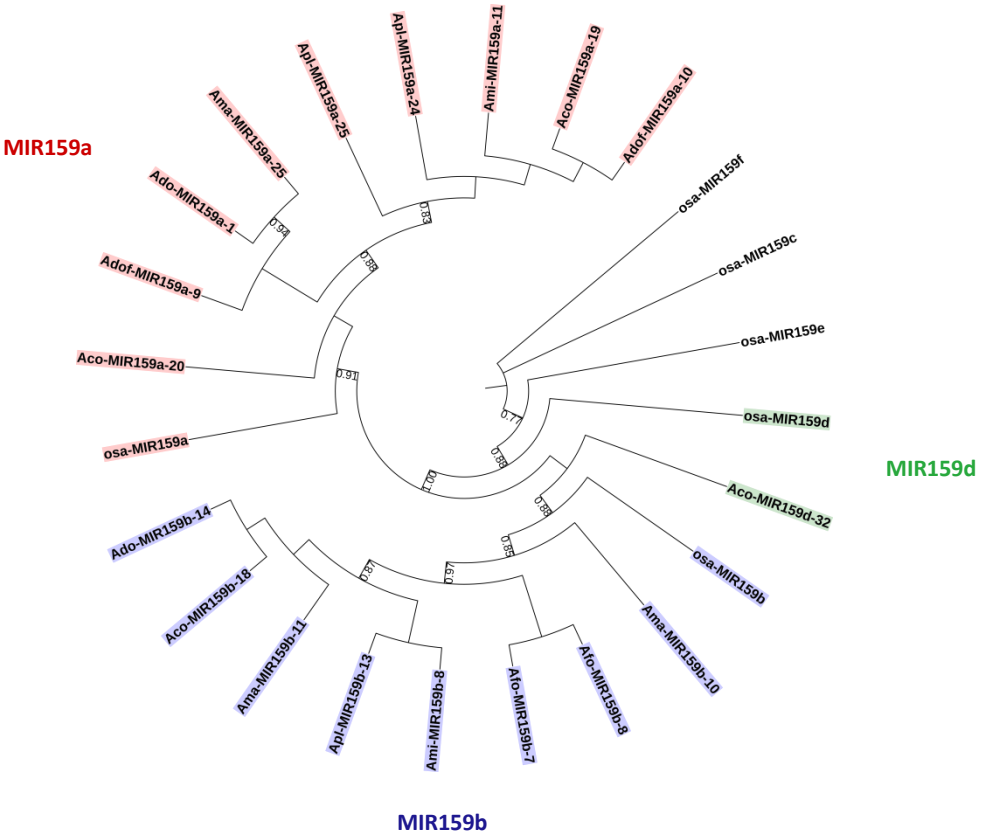



**Figure 2.4.** Phylogenetic reconstruction with precursor miRNAs of conserved MIR159 family from seven *Arundo* species and *Oryza Sativa*.

### 2.4.3 Target prediction of miRNAs and Functional annotation

In order to attain a more reliable set of true positive targets, the predictions were carried out with two popular programs, Targetfinder and psRNATarget. By considering only the common hits obtained with these two programs, we predicted a total of 164 mature miRNAs out of 235 mature miRNAs to regulate with high reliability 175 non-redundant target genes in *Arundo* genus (Supplementary Table 2.2). miRNAs, therefore, tend to regulate multiple distinct genes, which are involved in different biological process, cellular component and molecular function (Figure 2.5). These results indicated that miRNAs in *Arundo* species involved in multiple processes by regulating different target genes during metabolic process and developmental stages, as found in other plant species (Yuan et al., 2009; Dehury et al., 2013).

For functional annotation, gene ontology (GO) analysis was carried out for the predicted *Arundo* targets, which indicated their involvement in regulating diverse biological processes (Figure 2.5). Similar to the major role of microRNAs in the regulation of transcriptional cascades by targeting transcription factors (Nazarov et al., 2013), the main molecular functions associated to the predicted *Arundo* target genes were binding, catalytic activity and transcription regulator activity. For the cellular component category, the majority of the target genes were associated with cell part, organelle, membrane and protein-containing complex. Under the biological process category, target

transcripts were associated with biological regulation, metabolic process and cellular process. The functional annotation of the targets was completed by sequence similarity searches against the *Arabidopsis thaliana* proteins using the BLASTx algorithm, a total of 76 (43.43%) target genes corresponding to *Arabidopsis thaliana* orthologues were identified (Supplementary Table 2.3).

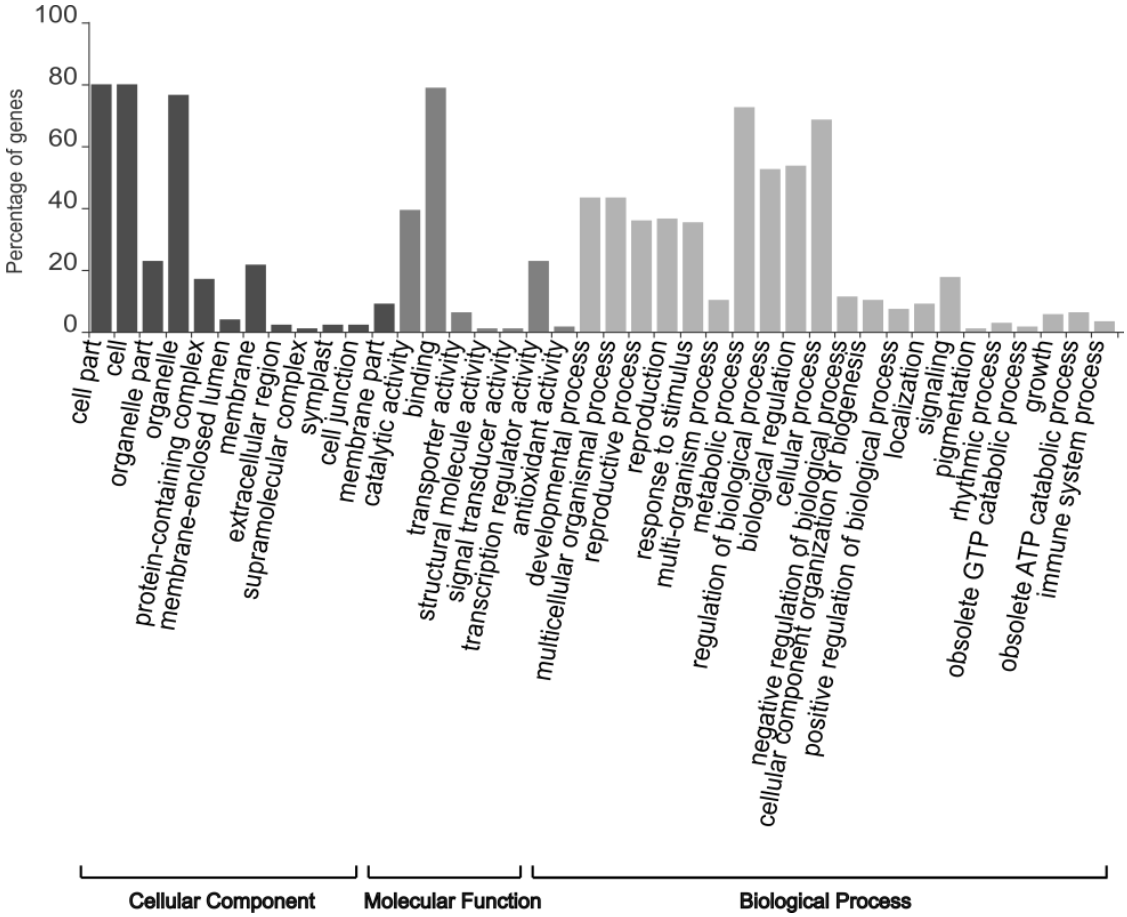
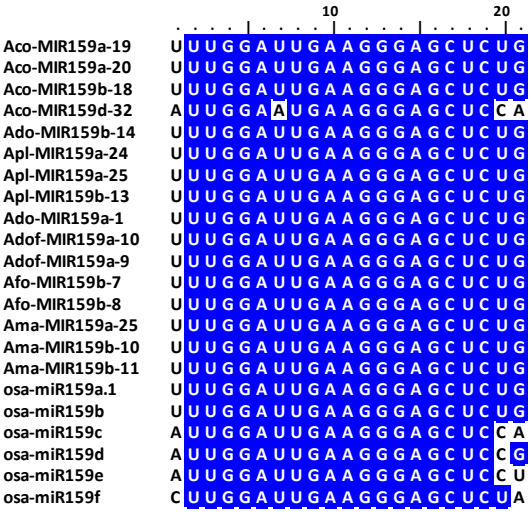



**Figure 2.5.** Gene ontology (GO) functional annotation of target genes of *Arundo* miRNAs.

### 2.4.4 Conservation of miRNA targets among *Arundo* genus

Multiple alignment software MUSCLE was utilized to align conserved family and their homolog target sequences from *Arundo* genus and *O. sativa*. The result is shown in Figure 2.7: conserved plant MIR159 family members regulate homologous targets at identical target sites in different species. In this study, the conserved miRNAs were confirmed to have recognition binding sites in homolog targets, these target sites are conserved among different species (Michael and David, 2005; Axtell and Bowman, 2008; Li et al., 2010; Lenz et al., 2011).

**A)**




**B)**

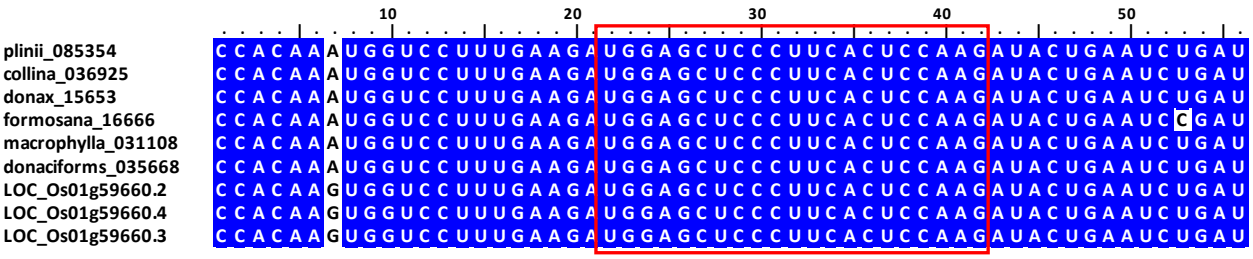



**Figure 2.6.** miRNA159 and their homolog targets across diverse plant species. Homolog targets were identified by local blast search, the alignment of MIR159 and fragments of their target mRNAs was carried out by MUSCLE alignment software. Blue shading indicates bases present in >80% of sequences. Red box represents the complementarity region of miRNA to mRNAs. Sequence of Plinii_085354 come from species Apl; Collina_036925 from Aco; donax_15653 from Ado; formosana_16666 from Afo; macrophylla_031108 from Ama; donaciforms_035668 from Adof; LOC_Os01g59660.2, LOC_Os01g59660.3 and LOC_Os01g59660.4 from osa.

## 2.5 Discussion

### 2.5.1 Identification and comparative analyses of conserved miRNAs in *Arundo* genus

Comparative analyses of miRNAs in seven *Arundo* species resulted in the identification of conserved miRNAs in *Arundo* leaves. Identification and comparative analyses of conserved microRNAs from *Arundo* genus may enhance our understanding of growth regulation, development and response to various stresses. Transcriptome sequencing is one of the most efficient and cost-effective approaches available for gene discovery, and it can provide massive and valuable information for the identification of

low abundance or species-specific miRNA and their targets (Xu et al., 2015; Ling et al., 2017; Chen et al., 2015; Fahlgren et al., 2007), for phylogenetic inferences, and for the characterization of polyploid speciation (Wang et al., 2017; J. A. BUGGS et al., 2012). The miRNA expression levels can vary in different species, but the sampling design should have kept sampling variance as low as reasonably possible. In this study, the plant material was collected from 3 individuals from each species, grown in a commen garden experiment at the Edmund Mach Foundation. Leaf tissues were sampled from the first and second fully developed leaves of each individual, in order to standardize as much as possible transcript sampling. A total of 235 miRNAs were identified in seven *Arundo* species, belonging to 37 conserved miRNAs families. The length distributions of predicted pre- and mature-microRNAs (from 40 to 317 nt and from 20 to 24 nt, respectively) are similar to those found in other species (Patanun et al., 2013; Wang et al., 2004; Zhang et al., 2008; Xuan et al., 2011), thus supporting the reliability of the identification. Although miRNAs are conserved in plant species, some of them can undergo gain or loss events (Ma et al., 2010). MIR444 family is ubiquitously expressed among *Arundo* species leaf, which is in line with the expression of miR444 family is leaf-biased in *Oryza sativa* (Wen et al., 2016). In this study, MIR159, MIR167 and MIR162 were identified across *Arundo* genus species, and this result was highly consistent with previous studies. MIR159 was highly conserved, MIR167 and MIR162 were moderately conserved miRNAs in plants (Zhang et al., 2006). However, some miRNA families specifically expressed in *Arundo* leaf were identified, for example, MIR1430 family were present in *Arundo donaciformis*, *Arundo collina* and *Arundo plinii*, but not in other *Arundo* species. Previous studies indicated that in rice MIR1430 post-transcriptionally regulate a gene encoding a member of the nuclear factor Y-A (NF-YA) gene family (Sun et al. 2015; Thirumurugan et al. 2008). MIR169 was found to be expressed in all *Arundo* species, but not in *Arundo macrophylla*, possibly due to the filtering strategy. This miRNA is interesting as it takes part in stress-induced early flowering in *Arabidopsis* (Xu et al. 2014). The MIR169/NF-YA (Nuclear factor Y, subunit A) is a well established regulatory module functioning in developmental processes and responding to environmental stresses (Sorin et al., 2014; Li et al., 2010), a trait that may be relevant for improving biomass species *Arundo donax* and its tolerance to drought (Fu et al. 2016).

### 2.5.2 Functional annotation of putative miRNA targets

Functional annotation indicated that the putative miRNA targets are mainly involved in binding and transcription regulation, in line with the major functions of miRNAs which preferentially target transcription factors (Cui et al., 2007). In addition, the most common biological process GO classes indicate that the set of microRNAs identified in *Arundo* species play important roles in the regulation of development, reproduction, metabolism as well as stress response. For example, MIR444 is conserved among these seven *Arundo* species, also in *Oryza sativa*, RNA-dependent RNA ploymerase1 (RDR1) is a central component in the antiviral RNA-silencing pathway, and MIR444 plays an important role in transducing the antiviral signal from virus infection to RDR1 expression (Wang et al., 2016). However, the MIR156 was not identified in the leaf

transcriptome of the *Arundo donax* and *Arundo plinii*, and MIR172 was not identified in *Arundo formosana* and *Arundo donaciformis*. Previous studies showed that MIR156 regulates developmental timing by repressing the expression of functionally distinct SPL transcription factors, while MIR172 regulates flowering time and flower formation by regulating the expression of AP2-like transcription factors (Wu et al., 2009; Zhu et al., 2011). MIR172 could be a useful tool to modulate flowering time in *Arundo*, similar to what observed in *Sorghum* (Calviño et al., 2011). MIR159 is a highly conserved miRNA family, also found in all *Arundo* genus, and it plays important role in flowering time control (Li et al., 2013). MIR167 and MIR162 are two moderately conserved miRNAs and were also expressed across *Arundo* genus. MIR167 is important in gene expression and involed in regulating reproduction, while MIR162 is regulates the dicer-like gene family, and both of these two miRNA families are associated with drought stress response (Wu et al., 2006; Barrera-Figueroa et al., 2011). MIR530 was found to be expressed only in *Arundo donax* compared to other *Arundo* speices, which was involved in stress condition in the rice inflorescences. This miRNA may be used to control leaf angles and leaf size, for increasing planting density and improving the production of biomass *Arundo donax* (Barrera-Figueroa et al., 2012).

## 2.6 Conclusion

In this study, a computational approach was utilized for the identification of miRNAs and corresponding targets from the leaf transcriptomic data of seven *Arundo* species. Meanwhile, miRNA target transcripts prediction showed that these target genes are mostly transcription factors involved in plant development. These findings improved our understanding of miRNAs in *Arundo* genus. The functional annotation of predicted miRNA targets may help to understand the mechanism of response to different environmental stresses, and will provide useful information for further functional analyses and validation in future experimental studies.

# CHAPTER 3

Phylogeny and adaptive trait evolution in the *Arundo* genus (Poaceae)

## 3.1 Abstract

*A. donax* L., also called "giant reed", is a perennial $C_3$ grass with fast growth and a high potential for the production of biomass in the Mediterranea area. This is as a result of its adaptability and resistance to most pests and pathogens, and ability to grow without significant P or N fertilization. Despite its relevance both as a model species and as a crop for bioenergy production, there are still not sufficient transcriptomic or genomic data available in public databases for understanding of the precise evolutionary relationships among the members of the *Arundo* genus (up to five species of *Arundo* may exist) and the evolutionary origins of *A. donax* are still debated. In this study, the RNA sequencing method (RNA-Seq) was used to *de novo* assemble the leaf transcriptomes of seven taxa/accessions from the *Arundo* genus and closely related Arundinoideae outgroups. Assembly yielded a total of 1,016,877 unigenes with average length ranging from 741 to 1065 bp. Functional annotation of unigenes was carried out by mapping to non-redundant protein database and Gene Ontology annotation. Phylogenomic reconstruction of the relationships among *Arundo* species with a total of 150 one-to-one orthologous groups were identified, showing that *A. formosana* is sister to the other members of the *Arundo* genus, and the sister group to *A. plinii* was *A. collina* as well as the sister group to *A. donax* was *A. macrophylla* with strong statistical support values. Probabilistic models analyses demonstrated that the ancestral haploid chromosome number of *Arundo* was thirty-six (likelihood and Bayesian framework), suggesting that demi-duplication is the mainly driving force for the chromosome evolution in the *Arundo* genus. Molecular evolution analysis identified some genes under positive selection, and functional annotation identified various biological processes these genes might be involved in. In summary, the study elucidated for the first time the precise evolutionary relationships within the *Arundo* genus and provided valuable insights into adaptive selection of the *Arundo* genus at the sequence level. These results will be valuable for future gene functional validation for improvement of the biomass species *A. donax*.

## 3.2 Introduction

Transcriptomics developed fast with the development of next-generation sequencing technologies, and RNA-seq has greatly advanced the understanding of the gene expression and regulation in model or non-model organisms (McGettigan, 2013). Transcriptome sequencing is an efficient and cost-effective way to produce large amounts of RNA transcripts data used for reconstructing phylogeny and gene discovery in eukaryotes, and it is particularly widely used in some non-model organisms (Lemmon and Lemmon, 2013). In recent studies, more and more researchers used transcriptome data for resolving the evolutionary relationships among different plants lineages, such as reconstructing the phylogeny and define the origin of land plants (Timme et al., 2012; Wickett et al., 2014), or resolving evolutionary relationships in crop families (e.g. Wen et al., 2013). In addition, comparative transcriptomics also has been used in many other areas of systematic biology (Wen et al., 2015). Meanwhile, comparative transcriptomics provide a new insight into horizontal gene transfer (HGT), which is an important way for microorganisms evolution (Zhang et al., 2014). Phylogenomics is the study of evolutionary relationships by analyzing comparative genome-scale or transcriptome data, which is important for understanding the diverse biological characteristics, for example, the origin and evolution of species, gene

function analysis, and reconstructing life tree. Phylogenomics reconstruction is based on large homologous data, and it can utilize comparative gene-by-gene data (Chan et al., 2009; Puigbo et al., 2010), concatenated gene sequences (Burki et al., 2012; Yutin et al., 2012) or whole-genome sequencing data (Rannala et al., 2008). Given the relative ease to obtain them, it is popular to use transcriptome data for phylogenomic analyses in cases where genome data are not available (Oliveira et al., 2012).

Polyploidy and chromosome number changes are considered as important features and used for investigating speciation, and many crops are relatively recent polyploids (Renny-Byfield & Wendel, 2014). Chromosome numbers evolution and polyploidization has attracted great attention in plants (Soltis et al., 2009; Cui et al., 2006). There are differences between polyploids and their ancestors in morphological, physiological, and life evolutionary history (Ramsey and Schemske, 2002), and these differences may be the cause of polyploid species' ability to adapt to new life environments. It is thus hypothesized that polyploidy in the *Arundo* genus may also have played an important role for adaptation to harsh environments. Recent tools have been developed that can help to reconstruct models of chromosome numbers evolution and polyploidization with rigorous statistical evaluation based on phylogenetic trees like the software ChromEvol (Glick et al., 2014). On the other hand, traits evolution is important for various morphological and physiological adaptations in plants. This is why identification of candidate genes with positive selection has been considered a main goal in evolutionary biology research. Some previous studies have reported positive selection in orthologous datasets in plants (e.g. Buschiazzo et al., 2012), but to date no such approach has been carried out in the *Arundo* genus.

Dwindling fossil energy reserves and global warming change urgently require the development of environmental-friendly renewable sources of energy (Ohlrogge et al., 2009). Biomass can mitigate the dependence from petrol and coal by providing a nearly carbon-neutral source of energy. In order to prevent the competition for land between food/feed and bioenergy crops, much attention has been recently devoted to plant species able to produce large amounts of biomass from marginal soils not suitable for agriculture. Perennial grasses has already been used as resource of bioenergy with many advantages in the Mediterranean and USA (Lewandowski, 2003). The *Arundo* genus and related Arundinoideae taxa is a clade of perennial grasses belong to the Poaceae family that has received extensive attention from researchers. Four perennial grasses have been proposed as the most promising bioenergy crops, namely, *A. donax* and *Phalaris arundinacea* ($C_3$ crops), *Miscanthus* and *switchgrass* ($C_4$ grasses; Lewandowski, 2003). Especially *A. donax L.*, also called "giant reed", is a perennial $C_3$ grass with fast growth (Rossa et al., 1998). It requires little management input and is resistant to most pests and pathogens. It can also grow without significant P or N fertilization, all features that allowed *A. donax* to become one of the most invasive riparian species in Southern USA (especially California). At the same time, this robustness makes it a very productive biomass species: in optimal conditions *A. donax* fields become productive after the second year and can provide dry biomass yields up to 40 tons/hectar/year, thus constituting one of the most promising species for the production of biomass in the Mediterranea area (Angelini et al., 2009). Previous genetic studies indicated that *A. donax* originated in Eastern Asia (Pilu et al., 2012), from where it spread, possibly by human intervention, into the Middle East and the Mediterranean. More recently, it was also introduced in Africa and even Australia. In the Arundineae tribe species of the *Arundo* genus are known to the non specialists because they encompass large and gorgeous perennial plants. Recent revisions of the genus indicated that up to five species of *Arundo* exist, but there is no whole genome sequences or transcriptome data, so their

precise phylogenetic relationships and evolutionary history are still debated (Hardion et al., 2012).

In this study, leveraging on the availability of the leaf transcriptomes for seven *Arundo* taxa/accessions plus three closely related outgroups from the Arundinoideae tribe of Poacea by an Illumina sequencing platform, the study carried out through phylogenomics the reconstruction of the relationships among *Arundo* species, and investigated chromosome number changes within the *Arundo* genus. The dissection of the patterns of evolution in this genus will support ongoing efforts to establish reverse genetics and functional genomics approaches in *A. donax*, thus contributing to provide promising cadidate genes for the improvement of this biomass species.

## 3.3 Materials and Methods

### 3.3.1 Transcriptome De novo Assembly

The plant material was collected from 3 individuals from all *Arundo* species and three closely related outgroups, grown in a commen garden experiment at the Edmund Mach Foundation. Leaf tissues were sampled from the first and second fully developed leaves of each individual, in order to standardize as much as possible transcript sampling, and have been sequenced with 100bp pair end Illumina reads on a HiSeq2000 sequencer. Quality assessments were conducted on these raw reads by FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). The Bowtie index was constructed with the bowtie-build utility from Bowtie v1.0.0 (Langmead et al., 2009) using the *Enterobacteria phage* phiX174 fasta file as reference. Adapters were cleaned by mapping of the reads to the phiX genome. Sequence adapters and low-quality regions reads were trimmed by Trimmomatic v 0.32 (Bolger et al., 2014), the reads length ≥100 bp were retained. Properly paired reads were found with the fastqutils properpairs from NGSUtils v 0.5.9 (Breese and Liu, 2013). Transcripts were assembled *de novo* using the Trinity Assembler (Grabherr et al., 2011) with the default settings. To lower the redundancy in the dataset, the transcipts were discarded using the CD-HIT-EST program (Li and Godzik, 2006) with an identity threshold of 95% and a word size of 10. The CAP3 (Huang and Madan, 1999) tool was used to generate unigenes for each cluster with default options and an identity of 99%.

### 3.3.2 Gene functional annotation

Assembled unigenes were annotated by performing a local BLASTx search against TAIR10 (https://www.arabidopsis.org/) and *Setaria italica* (http://www.uniprot.org/) databases with an E-value cutoff of 1e-10. Gene Ontology annotation was carried out using the web-based program FunctionAnnotator analyses (http://fa.cgu.edu.tw/index.php) (Chen et al., 2012) included four main parts (Homology searching against nr database; GO term assignment; Identification of domain and enzymes). Finally, transcripts annotation obtained from sequence homology search by LAST against NCBI-nr database and gene ontology were classified into molecular function, biological process, and cellular component.

### 3.3.3 Orthologous Groups Identification and supermatrix construction

The TransDecoder v2.0.1 (Haas et al., 2013) program was used to identify the candidate open reading frames (ORFs) within the transcripts with default parameters, and sequences with a protein length greater than 100 were retained for further analysis. The obtained ORFs were used for local BLAST (blastp) search against the protein database of

*Setaria italica* with a cutoff e-value of 1e-5, and all putative results were retained as coding regions by running the TransDecoder program. Translated sequences were clustered into multiple orthologous groups among *A. collina*, *A. donax*, *A. donaciformis*, *A. formosana*, *A. macrophylla*, *A. micrantha*, *A. plinii*, *Hakonechloa macra*, *Molinia caerulea* and *Phragmites australis* using OrthoFinder (Emms and Kelly, 2015) with default parameters, we retained only single-copy orthologs were identified by OrthoFinder for further analysis.

The phylogenetic supermatrices were reconstructed based on putative 150 one-to-one orthologous groups (OGs) were extracted from retained single-copy orthologs among 10 species. Each corresponding orthologous group of CDS was extracted with custom Python scripts via 'Gene ID' from CDS datasets predicted by TransDecoder. Multiple sequence alignments were performed for each orthologous group using MUSCLE v3.8.31(Edgar, 2004) with default parameters. The program Gblock v 0.9.1 (Castresana, 2000) was used to trim alignments with default parameters except for the 'with-half' gap positions option and multiple-alignments matrix were concatenated by the program Phyutility v 2.2.6 (Smith and Dunn, 2008), the steps of these independent scripts analyses have been implemented into Python packages by Ya Yang and Stephen A. Smith (Yang and Smith, 2014).

### 3.3.4 Phylogenomic reconstruction of *Arundo* species

Phylogenomic reconstruction were conducted based on constructed matrix (Set *H. macra*, *M. caerulea* and *P. australis* as outgroup) by Maximum-likelyhood (ML), Bayesian inference (BI) and Maximum parsimony (MP), respectively. Maximum-likelihood tree searches were conducted with RAxML v7.2.8 (Stamatakis, 2006), bootstrap resampling was performed for 200 replicates using a rapid bootstrapping algorithm (STAMATAKIS, 2008), The best-fitting model of protein evolution inferred by ProtTest 3.2 (Darriba et al., 2011) was JTT+G+F according to the Akaike information criterion (AIC) and Bayesian Information Criterion (BIC). Bayesian inference was carried out using The MPI version of the software MrBayes v3.2.6 (Ronquist and Huelsenbeck, 2003) under mixed model. Twenty parallel runs were performed with 1,000,000 generations, sampled every 100 generations, and the first 25% of trees from all runs were discarded as burn-in and excluded from the analysis, and the remaining trees were used to construct the consensus tree to represent posterior probabilities for each node. Maximum parsimony tree searches were conducted by PAUP*4.0 (Swofford, 2002). Heuristic tree searches were conducted using random taxon-addition with branch swapping tree-bisection-reconnection (TBR), non-parametric bootstrap analysis was performed by 1000 replicates with TBR branch-swapping. Maxtrees was set to 100,000 and then auto-increased by 100 until the searches were completed. The constructed 50% major-rule consensus MP tree are shown.

### 3.3.5 Inference of chromosome-number change

To infer chromosome numbers evolution and determine ploidy levels (diploid or polyploid) for *Arundo* taxa, the program ChromEvol v2.0 (Mayrose et al., 2010; Glick and Mayrose, 2014) was used with 100 randomly sampled MrBayes trees combined with chromosome median numbers for each *Arundo* species . The reliability of estimated ploidy levels were carried out by comparing ploidy inferences across phylogenies and by using a simulation-based approach among ten models. Species tree was reconstructed by using MrBayes v3.2.6 with mixed models, based on the constructed 150 one-to-one OGs supermatrices among 8 taxa after removing *A. collina* and *A. macrophylla*, which were the sister groups to *A. plinii* and *A. donax*, respectively. Chromosome median number counts for *A. donax*, *A. formosana*, *A. micrantha*, *A. plinii*, *H. macra*, *M. caerulea* and *P. australis* were obtained from the Chromosome Counts Database (CCDB, version 1.45) of plant

chromosome numbers (http://ccdb.tau.ac.il/home/) (Rice et al., 2015), while *A. donaciformis* was obtained from literature (Hardion et al., 2012). 10 models were used to assess the best-fitting model by Akaike information criterion (AIC) with a guide tree derived from the Bayesian consensus tree. The results obtained with the best-fitting model are shown and whole genome duplications, demi-duplications, base-chromosome number, and individual chromosome losses or gains along each branch of the phylogeny are reported.

### 3.3.6 Analysis of molecular evolution

Molecular evolution analyses were performed by HyPhy package (Pond et al., 2005). MACSE (Ranwez et al., 2011) was used for generating multiple sequence alignments of each CDS orthologous group. After removing stop codons, orthologs were assessed for recombination with the GARD method (Pond et al., 2006), which is implemented in HyPhy package and can find all the recombination breakpoints, fitting various model using the Akaike Information Criterion (AIC) and Shimodaira-Hasegawa test (SH test) for phylogenetic incongruence. The rate variation was implemented by a general discrete distribution algorithm. Evidence of episodic positive selection was verified by BUSTED (Murrell et al., 2015) (implemented in HyPhy), which uses the branch-site unrestricted statistical test for episodic diversification by estimating the proportion of selected codons across all branches of the phylogenetic tree. A false-discovery rate (at 5% FDR) correction was applied to reduce false positives. Detecting episodic diversifying selection of the specific branches in the phylogenetic trees by the adaptive branch-site random effects likelihood (aBSREL) (Pond et al., 2008) model (implemented in HyPhy), using the universal genetic code. We used the RELAX program (a general hypothesis testing framework ) (Wertheim et al., 2014) implemented in HyPhy for detecting relaxed selection. The selection intensity parameter (k) was introduced in the RELAX program to infer the relationship of ω (dN/dS) in the test and the reference backgroud branches. Site test for episodic diversification by a mixed effects model of evolution (MEME; P-value, <0.05) (Murrell et al., 2012) and Fast Unconstrained Bayesian Approximation (FUBAR; posterior probability, >=0.9) (Murrell et al., 2013), which are implemented in HyPhy, were used to identify episodic diversifying selection at the level of individual sites with universal genetic code.

Gene Ontology annotation for episodic positive selection genes was carried out using web-based program FunctionAnnotator (Chen et al., 2012), and the graphic of GO functional classifications were shown by WEGO (http://wego.genomics.org.cn/cgi-bin/wego/index.pl) (Ye et al., 2006). Further functional annotation of the genes under positive selection was carried out by performing local Blast searches against the *Arabidopsis thaliana* protein database (https://www.arabidopsis.org/).

## 3.4 Results

### 3.4.1 De novo assembly of Illumina reads

Total RNA extracted from leaves of all known *Arundo* species and three closely related outgroups have been sequenced with 100 bp pair end Illumina reads on a HiSeq2000 sequencer. After quality assessment and data filtering, retained sequence reads were used for the De novo assembly of full-length leaf transcripts by the Trinity program. Ultimately, *de novo* assembly yielded 1,016,877 unigenes with average length ranging from 741 to 1065 bp among *Arundo* genus and three outgroups species (Table

3.1). The majority of the transcripts lengths were distributed in the range of more than 1000 bp, which accounted for 351,767 unigenes (34.6%), a total of 39,015 unigenes (3.8%) had length distribution ranging from 800 to 900 bp and 35,371 unigenes (3.5%) had a length range of 900 to 1000bp are showed in Supplementary Figure 3.1.

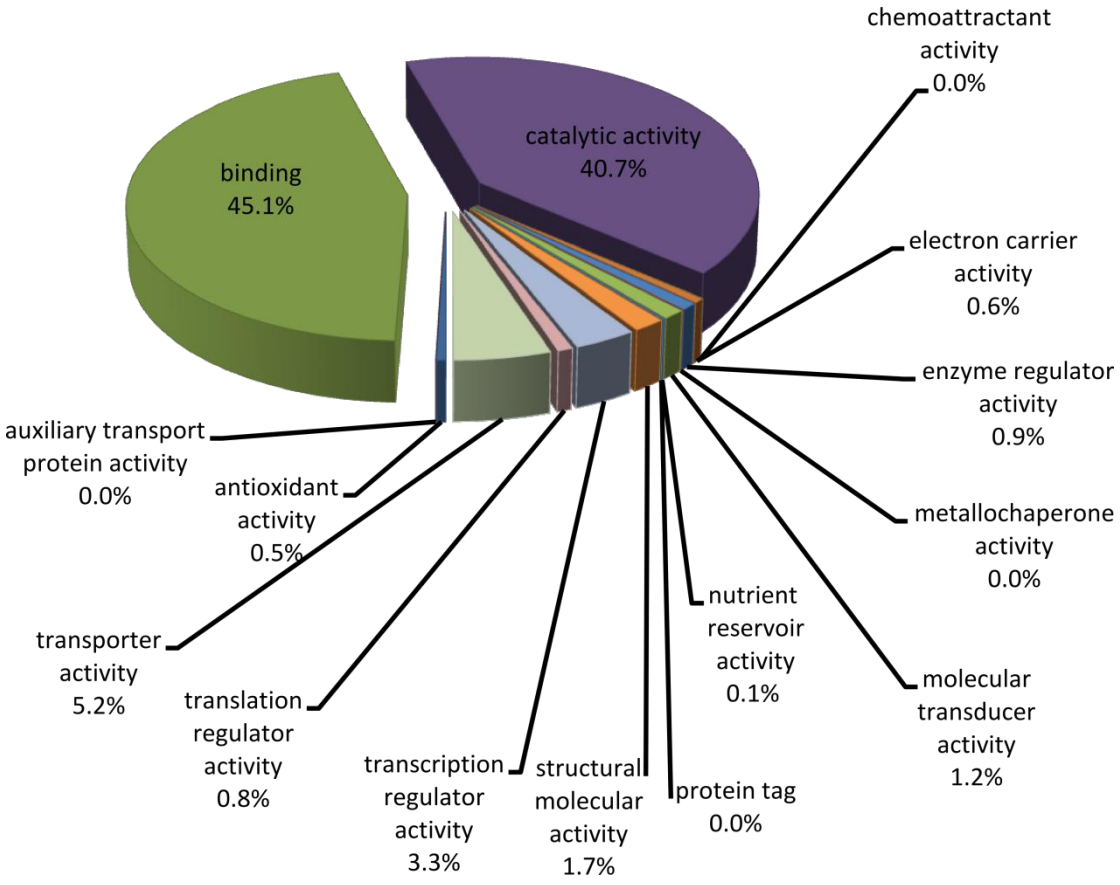
**Table 3.1.** Summary statistics of unigenes assembled by Trinity.

| Species | Number of Transcripts | Min_length (bp) | Max_length (bp) | Mean_length (bp) |
|---|---|---|---|---|
| *Arundo donax* | 76591 | 201 | 12981 | 921 |
| *Arundo macrophylla* | 104919 | 201 | 12989 | 901 |
| *Arundo formosana* | 60798 | 201 | 11616 | 959 |
| *Arundo donaciformis* | 115911 | 201 | 15201 | 984 |
| *Arundo micrantha* | 88524 | 201 | 15072 | 741 |
| *Arundo plinii* | 169584 | 201 | 15976 | 1065 |
| *Arundo collina* | 124291 | 201 | 13402 | 1010 |
| *Hakonechloa macra* | 52378 | 201 | 10952 | 901 |
| *Molinia caerulea* | 125280 | 201 | 15192 | 1009 |
| *Phragmites australis* | 98601 | 201 | 15198 | 1060 |

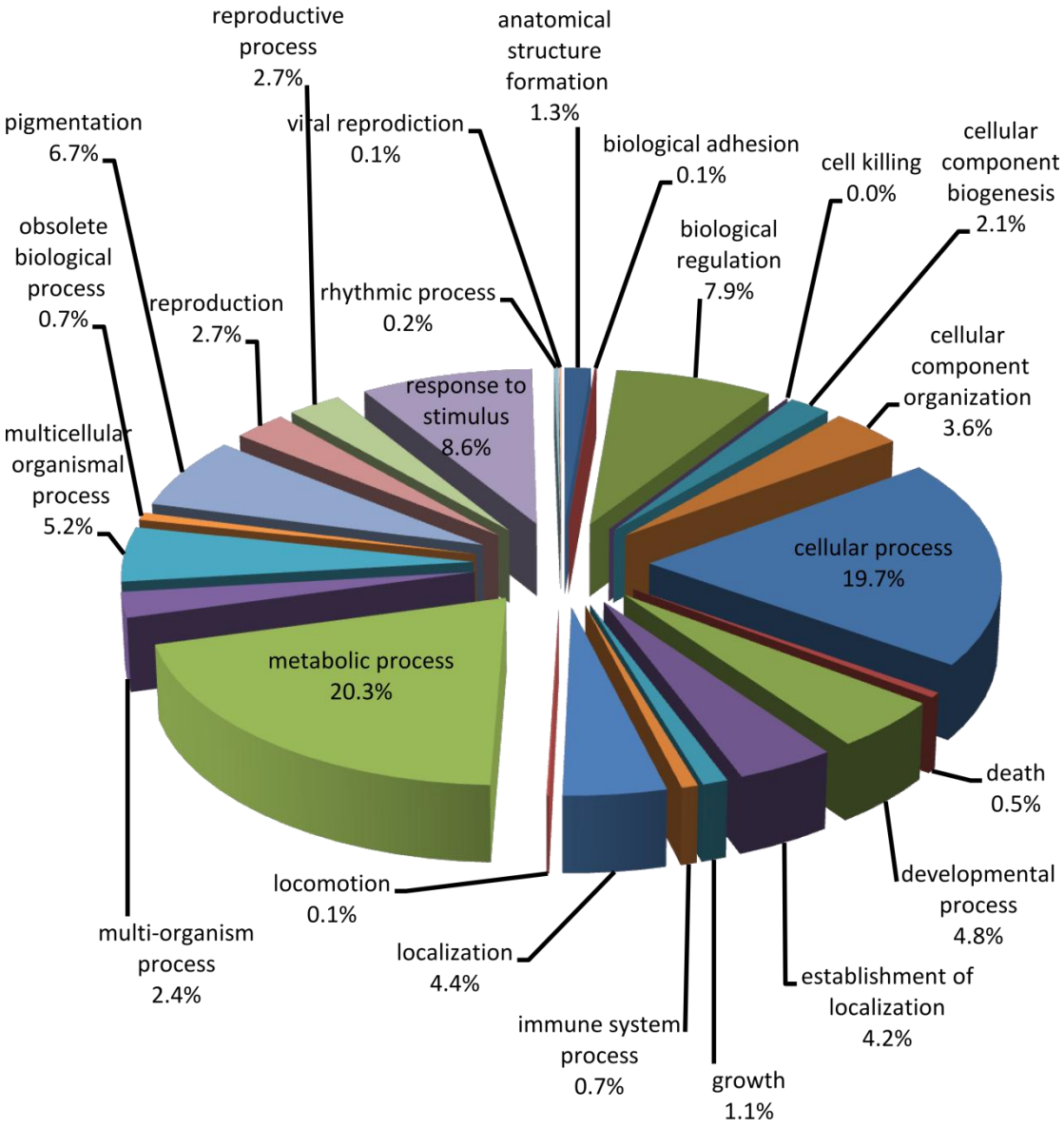
### 3.4.2 Functional annotation

For validation and annotation of assembled unigenes, sequence similarity search was carried out against the *Arabidopsis thaliana* and *Setaria italica* protein datasets using BLASTx algorithm and search against the NCBI non-redundant protein database (Nr) with FunctionAnnotator. This resulted in an average of 12.4%-24.0% , 17.7%-35.2% and 58.6%-74.7% unigenes with significant hits to these three databases, respectively. Gene ontology (GO) annotation assigned putative functions of transcripts into three categories: Biological process, Molecular function, and Cellular component as shown in Figure 3.1. The assigned functions of unigenes covered a broad range of GO categories corresponding to 671,330 transcripts (66.02%). Under the molecular function category 45.1% of the unigenes were associated with binding and 40.7% of the unigenes with catalytic activities represented the majorities of the category. Among the binding and catalytic activities part, ATP binding represented the most abundant classification, followed by ion binding, DNA binding and protein serine/threonine kinase activity. The biological process category showed that 20.3% of the genes were associated with metabolic process, 19.7% with cellular process, and 8.6% with response to stimulus. The cellular component category showed that genes associated with the cell and cell part were 27.1% and 27.1%, respectively, 22.9% were associated with organelle and 9.0% were associated with organelle part. The functional annotation of the unigenes also was performed by sequence similarity searches against the *A. thaliana* and *S. italica* proteins using the BLASTx algorithm. A total of 172,745 (16.99 %) unigenes functional hits correspond to *A. thaliana,* and a total of 256,783 (25.25%) unigenes functional hits correspond to *S. italica*.

**A**


chemoattractant activity 0.0%
catalytic activity 40.7%
binding 45.1%
electron carrier activity 0.6%
enzyme regulator activity 0.9%
auxiliary transport protein activity 0.0%
antioxidant activity 0.5%
metallochaperone activity 0.0%
nutrient reservoir activity 0.1%
transporter activity 5.2%
molecular transducer activity 1.2%
translation regulator activity 0.8%
transcription regulator activity 3.3%
structural molecular activity 1.7%
protein tag 0.0%

**B**


reproductive process 2.7%
anatomical structure formation 1.3%
pigmentation 6.7%
viral reprodiction 0.1%
biological adhesion 0.1%
cell killing 0.0%
cellular component biogenesis 2.1%
obsolete biological process 0.7%
biological regulation 7.9%
reproduction 2.7%
rhythmic process 0.2%
cellular component organization 3.6%
response to stimulus 8.6%
multicellular organismal process 5.2%
cellular process 19.7%
metabolic process 20.3%
death 0.5%
developmental process 4.8%
locomotion 0.1%
multi-organism process 2.4%
localization 4.4%
establishment of localization 4.2%
immune system process 0.7%
growth 1.1%

C

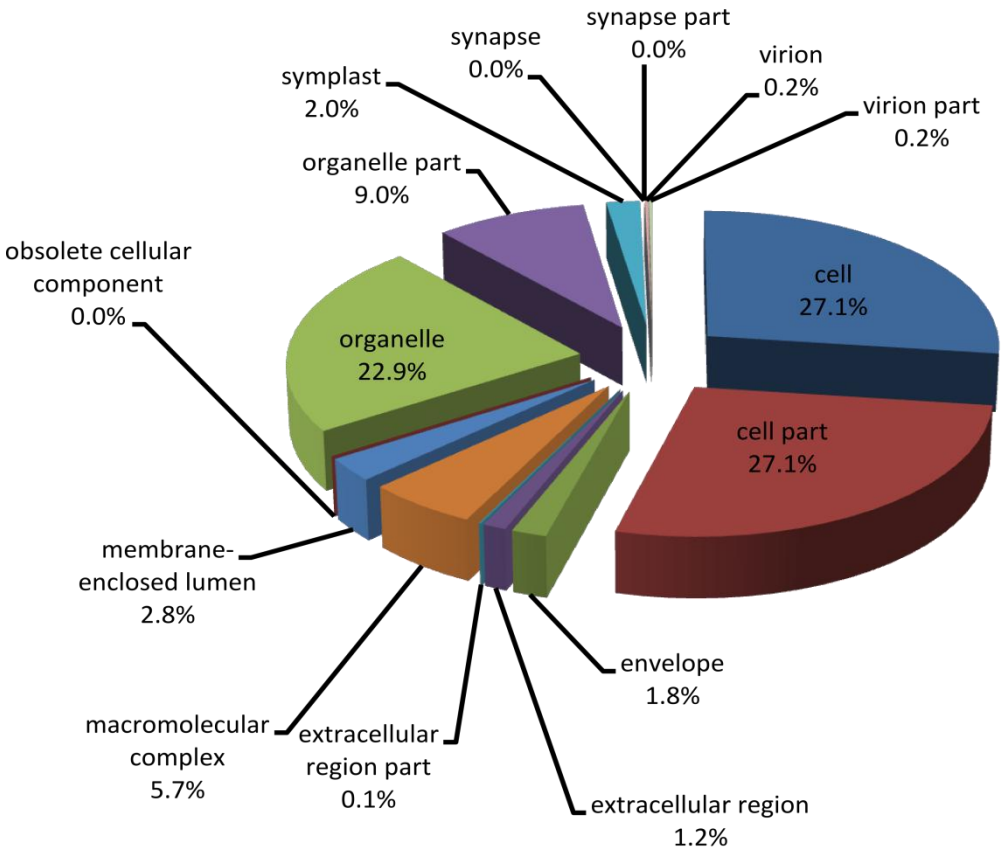



**Figure 3.1.** Gene Ontology (GO) assignment (2nd level GO terms) of the unigenes among 10 species.(A) Molecular functions; (B) Biological processes; (C) Cellular components. Numbers shown the percentage of each sub-category.

### 3.4.3 Phylogenomic reconstruction of *Arundo* species

A total of 150 pairs of unigenes that were putative one-to-one orthologs among 10 species, containing 747,841 characters and 77,955 aligned columns were concatenated for phylogenomic reconstruction of *Arundo* species. The species tree was reconstructed with high support value (almost all internal nodes having more than 90% bootstrap values and 1.0 posterior probabilities), and *M. caerulea*, *P. australis* and *H. macra* set as outgroup with three different reconstruction algorithms: Maximum Likelihood (ML), Maximum Parsimony (MP), Bayesian inference (BI). The resulting trees indicated *A. formosana* is sister to the other members of the *Arundo* genus with high support (bootstrap:100/posterior probability: 1.0). The sister group to *A. plinii* was *A. collina* as well as the sister group to *A. donax* was *A. macrophylla* with strong supporting values (bootstrap:100/posterior probability: 1.0), and *A. donaciformis* has a close evolutionary relationship to *A. plinii* and *A. collina*. *M. caerulea* and *P. australis* are sister groups (Figure 3.2). The least supported group (90% bootstrap support in ML reconstruction)

of the whole phylogeny is the clade constituted by *A. micrantha*, *A. donaciformis*, *A. plinii* and *A. collina*. However, both MP boostrap and BI posterior probability provide full support to the clade, confirming the basal position of *A. micrantha* as compared to the other taxa.

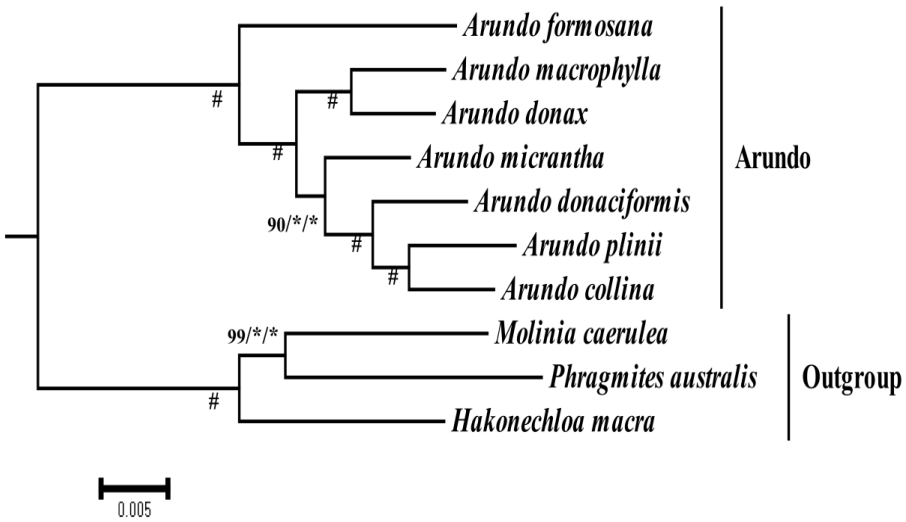



**Figure 3.2.** Phylogenetic relationship of *Arundo* genus. Phylogenomic trees were inferred by the concatenation analyses, PAUP, RAxML and MrBayes. ‘‘*’’ indicates support values of posterior probabilities(PP) = 1.0 and bootstrap (BP) = 100. ‘‘#’’ indicates all support values of PP = 1.0 and BP = 100. Support values are shown for nodes as maximum likelihood bootstrap/maximum parsimony bootstrap/ Bayesian inference posterior probability.

### 3.4.4 Chromosome evolution

Chromosome median numbers for *Arundo* taxa and three outgroups were previously reported: n = 18 for *M. caerulea*, n = 24 for *P. australis* and *H. macra*, n = 54 for *A. donax* and *A. donaciformis*, n = 36 for *A. formosana*, *A. micrantha* and *A. plinii*. The species tree was reconstructed using MrBayes, based on the constructed 150 one-to-one OGs concatenated data among 8 taxa after removing *A. collina* and *A. macrophylla*, which are likely synonyms of *A. plinii* and *A. donax*, the result was consistent with high support (posterior probability: 1.0) in each branch to the phylogenetic tree reconstructed with all taxa showed in Figure 3.2 (Supplementary Figure 3.2). Based on the phylogenetic relationships within *Arundo* resulting from MrBayes analysis, ChromEvol analyses suggested that the best fitting model of the process of chromosome evolution in *Arundo* was the hypothesis with constant gain, loss, no duplication and estimation of demi-polyploidizations (Table 3.2).

**Table 3.2.** Likelihood and AIC scores reckon for the data set analyzed for each model carried out by ChromEvol software.

| Models | Log-likelihood | AIC scores |
|---|---|---|
| **CONST_RATE_DEMI_EST*** | **-19.1553** | **46.3106** |
| CONST_RATE_DEMI | -22.2378 | 50.4756 |
| CONST_RATE | -24.1571 | 54.3141 |
| BASE_NUM | -23.7997 | 55.5995 |
| LINEAR_RATE_DEMI | -22.9247 | 55.8494 |
| LINEAR_RATE | -23.1375 | 56.275 |
| LINEAR_RATE_DEMI_EST | -22.3977 | 56.7955 |
| BASE_NUM_DUPL | -23.7997 | 57.5994 |
| CONST_RATE_NO_DUPL | -28.5152 | 61.0303 |
| LINEAR_RATE_NO_DUPL | -27.1864 | 62.3728 |

*Best fitting model

The evolutionary history for chromosome number changes inferred by ChromEvol for *Arundo* genus is shown in Figure 3.3. The ancestral chromosome numbers were consistent in each branch estimated with ML and Bayesian analyses. The haploid number inferred for the ancestor of *Arundo* species was n=36 (posterior probability (pp) = 0.76) and the most likely haploid number with the ML analysis also was n = 36. From the most recent common ancestor, haploid number n = 36 (pp = 0.95) increased to n = 54 in *A. donax* by demi-duplication event. The haploid chromosome number inferred for the branch leading to *A. micrantha, A.* plinii and *A. donaciformis* was n = 36 (pp = 0.98), while in the case of *A. plinii* and *A. donaciformis* branch the haploid chromosome number was n = 36 (pp = 0.99) and increased to n = 54 in *A. donaciformis* by demi-duplication event. The inferred ploidy level of *Arundo* taxa showed that all of them are polyploid species, as all *Arundo* genus showed simulation reliability pp larger than 0.95 (Supplementary Table 3.1). Under the most recent common ancestor haploid number n = 36, demi-duplication clearly results to have been the main driving force in chromosome evolution for the *Arundo* genus.

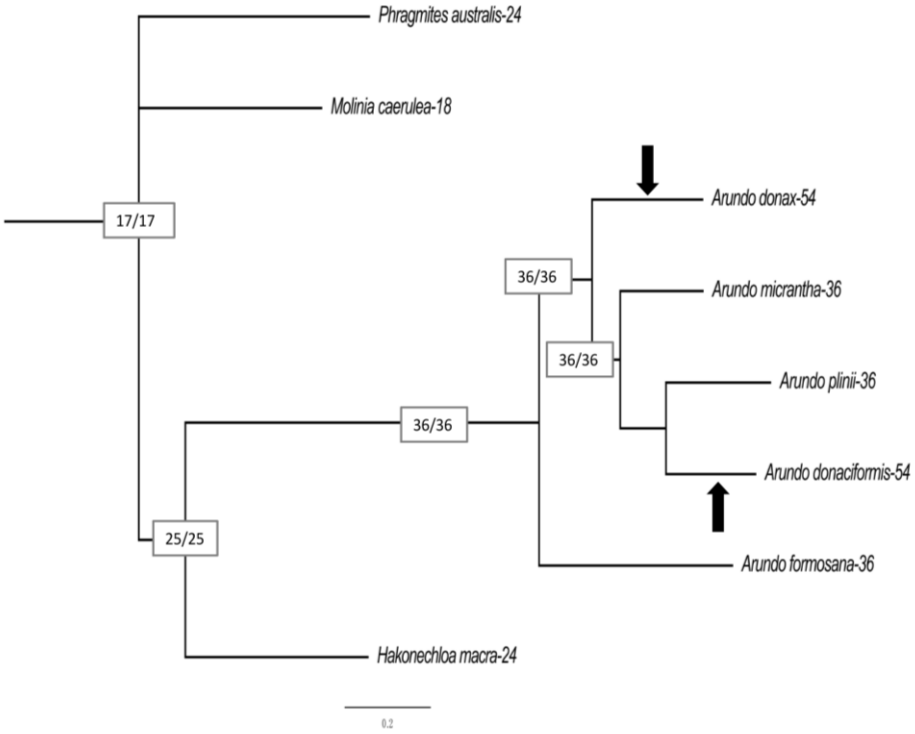



**Figure 3.3**. Chromosome number evolution and inferred ancestral chromosome state in the genus *Arundo* (black boxes) inferred under Maximum likelihood and Bayesian optimization, including other three outgroups. Boxes at the nodes present the inferred ancestral haploid chromosome number for each node by ML and Bayesian analysis, respectively. Numbers at the tips are the known haploid chromosome numbers of species. The arrow indicates that the demi-duplication event occurs in the branch.

### 3.4.5 Evolutionary analysis

GARD analyses, implemented in the HyPhy batch language, were used to detect the best-fit number and location for recombination breakpoints. In total, recombination breakpoints were identified among 106 orthologous gene sets, and fourteen orthologous gene sets showed significant evidence for recombination breakpoints (SH $p < 0.01$) using GARDProcessor (Table 3.3). As recombination can mislead phylogenetic analysis, To reduce noise the study considered all recombination breakpoints identified by GARD for sites positive selection analyses, and the results produced by GARD as input files used for MEME and FUBAR algorithm.

**Table 3.3.** Orthologous genes with significance evidence of recombination breakpoints.

| OGs_ID | *S* | Nucleotides / Diversity | | Recombination Fragments (bp) | Mean Tree Splits | Significance | |
|---|---|---|---|---|---|---|---|
| | | Variable Sites | (%) | | Identity % | Delta AICc | SH P<=0.01 |
| OG0018374 | 10 | 2982/374 | 6.13 | 96,1130,85,138,252,1281 | 4.67 | 468.09 | 3/5 |
| OG0018272 | 10 | 1680/143 | 4.22 | 192,350,304,265,332,000 | 14 | 151.03 | 1/5 |
| OG0018371 | 10 | 1344/167 | 7.39 | 315,105,432,215,172,000 | 9.33 | 240.75 | 1/5 |
| OG0017423 | 10 | 2148/234 | 7.99 | 219,858,129,527,415 | 13.05 | 188.84 | 1/4 |
| OG0018296 | 10 | 1749/179 | 4.59 | 195,300,861,132,261 | 22.67 | 204.26 | 1/4 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| OG0018353 | 10 | 1692/190 | 5.45 | 55,373,686,433,145 | 10.83 | 347.48 | 1/4 |
| OG0018354 | 10 | 2706/354 | 6.59 | 579,87,390,264,1386 | 12.67 | 269.37 | 1/4 |
| OG0018254 | 10 | 1863/317 | 7.14 | 156,526,671,510 | 15.56 | 319.15 | 1/3 |
| OG0018205 | 10 | 972/142 | 6.6 | 257,69,487,159 | 9.72 | 165 | 1/3 |
| OG0018112 | 10 | 1242/125 | 5.43 | 181,896,165 | 17.78 | 92 | 1/2 |
| OG0018250 | 10 | 1425/129 | 3.97 | 678,656,91 | 24.44 | 128 | 1/2 |
| OG0017710 | 7 | 369/44 | 5.07 | 330,39 | 0 | 24.47 | 1/1 |
| OG0017727 | 10 | 405/43 | 4.2 | 207,198 | 20 | 22.01 | 1/1 |
| OG0017994 | 9 | 465/74 | 7.65 | 38,427 | 0 | 34.07 | 1/1 |

Notes: A summary of output from GARD analysis to detect recombination breakpoints, including the recombination fragments, and the P-value from the Shimodaira and Hasegawa test (SH P<0.01 : proportion of break points found significant by the SH test on flanking trees).

In the present study, a total of 28 genes with evidence of positive selection (*p*-value < 0.05) were identified by Adaptive Branch-site REL method, among them, 13 genes were still significant after multiple test correction (at 5% FDR; Supplementary Table 3.2), GO annotation showed that these positive genes were mainly involved in metabolic process, function in binding and catalytic activity (Figure 3.5). A total of 4 genes showed significant evidence for relaxed selection along the test branches ( $k < 1$), while 15 genes showed significant evidence for intensified selection along the test branches ( $k > 1$, *p-value* < 0.05) present in Figure 3.4 (Supplementary Table 3.3; Supplementary Table 3.4).

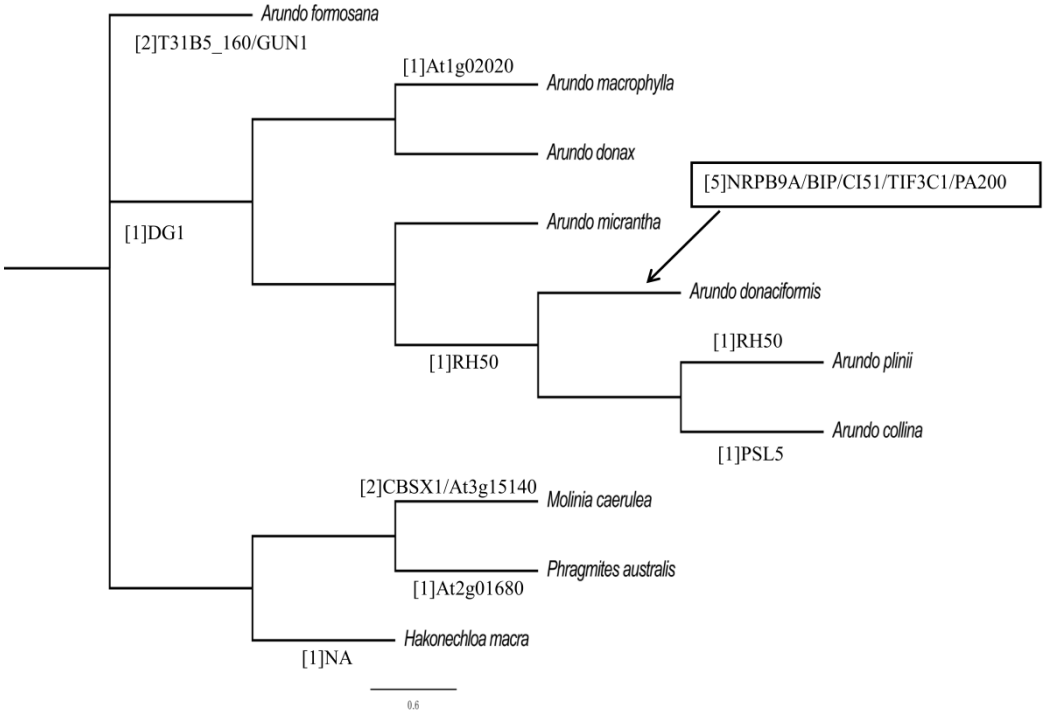



**Figure 3.4.** aBS-REL analysis of gene under positive selection. Branches where annotated with genes that were confirmed to be under episodic positive selection using the adaptive branchsite models. In square brackets the number of genes identified for each branch and their names are reported. NA-Uncharacterized Protein.

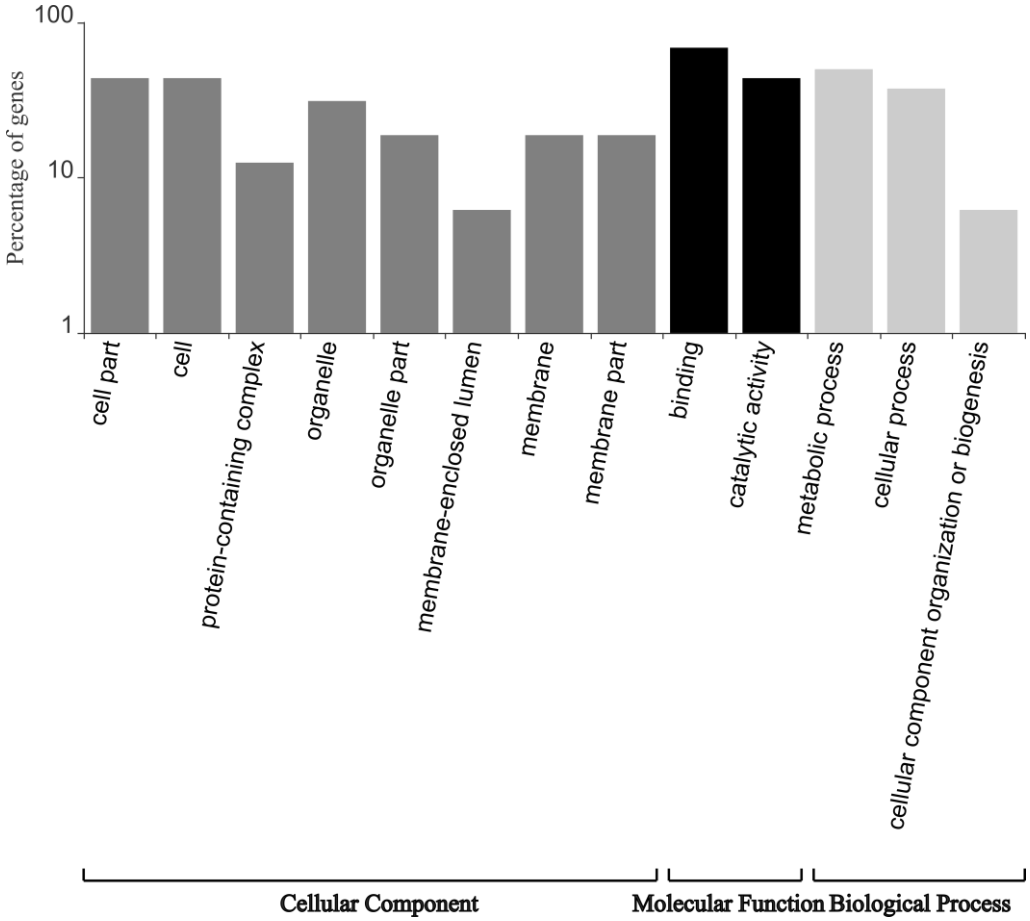



**Figure 3.5.** GO classification for 28 positive selection genes identified by aBSREL in *Arundo* speices.

A total of 34 genes with evidence of positive selection ($p < 0.05$) were identified by an alignment-wide approach using BUSTED. Functional annotation for a total of 34 positive selection genes by BLAST similar search against *Arabidopsis thaliana* protein database (Table 3.4), GO annotation showed that these positive genes were mainly involved in cellular process, metabolic process, response to stimulus and biological regulation (Figure 3.6). Out of the 28 genes identified by Branch-Site REL, 23 genes were also significant for positive selection using BUSTED, while the 11 additional genes were identified only by BUSTED with episodic positive selection. After correction for multiple testing ( at 5% FDR) a total of 25 genes were still significant. Out of the 13 multiple test correction positively selected genes identified by Branch-Site REL, 11 were also significant for multiple test correction using BUSTED, and 14 additional genes were identified only by BUSTED.

**Table 3.4.** Functional annotation of 34 genes under positive selection identified by sequences alignment-wide method.

| OGs_ID | P-VALUE | FDR | GENE_NAME | TAIR_ID | DESCRIPTION |
|---|---|---|---|---|---|
| OG0017436 | 0 | 0.0013 | DG1 | AT5G67570.1 | Tetratricopeptide repeat (TPR)-like superfamily protein |
| OG0017508 | 0.0005 | 0.0057 | NA | NA | NA |
| OG0017514 | 0 | 0.0017 | At5g26280 | AT5G26280.1 | TRAF-like family protein |
| OG0017532 | 0 | 0.002 | At1g02020 | AT1G02020.1 | nitroreductase family protein |

| | | | | | |
|---|---|---|---|---|---|
| OG0017575 | 0 | 0.0003 | F13I12.50 | AT3G47000.1 | Glycosyl hydrolase family protein |
| OG0017620 | 0.0214 | 0.009 | RH58 | AT5G19210.2 | P-loop containing nucleoside t riphosphate hydrolases superfamily protein |
| OG0017657 | 0.007 | 0.0083 | NRPB9A | AT3G16980.1 | RNA polymerases M/15 Kd subunit |
| OG0017710 | 0 | 0.0027 | CBSX1 | AT4G36910.1 | Cystathionine beta-synthase (CBS) family protein |
| OG0017739 | 0.0273 | 0.0107 | MHJ24.13 | AT5G64150.1 | RNA methyltransferase family protein |
| OG0017740 | 0 | 0.0037 | LACS9 | AT1G77590.1 | long chain acyl-CoA synthetase 9 |
| OG0017807 | 0.0015 | 0.0063 | RH50 | AT3G06980.1 | Putative DEAD-box ATP-dependent RNA helicase family protein |
| OG0017835 | 0 | 0.0023 | T31B5_160 | AT5G13340.1 | Arginine/glutamate-rich 1 protein |
| OG0017940 | 0.0031 | 0.0077 | At5g36230 | AT5G36230.1 | ARM repeat superfamily protein |
| OG0017952 | 0.0004 | 0.0053 | PVA42 | AT4G21450.3 | PapD-like superfamily protein |
| OG0018016 | 0.0019 | 0.0067 | CHL | AT3G47860.1 | Chloroplast lipocalin |
| OG0018051 | 0.0069 | 0.008 | RPL1 | AT3G63490.1 | Ribosomal protein |
| OG0018069 | 0.0263 | 0.0103 | BIP | AT5G42020.1 | Heat shock protein 70 (Hsp 70) family protein |
| OG0018070 | 0.0002 | 0.005 | BIP | AT5G42020.1 | Heat shock protein 70 (Hsp 70) family protein |
| OG0018102 | 0 | 0.0047 | APS1 | AT5G48300.1 | ADP glucose pyrophosphorylase 1 |
| OG0018124 | 0.0258 | 0.0097 | CLPP4 | AT5G45390.1 | ATP-dependent Clp protease proteolytic subunit |
| OG0018133 | 0.0315 | 0.0113 | SOX | AT3G01910.1 | sulfite-oxidase |
| OG0018157 | 0.0126 | 0.0087 | At4g35140 | AT4G35140.1 | Transducin/WD40-repeat-like-superfamily-prot ein |
| OG0018254 | 0.0012 | 0.006 | GUN1 | AT2G31400.1 | genomes-uncoupled-1 |
| OG0018329 | 0.0239 | 0.0093 | At2g13440 | AT2G13440.1 | glucose-inhibited-division-family-A-protein |
| OG0018339 | 0 | 0.003 | CUL4 | AT5G46210.1 | cullin4 |
| OG0018342 | 0 | 0.0043 | TIF3A1 | AT4G11420.1 | Eukaryotic translation initiation factor 3 subunit A |
| OG0018353 | 0.0259 | 0.01 | PSL5 | AT5G63840.1 | Glycosyl-hydrolases-family-31--protein |
| OG0018354 | 0 | 0.0033 | TIF3C1 | AT3G56150.2 | eukaryotic translation initiation factor 3C |
| OG0018357 | 0 | 0.0007 | SUD1 | AT4G34100.2 | RING/U-box-superfamily-protein |
| OG0018364 | 0.0019 | 0.007 | CAMTA5 | AT4G16150.1 | calmodulin-binding;transcription-regulators |
| OG0018373 | 0.0309 | 0.011 | GLU1 | AT5G04140.1 | glutamate synthase 1 |
| OG0018377 | 0 | 0.001 | At5g47690 | AT5G47690.1 | Binding protein |
| OG0018380 | 0 | 0.004 | TRX4 | AT1G19730.1 | Thioredoxin superfamily protein |
| OG0018383 | 0.0028 | 0.0073 | At3g61320 | AT3G61320.1 | Bestrophin-like-protein |

Note: NA-Uncharacterized Protein.

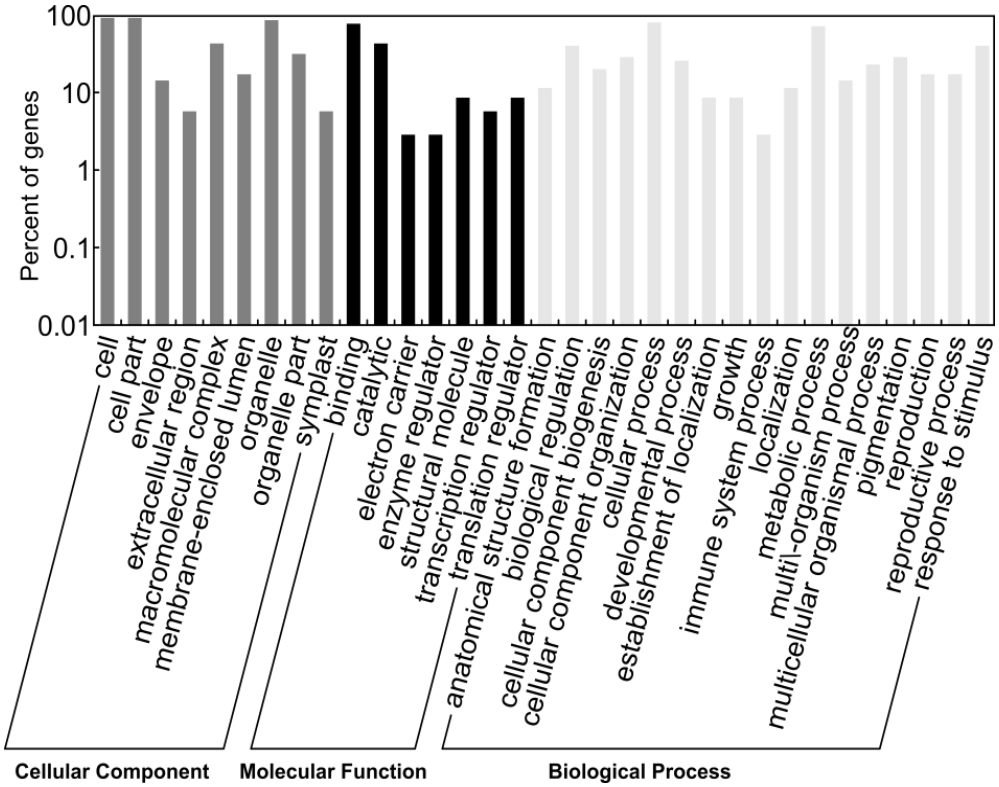



**Figure 3.6.** GO classification for 34 positive selection genes identified by BUSTED in *Arundo* speices.

A total of 114 positions involved in positive selection with *p*-value less than 0.05 were identified by MEME analysis and 108 sites under positive selection with posterior probability larger than 0.9 inferred by FUBAR analysis. Finally, 18 high-confidence, positively selected sites in 15 genes were identified by intersection of the FUBAR and MEME tests (Table 3.5; Supplementary Table 3.5), GO annotation showed that these positive genes were mainly involved in cellular process, metabolic process, response to stimulus, function in binding and catalytic activity (Figure 3.7).

**Table 3.5**. Sites of genes under positive selection identified by FUBAR and MEME test.

| OGs_ID | LRT | P-value*[a] | Post. Pr*[b] | Sites |
|---|---|---|---|---|
| OG0017439 | 4.910840728 | 0.039673917 | 0.964854608 | 130 |
| OG0017734 | 4.886642327 | 0.040172459 | 0.978015851 | 29 |
| OG0017740 | 12.02592948 | 0.001059838 | 0.968206388 | 698 |
| OG0017807 | 5.567516025 | 0.028291893 | 0.9106675 | 130 |
| OG0017866 | 7.357036058 | 0.011322496 | 0.906260864 | 313 |
| OG0017873 | 7.807895684 | 0.008998111 | 0.969347154 | 66 |
| OG0017952 | 7.270461903 | 0.011833683 | 0.951640614 | 22 |
| OG0018005 | 7.456835191 | 0.010760722 | 0.94859124 | 268 |
| OG0018090 | 4.480212728 | 0.049564056 | 0.978926353 | 6 |
| OG0018152 | 5.075037069 | 0.03645271 | 0.957773539 | 137 |
| OG0018354 | 8.338018877 | 0.006870342 | 0.946751769 | 187 |
| OG0018357 | 22.9553325 | 0.00000432 | 0.901402511 | 594 |
| OG0018364 | 5.675389459 | 0.02676671 | 0.971800775 | 156 |

| | | | | |
|---|---|---|---|---|
| OG0018364 | 5.159947683 | 0.034892199 | 0.900293898 | 167 |
| OG0018364 | 5.958537611 | 0.023146675 | 0.951457249 | 343 |
| OG0018374 | 18.39410022 | 0.0000428 | 0.964656214 | 891 |
| OG0018377 | 7.088597089 | 0.012984467 | 0.995937532 | 1277 |
| OG0018377 | 7.7754017 | 0.009148269 | 0.940865764 | 1446 |

Note: *a: MEME test with *p*-value <0.05; *b: FUBAR test with posterior probability > 90%.

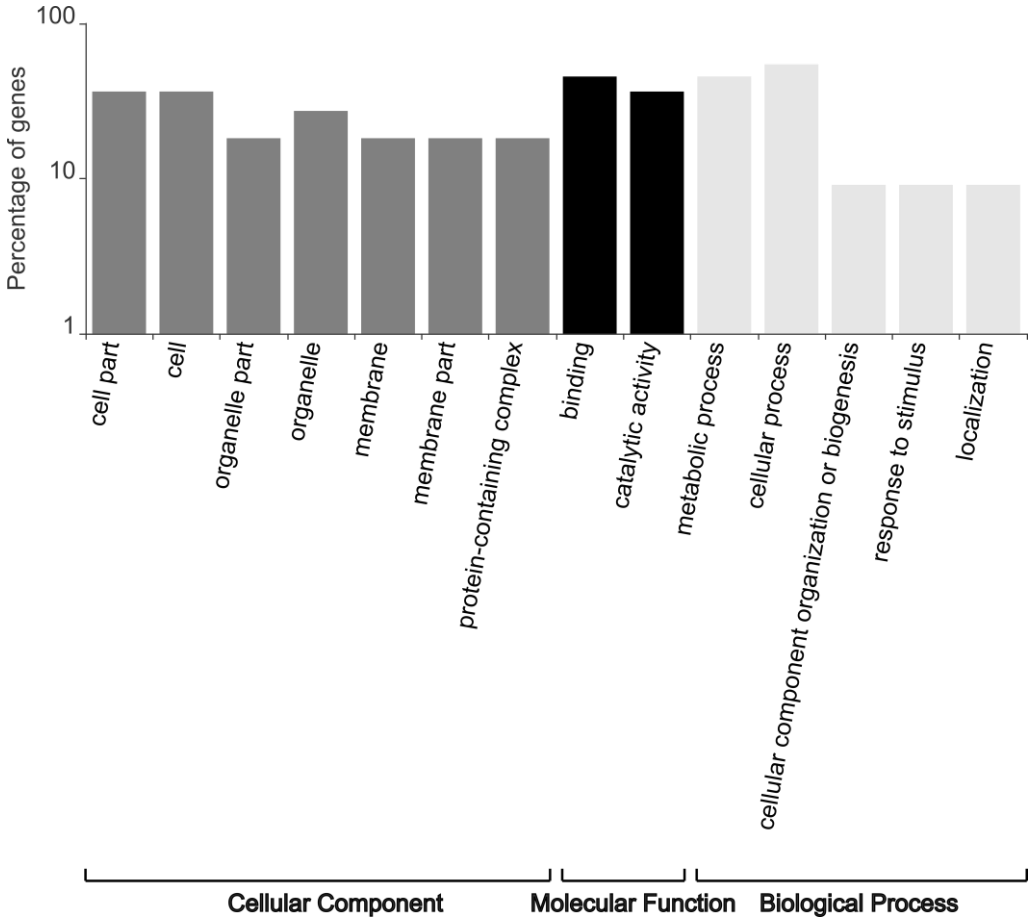



**Figure 3.6.** GO classification for 15 positive selection genes identified by FUBAR and MEME test in Arundo speices.

## 3.5 Discussion

### 3.5.1 *De novo* assembly of transcriptome data and function annotation

Transcriptome sequencing is an effective and accessible method used for comparative genomic analyses in non-model organisms when genome data are unavailable, as transcriptomes contain a large number of protein-coding genes that are most likely enriched for targets of natural selection. However, Illumina transcriptome or genome sequencing in the past was mainly limited to organism with reference genomes available (Rosenkranz et al., 2008). Former researchers confirmed that the relatively short reads can be effectively used for *de novo* assembly (Bentley et al., 2008), especially with the great advantages of paired-end Illumina sequencing method (Maher et al., 2009). The Illumina

transcriptome or whole genome sequencing and *de novo* assembly have been successfully used for model and non-model organisms, such as *A. thaliana* and maize (Weber et al., 2007; Wall et al., 2009; Vega-Arregu ń et al., 2009; Li et al., 2010; Collins et al., 2008; Wang et al., 2010). In this study, the first leaf transcriptomes for all taxa of the *Arundo* genus and three closely related outgroups were generated and annotated. Based on the illumina reads, *de novo* assembly yielded more than one million unigenes with an average length ranging from 741 to 1065bp, which suggested that the number of transcripts for all species and length were adequate for evolutionary and phylogenomics analyses.

Assessing all number of genes and the level of transcript coverage is an important issue for transcriptome sequencing projects. However, this is difficult without a reference genome. In this study thus the transcriptome coverage breadth was indirectly evaluated by determining the number of unique genes by BLAST annotation. A great number of unigenes could match to known proteins in public databases, which implied that the assembled unique genes were adequate from *Arundo* and three closely related outgroup species. GO annotation results showed that the majority of genes are associated with binding, catalytic activity and transporter activity under molecular function process category in each specie. Meanwhile, the majority of the genes are associated with cellular process, metabolic process and response to stimulus among all species. Compared to the leaf transcriptome of *Urochola humidicola* (Vigna et al., 2016), another non-model species from Poaceae, most of genes in *Arundo* species and *U. humidicola* have similar molecular functions in binding, transporter and catalytic activity. For cellular component process most of genes have similarity in that they are predicted as cell and organelle parts. Also under the biological process category the majority of the genes in both species are associated with cellular process, metabolism and response to stimulus. A large number of sequences and functional annotation comparison, thus, could provide sufficient transcriptomic information for discovering novel genes and phylogenomic reconstruction, and also confirm that high throughput Illumina paired-end sequencing is an efficient, inexpensive and reliable tool for transcriptome characterization and gene discovery in non-model species. At the same time, it confirmed that our assembled unigenes are appropriate for phylogenomic and evolutionary analyses, thanks to the total length of the assembly.

### 3.5.2 Phylogenomic reconstruction of *Arundo* species

Phylogenetic reconstruction of the relationships among *Arundo* species were previously proposed based on cpDNA sequences, suggesting that an Asian accession of *A. donax* was sister to the other members of the whole *Arundo* genus, with *A. donaciformis* and *A. plinii* being more closely related by plastid DNA analysis (Hardion et al., 2014(a); Hardion et al., 2014(b)). The inconsistence between these early phylogenies and our results might result from several factors: (1) the small number of chloroplast markers used in previous studies, (2) their maternal inheritance (Pillon et al., 2013; Yi et al., 2015; Zimmer and Wen, 2012) and (3) the still incomplete sampling of taxa in the *Arundo* genus. In this study, the species tree has been reconstructed from the transcriptome-based super-alignments of 150 one-to-one orthologous genes among 10 species, those genes were extracted from single-copy orthologs identified by OrthoFinder, using 3 different methods: Maximum Parsimony (MP), Maximum Likelyhood (ML), and Bayesian Inference (BI). The high number of informative and divergent characters used and their mainly nuclear origin provide extremely congruent phylogenies, suggesting that they accurately reflect the true evolutionary relationships among *Arundo* species, preventing the possible artefacts associated to phylogenies based on few maternally inherited chloroplast loci. This does not

exclude the possibility that another, very basal taxon belonging to the *Arundo* genus might exist in Asia and that the results from previous studies simply reflect differences in taxon sampling (Hardion et al., 2014(b)). According to our phylogenetic analysis, however, we can confidently conclude that *Arundo* is divided into three major linages: one is composed only by the species *A. formosana*; the second one comprises *A. macrophylla* and *A. donax*; the last one comprise *A. micrantha*, *A. donaciformis*, *A. plinii* and *A. collina*. These results are consistent with the conclusion put forth in precedence that *A. collina* is a synonym of *A. plinii* (Hardion et al., 2012) and that *A. macrophylla* is a variety with larger leaves of the invasive clone of *A. donax* present in the Mediterranean area (hort.purdue.edu/newcrop/duke_energy/Arundo_donax.html). Thus, *A. formosana* is sister to the other members of the *Arundo* genus as currently defined, and *A. micrantha* is the basal species of the clade encompassing also *A. donaciformis* and *A. plinii*. We thus confirm that *A. donaciformis* has a close evolutionary relationship to *Arundo plinii*, but we cannot either confirm or refute the hypothesis recently suggested that *A. donaciformis* is a derivative of *A. plinii* originating from a hemiploidization event stemming from fusion of reduced and unreduced gametes (Hardion et al., 2014(a)). While awaiting the solution of the possible existence of an Asian/Middle East *A. donax*-like taxon basal to the whole genus, our study suggests a possible scenario for the evolution and geografic migratory route followed by the known *Arundo* species during the genus radiation. According to the data currently available, *A. formosana*, currently endemic to Taiwan and few other Eastern Asia countries (Lin et al., 2006), probably originated in Eastern Asia, and later differentiated into the invasive clone of *A. donax* , which immigrated into middle Asia, to then finally spread into Mediterranean and all over the world (Mariani et al., 2010). Interestingly, western Mediterranean species of *Arundo* (*A. micrantha*, *A. donaciformis* and *A. plinii*) originated afterwards from another lineage at lower ploidy level than the invasive *A. donax* clone. They underwent independent demi-polyploidization historically in *A. donaciformis* through the same process still ongoing in some Italian populations of *A. plinii*, and either spreading by human intervention or being an event of marginal polyploidization (Hardion et al., 2014(a)).

Previous studies proposed different scenarios for the evolutionary origin of *A. donax*. In a first study, two contrasting hypotheses of auto-polyplodization *versus* allo-polyploidization were put forward for the origin of the *A. donax* invasive clone (Bucci et al., 2013). The first hypothesis (auto-polyplodization) was that *A. plinii* underwent a chromosome duplication and produced a fertile tetraploid crossed with a diploid *Arundo plinii* creating the sterile triploid *A. donax*. Another hypothesis (allo-polyploidization) was that the fertile tetraploid *A. plinii* crossed with *P. australis* to produce the sterile hybrid *A. donax*. Hybridization could indeed be possible at the relatively close evolutionary distances between the congeneric species *A. donax* and *A. plinii*, as reported for other genera (e.g. Brochmann et al., 2000). With the increase of evolutionary divergence among species, however, hybridization results progressively less common (Abbott et al., 2013). Therefore, hybridization among species from different tribes like *A. plinii* and *P. australis*, although in principle possible, is less likely than the former hypothesis. The results of this study, however, provide no evidence to support either of these two hypotheses. In fact, *A. donax* placement in the phylogenomic reconstruction is clearly neither derived from *A. plinii* as one would expect based on the first hypothesis nor related to *P. australis* as one would expect based on the second hypothesis.

### 3.5.3 Chromosome evolution

This study contributes to shed new light also on the evolutionary history of the *A.*

*donaciformis*/*A. plinii* clade. A previous study indicated that the demi-polyploidization event is very important in the origin of *Arundo donaciformis*, hypothesizing that it differentiated from *Arundo plinii* (Hardion et al., 2014(a)). Our results show that *A. donaciformis* is basal to *A. plinii*, demonstrating that they originated from a common ancestor, but suggesting to confute the direct derivation of *A. donaciformis* from current-day *A. plinii*. This result further indicates that *A. donaciformis* may possibly be a species separated from *A. plinii*, with important implications on the conservation priority of *A. donaciformis*. Re-evaluation of previous hypotheses on the origin of two of the species with the highest chromosome numbers in the *Arundo* genus is also relevant for the elucidation of the processes underlying its evolution. Chromosome number changes during the evolution of *Arundo* species, in fact, are likely to have played a major role in speciation. The study of chromosome number evolution allowed to infer the ancestral haploid numbers for the *Arundo* genus using an evolutionary model with a statistically robust approach. The ancestral chromosome number for *Arundo* genus obtained under the ML approach and Bayesian inference was n = 36, suggesting that demi-duplication is the main driving force for chromosome evolution of the *Arundo* genus. Demi-duplication took place at least 2 times independently during the genus radiation: (1) Once during the evolution of *A. donax* invasive clone; and (2) a second time in the lineage of *A. donaciformis*. The observation that this is still an ongoing process in *A. plinii* populations (Hardion et al., 2014(a)) further corroborates the fundamental role played by demi-duplication throughout the evolution of the *Arundo* genus. In the future, it will be interesting to test whether *A. donax* and *A. donaciformis* are autopolyploids, as suggested by the *A. plinii* demiploidization mechanism of fusion of reduced and unreduced gametes.

### 3.5.4 Evolutionary analysis

Positive selection is an important factor for evolutionary innovation and environmental adaptation. One of the aims of this study was to identify genes subject to positive selection. However, recombination can mislead phylogenetic analysis if ignored (Schierup and Hein, 2000), so recombination testing was performed before testing for positive selection. Recombination was found to be widespread among orthologous genes. A total of fifteen genes were found under strong positive selection along different branches in the phylogeny, and functional annotation suggested that these genes were involved in important biological functions, such as DNA-directed RNA polymerization (OG0017657), ubiquinone oxidoreduction (OG0018204), ribosomal translation (OG0018354) and RNA helicase activity (OG0017807), while one of the candidate genes (OG0017508) under positive selection still remains of uncharacterized function (Supplementary Table 3.4). Interestingly, five genes were found to be under positive selection along the *Arundo donaciformis* branch, but at present the reason for this branch-specific intensification of positive selection remains unknown. Thirty-four positively selected genes were identified by BUSTED, and the majority of them could be annotated with putative protein homologs, but for two of them without significant homology to other proteins no function could be inferred (Table 3.4). Many genes were identified under positive selection in this study, and this might be due to high rates of false positive results produced by using the branch-site unrestricted statistical test for episodic diversification (BUSTED) (Venkat et al., 2018). Two of the positively selected genes (OG0018069 and OG0018070) are homologous to the BiP (Luminal binding protein) gene. A previous study has shown that the *BiP2* gene (*Luminal-binding protein 2*) functions in polar nuclei fusion during endosperm nuclei proliferation (Xu et al., 2013). BiP can reduce the stress of the endoplasmic reticulum, and a previous study in soybean and tobacco showed that BiP functions in drought tolerance

and delaying leaf aging (Valente et al., 2009), which may be used as a powerful tool for improvement of the biomass in *A. donax*. The gene *CUL4* (OG0018339) has an E3 ubiquitin ligase protein function in mediating light control of plant development and is important for photomorphogenesis repression. A previous study showed that *CUL4* loss of function resulted in morphogenesis and light-regulated gene expression (Chen et al., 2006). CHL (OG0018016) is a thylakoid lumenal protein involved in thylakoidal membrane lipids protection (Levesque-Tremblay et al., 2009). TIF3A1 (OG0018342) is a translation initiation factor containing a conserved PCI or PINT motif at the protein N-terminus (Burks et al., 2001). A site of *TIF3C1* (OG0018354) gene is also under positive selection, indicating that this site might be an active site of the enzyme and involved in interaction with other proteins to regulate light control of plant development (Wei and Deng, 1999). In this study, mainly positively selected genes were identified and annotated, but other genes should also be worthy of functional analysis, which could provide further insights into the evolution of the *Arundo* species.

### 3.6 Conclusion

In this study, the first report is provided on the accurate reconstruction of the relationships among all currently recognized species of the *Arundo* genus by means of a phylogenomic approach using 150 one-to-one orthologous genes. The results confirmed that *A. formosana* is sister to the other sampled species of the *Arundo* genus rather than *A. donax*, which is in contrast with recent chloroplast-based phylogenomic trees possibly due to different taxa sampling and/or existence of cryptic *Arundo* species still to ascertain. Based on this study, however, previous hypotheses on *A. donax* and *A. donaciformis* evolution can be confidently refuted, further suggesting a likely autopolyploid origin for both taxa. The probabilistic models suggest that the ancestral haploid chromosome number of *Arundo* was 36 (likelihood and Bayesian framework) and suggested that independent demi-duplications were responsible for the evolutionary increases in chromosome numbers of *A. donax* and *A. donaciformis*. This study also found some genes under positive selection, which provide valuable insights into adaptive selection of the *Arundo* genus at the sequence level and hold great potential for future gene functional validation and improvement of the biomass species *A. donax*.

# CHAPTER 4

*In silico* identification and comparative analysis of lignin and cellulose biosynthesis gene families across the Arundinoideae (Poaceae)

## 4.1 Abstract

Cellulose, hemicellulose and lignin are three important components for plant cell wall, supporting various environmental stress responses. Meanwhile, they are also important renewable sources for the production of biofuels, making their study relevant from both applied and fundamental research. In this study, *in silico* identification and comparative analyses of lignin and cellulose biosynthesis gene families in Arundinoidea species were carried out. A total of 741 protein sequences from CesA/Csl and a total of 1118 genes from 10 lignin biosynthetic gene families were identified. Phylogenetic analysis of CesA/Csl proteins showed that CesA/Csl genes classified into 8 clades including CSLA, CSLC, CSLD, CSLE, CSLF, CSLH and CSLJ subfamilies and CESA gene family in Arundinoideae, and the phylogenetic tree also showed that the CSLA and CSLC subfamilies form an independent lineage to other CesA/Csl gene subfamilies, indicating that they are probably originated from a separate duplication event. Phylogenetic analysis using all lignin biosynthesis gene families showed that these genes are highly divergent between eudicots and monocots. Phylogenetic reconstruction of C3H, F5H and PAL proteins showed clear separation among *Amborella*, eudicots and monocots, respectively, indicating that these genes might have experienced expansion event after species differentiation. Other phylogenetic tree reconstruction of lignin biosynthesis proteins showed that these gene family divided into different classes based on reference species (rice, *Arabidopsis* and *Amborella*), indicating that diverse functions might exist in these genes families. The cellulose and lignin biosynthesis genes identified in this study will be helpful for establishing mutagenesis-based reverse genetics and functional genomics approaches in biomass species *A.donax*.

## 4.2 Introduction

Lignocellulosic biomass is an important and promising renewable source used for production of biofuel, which has attracted great attention. In recent years many efficient technologies have been proposed for pretreatment of lignocellulosic biomass to produce bioenergy, such as the micro-aerobic pretreatment (Foyle et al., 2006; Kumar et al., 2009; Jönsson et al., 2013; Wi et al., 2015; Amin et al., 2017). Lignocellulose is a complex constituted by three main components including cellulose, hemicellulose and lignin. More specifically, cellulose and hemicellulose are a class of heteroglycans used for bioconversion into biofuels, while lignin is a class of aromatic polymers that need to be removed from lignocellulose biomass (Mussatto and Teixeira, 2010). For improving the efficient utilization of lignocellulosic biomass by biological technology, more and more studies were carried out for identifying cellulose synthase (CesA), cellulose synthase-like (Csl) and lignin biosynthesis gene families from biofuel crops and other plants, such as *Setaria italica*, *Populus trichocarpa* and wheat (Muthamilarasan et al., 2015; Suzuki et al., 2006; Kaur et al., 2017). In a previous study, out of a total of 45 CesA/CSL genes identified from *O. sativa*, 11 were identified as CesA gene members, 34 as gene members of Cellulose synthase-like Family (including CSLA, CSLC, CSLD, CSLE, CSLF and CSLH subfamilies; Rice Genome Annotation Project, http://rice.plantbiology.msu.edu/) (Hazen et al., 2002; Wang et al., 2010). This dataset has provided a good reference for

computational identifying CesA/CSL gene family from biofuel crops. Lignin is an important component of plant cell walls, which is related to structural support and response to various environmental stresses (Liu et al., 2018). There are 10 gene families reported as involved in lignin biosynthesis, namely CAD (cinnamyl alcohol dehydrogenase), CCoAOMT (caffeoyl-CoA O-methyltransferase), 4CL (4-coumarate: CoA ligase), CCR (cinnamoyl-CoA reductase), PAL (phenylalanine ammonia-lyase), C4H (cinnamate 4-hydroxylase), HCT (hydroxycinnamoyl-CoA shikimate/Quinate hydroxycinnamoyl transferase), COMT (caffeic acid O-methyl transferase), C3H (p-coumarate 3-hydroxylase) and F5H (ferulate 5-hydroxylase) gene families (Liu et al., 2018; Raes et al., 2003; Xu et al., 2009). They have been all identified among the phenylpropanoid biosynthetic genes deposited into the cell wall genomics database (https://cellwall.genomics.purdue.edu/).

*A. donax* has been indicated among the most important and promising next-generation bioenergy crops (Angelini et al., 2009). Lignocelluloses content (cellulose, hemicellulose and lignin) in biomss *A. donax* as feedstocks has great potential for increasing production of biofuel, and interactions among these components have effect on slow steam pyrolysis of biomass (Giudicianni et al., 2014). Pretreatment of lignocelluloses is an important step to improve the productivity of biofuel in *A. donax*. For instance, microwave irradiation pretreatment has been utilized to improve cellulose hydrolysis and a previous study confirmed that efficient hydrolysis occured in hemicellulose at ideal conditions (Komolwanich et al., 2014). In recent years, physical mutagenesis also has been developed and carried out based on γ-irradiation for genetic improvement of this biomass *A. donax* (Valli et al., 2017), which lays solid bases for the identification of *A. donax* mutants with increased productivity.

In this study, computational identification and comparative analysis were carried out for the mining of biosynthetic enzymes involved in lignocellulose biosynthesis genes from *A. donax, A. macrophylla, A. formosana, A. donaciformis, A. micrantha, A. plinii, A. collina*, *H. macra*, *M. caerulea* and *P. australis* leaf unigenes. These identified gene families will provide a good opportunity to establish mutagenesis-based reverse genetics and functional genomics approaches in biomass species *A. donax*.

## 4.3 Materials and Methods

### 4.3.1 Computational identification of cellulose and lignin biosynthesis gene families

Assembled Unigenes of Arundinoideae leaf (see CHAPTER 3 of this thesis) were used for comparative analysis of lignocellulose biosynthesis-related gene families, and these unigenes were analysed for prediction of protein sequences by the GENSCAN program (Burge and Karlin, 1998). A total of 45 protein sequences namely CesA (Cellulose synthases) and Csl (Cellulose synthase-like gene family) of rice were retrieved from the Rice Genome Annotation Project website (http://rice.plantbiology.msu.edu/). A total of 63 protein sequences for 10 gene families namely PAL, C4H, 4CL, HCT, C3H, CCoAOMT, F5H, COMT, CCR and CAD of *Arabidopsis* were collected from the Cell Wall Genomics website (https://cellwall.genomics.purdue.edu/). Cellulose and lignin

biosynthesis homologous genes were identified from Arundinoideae leaf unigenes by application of the Blastp algorithm using the 45 rice cellulose biosynthetic genes and the 63 *Arabidopsis* lignin biosynthesis genes as reference queries with an E-value cut-off of 1e-5, score value greater than 50 (R. Pearson, 2013). In addition, we also identified CesA/Csl superfamily gene and lignin biosythesis gene family from *Amborella* dataset retrieved from Phytozome v12 (http://www.phytozome.net), and identified lignin biosythesis gene family from rice dataset retrieved from the Rice Genome Annotation Project website. Finally, the identified protein sequences were confirmed by HMMSCAN from Pfam (https://pfam.xfam.org/) for protein function domain comparison with known reference proteins under default parameters. As lignin biosythesis gene family is large, in order to reduce false positives, the identified lignin biosythesis proteins were then blasted against all *Arabidopsis* proteins, and only the best hits to one of 63 *Arabidopsis* lignin biosynthesis genes were retained for the following analysis.

### 4.3.2 Sequence alignment and phylogenetic analysis

The identified protein sequences were used for manually removing short sequences, as these short sequences have a big potential impact on the accuracy of phylogenetic reconstruction. Then, the remaining sequences were used for reconstruction of the phylogenetic trees. The protein sequences of the respective gene families were aligned by MUSCLE v3.8.31 (Edgar, 2004) and the program Gblock v0.91 (Castresana, 2000) was used to select conserved blocks with less strict parameters: maximum number of contiguous nonconserved positions = 8, minimum length of a block = 5, and allowed gap positions: with half. The best-fitting model of protein evolution was inferred by SMS web server (Smart Model Selection, web server: http://www.atgc-montpellier.fr/sms/) (Lefort et al., 2017), Maximum-likelihood (ML) tree was reconstructed by PhyML v3.0 (Guindon et al., 2010), and Branch Support was calculated using aLRT (approximate Likelihood-Ratio Test) (Anisimova and Gascuel, 2006).

## 4.4 Results

### 4.4.1 Identification of CesA/Csl gene superfamily and lignin biosynthetic gene families

Homology search and functional domain comparison were used for identifying lignin biosynthesis and CesA/Csl gene families, using the Pipeline shown in Figure 4.1. Previously, about 45 members of the CesA/Csl gene families have been reported in rice (Wang et al., 2010), and 63 members of lignin biosynthetic genes in *Arabidopsis* (Xu et al., 2009). These protein sequences from rice and *Arabidopsis* were used for lignin and cellulose biosynthetic gene families identification. The protein sequences not having clear functional domain for CesA/Csl gene superfamily were removed from candidate sequences. Finally, there were 53, 65, 51, 98, 48, 130, 90, 33, 113 and 60 CesA/Csl proteins identified from *A. donax*, *A. macrophylla*, *A. formosana*, *A. donaciformis*, *A. micrantha*, *A. plinii*, *A. collina*, *H. macra*, *M. caerulea* and *P. australis* leaf unigenes, respectively (Table 4.1). Meanwhile, 28 CesA/Csl protein sequences (11 CesA and 17 Csl) were identified in *Amborella* (CesA/Csl biosynthetic proteins id of rice, *Arabidopsis* and *Amborella* as reference species are shown in supplementary table 4.1). We renamed

the identified genes based on the known genes from rice and *Arabidopsis*, integrating them with previous reports and phylogenetic tree reconstruction (Schwerdt et al., 2015; Supplementary Figure 4.1). Finally, we identified 8 groups across the CSLA, CSLC, CSLD, CSLE, CSLF, CSLJ, CSLH and CesA gene families for the 10 Arundinoideae species. As one can see from table 4.1, CSLJ is the smallest gene family, containing only 16 genes, while CSLA and CesA contain together more than 100 genes across the 10 species. CesA was the biggest gene family with 233 members across all species. A total of 130 CesA/Csl biosynthetic genes were found in *A. plinii*, while only 33 genes were identified in *H. macra*. These differences between species leaves probably were related to the size of each species' gene family, sequencing dataset and genes expressed in other tissues.

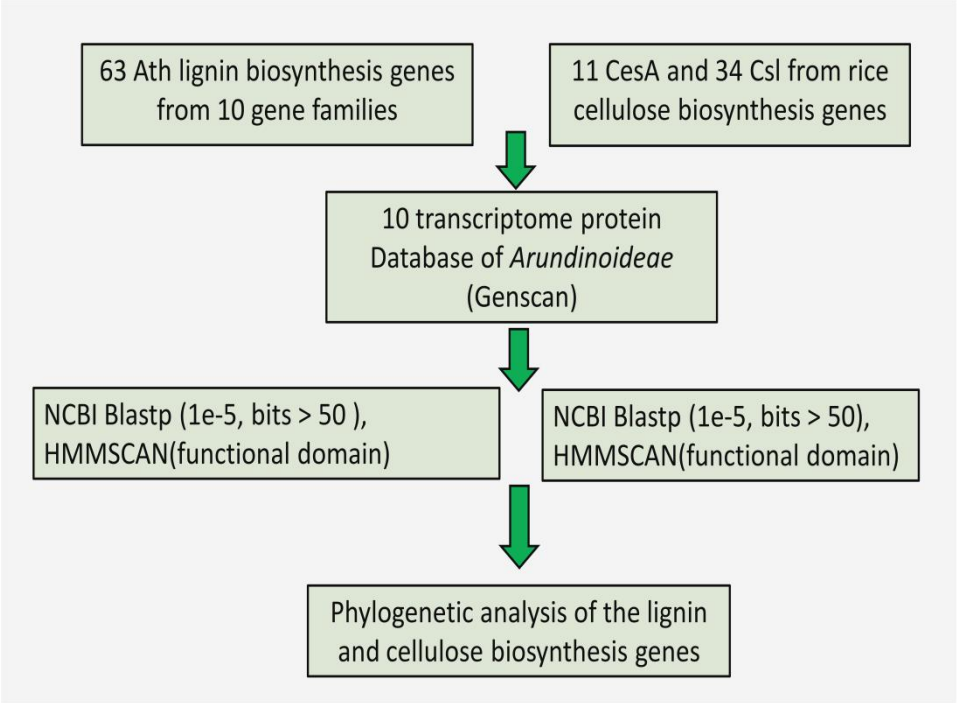



**Figure 4.1.** Schematic workflow of the study.

**Table 4.1.** Identification of CesA/Csl genes in each gene family across 10 species in Arundinoideae.

| Species | CesA | CSLA | CSLC | CSLD | CSLE | CSLF | CSLJ | CSLH | Total |
|---|---|---|---|---|---|---|---|---|---|
| *Arundo donax* | 18 | 24 | 0 | 1 | 2 | 6 | 0 | 2 | 53 |
| *Arundo macrophylla* | 28 | 14 | 1 | 3 | 4 | 10 | 2 | 3 | 65 |
| *Arundo formosana* | 25 | 7 | 1 | 5 | 2 | 5 | 3 | 3 | 51 |
| *Arundo donaciformis* | 34 | 30 | 1 | 2 | 6 | 13 | 2 | 10 | 98 |
| *Arundo micrantha* | 24 | 4 | 2 | 6 | 2 | 6 | 3 | 1 | 48 |
| *Arundo plinii* | 30 | 52 | 3 | 12 | 22 | 7 | 0 | 4 | 130 |
| *Arundo collina* | 17 | 29 | 2 | 8 | 7 | 22 | 1 | 4 | 90 |
| *Hakonechloa macra* | 16 | 2 | 0 | 3 | 8 | 0 | 1 | 3 | 33 |
| *Molinia caerulea* | 24 | 11 | 6 | 6 | 15 | 24 | 4 | 23 | 113 |
| *Phragmites australis* | 17 | 7 | 3 | 3 | 7 | 1 | 0 | 22 | 60 |
| **Total** | 233 | 180 | 19 | 49 | 75 | 94 | 16 | 75 | 741 |

The lignin biosynthetic genes identified across the 10 species are summarized in Table 4.2, while the gene IDs of the three reference species used (rice, *Arabidopsis* and *Amborella*) are shown in Supplementary Table 4.2. Finally, a total of 104, 111, 80, 169, 116, 150, 182, 38, 96 and 72 lignin biosynthetic genes were identified from *A. donax*, *A. macrophylla*, *A. formosana*, *A. donaciformis*, *A. micrantha*, *A. plinii*, *A. collina*, *H. macra*, *M. caerulea* and *P. australis* leaf unigenes, while 84 and 80 lignin biosynthetic genes were found in rice and *Amborella* protein datasets, respectively (Table 4.2). All 10 lignin biosynthetic gene families were detected in each Arundinoideae species except for the C4H and F5H genes family, which were not found in *H. macra* and the F5H genes family was not found in *A. plinii*, probably due its lack from the corresponding sequencing dataset. F5H was the smallest gene family with 15 gene members, and five gene families namely CAD, 4CL, CCR, HCT and COMT had more than 100 member genes across the 10 Arundinoideae species. A total of 38 lignin biosynthetic genes were identified in *H. macra*, and 182 genes were found in *A. collina*, as a possible consequence of the fact that both the sizes of each species' gene family, genes expressed in other tissues and sequencing dataset may cause differences in gene identification across Arundinoideae.

**Table 4.2.** Identification of lignin biosynthesis genes in each gene family across 10 species in Arundinoideae.

| Species | CAD | CCoAOMT | 4CL | CCR | PAL | C4H | HCT | COMT | C3H | F5H | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Arundo donax* | 12 | 7 | 18 | 15 | 9 | 5 | 8 | 26 | 3 | 1 | 104 |
| *Arundo macrophylla* | 9 | 10 | 15 | 20 | 7 | 5 | 15 | 24 | 3 | 3 | 111 |
| *Arundo formosana* | 12 | 5 | 10 | 20 | 7 | 3 | 8 | 10 | 2 | 3 | 80 |
| *Arundo donaciformis* | 16 | 8 | 32 | 49 | 6 | 3 | 21 | 28 | 4 | 2 | 169 |
| *Arundo micrantha* | 10 | 18 | 19 | 20 | 5 | 3 | 16 | 21 | 2 | 2 | 116 |
| *Arundo plinii* | 11 | 4 | 11 | 17 | 7 | 4 | 2 | 92 | 2 | 0 | 150 |
| *Arundo collina* | 25 | 13 | 27 | 67 | 6 | 9 | 19 | 11 | 4 | 1 | 182 |
| *Hakonechloa macra* | 3 | 3 | 7 | 10 | 3 | 0 | 3 | 7 | 2 | 0 | 38 |
| *Molinia caerulea* | 5 | 11 | 8 | 20 | 15 | 3 | 12 | 18 | 2 | 2 | 96 |
| *Phragmites australis* | 5 | 12 | 8 | 10 | 4 | 5 | 10 | 14 | 3 | 1 | 72 |
| **Total** | 108 | 91 | 155 | 248 | 69 | 40 | 114 | 251 | 27 | 15 | 1118 |

### 4.4.2 Phylogenetic analysis of CesA/Csl proteins

The unrooted phylogenetic tree reconstructed using multiple protein sequences alignment of CesA/Csl gene families from *A. donax*, *A. macrophylla*, *A. formosana*, *A. donaciformis*, *A. micrantha*, *A. plinii*, *A. collina*, *H. macra*, *M. caerulea* and *P. australis* with three reference species *O. sativa*, *Arabidopsis* and *Amborella* showed that CesA/Csl genes classified into 10 clades including CSLA, CSLB, CSLC, CSLD, CSLE, CSLF, CSLG, CSLH and CSLJ subfamilies and CESA gene family. CSLG and CSLJ were closely related in the phylogenetic tree (Figure 4.2). The CSLG gene subfamily was found in *Arabidopsis* and *Amborella*, while the CSLB group was only found in *Arabidopsis*. However, both of these two groups were not represented in the leaf transcriptomes of any Arundinoideae species, indicating that these two gene subfamily are probably expressed in other tissues or have been lost in this clade. The CSLA and CSLC gene families are broadly distributed in monocot and eudicot species, as well as in *Amborella*, indicating that both of these gene families originated before the monocot/eudicot divergence event, and the phylogenetic tree showed that the CSLA and CSLC subfamily clade is an independent lineage to other CesA/Csl gene families, indicating that they are probably originated from a separate ancestral duplication event.

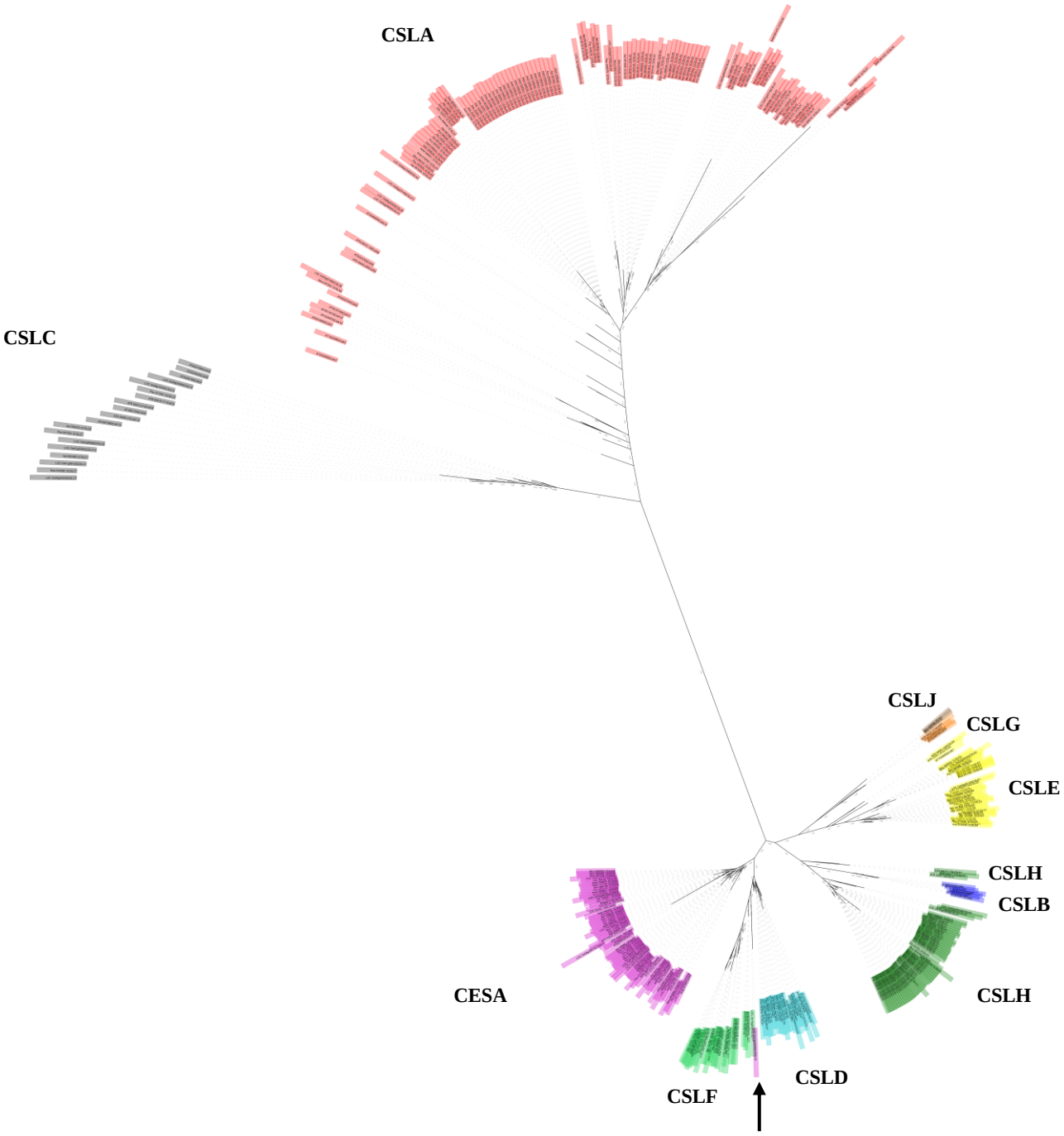


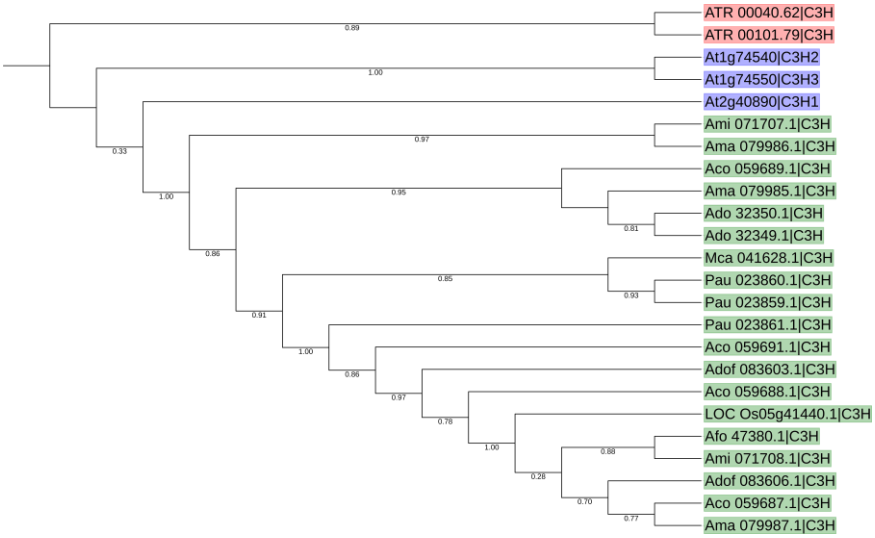

**Figure 4.2.** Reconstruction of unrooted phylogenetic tree with CesA/Csl proteins of *Oryza sativa* (LOC_Os), *Arabidopsis* (AT), *Amborella* (ATR) and Arundinoideae species. The arrow represents CESA10 (LOC_Os12g29300) in rice.

### 4.4.3 Phylogenetic analysis of lignin biosynthesis proteins

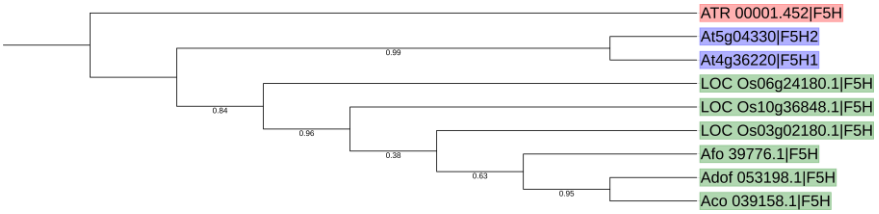
Phylogenetic analysis using each lignin biosynthesis gene family of Arundinoideae species, rice, *Arabidopsis* and *Amborella* showed that lignin biosynthesis genes were highly divergent between eudicots and monocots. Phylogenetic reconstruction of C3H, F5H and PAL proteins showed clear separation among *Amborella*, eudicots and monocots, indicating that these genes might have experienced expansion event after

species differentiation (Figure 4.3).

**(A)**




**(B)**




**(C)**

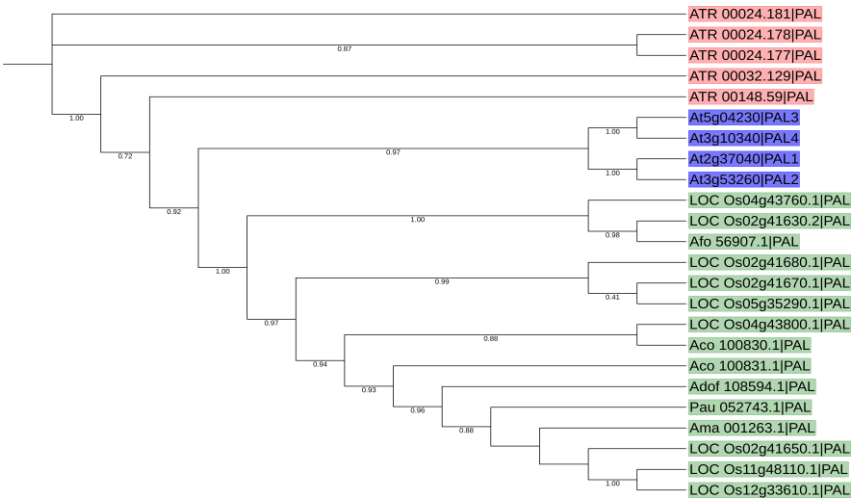



**Figure 4.3**. Phylogenetic analysis of C3H(A), F5H(B) and PAL(C) proteins of *Oryza sativa* (LOC_Os), *Arabidopsis* (At), *Amborella* (ATR) and Arundinoideae species, respectively.

Phylogenetic reconstruction of CAD, CCoAOMT, 4CL, CCR, HCT, COMT and C4H proteins showed that these gene family divided into different groups based on *Arabidopsis* and *Amborella*, indicating that diverse functions might exist in these gene families. For instance (Figure 4.4) homologs of putative CAD genes are distributed across *Amborella*, eudicots and monocots, and these genes clustered into three classes based on known CAD genes in *Arabidopsis* (Class I, Class II and Class III). Specifically, Class I included CAD2 and CAD6, Class II clustered with most CAD members, namely CAD1, CAD3, CAD4, CAD5, CAD7 and CAD8, while Class III contained CAD9. All classes encompass homologs from *Amborella*, eudicots and monocots.

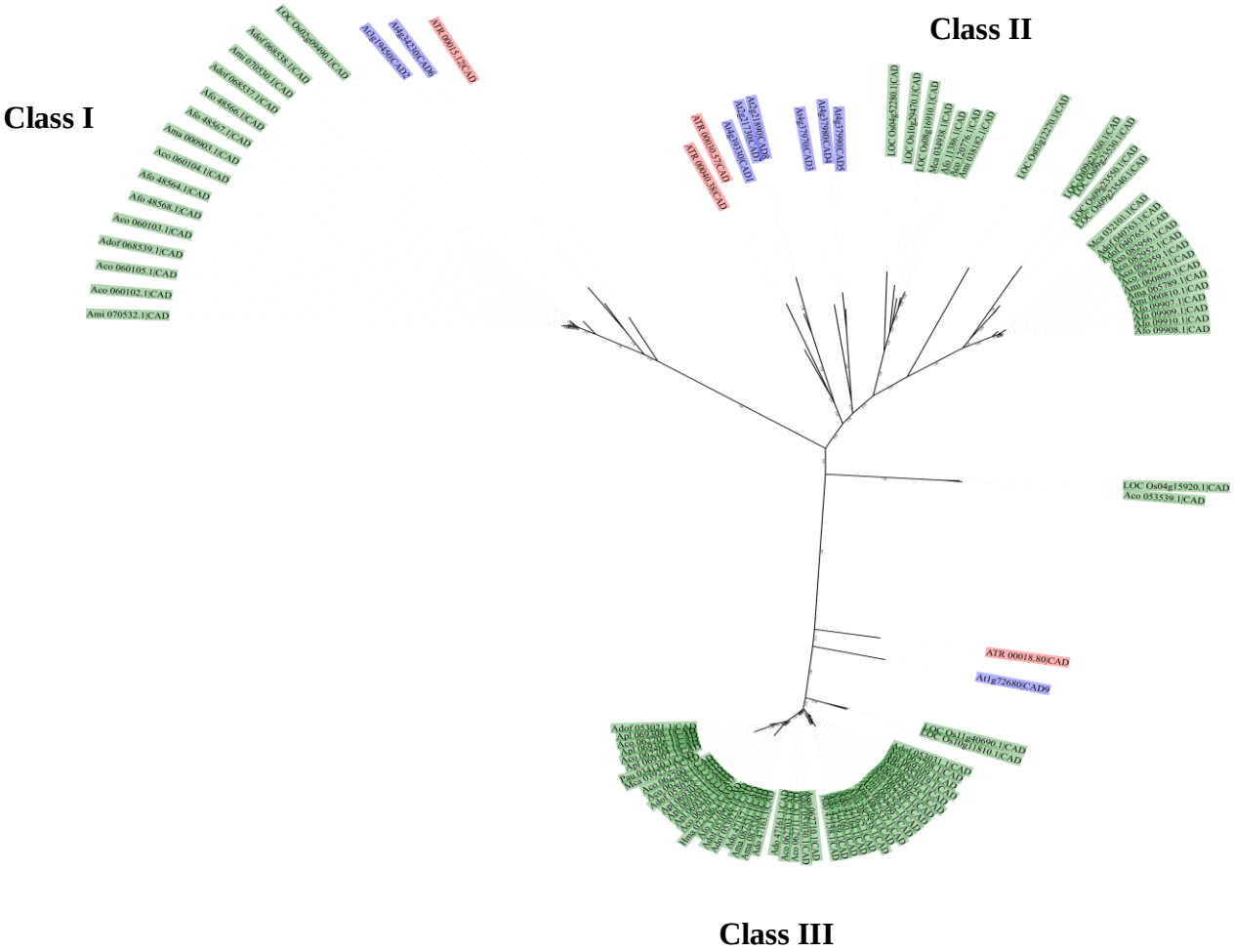



**Figure 4.4**. Phylogenetic reconstruction of CAD proteins of *Oryza sativa* (LOC_Os), *Arabidopsis* (At), *Amborella* (ATR) and Arundinoideae species.

## 4.5 Discussion

In this study, bioinformatic approaches and transcriptomic data were used for identification and comparative analysis of several cellulose and lignin biosynthesis gene families in 10 Arundinoideae species. These gene families are responsible for the biosynthesis of the most important components of plant cell walls, and are involved in plant development and growth (Hamann et al., 2004; Liu et al., 2018).

### 4.5.1 Identification and phylogenetic analysis of CesA/Csl gene families

A total of 233 CesA and 508 Csl genes were identified in Arundinoideae species, and all putative CesA proteins have cellulose synthase domain, as previously reported in rice. CESA10 (LOC_Os12g29300) in rice was short and only contained cellulose synthase domain without zinc finger domain, so it was excluded from CESA clade showed in Figure 4.2 and its function is still unclear (Wang et al., 2010). In previous studies, CesA1, CesA3 and CesA6 have been reported to be involved in primary cell wall cellulose synthesis (Persson et al., 2007), and these genes related to plant development. CesA4, CesA7 and CesA8 were found to be related with secondary cell wall development (Taylor et al., 2003), leading to the conclusion that these genes were important for cellulose synthesis. Homology-based functional characterization of these gene families indicated that CesA identified in Arundinoidea species may have potential functions in cell wall cellulose synthesis, such as CesA8 gene have found to be involved in cellulose synthesis in *Arabidopsis*, which can be applied to enhancing drought and osmotic stress tolerance (Chen et al., 2005). Previous studies also showed that Csl (Cellulose synthase-like) genes have sequences similarity to CesA genes (Richmond and Somerville, 2000), and are involved in hemicelluloses biosynthesis (Yin et al., 2009). The phylogenetic tree showed that 741 CesA/Csl genes in Arundinoidea were divided into 8 clades namely CESA, CSLA, CSLC, CSLD, CSLE, CSLF, CSLJ and CSLH based on CESA/CSL genes identified in rice, consistently to previously reported homologs in Grasses (Wang et al., 2010; Schwerdt et al., 2015). However, CSLG and CSLB were not represented in Arundinoidea species, probably being expressed in other tissues, as it was previously reported that at least CSLG is found in monocots (Yin et al., 2014). Phylogenetic tree reconstructed in Figure 4.2 showed that CSLB and CSLH clustered closely, and CSLH proteins were distributed in basal species *Amborella*, suggesting that CSLB clade should be more appropriately renamed as CSLH. CSLA and CSLC gene families were divergent from other CesA/Csl gene families, indicating that CSLA and CSLC families generated by duplication event from a common ancient gene (Yin et al., 2009). Previous studies reported that CSLA genes were related to glucomannan synthesis (Liepman et al., 2007), while CSLC genes encode β-1,4 glucan synthase (Cocuron et al., 2007), which is involved in xyloglucan biosynthesis. Cellulose synthase-like D (CSLD) subfamily has been demonstrated to play an important role in normal development and growth in *Arabidopsis* (Yin et al., 2011). It was reported in rice that Cellulose synthase-like F functions in (1,3;1,4)-β-D-Glucan synthesis (Burton et al., 2006). Cellulose synthase-like H gene is involved in the synthesis of (1,3;1,4)-β-D-Glucan (Doblin et al., 2009), and phylogenetic tree showed that CSLH subfamily is broadly distributed in monocots.

### 4.5.2 Identification and phylogenetic analysis of lignin biosynthesis gene families

A total of 1118 genes were identified from 10 monolignol biosynthesis gene families including CAD, CCoAOMT, 4CL, CCR, PAL, C4H, HCT, COMT, C3H and F5H among Arundinoideae species. A total of 104, 111, 80, 169, 116, 150, 182, 38, 96 and 72 proteins were identified in *A. donax*, *A. macrophylla*, *A. formosana*, *A. donaciformis*, *A.*

*micrantha*, *A. plinii*, *A. collina*, *H. macra*, *M. caerulea* and *P. australis* leaf unigenes, respectively. Phylogenetic analysis of lignin biosynthesis genes in Arundinoideae with *Amborella*, *Arabidopsis* and rice showed that lignin biosynthesis genes of Arundinoideae are highly divergent from *Arabidopsis* and *Amborella*, but all lignin biosynthesis genes are conserved among Arundinoideae and rice (Figure 4.3; Figure 4.4). Many previous papers report about the biological function of lignin biosynthesis genes. For instance, 4CL was reported in rice to function in reducing lignin content, affecting regulation of development of both rice and a pathogenic fungus (Liu et al., 2017); CCR gene family members were related with reduction of lignin content and changes in lignin deposition (Giordano et al., 2014); CAD and COMT were two important gene families in the lignin biosynthesis pathway, which can be used for modification of lignin structure and content and are potentially useful for improving the production of biomass (Acker et al., 2017; Trabucco et al., 2013). CCoAOMT was important for caffeoyl CoA methylation and hydroxycinnamates 5-methoxylation, which can be used for reduction of lignin content (Zhong et al., 2000), PAL gene family have been found to function in development, normal growth and response to abiotic stress in *Arabidopsis* (Huang et al., 2010). HCT gene family members are also involved in lignin biosynthesis pathway and responded to different environmental stress (Chowdhury et al., 2012); C4H, C3H and F5H genes encoding cytochrome P450 proteins are involved in lignin biosynthesis, content and structure and they can be used for down-regulation of lignin amount and to modulate its quality (Reddy et al., 2005; Goujon et al., 2003). Both of these lignin biosynthesis gene families have great potential for improving the production of *A. donax* biomass, such as γ -irradiation mutagenesis could be used in these genes modification for genetic improvement of biomass *A. donax* (Valli et al., 2017).

## 4.6 Conclusion

In summary, the putative proteins of CesA/Csl, CAD, CCoAOMT, 4CL, CCR, PAL, C4H, HCT, COMT, C3H and F5H gene families were first identified and characterized in Arundinoidea species by computational approaches. Orthologous groups identification and evolutionary analysis of the cellulose and lignin biosynthesis gene families have not been carried out yet, but they hold the promise to help in the elucidation of the evolution of lignification in Arundinideae. These gene families will further provide a good opportunity to establish mutagenesis-based reverse genetics and functional genomics approaches in biomass *A. donax*.

# Conclusion of the Thesis

Dwindling fossil fuel reserves and global climate change urgently require the development of environmental-friendly renewable sources of energy. Biomass can reduce the dependence from petrol and coal by providing a nearly carbon-neutral source of energy. *Arundo donax* L., is a perennial $C_3$ grass with fast growth, which is considered one of the important next generation bioenergy crops. This thesis focus on microRNAs identification from *Arundo donax* and comparative analysis of microRNAs among taxa of the *Arundo* genus, molecular evolution analysis and phylogenomic reconstruction of the relationships among *Arundo* species. For these purposes, a computational step-by-step workflow analysis for the identification of microRNAs and their targets was applied, and a phylogenetic framework carried out for exploring evolutionary relationships and analyze molecular evolution among *Arundo* species.

In the first chapter, the first time *in silico* identification of miRNAs and their targets in *Arundo donax* is presented, which was recently published in Scientific Reports (Jike et al., 2018). The evolutionary conservation of miRNA in plants is a powerful tool in miRNA identification using transcriptome data. The study identified 141 miRNAs belonging to 14 families and 462 high-confidence predicted targets leveraging on the reference transcriptome of *Arundo donax*. Gene ontology functional annotation showed that most putative miRNA targets may function in biological regulation, development and reproductive process. The analysis of position-specific nucleotide preferences showed that dominance of uracil at the first position of the 5' terminus may play an important role in miRNA biogenesis or RISC formation, while preference for cytosine at position 19 seems to be relevant for targeting RISC or Dicer-mediated cleavage to specific sites in pre-miRNAs. The capacity of *Arundo donax* to withstand viral infections may be related to the possible expansion of the MIR444 familiy in this species. The conserved miRNA families targets predicted in *Arundo donax* also had homologs in other species, confirming the overall conservation of miRNA targets among monocots and eudicots. This work, however, identified for the first time also novel targets (experimentally validated in Jike et al., 2018) suggesting a still incomplete elucidation of microRNA functions in Poaceae and/or ongoing evolution of targets within the family. Some identified miRNA, such as MIR172 may be employed as a useful tools to modulate flowering time in *Arundo donax*. The set of miRNAs identified in this study will pave the road to further miRNA research in bioenergy crop *Arundo donax* and hopefully contribute towards a better understanding of miRNA-mediated gene regulatory processes in other bioenergy crops.

In the second chapter, computational approaches and comparative analysis were used for identifying miRNAs and their targets, as these targets play an important role in understanding gene regulation. In the study (which already resulted in a manuscript now being finalized), a total of 235 miRNAs belonging to 37 miRNA families and 175 high-confidence targets were identified from *de novo* assembled *Arundo* leaf transcriptomes by using a computational pipeline. Comparative analysis of miRNA families expression among *Arundo* species showed that MIRNA444, MIR167, MIR159 and MIR162 are universally expressed among *Arundo* species. Phylogenetic analysis based on the highly conserved miRNA159 family indicated that different miRNAs

evolve with different rates in the *Arundo* genus, and confirmed that miRNAs are evolutionarily conserved in Arundinoideae. Gene functional annotation showed that most of putative targets act as transcription regulators, but they also function in metabolic processes, reproduction and response to various environmental cues. In this study, comparative analyses identified in most cases conserved miRNA and their homologous targets in *Arundo* taxa, suggesting that conserved miRNA regulate homologous targets at the conserved target sites among different species. These findings boost our understanding of miRNAs in the *Arundo* genus and the functional annotation of predicted miRNA targets may help to understand the mechanism of response to different environmental stresses, and have the potential to be further utilized for controlling secondary metabolism in order to improve the production and fermentation efficiency of biomass *Arundo donax* and other *Arundo* species.

In the third chapter, the first report is provided on the accurate reconstruction of the relationships among *Arundo* genus by means of a phylogenomic approach using 150 one-to-one orthologous genes. The study (which resulted in a third manuscript now nearly ready for submission) refined *Arundo formosana* as sister to the other members of the *Arundo* genus. It further confirmed the close relationship to *Arundo plinii* of *Arundo collina* as well as the close relationship to *Arundo donax* of *Arundo macrophylla*. Additionally, it validated the close evolutionary relationship of *Arundo donaciformis* to *Arundo plinii* and *Arundo collina.* In agreement to previous indications, the study points to an Eastern Asia origin of the *Arundo* species; after *A. formosana*, the most basal species in the genus are *Arundo donax* and *Arundo micrantha*, which first underwent differentiation and immigrated into middle Asia, then spread into Mediterranean. From here the invasive *A. donax* Mediterranean clone characterized in this study further spread all over the world in historical time by human intervention. Interestingly, *Arundo donaciformis* and *Arundo plinii* originated only after the invasive *A. donax* clone, confirming that human-driven dispersion of plant species is one of the major causes of biological invasions. There are many crops which are relatively recent polyploids, and thus the study of chromosome number evolution is useful to improve our understanding of crops improvement. In this study, probabilistic models suggested that the ancestral haploid chromosome number of *Arundo* was 36 and that repeated demi-duplication events were responsible for the chromosome number evolution in *Arundo* species. They further support that the origin of the *Arundo donax* clone is probably the result of fusion among unreduced gametes with different ploidy. However, it is still difficult to exactly identify any candidate parent species from this study, and this represents an important field of future investigations. Finally, the study identified some genes under positive selection, which provide valuable insights into adaptive selection mechanisms in the *Arundo* genus at the sequence level and will be valuable candidates for future functional validation.

In the last chapter, *in silico* identification and comparative analysis of lignin and cellulose biosynthesis gene families across the Arundinoideae (Poaceae) was carried out. A total of 741 CesA/Csl protein sequences and a total of 1118 lignin biosynthetic genes from 10 different gene families were identified. Phylogenetic analysis of CesA/Csl

proteins showed that CesA/Csl genes classified into 8 clades including CSLA, CSLC, CSLD, CSLE, CSLF, CSLH and CSLJ subfamilies and CESA gene family in Arundinoidea, and the phylogenetic tree also showed that CSLA and CSLC subfamilies constitute an independent lineage, indicating that they probably originated from a separate ancestral duplication event. Phylogenetic analysis using each lignin biosynthesis gene family showed that these genes were highly divergent between eudicots and monocots. Phylogenetic reconstruction of C3H, F5H and PAL proteins showed clear separation among *Amborella*, eudicots and monocots, indicating that these genes might have experienced expansion after species differentiation. The other reconstructions of phylogenetic trees of lignin biosynthesis proteins showed that these gene families divided early during evolution of land plants into different groups, indicating that diverse and conserved functions might exist in these genes families.

The comparative approach here proposed and the high-throughput transcriptomes sequencing utilized in this dissertation extends the analyses carried out in this focal species by introducing aspects of comparative genomics within the Arundinoideae. The dissection of the patterns of evolution in the *Arundo* genus will support ongoing efforts to establish reverse genetics and functional genomics approaches in *Arundo donax*, thus contributing to provide promising candidate genes for the improvement of this biomass species.

# References

Adams, Mark D, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, et al. 2000. “The Genome Sequence of Drosophila Melanogaster.” *Science* 287 (5461): 2185–95.

Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J.E. Baird, N. Bierne, J. Boughman, et al. 2013. “Hybridization and Speciation.” *Journal of Evolutionary Biology* 26 (2): 229–46.

Acker, Rebecca Van, Annabelle Déjardin, Sandrien Desmet, Lennart Hoengenaert, Ruben Vanholme, Kris Morreel, Francoise Laurans, et al. 2017. “Different Metabolic Routes for Coniferaldehyde and Sinapaldehyde with CINNAMYL ALCOHOL DEHYDROGENASE1 Deficiency.” *Plant Physiology* 175: 1018–39.

Adai, Alex, Cameron Johnson, Sizolwenkosi Mlotshwa, Sarah Archer-evans, Varun Manocha, Vicki Vance, and Venkatesan Sundaresan. 2005. “Computational Prediction of MiRNAs in Arabidopsis Thaliana.” *Genome Research* 15: 78–91.

Alves-Carvalho, Susete, Grégoire Aubert, Sébastien Carrère, Corinne Cruaud, Anne Lise Brochot, Françoise Jacquin, Anthony Klein, et al. 2015. “Full-Length de Novo Assembly of RNA-Seq Data in Pea (Pisum Sativum L.) Provides a Gene Expression Atlas and Gives Insights into Root Nodulation in This Species.” *Plant Journal* 84 (1): 1–19.

Amin, Farrukh Raza, Habiba Khalid, Han Zhang, Sajidu Rahman, Ruihong Zhang, Guangqing Liu, and Chang Chen. 2017. “Pretreatment Methods of Lignocellulosic Biomass for Anaerobic Digestion.” *AMB Express* 7 (72): 1–12.

Angelini, Luciana G, Lucia Ceccarini, Nicoletta Nassi o Di Nasso, and Enrico Bonari. 2009. “Comparison of Arundo Donax L. and Miscanthus x Giganteus in a Long-Term Field Experiment in Central Italy: Analysis of Productive Characteristics and Energy Balance.” *Biomass and Bioenergy* 33 (4): 635–43.

Anisimova, Maria, and Olivier Gascuel. 2006. “Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative.” *Systematic Biology* 55 (4): 539–52.

Archak, Sunil, and J. Nagaraju. 2007. “Computational Prediction of Rice (Oryza Sativa) MiRNA Targets.” *Genomics, Proteomics and Bioinformatics* 5 (3–4): 196–206.

Axtell, Michael J and David P. Bartel. 2005. “Antiquity of MicroRNAs and Their Targets in Land Plants.” *The Plant Cell* 17 (6): 1658–73.

Axtell, Michael J., and John L. Bowman. 2008. “Evolution of Plant MicroRNAs and Their Targets.” *Trends in Plant Science* 13 (7): 343–49.

Barrera-Figueroa, Blanca E., Lei Gao, Ndeye N. Diop, Zhigang Wu, Jeffrey D. Ehlers, Philip A. Roberts, Timothy J. Close, Jian Kang Zhu, and Renyi Liu. 2011. “Identification and Comparative Analysis of Drought-Associated MicroRNAs in Two Cowpea Genotypes.” *BMC Plant Biology*, no. 1: 1–11.

Barrera-Figueroa, Blanca E., Lei Gao, Zhigang Wu, Xuefeng Zhou, Jianhua Zhu, Hailing Jin, Renyi Liu, and Jian Kang Zhu. 2012. "High Throughput Sequencing Reveals Novel and Abiotic Stress-Regulated MicroRNAs in the Inflorescences of Rice." *BMC Plant Biology* 12 (1): 1–11.

Barrero, Roberto A, Felix D Guerrero, Paula Moolhuijzen, John A Goolsby, Jason Tidwell, Stanley E Bellgard, and Matthew I Bellgard. 2015. "Data in Brief Shoot Transcriptome of the Giant Reed , Arundo Donax." *Data in Brief* 3: 1–6.

Bartel, David P, Rosalind Lee, and Rhonda Feinbaum. 2004. "MicroRNAs : Genomics , Biogenesis , Mechanism , and Function." *Cell* 116: 281–97.

Baumberger, N., and D. C. Baulcombe. 2005. "Arabidopsis ARGONAUTE1 Is an RNA Slicer That Selectively Recruits MicroRNAs and Short Interfering RNAs." *Proceedings of the National Academy of Sciences* 102 (33): 11928–33.

Bo, Xiaochen, and Shengqi Wang. 2005. "TargetFinder : A Software for Antisense Oligonucleotide Target Site Selection Based on MAST and Secondary Structures of Target MRNA" *Bioinformatics* 21 (8): 1401–1402.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Genome Analysis Trimmomatic : A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

Bonanno, G, and R. Lo Giudice. 2010. "Heavy Metal Bioaccumulation by the Organs of Phragmites Australis (Common Reed) and Their Potential Use as Contamination Indicators." *Ecological Indicators* 10 (3): 639–45.

Bonnet, Eric, Jan Wuyts, Pierre Rouzé, and Yves Van de Peer. 2004. "Evidence That MicroRNA Precursors, Unlike Other Non-Coding RNAs, Have Lower Folding Free Energies than Random Sequences." *Bioinformatics* 20 (17): 2911–17.

Boualem, Adnane, Philippe Laporte, Mariana Jovanovic, Carole Laffont, Julie Plet, Jean-philippe Combier, Andreas Niebel, Martin Crespi, and Florian Frugier. 2008. "MicroRNA166 Controls Root and Nodule Development in Medicago Truncatula" *The Plant Journal* 54 (5): 876-887.

Bragato, Claudia, Hans Brix, and Mario Malagoli. 2006. "Accumulation of Nutrients and Heavy Metals in Phragmites Australis (Cav.) Trin. Ex Steudel and Bolboschoenus Maritimus (L.) Palla in a Constructed Wetland of the Venice Lagoon Watershed." *Environmental Pollution* 144 (3): 967–75.

Breese, Marcus R, and Yunlong Liu. 2013. "NGSUtils : A Software Suite for Analyzing and Manipulating next-Generation Sequencing Datasets." *Bioinformatics* 29 (4): 494–96.

Brochmann, Christian, Liv Borgen, and Odd E. Stabbetorp. 2000. "Multiple Diploid Hybrid Speciation of the Canary Island Endemic Argyranthemum Sundingii (Asteraceae)." *Plant Systematics and Evolution* 220 (1–2): 77–92.

Bruno, Barbosa, Sara Boléo, Sarah Sidella, Jorge Costa, Maria Paula Duarte, Benilde Mendes, Salvatore L. Cosentino, and Ana Luisa Fernando. 2015. "Phytoremediation of Heavy Metal-Contaminated Soils Using the Perennial Energy Crops Miscanthus Spp. and Arundo Donax L." *Bioenergy Research* 8 (4): 1500–1511.

Bucci, A, E Cassani, M Landoni, E Cantaluppi, and R Pilu. 2013. "Analysis of Chromosome Number and Speculations on the Origin of Arundo Donax L . ( Giant Reed )." *Cytology and Genetics* 47 (4): 237–41.

Buggs, Richard J.A., Simon Renny-Byfield, Michael Chester, Ingrid E. Jordon-Thaden, Lyderson Facio Viccini, Srikar Chamala, Andrew R. Leitch, et al. 2012. "Next-Generation Sequencing and Genome Evolution in Allopolyploids." *American Journal of Botany* 99 (2):372-382.

Burge, Christopher B., and Samuel Karlin. 1998. "Finding the Genes in Genomic DNA." *Current Opinion in Structural Biology* 8 (3): 346–54.

Burki, Fabien, Noriko Okamoto, and Patrick J Keeling. 2012. "The Evolutionary History of Haptophytes and Cryptophytes : Phylogenomic Evidence for Separate Origins." *Proc. R. Soc. B* 279: 2246–54.

Burks, Elizabeth A, Paula P Bezerra, Hahn Le, Daniel R Gallie, and Karen S Browning. 2001. "Plant Initiation Factor 3 Subunit Composition Resembles Mammalian Initiation Factor 3 and Has a Novel Subunit*." *The Journal of Biological Chemistry* 276 (3): 2122–31.

Burton, Rachel a, Sarah M Wilson, Maria Hrmova, Andrew J Harvey, Neil J Shirley, Anne Medhurst, Bruce a Stone, Edward J Newbigin, Antony Bacic, and Geoffrey B Fincher. 2006. "Cellulose Synthase – Like CslF Genes Mediate the Synthesis of Cell Wall(1,3;1,4)-b-D-Glucans." *Science* 311: 1940–42.

Buschiazzo, Emmanuel, Carol Ritland, Jörg Bohlmann, and Kermit Ritland. 2012. "Slow but Not Low : Genomic Comparisons Reveal Slower Evolutionary Rate and Higher DN / DS in Conifers Compared to Angiosperms." *BMC Evolutionary Biology* 12 (8): 1–14.

C. Vella Monica and J. Slack Frank. 2005. "C. Elegans MicroRNAs." *WormBook*, 1–9.

Calheiros, Cristina S C, Paula V B Quitério, Gabriela Silva, Luís F C Crispim, Hans Brix, Sandra C. Moura, and Paula M L Castro. 2012. "Use of Constructed Wetland Systems with Arundo and Sarcocornia for Polishing High Salinity Tannery Wastewater." *Journal of Environmental Management* 95 (1): 66–71.

Calviño, Martín, Rémy Bruggmann, and Joachim Messing. 2011. "Characterization of the Small RNA Component of the Transcriptome from Grain and Sweet Sorghum Stems " *BMC Genomics* 12 (1): 356.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, et al. 2011. "Whole-Genome Sequencing of

Multiple Arabidopsis Thaliana Populations." *Nature Genetics* 43 (10): 956–65.

Castresana, J. 2000. "Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis." *Mol. Biol. Evol* 17 (4): 540–52.

Chan, Cheong Xin, and Mark A Ragan. 2013. "Next-Generation Phylogenomics." *Biology Direct* 8 (3): 1–6.

Chan, Cheong Xin, Robert G Beiko, Aaron E Darling, and Mark A Ragan. 2009. "Lateral Transfer of Genes and Gene Fragments in Prokaryotes." *Genome. Biol. Evol*: 429–38.

Chen, Haodong, Yunping Shen, Xiaobo Tang, Lu Yu, Jia Wang, Lan Guo, and Yu Zhang. 2006. "Arabidopsis CULLIN4 Forms an E3 Ubiquitin Ligase with RBX1 and the CDD Complex in Mediating Light Control of Development." *The Plant Cell* 18: 1991–2004.

Chen, Min, Hai Bao, Qiuming Wu, and Yanwei Wang. 2015. "Transcriptome-Wide Identification of MiRNA Targets under Nitrogen Deficiency in Populus Tomentosa Using Degradome Sequencing" *Int J Mol Sci* 16(6):13937-58.

Chen, Ting-wen, Ruei-chi Richie Gan, Timothy H Wu, Po-jung Huang, Cheng-yang Lee, Yi-ywan M Chen, Che-chun Chen, and Petrus Tang. 2012. "FastAnnotator- an Efficient Transcript Annotation Web Tool." *BMC Genomics* 13: 1–8.

Chen, Zhizhong, Xuhui Hong, Hairong Zhang, Youqun Wang, Xia Li, Jian Kang Zhu, and Zhizhong Gong. 2005. "Disruption of the Cellulose Synthase Gene, AtCesA8/IRX1, Enhances Drought and Osmotic Stress Tolerance in Arabidopsis." *Plant Journal* 43 (2): 273–83.

Chowdhury, Emran Md, Bo Sung Choi, Sang Un Park, Hyoun Sub Lim, and Hanhong Bae. 2012. "Transcriptional Analysis of Hydroxycinnamoyl Transferase (HCT) in Various Tissues of Hibiscus Cannabinus in Response to Abiotic Stress Conditions." *Plant OMICS* 5 (3): 305–13.

Cocuron, Jean-Christophe, Olivier Lerouxel, Georgia Drakakaki, Ana P Alonso, Aaron H Liepman, Kenneth Keegstra, Natasha Raikhel, and Curtis G Wilkerson. 2007. "A Gene from the Cellulose Synthase-like C Family Encodes a Beta-1,4 Glucan Synthase." *Proceedings of the National Academy of Sciences of the United States of America* 104 (20): 8550–55.

Collins, Lesley J, Claudia Voelckel, Patrick J Biggs, and Simon Joly. 2008. "An approach to transcriptome analysis of non-model organisms using short-read sequences" *Genome Informatics* 21: 3–14.

Cui, Jie, Chenjiang You, and Xuemei Chen. 2017. "The Evolution of MicroRNAs in Plants." *Current Opinion in Plant Biology* 35:61-67

Cui, Liying, P Kerr Wall, James H Leebens-mack, Bruce G Lindsay, Douglas E Soltis, Jeff J Doyle, Pamela S Soltis, et al. 2006. "Widespread Genome Duplications

throughout the History of Flowering Plants." *Genome Research* 16: 738–49.

Cui, Qinghua, Youlian Pan, Enrico O Purisima, and Edwin Wang. 2007. "MicroRNAs Preferentially Target the Genes with High Transcriptional Regulation Complexity" *Biochem Biophys Res Commun* 352(3):733-8.

Dai, Xinbin, and Patrick Xuechun Zhao. 2011. "PsRNATarget : A Plant Small RNA Target Analysis Server" *Nucleic Acids Res* 39: 155–59.

Dalal, Ankit, and Ankur Atri. 2014. "An Introduction to Sequence and Series." *International Journal of Research* 1 (10): 1286–92.

Danin, A. 2004. "Arundo ( Gramineae ) in the Mediterranean Reconsidered." *Willdenowia* 34 (2): 361–369.

Danin, A., Raus, Th. & Scholz, H. 2002. "Contribution to the Flora of Greece: A New Species of Arundo (Poaceae)." *Willdenowia* 32: 191–94.

Darriba, Diego, Guillermo L Taboada, Ramón Doallo, and David Posada. 2011. "ProtTest 3 : Fast Selection of Best-Fit Models of Protein Evolution." *Bioinformatics* 27 (8): 1164–65.

David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P., Colin L. Barnes Smith, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456: 53–59.

Davis, Brandi N, and Akiko Hata. 2009. "Regulation of MicroRNA Biogenesis: A MiRiad of Mechanisms." *Cell Communication and Signaling* 7 (18): 1–22.

Dehury, Budheswar, Debashis Panda, Jagajjit Sahu, Mousumi Sahu, Kishore Sarma, Madhumita Barooah, Priyabrata Sen, and Mahendra Modi. 2013. "In Silico Identification and Characterization of Conserved MiRNAs and Their Target Genes in Sweet Potato (Ipomoea Batatas L.) Expressed Sequence Tags (ESTs)." *Plant Signaling & Behavior* 8 (12): 1–13.

Delmer, Deborah P. 1999. "CELLULOSE BIOSYNTHESIS: Exciting Times for A Difficult Field of Study." *Annual Review of Plant Physiology and Plant Molecular Biology* 50 (1): 245–76.

Devi, Karam Jayanandi, Sreejita Chakraborty, Bibhas Deb, and Ravi Rajwanshi. 2016. "Computational Identification and Functional Annotation of MicroRNAs and Their Targets from Expressed Sequence Tags (ESTs) and Genome Survey Sequences (GSSs) of Coffee (Coffea Arabica L.)." *Plant Gene* 6: 30–42.

Dhir, Sarwan, Kaye Knowles, C Livia Pagan, Justin Mann, and Shireen Dhir. 2010. "Optimization and Transformation of Arundo Donax L . Using Particle Bombardment" *African Journal of Biotechnology* 9 (39): 6460–6469.

Doblin, M. S., F. A. Pettolino, S. M. Wilson, R. Campbell, R. A. Burton, G. B. Fincher, E. Newbigin, and A. Bacic. 2009. "A Barley Cellulose Synthase-like CSLH Gene

Mediates (1,3;1,4)-D-Glucan Synthesis in Transgenic Arabidopsis." *Proceedings of the National Academy of Sciences* 106 (14): 5996–6001.

Donald L. Klass. 2004. "Biomass for Renewable Energy and Fuels." *The Encyclopedia of Energy*.

DONG, QING-HUA, Jian Han, Aying Li, Hong Liu, Xicheng Wen, Mizhen Zhao, Nadira Bilkish Korir, Nicholas Kibet Korir, Chen Wang, and Jinggui Fang. 2012. "Computational Identification of MicroRNAs in the Strawberry Expressed Sequence Tags and Validation of Their Precise Sequences by MiR-RACE." *Journal of Heredity* 103 (2): 268–77.

Edgar, Robert C. 2004. "MUSCLE : Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97.

Egan, Ashley N., Jessica Schlueter, and David M. Spooner. 2012. "Applications of Next-Generation Sequencing in Plant Biology." *American Journal of Botany* 99 (2): 175–85.

Emms, David M, and Steven Kelly. 2015. "OrthoFinder : Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16 (157): 1–14.

Essoussi, Nadia, Khaddouja Boujenfa, and Mohamed Limam. 2008. "A Comparison of MSA Tools." *Bioinformation* 2 (9): 452–55.

Fahlgren, Noah, Miya D Howell, Kristin D Kasschau, Elisabeth J Chapman, Christopher M Sullivan, Jason S Cumbie, and Scott A Givan. 2007. "High-Throughput Sequencing of Arabidopsis MicroRNAs : Evidence for Frequent Birth and Death of MIRNA Genes." *Plos One* 2 (2): 1–14.

Foyle, Thomas, Linda Jennings, and Patricia Mulcahy. 2007. "Compositional Analysis of Lignocellulosic Materials: Evaluation of Methods Used for Sugar Analysis of Waste Paper and Straw." *Bioresource Technology* 98 (16): 3026–36.

Fu, Chunxiang, Ramanjulu Sunkar, Chuanen Zhou, Hui Shen, Ji-yi Zhang, Jessica Matts, Jennifer Wolf, et al. 2012. "Overexpression of MiR156 in Switchgrass ( Panicum Virgatum L .) Results in Various Morphological Alterations and Leads to Improved Biomass Production." *Plant Biotechnology Journal* 10 (4): 443-52.

Fu, Yuan, Michele Poli, Gaurav Sablok, Bo Wang, Yanchun Liang, Nicola La Porta, and Violeta Velikova. 2016. "Dissection of Early Transcriptional Responses to Water Stress in Arundo Donax L . by Unigene-based RNA-seq." *Biotechnology for Biofuels* 9 (54): 1–18.

G.W. Heil and M. Bruggink 1987. "Competition for Nutrients between CMluna Vulgaris (L.) Hull and Molinia Caerulea (L.) Moench." *Oecologia* 73: 105–7.

Garcia-Seco, Daniel, Yang Zhang, Francisco J. Gutierrez-Maero, Cathie Martin, and Beatriz Ramos-Solano. 2015. "RNA-Seq Analysis and Transcriptome Assembly for

Blackberry (Rubus Sp. Var. Lochness) Fruit." *BMC Genomics* 16 (5): 1-11.

German, Marcelo A, Manoj Pillay, Dong-Hoon Jeong, Amit Hetawal, Shujun Luo, Prakash Janardhanan, Vimal Kannan, et al. 2008. "Global Identification of MicroRNA–target RNA Pairs by Parallel Analysis of RNA Ends." *Nature Biotechnology* 26 (8): 1–7.

Giordano, Andrea, Zhiqian Liu, Stephen N. Panter, Adam M. Dimech, Yongjin Shang, Hewage Wijesinghe, Karen Fulgueras, et al. 2014. "Reduced Lignin Content and Altered Lignin Composition in the Warm Season Forage Grass Paspalum Dilatatum by Down-Regulation of a Cinnamoyl CoA Reductase Gene." *Transgenic Research* 23 (3): 503–17.

Giudicianni, Paola, Giuseppe Cardone, Giancarlo Sorrentino, and Raffaele Ragucci. 2014. "Hemicellulose, Cellulose and Lignin Interactions on Arundo Donax Steam Assisted Pyrolysis." *Journal of Analytical and Applied Pyrolysis* 110 (1): 138–46.

Glick, Lior, and Itay Mayrose. 2014. "ChromEvol : Assessing the Pattern of Chromosome Number Evolution and the Inference of Polyploidy along a Phylogeny." *Mol. Biol. Evol* 31 (7): 1914–22.

Goujon, Thomas, Richard Sibout, Aymerick Eudes, John MacKay, and Lise Jouanin. 2003. "Genes Involved in the Biosynthesis of Lignin Precursors in Arabidopsis Thaliana." *Plant Physiology and Biochemistry* 41 (8): 677–87.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52.

Griffiths-jones, Sam, Harpreet Kaur Saini, Stijn Van Dongen, and Anton J Enright. 2008. "MiRBase : Tools for MicroRNA Genomics." *Nucleic Acids Research* 36: 154–58.

Guindon, Stéphane, Jean Franois Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." *Systematic Biology* 59 (3): 307–21.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, D Philip, Joshua Bowden, Matthew Brian Couger, et al. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq: Reference Generation and Analysis with Trinity." *Nat Protoc* 8 (8): 1–43.

Hall Tom , Ibis biosciences, Carlsbad, Ca. 2011. "BioEdit : An Important Software for Molecular Biology." *GERF Bulletin of Biosciences* 2 (1): 60–61.

Hamann, Thorsten, Erin Osborne, Heather L. Youngs, Julie Misson, Laurent Nussaume, and Chris Somerville. 2004. "Global Expression Analysis of CESA and CSL Genes in Arabidopsis." *Cellulose* 11 (3–4): 279–86.

Hardion, Laurent, Alex Baumel, Régine Verlaque, and Bruno Vila. 2014(a). "Distinct Evolutionary Histories of Lowland Biota on Italian and Balkan Peninsulas Revealed by the Phylogeography of Arundo Plinii (Poaceae)." *Journal of Biogeography* 41 (11): 2150–61.

Hardion, Laurent, Carole Barthélémy, Jean Noël Consales, Perrine Gauthier, John D Thompson, Régine Verlaque, and Bruno Vila. 2015. "An Endangered Reed, Arundo Donaciformis, in a Dynamic Urban Environment: The Need for Interdisciplinary Conservation Proposals." *Journal for Nature Conservation* 26: 20–27.

Hardion, Laurent, Re´gine Verlaque, Marcela Rosato, Josep A Rossello, and Bruno Vila. 2015. "Impact of Polyploidy on Fertility Variation of Mediterranean Arundo L. (Poaceae)." *Comptes Rendus Biologies* 338: 298–306.

Hardion, Laurent, Re´gine Verlaque, Kristin Saltonstall, Agathe Leriche1, and Bruno Vila. 2014. "Origin of the Invasive Arundo Donax ( Poaceae ): A Trans-Asian Expedition in Herbaria" *Annals of Botany* 114(3): 455–462.

Hardion, Laurent, Régine Verlaque, Alex Baumel, Marianick Juin, and Bruno Vila. 2012. "Revised Systematics of Mediterranean Arundo (Poaceae ) Based on AFLP Fingerprints and Morphology." *Taxon* 61 (6): 1217–26.

Hardion, Laurent, Régine Verlaque, Kristin Saltonstall, Agathe Leriche, and Bruno Vila. 2014(b). "Origin of the Invasive Arundo Donax (Poaceae): A Trans-Asian Expedition in Herbaria." *Annals of Botany* 114 (3): 455–62.

Hardion, Laurent, Régine Verlaque, Martin W Callmander, and Bruno Vila. 2012. "Arundo Micrantha Lam. (Poaceae), the Correct Name for Arundo Mauritanica Desf. And Arundo Mediterranea Danin." *Candollea* 67 (1): 131–35.

Harvey, Michael P., and Mark H. Brand. 2002. "Division Size and Shade Density Influence Growth and Container Production of Hakonechloa Macra Makino 'Aureola.'" *Hort Science* 37 (1): 196–99.

Hazen, Samuel P, John S Scott-Craig, and Jonathan D Walton. 2002. "Cellulose Synthase-like (CSL) Genes of Rice." *Plant Physiology* 128: 336–40.

Holland, Neta, Doron Holland, Tim Helentjaris, Kanwarpal S. Dhugga, Beatriz Xoconostle-Cazares, and Deborah P. Delmer. 2000. "A Comparative Analysis of the Plant Cellulose Synthase ( CesA ) Gene Family." *Plant Physiology* 123 (4): 1313–24.

Hrdlickova, Radmila, Masoud Toloue, and Bin Tian. 2017. "RNA - Seq Methods for Transcriptome Analysis." *Wiley Interdiscip Rev RNA* 8 (1): 1–24.

Huang, J., M. Gu, Z. Lai, B. Fan, K. Shi, Y.-H. Zhou, J.-Q. Yu, and Z. Chen. 2010. "Functional Analysis of the Arabidopsis PAL Gene Family in Plant Growth, Development, and Response to Environmental Stress." *Plant Physiology* 153 (4): 1526–38.

Huang, Xiaoqiu, and Anup Madan. 1999. “CAP3 : A DNA Sequence Assembly Program.” *Genome Research* 9: 868–77.

Ingwell, Laura L, Robert Zemetra, Carol Mallory-smith, and A. Bosque-Pe ŕez Nilsa. 2014. “Arundo Donax Infection with Barley Yellow Dwarf Virus Has Implications for Biofuel Production and Non-Managed Habitats” *Biomass and Bioenergy* 66: 426–33.

J. A. BUGGS r, Renny-Byfield S., et al. 2012. “Next-generation sequencing and genome evolution in allopolyploids.” *American Journal of Botany* 99 (2): 372–82.

Jike, Wuhe, Gaurav Sablok, Giorgio Bertorelle, Mingai Li, and Claudio Varotto. 2018. “In Silico Identification and Characterization of a Diverse Subset of Conserved MicroRNAs in Bioenergy Crop Arundo Donax L.” *Scientific Reports* 8: 1–13.

Jones-Rhoades, Matthew W., David P. Bartel, and Bonnie Bartel. 2006. “MicroRNAs AND THEIR REGULATORY ROLES IN PLANTS.” *Annual Review of Plant Biology* 57 (1): 19–53.

Jönsson, Leif J., Björn Alriksson, and Nils Olof Nilvebrant. 2013. “Bioconversion of Lignocellulose: Inhibitors and Detoxification.” *Biotechnology for Biofuels* 6 (1): 1–10.

Kanehisa, Minoru, and Peer Bork. 2003. “Bioinformatics in the Post-Sequence Era.” *Nature Genetics* 33: 305–10.

Kaur, Simerjeet, Kanwarpal S. Dhugga, Robin Beech, and Jaswinder Singh. 2017. “Genome-Wide Analysis of the Cellulose Synthase-like (Csl) Gene Family in Bread Wheat (Triticum Aestivum L.).” *BMC Plant Biology* 17 (1): 1–17.

Kerstens, Hindrik H.D., Richard P.M.A. Crooijmans, Albertine Veenendaal, Bert W. Dibbits, Thomas F.C. Chin-A-Woeng, Johan T. den Dunnen, and Martien A.M. Groenen. 2009. “Large Scale Single Nucleotide Polymorphism Discovery in Unsequenced Genomes Using Second Generation High Throughput Sequencing Technology: Applied to Turkey.” *BMC Genomics* 10: 1–11.

Kidner, Catherine A., and Robert A. Martienssen. 2005. “The Developmental Role of MicroRNA in Plants.” *Current Opinion in Plant Biology* 8 (1): 38–44.

Kocot, Kevin M., Johanna T. Cannon, Christiane Todt, Mathew R. Citarella, Andrea B. Kohn, Achim Meyer, Scott R. Santos, et al. 2011. “Phylogenomics Reveals Deep Molluscan Relationships.” *Nature* 477: 452–56.

Komolwanich, Tidarat, Patomwat Tatijarern, Sirirat Prasertwasu, Darin Khumsupan, Thanyalak Chaisuwan, Apanee Luengnaruemitchai, and Sujitra Wongkasemjit. 2014. “Comparative Potentiality of Kans Grass (Saccharum Spontaneum) and Giant Reed (Arundo Donax) as Lignocellulosic Feedstocks for the Release of Monomeric Sugars by Microwave/Chemical Pretreatment.” *Cellulose* 21 (3): 1327–40.

Kozomara, Ana, and Sam Griffiths-jones. 2014. “MiRBase : Annotating High Confidence

MicroRNAs Using Deep Sequencing Data." *Nucleic Acids Research* 42: 68–73.

Kumar, Parveen, Diane M. Barrett, Michael J. Delwiche, and Pieter Stroeve. 2009. "Methods for Pretreatment of Lignocellulosic Biomass for Eficient Hydrolysis and Biofuel Production." *Industrial and Engineering Chemistry* 48 (8): 3713–29.

Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. "MEGA7 : Molecular Evolutionary Genetics Analysis Version 7 . 0 for Bigger Datasets." *Mol. Biol. Evol* 33 (7): 1870–74.

Kurihara, Y., and Y. Watanabe. 2004. "From The Cover: Arabidopsis Micro-RNA Biogenesis through Dicer-like 1 Protein Functions." *Proceedings of the National Academy of Sciences* 101 (34): 12753–58.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3):1-10.

Lee, Yoontae, Kipyoung Jeon, Jun-tae Lee, Sunyoung Kim, and V Narry Kim. 2002. "MicroRNA Maturation : Stepwise Processing and Subcellular Localization." *The EMBO Journal* Vol 21 (17): 4663–70.

Lee, Yoontae, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. 2004. "MicroRNA Genes Are Transcribed by RNA Polymerase II." *The EMBO Journal* 23 (20): 4051–60.

Lefort, Vincent, Jean Emmanuel Longueville, and Olivier Gascuel. 2017. "SMS: Smart Model Selection in PhyML." *Molecular Biology and Evolution* 34 (9): 2422–24.

Lemmon, Emily Moriarty, and Alan R Lemmon. 2013. "High-Throughput Genomic Data in Systematics and Phylogenetics." *Annual Review of Ecology, Evolution, and Systematics* 44 (1): 99–121.

Lenz, Dorina, Patrick May, and Dirk Walther. 2011. "Comparative Analysis of MiRNAs and Their Targets across Four Plant Species." *BMC Research Notes* 4 (483): 1–7.

Levesque-Tremblay Gabriel, Michel Havaux, and Francois Ouellet. 2009. "The Chloroplastic Lipocalin AtCHL Prevents Lipid Peroxidation and Protects Arabidopsis against Oxidative Stress." *The Plant Journal* 60: 691–702.

Lewandowski, Iris, Jonathan M.O. Scurlock, Eva Lindvall, and Myrsini Christou. 2003. "The Development and Current Status of Perennial Rhizomatous Grasses as Energy Crops in the US and Europe." *Biomass and Bioenergy* 25 (4): 335–61.

Li, Aili, and Long Mao. 2007. "Evolution of Plant MicroRNA Gene Families." *Cell Research* 17 (3): 212–18.

Li, Ruiqiang, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, et al. 2010. "De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing." *Genome Res.* 20 (2): 265–72.

Li, Ruiqiang, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, et al. 2010. “The Sequence and de Novo Assembly of the Giant Panda Genome.” *Nature* 463 (21): 311–17.

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit : A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59.

Li, Xiaoyan, Hongwu Bian, Dafeng Song, Shengyun Ma, Ning Han, Junhui Wang, and Muyuan Zhu. 2013. “Flowering Time Control in Ornamental Gloxinia (Sinningia Speciosa) by Manipulation of MiR159 Expression.” *Annals of Botany* 111 (5): 791–99.

Li, Yan, Sheng-li Zhao, Jin-lu Li, Xiao-hong Hu, He Wang, Xiao-long Cao, Yong-ju Xu, et al. 2017. “Osa-MiR169 Negatively Regulates Rice Immunity against the Blast Fungus Magnaporthe Oryzae.” *Frontiers in Plant Science* 8: 1–13.

Li, Yanjie, Yaru Fu, Lusha Ji, Changai Wu, and Chengchao Zheng. 2010. “Plant Science Characterization and Expression Analysis of the Arabidopsis Mir169 Family.” *Plant Science* 178 (3): 271–80.

Li, Yong Fang, Yun Zheng, Charles Addo-Quaye, Li Zhang, Ajay Saini, Guru Jagadeeswaran, Michael J. Axtell, Weixiong Zhang, and Ramanjulu Sunkar. 2010. “Transcriptome-Wide Identification of MicroRNA Targets in Rice.” *Plant Journal* 62 (5): 742–59.

Liepman, A. H., C. J. Nairn, W. G.T. Willats, I. Sorensen, A. W. Roberts, and K. Keegstra. 2007. “Functional Genomic Analysis Supports Conservation of Function Among Cellulose Synthase-Like A Gene Family Members and Suggests Diverse Roles of Mannans in Plants.” *Plant Physiology* 143 (4): 1881–93.

Lin, Wen Tzu, Chao Yuan Lin, and Wen Chieh Chou. 2006. “Assessment of Vegetation Recovery and Soil Erosion at Landslides Caused by a Catastrophic Earthquake: A Case Study in Central Taiwan.” *Ecological Engineering* 28 (1): 79–89.

Lindow, Morten, and Anders Krogh. 2005. “Computational Evidence for Hundreds of Non-Conserved Plant MicroRNAs.” *BMC Genomics* 6: 1–9.

Ling, Li-zhen, Shu-dong Zhang, Fan Zhao, Jin-long Yang, and Wen-hui Song. 2017. “Transcriptome-Wide Identification and Prediction of MiRNAs and Their Targets in Paris Polyphylla Var . Yunnanensis by High-Throughput Sequencing Analysis.” *Int. J. Mol. Sci.* 18 (219): 1–12.

Lissner, Jørgen, and Hans-henrik Schierup. 1997. “Effects of Salinity on the Growth of Phragmites Australis.” *Aquatic Botany* 55 (4): 247–260.

Little, Alan, Julian G. Schwerdt, Neil J. Shirley, Shi F. Khor, Kylie Neumann, Lisa A. O’Donovan, Jelle Lahnstein, et al. 2018. “Revised Phylogeny of the Cellulose Synthase Gene Superfamily: Insights into Cell Wall Evolution.” *Plant Physiology*

177 (3): 1124–41.

Liu, Hao, Zhenhua Guo, Fengwei Gu, Shanwen Ke, Dayuan Sun, Shuangyu Dong, Wei Liu, et al. 2017. "4-Coumarate-CoA Ligase-Like Gene OsAAE3 Negatively Mediates the Rice Blast Resistance, Floret Development and Lignin Biosynthesis." *Frontiers in Plant Science* 7: 1–13.

Liu, Qing, Tifeng Yang, Ting Yu, Shaohong Zhang, Xingxue Mao, and Bin Liu. 2017. "Integrating Small RNA Sequencing with QTL Mapping for Identification of MiRNAs and Their Target Genes Associated with Heat Tolerance at the Flowering Stage in Rice." *Frontiers in Plant Science* 8: 1–15.

Liu, Qingquan, Le Luo, and Luqing Zheng. 2018. "Lignins: Biosynthesis and Biological Functions in Plants." *International Journal of Molecular Sciences* 19 (2): 1–16.

Ma, Zhaorong, Ceyda Coruh, and Michael J. Axtell. 2010. "Arabidopsis Lyrata Small RNAs : Transient MIRNA and Small Interfering RNA Loci within the Arabidopsis Genus." *The Plant Cell* 22: 1090–1103.

Maher, Christopher A, Nallasivam Palanisamy, John C Brenner, Xuhong Cao, and Shanker Kalyana-sundaram. 2009. "Chimeric Transcript Discovery by Paired-End Transcriptome Sequencing." *PNAS* 106 (30): 12353–58.

Mallory, Allison C, and Hervé Vaucheret. 2006. "Functions of MicroRNAs and Related Small RNAs in Plants." *Nature Genetics* 38 (6S): S31–36.

Mariani, C., R. Cabrini, A. Danin, P. Piffanelli, A. Fricano, S. Gomarasca, M. Dicandilo, F. Grassi, and C. Soave. 2010. "Origin, Diffusion and Reproduction of the Giant Reed (Arundo Donax L.): A Promising Weedy Energy Crop." *Annals of Applied Biology* 157: 191–202.

Mascia, F, G Fenu, R Angius, and G Bacchetta. 2013. "Arundo Micrantha, a New Reed Species for Italy, Threatened in the Freshwater Habitat by the Congeneric Invasive A. Donax." *Plant Biosystems* 147 (3): 717–29.

MAYROSE, Itay, Michal S. barkel, and Sarah P. otto. 2010. "Probabilistic Models of Chromosome Number Evolution and the Inference of Polyploidy." *Syst. Biol* 59 (2): 132–44.

McGettigan, Paul A. 2013. "Transcriptomics in the RNA-Seq Era." *Current Opinion in Chemical Biology* 17 (1): 4–11.

McKendry P. 2002. "Energy Production from Biomass (Part 1): Overview of Biomass." *Bioresource Technology* 83 (July 2001): 37–46.

Michael, J. Axtell and David, P. Bartel. 2005. "Antiquity of MicroRNAs and Their Targets in Land Plants." *The Plant Cell* 17 (6): 1658–73.

Moorthie, Sowmiya, Alison Hall, and Caroline F. Wright. 2013. "Informatics and Clinical Genome Sequencing: Opening the Black Box." *Genetics in Medicine* 15 (3):

165–71.

Murrell, Ben, Joel O Wertheim, Sasha Moola, Thomas Weighill, Konrad Scheffler, and Sergei L Kosakovsky Pond. 2012. “Detecting Individual Sites Subject to Episodic Diversifying Selection.” *PLoS Genetics* 8 (7): 1–10.

Murrell, Ben, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Sergei L Kosakovsky Pond, and Konrad Scheffler. 2013. “FUBAR : A Fast , Unconstrained Bayesian AppRoximation for Inferring Selection.” *Mol. Biol. Evol* 30 (5): 1196–1205.

Murrell, Ben, Steven Weaver, Martin D Smith, Joel O Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, et al. 2015. “Gene-Wide Identification of Episodic Selection.” *Mol. Biol. Evol* 32 (5): 1365–71.

Mussatto, Si, and Ja Teixeira. 2010. “Lignocellulose as Raw Material in Fermentation Processes.” *Applied Microbiology an Microbial Biotechnology* 2: 897–907.

Nagalakshmi, Ugrappa, Karl Waern, and Michael Snyder. 2010. “RNA-Seq: A Method for Comprehensive Transcriptome Analysis.” *Current Protocols in Molecular Biology* Chapter 4:Unit 4.11.1-13.

Naik, S. N., Vaibhav V. Goud, Prasant K. Rout, and Ajay K. Dalai. 2010. “Production of First and Second Generation Biofuels: A Comprehensive Review.” *Renewable and Sustainable Energy Reviews* 14: 578–597.

Nazarov, Petr V, Susanne E Reinsbach, Arnaud Muller, Nathalie Nicot, Demetra Philippidou, Laurent Vallar, and Stephanie Kreis. 2013. “Interplay of MicroRNAs , Transcription Factors and Target Genes : Linking Dynamic Expression Changes to Function.” *Nucleic Acids Research* 41 (5): 2817–31.

Numnark, Somrak, Wuttichai Mhuantong, Supawadee Ingsriswang, and Duangdao Wichadakul. 2012. “C-Mii : A Tool for Plant MiRNA and Target Identification.” *BMC Genomics* 13: 1–10.

Ohlrogge, John, Doug Allen, Bill Berguson, Dean Dellapenna, Yair Shachar-hill, and Sten Stymne. 2009. “Driving on Biomass.” *Science* 324 (5930): 1019–20.

Oliveira, Louisi S de, Gustavo Bueno Gregoracci, Genivaldo G.Z Silva, Leonardo Tavares Salgado, Gilberto Amado Filho, Marcio a Alves-Ferreira, Renato Crespo Pereira, and Fabiano L Thompson. 2012. “Transcriptomic Analysis of the Red Seaweed Laurencia Dendroidea (Florideophyceae, Rhodophyta) and Its Microbiome.” *BMC Genomics* 13 (1): 487.

Oliver, Gavin R., Steven N. Hart, and Eric W. Klee. 2015. “Bioinformatics for Clinical next Generation Sequencing.” *Clinical Chemistry* 61 (1): 124–35.

Pamela S. Soltis, Jeff J. Doyle, and Douglas E. Soltis. 1992. "Molecular Data and Polyploid Evolution in Plants" *Molecular Systematics of Plants*: 177-201.

Papazoglou, E. G., G. A. Karantounias, S. N. Vemmos, and D. L. Bouranis. 2005. "Photosynthesis and Growth Responses of Giant Reed (Arundo Donax L.) to the Heavy Metals Cd and Ni." *Environment International* 31 (2): 243–49.

Park, W., Junjie Li, Rentao Song, Joachim Messing, and Xuemei Chen. 2002. "CARPEL FACTORY, a Dicer Homolog, and HEN1, a Novel Protein, Act in MicroRNA Metabolism in Arabidopsis Thaliana." *Curr Biol*. 12 (17): 1484–95.

Patanun, Onsaya, and Manassawe Lertpanyasampatha. 2013. "Computational Identification of MicroRNAs and Their Targets in Cassava ( Manihot Esculenta Crantz .)." *Molecular Biotechnology* 53: 257–69.

Patuzzi, Francesco, Tanja Mimmo, Stefano Cesco, Andrea Gasparella, and Marco Baratieri. 2013. "Common Reeds (Phragmites Australis) as Sustainable Energy Source: Experimental and Modelling Analysis of Torrefaction and Pyrolysis Processes." *GCB Bioenergy* 5 (4): 367–74.

Pauly, Markus, and Kenneth Keegstra. 2008. "Cell-Wall Carbohydrates and Their Modification as a Resource for Biofuels." *Plant Journal* 54: 559–68.

Persson, S., A. Paredez, A. Carroll, H. Palsdottir, M. Doblin, P. Poindexter, N. Khitrov, M. Auer, and C. R. Somerville. 2007. "Genetic Evidence for Three Unique Components in Primary Cell-Wall Cellulose Synthase Complexes in Arabidopsis." *Proceedings of the National Academy of Sciences* 104 (39): 15566–71.

Pervez, Muhammad Tariq, Masroor Ellahi Babar, Asif Nadeem, Muhammad Aslam, Ali Raza Awan, Naeem Aslam, Tanveer Hussain, et al. 2014. "Evaluating the Accuracy and Efficiency of Multiple Sequence Alignment Methods." *Evolutionary Bioinformatics* 10: 205–17.

Pigott, C. D. 1956. "The Vegetation of Upper Teesdale in the North Pennines." *Journal of Ecology* 44 (2): 545–86.

Pillon, Yohan, Jennifer B Johansen, Tomoko Sakishima, Eric H Roalson, Donald K Price, and Elizabeth A Stacy. 2013. "Gene Discordance in Phylogenomics of Recent Plant Radiations , an Example from Hawaiian Cyrtandra (Gesneriaceae)." *Molecular Phylogenetics and Evolution* 69 (1): 293–98.

Pilu, R., Andrea Bucci, Francesco Cerino Badone and Michela Landoni. 2012. "Giant Reed (Arundo Donax L.): A Weed Plant or a Promising Energy Crop?" *African Journal of Biotechnology* 11 (38): 9163–9174.

Pond, Sergei L Kosakovsky, Ben Murrell, Mathieu Fourment, Simon DW Frost, Wayne Delport, and Konrad Scheffler. 2008. "A Random Effects Branch-Site Model for Detecting Episodic Diversifying Selection." *Mol. Biol. Evol* 24 (1): 1–13.

Pond, Sergei L Kosakovsky, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon D W Frost. 2006. "Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm." *Mol. Biol. Evol* 23 (10): 1891–1901.

Pond, Sergei L Kosakovsky, Simon D W Frost, and Spencer V Muse. 2005. “HyPhy : Hypothesis Testing Using Phylogenies.” *Bioinformatics* 21 (5): 676–79.

Prakash, Pravin, Dolly Ghosliya, and Vikrant Gupta. 2015. “Identification of Conserved and Novel MicroRNAs in Catharanthus Roseus by Deep Sequencing and Computational Prediction of Their Potential Targets.” *Gene* 554 (2): 181–95.

Prakash, Pravin, Raja Rajakani, and Vikrant Gupta. 2016. “Transcriptome-Wide Identification of Rauvolfia Serpentina MicroRNAs and Prediction of Their Potential Targets” *Computational Biology and Chemistry* 61: 62–74.

Puigbo Pere, Yuri I. Wolf, and Eugene V. Koonin. 2010. “The Tree and Net Components of Prokaryote Evolution.” *Genome Biol. Evol* 2: 745–56.

R. Pearson, William R. 2013. “An Introduction to Sequence Similarity (“Homology”) Searching.” *Curr Protoc Bioinformatics* 249: 1–15.

Raes, J, Antje Rohde, Jørgen Holst Christensen, Yves Van de Peer, and and Wout Boerjan. 2003. “Genome-Wide Characterization of the Lignification Toolbox in Arabidopsis.” *Plant Physiology* 133 (3): 1051–71.

Rai, Amit, Mami Yamazaki, Hiroki Takahashi, Michimi Nakamura, Mareshige Kojoma, Hideyuki Suzuki, and Kazuki Saito. 2016. “RNA-Seq Transcriptome Analysis of Panax Japonicus, and Its Comparison with Other Panax Species to Identify Potential Genes Involved in the Saponins Biosynthesis.” *Frontiers in Plant Science* 7: 1–20.

Ramsey, Justin, and Douglas W Schemske. 1998. “ Pathways, mechanisms, and rates of polyploid formation in flowering plants.” *Annual Review of Ecology and Systematics* 29 (1): 467–501.

Ramsey, Justin, and Douglas W Schemske. 2002. “Neopolyploidy in Flowering Plants.” *Annu. Rev. Ecol. Syst* 33: 589–639.

Rannala, Bruce, and Ziheng Yang. 2008. “Phylogenetic Inference Using Whole Genomes.” *Annu. Rev. Genomics Hum. Genet* 9: 217–31.

Ranwez Vincent, Sebastien Harispe, Frederic Delsuc, Emmanuel J. P. Douzery. 2011. “MACSE : Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons.” *Plos One* 6 (9): 1–10.

Raspolli Galletti, Anna Maria, Claudia Antonetti, Erika Ribechini, Maria Perla Colombini, Nicoletta Nassi o Di Nasso, and Enrico Bonari. 2013. “From Giant Reed to Levulinic Acid and Gamma-Valerolactone: A High Yield Catalytic Route to Valeric Biofuels.” *Applied Energy* 102: 157–62.

Reddy, M. S., F. Chen, G. Shadle, L. Jackson, H. Aljoe, and R. A. Dixon. 2005. “Targeted Down-Regulation of Cytochrome P450 Enzymes for Forage Quality Improvement in Alfalfa (Medicago Sativa L.).” *Proc Natl Acad Sci USA* 102 (46): 16573–78.

Renny-Byfield, Simon, and Jonathan F. Wendel. 2014. “Doubling down on Genomes:

Polyploidy and Crop Plants." *American Journal of Botany* 101 (10): 1711–25.

Rhoades, Matthew W, Brenda J Reinhart, Lee P Lim, Christopher B Burge, Bonnie Bartel, and David P Bartel. 2002. "Prediction of Plant MicroRNA Targets" *cell* 110(4):513-20.

Rice Anna, Lior Glick, Shiran Abadi, Moshe Einhorn, Ayelet Salman-Minkov Naama M. Kopelman, and Ofer Chay and Itay Mayrose Jonathan Mayzel. 2015. "The Chromosome Counts Database ( CCDB )–a Community Resource of Plant Chromosome Numbers." *New Phytologist* 206: 19–26.

Richmond, Todd A., and Chris R. Somerville. 2000. "The Cellulose Synthase Superfamily." *Plant Physiology* 124 (2): 495–98.

Roberto Pilu. 2012. "Giant Reed (Arundo Donax L.): A Weed Plant or a Promising Energy Crop?" *African Journal of Biotechnology* 11 (38): 9163–74.

Ronquist, Fredrik, and John P Huelsenbeck. 2003. "MrBayes 3 : Bayesian Phylogenetic Inference under Mixed Models." *Bioinformatics* 19 (12): 1572–74.

Rosenkranz, Ruben, Tatiana Borodina, Hans Lehrach, and Heinz Himmelbauer. 2008. "Genomics Characterizing the Mouse ES Cell Transcriptome with Illumina Sequencing." *Genomics* 92: 187–94.

Rossa, B., A. V. Tüffers, G. Naidoo, and D. J. VON Willert. 1998. "Arundo Donax L. (Poaceae) - a C3 Species with Unusually High Photosynthetic Capacity." *Botanica Acta* 111 (3): 216–21.

Rubinelli, Peter M, George Chuck, Xu Li, and Richard Meilan. 2013. "Constitutive Expression of the Corngrass1 MicroRNA in Poplar Affects Plant Architecture and Stem Lignin Content and Composition." *Biomass and Bioenergy* 54: 312–21.

Sablok, Gaurav, Yuan Fu, Valentina Bobbio, Marina Laura, Giuseppe L Rotino, Paolo Bagnaresi, Andrea Allavena, et al. 2014. "Fuelling Genetic and Metabolic Exploration of C 3 Bioenergy Crops through the First Reference Transcriptome of Arundo Donax L .," *Plant Biotechnology Journal* 12: 554–67.

Schena, M., Shalon, D., Davis, R. and Brown, P. 1995. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science* 270: 467–70.

Schierup, Mikkel H, and Jotun Hein. 2000. "Consequences of Recombination on Traditional Phylogenetic Analysis." *Genetics* 156: 879–91.

Schlüter, Urte, Alisandra K Denton, and Andrea Bräutigam. 2016. "Understanding Metabolite Transport and Metabolism in C4 Plants through RNA-Seq." *Current Opinion in Plant Biology* 31: 83–90.

Schwab, Rebecca, Javier F. Palatnik, Markus Riester, Carla Schommer, Markus Schmid, and Detlef Weigel. 2005. "Specific Effects of MicroRNAs on the Plant

Transcriptome." *Developmental Cell* 8 (4): 517–27.

Schwerdt, Julian G., Katrin MacKenzie, Frank Wright, Daniel Oehme, John M. Wagner, Andrew J. Harvey, Neil J. Shirley, et al. 2015. "Evolutionary Dynamics of the Cellulose Synthase Gene Superfamily in Grasses." *Plant Physiology* 168 (3): 968–83.

Scordia, Danilo, Salvatore L. Cosentino, Jae Won Lee, and Thomas W. Jeffries. 2011. "Dilute Oxalic Acid Pretreatment for Biorefining Giant Reed (Arundo Donax L.)." *Biomass and Bioenergy* 35 (7): 3018–24.

Sedaghatinia, Asieh, Rodziah Binti Atan, Khairinatajul Arifin, Masrah Azrifah, and Binti Azmi. 2009. "Comparison and Evaluation of Multiple Sequence Alignment Tools In Bininformatics." *International Journal of Computer Science and Network Security* 9 (7): 51–56.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13: 2498–2504.

Singh, Noopur, Swati Srivastava, and Ashok Sharma. 2016. "Identification and Analysis of MiRNAs and Their Targets in Ginger Using Bioinformatics Approach." *Gene* 575 (2): 570–76.

Smith, Stephen A, and Casey W Dunn. 2008. "Phyutility : A Phyloinformatics Tool for Trees, Alignments and Molecular Data." *Bioinformatics* 24 (5): 715–16.

Soltis, Douglas E, Victor A Albert, Jim Leebens-mack, Charles D Bell, Andrew H Paterson, Chunfang Zheng, David Sankoff, W Claude, P Kerr Wall, and Pamela S Soltis. 2009. "Polyploidy and angiosperm diversification." *American Journal of Botany* 96 (1): 336–48.

Soreng, Robert J, Paul M Peterson, Konstantin Romaschenko, Gerrit Davidse, Fernando O Zuloaga, Emmet J Judziewicz, Tarciso S Filgueiras, Jerrold I Davis, and Osvaldo Morrone. 2015. "A Worldwide Phylogenetic Classification of the Poaceae (Gramineae)." *Journal of Systematics and Evolution* 53 (2): 117–137.

Soreng, Robert J, Paul M Peterson, Konstantin Romaschenko, Gerrit Davidse, Jordan K Teisher, Lynn G Clark, Patricia Barberá, Lynn J Gillespie, and Fernando O Zuloaga. 2017. "A Worldwide Phylogenetic Classification of the Poaceae (Gramineae) II: An Update and a Comparison of Two 2015 Classifications. " *Journal of Systematics and Evolution* 55 (4): 259-290.

Sorin, Celine, Marie Declerck, Thomas Blein, Linnan Ma, Christine Lelandais-bri, Maria Fransiska Njo, Tom Beeckman, Martin Crespi, and Caroline Hartmann. 2014. "A MiR169 Isoform Regulates Specific NF-YA Targets and Root Architecture in Arabidopsis." *New Phytologist* 202: 1197–1211.

Srivastava, Prashant K, Taraka Ramji Moturu, Priyanka Pandey, Ian T Baldwin, and Shree P Pandey. 2014. “A Comparison of Performance of Plant MiRNA Target Prediction Tools and the Characterization of Features for Genome-Wide Target Prediction.” *BMC Genomics* 15 (348): 1–15.

STAMATAKIS A, Hoover P, Rougemont J.. 2008. “A Rapid Bootstrap Algorithm for the RAxML Web Servers.” *Syst. Biol* 57 (5): 758–771.

Stamatakis, Alexandros. 2006. “RAxML-VI-HPC : Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.” *Bioinformatics* 22 (21): 2688–90.

Sun Wei, Xiao Hui Xu, Xiu Wu, Yong Wang, Xingbo Lu, Hongwei Sun, and Xianzhi Xie. 2015. “Genome-Wide Identification of MicroRNAs and Their Targets in Wild Type and PhyB Mutant Provides a Key Link between MicroRNAs and the PhyB-Mediated Light Signaling Pathway in Rice.” *Frontiers in Plant Science* 6: 1–15.

Sunkar, Ramanjulu, Thomas Girke, Pradeep Kumar Jain, and Jian-kang Zhu. 2005. “Cloning and Characterization of MicroRNAs from Rice.” *The Plant Cell* 17: 1397–1411.

Suzuki, S., L. Li, Y.-H. Sun, and V. L. Chiang. 2006. “The Cellulose Synthase Gene Superfamily and Biochemical Functions of Xylem-Specific Cellulose Synthase-Like Genes in Populus Trichocarpa.” *Plant Physiology* 142 (3): 1233–45.

Swofford, David L. 2002. “PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). ” book.

Takahashi, Wataru, Tadashi Takamizo, Makoto Kobayashi, and Masumi Ebina. 2010. “Plant Regeneration from Calli in Giant Reed ( Arundo Donax L .).” *Grassland Science* 56: 224–29.

Tanzer, Andrea, and Peter F Stadler. 2004. “Molecular Evolution of a MicroRNA Cluster.” *Journal of Molecular Biology* 339 (2): 327–335.

Taylor, K, A P Rowland, and H E Jones. 2001. “Molinia Caerulea ( L .) Moench.” *Journal of Ecology* 89: 126–44.

Taylor, N. G., R. M. Howells, A. K. Huttly, K. Vickers, and S. R. Turner. 2003. “Interactions among Three Distinct CesA Proteins Essential for Cellulose Synthesis.” *Proceedings of the National Academy of Sciences* 100 (3): 1450–55.

Thirumurugan, Thiruvengadam, and Yukihiro Ito. 2008. “Identification , Characterization and Interaction of HAP Family Genes in Rice.” *Mol Genet Genomics* 279: 279–89.

Timme, Ruth E, Tsvetan R Bachvaroff, and Charles F Delwiche. 2012. “Broad Phylogenomic Sampling and the Sister Lineage of Land Plants.” *PLoS One* 7 (1): 1–8.

Trabucco, G M, D A Matos, S J Lee, A J Saathoff, H D Priest, T C Mockler, G Sarath, and S P Hazen. 2013. “Functional Characterization of Cinnamyl Alcohol Dehydrogenase and Caffeic Acid O-Methyltransferase in Brachypodium Distachyon.” *BMC Biotechnol* 13: 2–18.

Triplett, Jimmy K., Lynn G. Clark, Amanda E. Fisher, and Jun Wen. 2014. “Independent Allopolyploidization Events Preceded Speciation in the Temperate and Tropical Woody Bamboos.” *New Phytologist* 204 (1): 66–73.

Trumbo, Jennifer Lynn, Baohong Zhang, and Charles Neal Stewart. 2015. “Manipulating MicroRNAs for Improved Biomass and Biofuels from Plant Feedstocks.” *Plant Biotechnology Journal* 13: 337–54.

Valente, Maria Anete S., Jerusa A.Q.A. Faria, Juliana R.L. Soares-Ramos, Pedro A.B. Reis, Guilherme L. Pinheiro, Newton D. Piovesan, Angélica T. Morais, et al. 2009. “The ER Luminal Binding Protein (BiP) Mediates an Increase in Drought Tolerance in Soybean and Delays Drought-Induced Leaf Senescence in Soybean and Tobacco.” *Journal of Experimental Botany* 60 (2): 533–46.

Valli, Fabio, Daniele Trebbi, Walter Zegada-Lizarazu, Andrea Monti, Roberto Tuberosa, and Silvio Salvi. 2017. “In Vitro Physical Mutagenesis of Giant Reed (Arundo Donax L.).” *GCB Bioenergy* 9 (8): 1380–89.

Vega-arreguín, Julio C, Enrique Ibarra-laclette, Beatriz Jiménez-moraila, Octavio Martínez, Jean Philippe Vielle-calzada, Luis Herrera-estrella, and Alfredo Herrera-estrella. 2009. “Deep Sampling of the Palomero Maize Transcriptome by a High Throughput Strategy of Pyrosequencing.” *BMC Genomics* 10 (299): 1–10.

Velculescu, E, Bert Vogelstein, Kenneth W. Kinzler. 2000. “Analysing Uncharted Transcriptomes with SAGE.” *Trends Genet* 9525: 423–425.

Venkat, Aarti, Matthew W Hahn, and Joseph W Thornton. 2018. “Multinucleotide Mutations Cause False Inferences of Lineage-Specific Positive Selection.” *Nature Ecology & Evolution* 2 (8): 1280–1288.

Vigna, Bianca Baccili Zanotto, Fernanda Ancelmo de Oliveira, Guilherme de Toledo-Silva, Carla Cristina da Silva, Cacilda Borges do Valle, and Anete Pereira de Souza. 2016. “Leaf Transcriptome of Two Highly Divergent Genotypes of Urochloa Humidicola ( Poaceae ), a Tropical Polyploid Forage Grass Adapted to Acidic Soils and Temporary Flooding Areas.” *BMC Genomics* 17 (910): 1–19.

Vincent Ranwez, Sébastien Harispe, Frederic Delsuc, Emmanuel J. P. Douzery. 2011. “MACSE : Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons.” *Plos One* 6 (9): 1–10.

Wall, P Kerr, Jim Leebens-mack, Andr é S Chanderbali, Abdelali Barakat, Erik Wolcott, Haiying Liang, Lena Landherr, et al. 2009. “Comparison of next Generation Sequencing Technologies for Transcriptome Characterization.” *BMC Genomics* 10 (347): 1–19.

Wang zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nat Rev Genet* 10 (1): 57–63.

Wang, Huacai, Xiaoming Jiao, Xiaoyu Kong, Sadia Hamera, Yao Wu, Xiaoying Chen, and Rongxiang Fang. 2016. “A Signaling Cascade from MiR444 to RDR1 in Rice Antiviral RNA Silencing Pathway.” *Plant Physiology* 170: 2365–77.

Wang, Kai, Wei Hong, Hengwu Jiao, and Huabin Zhao. 2017. “Transcriptome Sequencing and Phylogenetic Analysis of Four Species of Luminescent Beetles.” *Scientific Reports* 7(1):1–11.

Wang, Lingqiang, Kai Guo, Yu Li, Yuanyuan Tu, Huizhen Hu, Bingrui Wang, and Xiaocan Cui. 2010. “Expression Profiling and Integrative Analysis of the CESA / CSL Superfamily in Rice.” *BMC Plant Biology* 10 (282): 1–16.

Wang, Xiao-wei, Jun-bo Luan, Jun-min Li, Yan-yuan Bao, Chuan-xi Zhang, and Shu-sheng Liu. 2010. “De Novo Characterization of a Whitefly Transcriptome and Analysis of Its Gene Expression during Development.” *BMC Genomics* 11 (400): 1–11.

Wang, Xiu-Jie, Jos é L Reyes, Nam-Hai Chua, and Terry Gaasterland. 2004. “Prediction and Identification of Arabidopsis Thaliana MicroRNAs and Their MRNA Targets.” *Genome Biology* 5 (9): R65.1-R65.15.

Weber Andreas P.M., Katrin L. Weber, Kevin Carr, Curtis Wilkerson, and John B. Ohlrogge. 2007. “Sampling the Arabidopsis Transcriptome with Massively Parallel Pyrosequencing.” *Plant Physiology* 144: 32–42.

Wei Ning, Deng Xing-Wang. 1999. “Making Sense of the COP9 Signalosome. A Regulatory Protein Complex Conserved from Arabidopsis to Human.” *Cell Press* 15 (3): 98–103.

Wen, Jun, Ashley N Egan, Rebecca B Dikow, and Elizabeth A Zimmer. 2015. “Utility of Transcriptome Sequencing for Phylogenetic Inference and Character Evolution.” *In Next-Generation Sequencing in Plant Systematics,* Chapter2.

Wen, Jun, Zhiqiang Xiong, Ze Long Nie, Likai Mao, Yabing Zhu, Xian Zhao Kan, Stefanie M. Ickert-Bond, Jean Gerrath, Elizabeth A Zimmer, and Xiao Dong Fang. 2013. “Transcriptome Sequences Resolve Deep Relationships of the Grape Family.” *PLoS One* 8 (9): 1–9.

Wen, Ming, Munan Xie, Lian He, Yushuai Wang, Suhua Shi, and Tian Tang. 2016. “Expression Variations of MiRNAs and MRNAs in Rice.” *Genome Biol. Evol* 8 (11): 3529–44.

Wendel, Jonathan F. 2000. “Genome Evolution in Polyploids.” *Plant Molecular Evolution* 42: 225–49.

Wertheim, Joel O, Ben Murrell, Martin D Smith, and Sergei L Kosakovsky Pond. 2014. “RELAX : Detecting Relaxed Selection in a Phylogenetic Framework.” *Molecular*

*Biology and Evolution,* 32 (3): 820-32.

Wi, Seung Gon, Eun Jin Cho, Dae Seok Lee, Soo Jung Lee, Young Ju Lee, and Hyeun Jong Bae. 2015. "Lignocellulose Conversion for Biofuel: A New Pretreatment Greatly Improves Downstream Biocatalytic Hydrolysis of Various Lignocellulosic Materials." *Biotechnology for Biofuels* 8 (228): 1–11.

Wickett, Norman J, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, et al. 2014. "Phylotranscriptomic Analysis of the Origin and Early Diversification of Land Plants." *Proceedings of the National Academy of Sciences* 111 (45) E4859-E4868.

Wu, Gang, Mee Yeon Park, Susan R Conway, Jia-wei Wang, Detlef Weigel, and R Scott. 2009. "The Sequential Action of MiR156 and MiR172 Regulates Developmental Timing in Arabidopsis." *Cell* 138 (4): 750–59.

Wu, M.-F., Q. Tian, and J. W. Reed. 2006. "Arabidopsis MicroRNA167 Controls Patterns of ARF6 and ARF8 Expression, and Regulates Both Female and Male Reproduction." *Development* 133 (21): 4211–18.

Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, et al. 2014. "SOAPdenovo-Trans: De Novo Transcriptome Assembly with Short RNA-Seq Reads." *Bioinformatics* 30 (12): 1660–66.

Xu Hua, Wenzhong Xu, Hongmei Xi, Wenwen Ma, Zhenyan He, Mi Ma. 2013. "The ER luminal binding protein (BiP) alleviates Cd2+-induced programmed cell death through endoplasmic reticulum stress–cell death signaling pathway in tobacco cells" *Journal of Plant Physiology* 170: 1434–1441.

Xu, Jian-hua, Fei Li, and Qiu-feng Sun. 2008. "Identification of MicroRNA Precursors with Support Vector Machine and String Kernel." *Genomics, Proteomics & Bioinformatics* 6 (2): 121–28.

Xu, Miao Yun, Lan Zhang, Wei Wei Li, Xiao Long Hu, Ming-bo Wang, Yun Liu Fan, and Chun Yi Zhang. 2014. "Stress-Induced Early Flowering Is Mediated by MiR169 in Arabidopsis Thaliana." *Journal of Experimental Botany* 65 (1): 89–101.

Xu, Yingchun, Lingling Chu, Qijiang Jin, Yanjie Wang, Xian Chen, and Hui Zhao. 2015. "Transcriptome-Wide Identification of MiRNAs and Their Targets from Typha Angustifolia by RNA-Seq and Their Response to Cadmium Stress." *Plos One* 10 (4): 1–22.

Xu, Zhanyou, Dandan Zhang, Jun Hu, Xin Zhou, Xia Ye, Kristen L. Reichel, Nathan R. Stewart, et al. 2009. "Comparative Genome Analysis of Lignin Biosynthesis Gene Families across the Plant Kingdom." *BMC Bioinformatics* 10: 1–15.

Xuan, Ping, Maozu Guo, Xiaoyan Liu, Yangchao Huang, Wenbin Li, and Yufei Huang. 2011. "PlantMiRNAPred : Efficient Classification of Real and Pseudo Plant Pre-MiRNAs." *Bioinformatics* 27 (10): 1368–76.

Yang, Ya, and Stephen A Smith. 2014. “Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes : Improving Accuracy and Matrix Occupancy for Phylogenomics.” *Mol. Biol. Evol* 31 (11): 3081–92.

Yang, Ziheng. 2005. “The Power Ofphylogenetic Comparison in Revealing Protein Function.” *PNAS* 102 (9): 3179–80.

Ye, Jia, Lin Fang, Hongkun Zheng, Yong Zhang, Jie Chen, Zengjin Zhang, Jing Wang, et al. 2006. “WEGO : A Web Tool for Plotting GO Annotations.” *Nucleic Acids Research* 34: 293–97.

Yi, Ting-shuang, Gui-hua Jin, and Jun Wen. 2015. “Chloroplast Capture and Intra- and Inter-Continental Biogeographic Diversification in the Asian–New World Disjunct Plant Genus Osmorhiza (Apiaceae).” *Molecular Phylogenetics and Evolution* 85: 10–21.

Yi, Xin, Zhenhai Zhang, Yi Ling, Wenying Xu, and Zhen Su. 2015. “PNRD: A Plant Non-Coding RNA Database.” *Nucleic Acids Research* 43: D982–89.

Yin, Lan, Yves Verhertbruggen, Ai Oikawa, Chithra Manisseri, Bernhard Knierim, Lina Prak, Jacob Krüger Jensen, et al. 2011. “The Cooperative Activities of CSLD2, CSLD3, and CSLD5 Are Required for Normal Arabidopsis Development.” *Molecular Plant* 4 (6): 1024–37.

Yin, Yanbin, Jinling Huang, and Ying Xu. 2009. “The Cellulose Synthase Superfamily in Fully Sequenced Plants and Algae.” *BMC Plant Biology* 9: 1–14.

Yin, Yanbin, Mitrick A. Johns, Huansheng Cao, and Manju Rupani. 2014. “A Survey of Plant and Algal Genomes and Transcriptomes Reveals New Insights into the Evolution and Function of the Cellulose Synthase Superfamily.” *BMC Genomics* 15 (1): 1–15.

Yin, Zujun, Chunhe Li, Xiulan Han, and Fafu Shen. 2008. “Identification of Conserved MicroRNAs and Their Target Genes in Tomato (Lycopersicon Esculentum).” *Gene* 414: 60–66.

Yu, Xiaoqing, Kishore Guda, Joseph Willis, Martina Veigl, Zhenghe Wang, Sanford Markowitz, Mark D. Adams, and Shuying Sun. 2012. “How Do Alignment Programs Perform on Sequencing Data with Varying Qualities and from Repetitive Regions?” *BioData Mining* 5: 1–12.

Yuan, Xiongying, Changning Liu, Pengcheng Yang, Shunmin He, Qi Liao, Shuli Kang, and Yi Zhao. 2009. “Clustered MicroRNAs’ Coordination in Regulating Protein-Protein Interaction Network.” *BMC Systems Biology* 10: 1–10.

Yutin, Natalya, Pere Puigbo, Eugene V. Koonin, and Yuri I. Wolf. 2012. “Phylogenomics of Prokaryotic Ribosomal Proteins.” *Plos One* 7 (5): 1–10.

Zerbino, Daniel R., and Ewan Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” *Genome Research* 18: 821–29.

Zhang, B H, X P Pan, S B Cox, G P Cobb, and T A Anderson. 2006. “Evidence That MiRNAs Are Different from Other RNAs.” *Cell. Mol. Life Sci* 63: 246–54.

Zhang, Baohong, Xiaoping Pan, and Edmund J Stellwag. 2008. “Identification of Soybean MicroRNAs and Their Targets.” *Planta* 229 (1): 161–82.

Zhang, Baohong, Xiaoping Pan, Charles H Cannon, George P Cobb, Todd A Anderson, and Texas Tech. 2006. “Conservation and Divergence of Plant MicroRNA Genes.” *The Plant Journal* 46: 243–59.

Zhang, Baohong, Xiaoping Pan, George P Cobb, and Todd A Anderson. 2006. “Plant MicroRNA : A Small Regulatory Molecule with Big Impact.” *Developmental Biology* 289 (1): 3–16.

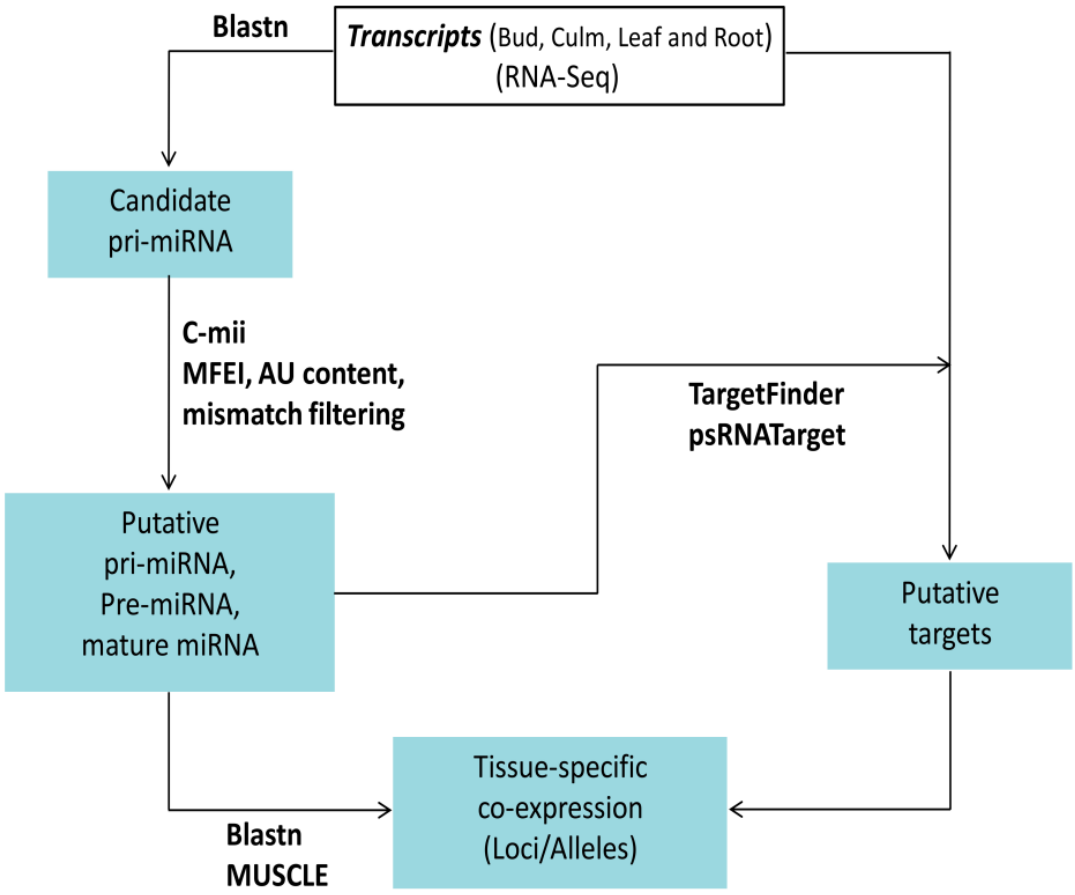
Zhang, Miaozhi, Michele de C. Pereira e Silva, Maryam Chaib De Mares, and Jan Dirk van Elsas. 2014. “The Mycosphere Constitutes an Arena for Horizontal Gene Transfer with Strong Evolutionary Implications for Bacterial-Fungal Interactions.” *FEMS Microbiology Ecology* 89 (3): 516–26.

Zhong, R, W H Morrison 3rd, D S Himmelsbach, F L Poole 2nd, and Z H Ye. 2000. “Essential Role of Caffeoyl Coenzyme A O-Methyltransferase in Lignin Biosynthesis in Woody Poplar Plants.” *Plant Physiology* 124 (2): 563–78.

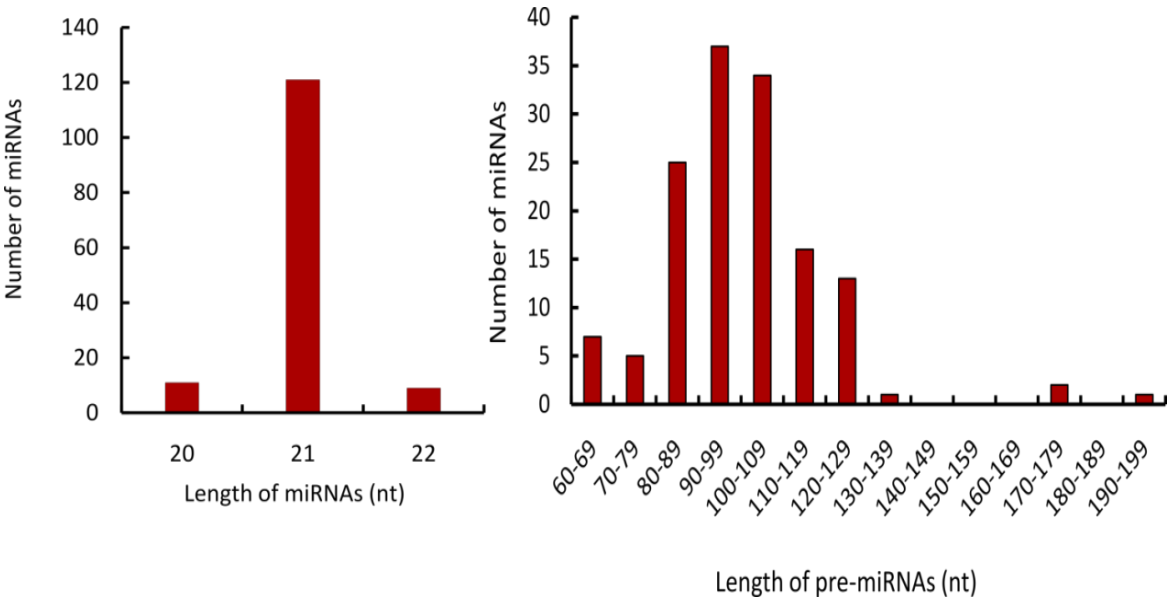
Zhu, Qian-hao, and Chris A Helliwell. 2011. “Regulation of Flowering Time and Floral Patterning by MiR172.” *Journal of Experimental Botany* 62 (2): 487–95.

Zimmer, Elizabeth A, and Jun Wen. 2012. “Using Nuclear Gene Data for Plant Phylogenetics: Progress and Prospects.” *Molecular Phylogenetics and Evolution* 65 (2): 774–85.

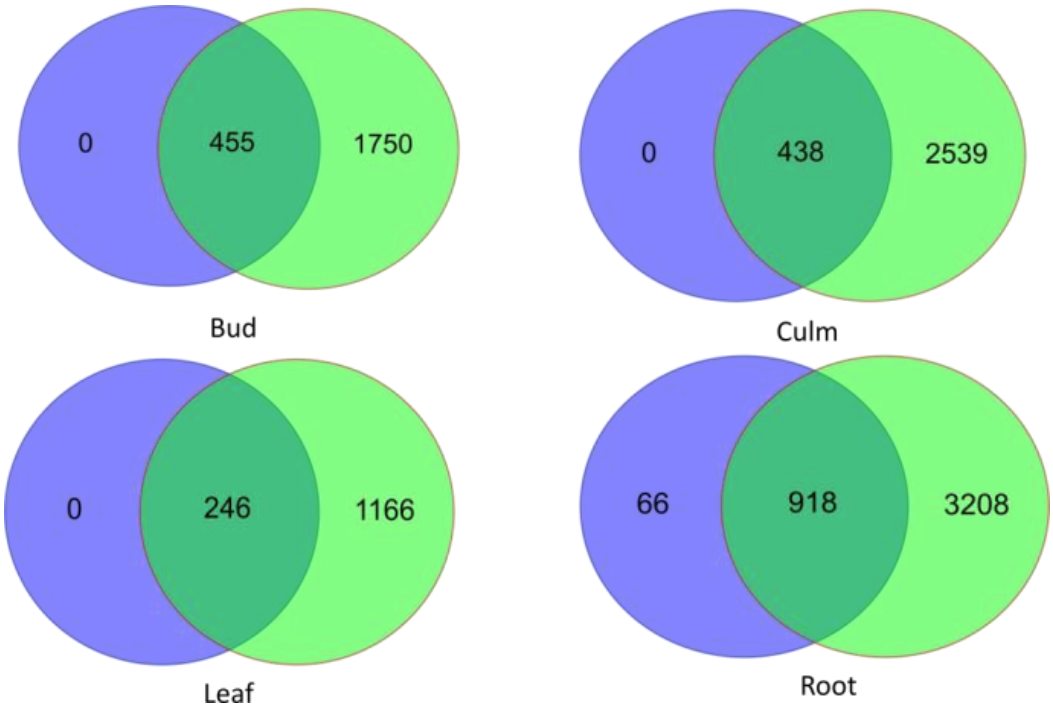
# Supplementary Figures and Tables



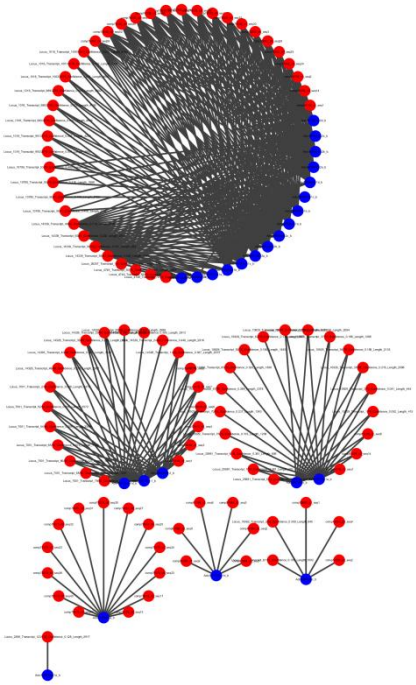

**Supplementary Figure 1.1**. Schematic workflow of the analyses for the identification of *A. donax* microRNAs and their targets.



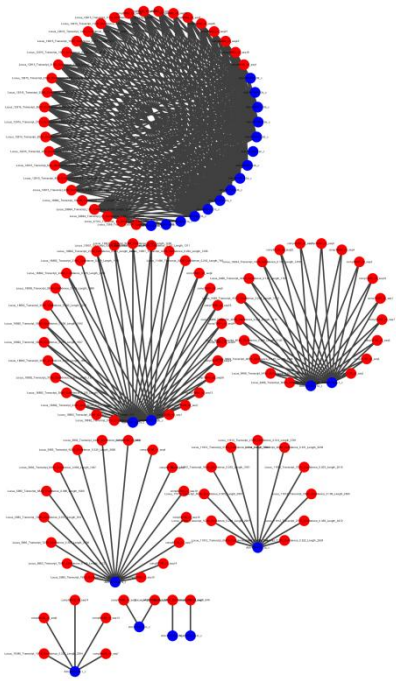

**Supplementary Figure 1.2**. Length distributions of mature miRNA and pre-miRNA.



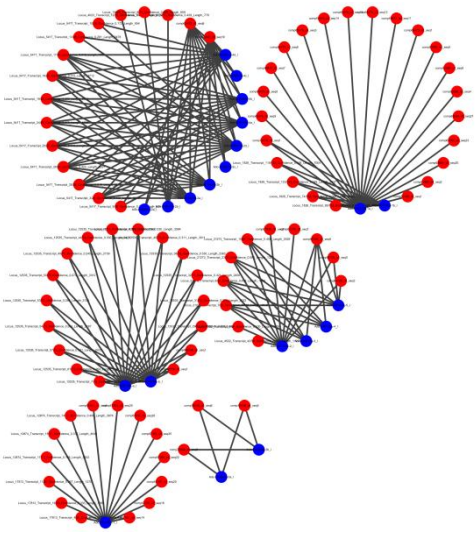

**Supplementary Figure 1.3**. Venn diagrams of putative targets predicted by both psRNAtarget and TargetFinder.



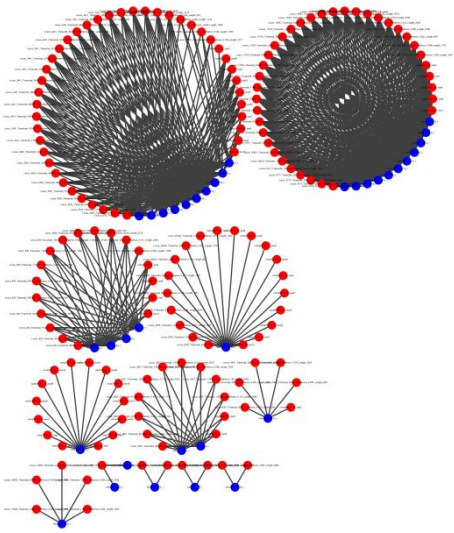
**Bud**



**Culm**



**Leaf**



**Root**

**Supplementary Figure 1.4.** Functional networks of *A. donax* microRNAs and their putative targets (Cytoscape). Blue dots represent miRNA genes, while red dots are their targets for each of four tissues (bud, culm, leaf and root). Lines represent miRNA-target relationships. Clusters of lines originating from single miRNAs indicate the size of the post-transcriptional gene silencing (PTGS) co-regulation for each miRNA gene.

**Supplementary Table 1.1.** Features of putative miRNA identified from *A. donax*.

| Putative locus | Primary miRNA | Precursor miRNA | mature miRNA | Precursor miRNA MFEI (kcal/mol) | Strand |
|---|---|---|---|---|---|
| Locus_1 | culm_36 | Ado-MIR1430-1_c | Ado-MIR1430-1_c | -0.931 | plus |
| | leaf_17 | Ado-MIR1430-1_l | Ado-MIR1430-1_l | -0.931 | plus |
| Locus_2 | culm_52 | Ado-MIR1430-2_c | Ado-MIR1430-2_c | -1.006 | minus |
| Locus_3 | leaf_24 | Ado-MIR1430-3_l | Ado-MIR1430-3_l | -0.997 | minus |
| Locus_4 | root_35 | Ado-MIR1430-4_r | Ado-MIR1430-4_r | -0.995 | plus |
| Locus_5 | root_86 | Ado-MIR156a_r | Ado-MIR156a-a_r | -0.952 | minus |
| | root_86 | Ado-MIR156a_r | Ado-MIR156a-b_r | -0.952 | minus |
| Locus_6 | root_55 | Ado-MIR156b_r | Ado-MIR156b-a_r | -1.037 | minus |
| | root_55 | Ado-MIR156b_r | Ado-MIR156b-b_r | -1.037 | minus |
| Locus_7 | root_29 | Ado-MIR156c_r | Ado-MIR156c_r | -0.926 | plus |
| Locus_8 | root_38 | Ado-MIR156g-1_r | Ado-MIR156g-1a_r | -0.864 | plus |
| | root_38 | Ado-MIR156g-1_r | Ado-MIR156g-1b_r | -0.864 | plus |
| Locus_9 | root_67 | Ado-MIR156g-2_r | Ado-MIR156g-2_r | -0.879 | minus |
| Locus_10 | root_99_part1 | Ado-MIR156g-3_r | Ado-MIR156g-3_r | -0.879 | plus |
| Locus_11 | root_68 | Ado-MIR156j_r | Ado-MIR156j-a_r | -0.888 | plus |
| | root_68 | Ado-MIR156j_r | Ado-MIR156j-b_r | -0.888 | plus |
| Locus_12 | leaf_58 | Ado-MIR156k_l | Ado-MIR156k-a_l | -0.852 | minus |
| | leaf_58 | Ado-MIR156k_l | Ado-MIR156k-b_l | -0.852 | minus |
| Locus_13 | culm_61 | Ado-MIR160b-1_c | Ado-MIR160b-1_c | -0.85 | minus |
| Locus_14 | root_15 | Ado-MIR160b-2_r | Ado-MIR160b-2_r | -0.85 | plus |
| Locus_15 | root_72 | Ado-MIR160b-3_r | Ado-MIR160b-3_r | -0.85 | plus |
| Locus_16 | Bud_67 | Ado-MIR160c_b | Ado-MIR160c_b | -0.977 | plus |
| Locus_17 | root_12 | Ado-MIR166a-1_r | Ado-MIR166a-1_r | -0.88 | plus |
| Locus_18 | root_13 | Ado-MIR166a-2_r | Ado-MIR166a-2_r | -0.88 | plus |
| Locus_19 | Bud_57 | Ado-MIR167d-1_b | Ado-MIR167d-1_b | -1.017 | plus |
| Locus_20 | leaf_49 | Ado-MIR167d-2_l | Ado-MIR167d-2_l | -1.017 | minus |
| Locus_21 | Bud_58 | Ado-MIR167g-1_b | Ado-MIR167g-1_b | -0.945 | plus |
| Locus_22 | culm_89 | Ado-MIR167g-2_c | Ado-MIR167g-2_c | -0.945 | minus |
| Locus_23 | root_37 | Ado-MIR167g-3_r | Ado-MIR167g-3_r | -0.945 | minus |
| Locus_24 | root_47 | Ado-MIR169a_r | Ado-MIR169a_r | -0.906 | minus |
| Locus_25 | Bud_32 | Ado-MIR169c-1-1_b | Ado-MIR169c-1a_b | -0.971 | plus |
| | Bud_32 | Ado-MIR169c-1-2_b | Ado-MIR169c-1b_b | -0.869 | minus |
| | culm_80 | Ado-MIR169c-1-1_c | Ado-MIR169c-1a_c | -0.971 | minus |
| | culm_80 | Ado-MIR169c-1-2_c | Ado-MIR169c-1b_c | -0.869 | plus |
| Locus_26 | leaf_41 | Ado-MIR169c-2_l | Ado-MIR169c-2_l | -0.874 | plus |
| Locus_27 | root_90 | Ado-MIR169c-3-1_r | Ado-MIR169c-3a_r | -0.951 | minus |
| | root_90 | Ado-MIR169c-3-2_r | Ado-MIR169c-3b_r | -0.913 | plus |
| Locus_28 | root_18 | Ado-MIR169i_r | Ado-MIR169i_r | -0.872 | minus |

| | | | | | |
|---|---|---|---|---|---|
| Locus_29 | leaf_36 | Ado-MIR169n-1_l | Ado-MIR169n-1_l | -0.86 | plus |
| Locus_30 | root_39 | Ado-MIR169n-2_r | Ado-MIR169n-2_r | -0.87 | minus |
| Locus_31 | root_51 | Ado-MIR169n-3_r | Ado-MIR169n-3_r | -0.862 | minus |
| Locus_32 | leaf_41 | Ado-MIR169p_l | Ado-MIR169p_l | -0.92 | minus |
| Locus_33 | Bud_48 | Ado-MIR169q-1_b | Ado-MIR169q-1_b | -0.951 | plus |
| | culm_60 | Ado-MIR169q-1_c | Ado-MIR169q-1_c | -0.951 | minus |
| Locus_34 | Bud_49 | Ado-MIR169q-2_b | Ado-MIR169q-2_b | -0.951 | plus |
| Locus_35 | leaf_07 | Ado-MIR169q-3_l | Ado-MIR169q-3_l | -0.858 | plus |
| Locus_36 | leaf_18 | Ado-MIR169q-4_l | Ado-MIR169q-4_l | -0.915 | plus |
| Locus_37 | leaf_47 | Ado-MIR169q-5_l | Ado-MIR169q-5_l | -0.951 | plus |
| Locus_38 | root_18 | Ado-MIR169q-6_r | Ado-MIR169q-6_r | -0.858 | minus |
| Locus_39 | root_57 | Ado-MIR169q-7_r | Ado-MIR169q-7_r | -0.951 | minus |
| Locus_40 | culm_45 | Ado-MIR171a_c | Ado-MIR171a_c | -0.927 | minus |
| Locus_41 | Bud_59 | Ado-MIR171c-1_b | Ado-MIR171c-1_b | -1.022 | minus |
| Locus_42 | culm_79 | Ado-MIR171c-2_c | Ado-MIR171c-2_c | -0.966 | minus |
| Locus_43 | root_58 | Ado-MIR171f_r | Ado-MIR171f_r | -0.938 | minus |
| Locus_44 | Bud_81 | Ado-MIR171i-1_b | Ado-MIR171i_b | -0.957 | minus |
| | culm_44 | Ado-MIR171i-2_c | Ado-MIR171i_c | -0.93 | minus |
| Locus_45 | Bud_82 | Ado-MIR172b_b | Ado-MIR172b_b | -1.019 | plus |
| Locus_46 | Bud_45 | Ado-MIR172d-1-1_b | Ado-MIR172d-1a_b | -1.012 | plus |
| | Bud_45 | Ado-MIR172d-1-2_b | Ado-MIR172d-1b_b | -0.924 | minus |
| | leaf_51 | Ado-MIR172d-1-1_l | Ado-MIR172d-1a_l | -1.012 | minus |
| | leaf_51 | Ado-MIR172d-1-2_l | Ado-MIR172d-1b_l | -0.924 | plus |
| | root_88 | Ado-MIR172d-1-1_r | Ado-MIR172d-1a_r | -1.012 | minus |
| | root_88 | Ado-MIR172d-1-2_r | Ado-MIR172d-1b_r | -0.924 | plus |
| Locus_47 | culm_69 | Ado-MIR172d-2-1_c | Ado-MIR172d-2a_c | -1.007 | minus |
| | culm_69 | Ado-MIR172d-2-2_c | Ado-MIR172d-2b_c | -0.929 | plus |
| Locus_48 | culm_55 | Ado-MIR319a-1-1_c | Ado-MIR319a-1a_c | -0.911 | plus |
| | culm_55 | Ado-MIR319a-1-2_c | Ado-MIR319a-1b_c | -0.853 | plus |
| Locus_49 | root_32 | Ado-MIR319a-2-1_r | Ado-MIR319a-2a_r | -0.911 | plus |
| | root_32 | Ado-MIR319a-2-2_r | Ado-MIR319a-2b_r | -0.853 | plus |
| Locus_50 | root_16 | Ado-MIR393a_r | Ado-MIR393a-a_r | -0.85 | plus |
| | root_16 | Ado-MIR393a_r | Ado-MIR393a-b_r | -0.85 | plus |
| Locus_51 | Bud_52 | Ado-MIR393b-1_b | Ado-MIR393b-1a_b | -0.866 | plus |
| | Bud_52 | Ado-MIR393b-1_b | Ado-MIR393b-1b_b | -0.866 | plus |
| | Bud_52 | Ado-MIR393b-1_b | Ado-MIR393b-1c_b | -0.866 | plus |
| Locus_52 | culm_53 | Ado-MIR393b-2_c | Ado-MIR393b-2a_c | -0.866 | minus |
| | culm_53 | Ado-MIR393b-2_c | Ado-MIR393b-2b_c | -0.866 | minus |
| | culm_53 | Ado-MIR393b-2_c | Ado-MIR393b-2c_c | -0.866 | minus |
| Locus_53 | leaf_37 | Ado-MIR393b-3_l | Ado-MIR393b-3a_l | -0.866 | minus |
| | leaf_37 | Ado-MIR393b-3_l | Ado-MIR393b-3b_l | -0.866 | minus |
| | leaf_37 | Ado-MIR393b-3_l | Ado-MIR393b-3c_l | -0.866 | minus |
| Locus_54 | culm_40 | Ado-MIR396a_c | Ado-MIR396a_c | -0.948 | minus |

| | | | | | |
|---|---|---|---|---|---|
| Locus_55 | culm_41 | Ado-MIR396b_c | Ado-MIR396b_c | -1.079 | minus |
| Locus_56 | Bud_18 | Ado-MIR444a-1-1_b | Ado-MIR444a-1a_b | -1.158 | plus |
| | Bud_18 | Ado-MIR444a-1-2_b | Ado-MIR444a-1b_b | -1.247 | plus |
| | culm_71 | Ado-MIR444a-2-1_c | Ado-MIR444a-2a_c | -1.298 | minus |
| | culm_71 | Ado-MIR444a-2-2_c | Ado-MIR444a-2b_c | -1.402 | minus |
| Locus_57 | culm_71 | Ado-MIR444a-2-2_c | Ado-MIR444a-2c_c | -1.402 | minus |
| | leaf_35 | Ado-MIR444a-2-1_l | Ado-MIR444a-2a_l | -1.298 | minus |
| | leaf_35 | Ado-MIR444a-2-2_l | Ado-MIR444a-2b_l | -1.402 | minus |
| | leaf_35 | Ado-MIR444a-2-2_l | Ado-MIR444a-2c_l | -1.402 | minus |
| | Bud_03 | Ado-MIR444c-1-1_b | Ado-MIR444c-1a_b | -0.902 | minus |
| | Bud_03 | Ado-MIR444c-1-2_b | Ado-MIR444c-1b_b | -1.165 | plus |
| | Bud_03 | Ado-MIR444c-1-3_b | Ado-MIR444c-1c_b | -1.267 | plus |
| | culm_34 | Ado-MIR444c-1-1_c | Ado-MIR444c-1a_c | -0.902 | minus |
| Locus_58 | culm_34 | Ado-MIR444c-1-2_c | Ado-MIR444c-1b_c | -1.165 | plus |
| | culm_34 | Ado-MIR444c-1-3_c | Ado-MIR444c-1c_c | -1.267 | plus |
| | root_66 | Ado-MIR444c-1-1_r | Ado-MIR444c-1a_r | -1.165 | plus |
| | root_66 | Ado-MIR444c-1-2_r | Ado-MIR444c-1b_r | -0.902 | minus |
| | root_66 | Ado-MIR444c-1-3_r | Ado-MIR444c-1c_r | -1.267 | plus |
| | Bud_70 | Ado-MIR444c-2-1_b | Ado-MIR444c-2a_b | -1.289 | minus |
| Locus_59 | Bud_70 | Ado-MIR444c-2-2_b | Ado-MIR444c-2b_b | -1.372 | minus |
| | Bud_70 | Ado-MIR444c-2-3_b | Ado-MIR444c-2c_b | -1.24 | plus |
| Locus_60 | root_04 | Ado-MIR444c-3-1_r | Ado-MIR444c-3a_r | -1.336 | minus |
| | root_04 | Ado-MIR444c-3-2_r | Ado-MIR444c-3b_r | -1.298 | minus |
| Locus_61 | culm_17 | Ado-MIR444c-4-1_c | Ado-MIR444c-4a_c | -1.298 | minus |
| | culm_17 | Ado-MIR444c-4-2_c | Ado-MIR444c-4b_c | -1.336 | minus |
| | Bud_13 | Ado-MIR444d-1-1_b | Ado-MIR444d-1a_b | -1.19 | minus |
| Locus_62 | Bud_13 | Ado-MIR444d-1-2_b | Ado-MIR444d-1b_b | -1.157 | minus |
| | Bud_13 | Ado-MIR444d-1-3_b | Ado-MIR444d-1c_b | -1.147 | minus |
| | Bud_13 | Ado-MIR444d-1-4_b | Ado-MIR444d-1d_b | -1.263 | minus |
| | Bud_62 | Ado-MIR444d-2-1_b | Ado-MIR444d-2a_b | -1.082 | plus |
| | Bud_62 | Ado-MIR444d-2-2_b | Ado-MIR444d-2b_b | -1.054 | plus |
| | Bud_62 | Ado-MIR444d-2-3_b | Ado-MIR444d-2c_b | -1.147 | plus |
| | Bud_62 | Ado-MIR444d-2-4_b | Ado-MIR444d-2d_b | -1.14 | plus |
| | culm_64 | Ado-MIR444d-2-1_c | Ado-MIR444d-2a_c | -1.082 | plus |
| Locus_63 | culm_64 | Ado-MIR444d-2-2_c | Ado-MIR444d-2b_c | -1.054 | plus |
| | culm_64 | Ado-MIR444d-2-3_c | Ado-MIR444d-2c_c | -1.147 | plus |
| | culm_64 | Ado-MIR444d-2-4_c | Ado-MIR444d-2d_c | -1.14 | plus |
| | leaf_09 | Ado-MIR444d-2-1_l | Ado-MIR444d-2a_l | -1.082 | plus |
| | leaf_09 | Ado-MIR444d-2-2_l | Ado-MIR444d-2b_l | -1.054 | plus |
| | leaf_09 | Ado-MIR444d-2-3_l | Ado-MIR444d-2c_l | -1.147 | plus |
| | leaf_09 | Ado-MIR444d-2-4_l | Ado-MIR444d-2d_l | -1.14 | plus |
| Locus_64 | culm_63 | Ado-MIR444d-3-1_c | Ado-MIR444d-3a_c | -1.19 | plus |
| | culm_63 | Ado-MIR444d-3-2_c | Ado-MIR444d-3b_c | -1.157 | plus |

| | | | | | |
|---|---|---|---|---|---|
| Locus_65 | culm_63 | Ado-MIR444d-3-3_c | Ado-MIR444d-3c_c | -1.147 | plus |
| | culm_63 | Ado-MIR444d-3-4_c | Ado-MIR444d-3d_c | -1.263 | plus |
| | leaf_45 | Ado-MIR444d-4-1_l | Ado-MIR444d-4a_l | -1.19 | minus |
| | leaf_45 | Ado-MIR444d-4-2_l | Ado-MIR444d-4b_l | -1.157 | minus |
| | leaf_45 | Ado-MIR444d-4-3_l | Ado-MIR444d-4c_l | -1.147 | minus |
| | leaf_45 | Ado-MIR444d-4-4_l | Ado-MIR444d-4d_l | -1.263 | minus |
| Locus_66 | root_23 | Ado-MIR444d-5-1_r | Ado-MIR444d-5a_r | -1.082 | plus |
| | root_23 | Ado-MIR444d-5-2_r | Ado-MIR444d-5b_r | -1.054 | plus |
| | root_23 | Ado-MIR444d-5-3_r | Ado-MIR444d-5c_r | -1.147 | plus |
| | root_23 | Ado-MIR444d-5-4_r | Ado-MIR444d-5d_r | -1.14 | plus |
| Locus_67 | root_33 | Ado-MIR444e-1_r | Ado-MIR444e-a_r | -1.235 | minus |
| | root_33 | Ado-MIR444e-2_r | Ado-MIR444e-b_r | -1.191 | minus |
| | root_33 | Ado-MIR444e-2_r | Ado-MIR444e-c_r | -1.191 | minus |
| Locus_68 | culm_73 | Ado-MIR529a_c | Ado-MIR529a-a_c | -0.924 | plus |
| | culm_73 | Ado-MIR529a_c | Ado-MIR529a-b_c | -0.924 | plus |
| Locus_69 | Bud_22 | Ado-MIR827-1_b | Ado-MIR827-a_b | -1.042 | minus |
| | root_59 | Ado-MIR827-2_r | Ado-MIR827-b_r | -1.042 | plus |

**Supplementary Table 1.2.** Distribution in different tissues of mature miRNA by family.

| miRNA families | Number of miRNA | Bud | Culm | Leaf | Root |
|---|---|---|---|---|---|
| MIR156 | 13 | 0 | 0 | 2 | 11 |
| MIR160 | 4 | 1 | 1 | 0 | 2 |
| MIR166 | 2 | 0 | 0 | 0 | 2 |
| MIR167 | 5 | 2 | 1 | 1 | 1 |
| MIR169 | 21 | 4 | 3 | 6 | 8 |
| MIR171 | 6 | 2 | 3 | 0 | 1 |
| MIR172 | 9 | 3 | 2 | 2 | 2 |
| MIR319 | 4 | 0 | 2 | 0 | 2 |
| MIR393 | 11 | 3 | 3 | 3 | 2 |
| MIR396 | 2 | 0 | 2 | 0 | 0 |
| MIR444 | 55 | 16 | 16 | 11 | 12 |
| MIR827 | 2 | 1 | 0 | 0 | 1 |
| MIR529 | 2 | 0 | 2 | 0 | 0 |
| MIR1430 | 5 | 0 | 2 | 2 | 1 |

**Supplementary Table 1.3.** Base composition of *A. donax* miRNAs.

| Position | Adenine | | Cytosine | | Guanine | | Uracil | |
|---|---|---|---|---|---|---|---|---|
| | No. | Percentage (%) | No. | Percentage (%) | No. | Percentage (%) | No. | Percentage (%) |
| 1 | 7 | 4.96 | 6 | 4.26 | 8 | 5.67 | 120 | 85.11 |
| 2 | 26 | 18.44 | 16 | 11.35 | 74 | 52.48 | 25 | 17.73 |
| 3 | 39 | 27.66 | 37 | 26.24 | 49 | 34.75 | 16 | 11.35 |
| 4 | 47 | 33.33 | 49 | 34.75 | 9 | 6.38 | 36 | 25.53 |
| 5 | 30 | 21.28 | 25 | 17.73 | 53 | 37.59 | 33 | 23.4 |
| 6 | 38 | 26.95 | 30 | 21.28 | 44 | 31.21 | 29 | 20.57 |
| 7 | 46 | 32.62 | 39 | 27.66 | 18 | 12.77 | 38 | 26.95 |
| 8 | 22 | 15.6 | 16 | 11.35 | 74 | 52.48 | 29 | 20.57 |
| 9 | 16 | 11.35 | 38 | 26.95 | 50 | 35.46 | 37 | 26.24 |
| 10 | 71 | **50.35** | 29 | 20.57 | 6 | 4.26 | 35 | 24.82 |
| 11 | 31 | 21.99 | 20 | 14.18 | 47 | 33.33 | 43 | 30.5 |
| 12 | 29 | 20.57 | 38 | 26.95 | 51 | 36.17 | 23 | 16.31 |
| 13 | 27 | 19.15 | 46 | 32.62 | 53 | 37.59 | 15 | 10.64 |
| 14 | 16 | 11.35 | 47 | 33.33 | 12 | 8.51 | 66 | 46.81 |
| 15 | 11 | 7.8 | 43 | 30.5 | 26 | 18.44 | 61 | 43.26 |
| 16 | 48 | 34.04 | 16 | 11.35 | 20 | 14.18 | 57 | 40.43 |
| 17 | 47 | 33.33 | 17 | 12.06 | 48 | 34.04 | 29 | 20.57 |
| 18 | 5 | 3.55 | 53 | 37.59 | 57 | 40.43 | 26 | 18.44 |
| 19 | 35 | 24.82 | 71 | **50.35** | 13 | 9.22 | 22 | 15.6 |
| 20 | 12 | 8.51 | 45 | 31.91 | 16 | 11.35 | 68 | 48.23 |
| 21 | 30 | 23.08 | 44 | 33.85 | 25 | 19.23 | 31 | 23.85 |
| 22 | 0 | 0 | 1 | 11.11 | 5 | 55.56 | 3 | 33.33 |
| **Overall** | **633** | **450.73** | **726** | **527.94** | **758** | **591.1** | **842** | **630.24** |

Base composition of *O. sativa* miRNAs

| Position | Adenine | | Cytosine | | Guanine | | Uracil | |
|---|---|---|---|---|---|---|---|---|
| | **No.** | **Percentage (%)** | **No.** | **Percentage (%)** | **No.** | **Percentage (%)** | **No.** | **Percentage (%)** |
| 1 | 4 | 6.78 | 3 | 5.08 | 15 | 25.42 | 37 | **62.71** |
| 2 | 7 | 11.86 | 13 | 22.03 | 29 | 49.15 | 10 | 16.95 |
| 3 | 20 | 33.9 | 11 | 18.64 | 17 | 28.81 | 11 | 18.64 |
| 4 | 12 | 20.34 | 25 | 42.37 | 7 | 11.86 | 15 | 25.42 |
| 5 | 22 | 37.29 | 7 | 11.86 | 17 | 28.81 | 13 | 22.03 |
| 6 | 15 | 25.42 | 17 | 28.81 | 18 | 30.51 | 9 | 15.25 |
| 7 | 20 | 33.9 | 10 | 16.95 | 9 | 15.25 | 20 | 33.9 |
| 8 | 14 | 23.73 | 12 | 20.34 | 26 | 44.07 | 7 | 11.86 |
| 9 | 9 | 15.25 | 17 | 28.81 | 21 | 35.59 | 12 | 20.34 |
| 10 | 21 | **35.59** | 16 | 27.12 | 7 | 11.86 | 15 | 25.42 |
| 11 | 8 | 13.56 | 11 | 18.64 | 24 | 40.68 | 16 | 27.12 |
| 12 | 12 | 20.34 | 18 | 30.51 | 14 | 23.73 | 15 | 25.42 |
| 13 | 11 | 18.64 | 18 | 30.51 | 19 | 32.2 | 11 | 18.64 |
| 14 | 7 | 11.86 | 18 | 30.51 | 10 | 16.95 | 24 | 40.68 |
| 15 | 7 | 11.86 | 17 | 28.81 | 14 | 23.73 | 21 | 35.59 |
| 16 | 23 | 38.98 | 6 | 10.17 | 9 | 15.25 | 21 | 35.59 |
| 17 | 15 | 25.42 | 8 | 13.56 | 23 | 38.98 | 13 | 22.03 |
| 18 | 6 | 10.17 | 23 | 38.98 | 15 | 25.42 | 15 | 25.42 |
| 19 | 27 | 45.76 | 24 | **40.68** | 1 | 1.69 | 7 | 11.86 |
| 20 | 13 | 22.03 | 17 | 28.81 | 6 | 10.17 | 23 | 38.98 |
| 21 | 11 | 22.45 | 13 | 26.53 | 15 | 30.61 | 10 | 20.41 |
| 22 | 0 | 0 | 4 | 50 | 3 | 37.5 | 1 | 12.5 |
| **Overall** | **284** | **485.13** | **308** | **569.72** | **319** | **578.24** | **326** | **566.76** |

**NOTE**: in **bold red** the interspecific differences at specific positions highlighted in the text.

Summary of composition

| **Specific position** | ***Arundo donax*** | ***Oryza sativa*** |
|---|---|---|
| Adenine | 21.39% | 22.96% |
| Uracil | 28.46% | 26.35% |
| Guanine | 25.62% | 25.79% |
| Cytosine | 24.54% | 24.90% |
| G+C | 50.15% | 50.69% |
| A+U | 49.85% | 49.31% |

**Supplementary Table 2.1.** Features of putative miRNA identified from *Arundo* genus.

| Predicted mature miRNA | Number of mismatches (miRNA and miRNA*) | Precursor miRNA MFE(kcal/mol) | Precursor miRNA MFEI(kcal/mol) | homolog mismatches | strand |
|---|---|---|---|---|---|
| Aco-MIR5161-1 | 4 | -15.8 | -0.686 | 2 | plus |
| Aco-MIR1430-2 | 3 | -49.2 | -1.093 | 1 | minus |
| Aco-MIR437-3 | 2 | -56.8 | -1.071 | 4 | plus |
| Aco-MIR2905-4 | 2 | -20.4 | -0.927 | 3 | plus |
| Aco-MIR169q-5a | 2 | -38.7 | -0.841 | 0 | minus |
| Aco-MIR169q-5b | 1 | -31.8 | -0.676 | 1 | plus |
| Aco-MIR172d-6a | 2 | -40.6 | -1.015 | 0 | minus |
| Aco-MIR172d-6b | 1 | -38.4 | -0.936 | 1 | plus |
| Aco-MIR1432-7 | 3 | -39.7 | -0.863 | 1 | minus |
| Aco-MIR169g-8 | 2 | -40.7 | -0.782 | 0 | plus |
| Aco-MIR160b-9 | 2 | -37.1 | -0.727 | 0 | plus |
| Aco-MIR408-10 | 3 | -67.4 | -0.694 | 0 | plus |
| Aco-MIR408-11 | 3 | -67.4 | -0.694 | 0 | plus |
| Aco-MIR156j-12 | 0 | -50.2 | -0.822 | 1 | minus |
| Aco-MIR444d-13a | 0 | -26.4 | -1.147 | 0 | plus |
| Aco-MIR444d-13b | 2 | -55.6 | -1.263 | 0 | plus |
| Aco-MIR444d-14a | 0 | -26.4 | -1.147 | 0 | plus |
| Aco-MIR444d-14b | 3 | -50.2 | -1.167 | 0 | plus |
| Aco-MIR444d-14c | 3 | -54.1 | -1.104 | 1 | plus |
| Aco-MIR444d-14d | 4 | -54.8 | -1.074 | 1 | plus |
| Aco-MIR162b-15 | 2 | -46.4 | -0.692 | 0 | minus |
| Aco-MIR156d-16 | 1 | -31.7 | -0.754 | 2 | minus |
| Aco-MIR166d-17 | 3 | -31.4 | -0.713 | 0 | minus |
| Aco-MIR159b-18 | 2 | -79.7 | -0.821 | 0 | plus |
| Aco-MIR159a-19 | 0 | -82.7 | -0.731 | 0 | plus |
| Aco-MIR159a-20 | 0 | -86.8 | -0.748 | 0 | plus |
| Aco-MIR399i-21 | 1 | -39.9 | -0.814 | 2 | minus |
| Aco-MIR528-22 | 2 | -39.8 | -0.796 | 0 | plus |
| Aco-MIR528-23 | 2 | -43.6 | -0.854 | 0 | plus |
| Aco-MIR5161-24a | 0 | -88.7 | -0.751 | 3 | plus |
| Aco-MIR5161-24b | 1 | -97.5 | -0.833 | 3 | minus |
| Aco-MIR5161-24c | 0 | -83.3 | -0.764 | 3 | minus |
| Aco-MIR168a-25 | 3 | -37.9 | -0.806 | 0 | plus |
| Aco-MIR168a-26 | 3 | -37.9 | -0.806 | 0 | plus |
| Aco-MIR818f-27 | 0 | -98.4 | -1.587 | 3 | minus |
| Aco-MIR818f-28a | 1 | -79.7 | -1.245 | 3 | minus |
| Aco-MIR818f-28b | 3 | -56.5 | -1.177 | 3 | minus |
| Aco-MIR812h-29 | 1 | -15.9 | -0.795 | 3 | minus |
| Aco-MIR167h-30 | 1 | -43.9 | -0.77 | 0 | minus |
| Aco-MIR167d-31 | 0 | -44.8 | -0.995 | 0 | minus |

| | | | | | |
|---|---|---|---|---|---|
| Aco-MIR159d-32 | 4 | -48.3 | -0.743 | 1 | plus |
| Aco-MIR5161-33 | 0 | -76.5 | -0.859 | 3 | plus |
| Aco-MIR164a-34 | 2 | -81.2 | -0.773 | 0 | plus |
| Adof-MIR408-1 | 3 | -67.4 | -0.694 | 0 | minus |
| Adof-MIR444d-2 | 0 | -26.4 | -1.147 | 0 | plus |
| Adof-MIR1862e-3 | 2 | -100.4 | -0.717 | 3 | plus |
| Adof-MIR169q-4a | 2 | -38.7 | -0.841 | 0 | minus |
| Adof-MIR169q-4b | 1 | -31.8 | -0.676 | 1 | plus |
| Adof-MIR167h-5 | 1 | -43.9 | -0.77 | 0 | minus |
| Adof-MIR162b-6 | 2 | -46.4 | -0.692 | 0 | plus |
| Adof-MIR444a-7a | 1 | -64.5 | -1.372 | 0 | minus |
| Adof-MIR444a-8b | 2 | -67.5 | -1.273 | 1 | minus |
| Adof-MIR159a-9 | 0 | -85.4 | -0.749 | 0 | minus |
| Adof-MIR159a-10 | 0 | -82.7 | -0.731 | 0 | minus |
| Adof-MIR169g-11 | 2 | -40.7 | -0.782 | 0 | minus |
| Adof-MIR444c-12a | 1 | -54.8 | -1.191 | 0 | minus |
| Adof-MIR444c-12b | 1 | -50.7 | -1.3 | 0 | minus |
| Adof-MIR444c-12c | 3 | -36.1 | -0.925 | 1 | plus |
| Adof-MIR168a-13 | 3 | -37.9 | -0.806 | 0 | minus |
| Adof-MIR169q-14 | 2 | -41.3 | -0.983 | 0 | plus |
| Adof-MIR1430-15 | 2 | -39.2 | -0.834 | 1 | plus |
| Adof-MIR1430-16 | 2 | -39.2 | -0.834 | 1 | plus |
| Adof-MIR166d-17 | 3 | -31.4 | -0.713 | 0 | minus |
| Adof-MIR528-18 | 2 | -39.8 | -0.796 | 0 | plus |
| Adof-MIR156j-19 | 2 | -46.6 | -0.847 | 1 | plus |
| Adof-MIR156j-20 | 0 | -50.2 | -0.822 | 1 | plus |
| Adof-MIR437-21a | 1 | -92.3 | -1.153 | 3 | minus |
| Adof-MIR437-22b | 1 | -89.7 | -1.121 | 3 | plus |
| Adof-MIR437-23 | 3 | -83.7 | -1.059 | 3 | plus |
| Adof-MIR5161-24 | 2 | -41.8 | -1.129 | 3 | minus |
| Ado-MIR159a-1 | 0 | -80.6 | -0.713 | 0 | minus |
| Ado-MIR444a-2a | 2 | -54.9 | -1.247 | 0 | minus |
| Ado-MIR444a-2b | 3 | -57.9 | -1.158 | 1 | minus |
| Ado-MIR162a-3 | 2 | -47.5 | -0.73 | 0 | minus |
| Ado-MIR169q-4a | 2 | -38.7 | -0.841 | 0 | plus |
| Ado-MIR169q-4b | 1 | -31.8 | -0.676 | 1 | minus |
| Ado-MIR169q-5a | 2 | -39.5 | -0.858 | 0 | plus |
| Ado-MIR169q-5b | 2 | -33.4 | -0.726 | 0 | minus |
| Ado-MIR169q-6a | 2 | -41.4 | -0.752 | 0 | plus |
| Ado-MIR169q-6b | 4 | -35.1 | -0.662 | 1 | minus |
| Ado-MIR167h-7 | 1 | -43.9 | -0.77 | 0 | minus |
| Ado-MIR166d-8 | 3 | -30.1 | -0.752 | 0 | minus |
| Ado-MIR169n-9 | 2 | -52.5 | -0.86 | 1 | plus |
| Ado-MIR530-10 | 2 | -36.4 | -0.887 | 1 | plus |
| Ado-MIR444d-11a | 0 | -26.4 | -1.147 | 0 | plus |
| Ado-MIR444d-11b | 3 | -51.3 | -1.14 | 0 | plus |

| | | | | | |
|---|---|---|---|---|---|
| Ado-MIR444d-11c | 2 | -55.2 | -1.082 | 1 | plus |
| Ado-MIR444d-11d | 3 | -55.9 | -1.054 | 1 | plus |
| Ado-MIR444d-12a | 0 | -44.1 | -0.678 | 0 | plus |
| Ado-MIR444f-12b | 3 | -69 | -0.793 | 0 | plus |
| Ado-MIR444f-12c | 2 | -72.9 | -0.783 | 1 | plus |
| Ado-MIR444f-12d | 3 | -73.6 | -0.774 | 1 | plus |
| Ado-MIR169q-13 | 2 | -40.9 | -0.951 | 0 | plus |
| Ado-MIR159b-14 | 2 | -79.7 | -0.821 | 0 | plus |
| Ado-MIR5143b-15 | 2 | -24 | -1.043 | 3 | plus |
| Ado-MIR167g-16 | 0 | -33.1 | -0.945 | 0 | plus |
| Ado-MIR167d-17 | 0 | -41.7 | -1.017 | 0 | plus |
| Ado-MIR437-18 | 2 | -86.6 | -1.332 | 4 | minus |
| Ado-MIR437-19 | 2 | -83.4 | -1.345 | 4 | minus |
| Ado-MIR444c-20a | 1 | -52.8 | -1.353 | 0 | plus |
| Ado-MIR444c-20b | 1 | -61.1 | -1.272 | 1 | plus |
| Ado-MIR5161-21 | 1 | -89.7 | -0.723 | 2 | plus |
| Ado-MIR172d-22a | 2 | -40.5 | -1.012 | 0 | minus |
| Ado-MIR172d-22b | 1 | -37.9 | -0.924 | 1 | plus |
| Ado-MIR818d-23 | 3 | -60.6 | -0.73 | 3 | plus |
| Ado-MIR1879-24 | 4 | -64 | -0.673 | 0 | minus |
| Ado-MIR393a-25a | 1 | -55.6 | -0.712 | 0 | minus |
| Ado-MIR393b-25b | 0 | -64.5 | -0.816 | 1 | plus |
| Afo-MIR162a-1 | 2 | -47.5 | -0.73 | 0 | minus |
| Afo-MIR393b-2 | 1 | -63.2 | -0.854 | 0 | plus |
| Afo-MIR528-3 | 2 | -39.8 | -0.796 | 0 | plus |
| Afo-MIR169g-4 | 2 | -37.9 | -0.758 | 0 | plus |
| Afo-MIR167d-5 | 0 | -42.1 | -0.979 | 0 | minus |
| Afo-MIR168a-6 | 3 | -38 | -0.791 | 0 | plus |
| Afo-MIR159b-7 | 3 | -76.9 | -0.769 | 0 | minus |
| Afo-MIR159b-8 | 3 | -76.9 | -0.769 | 0 | minus |
| Afo-MIR156k-9 | 1 | -35.3 | -0.692 | 1 | plus |
| Afo-MIR156g-10 | 0 | -53.6 | -0.864 | 0 | minus |
| Afo-MIR167h-11 | 1 | -43.9 | -0.828 | 0 | minus |
| Afo-MIR167h-12 | 1 | -43.9 | -0.828 | 0 | minus |
| Afo-MIR444c-13a | 1 | -54.8 | -1.165 | 0 | minus |
| Afo-MIR444c-13b | 1 | -50.7 | -1.267 | 0 | minus |
| Afo-MIR444c-13c | 3 | -36.1 | -0.902 | 1 | plus |
| Ama-MIR156j-1 | 0 | -54.4 | -0.863 | 1 | plus |
| Ama-MIR5143b-2 | 2 | -24 | -1.043 | 3 | minus |
| Ama-MIR399i-3 | 2 | -32.8 | -0.697 | 1 | minus |
| Ama-MIR444a-4a | 2 | -54.9 | -1.247 | 0 | minus |
| Ama-MIR444a-4b | 3 | -57.9 | -1.158 | 1 | minus |
| Ama-MIR393b-5 | 1 | -62.4 | -0.866 | 0 | plus |
| Ama-MIR166d-6 | 3 | -30.1 | -0.752 | 0 | plus |
| Ama-MIR162a-7 | 2 | -52.9 | -0.745 | 0 | minus |
| Ama-MIR172d-8a | 2 | -40.5 | -1.012 | 0 | minus |

| | | | | | |
|---|---|---|---|---|---|
| Ama-MIR172d-8b | 1 | -37.9 | -0.924 | 1 | plus |
| Ama-MIR167e-9 | 1 | -56.1 | -0.719 | 0 | minus |
| Ama-MIR159b-10 | 3 | -76.4 | -0.727 | 0 | minus |
| Ama-MIR159b-11 | 2 | -79.7 | -0.821 | 0 | minus |
| Ama-MIR156g-12 | 0 | -58.3 | -0.832 | 0 | minus |
| Ama-MIR156j-13 | 0 | -54.4 | -0.863 | 1 | plus |
| Ama-MIR160b-14 | 2 | -37.1 | -0.727 | 0 | plus |
| Ama-MIR528-15 | 2 | -39.8 | -0.796 | 0 | minus |
| Ama-MIR167h-16 | 1 | -43.9 | -0.828 | 0 | minus |
| Ama-MIR167h-17 | 1 | -43.9 | -0.828 | 0 | minus |
| Ama-MIR167h-18 | 1 | -43.9 | -0.77 | 0 | minus |
| Ama-MIR2905-19 | 1 | -30.3 | -1.122 | 2 | minus |
| Ama-MIR1435-20 | 2 | -44.5 | -1.171 | 2 | plus |
| Ama-MIR437-21 | 2 | -46.5 | -1.223 | 3 | plus |
| Ama-MIR437-22a | 2 | -64.6 | -1.133 | 2 | minus |
| Ama-MIR437-22b | 1 | -110.6 | -1.316 | 2 | minus |
| Ama-MIR5337b-22c | 2 | -55.2 | -0.665 | 2 | minus |
| Ama-MIR818e-23 | 1 | -94.5 | -1.453 | 3 | minus |
| Ama-MIR818f-24 | 2 | -64.7 | -1.617 | 3 | plus |
| Ama-MIR159a-25 | 0 | -80.6 | -0.713 | 0 | minus |
| Ama-MIR444c-26a | 1 | -52.8 | -1.353 | 0 | plus |
| Ama-MIR444c-26b | 1 | -61.1 | -1.272 | 1 | plus |
| Ama-MIR408-27 | 3 | -72.2 | -0.76 | 0 | minus |
| Ama-MIR1435-28 | 1 | -51.9 | -1.179 | 3 | plus |
| Ama-MIR2905-29 | 1 | -36.3 | -0.981 | 4 | minus |
| Ama-MIR5831-30 | 2 | -56.4 | -1.175 | 4 | minus |
| Ama-MIR437-31 | 2 | -49.1 | -1.067 | 3 | plus |
| Ami-MIR827-1 | 1 | -34.9 | -0.894 | 0 | minus |
| Ami-MIR169g-2 | 2 | -40.7 | -0.782 | 0 | plus |
| Ami-MIR408-3 | 3 | -67.4 | -0.694 | 0 | plus |
| Ami-MIR167d-4 | 0 | -38 | -0.926 | 0 | minus |
| Ami-MIR444c-5a | 1 | -54.9 | -1.372 | 0 | minus |
| Ami-MIR444c-5b | 1 | -63.2 | -1.289 | 1 | minus |
| Ami-MIR444c-5c | 1 | -49.6 | -1.24 | 1 | plus |
| Ami-MIR1866-6 | 4 | -19.3 | -0.689 | 4 | plus |
| Ami-MIR166d-7 | 3 | -31.4 | -0.713 | 0 | minus |
| Ami-MIR159b-8 | 2 | -79.7 | -0.821 | 0 | minus |
| Ami-MIR528-9 | 2 | -39.8 | -0.796 | 0 | minus |
| Ami-MIR172d-10a | 2 | -40.5 | -1.012 | 0 | minus |
| Ami-MIR172d-10b | 1 | -37.9 | -0.924 | 1 | plus |
| Ami-MIR159a-11 | 0 | -82.7 | -0.731 | 0 | plus |
| Ami-MIR156j-12 | 0 | -50.2 | -0.822 | 1 | minus |
| Ami-MIR162b-13 | 2 | -47.9 | -0.736 | 0 | minus |
| Ami-MIR5824-14 | 0 | -69.5 | -1.287 | 4 | plus |
| Ami-MIR818f-15 | 1 | -94.2 | -1.365 | 3 | plus |
| Ami-MIR818a-16 | 2 | -21.7 | -1.142 | 3 | minus |

| | | | | | |
|---|---|---|---|---|---|
| Ami-MIR168a-17 | 3 | -38 | -0.791 | 0 | minus |
| Ami-MIR11340-18 | 2 | -63.3 | -0.688 | 3 | minus |
| Ami-MIR5161-19a | 2 | -68.8 | -1.563 | 4 | plus |
| Ami-MIR5161-19b | 2 | -74 | -1.608 | 4 | plus |
| Apl-MIR812a-1 | 1 | -82.9 | -1.535 | 3 | minus |
| Apl-MIR169q-2 | 2 | -41.3 | -0.983 | 0 | minus |
| Apl-MIR396e-3a | 2 | -64.5 | -0.777 | 0 | minus |
| Apl-MIR396e-3b | 2 | -66 | -0.776 | 1 | minus |
| Apl-MIR169b-4 | 2 | -41.1 | -0.79 | 3 | minus |
| Apl-MIR2275d-5 | 1 | -34 | -0.918 | 1 | plus |
| Apl-MIR5161-6a | 2 | -41.6 | -1.155 | 3 | minus |
| Apl-MIR5161-6b | 2 | -42.8 | -1.156 | 3 | plus |
| Apl-MIR1430-7 | 4 | -45.5 | -1.011 | 1 | plus |
| Apl-MIR827-8 | 1 | -34.9 | -0.894 | 0 | plus |
| Apl-MIR1435-9 | 3 | -44.3 | -0.886 | 4 | minus |
| Apl-MIR399j-10 | 3 | -33.4 | -0.742 | 0 | plus |
| Apl-MIR169g-11 | 2 | -40.7 | -0.782 | 0 | minus |
| Apl-MIR162b-12 | 2 | -46.4 | -0.692 | 0 | plus |
| Apl-MIR159b-13 | 2 | -79.5 | -0.828 | 0 | minus |
| Apl-MIR167h-14 | 1 | -43.9 | -0.77 | 0 | minus |
| Apl-MIR167h-15 | 1 | -43.9 | -0.812 | 0 | minus |
| Apl-MIR399c-16a | 1 | -38.8 | -0.76 | 0 | minus |
| Apl-MIR399f-16b | 3 | -40.6 | -0.99 | 0 | plus |
| Apl-MIR399c-17 | 1 | -38.8 | -0.76 | 0 | minus |
| Apl-MIR166d-18 | 3 | -31.4 | -0.713 | 0 | minus |
| Apl-MIR444a-19a | 2 | -54.9 | -1.247 | 0 | minus |
| Apl-MIR444a-19b | 3 | -57.9 | -1.158 | 1 | minus |
| Apl-MIR160b-20 | 2 | -37.1 | -0.727 | 0 | plus |
| Apl-MIR172d-21a | 2 | -40.6 | -1.015 | 0 | plus |
| Apl-MIR172d-21b | 1 | -38.4 | -0.936 | 1 | minus |
| Apl-MIR172d-22a | 2 | -40.6 | -1.015 | 0 | plus |
| Apl-MIR172d-22b | 1 | -38.4 | -0.936 | 1 | minus |
| Apl-MIR172d-23 | 1 | -40.4 | -1.01 | 0 | minus |
| Apl-MIR159a-24 | 0 | -82.7 | -0.731 | 0 | plus |
| Apl-MIR159a-25 | 0 | -82.7 | -0.731 | 0 | plus |
| Apl-MIR5143b-26 | 2 | -24 | -1.043 | 3 | minus |
| Apl-MIR5161-27 | 1 | -97.5 | -0.826 | 2 | minus |
| Apl-MIR818d-28 | 1 | -75.1 | -0.79 | 1 | plus |
| Apl-MIR1439-29 | 1 | -44.6 | -1.311 | 3 | plus |
| Apl-MIR444c-30a | 1 | -54.8 | -1.191 | 0 | plus |
| Apl-MIR444c-30b | 1 | -50.7 | -1.3 | 0 | plus |
| Apl-MIR444c-30c | 3 | -36.1 | -0.925 | 1 | minus |
| Apl-MIR818d-31 | 1 | -63.6 | -0.978 | 2 | minus |
| Apl-MIR818d-32 | 1 | -63.6 | -0.978 | 2 | minus |
| Apl-MIR818d-33 | 2 | -40.8 | -0.728 | 2 | minus |
| Apl-MIR437-34 | 3 | -83.7 | -1.059 | 3 | minus |

| | | | | | |
|---|---|---|---|---|---|
| Apl-MIR812q-35 | 1 | -89.5 | -1.598 | 4 | plus |
| Apl-MIR818f-36 | 4 | -64.9 | -1.03 | 4 | minus |
| Apl-MIR818f-37a | 0 | -78.8 | -1.176 | 2 | minus |
| Apl-MIR818f-37b | 2 | -53.8 | -1.034 | 2 | minus |
| Apl-MIR1439-38 | 1 | -74.4 | -0.808 | 3 | minus |
| Apl-MIR1439-39 | 2 | -70.3 | -1.495 | 2 | plus |
| Apl-MIR5161-40 | 1 | -85.7 | -0.726 | 2 | minus |
| Apl-MIR408-41 | 3 | -67.9 | -0.7 | 0 | minus |
| Apl-MIR528-42 | 2 | -43.6 | -0.854 | 0 | minus |
| Apl-MIR171i-43 | 1 | -38.3 | -0.957 | 0 | plus |
| Apl-MIR818d-44 | 3 | -68.7 | -0.798 | 3 | minus |

**Supplementary Table 2.2.** Putative targets of *Arundo* miRNAs.

| miRNA name | Target name | Expectation | Unpaired Energy (kcal/mol) | Inhibition |
|---|---|---|---|---|
| Ado-MIR159a-1 | donax_20465 | 3 | 20.796 | Cleavage |
| Ado-MIR159b-14 | donax_20465 | 3 | 20.796 | Cleavage |
| Adof-MIR444a-7a | donaciforms_095716 | 2.5 | 18.167 | Cleavage |
| Adof-MIR444c-12b | donaciforms_095716 | 2 | 18.167 | Cleavage |
| Adof-MIR444c-12c | donaciforms_095716 | 3.5 | 18.167 | Cleavage |
| Apl-MIR1435-9 | plinii_133253 | 4 | 13.967 | Translation |
| Apl-MIR5161-6b | plinii_070980 | 4 | 11.083 | Translation |
| Aco-MIR444d-13b | collina_092944 | 3 | 19.005 | Cleavage |
| Aco-MIR444d-14b | collina_092944 | 3 | 19.005 | Cleavage |
| Ama-MIR172d-8a | macrophylla_093426 | 4 | 23.117 | Cleavage |
| Aco-MIR1432-7 | collina_065806 | 3.5 | 17.509 | Cleavage |
| Ami-MIR5161-19b | micrantha_033572 | 4 | 11.979 | Translation |
| Ami-MIR818a-16 | micrantha_033572 | 3.5 | 11.206 | Cleavage |
| Ami-MIR818f-15 | micrantha_033572 | 2.5 | 11.206 | Cleavage |
| Aco-MIR5161-1 | collina_101612 | 1.5 | 11.785 | Cleavage |
| Aco-MIR5161-24a | collina_101612 | 3.5 | 11.785 | Cleavage |
| Aco-MIR5161-24b | collina_101612 | 4 | 11.785 | Cleavage |
| Aco-MIR5161-33 | collina_101612 | 3 | 11.785 | Cleavage |
| Aco-MIR818f-27 | collina_101612 | 2.5 | 11.785 | Cleavage |
| Adof-MIR444a-7a | donaciforms_061648 | 4 | 19.302 | Translation |
| Ado-MIR5161-21 | donax_50601 | 2.5 | 15.365 | Translation |
| Apl-MIR444a-19a | plinii_108441 | 4 | 22.912 | Cleavage |
| Apl-MIR444c-30c | plinii_108441 | 3.5 | 22.912 | Cleavage |
| Apl-MIR444a-19a | plinii_108445 | 4 | 22.912 | Cleavage |
| Apl-MIR444c-30c | plinii_108445 | 3.5 | 22.912 | Cleavage |
| Adof-MIR1862e-3 | donaciforms_075193 | 4 | 18.798 | Cleavage |
| Adof-MIR5161-24 | donaciforms_075193 | 3.5 | 19.519 | Translation |
| Apl-MIR5161-6b | plinii_091022 | 3.5 | 17.48 | Translation |
| Aco-MIR156d-16 | collina_120304 | 3.5 | 13.274 | Translation |
| Afo-MIR156g-10 | formosana_57367 | 4 | 14.189 | Translation |
| Ama-MIR156g-12 | macrophylla_101942 | 4 | 10.224 | Translation |
| Adof-MIR528-18 | donaciforms_038175 | 3 | 14.53 | Cleavage |
| Afo-MIR528-3 | formosana_13913 | 2.5 | 16.281 | Cleavage |
| Ado-MIR5161-21 | donax_35717 | 2 | 13.165 | Cleavage |
| Apl-MIR5161-6b | plinii_133670 | 4 | 8.795 | Translation |
| Ami-MIR528-9 | micrantha_030175 | 3.5 | 17.042 | Cleavage |
| Apl-MIR1435-9 | plinii_064160 | 2.5 | 15.889 | Cleavage |
| Apl-MIR437-34 | plinii_064160 | 3.5 | 15.889 | Cleavage |
| Apl-MIR812a-1 | plinii_064160 | 4 | 15.889 | Cleavage |
| Ado-MIR172d-22b | donax_16384 | 4 | 22.38 | Cleavage |
| Ado-MIR172d-22b | donax_16385 | 4 | 22.341 | Cleavage |
| Afo-MIR156g-10 | formosana_48593 | 4 | 24.371 | Translation |
| Ama-MIR156g-12 | macrophylla_033555 | 4 | 23.333 | Translation |
| Apl-MIR408-41 | plinii_042988 | 4 | 19.626 | Translation |

| | | | | |
|---|---|---|---|---|
| Apl-MIR818d-31 | plinii_077901 | 3.5 | 16.3 | Cleavage |
| Apl-MIR818d-32 | plinii_077901 | 3.5 | 16.3 | Cleavage |
| Apl-MIR818d-33 | plinii_077901 | 3.5 | 16.3 | Cleavage |
| Ami-MIR818a-16 | micrantha_025393 | 3.5 | 24.188 | Cleavage |
| Adof-MIR166d-17 | donaciforms_045772 | 2 | 18.904 | Cleavage |
| Ado-MIR167d-17 | donax_55861 | 4 | 21.21 | Translation |
| Ado-MIR167g-16 | donax_55861 | 4 | 21.21 | Translation |
| Ado-MIR167h-7 | donax_55861 | 4 | 21.21 | Translation |
| Ama-MIR159a-25 | macrophylla_001307 | 2.5 | 17.12 | Cleavage |
| Ama-MIR159b-10 | macrophylla_001307 | 2.5 | 17.12 | Cleavage |
| Ama-MIR159b-11 | macrophylla_001307 | 2.5 | 17.12 | Cleavage |
| Apl-MIR159a-24 | plinii_135615 | 2.5 | 19.97 | Cleavage |
| Apl-MIR159a-25 | plinii_135615 | 2.5 | 19.97 | Cleavage |
| Apl-MIR159b-13 | plinii_135615 | 2.5 | 19.97 | Cleavage |
| Apl-MIR396e-3a | plinii_135615 | 4 | 14.969 | Translation |
| Apl-MIR396e-3b | plinii_135615 | 4 | 15.116 | Cleavage |
| Apl-MIR399c-16a | plinii_135615 | 2.5 | 21.519 | Cleavage |
| Apl-MIR399c-17 | plinii_135615 | 2.5 | 21.519 | Cleavage |
| Apl-MIR399f-16b | plinii_135615 | 3.5 | 21.519 | Cleavage |
| Apl-MIR399j-10 | plinii_135615 | 3 | 21.519 | Cleavage |
| Adof-MIR528-18 | donaciforms_098815 | 3.5 | 16.044 | Cleavage |
| Ama-MIR1435-28 | macrophylla_102106 | 4 | 13.81 | Translation |
| Adof-MIR1862e-3 | donaciforms_065611 | 3.5 | 10.477 | Cleavage |
| Adof-MIR437-22b | donaciforms_065611 | 4 | 10.118 | Cleavage |
| Adof-MIR5161-24 | donaciforms_065611 | 4 | 9.696 | Cleavage |
| Ado-MIR444d-11d | donax_26921 | 3.5 | 22.181 | Cleavage |
| Ado-MIR444f-12d | donax_26921 | 3.5 | 22.181 | Cleavage |
| Ado-MIR169n-9 | donax_32252 | 3.5 | 15.644 | Translation |
| Ado-MIR169q-13 | donax_32252 | 3 | 15.644 | Cleavage |
| Ado-MIR169q-4a | donax_32252 | 3 | 15.644 | Cleavage |
| Ado-MIR169q-5a | donax_32252 | 3 | 15.644 | Cleavage |
| Ado-MIR169q-6a | donax_32252 | 3 | 15.644 | Cleavage |
| Aco-MIR169g-8 | collina_047723 | 3 | 15.328 | Cleavage |
| Aco-MIR169q-5a | collina_047723 | 3 | 15.328 | Cleavage |
| Adof-MIR169g-11 | donaciforms_045618 | 3 | 15.874 | Cleavage |
| Adof-MIR169q-14 | donaciforms_045618 | 3 | 15.874 | Cleavage |
| Adof-MIR169q-4a | donaciforms_045618 | 3 | 15.874 | Cleavage |
| Aco-MIR169g-8 | collina_054425 | 3.5 | 18.734 | Cleavage |
| Aco-MIR169q-5a | collina_054425 | 3.5 | 18.734 | Cleavage |
| Afo-MIR169g-4 | formosana_31186 | 3.5 | 17.148 | Cleavage |
| Apl-MIR169g-11 | plinii_077500 | 3.5 | 18.935 | Cleavage |
| Apl-MIR169q-2 | plinii_077500 | 3.5 | 18.935 | Cleavage |
| Afo-MIR528-3 | formosana_50402 | 4 | 11.279 | Translation |
| Adof-MIR159a-10 | donaciforms_100283 | 3 | 16.641 | Cleavage |
| Adof-MIR159a-9 | donaciforms_100283 | 3 | 16.641 | Cleavage |
| Apl-MIR159a-24 | plinii_066011 | 3 | 16.641 | Cleavage |
| Apl-MIR159a-25 | plinii_066011 | 3 | 16.641 | Cleavage |
| Apl-MIR159b-13 | plinii_066011 | 3 | 16.641 | Cleavage |
| Aco-MIR169g-8 | collina_054429 | 3 | 22.082 | Cleavage |
| Aco-MIR169q-5a | collina_054429 | 3 | 22.082 | Cleavage |
| Apl-MIR169g-11 | plinii_016654 | 3 | 22.082 | Cleavage |

| | | | | |
|---|---|---|---|---|
| Apl-MIR169q-2 | plinii_016654 | 3 | 22.082 | Cleavage |
| Aco-MIR528-22 | collina_119641 | 4 | 12.243 | Translation |
| Aco-MIR528-23 | collina_119641 | 4 | 12.243 | Translation |
| Aco-MIR159d-32 | collina_108923 | 3.5 | 13.726 | Cleavage |
| Ado-MIR444d-11b | donax_56841 | 4 | 18.827 | Cleavage |
| Ado-MIR444f-12b | donax_56841 | 4 | 18.827 | Cleavage |
| Adof-MIR444a-8b | donaciforms_088528 | 4 | 23.719 | Cleavage |
| Ado-MIR444a-2b | donax_39600 | 4 | 23.719 | Cleavage |
| Ama-MIR444a-4b | macrophylla_076886 | 4 | 23.719 | Cleavage |
| Apl-MIR444a-19b | plinii_091954 | 4 | 23.719 | Cleavage |
| Aco-MIR159d-32 | collina_016313 | 4 | 21.203 | Cleavage |
| Adof-MIR528-18 | donaciforms_066613 | 3.5 | 17.656 | Translation |
| Aco-MIR156j-12 | collina_066272 | 4 | 19.597 | Cleavage |
| Adof-MIR5161-24 | donaciforms_083222 | 4 | 10.926 | Cleavage |
| Aco-MIR5161-1 | collina_074428 | 2.5 | 18.753 | Cleavage |
| Aco-MIR5161-24a | collina_074428 | 3 | 18.753 | Cleavage |
| Aco-MIR818f-27 | collina_074428 | 3.5 | 18.753 | Cleavage |
| Aco-MIR818f-28a | collina_074428 | 4 | 18.753 | Cleavage |
| Ama-MIR437-22b | macrophylla_053589 | 4 | 13.11 | Cleavage |
| Apl-MIR172d-21b | plinii_097413 | 3 | 17.19 | Cleavage |
| Apl-MIR172d-22b | plinii_097413 | 3 | 17.19 | Cleavage |
| Ado-MIR444d-11c | donax_56947 | 4 | 22.876 | Cleavage |
| Ado-MIR444f-12c | donax_56947 | 4 | 22.876 | Cleavage |
| Ado-MIR5161-21 | donax_56947 | 1 | 19.864 | Cleavage |
| Apl-MIR1439-38 | plinii_109943 | 4 | 9.716 | Cleavage |
| Apl-MIR1439-39 | plinii_109943 | 3.5 | 9.716 | Cleavage |
| Apl-MIR5161-27 | plinii_109943 | 3.5 | 9.716 | Cleavage |
| Apl-MIR5161-40 | plinii_109943 | 3.5 | 9.716 | Cleavage |
| Apl-MIR812a-1 | plinii_109943 | 3 | 9.716 | Cleavage |
| Apl-MIR812q-35 | plinii_109943 | 4 | 9.716 | Cleavage |
| Apl-MIR818f-36 | plinii_109943 | 3.5 | 9.716 | Cleavage |
| Ama-MIR1435-28 | macrophylla_098552 | 4 | 21.639 | Cleavage |
| Adof-MIR1862e-3 | donaciforms_060837 | 3 | 10.789 | Cleavage |
| Ama-MIR818e-23 | macrophylla_014006 | 4 | 19.843 | Translation |
| Ami-MIR818f-15 | micrantha_027349 | 4 | 13.593 | Translation |
| Apl-MIR2275d-5 | plinii_034356 | 4 | 19.24 | Cleavage |
| Ado-MIR393a-25a | donax_33641 | 4 | 20.161 | Cleavage |
| Adof-MIR528-18 | donaciforms_005316 | 4 | 15.74 | Translation |
| Aco-MIR818f-27 | collina_120043 | 4 | 15.92 | Cleavage |
| Ado-MIR444a-2a | donax_55564 | 4 | 20.611 | Translation |
| Ami-MIR156j-12 | micrantha_083676 | 2.5 | 20.491 | Cleavage |
| Ado-MIR169n-9 | donax_17482 | 1 | 19.108 | Cleavage |
| Ado-MIR169q-13 | donax_17482 | 2.5 | 19.108 | Cleavage |
| Ado-MIR169q-4a | donax_17482 | 2.5 | 19.108 | Cleavage |
| Ado-MIR169q-5a | donax_17482 | 2.5 | 19.108 | Cleavage |
| Ado-MIR169q-6a | donax_17482 | 2.5 | 19.108 | Cleavage |
| Ami-MIR444c-5b | micrantha_084995 | 3.5 | 17.177 | Translation |
| Ami-MIR5161-19b | micrantha_084995 | 3.5 | 16.301 | Cleavage |
| Apl-MIR172d-21a | plinii_129845 | 0.5 | 10.79 | Cleavage |
| Apl-MIR172d-21b | plinii_129845 | 2 | 10.79 | Cleavage |
| Apl-MIR172d-22a | plinii_129845 | 0.5 | 10.79 | Cleavage |

| | | | | |
|---|---|---|---|---|
| Apl-MIR172d-22b | plinii_129845 | 2 | 10.79 | Cleavage |
| Apl-MIR172d-23 | plinii_129845 | 1.5 | 10.79 | Cleavage |
| Ama-MIR444c-26a | macrophylla_044931 | 2.5 | 18.986 | Translation |
| Apl-MIR159a-24 | plinii_118381 | 3.5 | 21.764 | Translation |
| Apl-MIR159a-25 | plinii_118381 | 3.5 | 21.764 | Translation |
| Apl-MIR159b-13 | plinii_118381 | 3.5 | 21.764 | Translation |
| Ado-MIR169n-9 | donax_09498 | 2.5 | 10.685 | Translation |
| Ado-MIR169q-13 | donax_09498 | 2 | 10.685 | Cleavage |
| Ado-MIR169q-4a | donax_09498 | 2 | 10.685 | Cleavage |
| Ado-MIR169q-5a | donax_09498 | 2 | 10.685 | Cleavage |
| Ado-MIR169q-6a | donax_09498 | 2 | 10.685 | Cleavage |
| Adof-MIR169g-11 | donaciforms_041266 | 3.5 | 18.223 | Cleavage |
| Adof-MIR169q-14 | donaciforms_041266 | 3.5 | 18.223 | Cleavage |
| Adof-MIR169q-4a | donaciforms_041266 | 3.5 | 18.223 | Cleavage |
| Apl-MIR1430-7 | plinii_042961 | 3.5 | 15.148 | Translation |
| Apl-MIR169g-11 | plinii_042961 | 2 | 15.148 | Cleavage |
| Apl-MIR169q-2 | plinii_042961 | 2 | 15.148 | Cleavage |
| Ami-MIR169g-2 | micrantha_051143 | 2.5 | 17.998 | Cleavage |
| Afo-MIR444c-13a | formosana_53918 | 4 | 16.759 | Cleavage |
| Aco-MIR5161-1 | collina_034390 | 2 | 12.774 | Cleavage |
| Aco-MIR5161-24a | collina_034390 | 4 | 12.774 | Cleavage |
| Aco-MIR5161-24c | collina_034390 | 3.5 | 6.709 | Cleavage |
| Aco-MIR818f-27 | collina_034390 | 3 | 12.774 | Cleavage |
| Adof-MIR159a-10 | donaciforms_035668 | 2.5 | 14.379 | Cleavage |
| Adof-MIR159a-9 | donaciforms_035668 | 2.5 | 14.379 | Cleavage |
| Apl-MIR159a-24 | plinii_085354 | 2.5 | 14.379 | Cleavage |
| Apl-MIR159a-25 | plinii_085354 | 2.5 | 14.379 | Cleavage |
| Apl-MIR159b-13 | plinii_085354 | 2.5 | 14.379 | Cleavage |
| Aco-MIR159a-19 | collina_036925 | 2.5 | 14.503 | Cleavage |
| Aco-MIR159a-20 | collina_036925 | 2.5 | 14.503 | Cleavage |
| Aco-MIR159b-18 | collina_036925 | 2.5 | 14.503 | Cleavage |
| Aco-MIR159d-32 | collina_036925 | 2.5 | 14.503 | Cleavage |
| Ado-MIR159a-1 | donax_15653 | 2.5 | 16.731 | Cleavage |
| Ado-MIR159b-14 | donax_15653 | 2.5 | 16.731 | Cleavage |
| Afo-MIR159b-7 | formosana_16666 | 2.5 | 16.701 | Cleavage |
| Afo-MIR159b-8 | formosana_16666 | 2.5 | 16.701 | Cleavage |
| Ama-MIR159a-25 | macrophylla_031108 | 2.5 | 17.353 | Cleavage |
| Ama-MIR159b-10 | macrophylla_031108 | 2.5 | 17.353 | Cleavage |
| Ama-MIR159b-11 | macrophylla_031108 | 2.5 | 17.353 | Cleavage |
| Apl-MIR444a-19a | plinii_107897 | 4 | 17.765 | Cleavage |
| Apl-MIR444c-30c | plinii_107897 | 3.5 | 17.765 | Cleavage |
| Adof-MIR444a-7a | donaciforms_096132 | 4 | 16.22 | Cleavage |
| Adof-MIR444c-12c | donaciforms_096132 | 3.5 | 16.22 | Cleavage |
| Ama-MIR5831-30 | macrophylla_001889 | 3.5 | 23.843 | Cleavage |
| Apl-MIR444a-19a | plinii_107896 | 4 | 16.22 | Cleavage |
| Apl-MIR444c-30c | plinii_107896 | 3.5 | 16.22 | Cleavage |
| Adof-MIR166d-17 | donaciforms_085342 | 2 | 21.563 | Cleavage |
| Ado-MIR166d-8 | donax_57694 | 2 | 20.789 | Cleavage |
| Aco-MIR166d-17 | collina_119347 | 2 | 22.871 | Cleavage |
| Adof-MIR166d-17 | donaciforms_111958 | 2 | 24.308 | Cleavage |
| Ami-MIR166d-7 | micrantha_084466 | 2 | 24.303 | Cleavage |

| | | | | |
|---|---|---|---|---|
| Apl-MIR166d-18 | plinii_000545 | 2 | 24.308 | Cleavage |
| Ado-MIR393a-25a | donax_40321 | 2 | 21.446 | Cleavage |
| Ado-MIR393b-25b | donax_40321 | 2.5 | 21.446 | Cleavage |
| Ado-MIR444c-20b | donax_40321 | 4 | 23.114 | Translation |
| Ama-MIR393b-5 | macrophylla_001724 | 2 | 19.434 | Cleavage |
| Ama-MIR444a-4a | macrophylla_001724 | 4 | 20.433 | Cleavage |
| Ama-MIR444c-26b | macrophylla_001724 | 4 | 21.622 | Translation |
| Apl-MIR818f-37b | plinii_084686 | 2 | 11.429 | Cleavage |
| Adof-MIR156j-19 | donaciforms_022176 | 2 | 11.144 | Cleavage |
| Adof-MIR156j-20 | donaciforms_022176 | 2 | 11.144 | Cleavage |
| Adof-MIR162b-6 | donaciforms_111912 | 2.5 | 24.192 | Cleavage |
| Ado-MIR162a-3 | donax_56921 | 2.5 | 24.2 | Cleavage |
| Ama-MIR162a-7 | macrophylla_090789 | 2.5 | 24.2 | Cleavage |
| Ami-MIR162b-13 | micrantha_053616 | 2.5 | 24.192 | Cleavage |
| Apl-MIR162b-12 | plinii_135169 | 2.5 | 24.2 | Cleavage |
| Adof-MIR5161-24 | donaciforms_090100 | 4 | 16.854 | Cleavage |
| Aco-MIR156d-16 | collina_011227 | 3 | 10.739 | Cleavage |
| Aco-MIR156j-12 | collina_011227 | 2 | 10.739 | Cleavage |
| Afo-MIR156g-10 | formosana_06194 | 2 | 12.714 | Cleavage |
| Afo-MIR156k-9 | formosana_06194 | 1 | 12.714 | Cleavage |
| Ama-MIR160b-14 | macrophylla_073113 | 3.5 | 15.695 | Translation |
| Apl-MIR444c-30b | plinii_131158 | 2.5 | 12.26 | Translation |
| Ami-MIR166d-7 | micrantha_014460 | 2 | 18.92 | Cleavage |
| Aco-MIR5161-24a | collina_105252 | 4 | 23.013 | Cleavage |
| Aco-MIR1432-7 | collina_063086 | 2.5 | 18.45 | Cleavage |
| Ado-MIR393a-25a | donax_56827 | 1 | 20.419 | Cleavage |
| Ado-MIR393b-25b | donax_56827 | 1.5 | 20.419 | Cleavage |
| Adof-MIR444c-12a | donaciforms_109747 | 4 | 17.174 | Cleavage |
| Apl-MIR1439-39 | plinii_127784 | 4 | 9.384 | Cleavage |
| Apl-MIR5161-27 | plinii_127784 | 4 | 9.384 | Cleavage |
| Apl-MIR5161-40 | plinii_127784 | 4 | 9.384 | Cleavage |
| Apl-MIR5161-6a | plinii_127784 | 4 | 9.384 | Cleavage |
| Apl-MIR5161-6b | plinii_127784 | 3 | 8.939 | Translation |
| Apl-MIR812a-1 | plinii_127784 | 1.5 | 9.384 | Cleavage |
| Apl-MIR812q-35 | plinii_127784 | 2.5 | 9.384 | Cleavage |
| Apl-MIR818d-28 | plinii_127784 | 4 | 9.384 | Cleavage |
| Apl-MIR818f-36 | plinii_127784 | 3.5 | 9.384 | Cleavage |
| Ama-MIR437-22b | macrophylla_060909 | 4 | 12.17 | Cleavage |
| Aco-MIR156d-16 | collina_099015 | 4 | 7.342 | Cleavage |
| Adof-MIR168a-13 | donaciforms_111914 | 3.5 | 21.892 | Cleavage |
| Afo-MIR168a-6 | formosana_00679 | 3.5 | 18.534 | Cleavage |
| Ama-MIR160b-14 | macrophylla_067513 | 4 | 24.137 | Translation |
| Apl-MIR160b-20 | plinii_094868 | 4 | 23.613 | Translation |
| Apl-MIR399c-16a | plinii_092757 | 4 | 18.795 | Cleavage |
| Apl-MIR399c-17 | plinii_092757 | 4 | 18.795 | Cleavage |
| Apl-MIR399f-16b | plinii_092757 | 2.5 | 18.795 | Cleavage |
| Apl-MIR399j-10 | plinii_092757 | 4 | 18.795 | Cleavage |
| Ama-MIR172d-8a | macrophylla_082550 | 4 | 14.401 | Translation |
| Apl-MIR399c-16a | plinii_092755 | 4 | 18.803 | Cleavage |
| Apl-MIR399c-17 | plinii_092755 | 4 | 18.803 | Cleavage |
| Apl-MIR399f-16b | plinii_092755 | 2.5 | 18.803 | Cleavage |

| | | | | |
|---|---|---|---|---|
| Apl-MIR399j-10 | plinii_092755 | 4 | 18.803 | Cleavage |
| Ado-MIR444c-20b | donax_18394 | 4 | 18.26 | Cleavage |
| Ama-MIR818f-24 | macrophylla_100830 | 2 | 13.697 | Cleavage |
| Adof-MIR528-18 | donaciforms_054071 | 4 | 5.718 | Cleavage |
| Ama-MIR528-15 | macrophylla_023141 | 4 | 5.776 | Cleavage |
| Aco-MIR164a-34 | collina_107352 | 4 | 18.323 | Cleavage |
| Apl-MIR1439-38 | plinii_120607 | 4 | 23.904 | Cleavage |
| Apl-MIR1439-39 | plinii_120607 | 3.5 | 23.904 | Cleavage |
| Apl-MIR5161-27 | plinii_120607 | 3.5 | 23.904 | Cleavage |
| Apl-MIR5161-40 | plinii_120607 | 3.5 | 23.904 | Cleavage |
| Apl-MIR5161-6b | plinii_120607 | 3.5 | 22.431 | Translation |
| Apl-MIR812a-1 | plinii_120607 | 2 | 23.904 | Cleavage |
| Apl-MIR812q-35 | plinii_120607 | 3 | 23.904 | Cleavage |
| Apl-MIR818d-28 | plinii_120607 | 3.5 | 23.904 | Cleavage |
| Apl-MIR818f-36 | plinii_120607 | 3 | 23.904 | Cleavage |
| Adof-MIR156j-19 | donaciforms_081779 | 2 | 21.975 | Cleavage |
| Adof-MIR156j-20 | donaciforms_081779 | 2 | 21.975 | Cleavage |
| Aco-MIR156d-16 | collina_101707 | 2 | 18.027 | Cleavage |
| Aco-MIR156d-16 | collina_101728 | 2 | 16.128 | Cleavage |
| Aco-MIR156j-12 | collina_101707 | 1 | 18.027 | Cleavage |
| Aco-MIR156j-12 | collina_101728 | 1 | 16.128 | Cleavage |
| Aco-MIR444d-14c | collina_101707 | 4 | 17.903 | Cleavage |
| Aco-MIR444d-14c | collina_101728 | 4 | 17.903 | Cleavage |
| Ama-MIR156g-12 | macrophylla_084146 | 1 | 16.278 | Cleavage |
| Ama-MIR156j-1 | macrophylla_084146 | 1 | 16.278 | Cleavage |
| Ama-MIR156j-13 | macrophylla_084146 | 1 | 16.278 | Cleavage |
| Ami-MIR156j-12 | micrantha_078289 | 1 | 16.278 | Cleavage |
| Ama-MIR172d-8a | macrophylla_083546 | 0.5 | 15.541 | Cleavage |
| Ama-MIR172d-8a | macrophylla_083555 | 0.5 | 15.541 | Cleavage |
| Ama-MIR172d-8b | macrophylla_083546 | 2 | 15.541 | Cleavage |
| Ama-MIR172d-8b | macrophylla_083555 | 2 | 15.541 | Cleavage |
| Ama-MIR172d-8a | macrophylla_056355 | 2.5 | 14.861 | Cleavage |
| Ama-MIR172d-8a | macrophylla_056357 | 2.5 | 14.861 | Cleavage |
| Ama-MIR172d-8b | macrophylla_056355 | 3 | 14.861 | Cleavage |
| Ama-MIR172d-8b | macrophylla_056357 | 3 | 14.861 | Cleavage |
| Ami-MIR172d-10a | micrantha_069690 | 2.5 | 15.232 | Cleavage |
| Ami-MIR172d-10b | micrantha_069690 | 3 | 15.232 | Cleavage |
| Ado-MIR172d-22a | donax_27281 | 0.5 | 23.575 | Cleavage |
| Ado-MIR172d-22b | donax_27281 | 2 | 23.575 | Cleavage |
| Aco-MIR172d-6a | collina_085567 | 2.5 | 14.732 | Cleavage |
| Aco-MIR172d-6a | collina_085568 | 2.5 | 15.387 | Cleavage |
| Aco-MIR172d-6b | collina_085567 | 3 | 14.732 | Cleavage |
| Aco-MIR172d-6b | collina_085568 | 3 | 15.387 | Cleavage |
| Apl-MIR172d-21a | plinii_129846 | 0.5 | 10.79 | Cleavage |
| Apl-MIR172d-21b | plinii_129846 | 2 | 10.79 | Cleavage |
| Apl-MIR172d-22a | plinii_129846 | 0.5 | 10.79 | Cleavage |
| Apl-MIR172d-22b | plinii_129846 | 2 | 10.79 | Cleavage |
| Apl-MIR172d-23 | plinii_129846 | 1.5 | 10.79 | Cleavage |
| Ama-MIR156j-1 | macrophylla_097714 | 4 | 13.569 | Cleavage |
| Ama-MIR156j-13 | macrophylla_097714 | 4 | 13.569 | Cleavage |
| Ama-MIR5143b-2 | macrophylla_095635 | 3.5 | 17.162 | Cleavage |

| | | | | |
|---|---|---|---|---|
| Apl-MIR160b-20 | plinii_121545 | 0 | 23.068 | Cleavage |
| Apl-MIR160b-20 | plinii_121546 | 0 | 18.169 | Cleavage |
| Ado-MIR169n-9 | donax_43178 | 1.5 | 18.982 | Cleavage |
| Ado-MIR169q-13 | donax_43178 | 2.5 | 18.982 | Cleavage |
| Ado-MIR169q-4a | donax_43178 | 2.5 | 18.982 | Cleavage |
| Ado-MIR169q-5a | donax_43178 | 2.5 | 18.982 | Cleavage |
| Ado-MIR169q-6a | donax_43178 | 2.5 | 18.982 | Cleavage |
| Ami-MIR169g-2 | micrantha_075304 | 2.5 | 18.503 | Cleavage |
| Ama-MIR167e-9 | macrophylla_093386 | 3.5 | 19.481 | Cleavage |
| Ama-MIR167h-16 | macrophylla_093386 | 3.5 | 19.481 | Cleavage |
| Ama-MIR167h-17 | macrophylla_093386 | 3.5 | 19.481 | Cleavage |
| Ama-MIR167h-18 | macrophylla_093386 | 3.5 | 19.481 | Cleavage |
| Ado-MIR5161-21 | donax_16151 | 2 | 11.924 | Cleavage |
| Ado-MIR444d-11b | donax_32918 | 3.5 | 18.093 | Cleavage |
| Ado-MIR444f-12b | donax_32918 | 3.5 | 18.093 | Cleavage |
| Ama-MIR399i-3 | macrophylla_100914 | 3 | 11.827 | Translation |
| Ado-MIR1879-24 | donax_50068 | 4 | 14.988 | Cleavage |
| Ado-MIR5161-21 | donax_50068 | 4 | 21.19 | Cleavage |
| Apl-MIR396e-3a | plinii_120854 | 4 | 19.705 | Cleavage |
| Ado-MIR444a-2a | donax_17537 | 3.5 | 22.545 | Cleavage |
| Ama-MIR444a-4a | macrophylla_045741 | 3.5 | 22.545 | Cleavage |
| Ami-MIR818a-16 | micrantha_078361 | 4 | 12.66 | Cleavage |
| Ado-MIR5161-21 | donax_51764 | 3.5 | 12.144 | Cleavage |
| Afo-MIR393b-2 | formosana_00797 | 1 | 19.177 | Cleavage |
| Ado-MIR172d-22a | donax_49796 | 4 | 21.895 | Cleavage |
| Ama-MIR172d-8a | macrophylla_075085 | 4 | 20.36 | Cleavage |
| Apl-MIR172d-21a | plinii_117502 | 4 | 22.496 | Cleavage |
| Apl-MIR172d-22a | plinii_117502 | 4 | 22.496 | Cleavage |
| Aco-MIR444d-13a | collina_030711 | 3.5 | 14.375 | Translation |
| Aco-MIR444d-14a | collina_030711 | 3.5 | 14.375 | Translation |
| Ado-MIR437-18 | donax_41579 | 3 | 15.3 | Cleavage |
| Ado-MIR437-19 | donax_41579 | 3 | 15.3 | Cleavage |
| Ama-MIR528-15 | macrophylla_091673 | 4 | 16.61 | Cleavage |

**Supplementary Table 2.3.** Annotation of putative targets.

| miRNA | Putative target | Reference Id | Description |
|---|---|---|---|
| Aco-MIR1432-7 | collina_065806 | AT5G04460.3 | RING/U box superfamily protein |
| Aco-MIR156d-16 | collina_120304 | AT1G15290.1 | Tetratricopeptide repeat (TPR) like superfamily protein |
| Aco-MIR156d-16 | collina_099015 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Aco-MIR159a-19 | collina_036925 | AT3G11440.1 | myb domain protein 65 |
| Aco-MIR159a-20 | collina_036925 | AT3G11440.1 | myb domain protein 65 |
| Aco-MIR159b-18 | collina_036925 | AT3G11440.1 | myb domain protein 65 |
| Aco-MIR159d-32 | collina_108923 | AT4G14350.3 | AGC (cAMP dependent, cGMP |

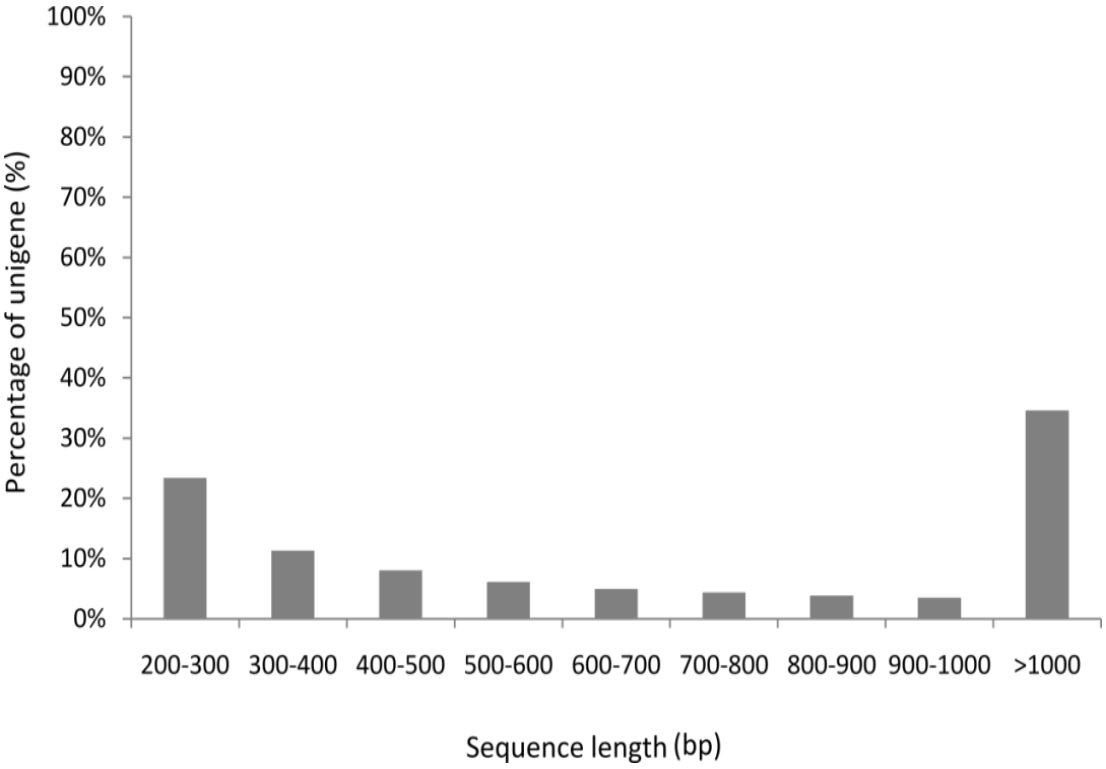
| | | | |
|---|---|---|---|
| | | | dependent and protein kinase C) kinase family protein |
| Aco-MIR159d-32 | collina_036925 | AT3G11440.1 | myb domain protein 65 |
| Aco-MIR159d-32 | collina_016313 | AT1G21250.1 | cell wall associated kinase |
| Aco-MIR164a-34 | collina_107352 | AT2G47070.1 | squamosa promoter binding protein like 1 |
| Aco-MIR166d-17 | collina_119347 | AT5G60690.1 | Homeobox leucine zipper family protein / lipid binding START domain containing protein |
| Aco-MIR444d-13a | collina_030711 | AT1G74600.1 | pentatricopeptide (PPR) repeat containing protein |
| Aco-MIR444d-14a | collina_030711 | AT1G74600.1 | pentatricopeptide (PPR) repeat containing protein |
| Aco-MIR5161-1 | collina_101612 | AT5G45560.1 | Pleckstrin homology (PH) domain containing protein / lipid binding START domain containing protein |
| Aco-MIR5161-24a | collina_101612 | AT5G45560.1 | Pleckstrin homology (PH) domain containing protein / lipid binding START domain containing protein |
| Aco-MIR5161-24b | collina_101612 | AT5G45560.1 | Pleckstrin homology (PH) domain containing protein / lipid binding START domain containing protein |
| Aco-MIR5161-33 | collina_101612 | AT5G45560.1 | Pleckstrin homology (PH) domain containing protein / lipid binding START domain containing protein |
| Aco-MIR818f-27 | collina_101612 | AT5G45560.1 | Pleckstrin homology (PH) domain containing protein / lipid binding START domain containing protein |
| Aco-MIR818f-27 | collina_120043 | AT2G39800.4 | delta1 pyrroline 5 carboxylate synthase 1 |
| Adof-MIR159a-10 | donaciforms_035668 | AT3G11440.1 | myb domain protein 65 |
| Adof-MIR159a-9 | donaciforms_035668 | AT3G11440.1 | myb domain protein 65 |
| Adof-MIR162b-6 | donaciforms_111912 | AT1G01040.1 | dicer like 1 |
| Adof-MIR166d-17 | donaciforms_111958 | AT5G60690.1 | Homeobox leucine zipper family protein / lipid binding START domain containing protein |
| Adof-MIR166d-17 | donaciforms_085342 | AT2G34710.1 | Homeobox leucine zipper family protein / lipid binding START domain containing protein |
| Adof-MIR168a-13 | donaciforms_111914 | AT1G48410.1 | Stabilizer of iron transporter SufD / Polynucleotidyl transferase |
| Adof-MIR1862e-3 | donaciforms_060837 | AT5G07630.1 | lipid transporters |
| Adof-MIR444c-12a | donaciforms_109747 | AT1G12820.1 | auxin signaling F box 3 |
| Adof-MIR5161-24 | donaciforms_083222 | AT3G19280.1 | fucosyltransferase 11 |
| Adof-MIR528-18 | donaciforms_054071 | AT2G32230.1 | proteinaceous RNase P 1 |
| Ado-MIR162a-3 | donax_56921 | AT1G01040.1 | dicer like 1 |
| Ado-MIR166d-8 | donax_57694 | AT2G34710.1 | Homeobox leucine zipper family protein / lipid binding START domain containing protein |
| Ado-MIR167d-17 | donax_55861 | AT4G01020.1 | helicase domain containing protein / IBR domain containing protein / zinc finger protein related |
| Ado-MIR167g-16 | donax_55861 | AT4G01020.1 | helicase domain containing protein / IBR domain containing protein / zinc |

| | | | |
|---|---|---|---|
| | | | finger protein related |
| Ado-MIR167h-7 | donax_55861 | AT4G01020.1 | helicase domain containing protein / IBR domain containing protein / zinc finger protein related |
| Ado-MIR172d-22a | donax_49796 | AT1G12820.1 | auxin signaling F box 3 |
| Ado-MIR172d-22a | donax_27281 | AT2G28550.3 | related to AP2.7 |
| Ado-MIR172d-22b | donax_27281 | AT2G28550.3 | related to AP2.7 |
| Ado-MIR393a-25a | donax_40321 | AT3G62980.1 | F box/RNI like superfamily protein |
| Ado-MIR393b-25b | donax_40321 | AT3G62980.1 | F box/RNI like superfamily protein |
| Ado-MIR444a-2a | donax_55564 | AT1G68830.1 | STT7 homolog STN7 |
| Ado-MIR444a-2b | donax_39600 | AT3G19490.1 | sodium:hydrogen antiporter 1 |
| Ado-MIR444c-20b | donax_18394 | AT1G55620.2 | chloride channel F |
| Ado-MIR444c-20b | donax_40321 | AT3G62980.1 | F box/RNI like superfamily protein |
| Ado-MIR444d-11b | donax_56841 | AT4G14350.2 | AGC (cAMP dependent, cGMP dependent and protein kinase C) kinase family protein |
| Ado-MIR444d-11b | donax_32918 | AT5G43900.1 | myosin 2 |
| Ado-MIR444d-11c | donax_56947 | AT5G03070.1 | importin alpha isoform 9 |
| Ado-MIR444f-12b | donax_56841 | AT4G14350.2 | AGC (cAMP dependent, cGMP dependent and protein kinase C) kinase family protein |
| Ado-MIR444f-12b | donax_32918 | AT5G43900.1 | myosin 2 |
| Ado-MIR444f-12c | donax_56947 | AT5G03070.1 | importin alpha isoform 9 |
| Ado-MIR5161-21 | donax_56947 | AT5G03070.1 | importin alpha isoform 9 |
| Afo-MIR156g-10 | formosana_57367 | AT1G15290.1 | Tetratricopeptide repeat (TPR) like superfamily protein |
| Afo-MIR168a-6 | formosana_00679 | AT1G48410.1 | Stabilizer of iron transporter SufD / Polynucleotidyl transferase |
| Afo-MIR528-3 | formosana_13913 | AT3G42170.1 | BED zinc finger ;hAT family dimerisation domain |
| Ama-MIR1435-28 | macrophylla_098552 | AT3G46960.1 | RNA helicase, ATP dependent, SK12/DOB1 protein |
| Ama-MIR156g-12 | macrophylla_101942 | AT1G15290.1 | Tetratricopeptide repeat (TPR) like superfamily protein |
| Ama-MIR159a-25 | macrophylla_031108 | AT5G06100.1 | myb domain protein 33 |
| Ama-MIR159a-25 | macrophylla_001307 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Ama-MIR159b-10 | macrophylla_031108 | AT5G06100.1 | myb domain protein 33 |
| Ama-MIR159b-10 | macrophylla_001307 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Ama-MIR159b-11 | macrophylla_031108 | AT5G06100.1 | myb domain protein 33 |
| Ama-MIR159b-11 | macrophylla_001307 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Ama-MIR160b-14 | macrophylla_073113 | AT1G27440.1 | Exostosin family protein |
| Ama-MIR160b-14 | macrophylla_067513 | AT1G71860.2 | protein tyrosine phosphatase 1 |

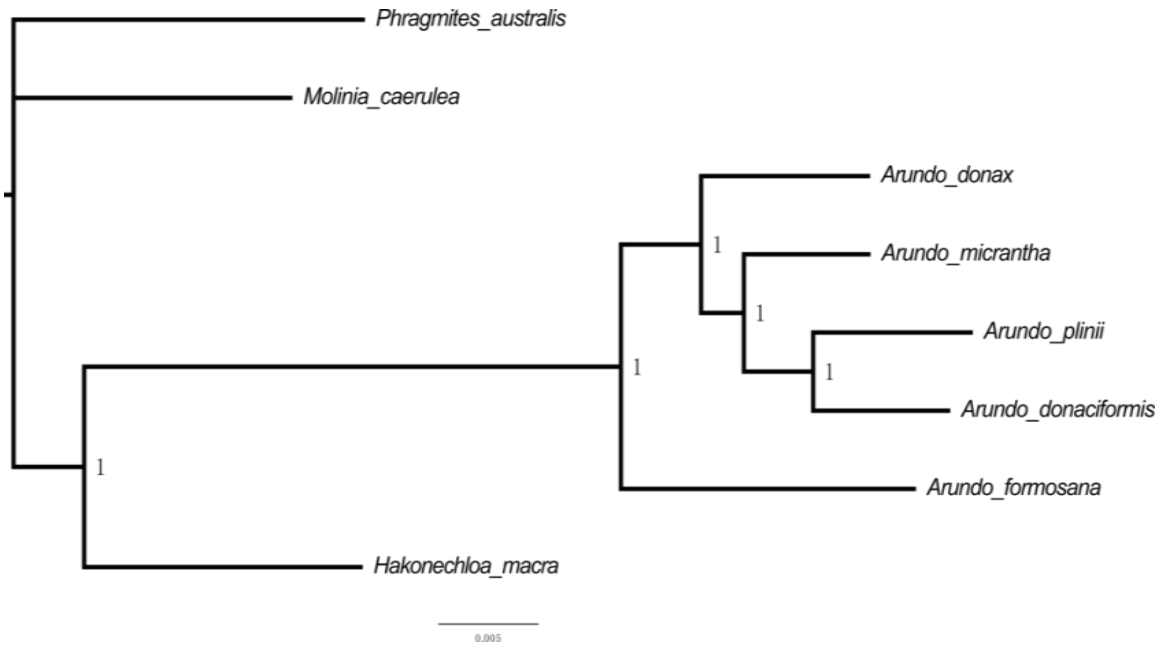
| | | | |
|---|---|---|---|
| Ama-MIR162a-7 | macrophylla_090789 | AT1G01040.1 | dicer like 1 |
| Ama-MIR172d-8a | macrophylla_082550 | AT5G64813.1 | Ras related small GTP binding family protein |
| Ama-MIR172d-8a | macrophylla_075085 | AT1G12820.1 | auxin signaling F box 3 |
| Ama-MIR393b-5 | macrophylla_001724 | AT4G03190.1 | GRR1 like protein 1 |
| Ama-MIR399i-3 | macrophylla_100914 | AT2G33770.1 | phosphate 2 |
| Ama-MIR444a-4a | macrophylla_001724 | AT4G03190.1 | GRR1 like protein 1 |
| Ama-MIR444a-4b | macrophylla_076886 | AT3G19490.1 | sodium:hydrogen antiporter 1 |
| Ama-MIR444c-26b | macrophylla_001724 | AT4G03190.1 | GRR1 like protein 1 |
| Ama-MIR5143b-2 | macrophylla_095635 | AT4G30080.1 | auxin response factor 16 |
| Ama-MIR528-15 | macrophylla_091673 | AT2G17510.1 | ribonuclease II family protein |
| Ama-MIR5831-30 | macrophylla_001889 | AT4G26000.1 | RNA binding KH domain containing protein |
| Ama-MIR818e-23 | macrophylla_014006 | AT3G14470.1 | NB ARC domain containing disease resistance protein |
| Ami-MIR156j-12 | micrantha_083676 | AT1G61030.1 | WAPL (Wings apart like protein regulation of heterochromatin) protein |
| Ami-MIR162b-13 | micrantha_053616 | AT1G01040.1 | dicer like 1 |
| Ami-MIR166d-7 | micrantha_084466 | AT5G60690.1 | Homeobox leucine zipper family protein / lipid binding START domain containing protein |
| Ami-MIR444c-5b | micrantha_084995 | AT3G01780.1 | ARM repeat superfamily protein |
| Ami-MIR5161-19b | micrantha_084995 | AT3G01780.1 | ARM repeat superfamily protein |
| Ami-MIR818a-16 | micrantha_078361 | AT5G13020.1 | Emsy N Terminus (ENT)/ plant Tudor like domains containing protein |
| Ami-MIR818a-16 | micrantha_025393 | AT3G14460.1 | LRR and NB ARC domains containing disease resistance protein |
| Ami-MIR818f-15 | micrantha_027349 | AT3G14470.1 | NB ARC domain containing disease resistance protein |
| Apl-MIR159a-24 | plinii_118381 | AT5G09810.1 | actin 7 |
| Apl-MIR159a-24 | plinii_085354 | AT3G11440.1 | myb domain protein 65 |
| Apl-MIR159a-24 | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR159a-25 | plinii_118381 | AT5G09810.1 | actin 7 |
| Apl-MIR159a-25 | plinii_085354 | AT3G11440.1 | myb domain protein 65 |
| Apl-MIR159a-25 | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR159b-13 | plinii_118381 | AT5G09810.1 | actin 7 |
| Apl-MIR159b-13 | plinii_085354 | AT3G11440.1 | myb domain protein 65 |
| Apl-MIR159b-13 | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR160b-20 | plinii_121546 | AT4G30080.1 | auxin response factor 16 |
| Apl-MIR160b-20 | plinii_121545 | AT4G30080.1 | auxin response factor 16 |

| | | | |
|---|---|---|---|
| Apl-MIR160b-20 | plinii_094868 | AT1G71860.3 | protein tyrosine phosphatase 1 |
| Apl-MIR162b-12 | plinii_135169 | AT1G01040.1 | dicer like 1 |
| Apl-MIR166d-18 | plinii_000545 | AT5G60690.1 | Homeobox leucine zipper family protein / lipid binding START domain containing protein |
| Apl-MIR172d-21a | plinii_117502 | AT3G26810.1 | auxin signaling F box 2 |
| Apl-MIR172d-21a | plinii_129845 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-21a | plinii_129846 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-21b | plinii_129845 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-21b | plinii_097413 | AT2G42620.1 | RNI like superfamily protein |
| Apl-MIR172d-21b | plinii_129846 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-22a | plinii_117502 | AT3G26810.1 | auxin signaling F box 2 |
| Apl-MIR172d-22a | plinii_129845 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-22a | plinii_129846 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-22b | plinii_129845 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-22b | plinii_097413 | AT2G42620.1 | RNI like superfamily protein |
| Apl-MIR172d-22b | plinii_129846 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-23 | plinii_129845 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR172d-23 | plinii_129846 | AT2G28550.3 | related to AP2.7 |
| Apl-MIR396e-3a | plinii_120854 | AT4G11810.1 | Major Facilitator Superfamily with SPX (SYG1/Pho81/XPR1) domain containing protein |
| Apl-MIR396e-3a | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR396e-3b | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR399c-16a | plinii_092757 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399c-16a | plinii_092755 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399c-16a | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR399c-17 | plinii_092757 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399c-17 | plinii_092755 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399c-17 | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR399f-16b | plinii_092757 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399f-16b | plinii_092755 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399f-16b | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside |

| | | | |
|---|---|---|---|
| | | | triphosphatases;transcription factor binding |
| Apl-MIR399j-10 | plinii_092757 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399j-10 | plinii_092755 | AT5G64813.1 | Ras related small GTP binding family protein |
| Apl-MIR399j-10 | plinii_135615 | AT1G67120.1 | ATPases;nucleotide binding;ATP binding;nucleoside triphosphatases;transcription factor binding |
| Apl-MIR408-41 | plinii_042988 | AT5G04500.1 | glycosyltransferase family protein 47 |
| Apl-MIR444a-19a | plinii_107896 | AT3G04610.1 | RNA binding KH domain containing protein |
| Apl-MIR444a-19b | plinii_091954 | AT3G19490.1 | sodium:hydrogen antiporter 1 |
| Apl-MIR444c-30c | plinii_107896 | AT3G04610.1 | RNA binding KH domain containing protein |




**Supplementary Figure 3.1**. Length distribution of assembled unigenes.




**Supplementary Figure 3.2**. Phylogenomic reconstruction of *Arundo* species.

**Supplementary Table 3.1.** Based on root node of ploidy level.

| Taxon name | Chromosome count | Phylogeny robustness | Simulation reliability | Ploidy inference |
|---|---|---|---|---|
| *Arundo donaciformis* | 54 | 1 | 1 | 1 |
| *Arundo donax* | 54 | 1 | 0.98 | 1 |
| *Arundo formosana* | 36 | 1 | 0.97 | 1 |
| *Arundo micrantha* | 36 | 1 | 0.97 | 1 |
| *Arundo plinii* | 36 | 1 | 0.98 | 1 |
| *Hakonechloa macra* | 24 | 1 | 0.89 | NA |
| *Molinia caerulea* | 18 | 1 | 0.99 | 0 |
| *Phragmites australis* | 24 | 1 | 1 | 1 |

Note: Simulationre reliabilty larger than 0.95 was considered reliable, number 1 represented in Ploidy inference is polyploidy, 0 is diploid, NA is undertermined.

**Supplementary Table 3.2.** Functional annotation of 28 genes under positive selection identified by Adaptive Branch-site REL method.

| OGs_ID | GENE_NAME | TAIR_ID | DESCRIPTION | E-value |
|---|---|---|---|---|
| OG0017436 | DG1 | AT5G67570.1 | Tetratricopeptide repeat (TPR)-like superfamily protein | 0 |
| OG0017508 | NA | NA | NA | NA |
| OG0017514 | At5g26280 | AT5G26280.1 | TRAF-like family protein | 0.69 |
| OG0017532 | At1g02020 | AT1G02020.1 | nitroreductase family protein | 1.00E-173 |
| OG0017575 | F13I12.50 | AT3G47000.1 | Glycosyl hydrolase family protein | 3.5 |
| OG0017620 | RH58 | AT5G19210.2 | P-loop containing nucleoside triphosphate hydrolases superfamily protein | 0 |
| OG0017657 | NRPB9A | AT3G16980.1 | RNA polymerases M/15 Kd subunit | 6.00E-50 |
| OG0017710 | CBSX1 | AT4G36910.1 | Cystathionine beta-synthase (CBS) family protein | 7.00E-91 |
| OG0017740 | LACS9 | AT1G77590.1 | long chain acyl-CoA synthetase 9 | 0 |
| OG0017807 | RH50 | AT3G06980.1 | DEA(D/H)-box RNA helicase family protein | 0 |
| OG0017835 | T31B5_160 | AT5G13340.1 | Arginine/glutamate-rich 1 protein | 3.00E-32 |
| OG0017952 | PVA42 | AT4G21450.1 | PapD-like superfamily protein | 8.00E-115 |
| OG0018017 | At1g64710 | AT1G64710.1 | GroES-like zinc-binding dehydrogenase family protein | 0 |
| OG0018069 | BIP | AT5G42020.1 | Heat shock protein 70 (Hsp 70) family protein | 0 |
| OG0018102 | APS1 | AT5G48300.1 | ADP glucose pyrophosphorylase 1 | 0 |
| OG0018170 | At2g01680 | AT2G01680.1 | Ankyrin repeat family protein | 0 |
| OG0018204 | CI51 | AT5G08530.1 | 51 kDa subunit of complex I | 0 |
| OG0018205 | At3g15140 | AT3G15140.1 | Polynucleotidyl transferase, ribonuclease H-like superfamily protein | 7.00E-22 |
| OG0018254 | GUN1 | AT2G31400.1 | genomes uncoupled 1 | 0 |
| OG0018339 | CUL4 | AT5G46210.1 | cullin4 | 0 |
| OG0018342 | TIF3A1 | AT4G11420.1 | eukaryotic translation initiation factor 3A | 0 |

| | | | | |
|---|---|---|---|---|
| OG0018353 | PSL5 | AT5G63840.1 | Glycosyl hydrolases family 31 protein | 0 |
| OG0018354 | TIF3C1 | AT3G56150.2 | eukaryotic translation initiation factor 3C | 0 |
| OG0018357 | SUD1 | AT4G34100.2 | RING/U-box superfamily protein | 0 |
| OG0018359 | PA200 | AT3G13330.1 | proteasome activating protein 200 | 0 |
| OG0018373 | GLU1 | AT5G04140.1 | glutamate synthase 1 | 0 |
| OG0018377 | At5g47690 | AT5G47690.1 | Binding protein | 0 |
| OG0018380 | TRX4 | AT1G19730.1 | Thioredoxin superfamily protein | 3.00E-22 |

**Supplementary Table 3.3.** Summary of Relax testing in positive selection genes identified by Adaptive Branch-site REL method.

| Test-branch | K-test | P-values | LR-values |
|---|---|---|---|
| OG0018357-ama | 0.75 | 0.17958736 | 1.8 |
| OG0017620-hm | 0.75 | 0.273474132 | 1.2 |
| OG0017575-pa | 0.87 | 0.89177901 | 0.02 |
| OG0018377-ap | 0.88 | 0.395510707 | 0.72 |
| OG0017575-node14 | 0.89 | 0.792479636 | 0.07 |
| OG0018357-mc | 0.91 | 0.800949701 | 0.06 |
| OG0017740-af | 1 | 0.980751586 | 0 |
| OG0017740-node9 | 1 | 0.997173906 | 0 |
| OG0018254-af | 28.28 | 4.66E-15 | 61.39 |
| OG0017657-adf | 50 | 6.51E-13 | 51.69 |
| OG0018353-ac | 5.98 | 8.29E-07 | 24.29 |
| OG0017835-af | 4.32 | 5.37147E-06 | 20.7 |
| OG0018205-mc | 15.84 | 8.60921E-05 | 15.42 |
| OG0017436-node3 | 2.44 | 0.000173285 | 14.1 |
| OG0017710-mc | 2.7 | 0.000464522 | 12.25 |
| OG0018170-pa | 6.23 | 0.001398769 | 10.21 |
| OG0017508-hm | 3.64 | 0.001973028 | 9.57 |
| OG0017807-ap | 50 | 0.00270952 | 8.99 |
| OG0017532-ama | 2.38 | 0.006875573 | 7.31 |
| OG0018069-adf | 43.22 | 0.007204733 | 7.22 |
| OG0017807-Node9 | 1.83 | 0.012355806 | 6.26 |
| OG0018354-adf | 1.69 | 0.012865503 | 6.19 |
| OG0018204-adf | 3.1 | 0.018170003 | 5.58 |
| OG0018359-adf | 2.31 | 0.042067067 | 4.13 |
| OG0018339-Node4 | 30.34 | 0.096575591 | 2.76 |
| OG0018380-pa | 1.6 | 0.120097172 | 2.42 |
| OG0018102-Node4 | 27.33 | 0.125251418 | 2.35 |
| OG0018342-Node14 | 1.22 | 0.153774937 | 2.03 |
| OG0018339-adf | 1.59 | 0.158029538 | 1.99 |
| OG0017575-hm | 34.03 | 0.16635404 | 1.92 |
| OG0018377-pa | 1.14 | 0.186222636 | 1.75 |
| OG0018380-mc | 1.47 | 0.19838976 | 1.65 |
| OG0018354-ac | 1.35 | 0.199537994 | 1.65 |
| OG0017952-ad | 23.28 | 0.245824236 | 1.35 |
| OG0018377-ama | 1.21 | 0.248349937 | 1.33 |
| OG0017710-Node14 | 2.38 | 0.302031658 | 1.07 |

| | | | |
|---|---|---|---|
| OG0017952-adf | 19.52 | 0.310020453 | 1.03 |
| OG0018017-ami | 1.54 | 0.527578951 | 0.4 |
| OG0017575-mc | 3.02 | 0.665216916 | 0.19 |
| OG0018373-ami | 1.09 | 0.728983779 | 0.12 |
| OG0018339-ac | 1.04 | 0.764432878 | 0.09 |
| OG0017514-node14 | 3.92 | 0.984494287 | 0 |

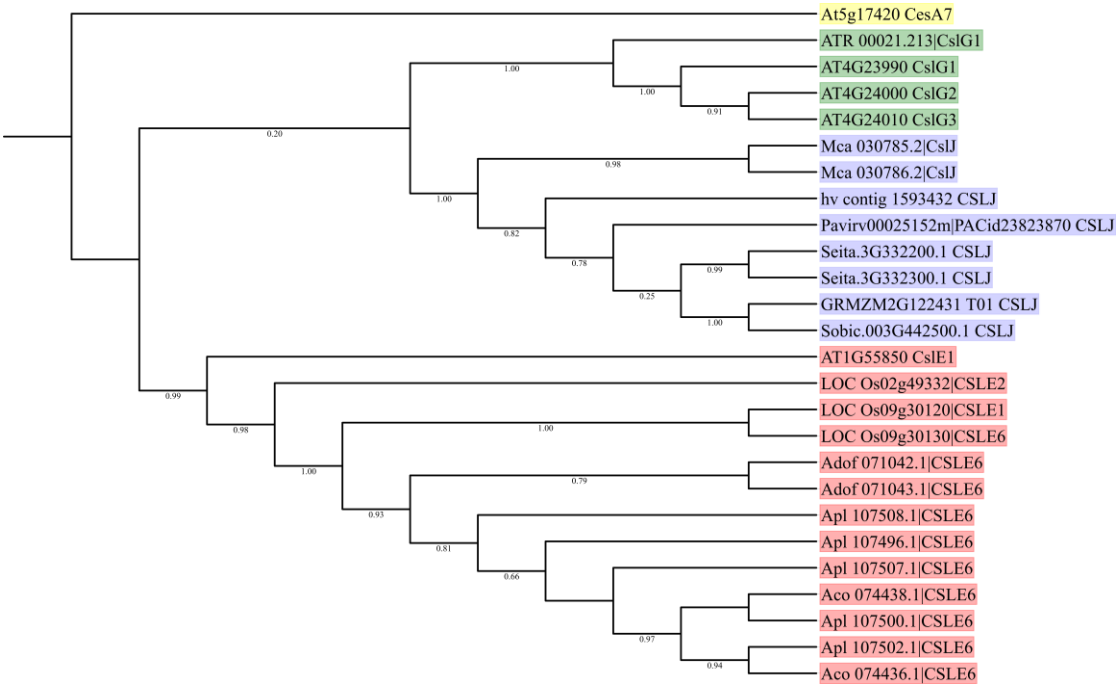
Note: intensified (k > 1) or relaxed (k < 1) selection compared with background branches, LR: likelihood ratio statistic, to compare the alternative and null models (significant difference).

**Supplementary Table 3.4.** Functional annotation of 15 intensify positive selection genes identified by Adaptive Branch-site REL method.

| OGs_ID | GENE NAME | BRANCH | K-test | P-value | LR | DESCRIPTION |
|---|---|---|---|---|---|---|
| OG0017436 | DG1 | node3 | 2.44 | 0.000173285 | 14.1 | Tetratricopeptide-repeat-(TPR) -like-superfamily-protein |
| OG0017508 | NA | hm | 3.64 | 0.001973028 | 9.57 | NA |
| OG0017532 | At1g02020 | ama | 2.38 | 0.006875573 | 7.31 | nitroreductase-family-protein |
| OG0017657 | NRPB9A | adf | 50 | 6.51E-13 | 51.69 | RNA-polymerases-M/15-Kd-subunit |
| OG0017710 | CBSX1 | mc | 2.7 | 0.000464522 | 12.25 | Cystathionine-beta-synthase-(CBS)-family-protein |
| OG0017807 | RH50 | ap | 50 | 0.00270952 | 8.99 | DEA(D/H)-box-RNA-helicase-family-protein |
| OG0017807 | RH50 | node9 | 1.83 | 0.012355806 | 6.26 | DEA(D/H)-box-RNA-helicase-family-protein |
| OG0017835 | T31B5_160 | af | 4.32 | 5.37E-06 | 20.7 | Arginine/glutamate-rich 1 protein |
| OG0018069 | BIP | adf | 43.22 | 0.007204733 | 7.22 | Heat-shock-protein-70-(Hsp-70)-family-protein |
| OG0018170 | At2g01680 | pa | 6.23 | 0.001398769 | 10.21 | Ankyrin repeat-containing protein |
| OG0018204 | CI51 | adf | 3.1 | 0.018170003 | 5.58 | 51 kDa subunit of complex I |
| OG0018205 | At3g15140 | mc | 15.84 | 8.61E-05 | 15.42 | Polynucleotidyl transferase, ribonuclease H-like superfamily protein |
| OG0018254 | GUN1 | af | 28.28 | 4.66E-15 | 61.39 | genomes-uncoupled-1 |
| OG0018353 | PSL5 | ac | 5.98 | 8.29E-07 | 24.29 | Glycosyl-hydrolases-family-31--protein |
| OG0018354 | TIF3C1 | adf | 1.69 | 0.012865503 | 6.19 | eukaryotic-translation-initiation-factor-3C |
| OG0018359 | PA200 | adf | 2.31 | 0.042067067 | 4.13 | Proteasome activator subunit 4 |

**Supplementary Table 3.5.** Functional annotation for 15 genes under positive selection identified by MEME and FUABR.

| OGs_ID | GENE_NAME | TAIR_ID | DESCRIPTION | E-value |
|---|---|---|---|---|
| OG0017439 | YUP8H12.6 | AT1G05320.3 | Myosin heavy chain, embryonic smooth protein | 0.001 |
| OG0017734 | WDL4 | AT2G35880.1 | TPX2 (targeting protein for Xklp2) protein family | 2.00E-49 |
| OG0017740 | LACS9 | AT1G77590.1 | long chain acyl-CoA synthetase 9 | 0 |
| OG0017807 | RH50 | AT3G06980.1 | DEA(D/H)-box RNA helicase family protein | 0 |
| OG0017866 | REIL1 | AT4G31420.1 | Zinc finger protein 622 | 3.00E-166 |
| OG0017873 | CM3,cm-3 | AT1G69370.1 | chorismate mutase 3 | 1.00E-138 |
| OG0017952 | PVA42 | AT4G21450.1 | PapD-like superfamily protein | 6.00E-114 |
| OG0018005 | FATA | AT3G25110.1 | fatA acyl-ACP thioesterase | 2.00E-171 |
| OG0018090 | PKp3 | AT1G32440.1 | plastidial pyruvate kinase 3 | 0 |
| OG0018152 | ABC1K7 | AT3G07700.2 | Protein kinase superfamily protein | 0 |
| OG0018354 | TIF3C1 | AT3G56150.2 | eukaryotic translation initiation factor 3C | 0 |
| OG0018357 | SUD1 | AT4G34100.2 | RING/U-box superfamily protein | 0 |
| OG0018364 | CAMTA5 | AT4G16150.1 | calmodulin binding;transcription regulators | 0 |
| OG0018374 | THO2 | AT1G24706.1 | THO2 | 0 |
| OG0018377 | At5g47690 | AT5G47690.1 | Binding protein | 0 |




**Supplementary Figure 4.1.** Phylogenetic reconstruction of CSLG and CSLJ genes from Arundinoideae species, *Amborella*, *Arabidopsis* and reference genes from the grasses (dataset selected from Schwerdt et al., 2015).

**Supplementary Table 4.1.** CesA/Csl biosynthetic genes information of rice, *Arabidopsis* and *Amborella*.

| Species | Gene_ID | Gene_Name | Species | Gene_ID | Gene_Name |
|---|---|---|---|---|---|
| *Arabidopsis* | AT1G02730 | CslD5 | *Oryza* | Os01g54620 | CESA4 |
| *Arabidopsis* | AT1G23480 | CslA3 | *Oryza* | Os01g56130 | CSLC1 |
| *Arabidopsis* | AT1G24070 | CslA10 | *Oryza* | Os02g09930 | CSLA1 |
| *Arabidopsis* | AT1G32180 | CslD6 | *Oryza* | Os02g49332 | CSLE2 |
| *Arabidopsis* | AT1G55850 | CslE1 | *Oryza* | Os02g51060 | CSLA6 |
| *Arabidopsis* | AT2g21770 | CesA9 | *Oryza* | Os03g07350 | CSLA4 |
| *Arabidopsis* | AT2G24630 | CslC8 | *Oryza* | Os03g26044 | CSLA5 |
| *Arabidopsis* | AT2g25540 | CesA10 | *Oryza* | Os03g56060 | CSLC9 |
| *Arabidopsis* | AT2G32530 | CslB1 | *Oryza* | Os03g59340 | CESA2 |
| *Arabidopsis* | AT2G32540 | CslB2 | *Oryza* | Os03g62090 | CESA5 |
| *Arabidopsis* | AT2G32610 | CslB3 | *Oryza* | Os04g35020 | CSLH2 |
| *Arabidopsis* | AT2G32620 | CslB4 | *Oryza* | Os04g35030 | CSLH3 |
| *Arabidopsis* | AT2G33100 | CslD1 | *Oryza* | Os05g08370 | CESA1 |
| *Arabidopsis* | AT2G35650 | CslA7 | *Oryza* | Os05g43530 | CSLC7 |
| *Arabidopsis* | AT3G03050 | CslD3 | *Oryza* | Os06g02180 | CSLD2 |
| *Arabidopsis* | AT3G07330 | CslC6 | *Oryza* | Os06g12460 | CSLA3 |
| *Arabidopsis* | AT3G28180 | CslC4 | *Oryza* | Os06g22980 | CSLD5 |
| *Arabidopsis* | AT3G56000 | CslA14 | *Oryza* | Os06g39970 | CESA11 |

| | | |
|---|---|---|
| *Arabidopsis* | AT4G07960 | CslC12 |
| *Arabidopsis* | AT4G13410 | CslA15 |
| *Arabidopsis* | AT4G15290 | CslB5 |
| *Arabidopsis* | AT4G15320 | CslB6 |
| *Arabidopsis* | AT4G16590 | CslA1 |
| *Arabidopsis* | AT4g18780 | CesA8 |
| *Arabidopsis* | AT4G23990 | CslG1 |
| *Arabidopsis* | AT4G24000 | CslG2 |
| *Arabidopsis* | AT4G24010 | CslG3 |
| *Arabidopsis* | AT4G31590 | CslC5 |
| *Arabidopsis* | AT4g32410 | CesA1 |
| *Arabidopsis* | AT4G38190 | CslD4 |
| *Arabidopsis* | AT4g39350 | CesA2 |
| *Arabidopsis* | AT5G03760 | CslA9 |
| *Arabidopsis* | AT5g05170 | CesA3 |
| *Arabidopsis* | AT5g09870 | CesA5 |
| *Arabidopsis* | AT5G16190 | CslA11 |
| *Arabidopsis* | AT5g17420 | CesA7 |
| *Arabidopsis* | AT5G22740 | CslA2 |
| *Arabidopsis* | AT5g44030 | CesA4 |
| *Arabidopsis* | AT5g64740 | CesA6 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.297 | CesA1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.298 | CesA1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00013.213 | CslC6 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00017.11 | CSLH1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00017.174 | CSLH1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00017.211 | CesA9 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00017.8 | CSLH1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00019.171 | CslC5 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00019.385 | CslD1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00021.213 | CslG1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00022.123 | CslD4 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00022.66 | CesA3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00040.259 | CslA9 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00045.156 | CesA8 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00048.59 | CslD3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00057.93 | CslD5 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00057.94 | CslD5 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00061.239 | CslG3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00061.240 | CSLE6 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00061.242 | CSLE6 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00065.27 | CslC12 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00067.211 | CesA4 |
| *Oryza* | Os06g42020 | CSLA9 |
| *Oryza* | Os07g03260 | CSLC10 |
| *Oryza* | Os07g10770 | CESA8 |
| *Oryza* | Os07g14850 | CESA6 |
| *Oryza* | Os07g24190 | CESA3 |
| *Oryza* | Os07g36610 | CSLF9 |
| *Oryza* | Os07g36630 | CSLF8 |
| *Oryza* | Os07g36690 | CSLF2 |
| *Oryza* | Os07g36700 | CSLF1 |
| *Oryza* | Os07g36740 | CSLF4 |
| *Oryza* | Os07g36750 | CSLF3 |
| *Oryza* | Os07g43710 | CSLA7 |
| *Oryza* | Os08g06380 | CSLF6 |
| *Oryza* | Os08g15420 | CSLC3 |
| *Oryza* | Os08g25710 | CSLD3 |
| *Oryza* | Os08g33740 | CSLA11 |
| *Oryza* | Os09g25490 | CESA9 |
| *Oryza* | Os09g25900 | CSLC2 |
| *Oryza* | Os09g30120 | CSLE1 |
| *Oryza* | Os09g30130 | CSLE6 |
| *Oryza* | Os10g20090 | CSLH1 |
| *Oryza* | Os10g20260 | CSLF7 |
| *Oryza* | Os10g26630 | CSLA2 |
| *Oryza* | Os10g32980 | CESA7 |
| *Oryza* | Os10g42750 | CSLD1 |
| *Oryza* | Os12g29300 | CESA10 |
| *Oryza* | Os12g36890 | CSLD4 |

| | | |
|---|---|---|
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00067.212 | CesA4 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00072.138 | CslA9 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00081.76 | CesA1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00103.76 | CesA8 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00137.15 | CesA7 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00170.1 | CesA8 |

**Supplementary Table 4.2.** Lignin biosynthetic genes information of rice, *Arabidopsis* and *Amborella*.

| Species | Gene_ID | Gene_Name | Species | Gene_ID | Gene_Name |
|---|---|---|---|---|---|
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00019.261 | 4CL2 | *Oryza* | Os02g08100.1 | 4CL2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00048.51 | 4CL3 | *Oryza* | Os06g44620.1 | 4CL2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00001.269 | 4CL-like1 | *Oryza* | Os08g14760.1 | 4CL2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00001.272 | 4CL-like1 | *Oryza* | Os08g34790.1 | 4CL2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00023.76 | 4CL-like1 | *Oryza* | Os02g46970.1 | 4CL3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00049.77 | 4CL-like1 | *Oryza* | Os03g04000.1 | 4CL-like1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00050.38 | 4CL-like1 | *Oryza* | Os07g44560.1 | 4CL-like1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00022.259 | 4CL-like4 | *Oryza* | Os04g24530.1 | 4CL-like5 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00025.296 | 4CL-like5 | *Oryza* | Os08g04770.1 | 4CL-like6 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00025.297 | 4CL-like5 | *Oryza* | Os10g42800.1 | 4CL-like6 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00048.90 | 4CL-like6 | *Oryza* | Os03g05780.1 | 4CL-like7 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00076.77 | 4CL-like7 | *Oryza* | Os01g67530.1 | 4CL-like8 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00076.80 | 4CL-like7 | *Oryza* | Os01g67540.1 | 4CL-like8 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00011.169 | 4CL-like8 | *Oryza* | Os07g17970.1 | 4CL-like8 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00023.75 | 4CL-like8 | *Oryza* | Os05g41440.1 | C3H1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00049.71 | 4CL-like8 | *Oryza* | Os10g12080.1 | C3H1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00049.72 | 4CL-like8 | *Oryza* | Os01g60450.1 | C4H |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00049.73 | 4CL-like8 | *Oryza* | Os02g26770.1 | C4H |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00049.76 | 4CL-like8 | *Oryza* | Os02g26810.1 | C4H |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00040.62 | C3H1 | *Oryza* | Os05g25640.1 | C4H |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00101.79 | C3H1 | *Oryza* | Os04g15920.1 | CAD1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00025.258 | C4H | *Oryza* | Os04g52280.1 | CAD1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00077.91 | C4H | *Oryza* | Os09g23550.1 | CAD1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00030.57 | CAD1 | *Oryza* | Os02g09490.1 | CAD2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00030.58 | CAD1 | *Oryza* | Os03g12270.1 | CAD3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00040.38 | CAD1 | *Oryza* | Os08g16910.1 | CAD3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00015.12 | CAD2 | *Oryza* | Os10g29470.1 | CAD3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00018.80 | CAD9 | *Oryza* | Os09g23530.1 | CAD4 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00036.174 | CCoAOMT1 | *Oryza* | Os09g23540.1 | CAD4 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00046.27 | CCoAOMT1 | *Oryza* | Os09g23560.1 | CAD4 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00036.168 | CCoAOMT2 | *Oryza* | Os10g11810.1 | CAD9 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00018.91 | CCoAOMT4 | *Oryza* | Os11g40690.1 | CAD9 |

| | | | | | |
|---|---|---|---|---|---|
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00036.171 | CCoAOMT6 | *Oryza* | Os06g06980.1 | CCoAOMT1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00036.170 | CCoAOMT7 | *Oryza* | Os08g38900.1 | CCoAOMT1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00065.159 | CCR1 | *Oryza* | Os08g38910.1 | CCoAOMT1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00099.162 | CCR-like2 | *Oryza* | Os08g38920.1 | CCoAOMT1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00022.133 | CCR-like3 | *Oryza* | Os09g30360.1 | CCoAOMT1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00022.134 | CCR-like3 | *Oryza* | Os08g05790.1 | CCoAOMT4 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00022.136 | CCR-like3 | *Oryza* | Os01g18110.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00022.137 | CCR-like3 | *Oryza* | Os01g18120.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00153.35 | CCR-like5 | *Oryza* | Os02g08420.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00153.36 | CCR-like5 | *Oryza* | Os02g56460.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00001.509 | COMT | *Oryza* | Os02g56680.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00003.285 | COMT | *Oryza* | Os02g56690.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00003.287 | COMT | *Oryza* | Os02g56700.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00003.288 | COMT | *Oryza* | Os08g17500.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00003.289 | COMT | *Oryza* | Os08g34280.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00003.290 | COMT | *Oryza* | Os09g04050.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00003.292 | COMT | *Oryza* | Os09g08720.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00003.299 | COMT | *Oryza* | Os09g25150.1 | CCR1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00009.209 | COMT | *Oryza* | Os02g56720.2 | CCR2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00058.148 | COMT | *Oryza* | Os01g74660.1 | CCR-like2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00062.194 | COMT | *Oryza* | Os06g41810.1 | CCR-like3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold02092.1 | COMT | *Oryza* | Os06g41840.1 | CCR-like3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold03062.1 | COMT | *Oryza* | Os08g08500.1 | CCR-like3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold03202.1 | COMT | *Oryza* | Os09g31490.1 | CCR-like3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold03713.1 | COMT | *Oryza* | Os09g31498.1 | CCR-like3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.374 | COMT-like11 | *Oryza* | Os09g31502.1 | CCR-like3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00741.1 | COMT-like6 | *Oryza* | Os09g31514.1 | CCR-like3 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00001.451 | F5H1 | *Oryza* | Os01g61230.1 | CCR-like5 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00001.452 | F5H1 | *Oryza* | Os03g60380.1 | CCR-like5 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.409 | HCT | *Oryza* | Os04g01470.1 | COMT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.410 | HCT | *Oryza* | Os04g09604.1 | COMT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.411 | HCT | *Oryza* | Os04g09654.1 | COMT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.413 | HCT | *Oryza* | Os08g06100.1 | COMT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00002.414 | HCT | *Oryza* | Os12g13800.1 | COMT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00038.22 | HCT | *Oryza* | Os02g57760.1 | COMT-like11 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00038.23 | HCT | *Oryza* | Os03g02180.1 | F5H1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00038.27 | HCT | *Oryza* | Os06g24180.1 | F5H1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00058.197 | HCT | *Oryza* | Os10g36848.1 | F5H1 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00058.198 | HCT | *Oryza* | Os02g39850.1 | HCT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00058.199 | HCT | *Oryza* | Os04g42250.2 | HCT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00137.34 | HCT | *Oryza* | Os06g08580.1 | HCT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00137.5 | HCT | *Oryza* | Os06g08640.1 | HCT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00727.2 | HCT | *Oryza* | Os09g25460.1 | HCT |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00148.59 | PAL1 | *Oryza* | Os02g41630.2 | PAL1 |

| | | |
|---|---|---|
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00024.177 | PAL2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00024.178 | PAL2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00024.181 | PAL2 |
| *Amborella* | evm_27.model.AmTr_v1.0_scaffold00032.129 | PAL4 |
| *Arabidopsis* | At1g51680 | 4CL1 |
| *Arabidopsis* | At3g21240 | 4CL2 |
| *Arabidopsis* | At1g65060 | 4CL3 |
| *Arabidopsis* | At3g21230 | 4CL4 |
| *Arabidopsis* | At1g20510 | 4CL-like1 |
| *Arabidopsis* | At1g20500 | 4CL-like2 |
| *Arabidopsis* | At1g20490 | 4CL-like3 |
| *Arabidopsis* | At1g20480 | 4CL-like4 |
| *Arabidopsis* | At1g62940 | 4CL-like5 |
| *Arabidopsis* | At4g19010 | 4CL-like6 |
| *Arabidopsis* | At4g05160 | 4CL-like7 |
| *Arabidopsis* | At5g63380 | 4CL-like8 |
| *Arabidopsis* | At5g38120 | 4CL-like9 |
| *Arabidopsis* | At2g40890 | C3H1 |
| *Arabidopsis* | At1g74540 | C3H2 |
| *Arabidopsis* | At1g74550 | C3H3 |
| *Arabidopsis* | At2g30490 | C4H |
| *Arabidopsis* | At4g39330 | CAD1 |
| *Arabidopsis* | At3g19450 | CAD2 |
| *Arabidopsis* | At4g37970 | CAD3 |
| *Arabidopsis* | At4g37980 | CAD4 |
| *Arabidopsis* | At4g37990 | CAD5 |
| *Arabidopsis* | At4g34230 | CAD6 |
| *Arabidopsis* | At2g21730 | CAD7 |
| *Arabidopsis* | At2g21890 | CAD8 |
| *Arabidopsis* | At1g72680 | CAD9 |
| *Arabidopsis* | At4g34050 | CCoAOMT1 |
| *Arabidopsis* | At1g24735 | CCoAOMT2 |
| *Arabidopsis* | At3g61990 | CCoAOMT3 |
| *Arabidopsis* | At3g62000 | CCoAOMT4 |
| *Arabidopsis* | At1g67990 | CCoAOMT5 |
| *Arabidopsis* | At1g67980 | CCoAOMT6 |
| *Arabidopsis* | At4g26220 | CCoAOMT7 |
| *Arabidopsis* | At1g15950 | CCR1 |
| *Arabidopsis* | At1g80820 | CCR2 |
| *Arabidopsis* | At1g76470 | CCR-like1 |
| *Arabidopsis* | At2g02400 | CCR-like2 |
| *Arabidopsis* | At2g33590 | CCR-like3 |
| *Arabidopsis* | At2g33600 | CCR-like4 |
| *Arabidopsis* | At5g58490 | CCR-like5 |
| *Oryza* | Os02g41650.1 | PAL1 |
| *Oryza* | Os02g41670.1 | PAL1 |
| *Oryza* | Os02g41680.1 | PAL1 |
| *Oryza* | Os04g43760.1 | PAL1 |
| *Oryza* | Os04g43800.1 | PAL1 |
| *Oryza* | Os05g35290.1 | PAL1 |
| *Oryza* | Os11g48110.1 | PAL1 |
| *Oryza* | Os12g33610.1 | PAL1 |

| | | |
|---|---|---|
| *Arabidopsis* | At4g35150 | COMT |
| *Arabidopsis* | At4g35160 | COMT |
| *Arabidopsis* | At5g54160 | COMT |
| *Arabidopsis* | At1g21100 | COMT-like1 |
| *Arabidopsis* | At1g77530 | COMT-like10 |
| *Arabidopsis* | At3g53140 | COMT-like11 |
| *Arabidopsis* | At5g37170 | COMT-like12 |
| *Arabidopsis* | At5g53810 | COMT-like13 |
| *Arabidopsis* | At1g21110 | COMT-like2 |
| *Arabidopsis* | At1g21120 | COMT-like3 |
| *Arabidopsis* | At1g21130 | COMT-like4 |
| *Arabidopsis* | At1g33030 | COMT-like5 |
| *Arabidopsis* | At1g51990 | COMT-like6 |
| *Arabidopsis* | At1g63140 | COMT-like7 |
| *Arabidopsis* | At1g76790 | COMT-like8 |
| *Arabidopsis* | At1g77520 | COMT-like9 |
| *Arabidopsis* | At4g36220 | F5H1 |
| *Arabidopsis* | At5g04330 | F5H2 |
| *Arabidopsis* | At5g48930 | HCT |
| *Arabidopsis* | At2g37040 | PAL1 |
| *Arabidopsis* | At3g53260 | PAL2 |
| *Arabidopsis* | At5g04230 | PAL3 |
| *Arabidopsis* | At3g10340 | PAL4 |

# Acknowledgment

I would like to thank my supervisor Prof. Giorgio Bertorelle for mentoring and support from the Doctoral programme in Evolutionary Biology and Ecology of Ferrara University. His understanding, discussion and suggestions during my Ph.D. study gave me great motivation to complete my doctoral project and study for my doctorate, and I am fortunate to have the opportunity to cooperate with him.

I am really grateful to Dr. Claudio Varotto for providing me the opportunity to work in his group at the Biodiversity and Molecular Ecology Department of Edmund Mach Foundation (FEM). His friendship and patient guidance greatly contributed to complete my doctoral project.

I also wish to express my gratitude to China Scholarship Council (CSC) for providing the fellowship for my Ph.D and many thanks to my friends and the colleagues specially Dr. Mingai Li, Mr. Enrico Barbaro, Dr. Roberto Biello, Dr. Bo Wang, Dr. Luca Stragliati, Dr. Huan Li and Dr. Mastaneh Ahrar for their friendships and kindly help during my Ph.D.

Finally, I would like to deeply express thanks to my family especially my parents and brothers for their unconditional support and encouragement.

**November 2018**

**Jike Wuhe**