Sónia Cristina Heleno Correia

Bachelor degree in Applied Chemistry

Molecular analysis of microRNA-target gene interactions in the pine seed

Dissertation to obtain the Master Degree in
Molecular Genetics and Biomedicine

Supervisor: Célia Miguel, PI, iBET & ITQB NOVA.

Jury:

Arguer: Dr. Pedro Miguel Rodrigues de Barros
President: Prof. Dra. Paula Maria Theriaga Mendes Bernardo Gonçalves
Supervisor: Prof. Dra. Célia Maria Rodrigues Miguel

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

March 2017

LOMBADA

Molecular analysis of microRNA-target gene interactions in the pine seed

Sónia Correia

2017

Sónia Cristina Heleno Correia
Bachelor degree in Applied Chemistry

Dissertation to obtain the Master Degree in Molecular Genetics and Biomedicine

Molecular analysis of microRNA-target gene interactions in the pine seed

Supervisor: Célia Miguel, PI, iBET & ITQB NOVA

Jury:
Arguer: Dr. Pedro Miguel Rodrigues de Barros
President: Prof. Dra. Paula Maria Theriaga Mendes Bernardo Gonçalves
Supervisor: Prof. Dra. Célia Maria Rodrigues Miguel

Universidade NOVA de Lisboa
Faculdade de Ciências e Tecnologia
Monte da Caparica, March 2017

Molecular analysis of microRNA-target gene interactions in the pine seed

Acknowledgements

To all my family, friends and colleagues who supported and believed in me, and went along with me in the good and bad moments, thank you!

To Célia Miguel who accepted and welcomed me for this project thesis, being perseverant and always available for doubts discussion and to guide me in the right direction, I'm so grateful!

To Andreia Rodrigues, my laboratory mother, who explained and taught almost everything in the lab, who dedicated me time of her PhD thesis, who spend whole nights and weekends in the hard final moments, and even in the Netherlands continued to help me in whatever I needed, thank you so, so much! You are a model who I have in great consideration a big kiss and luck to finish your thesis and in this step of your journey.

To Bruno Costa, the lab bioinformatitian, who taught me the first steps in command lines and how to deal with Linux, who has been available until the end, a big thank you and I owe you a coffee, beer and chocolates!

To Susana who performed many times of live mannequin so I could observe and absorb technics, who taught me how to maintain liquid cell cultures, how to isolate protoplasts, and to perform viability tests using the right fluorophores, thank you for your kindness and reasonable thinking.

To the rest of my lab colleagues Inês Chaves, Inês Modesto, Sofia Leal, Sofia Gonçalves, José Cerca, Cirenia Arias, Andreia Matos and Pedro, who created such a good work environment thank you for the support and contribution with ideas and discussion of work problems, and for the stories, teachings, laughs and joy in the pause moments, more than colleagues you became friends.

To my family who supported me and my special temperament in the last few times of writing, sorry and thank you. To my brother who always kept my feet on the ground, to my sister for her adolescent craziness and sweetness, to my cousin for the long talks, laughs and unconditional support, thank you dears I love you! To my grandfathers for all the contribute in my life, they have part in what I am now. And a giant thanks to my parents, for their love, patient and support, without them this wasn't possible.

To my friends I appreciated all the moments, and I'm grateful for all the support and motivation. To my master friends Ana, Sara, Mafalda, Prata, and João, thank you for all the dinners and coffees, which help me to keep my mental sanity, with a special kiss to Zapico. To Rita Correia, Martinho Luther, Rita Catarina, Diogo Freitas, and to all friends that somewhat participated in my live during this graduation thank you for the support and fun.

Resumo

Pinus pinaster é uma espécie de conífera com elevado valor ecológico e económico. O território ocupado por esta espécie, distribuída pela região mediterrânea, tem vindo a diminuir sobretudo devido a causas ambientais. A embriogénese somática possibilita a propagação em grande escala de plantas adaptadas a variados stresses permitindo contornar este problema. No entanto, esta técnica encontra-se pouco desenvolvida em *P. pinaster. O* estudo de reguladores da expressão génica como miARNs poderá contribuir para a sua optimização.

Os miARNs regulam a expressão génica a nível pós-transcripcional através da clivagem de mARNs ou da repressão da tradução. Neste trabalho, foram estudadas quatro interacções de clivagem entre miARN e mARN alvo. Primeiro, através da análise de dados de sequenciação com recurso a ferramentas bioinformáticas, foram previstos os mARNs alvo para uma lista de miARNs, dos quais sete pares miARN-alvo foram seleccionados. Seguidamente foi analisada a expressão de quatro mARNs alvo por RT-qPCR, em dois estádios embriogénicos, T4B (pré-cotiledonar) e T7 (maduro). Os quatro transcritos incluem *ARF* 10, 16, 17 (regulado pelo miR160) e *F-box* (regulado pelo miR482a); e dois clivados por novos miARNs: *LEA* (regulado pelo M09664) e *PPR* (regulado pelo M06658). Todos os transcritos apresentaram maior expressão em T4B.

Outros parâmetros como, conservação do miARN e do seu alvo, complementaridade entre miARN e alvo, e comparação de expressão entre miARNs e respectivos mARNs alvo, foram também abordados de forma a contribuir para a selecção de pares com elevada confiança. Apenas o par miARN160 e alvo *ARF* preencheu todas as condições.

Assim, foram identificados pela primeira vez potenciais alvos de miARNs em *P. pinaster*. Este trabalho contribuiu para o aumento do conhecimento sobre miRNAs em coníferas, apontando para potenciais funções no desenvolvimento embrionário. No entanto, são precisos mais estudos para validação das interacções entre os pares miARN-mARN alvo aqui estudados.

Palavras-chave: *Pinus pinaster*; regulação da embriogénese; miARN; mARN alvo; PARE; RT-qPCR.

Abstract

Pinus pinaster is a conifer species with high ecologic and economic value which covers a vast area of the Mediterranean region. However, the area occupied by *P. pinaster* has been decreasing mainly due to environmental causes. Somatic embryogenesis may contribute to circumvent this problem as it allows the large-scale propagation of plants adapted to several stresses. However, this technique is poorly understood in *P. pinaster* and the study of gene expression regulators like miRNAs may contribute to its improvement.

MiRNAs regulate gene expression at post-transcriptional level by mRNA cleavage or translation repression. In this work, a set of four miRNA-mRNA target cleavage interactions were studied. First, mRNAs targets were predicted for a list of miRNAs using bioinformatics tools to analyze available sequencing data, and seven miRNA-mRNA pairs were selected. Second, the expression of four of the mRNAs were analyzed by RT-PCR in two embryo developmental stages, T4B (pre-cotyledonary) and T7 (mature). The four transcripts included *ARF 10, 16 ou 17* (regulated by miR160), *F-box* (regulated by miR482a), *LEA* (regulated by miRM09664) and *PPR* (regulated by miRM06658). All transcripts show a higher expression in T4B than in T7 stage. Several parameters were studied to ensure that high-confidence pairs were selected, namely conservation of miRNA and respective target, complementarity between miRNA and respective target, relation between miRNA expression and target expression. Only miRNA160-*ARF* target fulfilled all the conditions.

It was the first time that potential miRNA targets were identified in *P. pinaster*. This work contributed to uncover the miRNA landscape in conifers, pointing to potential functions in embryo development. However, more experiments are needed to further validate the interaction between the miRNA-mRNA targets studied.

Keywords: *Pinus pinaster*; embryogenesis regulation; miRNA; mRNA target; PARE; RT-qPCR.

Contents

Index of Equations

Index of Figures

Index of Tables

List of abbreviations and acronyms

3'-UTR, 3' untranslated region;

5'-RLM-RACE, 5'-RNA-ligase-mediated rapid amplification of cDNA ends;

ARF, auxin regulator factor;

ATUB, α-tubulin;

EF1, elongation factor 1;

EST, expressed sequence tag;

dsRNA, double stranded RNA;

Fwd, forward;

GOI, gene of interest;

HISTO 3, histone 3;

LEA, late embryo abundant;

MGM, megagametophyte

MRE, microRNA response element;

NGS, next-generation silencing;

NTC, non-template control;

PARE, parallel analysis of RNA ends;

Pol, polymerase;

PTGS, post-transcriptional gene silencing;

Rev, reverse;

RG, reference gene;

RT, reverse transcriptase;

RT-qPCR, real time quantitative PCR;

SE, somatic embryogenesis;

se, standard error;

TAE, tris-acetate-EDTA;

Ta, annealing temperature;

Ubiq, ubiquitin;

ZE, zygotic embryo.

1. Introduction

1.1. *Pinus pinaster* - Distribution, ecologic and economic importance, deforestation and possible solutions

Pinus is a genus of the *Pinaceae* family and includes several species, among which are *Pinus pinaster* Aiton*,* also known, among others, as *Pinus maritima* (or "Pinheiro-bravo" in Portugal). This conifer forest tree occurs in pure afforestations, or mixed with other forest tree species. It is distributed along the Mediterranean basin, more specifically in regions of the Iberian Peninsula, southwest and southern Europe, and the north of African coast (Jalas and Suominen 1972; Farjon 2013). The largest continuous plantation forest in Europe is in France where maritime pine is the main species. *P. pinaster* forests provide habitats for countless species and ecosystems. It has been also planted for soil conservation purposes, contributing to a good structure and fertile soil. It also represents an important source of raw materials for human use, such as wood, resin and paper pulp (Abad Viñas et al. 2016; Farjon 2010).

P. pinaster has preference for winter rainfall regime and mild temperatures but, due to its fast growth characteristics and tolerance to poor soils, it adapts to a range of diverse climates and habitats. During the 20th century, temperature in the Mediterranean basin increased by 1.5-4°C depending on the sub-region (Ferragina and Quagliarotti 2008) which consequently led to a reduction of precipitation in this area. These changes in climate are expected to continue leading to dryer soils and an increased risk of forest fires, contributing also to exposure to new pests and diseases as well as reproduction limitations. Overall, forests have been decreasing (The world Bank 2016). In Portugal, from 1965 to 2010 the area occupied by this species decreased approximately 263 thousand ha, occupying in 2010 approximately 23% of the Portuguese forest land (ICNF-IFN6 2013). The fast global climatic change scenario is threatening the biological diversity of the forest ecosystems and, as a consequence, the available genetic variability of Mediterranean forest species such as *P. pinaster* may also be compromised. The loss of such diversity may lead in the future to huge environmental, social and economic losses.

Human interventions, such as transfer of forest reproductive material, genetic improvement/breeding programs, and large-scale vegetative propagation tools such as somatic embryogenesis (SE) may integrate strategies for increasing productivity and should be promoted to improve the adaptation of *P. pinaster* to the predicted climate scenario of extreme temperatures and higher water stress (EUFORGEN 2007).

1.2. Embryogenesis in Gymnosperms

Plant zygotic embryogenesis (ZE) comprehends the period of plant development that begins with the fertilized egg and culminates in the mature desiccated zygotic embryo surrounded by a protective seed coat (Mordhorst et. al 1997). This is valid for the seed plants groups, namely the angiosperms like the plant model *Arabidopsis thaliana*, and the

gymnosperms like *P. pinaster*. As a monoic species, *P. pinaster* female cones are found in the upper branches of the tree, where they may be fertilized by pollen brought by the wind from the male cones, which are also present in the lower branches of the same tree. Unlike angiosperms, conifers have a single fertilization event (Cairney and Pullman 2007) where the male gametophyte (pollen grain, n) migrates through the egg cytoplasm (female gametophyte, n) and fuses with the egg nucleus, originating the zygote (2n) which will develop in embryo (Von Arnold 2008) (see figure 1). The megagametophyte developed upon fertilization is of mother origin (n) only and surrounds and nourishes the embryo while it grows. This conifer species takes two years from pollination to seed maturation (Cairney and Pullman 2007).

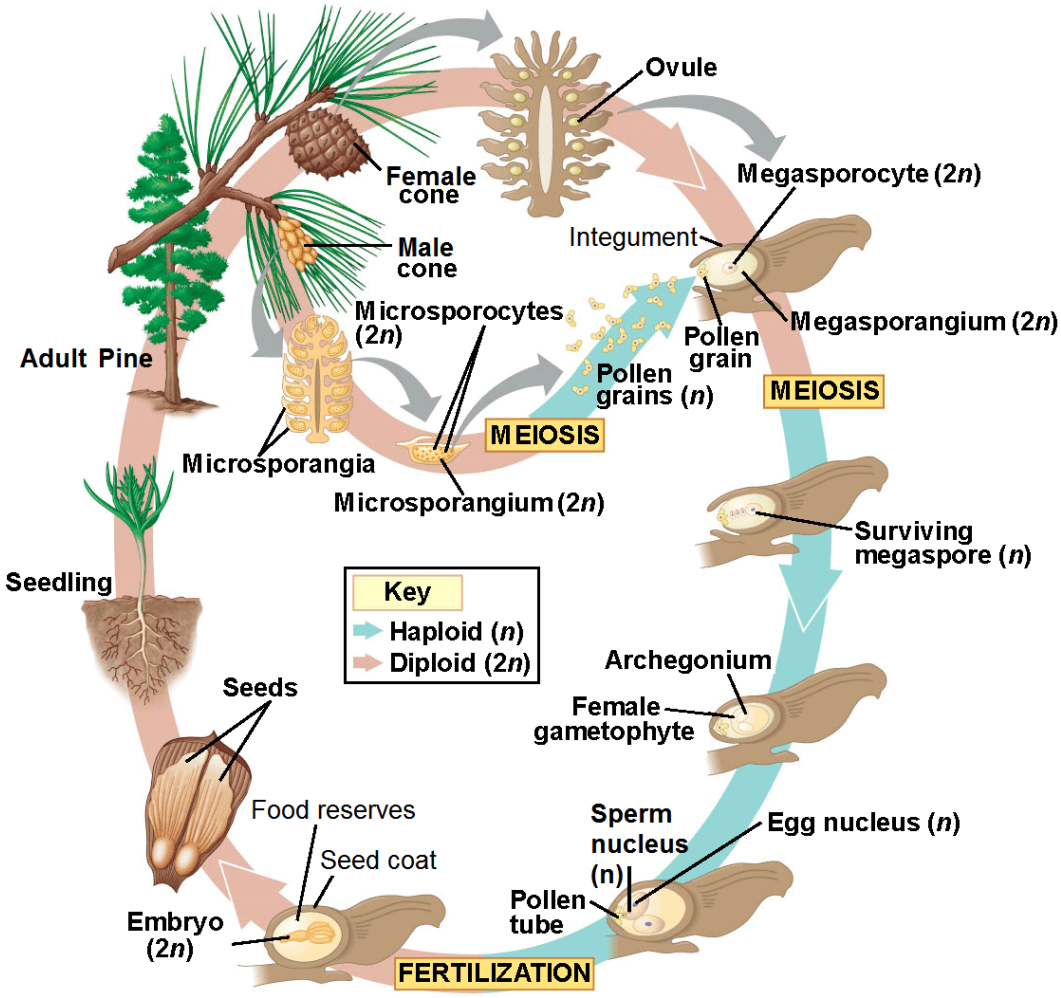


Figure 1.I Pine life cycle. Starting with the seed germination into a fertile adult tree, the development of its sexual organs and production of gametes, pollination, and fertilization to produce a new seed. Image adapted from (Campbell and Reece 2008).

In plant embryogenesis there is an initial morphogenetic phase followed by a maturation phase. The morphogenetic phase is characterized by cell division and the onset of cell differentiation, while maturation phase is characterized by an accumulation of major storage

products and preparation for seed desiccation, dormancy and germination (Meinke 1995). The morphogenetic phase, can be divided into three phases, according to Singh 1978, proembryogeny phase, early embryogeny and late embryogeny (see figure 2) (Von Arnold 2008).

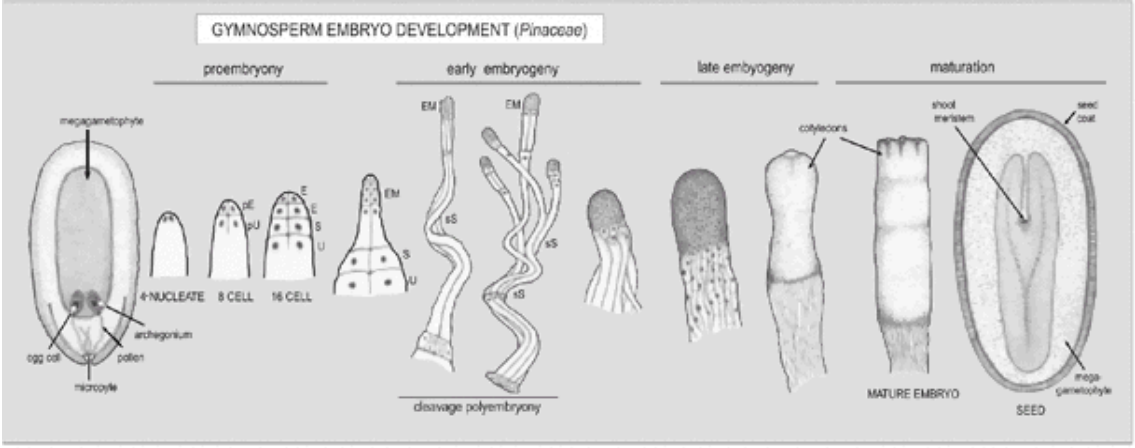


Figure 1.II Schematic overview of gymnosperm (*Pinaceae*) embryo development. Image adapted from (Von Arnold 2008). Abbreviation: EP-embryo proper; pU-primary upper tier; pE-prymary embryonal tier; E-embryonal tier; S-suspensor tier; U-upper tier; EM-embryonal mass; sS-secondary suspensor.

The proembryogeny consists in the asymmetric division of the zygote. For *Pinus spp.*, the fertilized nuclei divides to produce four free nuclei which later result in two initial cell tiers, the primary embryonal tier and the primary upper tier (See figure 2). These tiers divide again, and this phase ends with the elongation of the embryonal tier which will form the embryonal mass, and the remaining suspensor and upper tiers will give rise to the embryonal suspensor. During the early embryogeny, cells of the lower embryonal tier divide predominantly in a transverse plane and elongate, resulting in the embryonal mass and the thick secondary suspensor. The late embryogenesis is a period of histogenesis and organogenesis, were the root and shoot meristems are delineated and the plant apical-basal axis is established. The cotyledon primordia arise in a ring around the distal end of the embryo. At this stage provascular tissue and cortex also differentiate. By the end of the morphogenesis phase, the development program switches from pattern formation to the maturation phase, a period in which cell expansion and accumulation of storage reserves occur (Arnold 2008).

For simplicity, nine different stages along zygotic embryo development have been proposed in maritime pine by Gonçalves et al. (2005a) (figure 1.III) based on the staging system of Pullman and Webb (1994). These nine stages can be further grouped in early (T0, T1 and T2), middle (T3, T4 pre-cotyledonary embryos and T4B early-cotyledonary embryo) and late (T5, T6 cotyledonary embryos and T7 mature embryo) embryogenesis (de Vega-Bartol et al. 2013b)

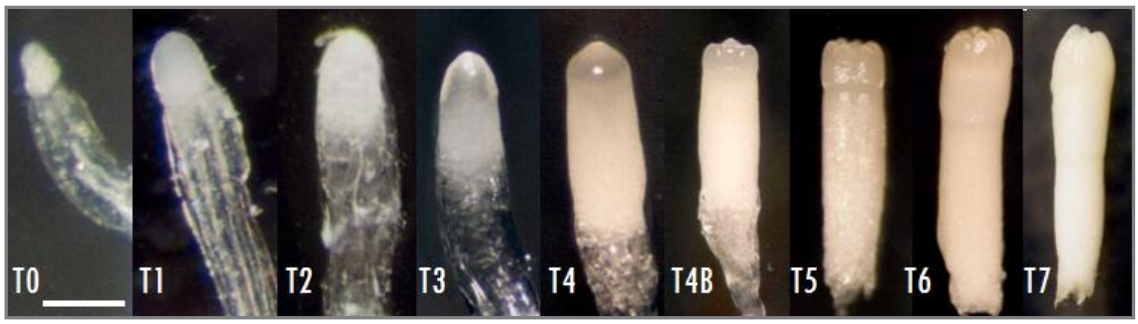


Figure 1.III *Pinus pinaster* zygotic embryo developmental stages, from T0 to T7, according to the staging system of Gonçalves et al, 2005a. Bar: T0 and T1 = 300 µm; T2, T3 and T4 = 400 µm; T4B = 800 µm; T5, T6 and T5 = 0.1 cm (Gonçalves et al. 2005b).

1.3. Somatic embryogenesis

Fertile plants can develop from a single somatic cell and not just from a fertilized egg. Somatic embryogenesis (SE) is a fascinating nonsexual propagation process whereby somatic cells differentiate into somatic embryos (Von Arnold 2008).

Somatic embryos pass through a series of developmental stages which are similar to ZE and it has been observed in both angiosperms (Mordhorst et al. 1997) and gymnosperms (Cairney and Pullman 2007). For this reason SE has also been used as a model system to study plant embryogenesis (Von Arnold 2008; Trontin et al. 2016). SE offers a great potential to be applied in clonal selection, allowing the propagation of elite genotypes (Klimaszewska et al. 2016; Park et al. 2016). It offers the capability for large-scale vegetative propagation (Merkle 2016) particularly because of the possibility to scale up the propagation by using bioreactors (Von Arnold 2008). It is also an ideal system for genetic transformation, due to the initiation of SE from single cells (Montalbán et al. 2016; Winkelmann 2016). Somatic embryogenesis complemented by the application of other biotechnological tools offers new strategies for forest trees improvement favoring productivity, quality, and production systems of elite plants adapted to different environmental conditions, and more resistant or tolerant to abiotic and biotic stresses. (Merkle 2016; Montalbán et al. 2016). However, *in vitro* somatic embryogenesis is still subjected to several limitations. The application of SE to forestry plantation is currently restricted to a few conifer species due to deficiencies in the current protocols, which affect the induction, maturation, germination and plantlet conversion steps of SE (Park et al. 2016). Problems such as low or asynchronous embryo production, abnormal morphology, or poor root development have been reported (Montalbán et al. 2016).

In conifers, little is known about gene expression during the early stages of embryogenesis, which is recognized to be critical for subsequent development studies. Also the elucidation of the molecular events regulating embryo development in trees, and particularly in conifers has been hindered by their large physical size, slow growth, long generation time, and very large genome (Trontin et al. 2016).

Efficient development of somatic embryos requires a number of critical physical and chemical treatments with proper timing and is highly dependent of genetic features, unlike zygotic embryogenesis (Winkelmann 2016). Studies are therefore required to increase our understanding of the basic mechanisms governing the highly complex embryonic phase in conifers and the massive gene regulation (Von Arnold et al. 2016; Trontin et al. 2016). Epigenetic mechanisms, including DNA methylation, histone modifications and RNA interference, play a major role in gene regulation and have attracted a lot of attention in recent years.

SmallRNA pathways in particular, and especially miRNAs, have been highlighted as important regulators of developmental processes, among others, in several species including conifers (Mahdavi-Darvari et al. 2014; Miguel et al. 2016; de Vega-Bartol et al. 2013b)

By taking zygotic embryos as the reference, learning from and mimicking the situation in seeds, somatic embryogenesis can be improved and optimized in order to make use of the enormous potential this regeneration pathway offers for plant propagation and breeding (Winkelmann 2016).

1.4. Small RNAs (in plants)

Small RNAs (sRNAs) expressed by eukaryotes are typically 20-30 nt non-coding RNA sequences that regulate several biological processes at the DNA or RNA level (Chen 2009).

There are a large diversity and complexity in small RNAs world. According to Axtell (2013) the small RNAs can be classified in two general groups according to their origin: hp-RNAs (hairpin small RNAs), derived from intermolecular self-complementary RNA hairpin structures, and siRNAs, derived from dsRNA precursors. A secondary classification could be stablished based on other biogenesis characteristics or small RNA function. Thus, hp-RNAs could be divided in miRNAs, the aim of this study (the characteristics are described in the next sub-chapters), and other hp-RNAs non-miRNAs, a group barely annotated. SiRNAs, in its turn, are divided in three groups: heterochromatic siRNAs, responsible for repressive chromatin modifications but not found in conifers; secondary siRNAs, which its biogenesis is triggered by one of the above small RNAs; and NAT-siRNAs (natural antisense siRNAs) which instead of derive from an overlapping RNA locus, they arise from two independent but complementary transcribed RNAs. This hierarchical classification is not static since new miRNAs could be found or new functions could be revealed, or even other characteristics could be taken into account. SiRNAs could, for instance, be classified into tasiRNA (trans-acting siRNAs), casiRNA (cis-acting siRNAs), exo-siRNA (exogenous siRNAs) and endo-siRNA (endogenous siRNAs) (Ghildiyal and Zamore 2009).

1.4.1. The biogenesis of miRNAs in plants

Plant miRNAs were first identified in 2002 (Reinhart et al 2002) and normally range from 20 to 24 nucleotides. They are mostly transcribed from independent or clustered, non-protein-

coding loci on the intergenic regions of the genome (Budak and Akpinar 2015; Jones-Rhoades et al. 2006). Besides these intergenic miRNA loci, intronic miRNAs are being increasingly reported, transcribed from the spliced out introns of the protein-coding genes (called mitrons) and, in spite of rare, exonic miRNAs, transcribed from the exons of protein-coding genes, have also been described in plants (Budak and Akpinar 2015).

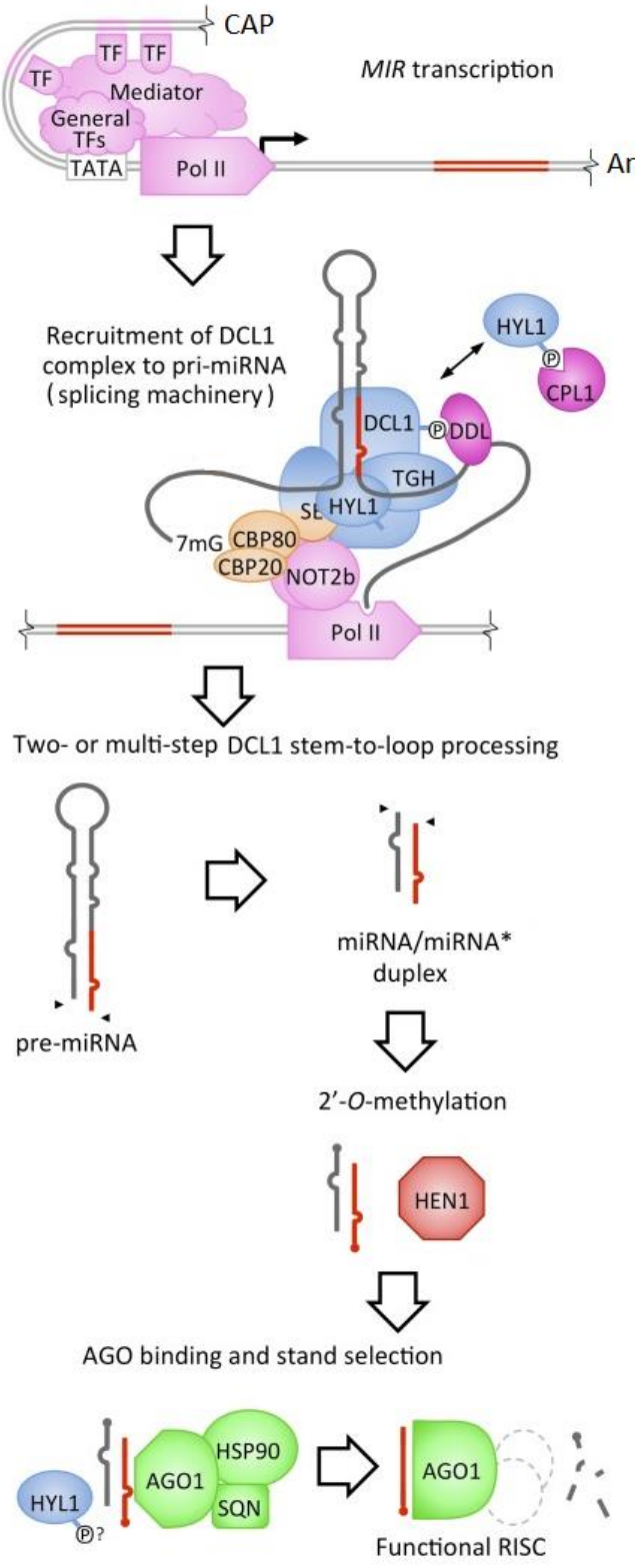


Figure 1.IV Summary of the major steps in miRNA biogenesis. Adapted from (Rogers and Chen 2013)

The precursor sequences of miRNAs are largely determined by its genomic sequences (Meng et al. 2011). RNA polymerase II is responsible for transcribing the majority of miRNA genes. The recruitment of Pol II to MIR promoters involves the iteration of several transcriptional activators and various sequence motifs, similarly to the protein-coding genes (Meng et al. 2011; Rogers and Chen 2013) (see figure 1.IV). The transcript is stabilized by polyadenylation (addition of a 3' poly-A tail), and capping (addition of 5' 7-methylguanosine cap) and folds into a hairpin structure termed as miRNA primary transcript (pri-miRNA).

After transcription, pri-miRNAs are subjected to DCL-mediated two-step cleavage with help of other proteins and motifs (see figure 1 IV). This process occurs inside the nucleus, within specialized compartments called Dicing-bodies (D-bodies) (Meng et al. 2011). Dicer-like proteins (DCLs) have RNase III activity and initially cleave the pri-miRNAs near to the base of the stem, into a shorter stem-loop structure called precursor miRNA (pre-miRNA) (Rogers and Chen 2013). A second subsequent cleavage event

occurs in the pre-miRNA releasing the miRNA/miRNA* duplex (Budak and Akpinar 2015). Different DCL family members give rise to miRNAs with distinct sizes. DCL1 is responsible for the biogenesis of most plant miRNAs that mainly fall in the 18–21nt size range (Rogers amd Chen 2013).

Once the miRNA-duplex form is ready, the methyltransferase HEN1 (Hua Enhancer 1) adds a methyl group to the 2' OH of the 3' terminal nucleotide, thus protecting this termination against uridylation and consequent degradation by exonucleases (Meng et al. 2011). The export mechanism (from nuclei to cytoplasm) for plant miRNAs remains unknown (Budak and Akpinar 2015; Park et al. 2005).

Finally, to exert their regulatory roles, the mature miRNAs must be associated to a RNA-induced silencing complex (RISC). First, mature miRNA guide strand is separated from the passenger strand miRNA* and then, through binding to AGO proteins, it is loaded into a RISC complex (Rogers and Chen 2013).

The majority of miRNAs possess a 5'-Uridine which is usually sorted into AGO1, however, some miRNAs are bound by other AGOs. Ten different AGO proteins were already identified in *Arabidopsis* (Budak and Akpinar 2015; Meng et al. 2011). The fully assembled RISC then binds to its mRNA target trough sequence complementarity with its mature miRNA strand and directs mRNA silencing.

1.4.2. The miRNA-mediated silencing pathways

In plants, miRNAs were described to regulate their targets by post-trancriptional gene silencing (PTGS), either cleaving or inhibiting the translation of the target mRNA sequence (Sun 2012).

Regulation through miRNA-mediated target cleavage seems to provide a major contribution to silencing by miRNAs and it is a relatively well understood mechanism (Meng et al. 2011; Huntzinger and Izaurralde 2011). MicroRNAs can guide AGO components of RISC to the complementary mRNAs to direct target RNA cleavage. Commonly, cleavage of the target mRNA occurs in a single phosphodiester bond mostly between the 10^{th} and 11^{th} position of the miRNA (Huntzinger and Izaurralde 2011; Jones-Rhoades et al. 2006) (figure 1.V). Minor cleavages have also been reported between 9-10th and 11-12th positions (revied in Zheng et al. 2012). Extensive complementarity of the mRNA sequence to the loaded guide miRNA strand is frequently reported to be required for effective target recognition and consequent slicing in plants (Axtell 2013; Wang et al. 2015).

After cleavage, mRNAs are degraded by a ribonuclease (Huntzinger and Izaurralde 2011; Iwakawa and Tomari 2015), while the fate of the miRNAs is not clear. MicroRNA stability could be affected by numerous dynamic factors leading to its degradation but the possibility of miRNA recycling has also been proposed (Meng et al. 2011).

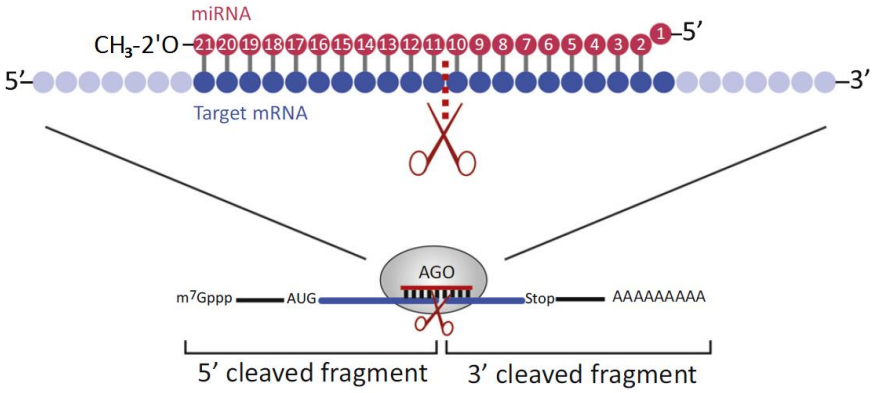


Figure 1.V The molecular mechanism behind the plant miRNAs' endonucleolytic activity. Image adapted from (Iwakawa and Tomari 2015). MicroRNAs recognize fully or nearly complementary binding sites on their targets. Usually, upon miRNA nucleotides 9–12 being engaged in Watson–Crick base pairing with their targets, the AGO cleaves the mRNA in the base-paired region; tipically between miRNA nucleotides 10 and 11. The slicing activity of the AGO resides in its *PIWI* domain. The 3′-most nucleotide of plant miRNAs is modified with a 2′O-methyl group that protects them from degradation. The miRNA 5′ terminal nucleotide is buried in the mid domain of AGOs and is not available for pairing with the target. Legend adapted from (Huntzinger and Izaurralde 2011)

The translational repression mechanism mediated by miRNAs is yet poorly understood. (Iwakawa and Tomari 2015)

One miRNA can target multiple mRNAs, which usually are members of the same gene family (Jones-Rhoades et al. 2006; Iwakawa and Tomari 2015). Some targets of the same miRNA are up-regulated while others are down-regulated (Willmann et al. 2011). Additionally, a target mRNA could be simultaneously regulated by both slicing and translational repression, however the relationship between these two modes of miRNA action or the prevalence of one process over the other is still unknown (Rogers and Chen 2013).

MicroRNAs take their action mostly in the cytoplasm. A recent study indicated that translation repression occurs in the endoplasmic reticulum (ER) (Ma et al. 2013), but in the case of mRNA cleavage, the location is still unknown. These dynamic subcellular localizations of miRNA-target mRNA might provide a mechanistic basis for the separation of translation repression and slicing (Meng et al. 2011; Rogers and Chen 2013). A few cases of nucleus-localized miRNAs and AGO proteins have been discovered (Meng et al. 2011; Rogers and Chen 2013), suggesting that it is also possible that miRNAs could target nuclear transcripts such as primary gene transcripts and the miRNA precursors in plants (Meng et al. 2011). More recently, a study in rice showed that plant miRNAs are also capable of transcriptional gene silencing (TGS) by triggering cytosine DNA methylation at both MIR and target loci (Wu L. et al. 2010).

Regulatory activities of the miRNAs are highly dynamic. They control a broad array of biological processes, including development, plant morphology, differentiation, proliferation and cell fate, and hormone signaling (Zhang, B. et al. 2005; Mallory and Vaucheret 2006; Sun 2012). They are also associated to a variety of environmental stress responses, including dehydration, mineral-nutrient stress, biotic stress (pathogens), mechanical stress or oxidative stress (Sunkar et al. 2012; Barciszewska-Pacak et al. 2015). Plant miRNAs also function in feedback regulation in small RNA pathway and in directing some siRNAs biogenesis (Chen 2009).

As regards to plant development, they have an essential role during all stages of the life cycle, operating from the first embryonic divisions up to the regulation of meiosis and gametogenesis, reproduction and genome reprogramming. Therefore the study of these regulatory RNAs may lead to a better understanding of plant embryogenesis eventually contributing to improvements in somatic embryogenesis.

1.4.3. miRNAs origin and conservation

It is known that some miRNAs are conserved among species. However, conclusions on evolutionary pathways should be drawn cautiously due to the fact that miRNAs are not extensively characterized in many plants and several reported novel miRNAs still require evidence verification. It is expected that some non-conserved miRNAs may be re-classified, as more miRNAs are being discovered. According to sequence similarity, structure configuration and shared function, miRNAs are organized in families.

Unlike animals where both mature miRNAs and their precursors are conserved, in plants only the mature sequence is conserved. Plant miRNA precursors also diverge more in structure and size. Their size may vary from 60 to 509 nucleotides, while in animals miRNA precursors are typically 70-80 nucleotides long (Zhang, B. et al. 2006a). These differences coupled with lack of sequence homology between miRNA of plants and animals, different biogenesis and mode of action suggests that miRNA families in plants and animals arose and evolved independently and convergently (Cuperus et al. 2011). However, recent findings about similarities in miRNA pathway between these two lineages and differences between evolutionary close species suggest a simpler hypothesis: the regulation by microRNAs have been inherited from a last common ancestor of plants and animals which already had a miRNA regulation system (Moran et al. 2017). None of the hypothesis could be ruled out and more studies are needed about this issue.

Through the expressed sequence tag (EST) analysis approach, Zhang, B. et al. (2006a) identified 481 miRNAs, representing 37 families in 71 plant species. They observed that not only some miRNA genes are conserved across all plant lineages, but that their targets are conserved in different plant families. These findings suggest gene regulation by miRNAs is an ancient evolutionary mechanism to control gene expression, believed to exist from more than 425 million years ago in the plant kingdom (Zhang, B. et al. 2006a).

A large number of miRNAs are species-specific or belong to closely related species which suggests that some miRNAs have been recently created and other lost (Cuperus et al. 2011). The events underlying the creation of novel miRNAs are not well understood (Budak and Akpinar 2015; Cuperus et al. 2011). However, it have been suggested that they evolve and prevail by a neutral path. Different mutations could be responsible for transcript affinity for biogenesis machinery and, at the same time, target transcripts co-evolve, forming a functional pair. Rarely when incorporated in regulatory networks it prevail in the organism (Cuperus et al. 2011).

Conservation is an indicator of miRNA function. Conserved miRNAs play important roles in conserved gene regulation such as flower and leaf development, while non-conserved or lowly conserved miRNAs may play more specific roles in specific plant species such as cotton fiber differentiation (Zhang, B. et al. 2006a). Novel miRNAs have lower abundance and fewer targets identified compared with the conserved miRNAs, suggesting that the majority of them may have no function, whereas the abundantly expressed conserved miRNAs may be the main small RNA regulators of SE (Wu X.M. et al. 2015).

1.4.4. miRNAs study in conifers

Large numbers of small RNAs have been identified by direct cloning and/or deep sequencing, with some targets being originally predicted via bioinformatics, based on either the perfect or nearly perfect sequence complementarity between a miRNA and the target mRNA, sequence conservation among different species in higher plants, or, more recently, using parallel analysis of RNA ends (PARE) (Yang et al. 2013). Currently, miRNAs have been reported for several species from the *Pinaceae* including *Picea abies; P. glauca; P. sitchensis; P. engelmanni; Pinus abies; P. banksiana; P. contorta; P. densata; P. pinaster;* and *P. taeda*, and deposited in mirBASE, mirNEST 2.0 and PMDR (Kozomara and Griffiths-Jones 2014; Szcześniak and Makałowska 2014; Zhang Z. et al. 2010). For *P. pinaster* five miRNAs were present in mirNEST 2.0 database, until now based on EST sequences, the miRNA482, miRNA839f, miRNA1255f, miRNA1314, miRNA1863. However, none of them has been experimentally validated or analyzed regarding expression in specific tissues or developmental stages.

Research in the last decade demonstrated that miRNAs have crucial roles during plant embryogenesis in both embryonic pattern formation (Seefried et al. 2014) and developmental timing (Willmann et al. 2011). Several miRNAs target transcription factors (Yang et al. 2013) and other key development regulators, influencing their spatial and temporal arrangement, and consequently different regions of the embryo require different levels of miRNAs for their patterning and cell specification (Seefried et al. 2014; Vashisht and Nodine 2014). Some miRNAs are only expressed or differentially expressed in embryogenic cells (Chen C.-J. et al. 2011; Luo et al. 2006; Li T. et al. 2012), unfortunately the role of several of them remain unknown.

During the last years, and considering that, from the technical point of view, is easier to get enough amounts of embryo tissues for expression analysis from SE than from ZE, (Vashisht and Nodine 2014) several works were published on the miRNA presence, regulation and expression patterns, during somatic embryogenesis of different plant species such as maize (Dinkova and Alejandri-Ramirez 2014), cotton (Yang et al. 2013), rice (Chen C.-J. et al. 2011; Luo et al. 2006), *Citrus* (Wu X.M. et al. 2011; Wu X. Met al. 2015), *Dimocarpus longan* (Lin & Lai 2013), *hydrid yellow poplar* (Li T. et al. 2012), or the model organism *Arabidopsis* (Willmann et al. 2011). Regarding conifer species, studies in 8 stages of embryo development during somatic embryogenesis of *Larix Kaempferi*, demonstrated the presence of 28 miRNAs validated by RT-qPCR, and 9 target transcripts validated by 5' RACE. Among the identified miRNA target pairs were the miRNA156, miRNA159, miRNA160 (targeting two ARF genes probably of ARF 10, 16 or 17), miRNA162 (targeting DCL1), miRNA164 (targeting a NAC member) miRNA166 (targeting two HD-ZipIII genes), miRNA167, miRNA168, miRNA169 (targeting NFYA), miRNA171, miRNA390 (targeting TAS3), miRNA397 (targeting laccase), miRNA398 (targeting plastocyanin) (Zhang J. et al. 2012).

Only a miRNA study was performed in conifers ZE until now, more specifically, it were identified several miRNAs in early to late zygotic embryogenesis of *Pinus taeda* and validated by RT-qPCR, including the miRNA159, miRNA 166, miRNA 167, miRNA 171, miRNA 172 (Oh et al. 2008). More studies of miRNA gene expression regulation on conifers ZE are imperative.

Few regulatory proteins are unique to gymnosperms embryogenesis which can explain the differences between gymnosperm and angiosperm morphology and development (Cairney and Pullman 2007). Additionally, the expression pattern of miRNAs observed for example, upon callus induction, between embryogenic and non-embryogenic callus, as well as during somatic embryogenesis and differentiation is dependent on the plant species (Dinkova and Alejandri-Ramirez 2014). However, given that ZE transcript profiles are highly correlated between *P. pinaster* and *A. thaliana* (Trontin et al. 2016) and that many miRNAs are highly conserved, insights gained from studies using *Arabidopsis* will probably lay the groundwork to formulate hypotheses about their analogous functions in less experimentally amenable plant systems, like conifers. (Vashisht and Nodine 2014).

1.5. Objectives

Embryogenesis is one the most important stages of the plant life-cycle, when the basic plant body plan is stablished. However, little is known about the basic mechanisms underlying embryogenesis in conifers, which directly contributes to the incapability of properly understanding and controlling somatic embryogenesis. The study of regulatory events or mechanisms underlying zygotic embryogenesis in pine, like miRNAs target regulation, will directly contribute to overcome the problems associated to somatic embryogenesis and its improvement.

Previously, in the host lab, several small RNA and mRNA libraries prepared from seeds of *P. pinaster* have been generated and sequenced (see ANNEX I) leading to the identification of several microRNA families expressed during embryo development as part of the European project ProCoGen (Procogen 2013).

This master thesis aims at characterizing a set of miRNAs selected from previously identified miRNA list, focusing on the identification and validation of putative targets involved in the regulation of *P. pinaster* embryogenesis, and the interaction between the miRNA and the predicted target. Therefore, targets were first predicted using bioinformatics tools based on PARE technology and with resource to the miRNA and mRNA libraries, and *P. pinaster* transcriptome. This was followed by a RT-qPCR analysis of the predicted targets expression in embryo.

This study, the regulation of gene expression by miRNAs during embryogenesis, will contribute to a better understanding of *P. pinaster* embryogenesis, and consequently for the improvement and valorization of this important species in the Mediterranean forests. Additionally, it will also elucidate some transversal features in SE, in conifers and other gymnosperms, and even in angiosperms.

2. Materials and Methods

2.1. Material

2.1.1. Prediction of miRNA-target pairs involved in seed development

For the prediction of the miRNAs target the following preexisting resources were used in bioinformatics analyses:

2.1.1.1. Genomic resources

- *P. pinaster* miRNA sequencing data;
- *P. pinaster* transcriptome data (unigene library, Sustain Pine DB, version 3.0, publicly available at http://www.scbi.uma.es/sustainpinedb/assemblies/28?tab=inf);
- *P. pinaster* degradome data.

2.1.1.2. Software

- PAREsnip version 2.3. (is available at Java at http://srna-workbench.cmp.uea.ac.uk/ tools/ analysis- tools/paresnip/)

- Cleaveland version 4.3,
(the code is freely available at http://sites.psu.edu/axtell/software/cleaveland4/).

2.1.2. Validation of the expression of target pairs in seed development

For the validation of the expression of the miRNAs targets it were used the following resources:

2.1.2.1. Biological material

Biological samples corresponded to *Pinus pinaster* Aiton embryos at different developmental stages which had been previously isolated from immature female cones and sampled according to the stages defined by Gonçalves et al. (2005b), (see figure 1.III). The female cones were collected in the Portuguese “Mata Nacional do Escaroupim” in “Salvaterra de Magos”.

2.1.2.2. Kits

- *RNA extraction:*

Plant/Fungi Total RNA Purification Kit, Norgen Biotek Corp, Canada.

- *RNA purification:*

TURBO DNA-*free*™ Kit, Catalog Number AM1907, Ambion®, Thermo Scientific™, Waltham, USA.

- cDNA synthesis:

Transcriptor High Fidelity cDNA Synthesis Kit - Version 8.0, Roche, Switzerland.

- Qubit RNA concentration measurement:

Qubit® RNA BR Assay Kit, Thermo Fisher Scientific, USA.

- cDNA amplification:

- GoTaq® G2 Flexi DNA Polymerase kit, Promega, USA.
- LightCycler® 480 SYBR Green I Master, version 13, Roche, Germany.

2.1.2.3. Oligonucleotides

Table 2.I List of primers used for RT-qPCR

Gene Name	SustainPineDB or GenBank accession number	Sequence (5' to 3')	Size (pb)	Amplicon size
1787PM	Unigene 1787 (middle region)	**FWD** AAAGGGAATGGGTCAGTCC	19	130
		REV GGGTTTCTTGACAGCGTC	18	
12150P3'	Unigene 12150 (3' region)	**FWD** GGAGAAGCCCAAAAGATTGA	20	217
		REV TCCCAACAAGTGAACTACAA	20	
806CP3'	Unigene 806 (3' region)	**FWD** GACATACATCAGAACAACATCC	22	134
		REV CACCTCAGACCTTTCAACC	19	
5940CPM	Unigene 5940 (middle region)	**FWD** TTTACCCAGTTTGAACAGAG	20	204
		REV TCCACAACCACAACATCC	18	
5940CP3'	Unigene 5940 (3' region)	**FWD** GTGGTTGTGGAGAATAGG	18	132
		REV CGAAATATGCCGAATCAC	18	
UBIQ[a] (Ubiquitin)	AF461687	**FWR** GATTTATTTCATTGGCAGGC	20	270
		REV AGGATCATCAGGATTTGGGT	20	
ATUB[b] (α-tubulin)	Unigene 80862	**FWD** ATCTGGAGCCGACTGTCA	18	75
		REV TGATAAGCTGTTCAGGATGGAA	22	
EF1[b] (Elongation factor-1α)	Unigene 14762	**FWR** AATGTTGCTGTTAAGGAT	18	99
		REV TATAATAACTTGAGCGGTAA	20	
HISTO3[b] (Histone 3)	BX682612	**FWD** GCTGAGGCTTACCTTGTG	18	94
		REV CCAGTTGTATATCCTTAGGCATAA	24	

a- From Sónia Gonçalves 2005 (Gonçalves et al. 2005b) ; b- Reference genes for RT-PCR from (de Vega-Bartol et al. 2013a). PM represents primers designed for amplification of the middle region of the gene; P3' represents primers designed for amplification of the 3' region of the gene (see chapter 2.2.2.3.1).

2.1.2.4. Chemicals and reagents

- Nuclease-Free Water;
- Agarose;
- TAE buffer 1x;
- 25 mM dNTPs NZYMix (MB08601);
- Loading Buffer:
- DNA Gel Loading Dye (6X), Thermo Scientific;
- DNA size marker:

GeneRuler™ DNA Ladder Mix, Range 100–10,000 pb, 5 x 50 µg, Catalog #SM0331, Thermo Scientific™, Waltham, USA.

GeneRuler™ 100bp DNA Ladder, Range 100-1,000 pb, # SM0241, Thermo Scientific™, Waltham, USA.

- Nucleic acid staining solution for agarose gel:

RedSafe™ Nucleic Acid Staining Solution.

2.1.2.5. Other Lab material

- Mortar and Pestle;
-Lightcycler® 480 multiwell plate 96, white, Roche, USA.

2.1.2.6. Equipments

- AccuTherm Microtube Shaking Incubator, 120V, Labnet international, Inc., USA.
- C1000 Thermal Cycler with Dual 48/48 Fast Reaction Module, Gradient Enabled, Bio-Rad Laboratories, Inc, USA.
- Gel Doc™ EZ Imager, with UV tray, Bio-Rad, EU.
- NanoDrop Spectrophotometer ND-1000;
- Qubit® 3.0 Fluorometer, ThermoFisher Scientific, Malaysia.
- LightCycler® 480 Instrument II, 96-well version, Roche, Switzerland.

2.1.2.7. Software

- Image Lab™ Software 5.2.1, Bio-Rad.
- LightCycler® 480 Software 1.5.1, Roche diagnostics.
- Microsoft Excel 2010.

2.2. Methods:

2.2.1. Prediction of miRNA-target pairs involved in seed development

2.2.1.1. In silico miRNA target prediction and miRNAs-target pairs selection for further characterization

MiRNA targets prediction was performed using an in-house pipeline which includes two similar bioinformatics tools, but which use different prediction algorithms, namely Cleaveland (Addo-Quaye et al. 2009) and PAREsnip (Folkes et al. 2012) software. These software require three FASTA-formatted input datasets: degradome, smallRNA sequences list (miRNAs in this case), and transcriptome database. Transcriptome is available at SustainPineDB, the *P. pinaster* gene expression database containing ESTs clustered as unigenes (Canales et al. 2014). As result for each analyzed miRNA, these programs give the predicted targets, as unigene numbers. Some characteristics of the prediction performance or results accuracy are also made available.

A PAREsnip analysis was performed as the first filtering step for miRNAs and targets selection (see figure 2.II(a)). Each degradome library was compared to the transcriptome library (mRNA) representing the same embryo developmental stage (see ANNEX II, table II.II). The parameters used were sRNA minimum abundance of 5, category value until 2, sRNA length between 18-26bp, 4 maximum mismatches, and p-value cutoff of 0.05. The miRNAs with identified targets were subsequently used as input to Cleaveland v4.3 software ("internalCleaveland01"), with p-value set for 1. The miRNAs and the respective identified targets resulting from "internalCleaveland01" analysis were compared with the results provided by LC Sciences (sequencing service provider company) using another analysis by Cleaveland, "CleavelandLC" (using Cleaveland v 4.3), with a p-value equal to 1. These two Cleaveland analyses differ in the degradome input which was differently filtered.

The resulting identified miRNAs and respective targets which were not common between the two sets of results were not taken into account for further analyses. The same applies for results from female and male cones. The proteins putatively encoded by the identified target sequences were searched in the SustainPineDB and NCBI Genbank databases. The selection of miRNA/target pairs was essentially based on whether the protein was complete in the SustainPineDB, and also on the interest of the proteins in terms of development and embryogenesis.

Novel miRNAs selection followed a different approach (see figure 2.II(b)). From the expression file with the read numbers of the sequenced smallRNAs, it were selected all the miRNAs which had a readnumber equal or above 1000 at least for one transcriptome library. These fragments were analyzed in a new Cleaveland 4.3 analysis the "internalCleaveland02". A protein search was made for the higher categories results within each degradome for these miRNAs, once again the miRNA and respective target were selected based on the interest of the protein.

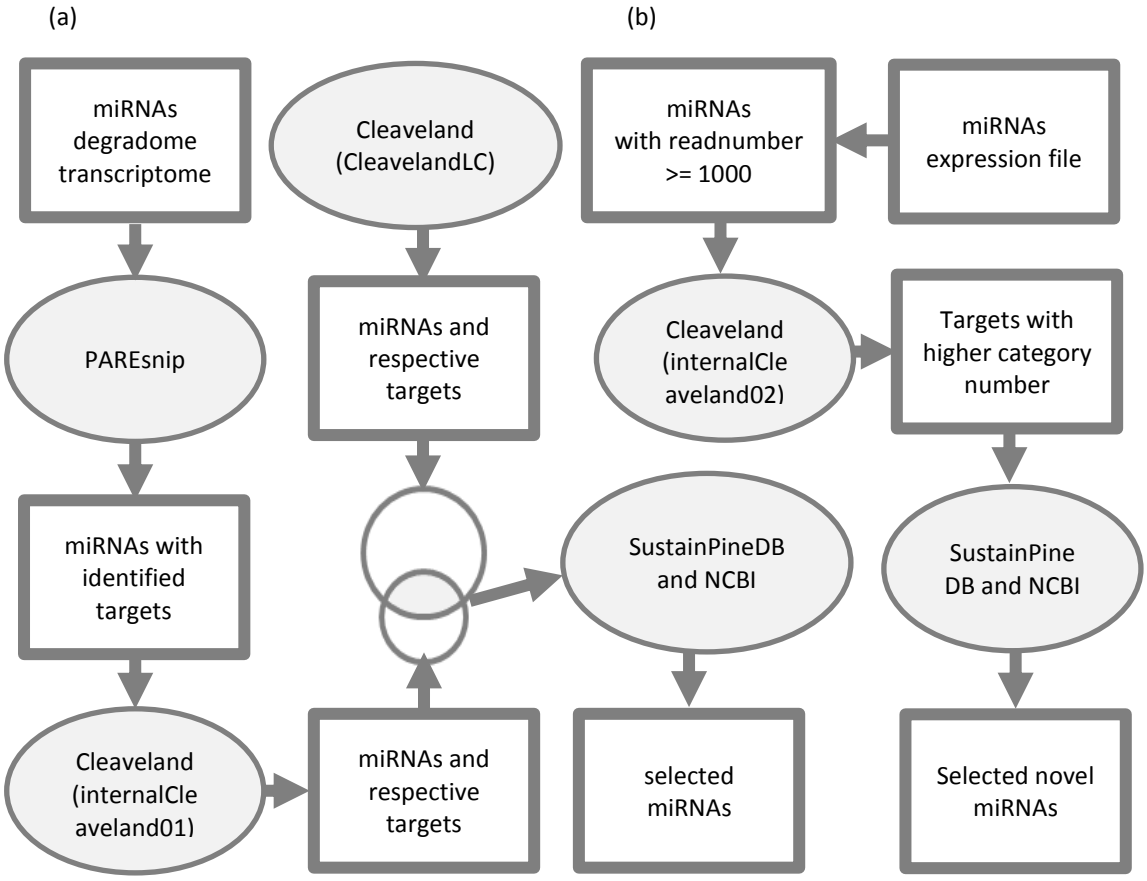


Figure 2.I Scheme of the target identification and selection of miRNAS and respective targets for further validation: (a) first analysis of the total miRNAs; (b) second analysis for novel microRNAs. *Higher category number*: For each miRNA target prediction, either PAREsnip or Cleaveland, give a "category number" from 0 to 4. Categories from 0-2 are considered to be of higher category, and consequently less probable to be random degradation products. More information can be found in (Folkes et al. 2012).

2.2.2. Validation of the expression of miRNA-target pairs in seed development

2.2.2.1. RNA extraction, purification and sample preparation

Before RNA extraction the embryos, stored at -80ºC, were grinded in liquid nitrogen, using previously cooled mortar and pestle. Pools of 9 to 60 embryos were grinded at a time, depending on the developmental stage (see ANNEX II, table II.II). Initial developmental stages required higher number of embryos, given their smaller size. RNA was extracted using the Plant/Fungi Total RNA Purification Kit following manufacturer's instructions. Several extraction reactions were performed for each developmental stage, from the five different stages of *P. pinaster* embryos.

The RNA concentration was determined by NanoDrop Spectrophotometer ND-1000 at 260nm, and purity was also checked based on 260nm/280nm and 260nm/230nm ratios.

Several RNA samples were combined to fulfill the minimum RNA quantity recommended for the DNase treatment. As a result two biological samples were obtained per developmental stage, with exception of T0/T1/T2 that only yielded one biological sample. Each RNA sample was visually checked for integrity after electrophoresis through 1% agarose gel stained with RedSafe.

Genomic DNA and other impurities were removed from the RNA samples by purification with TURBO DNA-*free*™ Kit, following the manufacturer's instructions. One reaction of 50 μl was performed for each embryo stage except for T0/T1/T2 where the total volume was 58 μl. After treatment each RNA sample was visually checked for the absence of DNA through an electrophoresis with 1% agarose gel stained with RedSafe. RNA concentration was measured by Qubit® 3.0 Fluorometer, using a Qubit® RNA BR Assay Kit.

A pool sample was prepared by addition of 500ng of RNA of each biological replicate of each stage, and 1000 ng of T0/T1/T2 that only had one biological replicate.

2.2.2.2. cDNA synthesis

Each synthesis of cDNA was performed using 500 ng of RNA using the Transcriptor High Fidelity cDNA Synthesis Kit, according to manufacturer's Standard Procedure for Quantitative RT-PCR. The summary of cDNA synthesis reactions could be found in ANNEX II, table II.III. Five hundred ng of RNA were used in each reaction. RT minus controls were included in the reactions for each stage. The quality of the cDNA synthesis was first verified with a PCR for amplification of the housekeeping *Ubiquitin* mRNA (results not showed). It was also performed an RT-qPCR for each cDNA synthesis in order to verify cDNA quality and purity, using GoTaq® G2 Flexi DNA Polymerase kit and primer for *HISTO3* gene, indispensable since this technique has a higher sensitivity.

2.2.2.3. Amplification of miRNA targets in the different embryo stages

2.2.2.3.1. Primers design

Fourteen primer pairs were designed as shown in the figure below (figure 2.III) for amplification of potential miRNA targets selected, one pair comprising the cleavage position and another to amplify the downstream region of the target.

For primer design several parameters were taken into consideration: primer length from 18 to 24 bp, GC content between 40-60%, primer melting temperature (Tm) in the range 50-60ºC, which could differ from each other 3^0C (ΔTm), similar amplicon sizes from 100 to 250 pb. It was also taken into account possible primer secondary structures that may compromise primer association to the target, following the suggestion given by PCR Primer Design Guidelines from Biosoft (http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html *accessed 15 December 2015*). In case of hairpins predicted, 3' end hairpins with a ΔG of -2 kcal/mol and internal hairpin with a ΔG of -3 kcal/mol were tolerated; for selfdimers, a 3' end self-dimer with ΔG of -5 kcal/mol and internal self-dimer with ΔG of -6 kcal/mol were tolerated;

and for possible cross dimers between the two primers (forward and reverse) 3' end cross dimers with a ΔG of -5 kcal/mol and internal cross dimers with a ΔG of -6 kcal/mol were accepted.

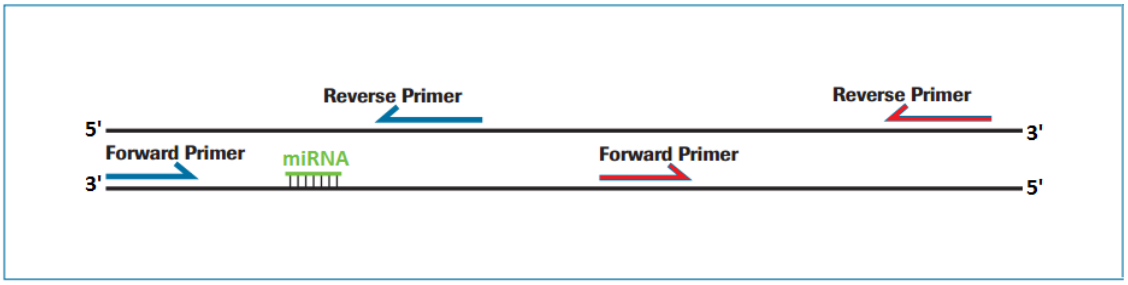


Figure 2.II Exemplification of a potential target of miRNA (mRNA). In green it is represented the miRNA:mRNA-target interaction region. For each mRNA target two pairs of primers were designed, a pair (in blue) to amplify the cleavage position region and another pair (in red) to amplify the downstream region of the mRNA target. This strategy of primers' design was adapted from a technique developed by *Oh et al.(2008)* to monitor the miRNA-directed cleavage of mRNAs called regional amplification quantitative RT-PCR (RA-PCR).

Most of the primers were designed manually and the parameters confirmed with Oligoanalizer 3.1 (http://eu.idtdna.com/calc/analyzer *accessed 16 December 2015*). However, the primers for 3' of unigene 11630, middle region of unigene 12150, and the 3' and middle region of the unigene 806 and unigene 8705, were designed by PerlPrimer Software v.1.1.21 (http://perlprimer.software.informer.com/1.1/ *accessed 18 December 201*5), due to the difficulty of finding adequate manually designed primers.

2.2.2.3.2. Optimization of primer amplification conditions

Table 2.II Optimized PCR reaction conditions common to all the selected primers.

Step	**Temperature (ºC)**	**Duration (s or min)**	
Initial denaturation	95	5'	
Denaturation	95	45''	x35 cycles
Annealing	Ta	45''	
Extension	72	15''	
Final extension	72	5'	

Primer annealing temperatures were optimized first by PCR with GoTaq® G2 Flexi DNA Polymerase for a range from 50º to 60ºC. The reactions were performed using the manufacturer instructions for a total volume of 15ul, with the following adjustments: 2uM of each primer, 2.5mM of $MgCl_2^-$, 1.65ng/µl of cDNA. The program used is described in table 2.IV. Amplification

products were qualitatively evaluated by electrophoresis in agarose gel (1%). Only the primers which had one quality amplification product were selected for RT-qPCR optimization.

The conditions for the RT-qPCR were optimized in a gradient cycler testing annealing temperatures from 55-57ºC and different primer concentrations from 1 to 4nM. Only the primers which yielded a single amplification product were selected for expression analysis. The selected primer sequences and amplicon size are in table 2.1.

2.2.2.3.3. RT-qPCR

RT-qPCR was used to determine the expression of the selected targets in different embryos stages. The five validated primers for the genes of interest (GOI), 1787PM, 12150P3', 806CP3', 5940CPM, 5940CP3', were used for amplification of cDNA from two embryo stages, namely the T4B and T7.

The RT-qPCR mix was prepared for a total volume of 20ul, with 10ul of all-in-one LightCycler SYBR Green I Master, 4µM of each primer (forward and reverse), 1ul of cDNA (dilution factor of 1:5), and H_2O. The protocol was followed as suggested by the manufacturer, with annealing temperature of 55ºC during 20seg in a ramp rate of 2.2ºC/s. To product purity verification, and distinguishing nonspecific products or primer dimers a melting curve was performed immediately after amplification, using continuous acquisition (5 acquisitions per ºC). The PCR and melting curves were performed in the LightCycler® 480 Instrument II, using 96-well plate. The Cts were determined by Second-derivation maximum method, an automatic tool from the "LightCycler480 SW 1.5.1".

For each reaction a non-template control (NTC) control was made replacing the cDNA with H_2O. Additionally, for each cDNA an RT negative control was performed, with the cDNA RT minus sample. Three technical replicates and two biological cDNA replicates were performed for each stage. Each sample and the respective calibrator were amplified in the same plate assay and using the same master mix.

2.2.2.3.3. Calculations and Statistical data analysis

Gene expression quantification method:

For analysis of gene expression quantification by RT-qPCR the most commonly used method is relative quantification. The gene expression is measured relative to the levels of a normalizer sample (reference gene), and the results are expressed as a relative abundance of the transcript, or target gene/reference gene ratio. As reference genes, the *Elongation factor-1*α*,* α*-tubulin* and *Histone 3* (see table 2.I) were chosen based on the work by Vega-Bartol et al. (2013a) that described their identification and adequacy for *P. pinaster* somatic embryogenesis.

The expression of the selected target genes was determined by RT-qPCR using the mathematical relative Pfaffl method with efficiency correction (Pfaffl 2001), which can be summarized in the equation:

$$RATIO = \frac{E_{GOI}^{\Delta Ct_{GOI}(Calibrator-Sample)}}{E_{RG}^{\Delta Ct_{RG}(Calibrator-Sample)}}$$

Equation 2.I Equation for expression ratio calculation (adapted from Pfaffl 2001). Legend: E (efficiency); GOI (gene of interest); RG (reference gene); Ct (threshold cycle).

For this method it is necessary to determine the primer efficiency and measure the Ct of the amplification qPCR curves. As calibrator sample, a pool of the eight different samples of embryo stages was used.

For the calculations an excel page was used with the equations adapted from (Hellemans et al. 2007), using the calibration factor and conversion of RQs into NRQs.

Amplification efficiency determination:

For efficiency estimation, calibration curves were performed using the dilution method. A standard curve was created from a range of 5 serial dilutions for each gene in each cDNA sample, (Bustin et al. 2009), for each GOI (1:5, 1:10, 1:20, 1:50, 1:100), and for the reference genes (1:5, 1:10, 1:50, 1:100, 1:500). The Ct means were plotted against the logarithm cDNA dilution factor. The efficiency was calculated using the equation $E=10^{(-1/SLOPE)}$ (Pfaffl 2004), being the slope estimated through the excel function "PROJ.LIN".

Statistical analyses:

The error was estimated using reported equations based on the first-order Taylor expansion (Hellemans et al. 2007) and propagated from the initial threshold cycles' measured mean to the expression ratio results. For efficiency obtained from a dilution calibration curve, the slope standard error ($SE_{(SLOPE)}$) was predicted using the excel function PROJ.LIN.

For each amplified gene a significance test was performed between the expression values obtained for T4B and T7 stages. A normal distribution of results was assumed. A parametrical t-test student with Welch's correction was performed using a free trial of GraphPad Prism 7 software (GraphPad Software, Inc., San Diego, CA, available in https://www.graphpad.com/scientific-software/prism/ *accessed Maio 2016*).

3. Results

3.1. Prediction of miRNA target genes involved in seed development

The first task of this work was to do the prediction of miRNA target genes involved in *P. pinaster* seed development and then reduce it to a list of candidate miRNA-mRNA interaction pairs with interest for further characterization. The list of miRNAs used in this work included 110 conserved and 10357 novel miRNAs, which had been previously identified by the in-house sRNA analysis pipeline "miRPursuit". The prediction of miRNAs targets was done based on *P. pinaster* degradome using bioinformatics tools. The number of miRNAs and respective targets predicted in preliminary analyses, was very high and therefore, it was necessary to consider extra criteria into account in order to reduce the list of results to a more robust and workable number of miRNAs and their respective targets. The different approaches used for prediction of target genes of conserved miRNAs *versus* novel miRNAs are herein described.

3.1.1. Target prediction of conserved miRNAs

The LC Sciences company, which sequenced the degradome libraries, had predicted an average of 5000 miRNA-mRNA pairs for each degradome library using Cleaveland pipeline (data not shown).

In order to reduce the list of predicted miRNA-mRNA pairs keeping only the high confidence results miRNA's target prediction analysis was repeated in the lab using a combination of three different predictions approaches. The detailed procedure is described in the Material and Methods section 2.2.1.1.

PAREsnip tool predicted a minimum of one target for each of the 8 conserved miRNAs and 12 novel miRNAs. These miRNAs with targets assigned were used as input for the subsequent Cleaveland analysis. Cleaveland analysis revealed one or more targets predicted for all the 8 conserved miRNAs, resulting in a total of 101 miRNA-target pairs for all the analyzed degradome libraries. No novel miRNA was present in this analysis results. The comparison between the lists of miRNA-mRNA target pairs predicted by this in-house Cleaveland analysis and the Cleaveland analysis from LC Sciences showed 51 common miRNA-mRNA pairs, corresponding to 7 conserved miRNAs.

The mRNA targets found in this final list were searched against NCBI and SustainPineDB databases to gather information concerning their putative biological roles. Four conserved miRNA-mRNA target pairs were selected based on the potential interest of the protein encoded by the target mRNA, typically associated directly with embryogenesis, like transcription factors or proteins associated to signaling pathways, or indirectly, like protective proteins. These miRNAs and respective target mRNAs are summarized in table 3.I. The miRNA-mRNA target hybrid structures predicted by Cleaveland software can be found in table 3.II.

3.1.2. Target prediction of novel miRNAs

Since no novel miRNAs were found in the previously described Cleaveland comparative analysis, it was necessary to use a different approach for novel miRNAs selection. Initially, 114 novel miRNAs were chosen based on their expression level, where only those with at least 1000 reads were selected. A new Cleaveland analysis performed with these novel miRNAs presented an average of 150 results for the reproductive cones, 300 for the megagamethophyte stages, 1700 for the zygotic embryo stages and 700 for the sample pool of zygotic embryo stages. Three novel miRNA-mRNA pairs were selected based on the availability of the complete unigene sequence and potential biological interest of the unigene in embryogenesis. The miRNA sequences and target mRNA information, and miRNA-mRNA hybrid structures can also be found in table 3.I and 3.II, respectively.

3.2. Expression of predicted targets in the *P. pinaster* embryo T4B and T7 stages

The degradome analysis is by itself an experimental validation of miRNA predicted target genes, associating cleaved mRNAs to a specific miRNA. From the degradome sequences, the bioinformatics programs used in this work give evidence of miRNA-mediated cleavage of a specific transcript. Since the degradome can be obtained from a specific tissue and/or stage of development, cleavage evidence can be found in more than one degradome, indicating that the specific miRNA is acting in one or more stages/tissues (table 3.I). This evidence may be in some cases corroborated by target and miRNA expression analysis for each of the stages through comparisons between the expression of the target gene and its regulating miRNA. Therefore, the next step of the work was to analyze the expression of the selected miRNA targets in two representative embryo stages, T4B (pre-cotyledonary embryo) and T7 (mature embryo). This type of analysis is also interesting for profiling the expression pattern of the miRNAs and targets during embryogenesis.

3.2.1. RNA extraction and cDNA synthesis

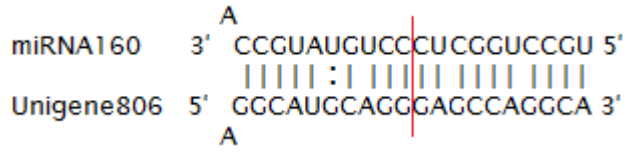
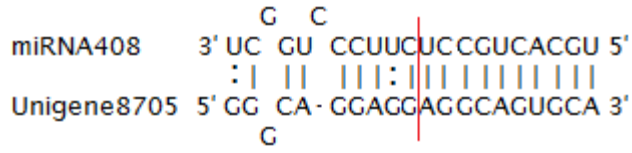
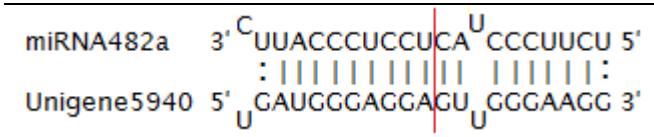
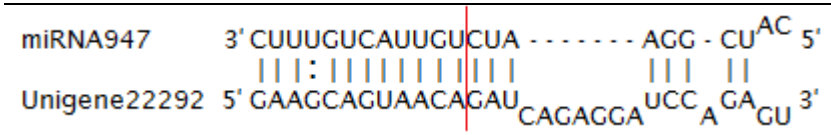
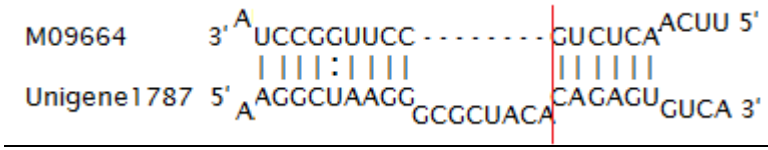
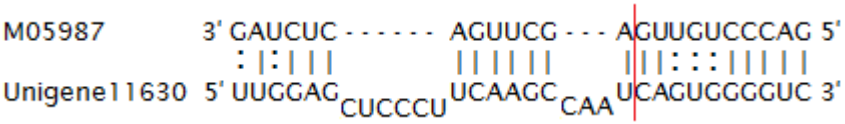
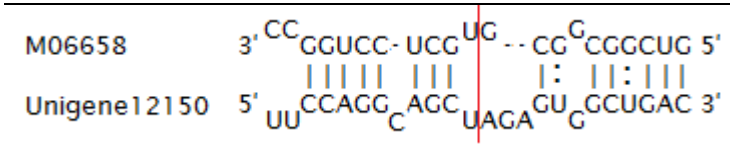
The spectrophotometric quantification of the total RNA from each extraction was done using the *NanoDrop* and it is shown in table 3.III. It is possible to observe that within the same embryo stage the yields are highly variable, which cannot be easily justified by technical issues since these extractions were performed at the same time. This variability is assumed to be intrinsic to the sample, because each extraction starts from a unique pool of embryos. Typically, the second elution samples present lower yields comparing to the first ones, as expected, being the reduction between 5.81% and 19.32%.

Table 3.I List of conserved and novel miRNAs selected, respective unigene targets and protein annotations

miRNA	miRNA sequence (5' to 3')	miRNA Length (bp)	Target unigene[a]	Target Length (bp)	Unigene annotation[a,b]	Species	Degradome
160	TGCCTGGCTCCCTGTATGCCA	21	**806**	2318	Auxin response factor (ARF)	Cycas rumphii;[a,b]	MG4B
408	TGCACTGCCTCTTCCCTGGCT	21	**8705**	815	2S albumin seed storage-like protein	Picea Glauca;[a,b]	ZET7, MG4B, MG7
482a	TCTTCCCTACTCCTCCCATTC	21	**5940**	1677	F-box family protein	Populus trichocarpa;[a]	ZET7, Pool
947	CATCGGAATCTGTTACTGTTTC	22	**22292**	1411	NAC transcription factor family (ATAF1-like protein)	Elaeis guineensis;[a] Picea mariana;[b]	ZET4B, MG4B
M09664	TTCAACTCTGCCTTGGCCTA	20	**1787**	850	Late embryogenesis abundant (LEA) protein coding ORF	Pseudotsuga menziesii;[a] Pinus tabuliformis;[b]	ZET4B
M05987	GACCCTGTTGAGCTTGACTCTAG	23	**11630**	1692	MYB transcription factor (R2R3)	Ricinus communis, Picea Glauca;[a]	Pool
M06658	GTCGGCGGCGTGCTCCTGGCC	21	**12150**	2301	Pentatricopeptide repeat-containing (PPR) protein	Arabidopsis thalia;[a]	ZET7

a - From SustainPine v3.0 annotations; b - From NCBI annotations.

Table 3.II MiRNA and respective predicted unigene target alignment, the MFE and respective unigene nucleotides start, stop and slice sites. Slice site always between the 10th and 11th position of the miRNA.

miRNA:mRNA target alignment	MFE	Start	Stop	Slice
	-45.40	945	965	956
	-34.30	476	495	486
	-34.30	880	900	891
	-35.4	109	138	121
	-32.1	365	392	383
	-35.5	946	977	968
	-40.2	1404	1427	1416

The measurement of the ratios $A_{260/280}$ and $A_{260/230}$ allows to conclude about RNA purity and the values are also present in table 3.III. A ratio $A_{260/280}$ close to ~2.00, which is observed in general for all the extraction samples, indicates a good purity of nucleic acids.

Also the majority of samples showed $A_{260/230}$ ratio around 2.00, indicating low contamination with polysaccharides. The exceptions are RNA sample T0/T1/T2 which has low ratios in almost every elution except the ext-2-I, T3/T4 ext-1-II and T4B ext-2-II. RNA samples extracted from middle and late stages of embryo of development present more consistent $A_{260/230}$ ratio around 2.00.

Samples in grey were discarded for having a lower ratio. Despite the samples T0/T1/T2 ext-2-II and T4B ext-2-II having low ratios, pointing to a slight contamination, they were acceptable to proceeded to the further step of purification.

Table 3.2III Quantification of total RNA samples extracted from *P. pinaster* embryos. Quantity and absorbance ratios measured by Nanodrop are shown.

Embryo Stage	Extraction	Elution	ng/µl elution volume (measured at 260nm)	260nm/280nm	260nm/230nm
T0/T1/T2	1	I	44.63	2.00	1.71
		II	5.19	2.03	1.01
	2	I	118.97	2.07	2.04
		II	11.52	1.93	1.60
T3/T4	1	I	168.29	2.10	2.14
		II	18.55	1.87	1.37
	2	I	414.69	2.09	2.27
		II	66.68	2.11	2.13
T4B	1	I	280.07	2.10	2.24
		II	55.87	2.08	2.10
	2	I	160.99	2.12	2.12
		II	17.69	1.96	1.65
	3	I	596.54	2.08	2.34
		II	104.7	2.09	2.22
T5	1	I	294.06	2.10	2.25
		II	31.21	2.07	1.93
	2	I	407.62	2.09	2.26
		II	58.26	2.10	2.15
T7	1	I	681.59	2.16	2.29
		II	196.07	2.13	2.26
	2	I	799.44	2.14	2.33
		II	257.4	2.12	2.35

In terms of yield the extraction method used showed low reproducibility, but in general it allows to obtain good RNA purity. Additionally, each RNA sample was run on electrophoresis gel in order to check RNA integrity and detect possible DNA contamination. Figure 3.I shows the electrophoretic separation of nucleic acids extracted. The presence of two characteristic RNA bands with no apparent degradation reflects the success of the extraction and the good quality of the RNA obtained. However, the presence of a high molecular size band, sometimes more intense than the RNA bands, indicates a considerable gDNA contamination. A DNAse treatment was performed in order to eliminate the observed gDNA and other impurities. In figure 3.II it is possible to see the result of a new electrophoretic separation of the samples after this treatment.

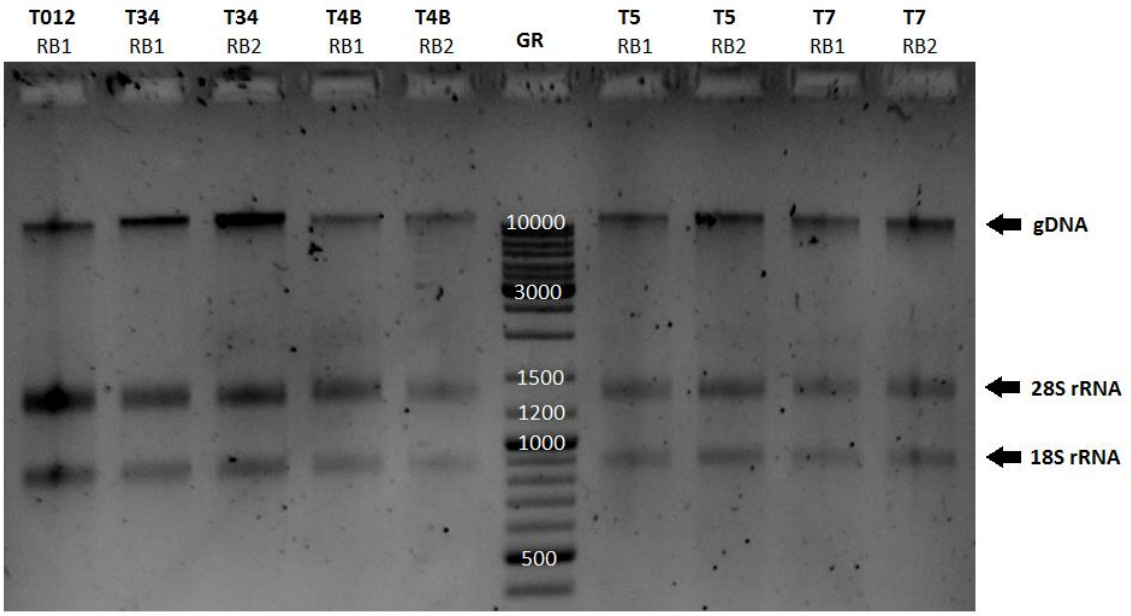


Figure 3.I Separation of the total RNA samples extracted from *P. pinaster* embryos in a 1% (w/v) agarose gel stained with RedSafe™ (0.025ul/ml). It is possible to distinguish the rRNA bands 28S (~1365pb) and 18S (~885) without degradation. A band above 100000pb indicative of gDNA contamination can also be observed. Run performed at 80V with ~200ng RNA (6:1 of loading buffer). Legend: gene ruler 100-10000 bp (GR); biological replicates 1 (RB1) and 2 (RB2); genomic DNA (gDNA); rRNA bands 28S (~1365pb) and 18S (~885).

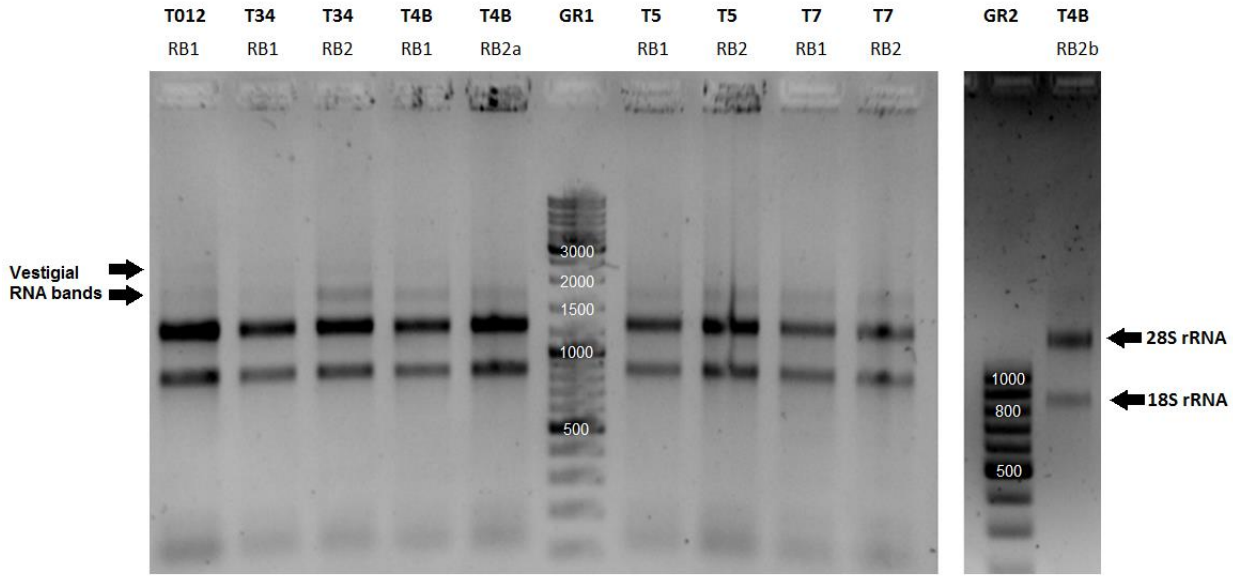


Figure 3.II Separation of total RNA samples after TURBO DNase treatment in a 1% agarose gel stained with RedSafe™ (0.025ul/ml). Run performed at 80V with ~200ng RNA (6:1 of loading buffer). No gDNA contamination is observed. Some samples present extra faded bands at ~1694 bp and ~2307 bp above the 28s (~1261bp) and 18S (~830bp) band which might represent other RNA bands. Legend: gene ruler 1 100-10000 bp (GR1) and gene ruler 2 100-1000 bp (GR2); biological replicates 1 (RB1) and 2 (RB2); rRNA bands 28S and 18S.

The band above 10000 denoting the presence of gDNA is no longer present in the electrophoresis gel run after TURBO DNAse treatment. The 18S and 28S rRNA bands are more intense in this electrophoresis gel than in the previous one, and some samples even show the presence of other RNA bands. The reason for this difference in the gels is that after eliminating gDNA contamination, the quantity of nucleic acids loaded in the gel is enriched in pure RNA.

RNA concentrations were measured again using Qubit, which is more accurate than NanoDrop for quantification of RNA, in order to rigorously quantify RNA quantities to be used in cDNA synthesis (see table 3.IV). As expected, after DNA digestion, the RNA concentrations significantly decreased.

Table 3.2IV Quantification of DNase-treated RNA samples with Qubit.

Embryo stage – Biological replicate	Quantification (ng/ul)
T0/T1/T2 – RB1	45.2
T3/T4 – RB1	90.8
T3/T4 – RB2	92.0
T4B – RB1	113
T4B – RB2	52.4
T5 – RB1	94.8
T5 – RB2	94.2
T7 – RB1	117
T7 – RB2	89.8
Pool – RB1	54.8

In this work, the availability of biological material (*P. pinaster* zygotic embryos) was limited given the seasonality of the collection periods. For that reason, it was only possible to prepare two cDNA biological replicates per developmental stage, whereas it was only possible to prepare one biological replicate of the pool RNA constituted by equal RNA amounts of both biological replicates of each of the five different embryo stages.

Amplification of HISTO3 mRNA in the different synthetized cDNA samples showed that the reaction was successfully achieved. Melting curves presented a single melting peak (Tm~78.00ºC) discarding the possibility of any amplified gene beyond HISTO3, and RT(-) showed no amplification under 45 cycles, which means that cDNA is free of gDNA contamination.

3.2.2. Validation of the expression profiles by RT-qPCR:

It is important to refer that the absence of any trace of gDNA was confirmed by the amplification of each candidate gene using as template the RT(-) cDNAs which were prepared in the same conditions as normal cDNA synthesis but where no Reverse Transcriptase enzyme was added to the mix. In such conditions none of the tests showed amplification under the 45 cycles. Additionally, for each RT-qPCR run the NTC was also clean. These were the initial criteria for PCR curve validation.

This section contains the results of the RT-qPCR analysis of five candidate targets expressed in stages T4B and T7 of the embryo development. As explained in materials and methods the Pfaffl

method with efficiency correction was applied. This method requires a measure of Ct in the qPCR amplification curves as well as the efficiencies of the primers designed to amplify the target genes.

Determination of amplification efficiency:

The primer efficiencies calculated (E ± standard error(E)) are presented in the table below. The slope (SLOPE ± standard error(slope)) values and respective graphical linear regression including the equation and some statistical values can be found in ANNEX III.

Table 3.V Primer efficiencies determined in the biological replicates and calculated according to the formula $e=10^{(-1/slope)}$. *ATUB*, *EF1* and *HISTO3* are the reference genes (RG*) and *ARF*, *LEA*, *PPR*, *F-box PM* and *F-box P3* are the genes of interest (GOI). Standard error (se).

Target gene	cDNA	E ± se(E)	Target gene	cDNA	E ± SE'(E)
	T4B_RB1	72.96±0.98		T4B_RB1	77.73±1.43
	T4B_RB2	73.74±2.23		T4B_RB2	75.84±0.68
ATUB*	T7_RB1	69.69±1.64	***HISTO3****	T7_RB1	66.68±1.04
	T7RB2	73.08±1.72		T7RB2	70.15±1.45
	POOL	74.59±0.82		POOL	68.80±1.06
	T4B_RB1	91.77±1.12		T4B_RB1	86.23±1.56
	T4B_RB2	89.37±1.17		T4B_RB2	81.32±1.11
EF1*	T7_RB1	92.55±4.58	***ARF***	T7_RB1	105.46±6.33
	T7RB2	88.41±1.78		T7RB2	114.89±5.83
	POOL	81.64±1.45		POOL	94.01±1.75
	T4B_RB1	114.54±3.44		T4B_RB1	88.57±2.63
	T4B_RB2	79.83±4.35		T4B_RB2	109.18±7.90
LEA	T7_RB1	98.27±4.74	***F-box PM***	T7_RB1	81.02±5.22
	T7RB2	97.86±1.46		T7RB2	80.47±5.52
	POOL	93.49±1.74		POOL	117.12±9.25
	T4B_RB1	101.17±2.48		T4B_RB1	88.45±2.60
	T4B_RB2	96.84±7.40		T4B_RB2	92.71±3.59
PPR	T7_RB1	78.47±7.20	***F-box P3***	T7_RB1	130.73±6.20
	T7RB2	113.95±10.02		T7RB2	81.75±4.23
	POOL	99.31±2.08		POOL	99.66±2.94

In general, there are evident variations in amplification efficiencies between the different embryo stages for each amplified gene, independent of the cDNA sample. These variations are higher for the GOI genes which also present the higher standard error.

The reference genes *ATUB* and *HISTO3* had the lowest amplification efficiencies (~72%). However, when compared to the literature the amplification efficiencies of these two genes are in agreement with the reported efficiencies (73% for ATUB and 75% for HISTO3). (de Vega-Bartol, Santos, et al. 2013) Only *EF1* has an average efficiency (~82%) above the described (76%), and this difference may due to the fact that the reported study was performed with somatic embryos and not with zygotic embryos as in this study.

The GOI showed higher efficiencies, sometimes above 100%, an overestimation commonly associated to this efficiency calculation method (Pfaffl 2004). A primer efficiency of 100% means that the number of template DNA sequences will double after each amplification cycle, which usually does not happen. The ideal range for efficiencies lay between 90%-110%. (Hellemans et al. 2007) In this work, only 44% of the target gene efficiency values meet these requirements. For some primer with efficiencies lower than 90% the amplification curves were repeated and the obtained values were similar. Due to the limited amount of biological material for testing additional qPCR conditions, the efficiencies used were those represented in the tables.

Pfaffl method assumes that the E_{SAMPLE} is the same of the $E_{CALIBRATOR}$, so the efficiency chosen for the fold change calculation could be either the amplification efficiency obtained for each sample or the efficiency obtained for the pool sample, representative of all stages. (Pfaffl 2004) Although the amplification efficiency of each gene in the pool sample was similar to the average efficiency in the samples of each embryo stage, except for *EF1* and *F-Box(PM)*, the amplification efficiencies of the samples were different from each other and from the calibrator. In order to avoid assumption errors, the PCR efficiency estimated for each sample was used.

Relative expression of target genes:

The relative expression of the candidate target genes in the two embryo stages T4B and T7 is shown next (see figure 3.III). In general, it is evident that each gene shows higher expression in the T4B stage when compared to the T7 stage.

For the F-box gene, two primer pairs were used, one that anneals in the middle region of the gene where miRNA was predicted to bind, and another one that anneals downstream of the binding region. It would be expected that the fragment of the middle region had less expression than the 3' fragment due to the lack of cleaved fragments amplification. However, no significant difference was observed between the expressions levels of these two fragments.

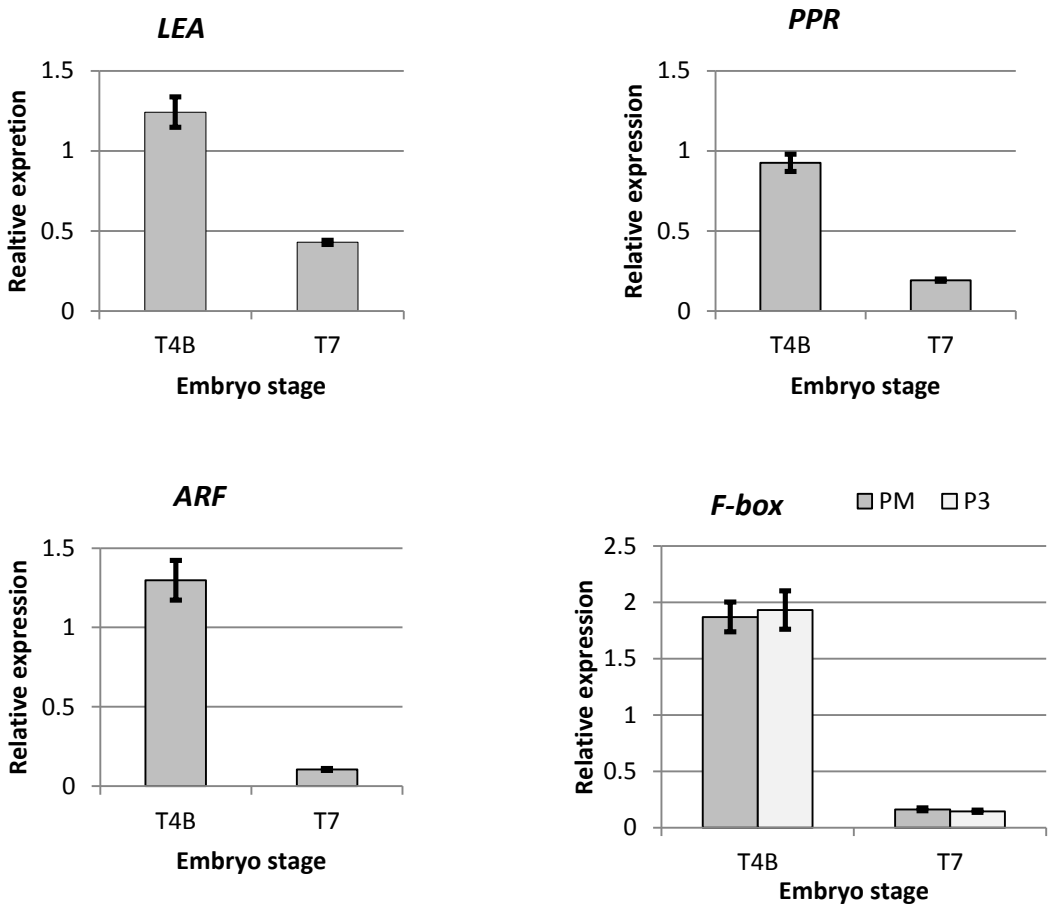


Figure 3.III Relative expression of *LEA, PPR, ARF* and *F-box* genes in the early cotyledonary embryo stage (T4B) and the mature embryo stage (T7). The columns represent the relative expression and consist in the fold change values calculated by the Pfaffl method with efficiency corrections. The results are normalized to the cDNA of the pool sample, being set to the value of 1. The black bars represent the standard error of the mean (SEM) of the technical and biological replicates. For the first 3 genes a single amplicon of the unigene was quantified; the numerical results for each gene are (T4B: 1.24±0.095; T7: 0.43±0.013) for *LEA*, (T4B: 0.92±0.054; T7: 0.19±0.007) for *PPR*, and (T4B: 1.30±0.125; T7: 0.11±0.006) for *ARF*. For the *F-box*, two amplicons were independently quantified, one in the potential miRNA binding region (PM) (T4B: 1.87±0.132; T7:0.16±0.013) and the other in the downstream region (P3) (T4B:1.93±0.172; T7: 0.15±0.010). P-value between expression in T4B and T7 cDNAs: 0.0696 (*LEA*); 0.0429 (*PPR*); 0.0661 (*ARF*); 0.0471 (*F-box PM*); 0.1198 (*F-box P3'*).

4. Discussion

4.1. Prediction of miRNA target genes involved in seed development

Given the critical roles of plant miRNAs in gene regulation, efforts to identify miRNAs and respective targets are imperative. Although the first miRNAs were discovered by genetic screening approaches, experimental approaches were limited by their low efficiency, time consuming, and high cost (Zhang et. al 2006b). Additionally, target prediction and validation were also based on expensive experimental approaches (Thomson et al. 2011). As alternative, in the last decade, several high throughput technologies and bioinformatics programs have been developed for miRNA discovery and identification of their target genes.

4.1.1. miRNA target prediction and selection

The existent bioinformatics tools used for miRNA targets identification are mainly based on sequence complementarity between the miRNA and target mRNA and the stability of the interaction, since in plants a high degree of complementarity exists between the miRNA and its target. There are several computer software programs publicly available for predicting plant miRNAs targets, such as RNAhybrid (Krüger and Rehmsmeier 2006); Targetfinder (Fahlgren et al. 2007); TAPIR (Bonnet et al. 2010); Target-align (Xie and Zhang 2010); miRU (upgraded to psRNAtarget) (Zhang, Y. 2005; Dai & Zhao 2011); p-TAREF (Jha and Shankar 2011); C-Mii (Numnark et al. 2012); UEA-sRNA (Stocks et al. 2012); starBASE (Li J. H. et al. 2014). These programs are efficient in searching all possible target candidates, resulting in a huge amount of possible targets. However, these predictions include a high proportion of false-positives.

The PARE technology in association with bioinformatics software tools for target prediction makes target identification more reliable, as in this case target prediction is accomplished using degradome data, which is a collection of sequenced 5'-ends of uncapped mRNAs. These fragments are considered as the products of mRNA cleavage guided by miRNAs (Dai et al. 2011). This improves the robustness of computational analysis by pointing out potential miRNA cleavage sites.

There are at least five bioinformatics tools for plant target prediction based on degradome. Cleaveland was one of the first and most cited target predictor software (Addo-Quaye et al. 2009). The original version has however strict rules concerning the pairing between the miRNA and its target. It allows only 4 or less mismatches and less in the seed region, and has more restrictive rules in the 10-11th positions (from 5' of miRNA), whereas slicing occurs, assuming positive correlation between canonical complementarity seed region and probability of actual cleavage. Experimental validation of miRNA-mRNA target duplexes with mismatches or G:U pairs at the 10-11th positions and also with more than 4 mismatches have been reported (Zheng et al. 2012) additionally, mismatches at canonical seed region positions, like a bulge, have also been reported in plants (Brousse et al. 2014) suggesting that cleavage of potential targets can occur even with poor complementary in the seed region. Based on these studies, programs with too stringent criteria may be omitting many genuine targets. Some authors mentioned the advantage of giving more emphasis to the alignment MFE than to the mismatch and gap penalties align score (Brousse et al. 2014). Therefore, more recent versions

of Cleaveland allow more extended criteria, such as version 4.3, (first cited in Brousse et al. 2014). This version organizes the predicted pairs according to Allen score or MFE (set by the user), been only limited by the p-value (also settled by the user), which consequently allows more mismatches and bulges between the miRNA and target. The main disadvantage of using this program is that the underlying algorithm makes the analysis very slow, which could only be overcome using parallelization across multiple machines.

PAREsnip uses a different algorithm for target predictions which allows it to be faster than Cleaveland. The algorithm is based on alignment prohibitions between the miRNA and potential target, therefore the program shares similar restrictions to the original version of Cleaveland. Additionally, it gives to the user the chance to customize certain parameters such as mismatches in the 11th position, the number of total mismatches allowed and p-value (Folkes et al. 2012).

SeqTar is another program which innovates by broadening and relaxing the complementarity criteria, but the speed of the prediction was not improved (Zheng et al. 2012).

However, these three programs focus exclusively on the annotated portion of the genome, utilizing cDNA sets (transcripts) as their inputs. Two other programs are available to extend this prediction to the whole genome (annotated and unannotated): sPARTA and StarScan. SPARTA accepts multiple species at the same time (additional program “comPARE”) and is fast due to true parallel computing (Kakrana et al. 2014). StarScan can be used for both plants and animals, but also for other small RNA classes such as piRNAs and endo-siRNAs targets prediction (Liu, S. et al. 2015).

The aim of this master thesis was to know more about the regulation performed by miRNAs during *P. pinaster* seed development therefore, the first thing to do was to select some potentially interesting miRNA-mRNA pairs. From the five programs available and already referred, Cleaveland4 and PAREsnip seemed to be the best choice because they give a more restrictive list of results. From what is published, the majority of cleavage events seem to occur between 10-11th position (Zheng et al. 2012) which is the slicing site assumed by these programs. Even though Cleaveland analysis takes a long time, it is the tool with more predicted targets, and the one which is integrated in “miRPursuit”, the in-house sRNA analysis pipeline.

Due to the large number of conserved miRNA:mRNA pair results provided by the sequencing company (LC Sciences) using Cleaveland, in this work we devised an alternative strategy to shorten this list of results and keeping only the high confidence ones by combining three types of analyses after different steps of data processing. Two of such analyses were performed with Cleaveland program and one with PAREsnip.

Because of their different algorithms, different p-value calculation system, and different customizable parameters, different sets of results were consequently obtained between Cleaveland and PAREsnip analysis. For instance, it was observed that Cleaveland retrieved more and different results than PAREsnip. This can be explained by the loosen permissions of Cleaveland, such as mismatch at position 10 (from 5' of miRNA), multiple gaps and more than 2.5 mismatches or adjacent mismatches within the seed region (first 12-nt from the 5'-end of miRNA), which PAREsnip does not allow (Folkes et al. 2012). Their different limitations are the reason why it was chosen to work with both tools consecutively, keeping only the common results given by both analyses.

It was chosen to perform the first analysis with PAREsnip because it has more restrictive parameters. It were not predicted any target for novel miRNAs list inputted, by "internalcleaveland01". Two factors may be beyond this result: First it were used tight parameters in PAREsnip analysis, (specially the category cutoff of 2 and the p-value cutoff of 0.5), which reduced the huge list of miRNAs to 8 conserved miRNAs and 12 novel miRNA; And, "internalcleaveland01", in spite of the stretched parameters, (respectively category cutoff of 4 and p-value of 1), it has different allowed parameters from PAREsnip (described above). Secondly it was difficult to predict targets for novel miRNAs than for that conserved ones. This may be to the fact they can obey target interaction rules different from what are known (Axtell 2008).

The comparison of “internalcleaveland01” results with the “CleavelandLC” results was performed in order to reduce even more the predicted target results.

Since no candidate novel miRNAs were found using the approach described above, it was necessary to do another novel miRNA target prediction. In this case, the novel miRNAs were first selected based on the sequenced reads number, which downsized miRNAs input and consequently reduced the time taken by the program to deliver the results. The shorter results list made it unnecessary to perform a second computational analysis. So for non-conserved miRNAs target prediction only the Cleaveland software was used. It is important to refer that these solitary *in situ* analysis give confident results since they are based on degradome data and the inputted miRNA sequences were pre-selected based on expression (reads number). The use of multiple analysis in the first miRNA target analysis was only an approach to splay the results.

The main disadvantage of using the discussed target prediction methods is that they only predict miRNA targets which undergo endonucleotide mRNA cleavage, ignoring miRNAs which act through translational repression. Fortunately, mRNA cleavage is the major miRNA mechanism in plants. However, if the goal is to predict targets without this limitation one should use bioinformatics tools that do not take into account degradome data, as those previously mentioned in this section (4.1.1), first paragraph.

Due to the considerable number of potential targets predicted and the application of enlarged criteria, the resulting miRNA-mRNA target pairs were manually inspected, for both of the analysis, by checking not only the complementarity between miRNA and its target, but also other criteria such as binding site evolutionary conservation and multiplicity of target sites (Dai et al. 2011). In the next sections the analysis of these parameters for each miRNA-mRNA target selected is discussed.

Target site accessibility for the miRNA is also an important feature which contributes for the for the target validation. But it will not take in consideration due to the fact that *in vivo* there are proteins and other binding motifs which could influence the mRNA secondary structure, and which at present are very hard to accurately predict *a priori*.

Complementarity between miRNA and predicted target

The alignments found between the conserved miRNAs 160a, 408 and 482a and their respective targets are in accordance with the canonical restrictive principles of high complementary (Rhoades et al. 2002). They have all less than 4 mismatches (including G:U) and no mismatch in the

10-11th positions, where the program predicted the cleavage. MiRNA 160a and 408 present mismatches in the 3' end of the miRNA, the less critical site for a mismatch when compared to the central or 5' miRNA region (reviewd in Liu Q. et al. 2014). In what regards the miR482a, in addition to the mismatch in the 3' end, it presents two mismatches in the seed region, one in the 8th position and one G:U in the 1st position. Although evidences had shown that mismatches in the 5' and center regions of the miRNA, but not in the 3' region, are critical for miRNA function (Parizotto et al. 2004), a recent article showed that 5' region of the miRNA does not have to be strictly homologous to the target (Brousse et al. 2014). The authors discovered that miRNA398, which cleaves its target *BCBP* at the 10-11th position, as confirmed by 5'RACE PCR analysis, presents a bulge of six nucleotides in the mRNA region opposite to the miRNA 5'region (6-7th positions of the miRNA) (Brousse et al. 2014). They tried to identify more miRNAs in *Arabidopsis* with similar characteristics but could not find any (Brousse et al. 2014). Such possibility may also exist in *P. pinaster*, as exemplified here by the conifer specific miRNA 947 which, according to the degradome analysis, pairs with a bulge of 7-nt in the 5'-end of the miRNA, between the 7-8th positions of the miRNA, representing another case of low complementary in the 5' end of the miRNA.

Novel miRNAs M09664 and M05987 show a bulge in the middle region of the miRNA, which has not yet been reported in plants. The M06658 has the most critical lack of complementarity, with mismatches in the 3'region (20th and 21st positions) and in the 5'region (7th position), small bulge of 1-nt between the 14-15th positions, and a 4-nt bulge in the central seed region (9th-12th positions).

Compared to conserved miRNAs, the novel ones show less complementary to the predicted targets. Some hypotheses have been proposed regarding the origin of miRNAs and their targets (Budak and Akpinar 2015; Cuperus et al. 2011), including the hypothesis that the miRNA and its target could arise independently (Fahlgren et al. 2007; Felippes et al. 2008). It have been proposed, for instance, the existence of a pool of transient individual miRNAs with no biological significance which can evolve to target a specific mRNA gene, thought the accumulation of mutations (Axtell 2008). The tendency for low expression, poor conservation, imprecise biogenic processing, and the existence of few, if any, targets found for newly born miRNAs corroborates this hypothesis (Axtell 2008; Cuperus et al. 2011). The novel miRNAs in this study may be in a period of evolving along with their targets towards a better interaction, which justifies the less complementarity between them. However, until a detailed phylogenetic analysis is performed with the miRNAs and the respective target sequences under study, the relative position of novel miRNAs in terms of evolution remains to be determined. If they are found also in distantly related species (in term of evolution) another hypothesis could be considered. The mispairing could be justified by an evolutionary tendency of the miRNA to be less effective in targeting the mRNA until it stops being functional through mutational drift or negative selection (Axtell and Bowman 2008). The veracity of this hypothesis is very difficult to demonstrate since it only can be supported by the absence of targets. Another possibility is in the case of miRNAs cooperation to target the same mRNA, where is possible that some of them have less complementary to the target.

Despite the majority of miRNA-mRNA target pairs experimentally validated in plants until now show a very high complementarity, this does not necessarily mean that plant miRNAs cannot

recognize target sites with several mismatches. Moreover, recent studies have proved that there are cleavage pairs with less complementarity. It should not be ruled out that some plant miRNAs, like in animals, may not need perfect complementarity to regulate their targets. However, because target sites with near perfect complementarity do exist, these have been extensively studied and little effort has been made to explore the potential targets with several mismatches (Dalmay 2013).

miRNA target conservation

Some miRNAs families are well conserved between species who share the same ancestor, also some targets are found to be conserved among close phylogenetic species for the same miRNA (Axtell and Bartel 2005). The conservation of miRNAs and targets contributes respectively for the confidence of the predicted conserved miRNAs and their target results. In this section it will be explored the conservation of conserved miRNA among conifers and if the respective predicted target has already been identified for any miRNA.

For miRNA160a a target was identified in the *P. pinaster* transcriptome with high similarity to *ARF* (Auxin Response Factor – Transcription Factor) gene identified for *Cycas rumphii* (NCBI: FN433183.1) which could be *ARF10*, *ARF16* or *ARF17*. The *ARF10*, *ARF16* and *ARF17* target genes have been identified as targets for miRNA160 in different species (Sun 2012), such as *Arabidopsis thaliana*, (Rhoades et al. 2002) or the close phylogenetic species *P. taeda* (Lu et al. 2007) showing an evolutionary conservation of the binding site. This bioinformatics analysis is in agreement with published results that show *ARFs* to be post-transcriptionally regulated by miRNA160 and, therefore its function seems to be conserved in *P. pinaster* too.

The remaining selection of miRNA-mRNA regulatory pairs presented in this thesis is new in the light of what has been published so far in plant species, which could be due to the fact that there are still few validation studies exploring the existence of new miRNAs targets beyond those documented and conserved ones. Moreover, this may also reflect the novelty of the work in terms of the identification of *P. pinaster*-specific miRNAs and respective targets.

The miRNA408 target genes known so far include genes encoding a copper ion binding proteins, a laccase and a plantacyanin (Jones-Rhoades et al. 2006; Sun 2012). Other species-specific targets were also identified (Zhang B. et al 2006b; Archak and Nagaraju 2007). However, in this study it was identified an mRNA with high similarity to a 2S-ASP (2S albumin storage protein) encoding transcript of *Picea Glauca* (GeneBank: AF074939.1). This was the first time a transcript encoding a storage protein was indicated as miRNA 408 target. Possibly, these less conserved miR408 targets, such as the *2S-ASP* transcript in *P. pinaster*, are specific of the species they were detected in.

For the first time, an *F-box* gene was identified as potential target of miRNA 482a although, it has been identified before as target for other miRNAs (Sun 2012). Until now, the list of documented miRNA482a targets included: cytochrome p450 encoding transcript, disease resistance proteins encoding transcripts (Sun 2012) and histone deacetylase genes, the last one identified in *P. densata* (Wan et al. 2012).

Currently, miRNA 947 seems to be a conifer-specific miRNA family because it has been only identified in few conifer species, such as *P. densata* (Wan et al. 2012) and *P. taeda* (Quinn et al.

2014). In this work, the mRNA coding for ATAF1 like protein, similar to *Picea mariana* mRNA sequence (GeneBank: AF051222.1), an element of the transcription factor family NAC, was identified as target of miR947. NAC transcription factors encoding transcripts have been already identified as targets of few miRNAs: miRNA 164, miRNA1218, miRNA1223 and miRNA1514, in different angiosperms species (Martin et. al. 2012; Sun 2012), but this is the first evidence in gymnosperms.

For the novel miRNAs target predictions, genes encoding R2R3 MYB domain proteins and *PPR* gene have already been identified as targets of conserved miRNAs. For example, miRNA159/319 regulate *Myb33 and Myb65* and, *PPR* is regulated by a variety of different miRNAs 161, 414, 474, 475, 476 and 1049, different for each species (Jones-Rhoades et al. 2006; Sun 2012).

From the list of selected miRNA targets in this work only *LEA* protein encoding transcript has not been previously identified as miRNA target.

Although it has been suggested that novel miRNAs target mostly species-specific transcript sequences (Zhang et al. 2006a), the miRNAs currently being annotated as novel miRNAs may be found to be conserved as more studies in this area are being published. Not only the identification of miRNAs and their targets in different plant species is important for the possible reclassification of current novel miRNAs as conserved ones, but also it will contribute to the improvement of criteria for establishment of parameters towards the identification of further miRNAs.

4.2. Validation of the expression of miRNA-target pairs in seed development

MiRTarBase has collected over 360,000 miRNA-target interactions from the literature, including data from plants (rice and *Arabidopsis*) and animals, and the statistics of miRNA-target interaction data collected so far reveal that most miRNA-target interactions are validated by NGS, then by microarrays, luciferase assay, western blot and finally pSILAC (Chou et al. 2016). Clearly, plant species are underrepresented in these statistics but it seems clear from the Arabidopsis data that NGS has been given the highest contribution to the validation of miRNA-target interactions also in plants (Chou et al. 2016).

Degradome sequencing is by itself a high-throughput experimental validation method which successfully identifies potential targets cleaved by miRNAs. However, its high sensitivity may introduce some false-positive results into bioinformatics analysis output (Ding et al. 2012; Thomson et al. 2011). For that reason, it is recommended that such method be complemented by other experimental techniques.

One of the goals of this work was to validate some miRNAs-mRNA interaction pairs in different stages of the embryo development in order to understand their possible roles. Three techniques can be used to validate individual miRNA:mRNA target interactions: Western-blot, RT-qPCR or luciferase reporter assay (Ding et al. 2012; Thomson et al. 2011). Western-blot is a technique that analyses the effect on target gene expression by measures at protein level, advisable in cases of PTGS by inhibition of mRNA translation (Kuhn et al. 2008). Luciferase reporter assay reveals miRNA/mRNA target direct interaction, but is a very labor intensive technique, with complex protocols that are both time and resource consuming (Thomson et al. 2011). Since the targets selected here for further study are the ones validated by PARE technique at mRNA level, it was decided, in this case, to determine

the expression profile of the miRNA targets by RT-qPCR analysis as a validation of the expression level obtained by sequencing and as a possible additional evidence of miRNA-target interaction. RT-qPCR requires low amounts of starting material due to its high sensibility and it is faster, less labor, and relatively cheaper than the two techniques described above (Martinez-Sanchez and Murphy 2013).

The mRNA targets expression was analyzed in the same embryo stages as the ones used for degradome sequencing, namely the early cotyledonary embryo (T4B) and the mature embryo (T7). During the embryogenesis there are a different expression pattern of the genes (Gonçalves et al. 2005b). The T4B stage precedes one of the main differentiation events during embryo development (cotyledon formation), and it is expected that during this stage the genes related to cellular division, growth and energy consumption are highly expressed to support cellular differentiation and growth. In the T7 developmental stage the embryo completes the synthesis of storage materials, undergoes hormonal changes, so it is expected to express new sets of genes to prepare embryo for dormancy and desiccation. The timing expression of the miRNAs which are responsible of gene regulation are also important to embryonic development (Martin et. al. 2012) however, their expression were not study in this work.

From the seven initially selected miRNA-mRNA pairs, only four passed all the optimization steps necessary to proceed with RT-qPCR, mainly due to the fact that the other primers amplified unspecific products despite the tentative improvement of PCR reactions.

The RT-qPCR analysis revealed that the four genes have increased expression in the T4B embryo stage in comparison to the T7 mature stage which might be associated to regulation mediated by miRNA cleavage in the T7 stage. The relative expression results are an average obtained from two biological replicates, which in some cases have differences between them, as it can be seen in ANNEX V. *ARF* and *PPR* encoding transcripts have the most concordant biological replicates (Annex V), for that reason the respective miRNA-mRNA target pairs are suggested to be the focus of further characterization. More biological replicates should be performed, in order to be possible to discard the more distant ones. However, since the biologic resources were limited, it was not possible.

Additionally, the p-value calculated for the difference between the expression of the two embryonic stages for each of the mRNA encoding genes revealed that only the study genes *PPR* and *F-box* have a significant difference between the expression of the two tissues (they show a p-value inferior to 0.05). The 0.05 value is the classic limit imposed to p-value values. *ARF* and *LEA* genes have a p-value close to 0.05, respectively 0.0661 and 0.0696, they are considered almost significant.

ARF

ARF proteins have a key role as transcription factors elements in the auxin signaling. They are mediators of auxin responses and can activate or repress gene expression through binding to auxin response elements in mRNA. Thus, they are indispensable for embryogenesis. All the embryonic cells express at least one *ARF* gene or unique combinations of *ARF* genes. These expression patterns are dynamic during development (Zažímalová et. al. 2014). The *ARF 10*, *ARF 16* and *ARF17* belong to the same clade in terms of evolution, and there are evidences that these genes are regulated by

miR160 in the last common ancestor of the extant land plants (Finet et al. 2013). Unlike angiosperms, where all *ARFs* at different embryonic development stages have been mapped, (Rademacher et al. 2011) little is known in gymnosperms. The identification of this potential *ARF* in the embryo T4B and T7 stages contributes to the confirmation that this *ARF* clade is important to embryogenesis also in *Pinus*. The study of a putative *P. pinaster ARF16* (unigene 2451) shows a decrease in expression from early to pre-cotyledonary stage, and an increase from early cotyledonary to mature embryos, consistent to *Arabidopsis* expression pattern (de Vega-Bartol, et al. 2013b). However, in this study the expression results show the opposite trend, mature embryo having less expression of the putative ARF clade than early cotyledonary embryo. The reason may be that the *ARF* gene in study is different from the *ARF16* gene studied by de Vega-Bartol, suggesting that in spite of belong to the same clade they could be different expressed.

F-box

There are a large number of F-box proteins, and the functions of most of them have not yet been defined. All these proteins have in common an *F-box* motif required for protein-protein interaction, which is generally found in the amino-terminal half of proteins and is often coupled with other motifs in the carboxyl-terminal part of the protein (Kipreos and Pagano 2000). As a family, these proteins are essential for plant growth and development since they may participate in cell cycle control, hormonal signal transduction and some of them have been identified as regulators of protein ubiquitination and degradation (Kuroda et al. 2002). The transcript analyzed in this study which putatively codes for the F-box family protein is more expressed in the T4B stage, where the major synthesis and cell cycle reactions occur, than in the mature stage T7. When comparing the expression of the fragment of the cleavage region with the expression of the 3' fragment no significant difference is observed. This may be due to the presence of additional regulation mechanisms or to the presence of miRNA and its target in different cells of the embryo.

LEA

Late Embryogenesis Abundant (LEA) proteins are a group of hydrophilic proteins with characteristic amino acid composition. They are mainly expressed in dry seeds and, as the name suggests, mainly accumulated during late embryo development stage when desiccation occurs, despite existing in several tissues (Amara et al. 2014). They have been proposed to have various functions, including protection of cellular structures from the effects of water loss and desiccation, putatively through safeguarding enzymatic function and prevention of aggregation in times of dehydration/heat due to the ability to remain soluble, protection of proteins from stress-induced damage, sequestration of ions, and folding of denatured proteins, they can also act as chaperone proteins to resist cellular damage (Amara et al. 2014). Consequently, LEA proteins play crucial roles during different cellular processes, like embryonic development for instance, and in a tissue or time specific manner (Olvera-Carrillo et al. 2011).

Considerably few *LEA* genes have been identified in gymnosperms, mostly due to limited genome information. However, expression profiles of some proteins of this family have been already

studied in embryo stages of *Pinus* species. For example there was observed an increased expression of *LEA* 4 group from early embryogenesis to late cotyledonary phase in ZE of *Pinus taeda* and *Pinus oocarpa* (Lara-Chavez et al. 2012). Other studies in *P. pinaster* have been reported (Dubos et. al. 2003; Gonçalves et al. 2005b). Our expression results show the opposite from what was expected for LEA proteins in the light of what is known so far, and contradict previous studies in *P. pinaster* that demonstrate there is an increase of *LEA* transcripts expression from middle to late stages of embryo development (T4 to T7) (Gonçalves et al. 2005b). There is no similarity between the studied mRNA sequence and the sequence studied by Gonçalves et. al. (GenBank: AJ297302), confirmed through blast searches either in NCBI or in SustainPineDB, and through MultAlin version 5.4.1 an identity of 52.69% was obtained. Thus, the transcript studied here does not seem to code for the same protein, and therefore it may have a somehow different function consistent with a role in the T4B stage of embryo development. However, unigene 1787 has high similarity to *Pinus tabuliformis* LEA4-2 protein (GenBank accession: KM521229.1; Cover: 34%; Identity: 99%), which its expression were not detected in any of the studied organs (root, phloem, bud, needle or 2 weeks-old germinates or 2 weeks-old germinates) (Gao and Lan 2016). This may be because they are expressed at levels non detectable or because there are expressed in tissues not studied, like the embryo. The expression of this gene in *P. pinaster* embryo points to the possibility of also been present in *P. tabuliformis* embryo. It could be an embryo-specific gene.

PPR

Pentatricopeptide repeat (PPR) proteins are a large family of modular RNA-binding proteins, with *PPR* motifs in common between each other that play a multitude of roles in organelle gene expression and organism physiology. Typically they are one of the major mediators of post-transcriptional control in mitochondria or chloroplasts by processing, splicing, editing, turnover, stability and translation of RNAs. Their combined action has profound effects on organelle biogenesis and function and, consequently, on photosynthesis, respiration, plant development, and environmental responses (Barkan and Small 2014; Manna 2015). They also can compensate for a variety of genome-level defects like point mutations (Barkan and Small 2014). Because PPR proteins bind their cognate RNA substrates in a sequence-specific manner they are considered to have great versatility for engineering. RNA editing of miRNA gene products can also be carried by PPR protein family, modifications that could be another mediator of miRNA biogenesis and action (Meng et al. 2011).

This study reveals that *PPR* transcript, which possibly codes for a PPR protein, has more expression in the pre-cotiledonary stage of embryo development than in the mature stage.

MiRNA-target expression analysis

In the small RNAs libraries, sequenced by LC Sciences, the read count is also provided for each of the sequenced transcript. This read count is the number of identical sequences which reflects the smallRNA expression and can be compared to the target expression results. Read count values for the selected miRNAs can be found in ANNEX VI, table VI.I.

Overall, if the target is cleaved by the miRNA and, at the same time, no further mechanisms are regulating the target expression, it is expected a reduction of the target expression in the periods when more miRNA is expressed.

For example, the miRNA 160 has a low read count in the T7 sample and it was not detected in T4B sample. The presence of miRNA 160 in the T7 sample supports the decreasing of the target expression in this stage, when comparing to the initial stage T4B. It is possible that the miRNA 160 is targeting the ARF encoding transcript for cleavage in the T7 sample, which reduces its expression.

The mRNA target expression can be also compared with the bioinformatic target prediction (see ANNEX VI, table VI.II). In the case of miRNA 160, however, the pair miRNA:mRNA target was only predicted in MG4B tissue sample, so it was not possible to compare with the expression results, since no expression study was performed in this sample.

Regarding the conserved miRNA 482a, the functional miRNA:target pair, was equally predicted in the T4B and T7 samples, but its target (*F-box* gene) is less expressed in the T7 stage than in the T4B. The miRNAs, mRNAs sequences and additional the cDNA for expression study were isolated from entire embryos comprising different cell types. And it is possible that the different *F-box* transcript regulation mechanisms exist between different cells. For example, some cells could have the presence of *F-box* gene regulated by miRNA 482a, which justifies the prediction of the target in the two stages by the bioinformatics tools. While in other cells the *F-box* transcript could be expressed without the negative regulation of the miRNA in study, or the *F-box* gene could be different regulated by other mechanisms, which could justify the higher expression of the transcript in the T7 than in T4B when the target were identified equal in both of the stages. These different regulation specific of each type of cell contributes to a nonlinear relation between the miRNA expression and target expression. The miRNA 482a has a low expression and was only identified for female cones (result not presented), for that reason, it is not possible to compare the miRNA read count with target expression results since these were identified in the same sample.

For novel miRNAs there is no data available of the cleaved targets read count, so its comparison to the target expression is not possible. Additionally, the relation between the miRNA read count with the target expression analysis data is not as linear as it was found for the miRNA 160.

The M06658 has a high expression across all the small RNA transcriptome libraries, particularly, it presents a higher read count in the stage T4B than the T7 (see ANNEX VI, table VI.II). Its target, PPR encoding transcripts, shows higher expression in the T4B than the T7 sample as well, which is again inconsistent with the miRNA expression. But which could, for instance, be justified by the different regulation of the target in the different cells that constitute the embryo analyzed, like previously explained for F-box encoding transcript regulated by miRNA 482a.

The M09664-target pair was predicted in the T4B sample (see ANNEX VI, table VI.II). It was expected less expression in the T4B than in T7 stage since the cleavage pair was found in this sample however, the opposite was observed. This result may be due to other regulation mechanisms, including target regulation by other miRNAs, which are explained in more detail in the next section “Multiplicity of target sites”. The M09664 was only identified in the megagametophytes and ZET5

samples where no target (*LEA*) expression was quantified, consequently is not possible to compare the miRNA expression with the target expression.

It will be interesting to study the expression of the miRNAs in study in the embryo stages T4B and T7, by RT-PCR, in order to confirm the miRNAs read count number and verify a relation between the expression results of each miRNA and respective target.

Multiplicity of target sites

Just like one miRNA can target more than one gene (multiplicity), one gene can be controlled by more than one miRNA (cooperation). RT-qPCR method does not allow to distinguish between direct or secondary interaction between the miRNA and target, or which miRNA is cleaving the target in study. And beyond being possible that other mechanisms are regulating the mRNA target, it is possible to other miRNAs or small RNAs to act on the same target (Peter 2010).

For conserved miRNAs, and from the first analysis performed by the sequencing service provider using Cleaveland, it is possible to identify other miRNAs which possible target the same miRNA targets study in this work. This way, for ARF encoding transcript it was identified a second novel miRNA targeting this gene for the pool degradome library (see ANNEX IV - Table IV.III). For F-box encoding transcript, and in addition to the mir482a identified for MG4 and MG7 degradomes, it was identified, for the pool degradome library, a novel miRNAs which also target it. Despite the targets 2S-ASP and NAC encoding transcripts were not study in terms of expression, they are presented. The *2S-ASP* gene had more miRNAs targeting it, 5 novel miRNAs identified for pool degradome, 18 novel miRNAs and 1 conserved miRNA identified for MG4B degradome library and 23 novel miRNAs plus 1 conserved miRNAs identified MG7 degradome library. Differently there were not identified other miRNAs, behind the conserved miRNA 947, targeting NAC encoding transcript.

These results are an evidence that the target mRNAs in study may not be exclusively regulated by the miRNA stated and, therefore, the expression of *ARF* and *F-box* in ZE T4B and T7 stages cannot be interpreted as result of the direct regulation of the respective miRNA in study, but maybe as result of a complex regulation.

For novel miRNAs this comparison was not possible, since the Cleaveland analyses were not performed with all miRNAs, but instead with a list of pre-selected novel miRNA.

In other hand, high confidence miRNA targets tend to have multiple target sites instead of one single site. In reported plant miRNA target prediction tools, the importance of the target site multiplicity was generally underestimated (Ding, et al. 2012). In this study multiplicity was only reported but it was not considered in the prediction.

4.4. Conclusion and Future perspectives

In this work four miRNA-target interactions were identified with high confidence, and the expression of the identified targets was studied in an attempt to further support the validation of the interactions as evidenced by degradome sequencing analysis. The identification of the miRNA targets was made by PARE analysis associated with bioinformatics tools. This study represents the first report of experimental identification of miRNA targets in *P. pinaster*. The only studies reported until now in *P.*

pinaster related to miRNA research focused only on the prediction of miRNAs, and were performed by mirNEST 2.0 and microPC bioinformatics tools (Mhuantong and Wichadakul 2009; Szcześniak and Makałowska 2014), with no experimental validation.

One of the interactions relates to a conserved target regulated by miRNA160, an *ARF* gene, which was the only conserved target found taking into consideration the 4 conserved miRNAs analyzed. The other 5 target transcripts were found for the first time as regulated by the specific miRNAs under study, corresponding to the mRNA putatively encoding NAC, 2S albumin, LEA, F-BOX and PRR proteins. However, these protein families have already been identified as encoded by targets of other miRNAs, which can indicate a different or specific regulatory mechanism in *P. pinaster*, or a network of cooperative regulation of the target. Based on the scarce information on miRNA targets known so far, it is difficult to formulate a conclusion.

From the six targets studied here, four were tested by RT-PCR in order to evaluate their expression, namely those coding for LEA, ARF, PRR and F-box proteins. The expression profiles reveal that all these genes are more expressed in T4B stage than in T7. This is not in agreement with some previous studies in which the expression of genes encoding proteins of the same families, ARF and LEA, had an opposite trend. However, since these protein families have a large number of members, the target identified here may code for different proteins, with different functions, and consequently different embryonic expression patterns. The T4B stage of embryo development is associated to more dynamic processes of protein synthesis, cell cycle regulation and signaling than the mature embryo stage (T7), therefore it is not surprising to observe higher expression of transcription factors associated with signaling processes such as ARF, or F-Box or PPR proteins. Nonetheless, it is not common to observe a higher accumulation of transcripts encoding LEA proteins in pre-cotyledonary embryo stage than in mature stages. Since the miRNA expression levels were not yet quantified by RT-qPCR, and the relative number of reads resulting from degradome or miRNA sequencing in each embryo developmental stage still require validation, additional experimental studies are needed for further conclusions about the miRNA-target interactions analyzed here.

Given the current scarce knowledge on gymnosperm miRNAs and their regulatory functions, together with the lack of a publicly available genome in *P. pinaster,* and proper annotation of many transcripts, this work represents a significant contribution to improve our knowledge on miRNA landscape in *Pinus* species and points to putative molecular functions of specific miRNAs in *P. pinaster*.

As future work it would be important to perform additional experimental studies to validate *in vivo* the miRNA-target interactions. This could be accomplished for instance thorough a luciferase reporter assay, which allows to conclude if the interaction between the miRNA and respective target effectively takes place in vivo, and if it results in the cleavage of the target transcript (Confraria and Baena-González 2016). Initial experiments were performed in order to test this assay, however, the experiments could not be completed in time for inclusion in this thesis. The validation of miRNA:target pairs by luciferase assay, will also allow to understand if it should be give preference to the MFE score over the alignment score, in the evaluation thorough bioinformatics tools, as it has been recently suggested.

It would be interesting to extend the study of miRNA to other stages of embryo development and validate a higher number of miRNA:mRNA target interactions, using different bioinformatics filtering strategies, which could include searching for miRNAs already identified for *Pinaceae*, and as potential targets use only unigenes with high similarity to already annotated genes. Additionally, in order to discovery miRNA:mRNA pairs that act by translation repression and not only by cleavage, it could be extend the target prediction to other bioinformatics programs not based on degradome using.

However demonstrating individual miRNA:mRNA interactions misses the capacity for miRNAs to regulate complex gene networks. In the future, it would be interesting to increasingly focus on miRNA-regulated networks (Peter 2010; Thomson et al. 2011).

Since so little is known about *P. pinaster* genetics it will be also interesting to study the expression and function of these miRNAs in different tissues and, like miRNAs have different expression profiles under stress conditions, maybe extend the study to evaluate responses under diverse stress conditions.

References

Abad Viñas, R. et al. **2016**. *Pinus pinaster* in Europe: distribution, habitat, usage and threats. European Atlas of Forest Tree Species, 126–127.

Addo-Quaye, C., et al. **2009**. CleaveLand: A pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics*, 25(1):130–131.

Amara, I. et al. **2014**. Insights into Late Embryogenesis Abundant (LEA) Proteins in Plants: From Structure to the Functions. *American Journal of Plant Sciences*, 5:3440–3455.

Archak, S. and Nagaraju, J. 2007. Computational Prediction of Rice (Oryza sativa) miRNA Targets. *Genomics, Proteomics and Bioinformatics*, 5(3–4):196–206.

Axtell, M.J. **2008**. Evolution of microRNAs and their targets: Are all microRNAs biologically relevant? *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1779(11):725–734.

Axtell, M.J. **2013**. Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology*, 64(January):137–159.

Axtell, M.J. and Bartel, D.P. **2005**. Antiquity of microRNAs and their targets in land plants. *The Plant cell*, 17(6):1658–1673.

Axtell, M.J. and Bowman, J.L. **2008**. Evolution of plant microRNAs and their targets. *Trends in Plant Science*, 13(7):343–349.

Barciszewska-Pacak, M. et al. 2015. Arabidopsis microRNA expression regulation in a wide range of abiotic stress responses. Frontiers in plant science, 6(June):410.

Barkan, A. and Small, I. **2014**. Pentatricopeptide repeat proteins in plants. *Annual Review of Plant Biology*, 65(January):415–42.

Bonga, J.M. **2016**. Conifer clonal propagation in tree improvement programs. *In* Vegetative propagation of forest trees (Y-S. Park, J. M. Bonga, H-K. Moon eds), National Institute of Forest Science (NIFoS), pp.9–31, korea.

Bonnet, E. et al. **2010**. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics*, 26(12):1566–1568.

Brousse, C. et al. **2014**. A non-canonical plant microRNA target site. *Nucleic Acids Research*, 42(8):5270–5279.

Budak, H. and Akpinar, B.A. **2015**. Plant miRNAs: biogenesis, organization and origins. *Functional & integrative genomics*, 15(5): 523–31.

Bustin et al. **2009**. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry*, 55(4):611–622.

Cairney, J. and Pullman, G.S. **2007**. The cellular and molecular biology of conifer embryogenesis. *New Phytologist*, 176: 511–536.

Campbell, N. A. and Reece J. **2008**. Figure 30.6 The live cycle of a pine. *In* Biology, 8th edition, p. 624.

Canales, J. et al., **2014**. De novo assembly of maritime pine transcriptome: Implications for forest breeding and biotechnology. *Plant Biotechnology Journal*, 12(3): 286–299.

Chen, C.-J. et al. **2011**. Genome-wide discovery and analysis of microRNAs and other small RNAs from rice embryogenic callus. *RNA Biology*, 8(3):538–547.

Chen, X. **2009**. Small RNAs and their roles in plant development. *Annual review of cell and developmental biology*, 25:21–44.

Chou, C.H. et al. **2016**. miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research*, 44(D1):D239–D247.

Confraria, A. and Baena-González, E. **2016**. Stress Responses in Arabidopsis Mesophyll Protoplasts. *In* Environmental Responses in Plants: Methods and Protocols, Methods in Molecular Biology (Duque, P. ed), pp. 247–269, Springer Science+Business Media, New York.

Cuperus, J.T., et. al. **2011**. Evolution and functional diversification of miRNA genes. *The Plant Cell*, 23(2):431–442.

Dai, X., et. al. 2011. Computational analysis of miRNA targets in plants: Current status and challenges. Briefings in Bioinformatics, 12(2): 115–121.

Dai, X. and Zhao, P.X. **2011**. PsRNATarget: A plant small RNA target analysis server. *Nucleic Acids Research*, 39(Web Server issue): 155–159.

Dalmay, T. **2013**. Mechanism of miRNA-mediated repression of mRNA translation. *Essays Biochemistry*, 54:29–38.

Dinkova T.D. and Alejandri-Ramirez, N. D.. **2014**. MicroRNA expression and regulation during plant somatic embryogenesis (Chapter 7). *In* Epigenetics in Plants of Agronomic Importance: Fundamentals and Applications - Transcriptional Regulation and Chromatin Remodelling in Plants (R. Alvarez-Venegas, C. De La Pena, and J. A. Casas-Mollano eds), 111-123, Springer, Switzerland.

Ding, J., et. al. **2012**. Finding MicroRNA Targets in Plants: Current Status and Perspectives. *Genomics, Proteomics and Bioinformatics*, 10(5):264–275.

Dubos, C., et. al. **2003**. Identification of water-deficit responsive genes in maritime pine (Pinus pinaster Ait.) roots. *Plant Molecular Biology*, 51(2):249–262.

EUFORGEN. **2007**. Climate change and forest genetic diversity - Implications for sustainable forest management in Europe. [pdf]. J. Koskela, A. Buck and E. Teissier du Cros eds, Bioversity ~international, Italy.

Fahlgren, N. et al. **2007**. High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLoS ONE*, 2(2):e219.

Farjon, A. **2010**. *Pinus pinaster* Aiton. *In* A handbook of the world's conifers Volume I, pp 733-735, Brill, USA.

Farjon, A. and Filer D. **2013**. *Pinus*. *In* An atlas of the world's conifers - An analysis of their distribution, biogeography, diversity and conservation status, 1st ed., pp 143-150, Brill, USA.

Felippes, F. F. et al. **2008**. Evolution of *Arabidopsis* thaliana microRNAs from random sequences. *Rna*, 14(12):2455–2459.

Ferragina, E. and Quagliarotti, D. **2008**. Climatic Change in the Mediterranean Basin: Territorial Impact and Search for a Common Strategy. *New Medit*, 4: 4-12.

Finet, C., et al. **2013**. Evolution of the *ARF* gene family in land plants: Old domains, new tricks. *Molecular Biology and Evolution*, 30(1): 45–56.

Folkes, L. et al., **2012**. PAREsnip: A tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic Acids Research*, 40(13):e130.

Gao, J. and Lan, T. **2016**. Functional characterization of the late embryogenesis abundant (LEA) protein gene family from *Pinus tabuliformis* (*Pinaceae*) in *Escherichia coli*. *Scientific Reports*, 6(January):19467.

Ghildiyal, M. and Zamore, P.D. **2009**. Small silencing RNAs: an expanding universe. *Nat Rev Genet*, 10(2):94–108.

Gonçalves, S., et al. **2005a**. Evaluation of control transcripts in real-time RT-PCR expression analysis during maritime pine embryogenesis. *Planta*, 222(3):556–63.

Gonçalves, S., et al. **2005b**. Identification of Differentially Expressed Genes During Embryogenesis in Maritime Pine (*Pinus pinaster*). *Silva Lusitana*, 13(2):203–216.

Hellemans, J. et al. **2007**. qBase relative quantification framework and software for management and

automated analysis of real-time quantitative PCR data. *Genome Biol*, 8(2):R19.

Huntzinger, E. and Izaurralde, E. **2011**. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature reviews, Genetics*, 12(2):.99–110.

ICNF. **2013**. IFN6 - Áreas dos usos do solo e das espécies florestais de Portugal continental 1995|2005|2010 - Resultados preliminares - fevereiro' 2013. 6o inventário florestal nacional. [pdf] Instituto da Conservação da Natureza e das Florestas, Lisboa.

Iwakawa, H. O. and Tomari, Y. **2015**. The Functions of MicroRNAs: mRNA Decay and Translational Repression. *Trends in Cell Biology*, 25(11):651–665.

Jalas, J. and Suominen, J. **1972**. Atlas florae Europaeae - distribution of vascular plants in Europe. Vol.2 Gymnospermae (Pinaceae to Ephedraceae). C. for M. the F. of E. & S. B. F. Vanamo, eds, 1st ed, Helsinki.

Jha, A. and Shankar, R. **2011**. Employing machine learning for reliable miRNA target identification in plants. *BMC Genomics*, 12(1):636.

Jones-Rhoades, M.W., et. al. **2006**. MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology*, 57(1): 19–53.

Kakrana, A. et al. **2014**. SPARTA: A parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Research*, 42(18):e139.

Kipreos, E.T. and Pagano, M. **2000**. The F-box protein family. *Genome Biology*, 1(5):reviews3002.1–3002.7.

Klimaszewska, K., et. al. **2016**. Advances in Conifer Somatic Embryogenesis Since Year 2000. *In* In vitro embryogenesis in higher plants (M. A. Germana and M. Lambardi eds), volume 1359 of the series Methods in Molecular Biology, 131–166.

Kozomara, A. and Griffiths-Jones, S. **2014**. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):68–73.

Krüger, J. and Rehmsmeier, M. **2006**. RNAhybrid: MicroRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(Web Serv Iss):W451–W454.

Kuhn, D.E. et al. **2008**. Experimental validation of miRNA targets. Methods, 44(1):47–54.

Kuroda H, et. al. **2002**. Classification and expression analysis of Arabidopsis F-box-containing protein genes. *Plant Cell Physiology*, 43(10):1073–1085.

Lara-Chavez, A., et. al. **2012**. Comparison of gene expression markers during zygotic and somatic

embryogenesis in pine. *In Vitro Cellular and Developmental Biology - Plant*, 48(3):341–354.

Li, J.H. et al. **2014**. StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(D1):D92–D97.

Li, T. et al. **2012**. Deep Sequencing and Microarray Hybridization Identify Conserved and Species-Specific MicroRNAs during Somatic Embryogenesis in Hybrid Yellow Poplar. *PLoS ONE*, 7(8):1–12.

Lin, Y. and Lai, Z. **2013**. Comparative Analysis Reveals Dynamic Changes in miRNAs and Their Targets and Expression during Somatic Embryogenesis in Longan (Dimocarpus longan Lour.). *PLOS ONE*, 8(4):15–18.

Liu, Q., et.al. **2014**. Analysis of complementarity requirements for plant MicroRNA targeting using a Nicotiana benthamiana quantitative transient assay. *The Plant cell*, 26(2):741–53.

Liu, S. et al. **2015**. StarScan: A web server for scanning small RNA targets from degradome sequencing data. Nucleic Acids Research, 43(W1):480-486.

Lu, S. et al. **2007**. MicroRNAs in loblolly pine (Pinus taeda L.) and their association with fusiform rust gall development. *Plant Journal*, 51(6): 1077–1098.

Luo, Y-C. et al. **2006**. Rice embryogenic calli express a unique set of microRNAs, suggesting regulatory roles of microRNAs in plant post-embryogenic development. *FEBS Letters*, 580(21):5111–5116.

Ma, X. et al. **2013**. Trip to ER: MicroRNA-mediated translational repression in plants. *RNA biology*, 10(10):1586–92.

Mahdavi-Darvari, F., et al. **2014**. Epigenetic regulation and gene markers as signals of early somatic embryogenesis. *Plant Cell*, Tissue and Organ Culture, 120(2): 407–422.

Mallory, A. C. and Vaucheret, H. **2006**. Functions of microRNAs and related small RNAs in plants. *Nature Genetics*, 38 Suppl(June): S31–S36.

Manna, S. **2015**. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie*, 113:93–99.

Martin, R. C., et. al. **2012**. Role of miRNAs in seed development. *In* MicroRNAs in Plant development and stress responses- Signaling and Communication in Plants 15 (R. Sunkar ed.), pp.109–121, Springer-Verlag Berlin Heidelberg.

Martinez-Sanchez, A. and Murphy, C. L. **2013**. MicroRNA Target Identification-Experimental Approaches. *Biology*, 2(1):189–205.

Meinke, D.W. **1995**. Molecular Genetics of Plant Embryogenesis. *Annual Review of Plant Physiology and Plant Molecular Biology*, 446(1): 369–394.

Meng, Y. et al. **2011**. The Regulatory Activities of Plant MicroRNAs: A More Dynamic Perspective. *Plant Physiology*, 157(4): 1583–1595.

Merkle, S.A. **2016**. Application of somatic embryogenesis and transgenic technology to conserve and restore threatened forest tree species. *In* Vegetative Propagation of Forest Trees (Y-S Park, J. M. Bonga and H-K Moon eds), pp.261–278, National Institute of Forest Science (NIFoS), Korea.

Mhuantong, W. and Wichadakul, D. **2009**. MicroPC (microPC): A comprehensive resource for predicting and comparing plant microRNAs. *BMC genomics*, 10:366. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2907689&tool=pmcentrez&rendertype=abstract [Accessed November 2016].

Miguel, C.M. et al. **2016**. Impact of molecular studies on somatic embryogenesis development for implementation in conifer multi-varietal forestry. *In* Vegetative Propagation of Forest Trees (Y-S Park, J. M. Bonga and H-K Moon eds), pp.373–421, National Institute of Forest Science (NIFoS), Korea.

Montalbán, I.A., et. al. **2016**. Somatic Embryogenesis in *Pinus spp*. *In* In vitro embryogenesis in higher plants (M. A. Germana and M. Lambardy eds), volume 1359 of the series Methods in Molecular Biology, pp. 405–415.

Moran, Y., et al. **2017**. The evolutionary origin of plant and animal microRNAs. *Nature Ecology & Evolution*, 1(3):0027.

Mordhorst, A.P., et. al. **1997**. Plant embryogenesis. *Critical Reviews in Plant Sciences*, 16(6): 535–576.

Numnark, S. et al. **2012**. C-mii: a tool for plant miRNA and target identification. *BMC Genomics*, 13(Suppl 7):S16.

Oh, T.J. et al., **2008**. Evidence for stage-specific modulation of specific microRNAs (miRNAs) and miRNA processing components in zygotic embryo and female gametophyte of loblolly pine (*Pinus taeda*). *The New phytologist*, 179(1):67–80.

Olvera-Carrillo, Y., et. al. **2011**. Late embryogenesis abundant proteins. *Plant Signaling & Behavior*, 6(4):586–589.

Parizotto, E.A. et al. **2004**. *In vivo* investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA. *Genes & Dev*., 18(18):2237–2242.

Park, M.Y. et al. **2005**. Nuclear processing and export of microRNAs in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10): 3691–3696.

Peter, M. **2010**. Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene*, 29(15):2161–2164.

Pfaffl, M. W. **2001**. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Research*, 29(9):e45.

Pfaffl, M. W., **2004**. Quantification strategies in real-time PCR (Chapter 3). *In* A-Z of quantitative PCR (S. A. Bustin ed), pp.87-112, International University Line (IUL), La Jolla, CA, USA.

Procogen. **2013**. Promoting a Functional and Comparative Understanding of the Conifer Genome Implementing Applied Aspects for More Productive and Adapted Forests. Available at: http://www.procogen.eu/ [Accessed at September 2016]

Quinn, C.R., et. al. **2014**. Expression patterns of conserved microRNAs in the male gametophyte of loblolly pine (*Pinus taeda*). *Plant Reproduction*, 27(2):69–78.

Rademacher, E.H. et al. **2011**. A cellular expression map of the *Arabidopsis AUXIN RESPONSE FACTOR* gene family. *Plant Journal*, 68(4):597–606.

Reinhart, B. J., et al. **2002**. MicroRNAs in plants. *Genes Dev*, 16(13):1616–1626.

Rhoades, M.W., et al. **2002**. Prediction of plant microRNA targets. *Cell*, 110(4):513–520.

Rogers, K. and Chen, X. **2013**. Biogenesis, turnover, and mode of action of plant microRNAs. *The Plant cell*, 25(7):2383–99.

Seefried, W.F. et al. **2014**. Global regulation of embryonic patterning in *Arabidopsis* by microRNAs. *Plant physiology*, 165(June): 670–687.

Stocks, M.B. et al. **2012**. The UEA sRNA workbench: A suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, 28(15):2059–2061.

Sun, G. **2012**. MicroRNAs and their diverse functions in plants. *Plant Molecular Biology*, 80(1):17–36.

Sunkar, R., et. al. **2012**. Functions of microRNAs in plant stress responses. *Trends in Plant Science*, 17(4): 196–203.

Szcześniak, M.W. and Makałowska, I. **2014**. MiRNEST 2.0: A database of plant and animal microRNAs. *Nucleic Acids Research*, 42(Database):74–77.

The world Bank. **2016**. Forest area (% of land area) 1990-2015. Available at: http://data.worldbank.org/indicator/AG.LND.FRST.ZS [Accessed July 2016].

Thomson, D.W., et. al. **2011**. Experimental strategies for microRNA target identification. *Nucleic Acids* Research, 39(16):6845–6853.

Trontin, J.-F. et al., **2016**. Molecular Aspects of Conifer Zygotic and Somatic Embryo Development: A Review of Genome-Wide Approaches and Recent Insights. *In* In vitro embryogenesis in higher plants (M. A. Germana and M. Lambardi eds), volume 1359 of the series Methods in Molecular Biology, 167–207.

Vashisht, D. and Nodine, M. D. **2014**. MicroRNA functions in plant embryos. *Biochem Soc Trans*, 42(2):352–7.

de Vega-Bartol, J.J., et al. **2013a**. Normalizing gene expression by quantitative PCR during somatic embryogenesis in two representative conifer species: *Pinus pinaster* and *Picea abies*. *Plant cell reports*, 32(5):715–29.

de Vega-Bartol, J.J., et al. **2013b**. Transcriptomic analysis highlights epigenetic and transcriptional regulation during zygotic embryo development of *Pinus pinaster*. *BMC plant biology*, 13:123.

Von Arnold, S. **2008**. Somatic Embryogenesis (Chapter 9). *In* Plant Propagation by Tissue Culture Volume 1 (E. F. George, M. A. Hall, G-J De Klerk, eds.), 3rd Edition, pp. 335–354, The Background.

Von Arnold, S. et al. **2016**. Norway spruce as a model for studying regulation of somatic embryo development in conifers. *Vegetative Propagation of Forest Trees*, 1(1):351–372.

Wan, L.-C. et al. **2012**. Transcriptome-wide identification and characterization of miRNAs from *Pinus* densata. *BMC genomics*, 13(1):132.

Wang, F., et. al. **2015**. More than meets the eye? Factors that affect target selection by plant miRNAs and heterochromatic siRNAs. *Current opinion in plant biology*, 27:118–124.

Willmann, M.R. et al. **2011**. MicroRNAs regulate the timing of embryo maturation in Arabidopsis. *Plant physiology*, 155(4): 1871–1884.

Winkelmann, T. **2016**. Somatic Versus Zygotic Embryogenesis: Learning from Seeds. *In* In vitro embryogenesis in higher plants (M. A. Germana and M. Lambardy eds), volume 1359 of the series Methods in Molecular Biology, pp.25–46.

Wu, L. et al. **2010**. DNA Methylation Mediated by a MicroRNA Pathway. *Molecular Cell*, 38(3):465–475.

Wu, X. M. et al., **2011**. Stage and tissue-specific modulation of ten conserved miRNAs and their targets during somatic embryogenesis of Valencia sweet orange. *Planta*, 233(3):495–505.

Wu, X. M. et al. **2015**. Genomewide analysis of small RNAs in nonembryogenic and embryogenic tissues of citrus: microRNA- and siRNA-mediated transcript cleavage involved in somatic embryogenesis. *Plant Biotechnology Journal*, 13(3):383–394.

Xie, F. and Zhang, B. **2010**. Target-align: A tool for plant microrna target identification. *Bioinformatics*, 26(23):3002–3003.

Yang, X. et al. **2013**. Small RNA and degradome sequencing reveal complex miRNA regulation during cotton somatic embryogenesis. *Journal of Experimental Botany*, 64(6):1521–1536.

Zažímalová, E. et. al. **2014**. Auxin and its role in plant development, Springer-Verlag Wien.

Zhang, B. et al. **2005**. Plant microRNA: A small regulatory molecule with big impact. *Developmental Biology*, 289(1):3–16.

Zhang, B., et al. **2006a**. Conservation and divergence of plant microRNA genes. *Plant Journal*, 46(2):243–259.

Zhang, B., et. al. **2006b**. Identification of 188 conserved maize microRNAs and their targets. *FEBS Letters*, 580(15):3753–3762.

Zhang, J. et al. **2012**. Genome-wide identification of microRNAs in larch and stage-specific modulation of 11 conserved microRNAs and their targets during somatic embryogenesis. *Planta*, 236(2):647–657.

Zhang, Y. **2005**. miRU: An automated plant miRNA target prediction server. *Nucleic Acids Research*, 33(Web Server issue):701–704.

Zheng, Y. et. al., **2012**. SeqTar: an effective method for identificatifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic Acids Research*, 40(4):e28.

Zhang, Z. et al., **2010**. PMRD: Plant microRNA da tabase. *Nucleic Acids Research*, 38(Database):806–813.

Annexes

ANNEX I – Previous work

Construction of transcriptome, small RNA, degradome and miRNA libraries

Small RNA and degradome libraries were previously generated to cover different developmental stages of embryo, megagametophyte and reproductive cone tissues which were independently sampled. (see Table 1.I and 1.II for the tissues and libraries considered in this thesis).

Table I List of small RNAs libraries and respective tissues of origin

Small RNAs library	Tissue and/or stage of the seed
1	ZE stages T0/TT2
2 and 3	ZE stages T3/T4 (2 replicates)
4 and 5	ZE stage T4B (2 replicates)
6 and 7	ZE stage T5 (2 replicates)
8 and 9	ZE stage T7 (2 replicates)
10	Megagametophyte stage T0/T1/T2
11 and 12	Megagametophyte stage T4B (2 replicates)
13 and 14	Megagametophyte stage T7 (2 replicates)
15	Female cone
16	Male cone

Table I.II List of degradome libraries and respective tissues of origin

Degradome library	Tissue and/or stage of the seed
1	Pool of ZE at T0-T7 stages
2	Stage T4B
3	Stage T7
4	Megagamethophyte stage T4B
5	Megagamethophyte stage T7
6	Female cone
7	Male cone

Lists of annotated miRNAs derived from the analysis of the small RNA libraries done with the in-house sRNA analysis pipeline “miRPursuit”, which uses MirCAT bioinformatics tool. (Stocks et al. 2012)

The transcriptome of reference for *Pinus pinaster* Ait. used was the one published by (Canales et al. 2014), and is public available in Sustainpine version 3.0. (http://www.scbi.uma.es/sustainpinedb/home_page accessed at September 2015).

ANNEX II - Support tables for chapter 2.2 Methods

Table II.I Degradome vs transcriptome combinations used in bioinformatics analyses.

Degradome library	Transcriptome library (mRNA)
zygotic embryo pool	ZE stages T0/T1/T2; ZE stages T3/4 (2 replicates); ZE stage T4B (2 replicates); ZE stage T5 (2 replicates);
stage T4B	ZE stage 4B (2 replicates)
stage T7	ZE stage T7 (2 replicates)
Megagamethophyte stage T4B	Megagametophyte stage T4B (2 replicates)
Megagamethophyte stage T7	Megagametophyte stage T7 (2 replicates)
Female cone	Female cone
Male cone	Male cone

Table II.II Summary of RNA extraction reactions.

Stage	Extraction number	Embryos quantity
T0/T1/T2	1	60
	2	60
	3	60
	4	50
T3/T4	1	30
	2	32
T4B	1	19
	2	21
	3	9
T5	1	15
	2	14
T7	1	20
	2	21

Table II.III Summary of cDNA synthesis reactions. Legend: RT+, reverse transcription reaction; RT-, reverse transcription minus control (reverse transcription reaction prepared with water instead of reverse transcriptase enzyme).

Embryo stage	Biological replicate	Reactions
T4B	1	4RT+&1RT-
T4B	2	4RT+&1RT-
T7	1	(3RT+&1RT-); (3RT+&1RT-);
T7	2	(3RT+&1RT-); (2RT+&1RT-);
Pool	1	5RT+ & RT-

ANNEX III - Determination of RT-qPCR primer amplification efficiency for the reference and target genes among the different tissues and biological replicates

Reference Gene ATUB:

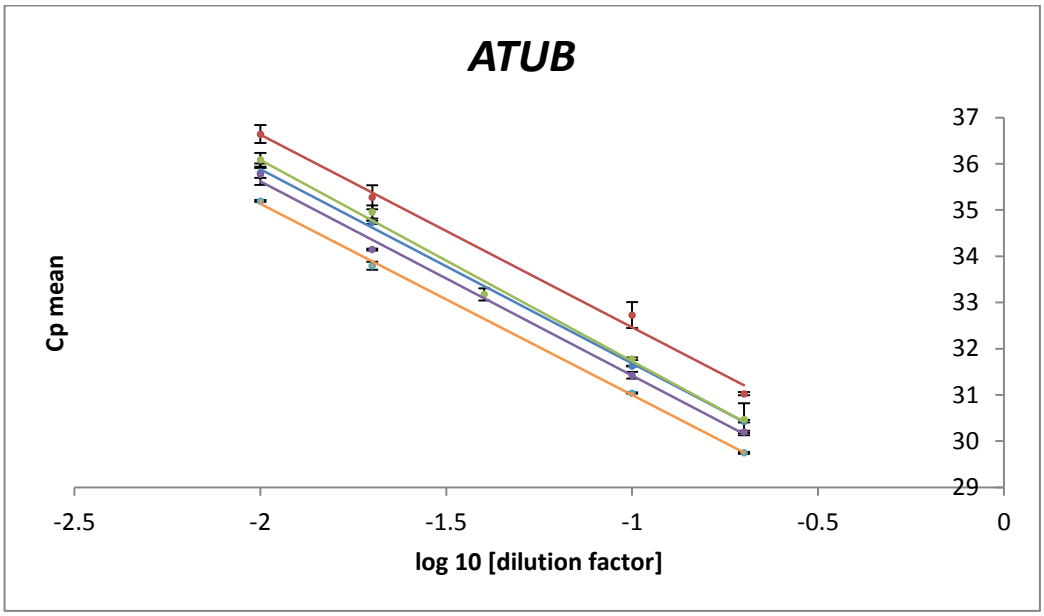


Figure III.I Determination of RT-qPCR efficiencies of reference gene *ATUB*. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points corresponds to different cDNA samples, for a different stage and biological replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.I Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with its calculation, for the reference gene ATUB.

ATUB	SLOPE	SE(SLOPE)	Y intercepts	R^2	Efficiency	SE(Ef)
T4B_RB1	-4.20	0.13	27.48	1.00	72.96	0.98
T4B_RB2	-4.17	0.23	28.30	0.99	73.74	2.23
T7_RB1	-4.35	0.19	27.38	0.99	69.69	1.64
T7_RB2	-4.20	0.18	27.22	1.00	73.08	1.72
POOL	-4.13	0.08	26.87	1.00	74.59	0.82

Reference Gene EF1:

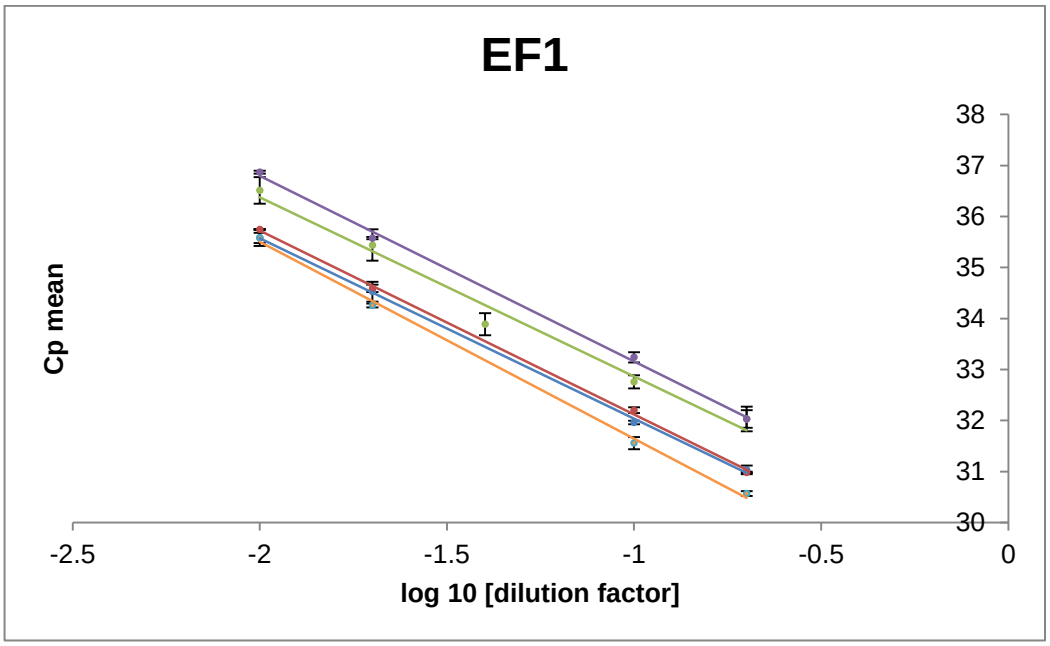


Figure III.II Determination of RT-qPCR efficiencies for reference gene *EF1*. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points correspond to different cDNA samples, for a different stage and biological replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.II Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with its calculation, for the reference gene EF1.

EF1	SLOPE	SE(SLOPE)	Y intercepts	R^2	Efficiency	SE(Ef)
T4B_RB1	-3.54	0.07	28.50	1.00	91.77	1.12
T4B_RB2	-3.61	0.07	28.51	1.00	89.37	1.17
T7_RB1	-3.51	0.27	29.35	0.98	92.55	4.58
T7_RB2	-3.64	0.12	29.52	1.00	88.41	1.78
POOL	-3.86	0.12	27.78	1.00	81.64	1.45

Reference Gene HISTO3:

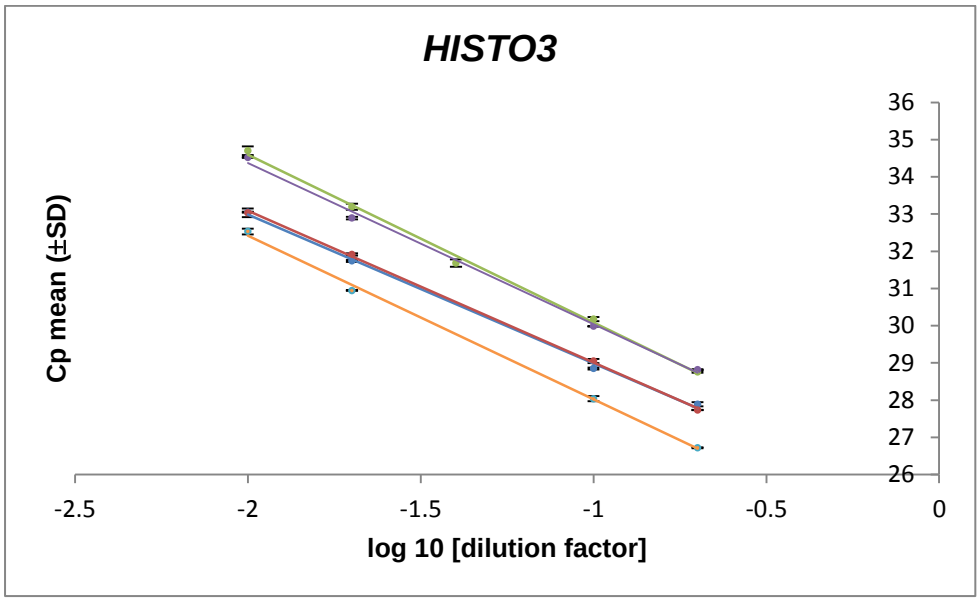


Figure III.III Determination of RT-qPCR efficiencies of reference gene *HISTO3*. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points correspond to different cDNA samples, for a different stage and biological replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.III Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with it calculation, for the reference gene HISTO3.

HISTO3	SLOPE	SE(SLOPE)	Y intercepts	R^2	Efficiency	SE(Ef)
T4B_RB1	-4.00	0.13	24.98	1.00	77.73	1.43
T4B_RB2	-4.08	0.07	24.93	1.00	75.84	0.68
T7_RB1	-4.51	0.14	25.58	1.00	66.68	1.04
T7_RB2	-4.33	0.17	25.71	1.00	70.15	1.45
POOL	-4.40	0.13	23.62	1.00	68.79	1.06

ARF

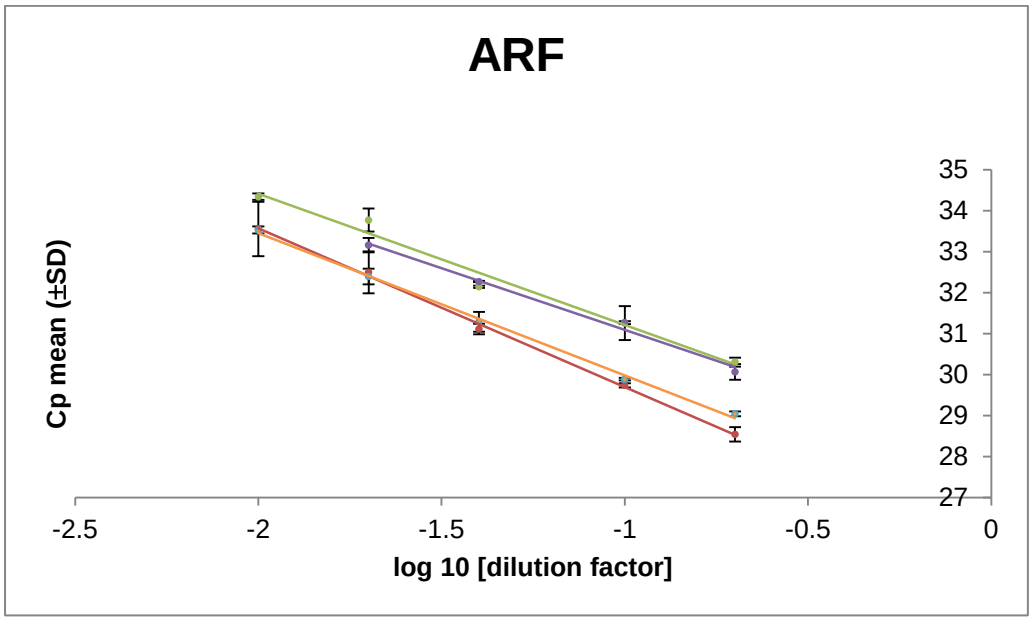


Figure III.IV Determination of RT-qPCR efficiencies of the ARF encoding transcript. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points correspond to different cDNA samples, for a different stage and biological replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.IV Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with it calculation, for the ARF encoding transcript.

ARF	SLOPE	SE(SLOPE)	Y intercepts	R^2	Efficiency	SE(Ef)
T4B_RB1	-3.70	0.11	25.42	1.00	86.23	1.56
T4B_RB2	-3.87	0.09	25.83	1.00	81.32	1.11
T7_RB1	-3.20	0.27	28.02	0.98	105.46	6.33
T7_RB2	-3.01	0.20	28.08	0.99	114.89	5.83
POOL	-3.47	0.10	26.51	1.00	94.01	1.75

LEA

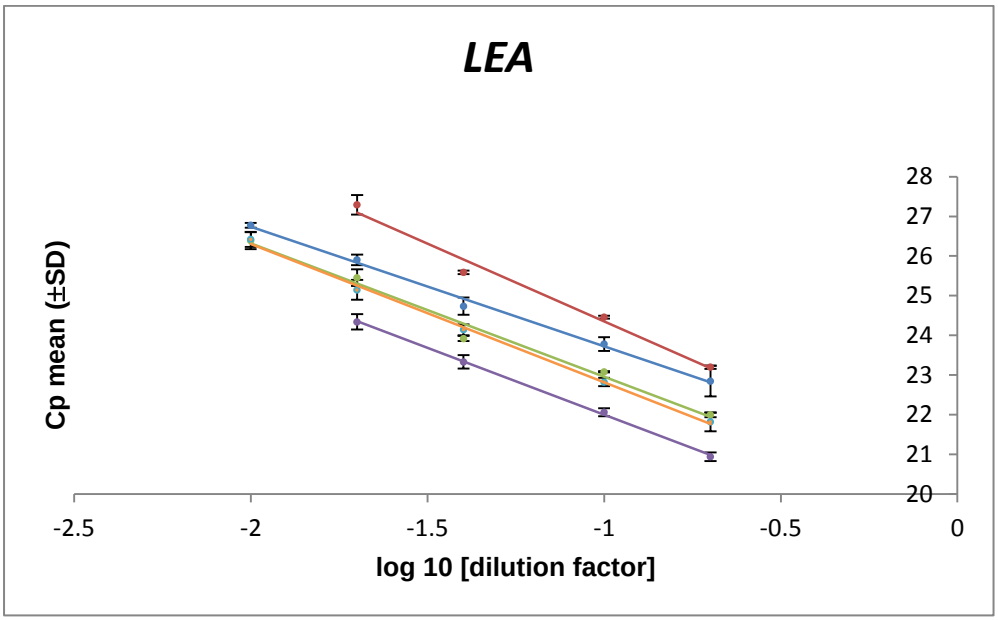


Figure III.V Determination of RT-qPCR efficiencies of LEA encoding transcript. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points correspond to different cDNA samples, for a different stadium and biologic replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.V Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with it calculation, for the LEA encoding transcript.

LEA	SLOPE	SE(SLOPE)	Y intercepts	R^2	Efficiency	SE(Ef)
T4B_RB1	-3.02	0.12	20.71	0.995	114.54	3.44
T4B_RB2	-3.92	0.36	20.43	0.983	79.83	4.35
T7_RB1	-3.36	0.24	19.59	0.985	98.27	4.74
T7_RB2	-3.37	0.07	18.62	0.999	97.86	1.46
POOL	-3.49	0.10	19.33	0.998	93.49	1.74

PPR

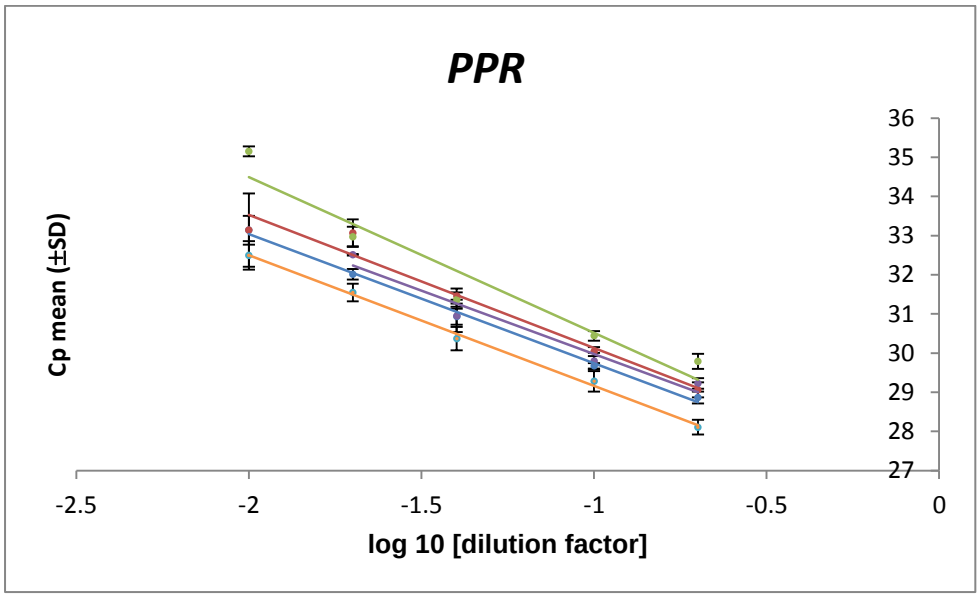


Figure III.VI Determination of RT-qPCR efficiencies of PPR encoding transcript. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points correspond to different cDNA samples, for a different stadium and biologic replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.VI Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with it calculation, for the PPR encoding transcript.

12150P3	**SLOPE**	**SE(SLOPE)**	**Y intercepts**	**R^2**	**Efficiency**	**SE(Ef)**
T4B_RB1	-3.29	0.12	101.17	2.48	101.17	0.12
T4B_RB2	-3.40	0.38	96.79	7.40	96.84	0.38
T7_RB1	-3.98	0.63	78.46	7.20	78.47	0.63
T7_RB2	-3.23	0.47	103.77	10.71	113.95	0.47
POOL	-3.34	0.10	99.31	2.08	99.31	0.10

F-Box PM

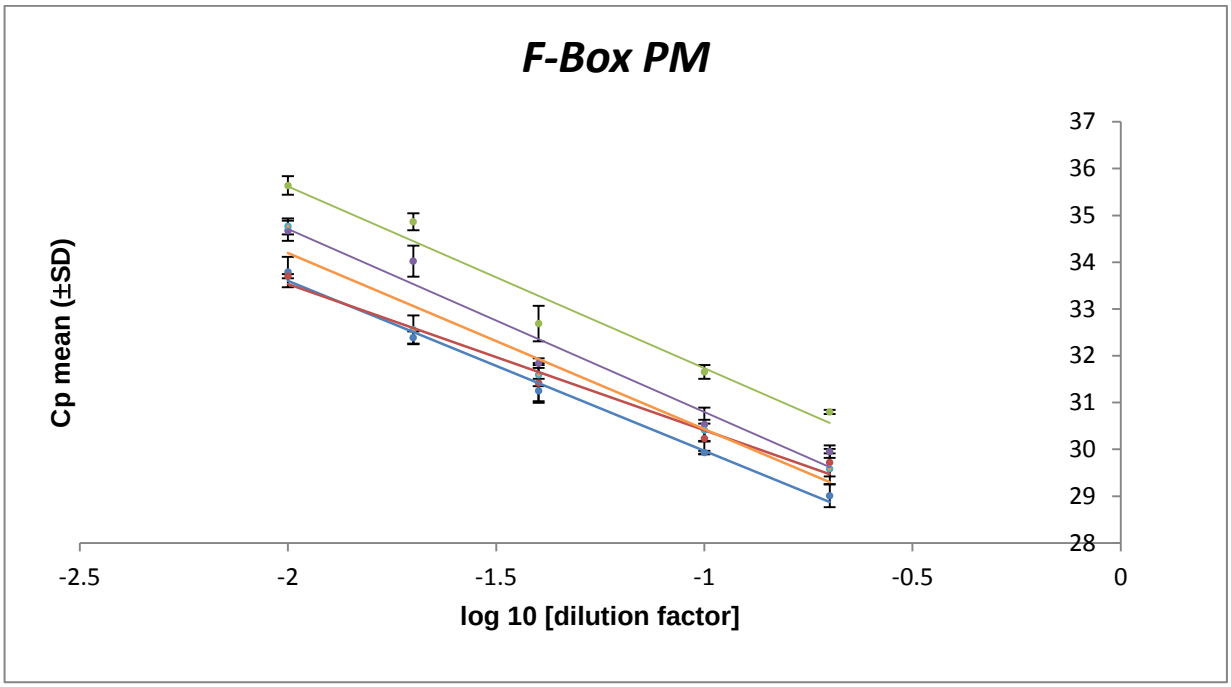


Figure III.VII Determination of RT-qPCR efficiencies of F-box PM encoding transcript. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points correspond to different cDNA samples, for a different stadium and biologic replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.VII Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with it calculation, for the F-box PM encoding transcript.

F-Box PM	SLOPE	SE(SLOPE)	Y intercepts	R^2	Efficiency	SE(Ef)
T4B_RB1	-3.63	0.17	26.33	0.99	88.44	2.63
T4B_RB2	-3.12	0.31	27.29	0.98	109.07	7.89
T7_RB1	-3.89	0.42	27.84	0.97	80.77	5.21
T7_RB2	-3.91	0.46	26.89	0.96	80.18	5.52
POOL	-3.77	0.49	26.66	0.95	84.31	6.66

F-Box P3

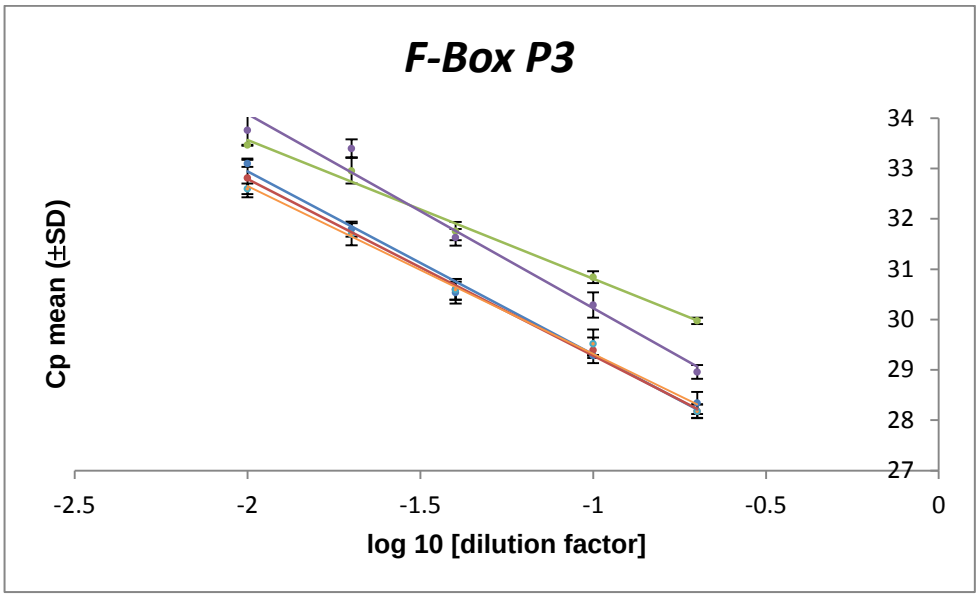


Figure III.VIII Determination of RT-qPCR efficiencies of F-box P3 encoding transcript. Ct means were plotted versus the log [cDNA dilution factor] to slope estimation for efficiency calculation. The linear regression was performed with excel tool. The error bars represent the SD of the Ct mean. Each set of color points correspond to different cDNA samples, for a different stadium and biologic replicate, namely T4B_RB1 (blue line), T4B_RB2 (red line), T7_RB1 (green line), T7_RB2 (purple line) and the control pool (orange line).

Table III.VIII Table of the linear regression constants, statistics of the linear regression r^2 and SE(SLOPE), efficiency values and the standard error associated with it calculation, for the F-box PM encoding transcript.

5940P3	SLOPE	SE(SLOPE)	Y intercepts	R^2	Efficiency	SE(Ef)
T4B_RB1	-3.63	0.17	25.67	0.99	88.45	2.60
T4B_RB2	-3.58	0.13	25.73	1.00	90.37	2.05
T7_RB1	-2.75	0.16	28.06	0.99	130.72	6.20
T7_RB2	-3.85	0.33	26.37	0.98	81.75	4.23
POOL	-3.33	0.14	25.99	0.99	99.61	2.94

ANNEX IV – Cq data for relative expression calculations

Table IV.I Cp average values and other statistic data from RT-qPCR analysis.

Target Gene	Stadium	Cpm	SD (Cp)	ΔCp
LEA	T4B_RB1*	22.35	0.09	0.16
	T4B_RB2	23.21	0.34	0.67
	T7_RB1	21.79	0.12	0.23
	T7_RB2	21.59	0.04	0.07
	Pool	21.83	0.03	0.06
	Pool*	22.43	0.10	0.19
F-Box PM	T4B_RB1	28.87	0.11	0.19
	T4B_RB2	29.99	0.17	0.32
	T7_RB1	31.01	0.32	0.6
	T7_RB2*	30.37	0.31	0.61
	Pool*	28.87	0.13	0.26
	Pool	29.56	0.02	0.03
F-Box P3	T4B_RB1	29.01	0.24	0.48
	T4B_RB2	29.72	0.30	0.59
	T7_RB1*	30.15	0.22	0.31
	T7_RB2*	29.24	0.20	0.4
	Pool*	28.21	0.14	0.2
	Pool	29.58	0.33	0.61
PPR	T4B_RB1*	28.86	0.15	0.3
	T4B_RB2	28.93	0.03	0.06
	T7_RB1	29.03	0.12	0.2
	T7_RB2*	29.22	0.13	0.24
	Pool*	28.11	0.19	0.35
	Pool	28.04	0.11	0.22
ARF	T4B_RB1	27.55	0.34	0.48
	T4B_RB2	28.32	0.36	0.51
	T7_RB1	29.54	0.15	0.21
	T7_RB2*	30.64	0.11	0.22
	Pool*	28.26	0.06	0.12
	Pool	28.21	0.10	0.14

Table IV.II RG cp average values and other statistic data from RT-qPCR analysis.

Target Gene	Stadium	Cpm	SD (Cp)	ΔCp
ATUB	T4B_RB1**	31.57	0.65	0.92
	T4B_RB2*	30.65	0.05	0.1
	T7_RB1	29.41	0.14	0.27
	T7_RB2	28.51	0.27	0.53
	Pool	35.15	0.30	0.58
	Pool*	29.50	0.02	0.03
	Pool**	29.06	0.16	0.28
EF1	T4B_RB1**	30.26	0.19	0.37
	T4B_RB2*	30.84	0.08	0.15
	T7_RB1	31.11	0.06	0.12
	T7_RB2	31.15	0.05	0.09
	Pool	31.47	0.15	0.29
	Pool*	30.27	0.05	0.09
	Pool**	30.26	0.06	0.12
HISTO3	T4B_RB1	26.79	0.02	0.03
	T4B_RB2	27.12	0.05	0.1
	T7_RB1	28.54	0.09	0.16
	T7_RB2	28.50	0.18	0.33
	Pool	26.50	0.03	0.06

Cp average values from RT-PCR, for each GOI and RG, and each of the studied zygotic embryo stage samples, including the calibrator sample, namely the pool. The reason why some genes have more than one calibrator sample (pool) is because the tests were divided in more than one plate, the correspondence of the ZE of stage sample to the respective pool sample are done by *. There are also other statistical data available in this table, specifically SD(Cp) (Cps standard deviation of the tree technical replicates) and ΔCp (range of the Cp values between the tree technical replicates).

ANNEX V - Relative expression between the two biological replicates

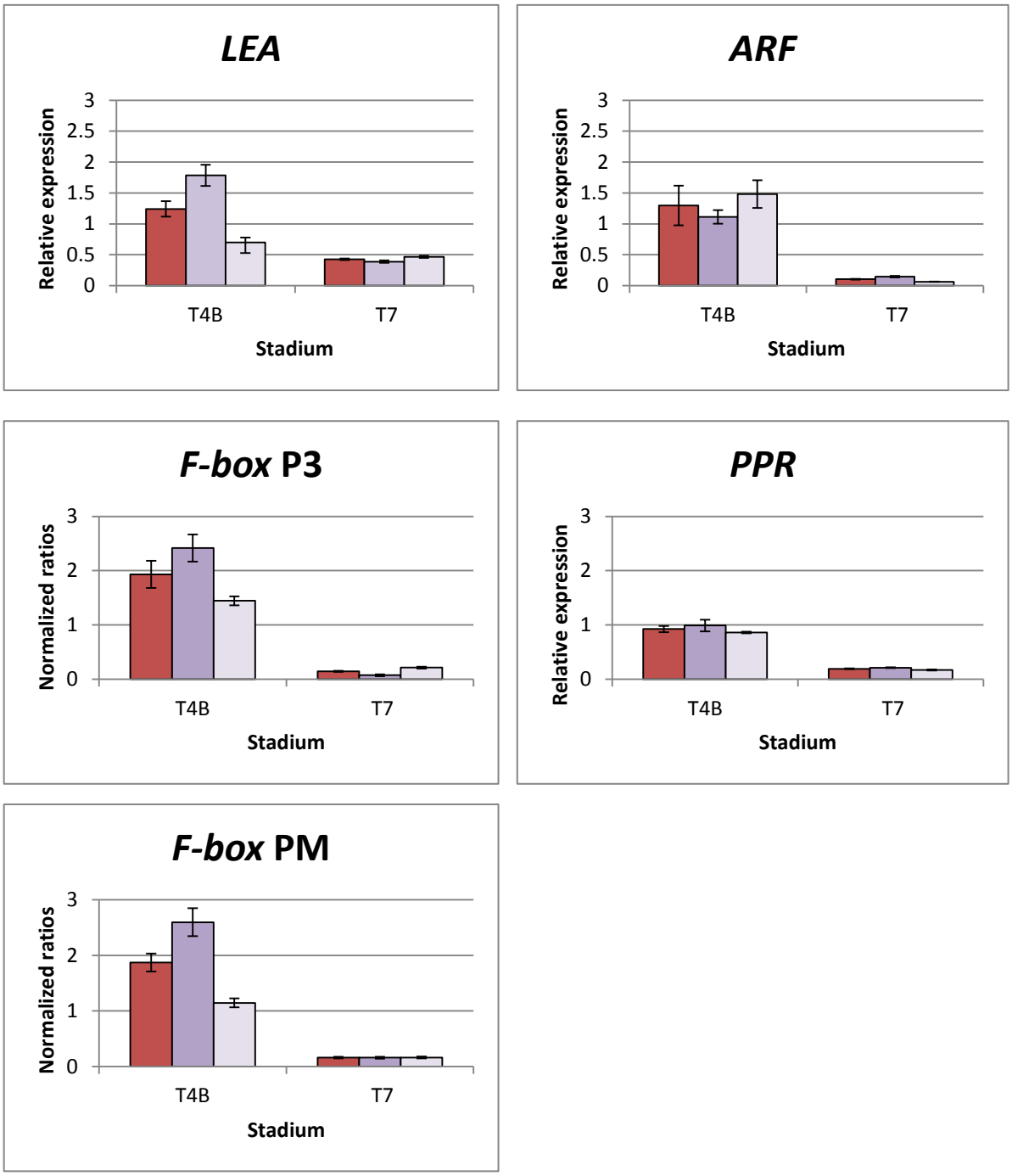


Figure V.IX Relative expression of four different genes among two different stages, the early-cotiledary stage (T4B) and the mature embryo (T7). Biologic replicate 1; biologic replicate 2; resulting expression (mean of the two biological replicates).

ANNEX VI – MiRNAs and degradome transcripts read counts from sequencing

Table VI.I Read counts of the miRNAs studied in this work, obtained from sequencing libraries prepared from the different ZE and MGM stages. The data are from the sequencing service performed by LC Science. For all the samples two biological replicates (RB) were performed, except for "stage T0/T1/T2" that only had one sequenced sample since the biological material was limited.

miRNA	Zygotic embryo								
	Stage T0/T1/T2	Stage T3/T4		Stage T4B		Stage T5		Stage T7	
	RB1	RB1	RB2	RB1	RB2	RB1	RB2	RB1	RB2
miR160	0	7	5	0	0	13	7	14	6
mir408	0	0	0	0	0	0	0	24	0
miR482	0	0	0	0	0	0	0	0	0
mir947	40644	36456	67338	52630	66586	72085	107727	69277	18573
M09664	0	0	0	0	0	6	5	0	0
M05987	717	477	343	159	345	174	466	85	95
M06658	11693	8612	62042	33249	33923	15125	25261	20905	15100

miRNA	Megagametophyte				
	Stage T0/T1/T2	Stage T4B		Stage T7	
	RB1	RB1	RB2	RB1	RB2
miR160	5	6	5	7	0
mir408	11	82	42	252	136
miR482	0	0	0	0	0
mir947	27420	60180	148612	89164	91744
M09664	0	21	5	12	13
M05987	648	2196	1913	3311	1374
M06658	17820	38187	16024	40179	28459

Table VI.II Read count of the selected cleaved target predicted by CleavelandLC for each degradome library sequenced.

miRNA	mRNA (binding site)	Target transcript	Embryo stage or tissue				
			Pool	T4B	T7	MG4B	MG7
miR160	unigene806:956	*ARF*	0	0	0	1	0
miR408	unigene8705:486	*2S Albumin*	0	0	2	2	9
miR482a	unigene5940:891	*F-Box*	4	1	1	1	1
miR947	unigene22292:121	*NAC*	0	1	0	1	0

Table VI.III Number of alternative miRNAs which have as target the same transcripts as the conserved miRNAs studied in this work, for each of the degradome libraries (Pool, T4B, T7, MG4, MG7).

Target	**Pool**	**T4B**	**T7**	**MG4**	**MG7**
ARF	1	0	0	0	0
2S albumin	5	0	4	19	24
F-Box	1	0	0	0	0
NAC	0	0	0	0	0