IN SUPREMÆ DIGNITATIS

1343

UNIVERSITÀ DEGLI STUDI DI PISA

Department of Agriculture, Food and Environment

Master of Science
in
Plant and Microbe Biotechnology

Long non-coding RNAs, a novel class of regulatory RNAs: identification and characterization in the model species *Brachypodium distachyon*

Candidate
Alice Pieri

Supervisors
Prof. Mario Enrico Pè
Dr. Rodolfo Bernardi

Co-Supervisor
Prof. Andrea Cavallini

Academic Year 2014/2015

Table of contents

Abstract

Ninety percent of the eukaryotic genome is transcribed although only a small part corresponds to protein coding mRNAs, suggesting that a large proportion of transcribed RNAs do not code for proteins, hence classified as non-coding RNAs (ncRNAs). High-throughput sequencing technology has allowed the identification and characterization of several classes of ncRNAs with key roles in various biological processes. Among ncRNAs, long ncRNAs (lncRNAs) are transcripts typically longer than 200 nucleotides that tend to be expressed at low levels and exhibit tissue-specific/cell-specific or stress responsive expression profiles. LncRNAs have been identified in animals and in plants as well, where they are involved in different regulatory pathways both in development and stress responses, even if the understanding of molecular basis of these mechanisms remains largely unexplored.

My thesis project aims at identifying lncRNAs in *Brachypodium distachyon* (Bd)*,* a wild grass belonging to the *Pooideae* and a model species for temperate cereals, such as wheat and barley.

A whole-genome annotation and a detailed analysis of lncRNAs expression patterns have been performed for the first time in *Brachypodium.* Moreover the potential lncRNA targets were investigated to highlight new regulatory networks and cross-talk between different RNA molecules.

Public and proprietary RNA-Seq data sets from 15 different experiments conducted in the reference inbred line Bd21 were analysed in this study. Public RNA-Seq data from different experiments, including several plant organs, were downloaded from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra). Proprietary RNA-Seq libraries were previously produced by the lab from three developmental leaf areas: proliferation, expansion and mature, grown in control and drought stress conditions. For each proprietary RNA-Seq sample, three biological replicates were produced.

This dataset is characterized by a total of 705 millions reads, which were subjected to a quality analysis. Each experiment was aligned independently to the Bd21 reference genome (v.2.1) using the spliced read aligner TopHat2 and, successively, for each experiment the transcriptome was *de novo* assembled using Cufflinks.

In order to identify Bd lncRNAs an in house bioinformatic pipeline was used. Briefly, this pipeline applies five filters based on the main lncRNA features: size selection,

Open Reading Frame filter, known protein domain filter, Coding Potential Calculator, filter of housekeeping lncRNAs and precursors of small RNAs. Starting from the whole set of loci/isoforms (99141) *de novo* reconstructed, 2507 *bona fide* lncRNAs were identified.

Bona fide lncRNAs differential expression analysis was taken into account for datasets with replicates, *i.e*. proprietary libraries from different developing areas of the third leaf. This analysis revealed that several lncRNAs are differentially expressed during leaf cell differentiation and during drought treatment. Some lncRNAs resulted more abundant in specific plant stages, tissues or organs.

Moreover, a computational method developed to identify endogenous microRNA target mimic (eTM) allowed to investigate the link between lncRNAs and microRNAs through target mimicry, a regulatory mechanism for miRNA functions in plants in which the decoy RNAs bind to miRNAs via complementary sequences and therefore could interfere with the interaction between miRNAs and their authentic targets.

1 Introduction

1.1 RNA world

RNA plays a central role in the pathway from DNA to proteins, known as the "central dogma" of molecular biology. The classic view of the central dogma of biology states that "the coded genetic information hard-wired into DNA is transcribed into individual transportable cassettes, composed of messenger RNA (mRNA); each mRNA cassette contains the program for synthesis of a particular protein (or small number of proteins)" (Lodish *et al.*, 2000).

As a general rule, the classic view of central dogma of biology reflects how the sequence information is organized and transferred between information-carrying biopolymers (DNA, mRNA and proteins) in living organisms. However, many exceptions to this dogma are now known as a result of genomic studies performed during the last decade and supported by new sequencing technologies.

Nowadays we know that up to 90% of eukaryotic genome is transcribed into both protein-coding and non protein-coding RNAs (ncRNAs). Until recently, discrimination between these two categories was relatively straightforward. Most transcripts were clearly identifiable as protein-coding messenger RNAs, which convey the genetic information from DNA to ribosome, and were readily distinguished from the small number of well characterized ncRNAs, with a completely different structure and, anyway, involved in protein biosynthesis, such as transfer, ribosomal and spliceosomal RNAs.

Recently, genome-wide studies have revealed the existence of thousands of non-coding transcripts (Dinger *et al.*, 2008). First of all, small regulatory RNAs (microRNAs and small interfering RNAs) were discovered and classified in different categories on the basis of length, function, biogenesis, structural features and protein binding partners (Farazi *et al.*, 2008). Small RNAs were found to perform diverse biological functions by guiding sequence-specific gene silencing at transcriptional and/or post-transcriptional level, known as RNA interference (RNAi) (Farazi *et al.*, 2008). Actually, a phenomenon of RNAi was reported for the first time in 1990, by Napoli *et al.* trying to deepen the colour of petunias. Overexpressing a key enzyme in flavonoid biosynthesis, they obtained white petunias, as result of the turning off of the gene (Napoli *et al.*, 1990). Nowadays, RNAi is widely used for systematic analysis of gene function and is under investigation for its potential therapeutic applications (Haussecker, 2014).

Attention is now shifting toward a novel class of ncRNAs, long non-coding RNAs (lncRNAs), increasingly recognized as functional regulatory component in eukaryotic gene regulation. They were discovered in 90s in humans (Brown *et al.*, 1992) and only in 2007 in plants (Franco-Zorrilla *et al.*, 2007).

The FANTOM consortium pioneered the genome-wide discovery of lncRNAs in mouse in 2000s, based on cDNA sequencing (Maeda *et al.*, 2006). ENCODE (Derrien *et al.*, 2012) and NONCODE (Xie *et al.*, 2014) projects followed, especially thanks to RNA Sequencing (RNA-Seq) technology advent, allowing the annotation of novel human and mouse lncRNAs. In plants, the study of lncRNAs is still in its infancy, with a fine annotation only for rice (Zhou *et al.*, 2009), maize (Li *et al.*, 2014), cotton (Wang *et al.*, 2015), *Populus* (Chen *et al.*, 2015; Shuai *et al.*, 2014) and tomato (Zhu *et al.*, 2015).

1.2 Non-coding RNA

The so called "dark matter" of the genome, *i.e.* non-coding genome, comprises a diverse group of transcripts:

- "Housekeeping" ncRNAs (ribosomal RNAs, transfer RNAs, small nuclear RNAs and small nucleolar RNAs).
- "Regulatory" ncRNAs:
    - Small regulatory RNAs, such as micro RNAs (miRNAs) and small interfering RNAs (siRNAs).
    - Long non-coding RNAs (lncRNAs) (Kim and Sung, 2012).

Small regulatory RNAs (sRNAs) are tiny molecules of approximately 20-24 nucleotides in length. They act to fine-tuning gene expression through sequence complementary-dependent mechanisms at both transcriptional and post-transcriptional level, binding complementary target mRNAs, inhibiting their translation and interacting with epigenetic DNA-methylation for RNA-directed DNA methylation (RdDM).

sRNAs are generated via processing of longer double-stranded RNA (dsRNA) precursors by a key RNaseIII-like enzyme termed Dicer, evolutionary conserved in different taxa.

Depending on their biogenesis and functions, sRNAs are classified as micro RNAs and small interfering RNA (Finnegan and Matzke, 2003).

siRNAs were first detected in plants in 1999 (Hamilton and Baulcombe, 1999). In general, siRNAs can be derived from all regions of perfect duplex RNAs and, at least in plants, they accumulate in both sense and antisense polarities. Perfect duplex RNAs can be synthetic RNAs, replicating viruses or even the result of the transcription of nuclear genes.

In plants, siRNAs have a variety of functions that can be grouped in at least two broad categories: those that trigger changes in the chromatin state of elements from which they derive and those that derive from and defend against exogenous RNA sequences such as viruses or sense transgene transcripts (Bonnet *et al.*, 2006).

miRNAs originate from specific endogenous genes called MIR genes, preferentially localized within intergenic regions. They derive via Dicer cleavage of imperfect duplex stem-loop RNAs, ~70-200 nt in length.

The biogenesis of plant miRNAs happens within specialized regions of the nucleus, called D-bodies, where RNA polymerase II mediate the production of a primary miRNA (pri-miRNA) transcript. The pri-miRNA is then processed into a shorter stem-loop precursor-miRNA (pre-miRNA) formed by base-paring between self-complementarity regions. Pre-miRNA is processed again, releasing a miRNA/miRNA* duplex, later exported to the cytoplasm, possibly by HASTY (HST), where the miRNA* is usually degraded and the mature miRNA is recruited by RNA-induced silencing complex (RISC) (Voinnet, 2009) (Figure 1.2).

Both siRNAs and miRNAs are loaded into a RISC, and associated with a member of the Argonaute protein family, which has RNA-binding ability.

Through this complex, siRNAs will then bind to the same messenger RNA from which they originate, and cleave the mRNA, silencing its expression. Whereas miRNAs will bind specifically to a target messenger RNA, and guide its cleavage (in most of the cases) or will repress its translation (Bonnet *et al.*, 2006) (Figure 1.1, 1.2).

Plant miRNAs were discovered for the first time in *Arabidopsis thaliana* in 2002. Some plant miRNAs were found to be conserved in many plant genomes such as those of *Oryza sativa*, *Zea mays* and those of more ancient vascular plant genera such as ferns or even nonvascular plants such as mosses (Bonnet *et al*., 2006). Surprisingly, a miRNA family, 854, has been shown to be expressed in *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Mus musculus* and *Homo sapiens*. Interestingly, across these diverse species, miRNA-854 targets the same mRNA, uridylate binding protein 1b (*UBP1b*), which normally encodes a member of a heterogeneous nuclear RNA binding protein-1 (hnRBP-1) gene family. This indicates an evolutionary common origin of miRNA-854 as a regulator of the basal eukaryotic transcription mechanism in both plants and animals for many hundreds of millions of years (Pogue *et al*., 2014). Nevertheless, also non-conserved miRNAs exist. It seems that plant genomes encode more non-conserved miRNA families than conserved miRNA families, such as those of *Arabidopsis* or those that control cotton fiber differentiation and elongation (Fahlgren *et al*., 2007; Rajagopalan *et al*., 2006; Zhang *et al*., 2006). The non-conserved plant miRNAs presumably emerged and dissipated in short evolutionary time scales (Sunkar and Jagadeeswaran, 2008).

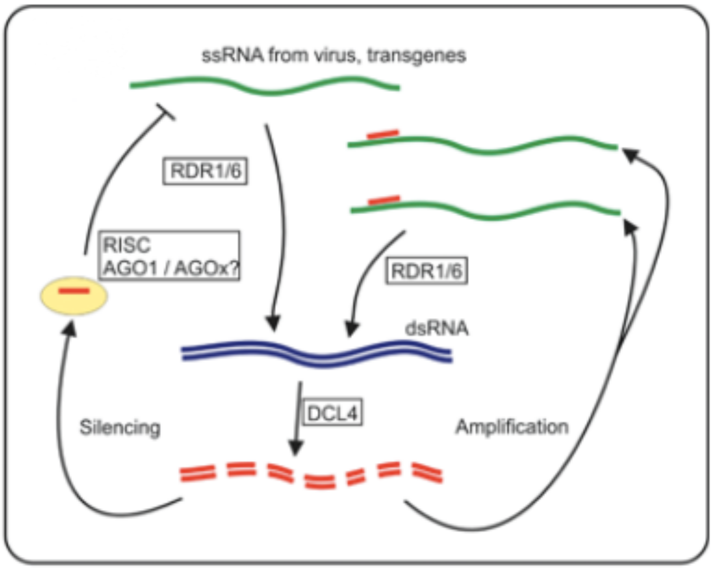


Figure 1.1 Small interfering RNAs (siRNAs). Long double-stranded RNAs (dsRNAs) from diverse origins (viruses, transpososons, transgenes, etc.) are converted into 21 nt long siRNAs by DICER enzymes. These small RNAs are then loaded into RISC and associated with AGO4 or another Argonaute protein. The complex will then bind to the same messenger RNA from which they originate, and cleave the mRNA, silencing its expression. Small interfering RNAs can also bind to the mRNA and initiate the transformation of single-stranded RNA (ssRNA) into dsRNA, thus amplifying siRNA production. RDR, RNA-dependent RNA polymerase. (Bonnet *et al*., 2006)

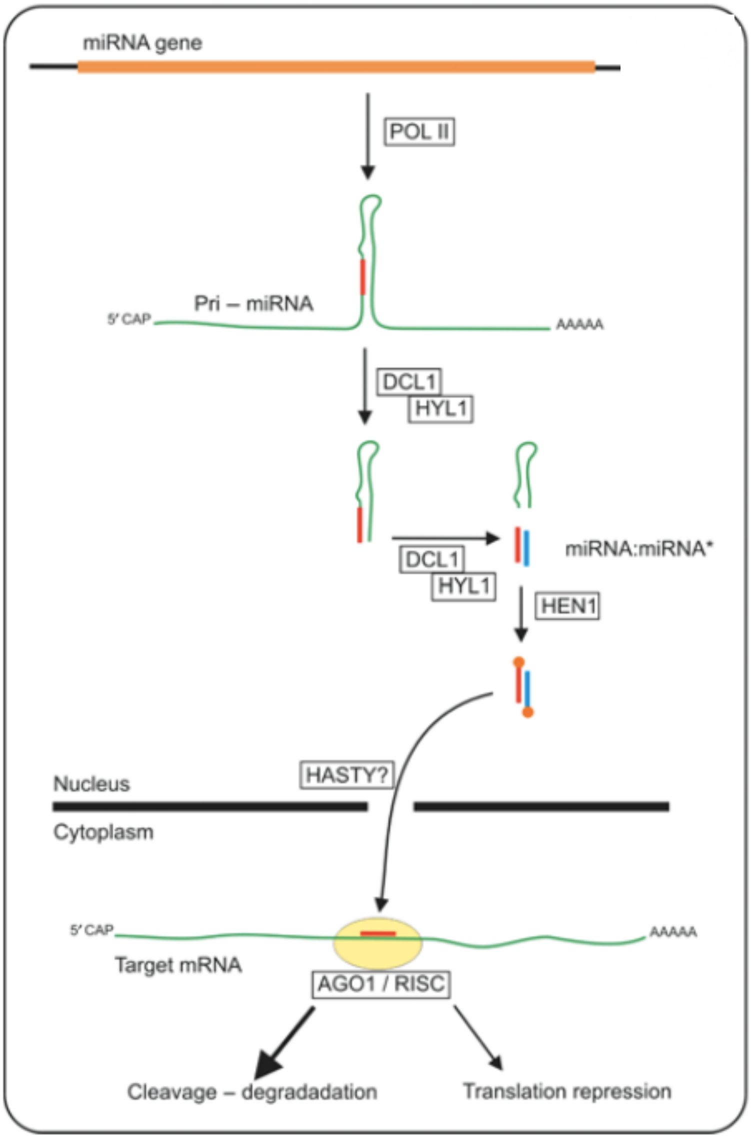


Figure 1.2 Plant microRNA (miRNA) biogenesis. MicroRNA genes are transcribed from their own locus by pol-II. The hairpin-like secondary structure is further processed by DICER in several steps to produce miRNA:miRNA* duplexes. The duplexes are then methylated by HEN1, before being exported to the cytoplasm, possibly by HASTY. Here the duplex is unwound and the miRNA is associated with AGO1. This complex, known as RISC, will bind specifically to a target messenger RNA, and guide its cleavage (in most of the cases) or will repress its translation. DCL, Dicer-like; HYL, HYPONASTIC LEAVES. (Bonnet *et al*., 2006).

1.3 Long non-coding RNAs

Small RNAs have been largely studied and are well known for their important roles in transcriptional and post-transcriptional regulation. Whereas, only recently, lncRNAs have been identified and characterized in several animal and plant species. In plants, first studies were conducted in *Arabidopsis thaliana* and published around 2011, whereas the fine annotation of lncRNAs in other plant spices, such as maize and rice, was reported only in 2014 and 2015 (Ding, *et al*., 2012a; Liu *et al*., 2012; Lin Li *et al*., 2014; Zhou *et al*., 2009; H. Wang *et al*., 2014a; M. Wang *et al*., 2015; B. Zhu *et al*., 2015; Derrien *et al*., 2012; Iyer *et al*., 2015). The long time employed in lncRNA study in plant kingdom is understandable if we think about the limited availability of sequenced and annotated genomes (Phytozome v.10.3 http://phytozome.jgi.doe.gov/).

LncRNAs are non-coding RNAs longer than 200 nucleotides that, unlike sRNAs, are able to act as regulatory RNA without being processed.

Initial studies about lncRNAs were conducted in animals and, in 90s, Xist (X-inactive specific transcript) was the first identified lncRNA (Brown *et al*., 1992).

Xist, a lncRNA of 17kb in mouse or 19kb in human, is one of the earliest examples that lncRNAs regulate gene transcription by modifying the chromatin status. In female mammals, one of the two copies of the X chromosome (Xi) is inactivated to maintain the same dosage of gene products as males. Xist, is a major effector in X chromosome silencing (Zhang *et al*., 2013b). It is specifically transcribed from and coats the Xi in somatic cells. Xist was recently shown to directly interact with Ezh2, the catalytic subunit of the Polycomb repressive complex 2 (PRC2), through a Repeat A motif, resulting in recruitment of PRC2 to the chromosome. PRC2 spreads across and silences genes along the Xi through catalysing the repressive trimethylation of lysine 27 on histone H3 (H3K27). During this process, Xist seems to alter the three-dimensional architecture of Xi to facilitate repositioning of active genes into the repressive compartment (Bergmann and Spector, 2014).

Another important lncRNA characterized in animals is MALAT1, originally identified amongst several genes up-regulated in metastatic non-small cell lung cancer (Ji *et al*., 2003). Recent studies identified mis-expression and mutations of MALAT1 in several other cancers, rendering MALAT1 a tumor marker with potential as a prognostic or therapeutic target (Gutschner *et al*., 2013).

The advent of RNA-Seq technology has largely stimulated lncRNAs annotation, in fact, the GENCODE consortium within the framework of the ENCODE project recently reported 9277 manually annotated genes producing 14880 transcripts (Derrien *et al*., 2012). A even more recent study about lncRNAs in human transcriptome revealed 58648 lncRNAs if which 79% were previously unannotated (Iyer *et al*., 2015).

Despite the large number of lncRNAs identified, there is a strong imbalance between the number of lncRNAs identified and those that have been functionally annotated, such as Xist, MALAT1, Air, KCNQ1ot1, HOTAIR, Frigidair, HOTTIP, PANDA, COOLAIR, COLDAIR, TERRA, DHFR-Minor and few others (Ma *et al*., 2012).

Compared with human and animals, the study of lncRNAs in plants is still in its infancy.

In 2007 the first lncRNA in plants was discovered, Induced by Phosphate Starvation 1 (IPS1), involved in phosphate uptake through target mimicry. IPS1 is a noncleavable lncRNA that forms a nonproductive interaction with the partially complementary miR-399, preventing it from cleaving its target, *PHO2* RNA, which negatively affects shoot Pi content and Pi remobilization (Franco-Zorrilla *et al*., 2007) (Figure 1.3C).

After that work, there are only few other examples of functionally characterized lncRNAs, such as COOLAIR, COLDAIR and LDMAR (Swiezewski *et al*., 2009; Heo and Sung, 2011; Ding *et al*., 2012a).

COOLAIR (cold induced antisense intragenic RNA) is a group of capped, polyadenylated and alternatively spliced lncRNAs. This group comprises cold induced antisense transcripts covering the entire *Flowering Locus C* (*FLC*), a master repressor of flowering in *Arabidopsis*. They react to cold earlier than Vernalization insensitive 3 (VIN3), the earliest factor in the polycomb silencing mechanism. COOLAIR is believed to negatively regulate *FLC* sense transcription in a polycomb-independent manner (Figure 1.3 A).

However, COOLAIR only transiently suppresses the *FLC*, and polycomb machinery is indispensable for the construction of epigenetic memory of *FLC* inactivation. It has been demonstrated that COOLAIR suppresses *FLC* through transcription interference, in particular promoter interference.

COLDAIR is another class of lncRNA derived from *FLC* locus. Unlike COOLAIR, COLDAIR is a sense transcript approximately 1100 nucleotides long with 5' cap but no polyA tail, these features could lead one to think that these lncRNAs are transcribed by RNA polymerase V and IV. Nevertheless, COLDAIR is transcribed by RNA polymerase II like many other lncRNAs in mammals. The COLDAIR expression is induced by cold exposure. COLDAIR specifically interact with CLF (Curly Leaf), a key component of PRC2 complex. Hence, it is very likely that COLDAIR negatively modulates *FLC* via a polycomb-dependent model (Figure 1.3 B). COLDAIR may play a role in the recruitment of PRC2 to *FLC* chromatin to trigger the epigenetic memory establishment of *FLC* silencing by vernalization (Zhang *et al*., 2013b).

LDMAR is a lncRNA that controls Photo-Sensitive Male Sterility (PSMS) in rice. Originated from an elite *japonica* rice variety Nongken 58N (NK58N), Nongken 58S (NK58S) was a spontaneous mutant exhibiting PSMS, *i.e.* its pollen becomes completely sterile when grown under long-day conditions, whereas the pollens are viable under short-day growth conditions. The PSMS in NK58N is caused by a C-to-G mutation in the LDMAR gene (Ding *et al*., 2012a). Probably, the C-to-G mutation altered the secondary structure of LDMAR in NK58S, and the structural alteration brought DNA methylation in the promoter region, which suppressed the LDMAR expression, and the insufficient LDMAR eventually led to the sterility of NK 58S under long-days (Figure 1.3 D). Later, Psi-LDMAR, a siRNA derived from the sense strand of LDMAR promoter region was also found to be responsible for the regulation of DNA methylation in this region (Ding *et al*., 2012b).

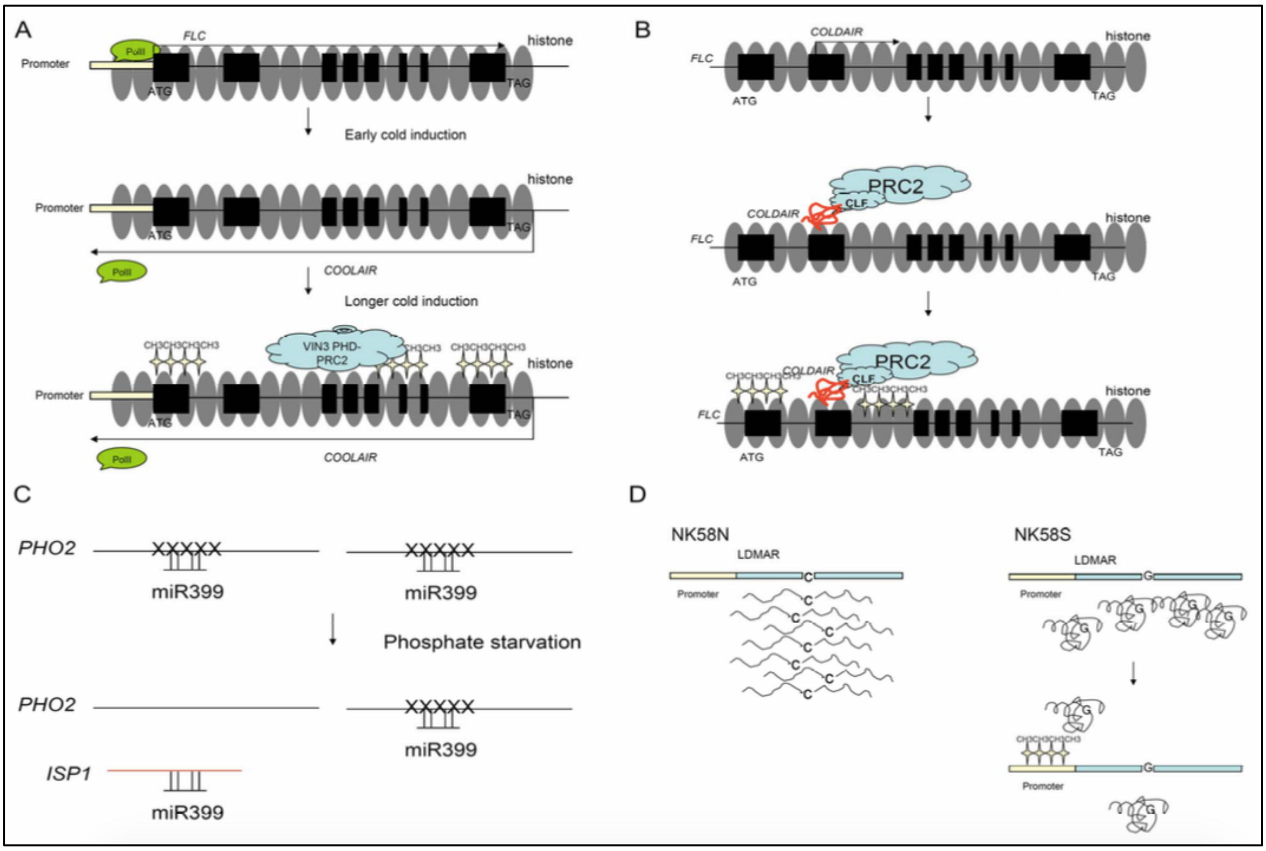
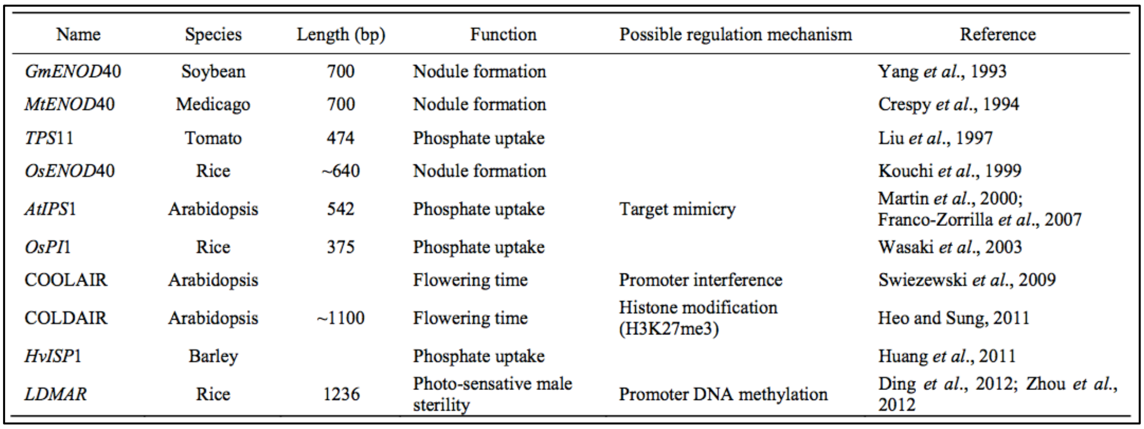


Figure 1.3 Schematic representation of four types of lncRNA regulation mechanism in plants. (A) COOLAIR regulates *FLC* in a transcription interference model. Top: *FLC* is transcribed by RNA Polymerase II before prolonged cold exposure; Middle: early cold treatment induced the expression of COOLAIR, which interferes with the RNA polymerase II binding to the *FLC* promoter, thus transiently suppresses the *FLC* expression; Bottom: after longer cold exposure, VIN3 recruits PRC2 complex to deposit H3K27me3 modification on *FLC* loci; (B) COLDAIR regulates *FLC* in a PRC2 associated histone modification model. Top: COLDAIR is induced by cold treatment. Middle: COLDAIR recruit PRC2 complex to the *FLC* loci; Bottom: PRC2 complex deposit H3K27me3 on the *FLC* loci; (C) *ISP1* regulates PHO2 in a target mimicry model. Top: under normal growth condition, miR399 specifically binds to PHO2 and degrades PHO2 mRNA; Bottom: *ISP1* competitively binds with miR399 to arrest its degradation function on PHO2; (D) LDMAR regulates the transcription of itself by a DNA methylation model. In NK58N, LDMAR is normally expressed. In NK58S, the C-to-G mutation altered the secondary structure of LDMAR and leads to the promoter DNA methylation, which reduced the LDMAR expression responsible for PSMS in NK58S. (Zhang *et al*., 2013b)

In addition to the lncRNAs described above, a few other lncRNAs from plant were also characterized, as reported in Table 1.1.



Name	Species	Length (bp)	Function	Possible regulation mechanism	Reference
*GmENOD*40	Soybean	700	Nodule formation		Yang *et al.*, 1993
*MtENOD*40	Medicago	700	Nodule formation		Crespy *et al.*, 1994
*TPS*11	Tomato	474	Phosphate uptake		Liu *et al.*, 1997
*OsENOD*40	Rice	~640	Nodule formation		Kouchi *et al.*, 1999
*AtIPS*1	Arabidopsis	542	Phosphate uptake	Target mimicry	Martin *et al.*, 2000; Franco-Zorrilla *et al.*, 2007
*OsPI*1	Rice	375	Phosphate uptake		Wasaki *et al.*, 2003
COOLAIR	Arabidopsis		Flowering time	Promoter interference	Swiezewski *et al.*, 2009
COLDAIR	Arabidopsis	~1100	Flowering time	Histone modification (H3K27me3)	Heo and Sung, 2011
*HvISP*1	Barley		Phosphate uptake		Huang *et al.*, 2011
LDMAR	Rice	1236	Photo-sensative male sterility	Promoter DNA methylation	Ding *et al.*, 2012; Zhou *et al.*, 2012

Table 1.1 Summary of the reported lncRNA genes in plants. (Zhang *et al*., 2013b)

Except this few examples of functionally characterized lncRNAs, other studies focused only on annotation of these molecules, in *Arabidopsis* (6480 lincRNAs and, more recently, 37238 lncNATs) (Liu, Wang, and Chua, 2015; H. Wang *et al*., 2014a), in rice (7142 lncNATs originating small RNAs) (Zhou *et al*., 2009), in maize (20163 putative lncRNAs, of which 1704 high-confidence lncRNAs) (Li *et al*., 2014), in cotton (50566 lincRNA and 5826 lncNAT transcripts) (Wang *et al*., 2015), in *Populus* (*P. trichocarpa* 2542 lincRNAs; *P. tomentosa* 1377 lncRNAs) (Chen *et al*., 2015; Shuai *et al*., 2014) and tomato (3679 lncRNAs) (Zhu *et al*., 2015). Recently, in rice, new lncRNAs have been identified, as competing endogenous RNAs (ceRNAs), which sequester miR160 or miR164 in a type of target mimicry, and one lncRNA, XLOC_057324, demonstrated to play a role in panicle development and fertility (Zhang *et al*., 2014).

LncRNAs are generally transcribed by RNA polymerase II, so they are always capped, polyadenylated and frequently spliced (Ulitsky and Bartel, 2013).

However many novel lncRNAs have been found to be transcribed by RNA polymerase III, which was previously thought to only transcribe housekeeping RNAs like tRNA and 5S RNA (Zhang *et al*., 2013b).

Some lncRNAs are generated by plant-specific polymerase V, capped at the 5’ end and lacking apparent poly(A) tails. These lncRNAs function as a scaffold for the RdDM pathway (Kim and Sung, 2012).

These features are important to choose the approach for lncRNAs identification, for example using RNA-Seq poly(A) libraries.

On the basis of their genomic origins, lncRNAs can be classified into:

- Long intergenic ncRNAs (lincRNAs)
- Intronic ncRNAs (incRNAs)
- Natural antisense transcripts (NATs), referred to the antisense transcripts of protein-coding transcripts. NATs are broadly grouped into two categories based on whether they act in *cis* or in *trans*. The so-called *cis*-NATs are transcribed from the same loci as sense transcripts and therefore have perfect match with the sense transcripts. On the contrary, *trans*-NATs are transcribed from different genomic loci and usually display only partial complementarity with the sense transcript (Zhang *et al.*, 2013b; Rinn and Chang, 2012).

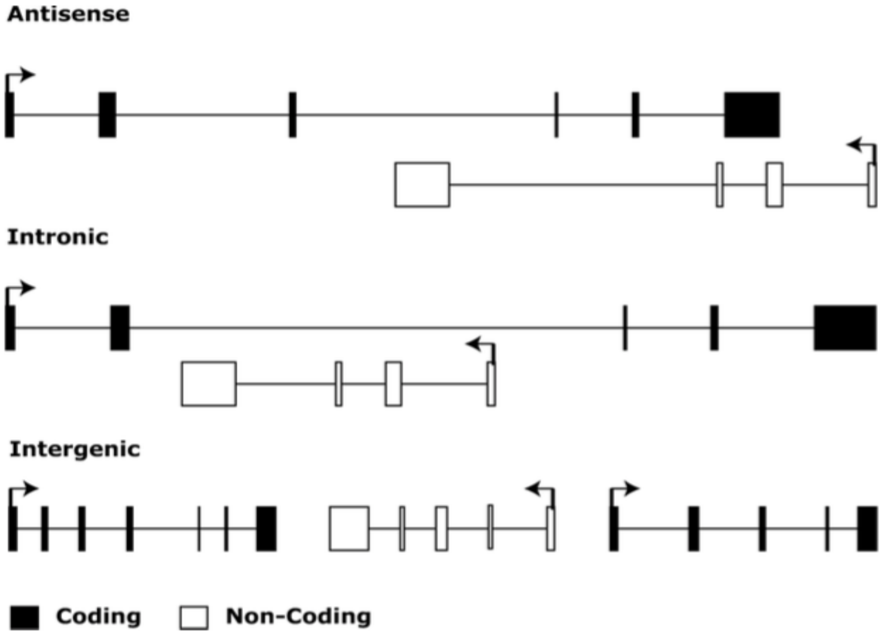


Figure 1.4 Anatomy of lncRNA loci. lncRNAs are often defined by their location relative to nearby protein coding genes. Antisense lncRNAs are lncRNAs that initiate inside of a protein coding gene and transcribe in the opposite direction that overlaps coding exons. Intronic lncRNAs are lncRNAs that initiate inside of an intron of a protein coding gene in either direction and terminates without overlapping exons. Intergenic lncRNAs are lncRNAs with separate transcriptional units from protein coding genes. (Rinn and Chang, 2012)

Compared with protein-coding genes and even small noncoding RNAs, most lncRNAs lack strong sequence conservation between species and they do not contain evolutionarily conserved long ORF. They are usually expressed at low levels, developmentally regulated and often exhibit tissue-specific and cell type-specific patterns. A significant proportion of lncRNAs are located exclusively in the nucleus with a few exceptions that are localized in cytosolic fractions (Wang *et al.*, 2015).

From the recent literature it's emerging that lncRNAs are potent regulators involved in several biological processes in eukaryotic cells. The greatest part of information about lncRNA function derives from animal world, in which the study of lncRNAs began and developed rapidly, especially thanks to the availability of sequenced genomes and the growing importance that lncRNAs have been found to have in many diseases (Brown *et al.*, 1992; Dong *et al.*, 2014; Engreitz *et al.*, 2013; Gao *et al.*, 2015; Li *et al.*, 2015; Shi *et al.*, 2015; Yao *et al.*, 2015).

In particular, it has been shown that lncRNAs can regulate gene expression at different levels: transcriptional, post-transcriptional and post-translational (Liu *et al.*, 2015). At transcriptional level, they can regulate the polymerase II transcription machinery in many ways (Figure 1.5), e.g. they can regulate the DNA-binding activity of transcription factor (TF) or can regulate mediator complex formation (Lai *et al.*, 2013). There is also a class of animal lncRNAs called "enhancer" RNAs, that are transcribed from the enhancer domains and/or transcription factor binding sites of genes, and they may regulate transcription activities of their flanking genes by recruiting transcription activators/repressors and/or controlling chromatin topology. One mode of action, typical of plant lncRNAs instead, is to trigger the formation of a stable RNA–DNA triplex so as to control TF binding specificity on promoter regions (Liu *et al.*, 2015) (Figure 1.5).

Furthermore, many lncRNAs have been shown to play a role in modifying chromatin marks and some can directly interact with chromatin modifiers, promoting or repressing transcription activity (Wang *et al.*, 2011).

In plants, lncRNAs activity is also correlated to small RNAs. Double-stranded RNAs could be processed into 21- to 24-nt small RNAs, which may initiate post-transcriptional or transcriptional gene silencing. LncRNAs and mRNAs can form double-stranded RNA duplexes with natural antisense transcripts to carry out their functions (Figure 1.6 A).

Concerning post-transcriptional regulation, a considerable number of repeat-containing lncRNAs (RC-lncRNA), originated from genes overlapping with transposable elements or repeat, can generate small RNAs which, in turn, can participate in post transcriptional gene silencing (PTGS) and/or RdDM pathways (Liu *et al.*, 2012). LncRNAs can also interact with another class of regulatory RNAs, microRNAs, through the so called target mimicry (Figure 1.6 B). It is a regulatory mechanism for miRNA functions in plants in which the decoy RNAs bind to miRNAs via partially complementary sequences and therefore block the interaction between miRNAs and their authentic targets. In this way, lncRNAs can act as sponges of miRNAs (Franco-Zorrilla *et al.*, 2007; Wu *et al.*, 2013).

In post-translational regulation, lncRNAs are able to regulate protein-protein interaction, interacting with RNA-binding proteins (RBPs) (Wang *et al.*, 2014b) (Figure 1.6 C). For instance, *Arabidopsis* genome encodes many genes encoding RBPs, which can potentially associate with lncRNAs to execute their functions, such as the regulation of ABA signalling pathway (Liu, Wang, and Chua, 2015). Also protein subcellular location can be regulated by lncRNAs (Figure 1.6 F). A case of a plant lncRNA altering the subcellular localization of a protein was reported some years ago, in *M. truncatula*, where the *endo40* RNA encodes only a short open reading frame and it colocalizes with RNA-Binding Protein 1 (MtRBP1). During nodule development, MtRBP1 together with *enod40* translocates from nuclear speckles to the cytosol in root cells, whereas in the cells without expressed *enod40*, MtRBP1 is retained in the nucleus (Campalans *et al.*, 2004). Recent studies found that several lncRNAs act as scaffolds (Figure 1.6 E), which have the capacity to bind some protein partners, serving as adaptors to form the functional protein complexes, e.g. TERRA and HOTAIR (Ma *et al.*, 2012).

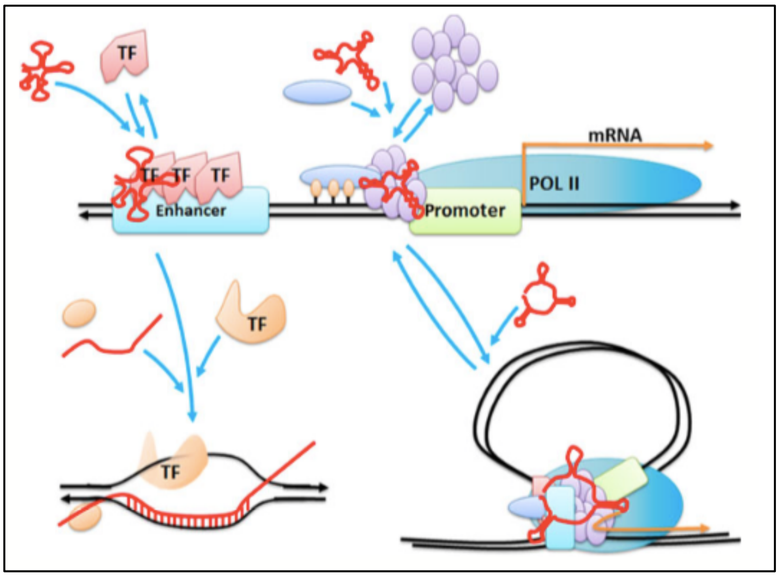


Figure 1.5 Schematic diagram of lncRNAs in transcriptional regulation. LncRNAs (red) can regulate transcription machinery and chromatin modification on promoter regions, and transcription-factor-binding affinity on enhancer regions in the nucleus. LncRNAs can also regulate RNA-DNA hybridization and chromatin topological structures. TF, transcription factors. (Liu, Wang, and Chua, 2015)

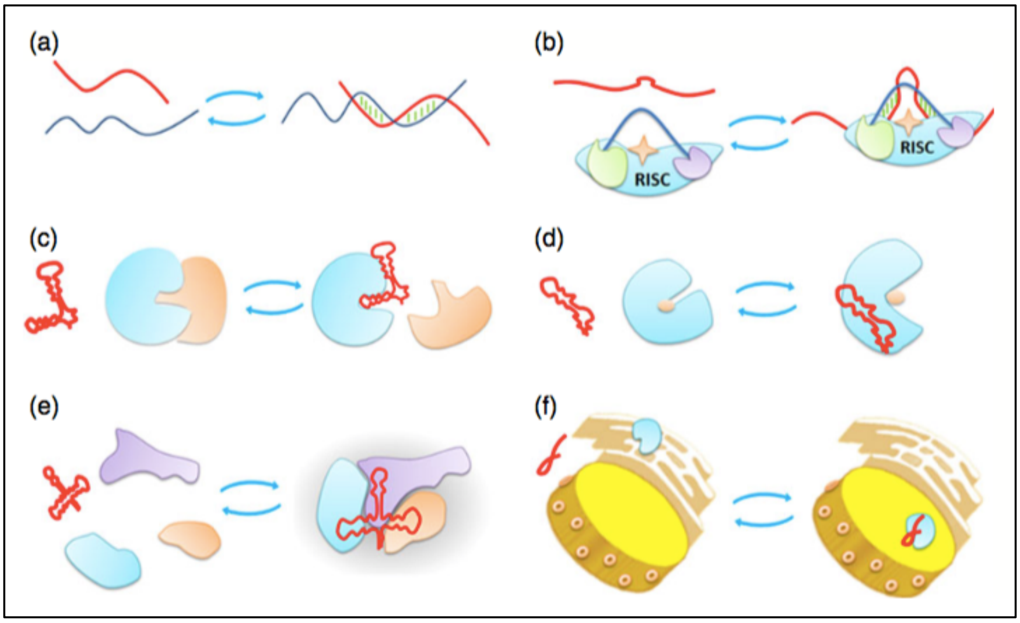


Figure 1.6 Schematic diagram of lncRNAs in post-transcriptional and post-translational regulation. (a) LncRNA (red) can form double-stranded RNA duplexes by complementary sequence to its targeted RNA. (b) lncRNA as a target mimic to down-regulate miRNA activity. RISC, RNA-induced silencing complex. (c) lncRNA regulates protein-protein interaction. (d) lncRNA changes protein structure to expose/protect specific amino acid for modification. (e) lncRNA as a scaffold to regulate assembly of protein complex subunits. (f) lncRNA guides RNA-binding protein relocation. (Liu, Wang, and Chua, 2015)

1.3.1 LncRNAs identification: genomics approaches

Over the years, different approaches in lncRNA study have been used, from microarrays onward to RNA-Seq. The identification of lncRNAs relies on the detection of transcription from genomic regions that are not annotated as protein coding (Fatica and Bozzoni, 2014).

Although traditional microarrays are incapable of identifying novel lncRNAs and are not able to distinguish different splicing variants, in past years, they were the first choice in many applications. In particular, since the identification of several lncRNA loci as part of the ENCODE project, the completeness of microarrays for human lncRNAs has been drastically improved. However, microarray is not sensitive enough to detect RNA transcripts with low-expression level. Thus the use of microarray to identify lncRNAs is limited due to the low expression level of many lncRNAs (Lee and Kikyo, 2012; Ma *et al.*, 2012).

As variation of traditional microarrays, DNA tiling arrays contain overlapping oligonucleotides that cover an entire length of a defined DNA region. A major advantage of using tiling arrays is their capacity to identify novel lncRNAs in a selected DNA region without prior knowledge of their precise locations within the region. This methodology allows the analysis of global transcription from specific genomic regions and was initially used for both identification and expression analysis of lncRNAs (Fatica and Bozzoni, 2014). Interesting discoveries have been achieved by employing tilling arrays. For instance, Rinn *et al.* (2007) focused on lncRNAs expressed in the region of the human HOX genes and compared skin fibroblasts isolated from different anatomical regions of the body. They printed 400000 probes of 50 bases in length with each probe overlapping the next one by 45 bases to cover all four human HOX gene clusters. Polyadenylated RNAs prepared from fibroblasts were then hybridized to the tiling arrays, resulting in the discovery of the lncRNA HOTAIR transcribed from an intergenic region within the HOXC cluster (Rinn *et al.*, 2007). A similar HOX tiling array was used to identify lncRNAs specifically expressed in metastatic breast carcinoma (Gupta *et al.*, 2010).

SAGE (serial analysis of gene expression) technology is based on the generation of short sequence tags of unbiased cDNA sequence by restriction enzymes. SAGE tags are concatenated before cloning and sequencing. This methodology allows both the quantification of transcripts throughout the transcriptome and the identification of new transcripts and it has been used and proved to be an efficient approach in studying lncRNAs. For example, Gibb *et al.* compiled 272 human SAGE libraries. By passing over 24 million tags they were able to generate lncRNA expression profiles in human normal and cancer tissues (Gibb *et al*., 2011). Lee *et al.* also used SAGE to identify potential lncRNA candidates in male germ cell (Lee *et al*., 2012). However, SAGE is much more expensive than microarray, therefore is not widely employed in large-scale studies (Fatica and Bozzoni, 2014; Ma *et al*., 2012).

CAGE (cap analysis of gene expression) is a technique similar to SAGE, however, unlike it, CAGE sequence tags are not unbiased but originate from the 5' end of a transcript. Therefore CAGE can be used to locate an exact transcription start site in the genome, in addition to quantifying expression level (Kawaji *et al*., 2006).
For instance, CAGE tags were used for experimental validation of the annotated start sites and expression level of lncRNAs in GENCODE project (Derrien *et al*., 2012).

To date, with the development of next generation sequencing (NGS) technology, the most widely used methodology to study lncRNAs is certainly RNA-Seq. Sequencing of transcriptomes by RNA-Seq is one of the most powerful methodologies for *de novo* discovery and expression analyses of lncRNAs. In this method, total RNA is converted to a cDNA library that is directly sequenced by high-throughput sequencing instruments. There are several types of sequencing technologies but Illumina platforms are currently the most commonly used for RNA-Seq experiments. A single sequencing run produces billions of reads that are subsequently aligned to a reference genome.
The basic workflow for lncRNA identification using RNA-Seq is shown in Figure 1.7 (Fatica and Bozzoni, 2014; Ma *et al*., 2012).

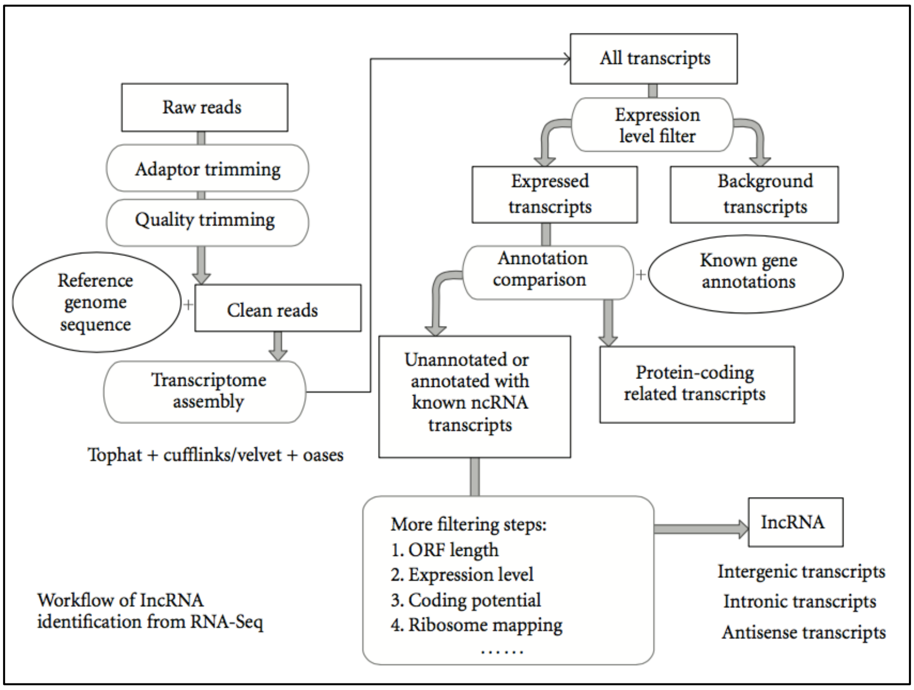


Figure 1.7 Workflow of lncRNA identification from RNA-Seq. (Ma *et al*., 2012)

Compared to traditional microarray technology, RNA-Seq has many advantages in studying gene expression. It is more sensitive in detecting less-abundant transcripts, and identifying novel alternative splicing isoforms and novel ncRNA transcripts.

RNA-Seq is able to detect transcripts that are missing or incomplete in the reference genome and allows for accurate quantification of expression levels, making it an ideal approach for lncRNA discovery. With an ultra sequencing depth RNA-Seq can be used to discover rare transcripts that are expressed in just a few cells within a tissue (Ma *et al*., 2012; Zhu and Wang, 2012).

There is also a method, chromatin immunoprecipitation (ChIP), that uses chromatin signatures, to study actively transcribed genes including lncRNAs. When combined with DNA sequencing (that is ChIP-Seq), this method can infer the genomic distribution of either proteins or histone modifications (Fatica and Bozzoni, 2014; Ma *et al*., 2012). Analysis of loci with specific histone modifications that characterize active transcription such as H3K4me3 (the marker of active promoters) and H3K36me3 (the marker of transcribed region), allowed an indirect identification of many unknown lncRNAs, For example, Guttman *et al*. identified 1600 large multiexonic lncRNAs that are regulated by key transcription factors such as p53 and NFkB (Guttman *et al*., 2009).

In plant genomes, initial analyses for lncRNA identification were based on the bioinformatics search for RNAs with poor coding capacity in cDNA databases. Through this approach, some lncRNAs were identified in *Arabidopsis* and *Medicago truncatula* (Ben Amor *et al*., 2009; Hirsch *et al*., 2006; Wen *et al*., 2007).

Also the analysis of expressed sequence tag (EST) databases helped in some lncRNAs identification, such as in wheat affected by *Puccinia striiformis*, suggesting their participation in pathogen-defense responses (Zhang *et al*., 2013a).

Just as in animal world, so in plants, tilling array provided a reach source for lncRNA discovery, for example in rice and *Arabidopsis* (Li *et al*., 2006; Matsui *et al*., 2010; Rehrauer *et al*., 2010).

Nevertheless, the greatest contribution in plant lncRNA study derives from RNA-Seq, especially for the annotation of these molecules (Liu, Wang, and Chua, 2015; H. Wang *et al*., 2014a; Zhou *et al*., 2009; Li *et al*., 2014; Wang *et al*., 2015; Zhu *et al*., 2015; Zhang *et al*., 2014).

2 Research objectives

The present project aims at performing, for the first time, a whole-genome annotation and a detailed analysis of lncRNAs in *Brachypodium distachyon* (Bd), a wild grass belonging to the *Pooideae* and an important model species for temperate cereals, such as wheat and barley.

The project takes advantages of public and proprietary RNA-Seq data sets from 15 different experiments conducted in the reference inbred line Bd21. Public RNA-Seq libraries consist of different plant tissues (see Table 3.1). Proprietary RNA-Seq libraries were previously produced in our lab from three developmental leaf areas: proliferation, expansion and mature, grown in control and drought stress conditions, considering three biological replicates for each sample.

The application of a modified *in silico* pipeline, described in Li *et al.* (2014), to the dataset provided a comprehensive systematic annotation of Bd lncRNAs.

For those libraries with biological replicates, a differential expression analysis was performed, with the aim of finding developmental, tissue-specific and cell specific lncRNAs.

Moreover the potential lncRNA target are investigated to highlight new regulatory networks and cross-talk between different RNA molecules. For instance, the algorithm developed by Wu *et al.* (2013) allow to investigate the link between lncRNAs and microRNAs through target mimicry, which is, as described above, a regulatory mechanism for miRNA functions in plants in which the decoy RNAs bind to miRNAs via complementary sequences and therefore block the interaction between miRNAs and their authentic targets.

3 Materials and methods

3.1 RNA-Seq libraries

A total of 15 RNA-Seq libraries generated from the reference inbred line Bd21, both public and proprietary data, were selected. Public RNA-Seq data were downloaded from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra). These samples were produced for the "Conserved Poaceae Specific Genes project" by Davidson *et al.* (2012) and comprise nine different plant tissue/organs grown in the greenhouse under long-day conditions (14-16 h light). As reported in Table 3.1, these libraries are poly(A^+) selected, single end (SE), generated with Illumina technology.

Proprietary libraries originated from a drought stress experiment conducted in January 2011 in Milano (Italy), in which *Brachypodium distachyon* inbred line 21 was grown under control and severe drought stress conditions, according to the protocol described in Verelst *et al.* (2013). In particular, plants were grown into a growth chamber, with controlled conditions of light, temperature and humidity. Control plants were maintained at 1.82 g of water per g dry soil, while stressed plants were dried down to 0.45 g water/g dry soil (severe drought stress).

All the samples were collected from the third leaf, about 24 hours after the emergence and each leaf was divided in 3 developmental zones: proliferation, expansion and mature zone.

RNA-Seq libraries were generated from each developmental zone for a total of 18 samples: three types of leaf cells (proliferation cells, expansion cells and mature cells) grown in control and drought condition, considering three biological replicates. For the analyses of the present thesis work, the biological replicates were merged in a single file, making a total of 6 libraries (proliferation, expansion, and mature cells in control and stress conditions).

As reported in Table 3.1, these libraries are poly(A^+) selected, single end (SE), generated with Illumina technology.

Samples	SRA accession	Raw reads	Reads length	Experiment	Project	References
Leaves 20 DAS	SRR349785 SRR352143	56673394	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Emerging inflorescence	SRR349787	23137165	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Early inflorescence	SRR349786	17601221	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Anther	SRR352140	26059840	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Pistil	SRR352137	17712829	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Seed 5 DAP	SRR352139	25922555	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Seed 10 DAP	SRR352141	25766517	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Embryo 25 DAP	SRR352138 SRR352144	48681058	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
Endosperm 25 DAP	SRR352142	27365511	35 SE	Poly (A) RNA-Seq	Conserved Poaceae Specific Genes Project	Davidson *et al.*, 2012
#3 leaf Proliferation ctrl (Pc)	-	76248037	50 SE	Poly (A) RNA-Seq	-	Bertolini *et al.* Unpublished
#3 leaf Proliferation stress (Ps)	-	82616294	50 SE	Poly (A) RNA-Seq	-	Bertolini *et al.* Unpublished
#3 leaf Expansion ctrl (Ec)	-	77830217	50 SE	Poly (A) RNA-Seq	-	Bertolini *et al.* Unpublished
#3 leaf Expansion stress (Es)	-	72602295	50 SE	Poly (A) RNA-Seq	-	Bertolini *et al.* Unpublished
#3 leaf Mature ctrl	-	58589296	50 SE	Poly (A) RNA-Seq	-	Bertolini *et al.* Unpublished
#3 leaf Mature stress	-	68671998	50 SE	Poly (A) RNA-Seq	-	Bertolini *et al.* Unpublished

Table 3.1 Summary of the main characteristics of the public (first 9) and the proprietary (last 6) libraries. DAS = Days after sowing; DAP = Days after pollination.

3.2 Bioinformatic analysis

First, a quality check of all the samples was assessed using FastQC tool (Andrews, 2015), than the mRNA reads were trimmed using ERNE-FILTER (v.1.3) (Del Fabbro *et al.*, 2013). For libraries containing adapter contaminants, Cutadapt (v.1.2.1) (Martin, 2011) was applied. ERNE-FILTER was used to remove low quality bases from the ends of the reads. Due to the different characteristics of public and proprietary libraries, different parameters of ERNE-FILTER were applied for the two data sets: a minimum PHRED score of 30 and a minimum read length of 35 bp.

After the quality and trimming filters, each experiment was aligned independently to the reference Bd21 genome (v.2.1), using a method of two mapping iterations, with the spliced read aligner TopHat2 version 2.0.9 (Trapnell *et al.*, 2012). This method was first proposed by Cabili *et al.* (2011) and it was used also in the lncRNA identification in maize (Li *et al.*, 2014). This approach allows maximizing the mapping efficiency by using the spliced site information derived from the first mapping iteration in all the samples.

Hence, after the first alignment, each experiment was re-aligned for a second time using the pooled splice sites file, considering the following parameters:

- Maximum number of mismatches = 0
- Minimum intron length = 10
- Maximum intron length = 500000
- Library type = fr-firststrand

Successively, for each experiment the transcriptome was *de novo* assembled using Cufflinks version 2.0.9 (Trapnell *et al.*, 2010).

All the transcripts, assembled with Cufflinks, were then merged with Cuffmerge to remove all the redundant transcripts and obtain only unique sequences. Finally, Gffread was used to generate a multi FASTA file. This FASTA file, containing both novel transcripts and previously annotated transcripts, was used to identify *bona fide* lncRNAs.

3.3 LncRNAs identification

To identify lncRNAs a modified bioinformatic pipeline “LncRNA_Finder”, described in Li *et al.* (2014), was used.

Particularly this pipeline applies five filters on the basis of the main lncRNA characteristics:

- *Size selection*: minimum lncRNA length was set at 200 nucleotides.
- *Open Reading Frame (ORF) filter*: maximum potential ORF length was kept as default at 100.
- *Known protein domain filter*: transcripts were aligned to a protein database of all Angiosperms downloaded from Plaza v.2.5 (http://plaza.psb.ugent.be/), the e-value parameter was set at 0.001.
- *Coding Potential Calculator (CPC)*: this parameter was used to evaluate the quality, completeness and homology of ORFs.
- *Filter of housekeeping RNAs and precursors of small RNAs*: since putative lncRNAs may contain precursors of housekeeping RNAs and small RNAs, filtered transcripts were compared to housekeeping RNAs, from Pfam databases (http://pfam.xfam.org/), and to public and proprietary small RNAs databases, 19 libraries from http://mpss.udel.edu/ and 8 libraries from Bertolini *et al*. (2013) from different tissue and cell types. The number of mismatches in the alignment with small RNA was kept as default at 0.

After each filter an output file, that can ben checked, is obtained.

At the end, putative lncRNAs with sequence similarity with small RNAs were classified as pre-lncRNAs, whereas transcripts that do not have similarity with any class of non-coding RNAs were defined as High Confidence lncRNAs (HC-lncRNAs).

3.4 Expression profiles and differential expression analysis

Considering the wide variety of libraries present in this study, it is interesting to see how lncRNAs expression changes across different stages, tissues, organs or conditions.

For this purpose, for all the libraries, even those with no replicates (that cannot be the subject of differential expression analysis), lncRNAs expression was expressed in

Reads Per Kilobase per Million mapped reads (RPKM), calculated using the CPM function of edgeR, a Bioconductor package (Robinson *et al.*, 2010).

Proprietary libraries, from different developing areas of the third growing leaf, were used to conduct a differential expression analysis. The R package DESeq2, based on the negative binomial distribution, which makes possible the evaluation of the raw variance of data from experimental design with small numbers of biological replicates, was applied (Love *et al.*, 2014).

To identify differentially expressed loci between drought and control conditions in the same developing zone and during leaf development in the proliferating and expansion conditions, a series of pairwise comparisons were set up, as reported in Table 3.2.

	Comparison
Drought stress	Ps vs Pc
	Es vs Ec
Leaf development	Pc vs Ec
	Ps vs Es

Table 3.2 List of pairwise comparison performed during the differential expression analysis. (P) proliferation, (E) expansion, (C) control and (S) drought stress conditions.

3.5 Interaction with microRNAs (target mimicry)

One of the most interesting biological function of lncRNAs, recently reported in plants (Franco-Zorrilla *et al.*, 2007; Wu *et al.*, 2013), is the action as endogenous target mimics (eTMs) of miRNAs, or target mimicry, via partially complementary sequences, blocking the interaction between miRNAs and their authentic targets (Wu *et al.*, 2013).

To identify HC-lncRNAs with potential target mimicry properties, the computational method described in Wu *et al.* (2013) was used. Briefly, a base-pairing interaction between Bd miRNAs and the annotated lncRNAs was predicted, based on: (1) bulges were only permitted at the 5' end 9th to 12th positions of miRNA sequence; (2) the bulge in eTMs should be composed of only three nucleotides; (3) perfect nucleotide pairing was required at the 5' end from second to eighth positions of miRNA sequence; (4) except for the central bulge, the total mismatches and G/U pairs allowed was set to 3.

3.6 RNA extraction and reverse transcription

To validate the identified HC-lncRNAs, it is important to conduct a Real-Time PCR of some putative lncRNAs. Therefore, cDNA of some of the datasets is needed.

So, RNA from proliferation and expansion zone, both in control and stress conditions, was extracted with the Plant/Fungi Total RNA Purification Kit (Norgen Biotek Corporation). The samples were flash-frozen in liquid nitrogen and conserved in a -80°C freezer. They were then grinded rapidly with mortar and pestle, using liquid nitrogen to ensure that the integrity of the RNA was not compromised. For the extraction, a maximum of 50 mg of starting material was needed. Briefly, the process involves a column with a resin that can bind RNA specifically, then washed with a provided wash solution and eluted with 80 µl of DEPC water.

The purified RNA samples were then retrotranscribed, starting from 1 µg of material, using iScript™ cDNA Synthesis Kit (Bio-Rad), following the manufactures' protocol.

3.7 LncRNA validation using Real-Time PCR

To validate the expression patterns of lncRNAs, Real-Time PCR experiments will be conducted in the near future. 10 lncRNAs have been selected among the up-regulated and down-regulated lncRNAs in control and stress condition between Pc/Ec and Ps/Es, as result of the differential expression analysis (Table 3.3).

Primers have also already been designed with Primer3 version 4.0.0 (Table 3.4).

LncRNA	Length	log2FoldChange	Expression
TCONS_00090190	358	-2.391265344	CTRL_down
TCONS_00089035	486	-2.255784059	CTRL_down
TCONS_00088496	2371	-3.057039273	CTRL_down
TCONS_00081033	1199	-2.894632964	CTRL_down
TCONS_00020495	358	-3.217483298	CTRL_down
TCONS_00089348	301	3.139134476	CTRL_up
TCONS_00086597	533	3.323773011	CTRL_up
TCONS_00030330	1749	2.971583098	CTRL_up
TCONS_00009992	1457	3.656548937	CTRL_up
TCONS_00019564	1512	-4.318542051	STRESS_down

Table 3.3 List of lncRNAs selected for Real-Time PCR validation. log2FoldChange indicates lncRNAs expression values generated with DESeq2.

Primers	Sequence (5’-3’)	Length (nt)	Tm (°C)	LncRNA
Forward	GGAACCAGAGAAGAAGAGGAGG	22	59.50	**TCONS_00090190**
Reverse	ATAGACAGGGAAGAGCTTGGAC	22	59.23	
Forward	AGTGGTAGAGTGGAGTGGAGT	21	59.57	**TCONS_00089035**
Reverse	GCGTCGTATATGTTGTCAGCG	21	59.81	
Forward	TGGATCGGTGTAGAATCGACC	21	59.32	**TCONS_00088496**
Reverse	TGATCTCCTACCAATCTGCGG	21	59.31	
Forward	CAAGCACTGAGGAATCTGACG	21	59.00	**TCONS_00081033**
Reverse	TGAGGAGTGTATGCCAACTCG	21	59.80	
Forward	ATCCAGTGACCTACAGCTGC	20	59.46	**TCONS_00020495**
Reverse	GGACCACGCATGTCACTAGT	20	59.75	
Forward	CCTACGCAACCCAAGCTATCT	21	59.86	**TCONS_00089348**
Reverse	CAACCGACCCAGTATAAACGC	21	59.34	
Forward	CATTGACACTGTCCTGGGTTTC	22	59.45	**TCONS_00086597**
Reverse	TACCCTAAAAGATACTCCGGCC	22	59.03	
Forward	GGGGTTCTTAGTCTTCGGGTT	21	59.37	**TCONS_00030330**
Reverse	TCCTCGGTTATTCACAGCTCC	21	59.52	
Forward	GCCATCCCATACTTCCTAGCC	21	60.00	**TCONS_00009992**
Reverse	CAAACAATCCACCCGCTTCTC	21	59.80	
Forward	CCGTTGTTTTGGTAGCTGCAA	21	59.93	**TCONS_00019564**
Reverse	CAAGGATTTGTCGGTGCGTAC	21	59.87	

Table 3.4 List of primers generated for the Real-Time PCR experiment.

4 Results

4.1 Experimental data set

Initial dataset for lncRNAs identification in *Brachypodium distachyon* was created selecting a total of 15 RNA-Seq libraries generated from the reference inbred line Bd21. Nine libraries were public and produced for the "Conserved Poaceae Specific Genes project" by Davidson *et al.* (2012), originating from nine different plant tissues/organs: leaves, early and emerging inflorescence, anther, pistil, seeds 5 and 10 days after pollination, embryo and endosperm. These libraries were downloaded from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra).

Six proprietary libraries originated from a drought stress experiment conducted on *Brachypodium distachyon* in January 2011 in Milano (Italy). RNA-Seq libraries were generated from three developmental zones of the third leaf collected about 24 hours after the emergence: proliferation, expansion and mature zone. A total of 18 samples were available: three types of leaf cells (proliferation cells, expansion cells and mature cells) grown in control and drought condition, considering three biological replicates. For the analyses of the present thesis work, the biological replicates were merged in a single file, making a total of 6 libraries (proliferation, expansion, and mature cells in control and stress conditions).

The libraries are poly(A^+) selected, single end (SE), generated with Illumina technology, making a total of 705478227 reads.

4.2 Quality check

Each library was subject to quality control that allows measuring the quality based on ten metrics: (1) Per base sequence quality; (2) Per sequence quality scores; (3) Per base sequence content; (4) Per base GC content; (5) Per sequence GC content; (6) Per base N content; (7) Sequence length distribution; (8) Sequence duplication levels; (9) Overrapresented sequences; (10) Kmer content. Some examples are reported in Figure 4.1.

Public and proprietary libraries had 35 and 50 bp long reads, respectively. Both had an average PHRED score of 30. The GC content was around 50% for all the libraries. Only in some libraries (SRR349786, SRR349787, SRR352137, SRR352138, SRR352142, SRR352143 and Mature ctrl) overrepresented sequences were found, referred to adapter contaminants.

Despite the high quality of the raw libraries, a quality trimming step was performed to remove the few Ns and the low quality nucleotide in the reads, as well as the sequencing adapters present in the sequence reads, with the aim to obtain a better alignment to the reference genome. Quality of trimmed reads was assessed, showing an improved quality in terms of PHRED (Q) scores: public and proprietary libraries had Q score greater than 30 and 33, respectively (Figure 4.2). Statistic of reads trimming is shown in Table 4.1.

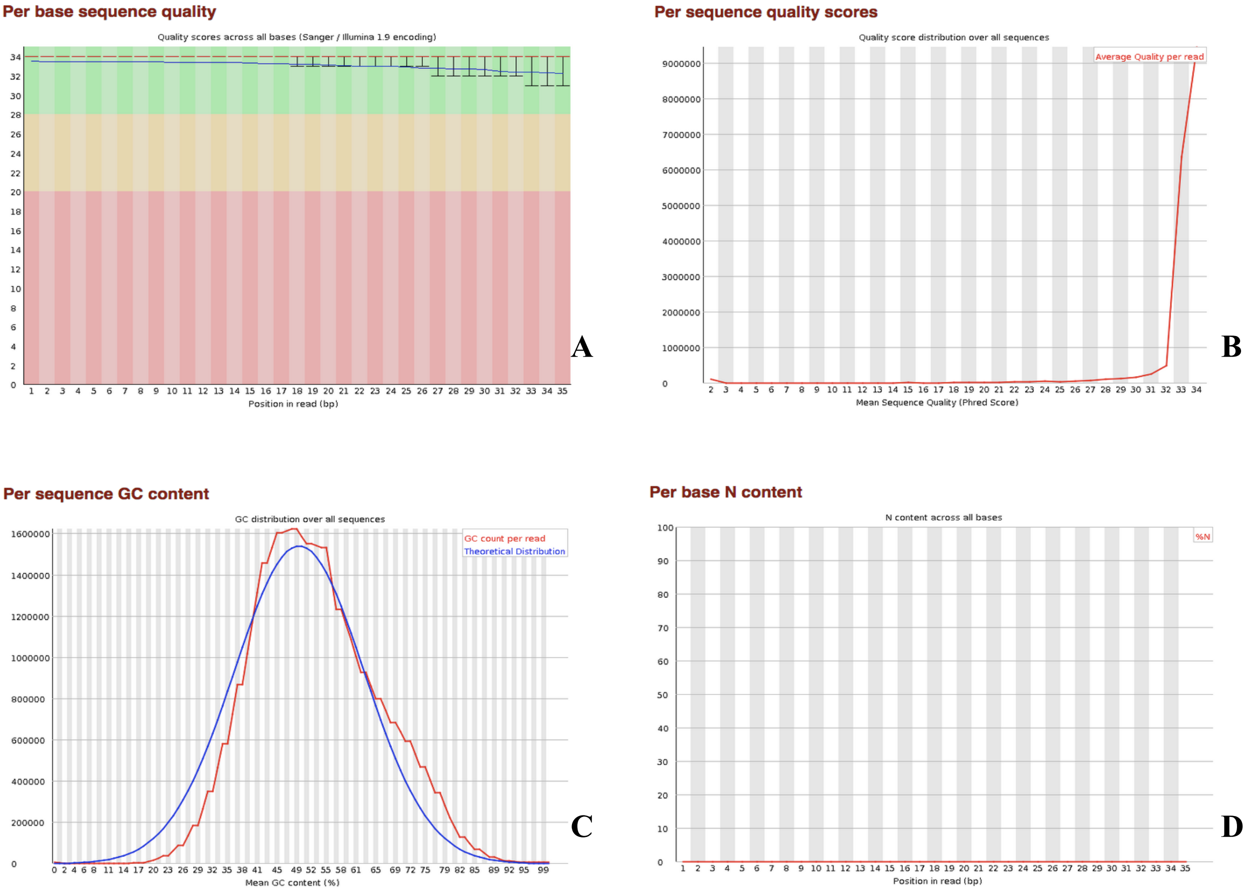


Figure 4.1 Visual output from FastQC of the library SRR352137. **A** Per base sequence quality plot. X-axis shows the base positions in the reads, Y-axis shows quality score (Q score). **B** Per sequence quality scores. Mean sequence quality are shown on x-axis. **C** Per sequence GC content. **D** Per base N content.

Samples	SRA accession	Raw reads	Reads trimmed	% reads trimmed
Leaves 20 DAS	SRR349785 SRR352143	56673394	53118138	93.73
Emerging inflorescence	SRR349787	23137165	21880389	95.29
Early inflorescence	SRR349786	17601221	16771555	94.57
Anther	SRR352140	26059840	23226658	89.13
Pistil	SRR352137	17712829	16739912	94.51
Seed 5 DAP	SRR352139	25922555	22737023	87.71
Seed 10 DAP	SRR352141	25766517	24540576	95.24
Embryo 25 DAP	SRR352138 SRR352144	48681058	44986617	92.41
Endosperm 25 DAP	SRR352142	27365511	24533458	89.65
#3 leaf Proliferation ctrl (Pc)	-	76248037	70970362	93.08
#3 leaf Proliferation stress (Ps)	-	82616294	77234409	93.49
#3 leaf Expansion ctrl (Ec)	-	77830217	71996832	92.50
#3 leaf Expansion stress (Es)	-	72602295	67857125	93.46
#3 leaf Mature ctrl	-	58589296	54766323	93.47
#3 leaf Mature stress	-	68671998	63660599	92.70
	Total	705478227	650613914	92.22

Table 4.1 Statistic of trimmed reads in each library.

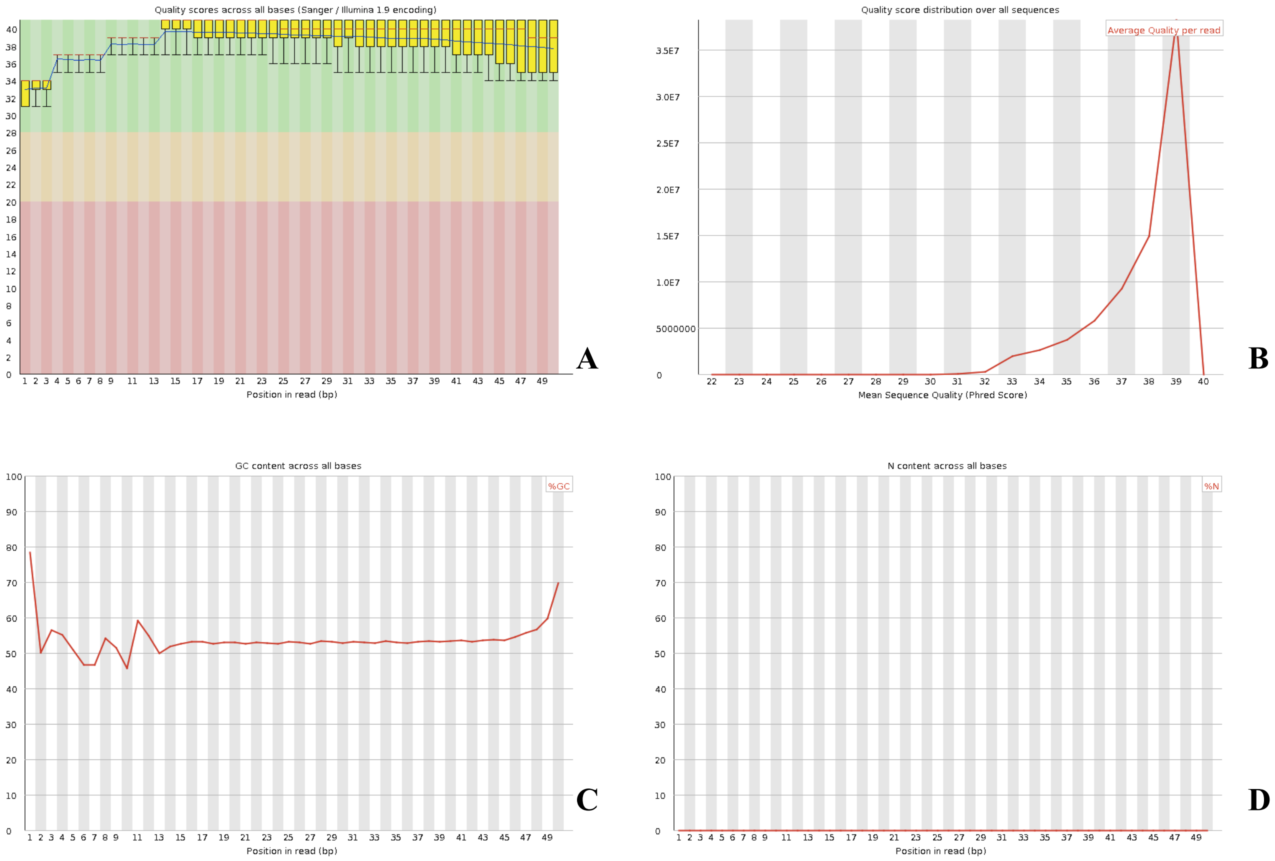


Figure 4.2 Visual output from FastQC of the trimmed library Ps. **A** Per base sequence quality plot. X-axis shows the base positions in the reads, Y-axis shows quality score (Q score). **B** Per sequence quality scores. Mean sequence quality are shown on x-axis. **C** Per sequence GC content. **D** Per base N content.

4.3 Alignment to the reference genome

Each library was independently mapped against Bd21 reference genome version 2.1, downloaded from Phytozome v.10.3 (http://phytozome.jgi.doe.gov/), using TopHat2 (Trapnell *et al*., 2012). This aligner program was designed to map reads from RNA-Seq experiments to a reference genome, allowing to identify exon-exon splice junctions. It is built on the ultrafast unspliced short read mapping program Bowtie, based on the exon-first approach (Langmead *et al*., 2009). Briefly, TopHat first maps reads to the genome, identifying potential exons using Bowtie. All reads that do not map to the genome are set aside as "Initially Unmapped Reads" (IUM reads). In this way TopHat builds a database of possible splice junctions and then maps the IUM reads against the junctions with a seed-and-extend strategy (Figure 4.3). This approach allows revealing new alternative spliced transcripts and isoforms and increase the overall percentage of mapped reads (Trapnell *et al*., 2009).

To optimize the alignment, the Bd21 annotation file of genes and exons (GFF3), version 2.1, available from Phytozome v.10.3 (http://phytozome.jgi.doe.gov/) was used. This new annotation file contains an improved annotation with a greater number of annotated genes and exons.

To maximize the use of splice site information derived from all samples, two iterations method first proposed by Cabili *et al*. (2011) was used. Therefore, after the first alignment, each experiment was re-aligned to the reference genome providing the sorted and non-redundant pooled splice sites file to the program.

Statistics of two iterations reads alignment are shown in Table 4.2. On average, the 96% of the reads was mapped to the reference genome.

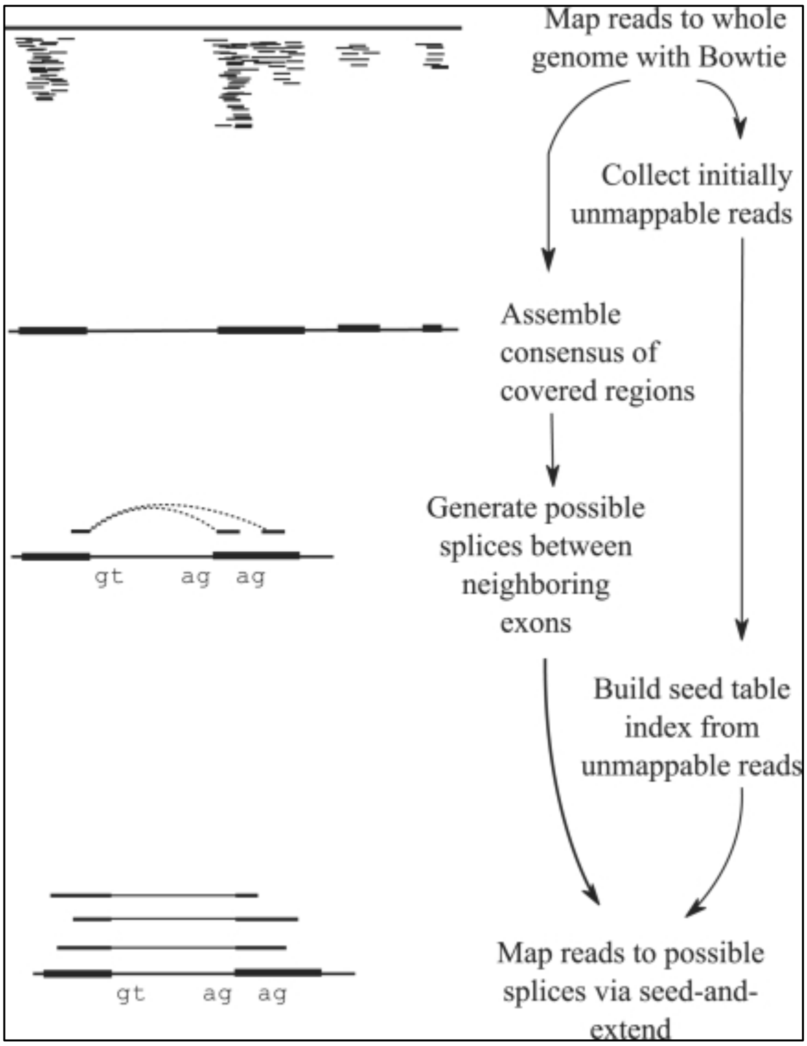

Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt
ag
ag
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt
ag
ag

Figure 4.3 The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. The IUM reads are indexed and aligned to these splice junction sequences. (Trapnell *et al*., 2009)

Samples	SRA accession	Reads trimmed	Mapped reads	% mapped reads	Unmapped reads	% unmapped reads
Leaves 20 DAS	SRR349785 SRR352143	53118138	51605492	97.2	1512646	2.8
Emerging inflorescence	SRR349787	21880389	21315167	97.4	565222	2.6
Early inflorescence	SRR349786	16771555	16191851	96.5	579704	3.5
Anther	SRR352140	23226658	22380996	96.4	845662	3.6
Pistil	SRR352137	16739912	15759963	94.1	979949	5.9
Seed 5 DAP	SRR352139	22737023	22094542	97.2	642481	2.8
Seed 10 DAP	SRR352141	24540576	21938177	89.4	2602399	10.6
Embryo 25 DAP	SRR352138 SRR352144	44986617	42581740	94.7	2404877	5.3
Endosperm 25 DAP	SRR352142	24533458	23314366	95.0	1219092	5.0
#3 leaf Proliferation ctrl (Pc)	-	70970362	68887254	97.1	2083108	2.9
#3 leaf Proliferation stress (Ps)	-	77234409	75050280	97.2	2184129	2.8
#3 leaf Expansion ctrl (Ec)	-	71996832	70274662	97.6	1722170	2.4
#3 leaf Expansion stress (Es)	-	67857125	65991118	97.3	1866007	2.7
#3 leaf Mature ctrl	-	54766323	53460628	97.6	1305695	2.4
#3 leaf Mature stress	-	63660599	62116512	97.6	1544087	2.4
	Total	650613914	632962748	97.3	22057228	3.4

Table 4.2 Statistic of reads alignment to the reference genome. The number of aligned reads is referred to the second alignment, with the sorted and non-redundant pooled splice sites file.

Successively, for each experiment the transcriptome was de novo assembled. The redundant transcripts were removed to select only unique sequences, obtaining a whole set of 99141 loci/isoforms.

4.4 Identified lncRNAs

From the 15 libraries analysed, a whole set of 99141 loci/isoforms was obtained. Among these loci/isoforms, known transcripts were discarded, obtaining 6590 new transcripts (Figure 4.5).

To distinguish lncRNA candidates, five sequential stringent filters to the 6590 transcripts were employed, according to the modified bioinformatic pipeline described by Li *et al.* (2014). First, transcripts shorter than 200 nucleotides and with ORFs longer than 100 amino acids were discarded and 3874 transcripts were retained (Figure 4.5). Transcripts passing this selection were aligned to protein database of all Angiosperms, downloaded from Plaza v.2.5 (http://plaza.psb.ugent.be/) to eliminate transcripts encoding protein. Next, the CPC was used to assess the protein-coding potential in order to eliminate other possible coding transcripts, and 2516 transcripts were obtained (Figure 4.5). After employing these criteria, the 2516 transcripts were considered as putative lncRNAs. Since putative lncRNAs may contain precursors of housekeeping RNAs and small RNAs, filtered transcripts were compared to a database of housekeeping RNAs and to a small RNAs databases (for details see the Materials and methods). Thus, putative lncRNAs with sequence similarity with small RNAs were classified as pre-lncRNAs, whereas the 2507 transcripts that do not have similarity with any class of non-coding RNAs were defined as High Confidence Bd lncRNAs (HC-lncRNAs) (Figure 4.5).

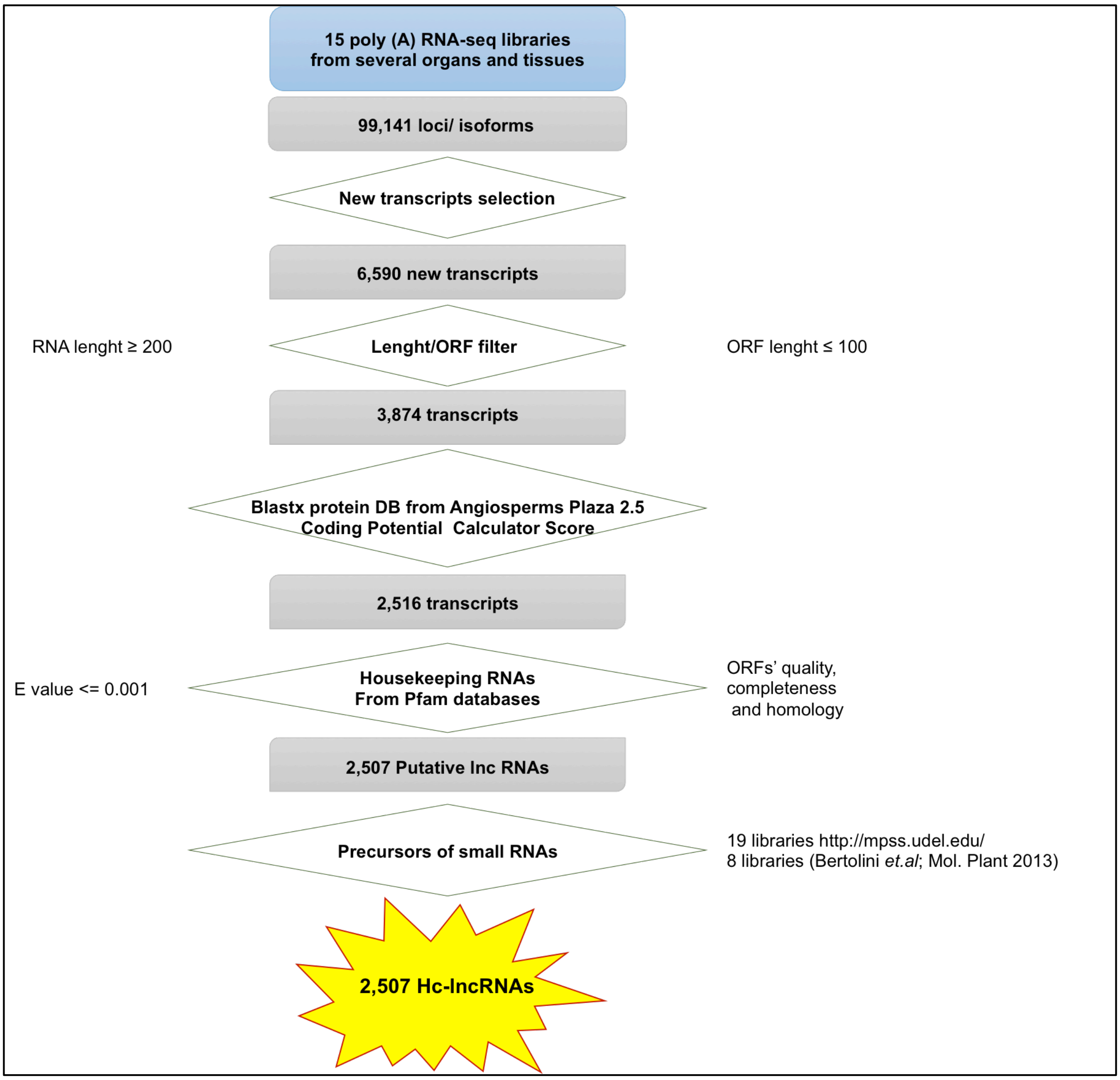


Figure 4.5 Detailed schematic diagram of the bioinformatic pipeline, described in Li *et al.* (2014), for identification of Bd lncRNAs. New transcripts were filtered with five criteria for identification of lncRNAs. (1) length ≥200 nucleotides and ORF ≤100 amino acids; (2) not encoding known protein domains; (3) little coding potential; (4) not housekeeping ncRNAs; and (5) not small RNA precursors.

Then the main features of identified HC-lncRNAs were examined, considering average length, %GC, number of exons and genome distribution (Figure 4.6).

Similarly to lncRNAs of other plants, such as tomato and maize (Li *et al*., 2014; Zhu *et al*., 2015), the greatest part of HC-lncRNAs (~70%) are shorter than 1000 nucleotides, with a median length of 300 nucleotides (Figure 4.6 A). Whereas, Bd lncRNAs are shorter than rice and human lncRNA transcripts (Derrien *et al*., 2012; Zhang *et al*., 2014). In accordance with previous studies, which have shown that both plant and animal lncRNAs are shorter and harbour fewer exons than protein-coding genes (Ding *et al*., 2012a; Li *et al*., 2014; Zhu *et al*., 2015), most of the genes encoding Bd lncRNAs only contained one exon (Figure 4.6 B).

A Circular plot clearly showed that Bd lncRNAs were pervasively transcribed across all the genome (Figure 4.6 D), similarly to the distribution observed in *S. lycopersicum* (Zhu *et al*., 2015).

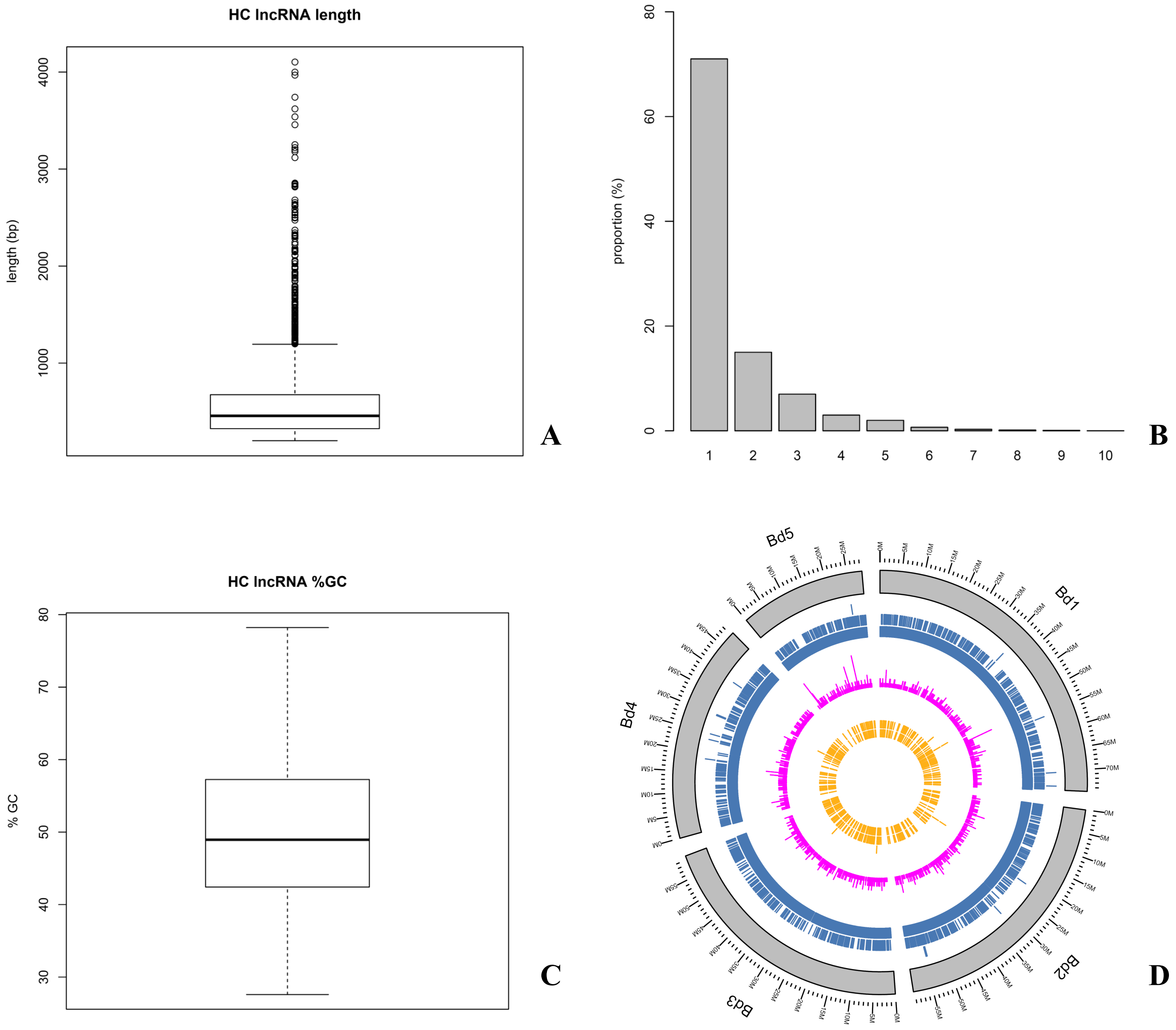


Figure 4.6 The main features of identified lncRNAs. **A** Bd lncRNAs length distribution. **B** Number of exons in Bd lncRNAs. **C** %GC in Bd lncRNAs. **D** Bd lncRNAs genome distribution; lncRNAs (magenta), miRNAs (yellow), coding genes (light blue).

4.5 LncRNAs expression analysis

Considering all different samples within the dataset, we were able to study the lncRNAs expression levels across developing stages, conditions, tissues and organs. An heat map was created, representing the HC-lncRNAs expression in Reads Per Kilobase per Million mapped reads (RPKM) (Figure 4.7). We observed that Bd lncRNAs are expressed in each sample and some of them have a uniform expression. Some lncRNAs appear to be specifically expressed in different stages of leaf development or reproductive organs, particularly in anthers.

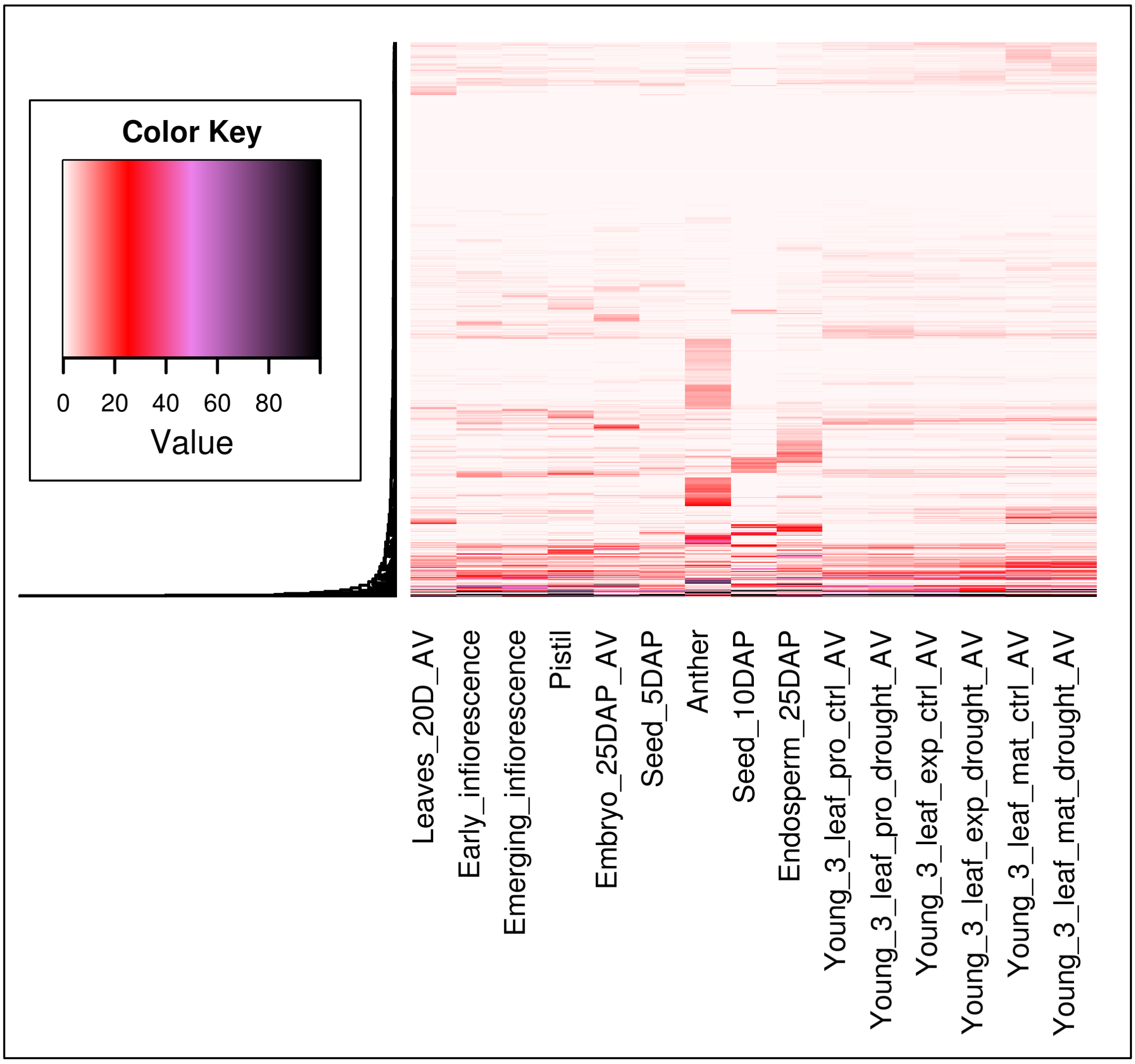


Figure 4.7 HC-lncRNAs abundance in 15 different Bd stages/conditions/tissues/organs. Expression in Reads Per Kilobase per Million mapped reads (RPKM). AV means average from 2 or 3 biological replicates.

To assess whether lncRNA expression changes during leaf development in control and drought stress conditions, between proliferation and expansion zone, a differential expression analysis was performed. Two classes of comparisons were performed:

	Comparison
Drought stress	Ps vs Pc
	Es vs Ec
Leaf development	Pc vs Ec
	Ps vs Es

Table 3.2 List of pairwise comparison performed during the differential expression analysis. (P) proliferation, (E) expansion, (C) control and (S) drought stress conditions.

- During drought stress, evaluating the expression profile of the same cell type in different growing conditions (Ps vs Pc, Es vs Ec).
- During leaf development, comparing the expression profiles of different cell types in the same growing conditions (Pc vs Ec, Ps vs Es).

Based on p-value < 0.05, in the first comparison, we retrieved a low number of differentially expressed lncRNAs in drought and control conditions, both in proliferation and expansion cells, with low fold change and no differentially expressed lncRNA in common between the two developing areas (Figure 4.8 A). On the other hand, the second comparison addressed many differentially expressed lncRNAs during leaf development in the same growing conditions, especially in drought stress (Figure 4.8 B), even if the values of fold change were still quite low, between -3 and 3. In particular, as we can see from the Venn diagram, 49 lncRNAs were differentially expressed from proliferation to expansion state in control conditions (31 up-regulated and 18 down-regulated), whereas 101 lncRNAs were differentially expressed from proliferation to expansion in drought stress conditions (72 up-regulated and 29 down-regulated). 89 were the lncRNAs differentially regulated during leaf development, independently from growing conditions (64 up-regulated and 25 down-regulated).

It seems clear that lncRNA expression profile is drastically influenced by leaf development, both in control and stress conditions. Whereas, drought treatment does not induce a drastic change in lncRNA expression profile.

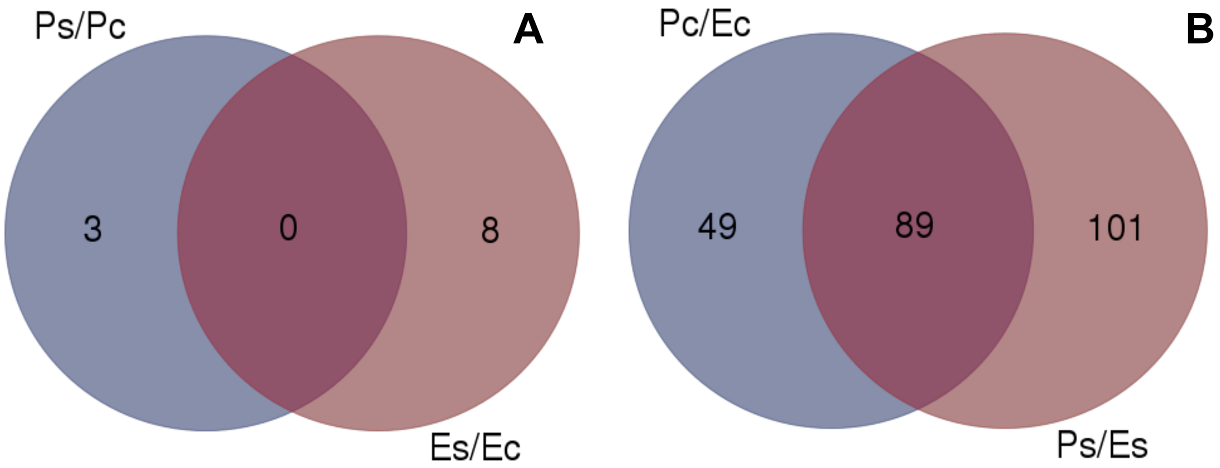


Figure 4.8 Differentially Expressed (DE) HC lncRNAs during the drought treatment (**A**) and cell differentiation (**B**) with p-value < 0.05 and p-adj < 0.05. DE analysis was done using DESeq2 package.

lncRNAs	log2FoldChange	p-value
TCONS_00015018	-0.391568158552012	0.00169553043101715
TCONS_00043107	-0.625683331815137	1,16E-08
TCONS_00027965	-0.826017424806412	2,85E-07

Table 4.3 List of the 3 *Brachypodium* lncRNAs differentially expressed during the drought treatment in proliferation zone (Ps/Pc), with their expression levels and the p-value.

lncRNAs	log2FoldChange	p-value
TCONS_00060391	0.998810486372624	0.0049834190480159
TCONS_00060390	0.813905722581128	0.0012339662696178
TCONS_00031988	0.798653873405099	0.000250472119968534
TCONS_00097102	0.652341793613137	0.00128858846016998
TCONS_00036885	-0.582357738937515	0.000729379014777919
TCONS_00026709	-0.792345171470535	0.00839529436403813
TCONS_00059484	-0.800763011678796	0.00656536533458766
TCONS_00039063	-0.948541558128087	0.00055381872010678

Table 4.4 List of the 8 *Brachypodium* lncRNAs differentially expressed during the drought treatment in expansion zone (Es/Ec), with their expression levels and the p-value.

lncRNAs	log2FoldChange	p-value
TCONS_00009992	3.656548937	0.00014596
TCONS_00086597	3.323773011	0.002914647
TCONS_00029729	3.222174	0.004100552
TCONS_00089348	3.139134476	0.005321221
TCONS_00085639	3.092605501	0.006158615
TCONS_00033709	3.02135064	0.007628533
TCONS_00030330	2.971583098	0.001142336
TCONS_00025737	2.906163663	0.001327771
TCONS_00032424	2.857482421	0.012231616
TCONS_00006657	2.824649738	0.013417575
TCONS_00056633	2.786979038	0.006392135
TCONS_00093704	2.691365811	0.019101252
TCONS_00086707	2.599974482	0.014078321
TCONS_00047695	2.568953676	0.015343956
TCONS_00041693	2.551672383	0.027224927
TCONS_00049972	2.508940416	0.002823506
TCONS_00078224	2.474158753	0.000782508
TCONS_00022889	2.408415504	0.025041605
TCONS_00034335	2.364881761	0.02640585
TCONS_00066957	2.223538429	0.025227778
TCONS_00067472	2.154397904	0.020800242
TCONS_00023948	2.139301883	0.013097427
TCONS_00091642	1.982484055	0.006617143
TCONS_00018012	1.865517047	0.020079112
TCONS_00002587	1.849697186	0.007367172
TCONS_00014615	1.818744983	0.00714711
TCONS_00026187	1.744962494	0.017015
TCONS_00024424	1.497581996	0.002233536
TCONS_00006953	1.348189905	0.009298903
TCONS_00033579	1.151785414	0.000625769
TCONS_00027455	0.933105045	0.029042562
TCONS_00024075	-0.635140208	0.000498325
TCONS_00036795	-0.867649303	0.007381457
TCONS_00041304	-1.008052937	0.004634515
TCONS_00007449	-1.340459316	0.004796425
TCONS_00002033	-1.433879341	0.028108367
TCONS_00083137	-1.584835553	1.76E-05

lncRNAs	log2FoldChange	p-value
TCONS_00045386	-1.744410134	0.004934379
TCONS_00068268	-2.212474045	0.008144458
TCONS_00029851	-2.239311078	0.024782165
TCONS_00089035	-2.255784059	0.024282922
TCONS_00095877	-2.294363526	0.022446944
TCONS_00069571	-2.313254181	0.00690456
TCONS_00090190	-2.391265344	0.027142375
TCONS_00081033	-2.894632964	0.011320595
TCONS_00079290	-2.994032098	0.008433052
TCONS_00088496	-3.057039273	0.002686827
TCONS_00020495	-3.217483298	0.004194486
TCONS_00095737	-3.229296607	8.81E-05

Table 4.5 List of the 49 *Brachypodium* lncRNAs differentially expressed during cell differentiation from proliferation zone to expansion zone in control conditions (Pc/Ec), with their expression levels and the p-value.

lncRNAs	log2FoldChange	p-value
TCONS_00010022	4.061441166	0.000656853
TCONS_00090462	3.940147872	0.001073887
TCONS_00069261	3.867641484	0.001365797
TCONS_00014199	3.801186541	0.001719713
TCONS_00098571	3.784870551	0.001839522
TCONS_00096683	3.774051483	0.001907477
TCONS_00011006	3.470338808	0.005067086
TCONS_00074624	3.470338808	0.005067086
TCONS_00091956	3.367282686	0.006916952
TCONS_00075163	3.307373065	0.008286416
TCONS_00096507	3.252918913	0.000577381
TCONS_00095053	3.215403956	0.010627474
TCONS_00090463	3.212647205	0.002863824
TCONS_00077606	3.111577863	0.013921259
TCONS_00066117	3.106575601	0.014108483
TCONS_00092902	3.085122706	0.004691338
TCONS_00061918	3.048669908	0.016186633
TCONS_00057658	3.041745887	0.016508948
TCONS_00035378	3.027309118	0.017484996
TCONS_00041986	2.99632893	0.006028321
TCONS_00067205	2.912587622	0.022471083
TCONS_00053243	2.907545157	0.022764093
TCONS_00083270	2.903248963	0.0009035
TCONS_00064998	288269803	0.024208076
TCONS_00059612	2.877046177	0.010699745
TCONS_00055078	2.876618204	0.003384124
TCONS_00024261	2.868211591	0.025003062
TCONS_00043395	2.854055767	0.010975499
TCONS_00053581	2.85150193	0.010735636
TCONS_00093117	2.819927488	0.012819446
TCONS_00048070	2.60139605	0.010773407
TCONS_00028837	2.533565899	0.02700719
TCONS_00034780	2.515422215	0.028744521
TCONS_00001781	2.505929779	0.030702203
TCONS_00078722	2.487822687	0.015926272
TCONS_00037644	2.469291459	0.007411618
TCONS_00040049	2.459751902	0.016250629

lncRNAs	log2FoldChange	p-value
TCONS_00045724	2.458783847	0.0041372
TCONS_00058363	2.438918763	0.00950643
TCONS_00020415	2.424007864	0.018633207
TCONS_00017860	2.289544825	0.017072111
TCONS_00050763	2.090988583	0.015524565
TCONS_00034664	2.080038744	0.029257967
TCONS_00006477	2.067507216	0.001404001
TCONS_00058731	2.049844265	0.0247089
TCONS_00093377	2.01135773	0.029539382
TCONS_00024575	1.999054099	0.026940126
TCONS_00007625	1.997480291	0.003379831
TCONS_00031782	1.987362179	0.001817409
TCONS_00036436	1.97297364	0.014681578
TCONS_00036246	1.956736729	0.009139925
TCONS_00079580	1.948932007	0.002386967
TCONS_00051392	1.948004953	0.018594921
TCONS_00005157	1.821767894	0.025673889
TCONS_00077038	1.720655378	0.004312795
TCONS_00096981	1.705713274	0.007928843
TCONS_00051674	1.681277602	0.021309372
TCONS_00064106	1.515343028	0.000537529
TCONS_00085493	1.502855173	0.014197522
TCONS_00077008	1.381654501	0.032681592
TCONS_00025709	1.376851016	0.015729498
TCONS_00027227	1.20199244	0.00517799
TCONS_00015899	1.189341095	0.01277566
TCONS_00053949	1.134139889	0.000723392
TCONS_00024810	0.986449005	0.000418507
TCONS_00072113	0.984204904	0.023467781
TCONS_00031564	0.980892108	0.024308869
TCONS_00072826	0.961556677	0.002137194
TCONS_00082062	0.893562248	0.000308489
TCONS_00046487	0.822703035	0.008742571
TCONS_00054918	0.809850396	0.020455035
TCONS_00020854	0.712465393	0.000552369
TCONS_00097102	-0.527302567	0.011380596
TCONS_00019452	-1.038388655	0.003521237

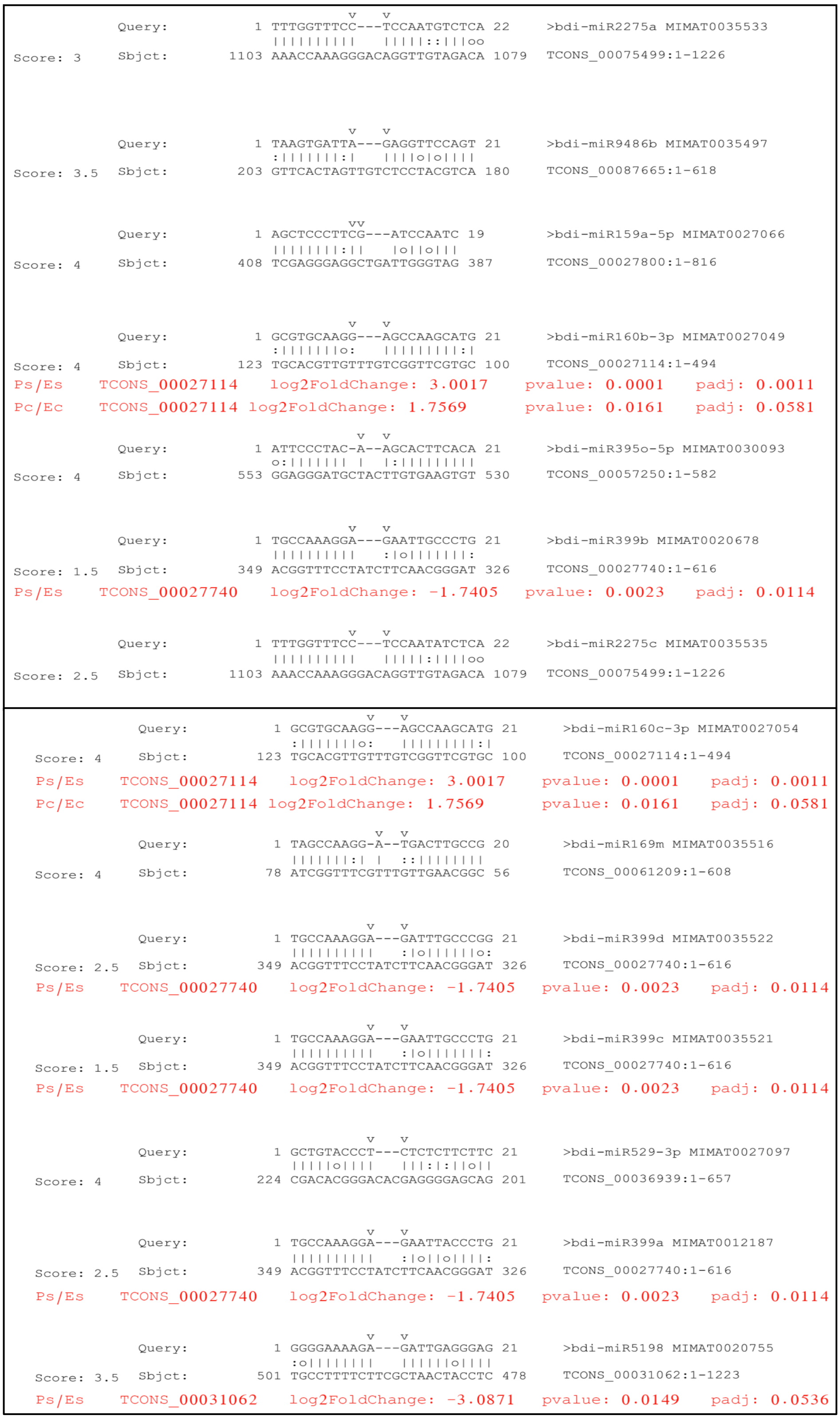
lncRNAs	log2FoldChange	p-value
TCONS_00071785	-1.113682118	0.001209529
TCONS_00054908	-1.156132672	0.021601069
TCONS_00032690	-1.402635866	0.019907418
TCONS_00000264	-1.470476666	0.022623468
TCONS_00094461	-1.603950274	0.007549635
TCONS_00027740	-1.740526527	0.00231951
TCONS_00034147	-1.836396266	0.006552438
TCONS_00056034	-2.168557034	0.00456031
TCONS_00056069	-2.615128614	0.023184802
TCONS_00005473	-2.703811919	0.018637201
TCONS_00045188	-2.886530891	0.010092372
TCONS_00055567	-2.917891081	0.022516386
TCONS_00045290	-2.949193411	0.020907554
TCONS_00081553	-3.071761806	0.015537071
TCONS_00031062	-3.087103004	0.014953207
TCONS_00062410	-3.105976014	0.014260342
TCONS_00029713	-3.236335426	0.010142415
TCONS_00008885	-3.242379641	0.009957766
TCONS_00074339	-3.258080461	0.009662622
TCONS_00069697	-3.526687326	0.004399436
TCONS_00058251	-3.538129281	0.004244907
TCONS_00073220	-3.565768162	0.003904556
TCONS_00024842	-3.645533959	0.003005665
TCONS_00098380	-3.680929291	0.002683127
TCONS_00049444	-3.748925425	0.002138603
TCONS_00070535	-4.13605239	0.000507366
TCONS_00019564	-4.318542051	0.00024005

Table 4.6 List of the 101 *Brachypodium* lncRNAs differentially expressed during cell differentiation from proliferation zone to expansion zone in drought stress conditions (Ps/Es), with their expression levels and the p-value.

4.6 Predicted endogenous lncRNAs: target mimicry

The computational method described in Wu *et al*. (2013) was used to identify HC-lncRNAs with potential target mimicry properties, *i.e.* partially complementary sequences that can block the interaction between miRNAs and their authentic targets (Wu *et al*., 2013).

14 eTMs were predicted, three of which were differentially expressed from proliferation to expansion in control and stress conditions: TCONS_00027114, up-regulated and targeting bdi-miR160, TCONS_00027740, down-regulated and targeting bdi-miR399 and TCONS_00031062, down-regulated and targeting bdi-miR5198.



```
                                       v   v
            Query:          1 TTTGGTTTCC---TCCAATGTCTCA 22     >bdi-miR2275a MIMAT0035533
                              ||||||||||   |||||::|||oo
Score: 3    Sbjct:       1103 AAACCAAAGGGACAGGTTGTAGACA 1079   TCONS_00075499:1-1226


                                       v   v
            Query:          1 TAAGTGATTA---GAGGTTCCAGT 21      >bdi-miR9486b MIMAT0035497
                              :|||||||:|   ||||o|o||||
Score: 3.5  Sbjct:        203 GTTCACTAGTTGTCTCCTACGTCA 180     TCONS_00087665:1-618


                                       vv
            Query:          1 AGCTCCCTTCG---ATCCAATC 19        >bdi-miR159a-5p MIMAT0027066
                              ||||||||:||   |o||o|||
Score: 4    Sbjct:        408 TCGAGGGAGGCTGATTGGGTAG 387       TCONS_00027800:1-816


                                       v   v
            Query:          1 GCGTGCAAGG---AGCCAAGCATG 21      >bdi-miR160b-3p MIMAT0027049
                              :|||||||o:   |||||||||:|
Score: 4    Sbjct:        123 TGCACGTTGTTTGTCGGTTCGTGC 100     TCONS_00027114:1-494
Ps/Es    TCONS_00027114    log2FoldChange: 3.0017    pvalue: 0.0001    padj: 0.0011
Pc/Ec    TCONS_00027114 log2FoldChange: 1.7569       pvalue: 0.0161    padj: 0.0581
                                        v  v
            Query:          1 ATTCCCTAC-A--AGCACTTCACA 21      >bdi-miR395o-5p MIMAT0030093
                              o:||||||| |  |:|||||||||
Score: 4    Sbjct:        553 GGAGGGATGCTACTTGTGAAGTGT 530     TCONS_00057250:1-582


                                       v   v
            Query:          1 TGCCAAAGGA---GAATTGCCCTG 21      >bdi-miR399b MIMAT0020678
                              ||||||||||   :|o|||||||:
Score: 1.5  Sbjct:        349 ACGGTTTCCTATCTTCAACGGGAT 326     TCONS_00027740:1-616
Ps/Es    TCONS_00027740    log2FoldChange: -1.7405   pvalue: 0.0023    padj: 0.0114

                                       v   v
            Query:          1 TTTGGTTTCC---TCCAATATCTCA 22     >bdi-miR2275c MIMAT0035535
                              ||||||||||   |||||:||||oo
Score: 2.5  Sbjct:       1103 AAACCAAAGGGACAGGTTGTAGACA 1079   TCONS_00075499:1-1226

                                       v   v
            Query:          1 GCGTGCAAGG---AGCCAAGCATG 21      >bdi-miR160c-3p MIMAT0027054
                              :|||||||o:   |||||||||:|
Score: 4    Sbjct:        123 TGCACGTTGTTTGTCGGTTCGTGC 100     TCONS_00027114:1-494
Ps/Es    TCONS_00027114    log2FoldChange: 3.0017    pvalue: 0.0001    padj: 0.0011
Pc/Ec    TCONS_00027114 log2FoldChange: 1.7569       pvalue: 0.0161    padj: 0.0581

                                        v  v
            Query:          1 TAGCCAAGG-A--TGACTTGCCG 20       >bdi-miR169m MIMAT0035516
                              |||||||:| |  ::||||||||
Score: 4    Sbjct:         78 ATCGGTTTCGTTTGTTGAACGGC 56       TCONS_00061209:1-608


                                       v   v
            Query:          1 TGCCAAAGGA---GATTTGCCCGG 21      >bdi-miR399d MIMAT0035522
                              ||||||||||   :|o||||||o:
Score: 2.5  Sbjct:        349 ACGGTTTCCTATCTTCAACGGGAT 326     TCONS_00027740:1-616
Ps/Es    TCONS_00027740    log2FoldChange: -1.7405   pvalue: 0.0023    padj: 0.0114

                                       v   v
            Query:          1 TGCCAAAGGA---GAATTGCCCTG 21      >bdi-miR399c MIMAT0035521
                              ||||||||||   :|o|||||||:
Score: 1.5  Sbjct:        349 ACGGTTTCCTATCTTCAACGGGAT 326     TCONS_00027740:1-616
Ps/Es    TCONS_00027740    log2FoldChange: -1.7405   pvalue: 0.0023    padj: 0.0114


                                       v   v
            Query:          1 GCTGTACCCT---CTCTCTTCTTC 21      >bdi-miR529-3p MIMAT0027097
                              |||||o||||   |||:|:||o||
Score: 4    Sbjct:        224 CGACACGGGACACGAGGGGAGCAG 201     TCONS_00036939:1-657


                                       v   v
            Query:          1 TGCCAAAGGA---GAATTACCCTG 21      >bdi-miR399a MIMAT0012187
                              ||||||||||   :|o||o||||:
Score: 2.5  Sbjct:        349 ACGGTTTCCTATCTTCAACGGGAT 326     TCONS_00027740:1-616
Ps/Es    TCONS_00027740    log2FoldChange: -1.7405   pvalue: 0.0023    padj: 0.0114

                                       v   v
            Query:          1 GGGGAAAAGA---GATTGAGGGAG 21      >bdi-miR5198 MIMAT0020755
                              :o||||||||   ||||||o||||
Score: 3.5  Sbjct:        501 TGCCTTTTCTTCGCTAACTACCTC 478     TCONS_00031062:1-1223
Ps/Es    TCONS_00031062    log2FoldChange: -3.0871   pvalue: 0.0149    padj: 0.0536
```

Figure 4.9 Predicted base-pairing interaction between Bd miRNAs and HC-lncRNAs, according to the computational method described in Wu *et al*. (2013). In red Differentially Expressed (DE) HC-lncRNAs during the drought treatment and cell differentiation, with their p-value and p-adj.

5 Discussion

With technical advances in studying the eukaryotic transcriptome, the increasing complexity of eukaryotic genome expression was revealed (Dinger *et al*., 2009). The existence of non-protein coding transcripts, including sRNAs and lncRNAs was discovered. sRNAs are relatively well characterized and their importance in transcriptional and post-transcriptional regulation of expression of other genes is well understood (Bonnet *et al*., 2006; Farazi *et al*., 2008; Finnegan and Matzke, 2003; Ghildiyal and Zamore, 2009; Zhang *et al*., 2006), whereas, lncRNAs have not been as comprehensively identified and well studied in many plant species.

This work focuses on the identification and characterization of a large set of long noncoding RNAs in the model species *Brachypodium distachyon*. Providing detailed information about their genomic distribution and expression patterns across different tissues/organs and identifying potential link between lncRNAs and miRNAs, through target mimicry features.

5.1 *Bona fide* lncRNAs

In this study, a total of 2507 HC-lncRNAs was identified in *Brachypodium distachyon*. For this purpose, the starting set of loci/isoforms, *de novo* reconstructed, was subjected to a strict criteria bioinformatic pipeline, described in Li *et al*. (2014), which takes into account all the major characteristics of long noncoding RNAs described so far in previous studies (Ben Amor *et al*., 2009; Wang *et al*., 2015; Zhang *et al*., 2014; Zhu *et al*., 2015; Maeda *et al*., 2006; Derrien *et al*., 2012; Xie *et al*., 2014). The analysis generated a robust list of potential lncRNAs in *Brachypodium*, which will be useful for the scientific community. The large number of identified lncRNAs is comparable to other annotation studies carried out in several plant species (Zhou *et al*., 2009; Li *et al*., 2014; Chen *et al*., 2015; Shuai *et al*., 2014; Zhu *et al*., 2015). Nevertheless, there are some potential limitations to the list of lncRNAs obtained that are worth noting. First, all the available RNA-Seq data were produced selecting polyadenylated transcripts, but it is possible that some lncRNAs lack polyadenylation (Di *et al*., 2014), so we have not been able to annotate them. Relatively strict criteria were also employed by requiring that the putative lncRNAs lack the ability to encode peptides of more than 100 amino acids or only have a weak coding potential. In fact, there are examples of previously characterized lncRNAs from other species with the potential to encode peptides >100 amino acids, such as *HOTAIR*, *XIST* and *KCNQ1OT* (Duret *et al*., 2006; Kanduri, 2011; Rinn *et al*., 2007).

In summary, although some Bd lncRNAs might be excluded due to the sequencing limitations and strict criteria, a relatively reliable list of Bd lncRNAs is provided and will be useful for the research community.

Concerning the characteristics of Bd lncRNAs annotated in this study, our results are in accordance with previous evidences made in different plant species (Li *et al*., 2014; Wang *et al*., 2015; Zhu *et al*., 2015). Length, percentage of GC and number of exons of HC-lncRNAs were exanimated. The length of the greatest part of HC-lncRNAs ranges from 200 to 1000 nucleotides. Also, as in other plant species, most of the genes encoding Bd lncRNAs only contained one exon (Chekanova, 2015).

In contrast with the observations in maize plants (Li *et al*., 2014), where lncRNAs are concentrated in chromosomes 1 and 5, in *Brachypodium distachyon* we detected a pervasive transcription of lncRNAs in every chromosome, similarly to the results obtained in the species *S. lycopersicum* (Zhu *et al*., 2015).

5.2 Expression patterns and differential expression analysis

To highlight specific expression profiles among all the samples considered in this study, data counts were expressed in Reads Per Kilobase per Million mapped reads (RPKM). As can be seen from the heat map in Figure 4.7, Bd lncRNAs are expressed in all of these samples and some of them seem to have a uniform expression in each sample. In accordance with literature, the expression levels of lncRNAs are, generally, very low (RPKM between 0 and 80), nevertheless there are also some lncRNAs highly expressed in specific samples, leading to think that lncRNAs are specifically involved in the development of some tissues. For instance, some lncRNAs appear to be specifically highly expressed in the different stages of leaf development, with no appreciable difference between control and stress conditions. Interestingly, a high expression level of specific lncRNAs characterizes reproductive organs and, particularly, anthers. This observation reflects the results of a previous study in rice (Zhang *et al*., 2014), in which several lncRNAs have been found to be highly expressed in reproductive organs and, in particular, in anthers, where they might play roles in germ cell development or meiosis. It would be interesting to investigate the role of reproductive organ-specific lncRNAs in *Brachypodium* too, in particular, as already done in rice, an interaction with miRNAs should be assessed, as many miRNAs have been reported to regulate reproduction in plants (Luo *et al*., 2013).

To better investigate the modulation of lncRNAs during drought stress condition and leaf development, a differential expression analysis was conducted, taking advantage of the experimental design concerning proprietary RNA-Seq libraries, where three biological replicates were generated for each developmental zone (proliferation, expansion, mature) and growing conditions (control, drought stress).

To further explore the modulation of lncRNAs during leaf differentiation and, in particular, between cell proliferation and cell expansion (the first two steps leading to differentiation), pairwise comparison between different cell type grown in control and drought stress conditions (hence Pc vs Ec and Ps vs Es) were investigated. Moreover, we were able to investigate the response of different cell type to drought stress (pairwise comparison: Ps vs Pc, Es vs Ec). Although previous studies have shown the modulation of lncRNAs in response to biotic and abiotic stresses in plants, such as *Arabidopsis*, wheat (Liu *et al*., 2012; Xin *et al*., 2011) and, recently, some drought-responsive lincRNAs have been characterized in *Populus trichocarpa* (Shuai *et al*., 2014); our results show a moderate modulation of lncRNAs in response to drought. Only a small number of differentially expressed lncRNAs between drought and control conditions was obtained but we cannot exclude that lncRNAs are involved in drought stress response, in fact, also the expression level of coding genes related to cell division in leaf's proliferation zone is not so influenced by drought stress and only the genes related to cytokinin have high transcript levels in the expansion zone of drought-stressed *Brachypodium* leaves (Verelst *et al*., 2013). Moreover we have to bear in mind that the expression levels of lncRNAs were observed only in specific plant cells and not in the entire plant system, hence we cannot appreciate the behaviour of these molecules in the global mechanism of response to drought stress.

On the contrary, the number of differentially expressed Bd lncRNAs obtained during leaf development indicates a straightforward participation of lncRNAs in cell differentiation, in particular 49 lncRNAs were differentially expressed from proliferation to expansion state in control conditions (31 up-regulated and 18 down-regulated), whereas 101 lncRNAs were differentially expressed from proliferation to expansion in drought stress conditions (72 up-regulated and 29 down-regulated). 89 were the lncRNAs differentially regulated during leaf development, independently from growing conditions (64 up-regulated and 25 down-regulated). The high number of lncRNAs differentially expressed in drought stress conditions may indicate a response of cell differentiation to the perturbation of normal physiological conditions, which would be interesting to deeply investigate in the future. In fact, cell cycle and cell differentiation regulation are of pivotal importance for plant growth and development and, even if the majority of molecular actors are well known, the role of non-coding genome in these processes is still unexplored. Moreover, recently, several studies in animal models revealed the importance of lncRNAs in cell proliferation in cancer conditions, such as MALAT1 (Dong *et al*., 2014; Gutschner *et al*., 2013; Ji *et al*., 2003). It isn’t hard to think that plant lncRNAs may have similar functions participating in the fine tuning of gene expression in several biological pathways.

5.3 lncRNA-miRNA interaction

Target mimicry is a newly identified miRNA regulation mechanism, first studied in *Arabidopsis* (Franco-Zorrilla *et al*., 2007). According to this mechanism, over-expression of the decoy RNAs that bind to miRNAs, through partially complementary sequences, block the interaction between miRNAs and their authentic targets. In this way, lncRNAs increase the expression of the miRNA target. No target mimics have yet been identified in *Brachypodium*.

In this study, potential target mimic lncRNAs for 14 miRNAs were identified. Intriguingly, the majority of lncRNAs involved in these interactions are differentially expressed during cell differentiation, both in physiological and drought conditions, indicating that lncRNAs might cooperate with miRNAs to regulate cell differentiation process. In particular, bdi-miR399 is targeted by a lncRNA down-regulated during leaf cell differentiation in stress conditions and a lncRNA up-regulated during leaf development in both control and stress conditions is a potential target mimic for bdi-miR160. To date, miR399 family is known to play an important role in phosphate homeostasis regulation in *Arabidopsis* (Franco-Zorrilla *et al*., 2007; Kuo and Chiou, 2011), particularly, miR399 can guide the cleavage of PHO2 RNA, which negatively affects shoot Pi content and Pi remobilization. On the basis of the obtained results, it's possible to speculate that, being phosphorous an essential nutrient for plant life, important in cell division and development of new tissues, the down-regulation of a decoy lncRNA for miRNA399, which positively regulate phosphorous homeostasis, is a possible mechanism involved in leaf development in stress conditions.

The interaction between a Bd lncRNA and bdi-miR160 could be an even more interesting mechanism involved in leaf development. In fact, miRNA160 targets several mRNAs implicated in auxin responses (auxin response transcription factors, ARFs) (Liu *et al*., 2007; Mallory *et al*., 2005; Turner *et al*., 2013). The phytohormone auxin is a major regulator of plant growth and development, which are sustained by coordinated cellular behaviors: cell division, expansion and differentiation. Auxin has been seen to participate in every one of these processes (Perrot-Rechenmann, 2010). In this way, lncRNAs may become part of the network of molecular actors involved in leaf development in control and stress conditions, regulating auxin pathway through target mimicry. Certainly, this is an interesting issue worth to be broadened.

6 Conclusions

In this study, from 15 RNA-Seq libraries of *Brachypodium distachyon*, 2507 lncRNAs have been identified, in different organs, tissues, conditions and leaf developmental stages. Of these, some have high expression levels in reproductive organs, such as anthers, indicating a possible involvement in reproduction.

A differential expression analysis provided a list of lncRNAs down- and up-regulated in third leaf expansion and proliferation areas, during drought stress and/or leaf differentiation, showing a straightforward implication of lncRNAs in cell differentiation.

In addition, an interaction network between lncRNAs and miRNAs was highlighted, giving clue to the interplay between lncRNAs and miRNAs through target mimicry.

This study provides the first comprehensive annotation of lncRNAs in *Brachypodium distachyon*, which can be considered an important step for the research community, to strengthen future functional genomics studies in this interesting model species and its relative crop species.

7 References

Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data.

Ben Amor, B., Wirth, S., Merchan, F., Laporte, P., d'Aubenton-Carafa, Y., Hirsch, J., Maizel, A., Mallory, A., Lucas, A., Deragon, J.M., Vaucheret, H., Thermes, C., and Crspi, M. (2009). Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. Genome Res. *19*, 57–69.

Bergmann, J.H., and Spector, D.L. (2014). Long non-coding RNAs: modulators of nuclear structure and function. Curr. Opin. Cell Biol. *26*, 10–18.

Bertolini, E., Verelst, W., Horner, D.S., Gianfranceschi, L., Piccolo, V., Inzé, D., Pè, M.E., and Mica, E. (2013). Addressing the Role of microRNAs in Reprogramming Leaf Growth during Drought Stress in Brachypodium distachyon. Mol. Plant *6*, 423–443.

Bonnet, E., Van de Peer, Y., and Rouzé, P. (2006). The small RNA world of plants. New Phytol. *171*, 451–468.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527–542.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Campalans, A., Kondorosi, A., and Crespi, M. (2004). Enod40, a Short Open Reading Frame–Containing mRNA, Induces Cytoplasmic Localization of a Nuclear RNA Binding Protein in Medicago truncatula. Plant Cell *16*, 1047–1059.

Chekanova, J.A. (2015). Long non-coding RNAs and their functions in plants. Curr. Opin. Plant Biol. *27*, 207–216.

Chen, J., Quan, M., and Zhang, D. (2015). Genome-wide identification of novel long non-coding RNAs in Populus tomentosa tension wood, opposite wood and normal wood xylem by RNA-seq. Planta *241*, 125–143.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.H., Jiang, N., and Robin Buell, C. (2012). Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J. *71*, 492–502.

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F.M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE *8*, e85024.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., and Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Di, C., Yuan, J., Wu, Y., Li, J., Lin, H., Hu, L., Zhang, T., Qi, Y., Gerstein, M.B., Guo, Y., and Lu, Z.J. (2014). Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. Plant J. Cell Mol. Biol. *80*, 848–861.

Ding, J., Lu, Q., Ouyang, Y., Mao, H., Zhang, P., Yao, J., Xu, C., Li, X., Xiao, J., and Zhang, Q. (2012a). A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proc. Natl. Acad. Sci. *109*, 2654–2659.

Ding, J., Shen, J., Mao, H., Xie, W., Li, X., and Zhang, Q. (2012b). RNA-directed DNA methylation is involved in regulating photoperiod-sensitive male sterility in rice. Mol. Plant *5*, 1210–1216.

Dinger, M.E., Pang, K.C., Mercer, T.R., and Mattick, J.S. (2008). Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. PLoS Comput Biol *4*, e1000176.

Dinger, M.E., Amaral, P.P., Mercer, T.R., and Mattick, J.S. (2009). Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief. Funct. Genomic. Proteomic. *8*, 407–423.

Dong, Y., Liang, G., Yuan, B., Yang, C., Gao, R., and Zhou, X. (2014). MALAT1 promotes the proliferation and metastasis of osteosarcoma cells by activating the PI3K/Akt pathway. Tumor Biol. *36*, 1477–1486.

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science *312*, 1653–1655.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., Plath, K., and Guttman, M. (2013). The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. Science *341*, 1237973.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., and Carrington, J.C. (2007). High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. PLoS ONE *2*.

Farazi, T.A., Juranek, S.A., and Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. Development *135*, 1201–1214.

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. *15*, 7–21.

Finnegan, E.J., and Matzke, M.A. (2003). The small RNA world. J. Cell Sci. *116*, 4689–4693.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J.A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet. *39*, 1033–1037.

Gao, Y., Meng, H., Liu, S., Hu, J., Zhang, Y., Jiao, T., Liu, Y., Ou, J., Wang, D., Yao, L., Liu, S., Hui, N. (2015). LncRNA-HOST2 regulates cell biological behaviors in epithelial ovarian cancer through a mechanism involving microRNA let-7b. Hum. Mol. Genet. *24*, 841–852.

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

Gibb, E.A., Vucic, E.A., Enfield, K.S.S., Stewart, G.L., Lonergan, K.M., Kennett, J.Y., Becker-Santos, D.D., MacAulay, C.E., Lam, S., Brown, C.J., and Lam, W.L. (2011). Human Cancer Long Non-Coding RNA Transcriptomes. PLoS ONE *6*, e25915.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.C., Hung, T., Argani, P., Rinn, J.L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R.B., Van De Vijver, M.J., Sakumar, S., and Chang, H.Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

Gutschner, T., Hämmerle, M., and Diederichs, S. (2013). MALAT1 — a paradigm for long noncoding RNA function in cancer. J. Mol. Med. *91*, 791–801.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., and Lander, E.S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Hamilton, A.J., and Baulcombe, D.C. (1999). A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. Science *286*, 950–952.

Haussecker, D. (2014). Current issues of RNAi therapeutics delivery and development. J. Controlled Release *195*, 49–54.

Heo, J.B., and Sung, S. (2011). Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA. Science *331*, 76–79.

Hirsch, J., Lefort, V., Vankersschaver, M., Boualem, A., Lucas, A., Thermes, C., d'Aubenton-Carafa, Y., and Crespi, M. (2006). Characterization of 43 Non-Protein-Coding mRNA Genes in Arabidopsis, Including the MIR162a-Derived Transcripts. Plant Physiol. *140*, 1192–1204.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S.M., Wu, Y.M., Robinson D.R., Beer, D.G., Feng, F.Y., Iyer, H.K., and Chinnaiyan, A.M. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W.E., Serve, H., and Müller-Tidow, C. (2003). MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031–8041.

Kanduri, C. (2011). Kcnq1ot1: a chromatin regulatory RNA. Semin. Cell Dev. Biol. *22*, 343–350.

Kawaji, H., Kasukawa, T., Fukuda, S., Katayama, S., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. Nucleic Acids Res. *34*, D632–D636.

Kim, E.D., and Sung, S. (2012). Long noncoding RNA: unveiling hidden layer of gene regulatory networks. Trends Plant Sci. *17*, 16–21.

Kuo, H.F., and Chiou, T.J. (2011). The Role of MicroRNAs in Phosphorus Deficiency Signaling. Plant Physiol. *156*, 1016–1024.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lee, C., and Kikyo, N. (2012). Strategies to identify long noncoding RNAs involved in gene regulation. Cell Biosci. *2*, 37.

Lee, T.L., Xiao, A., and Rennert, O.M. (2012). Identification of novel long noncoding RNA transcripts in male germ cells. Methods Mol. Biol. Clifton NJ *825*, 105–114.

Li, L., Wang, X., Stolc, V., Li, X., Zhang, D., Su, N., Tongprasit, W., Li, S., Cheng, Z., Wang, J., and Deng, X.W. (2006). Genome-wide transcription analyses in rice using tiling microarrays. Nat. Genet. *38*, 124–129.

Li, L., Eichten, S.R., Shimizu, R., Petsch, K., Yeh, C.T., Wu, W., Chettoor, A.M., Givan, S.A., Cole, R.A., Fowler, J.E., Evans, M.M., Scanlon, M.J., Yu, J., Schnable, P.S., Timmermans, M.C., Springer, N.M., and Muehlbauer, G.J. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. *15*, R40.

Li, Z., Zhao, X., Zhou, Y., Liu, Y., Zhou, Q., Ye, H., Wang, Y., Zeng, J., Song, Y., Gao, W., Zheng S.Y., Zhuang, B., Chen, H., Li, W., Li, H., Li, H., Fu, Z., and Chen, R. (2015). The long non-coding RNA HOTTIP promotes progression and gemcitabine resistance by regulating HOXA13 in pancreatic cancer. J. Transl. Med. *13*, 84.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell *24*, 4333–4345.

Liu, J., Wang, H., and Chua, N.H. (2015). Long noncoding RNA transcriptome of plants. Plant Biotechnol. J. *13*, 319–328.

Liu, P.P., Montgomery, T.A., Fahlgren, N., Kasschau, K.D., Nonogaki, H., and Carrington, J.C. (2007). Repression of AUXIN RESPONSE FACTOR10 by microRNA160 is critical for seed germination and post-germination stages. Plant J. Cell Mol. Biol. *52*, 133–146.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Molecular Cell Biology (W. H. Freeman).

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Luo, Y., Guo, Z., and Li, L. (2013). Evolutionary conservation of microRNA regulatory programs in plant flower development. Dev. Biol. *380*, 133–144.

Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J., and Tian, W. (2012). Molecular Mechanisms and Function Prediction of Long Noncoding RNA. Sci. World J. *2012*, e541786.

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engström, P.G., Lenhard, B., Aturaliya, R.N., Batalov, S., Beisel, K.W., Bult, C.J., Fletcher, C.F., Forrest, A.R.R., Furuno, M., Hill, D., Itoh, M., Kanamori-Katayama, M., Katayama, S., Katoh, M., Kawashima, T., Quackenbush, J., Ravasi, T., Ring, B.Z., Shibata, K., Sugiura, K., Takenaka, Y., Teasdale, R.D., Wells, C.A., Zhu, Y., Kai, C., Kawai, J., Hume, D.A., Carninci, P., and Hayashizaki, Y. (2006). Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs. PLoS Genet *2*, e62.

Mallory, A.C., Bartel, D.P., and Bartel, B. (2005). MicroRNA-Directed Regulation of Arabidopsis AUXIN RESPONSE FACTOR17 Is Essential for Proper Development and Modulates Expression of Early Auxin Response Genes. Plant Cell *17*, 1360–1375.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, pp. 10–12.

Matsui, A., Ishida, J., Morosawa, T., Okamoto, M., Kim, J.M., Kurihara, Y., Kawashima, M., Tanaka, M., To, T.K., Nakaminami, K., Kaminuma, E., Endo, T.A., Mochizuki, Y., Kawaguchi, S., Kobayashi, N., Shinozaki, K., Toyoda, T., and Seki, M. (2010). Arabidopsis tiling array analysis to identify the stress-responsive genes. Methods Mol. Biol. Clifton NJ *639*, 141–155.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell *2*, 279–289.

Perrot-Rechenmann, C. (2010). Cellular Responses to Auxin: Division versus Expansion. Cold Spring Harb. Perspect. Biol. *2*.

Pogue, A.I., Clement, C., Hill, J.M., and Lukiw, W.J. (2014). Evolution of microRNA (miRNA) Structure and Function in Plants and Animals: Relevance to Aging and Disease. J. Aging Sci. *2*.

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

Rehrauer, H., Aquino, C., Gruissem, W., Henz, S.R., Hilson, P., Laubinger, S., Naouar, N., Patrignani, A., Rombauts, S., Shu, H., Van De Peer, Y., Vuylsteke, M., Weigel, D., Zeller, G., and Hennig, L. (2010). AGRONOMICS1: a new resource for Arabidopsis transcriptome profiling. Plant Physiol. *152*, 487–499.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Shi, Y., Li, J., Liu, Y., Ding, J., Fan, Y., Tian, Y., Wang, L., Lian, Y., Wang, K., and Shu, Y. (2015). The long noncoding RNA SPRY4-IT1 increases the proliferation of human breast cancer cells by upregulating ZNF703 expression. Mol. Cancer *14*, 51.

Shuai, P., Liang, D., Tang, S., Zhang, Z., Ye, C.Y., Su, Y., Xia, X., and Yin, W. (2014). Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in Populus trichocarpa. J. Exp. Bot. eru256.

Sunkar, R., and Jagadeeswaran, G. (2008). In silico identification of conserved microRNAs in large number of diverse plant species. BMC Plant Biol. *8*, 37.

Swiezewski, S., Liu, F., Magusin, A., and Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature *462*, 799–802.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat. Biotechnol. *28*, 511–515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Turner, M., Nizampatnam, N.R., Baron, M., Coppin, S., Damodaran, S., Adhikari, S., Arunachalam, S.P., Yu, O., and Subramanian, S. (2013). Ectopic expression of miR160 results in auxin hypersensitivity, cytokinin hyposensitivity, and inhibition of symbiotic nodule development in soybean. Plant Physiol. *162*, 2042–2055.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. Cell *154*, 26–46.

Verelst, W., Bertolini, E., De Bodt, S., Vandepoele, K., Demeulenaere, M., Enrico Pè, M., and Inzé, D. (2013). Molecular and Physiological Analysis of Growth-Limiting Drought Stress in Brachypodium distachyon Leaves. Mol. Plant *6*, 311–322.

Voinnet, O. (2009). Origin, Biogenesis, and Activity of Plant MicroRNAs. Cell *136*, 669–687.

Wang, H., Chung, P.J., Liu, J., Jang, I.C., Kean, M.J., Xu, J., and Chua, N.H. (2014a). Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. Genome Res. *24*, 444–453.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., Wysocka, J., Lei, M., Dekker, J., Helms, J.A., and Chang, H.Y. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wang, M., Yuan, D., Tu, L., Gao, W., He, Y., Hu, H., Wang, P., Liu, N., Lindsey, K., and Zhang, X. (2015). Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). New Phytol.

Wang, P., Xue, Y., Han, Y., Lin, L., Wu, C., Xu, S., Jiang, Z., Xu, J., Liu, Q., and Cao, X. (2014b). The STAT3-Binding Long Noncoding RNA lnc-DC Controls Human Dendritic Cell Differentiation. Science *344*, 310–313.

Wen, J., Parker, B.J., and Weiller, G.F. (2007). In Silico identification and characterization of mRNA-like noncoding transcripts in Medicago truncatula. In Silico Biol. *7*, 485–505.

Wu, H.J., Wang, Z.M., Wang, M., and Wang, X.J. (2013). Widespread Long Noncoding RNAs as Endogenous Target Mimics for MicroRNAs in Plants. Plant Physiol. *161*, 1875–1884.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. *42*, D98–D103.

Xin, M., Wang, Y., Yao, Y., Song, N., Hu, Z., Qin, D., Xie, C., Peng, H., Ni, Z., and Sun, Q. (2011). Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. BMC Plant Biol. *11*, 61.

Yao, Y., Ma, J., Xue, Y., Wang, P., Li, Z., Liu, J., Chen, L., Xi, Z., Teng, H., Wang, Z., Li, Z., and Liu, Y. (2015). Knockdown of long non-coding RNA XIST exerts tumor-suppressive functions in human glioblastoma stem cells by up-regulating miR-152. Cancer Lett. *359*, 75–86.

Zhang, B., Pan, X., Cobb, G.P., and Anderson, T.A. (2006). Plant microRNA: A small regulatory molecule with big impact. Dev. Biol. *289*, 3–16.

Zhang, H., Chen, X., Wang, C., Xu, Z., Wang, Y., Liu, X., Kang, Z., and Ji, W. (2013a). Long non-coding genes implicated in response to stripe rust pathogen stress in wheat (Triticum aestivum L.). Mol. Biol. Rep. *40*, 6245–6253.

Zhang, J., Mujahid, H., Hou, Y., Nallamilli, B.R., and Peng, Z. (2013b). Plant Long ncRNAs: A New Frontier for Gene Regulatory Control. Am. J. Plant Sci. *04*, 1038–1045.

Zhang, Y.C., Liao, J.Y., Li, Z.Y., Yu, Y., Zhang, J.P., Li, Q.F., Qu, L.H., Shu, W.S., and Chen, Y.Q. (2014). Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol. *15*, 512.

Zhou, X., Sunkar, R., Jin, H., Zhu, J.K., and Zhang, W. (2009). Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa. Genome Res. *19*, 70–78.

Zhu, Q.H., and Wang, M.B. (2012). Molecular Functions of Long Non-Coding RNAs in Plants. Genes *3*, 176–190.

Zhu, B., Yang, Y., Li, R., Fu, D., Wen, L., Luo, Y., and Zhu, H. (2015). RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. J. Exp. Bot. *66*, 4483–4495.

Ringraziamenti

Sul finale di questo bellissimo percorso, ci sono diverse persone a cui devo i miei ringraziamenti.

Innanzitutto desidero ringraziare i miei due relatori, il Prof. Mario Enrico Pè e il Dr. Rodolfo Bernardi, che mi hanno permesso di realizzare questo lavoro, facendomi crescere e insegnandomi molto.

Un ringraziamento speciale va a Edoardo Bertolini, il mio tutor, che mi ha seguito durante tutto il lavoro, guidandomi sapeintemente e motivandomi ad ogni passo.

Grazie anche a chi mi è stato affianco in questa esperienza, i miei compagni di corso e di laboratorio, con cui ho passato due anni bellissimi. In particolare, ringrazio Rebecca Fiorella Talini, "Cioppi", che ha condiviso con me praticamente ogni istante di questi anni, rendendoli più felici e divertenti. Un grazie anche a Stefania De Quattro, la mia "capa", che mi ha aiutato e consigliato su ogni dubbio e problema.

Ringrazio enormemente i miei genitori e la mia famiglia, che mi ha appoggiato in ogni scelta, credendo in me e dandomi sostegno in ogni modo possibile, senza di loro tutto ciò non sarebbe stato possibile.
Un ringraziamento particolare va al mio ragazzo, Maurizio, che ha vissuto con me ogni momento, bello e meno bello, di questo percorso, supportandomi e sopportandomi.

Grazie mille a tutti voi, anche a chi non ho menzionato, ma senza cui tutto ciò non sarebbe stato lo stesso.