


UNIVERSITY OF TUSCIA

Department for Innovation in Biological, Agro-food and Forest systems – DIBAF

Ph.D. in Science, Technology and Biotechnology for Sustainability
(Biological Systems/Bioindustries)
XXIX Cycle

De novo assembly of giant reed (*Arundo donax* L.) leaf transcriptome using RNA-Seq provides insight into drought response, gene discovery and genetic marker identification

Ph.D. Candidate

Chiara Evangelistella

(AGR/05)

Tutor
Prof. Antoine Harfouche

Coordinator
Prof. Mauro Moresi

Academic year
2016-2017

This thesis was submitted to the Department for Innovation in Biological, Agrofood and Forest systems as a doctoral dissertation in partial fulfillment of the requirements for the degree of PhD in Science, Technology and Biotechnology for Sustainability (Biological Systems/Bioindustries) at the University of Tuscia, Viterbo, Italy.

Head of Department: Prof. Dr. Giuseppe Scarascia Mugnozza

Coordinator of the PhD program: Prof. Mauro Moresi

Members of the examination committee:

Chairman: Prof. Dr. Giuseppe Scarascia Mugnozza

Examiner: Prof. Dr. Alessio Valentini

Examiner: Prof. Dr. Maurizio Sabatti

Thesis Supervisor: Prof. Dr. Antoine Harfouche

CIRCULATION RESTRICTION

To Whom It May Concern,

I would like to request that a circulation restriction be placed on my dissertation "*De novo* assembly of giant reed (*Arundo donax* L.) leaf transcriptome using RNA-Seq provides insight into drought response, gene discovery and genetic marker identification" for a period of 1 year, because an article on this subject has been submitted for publication and others are being written. I have discussed this with my professor and we both agree. I understand that one electronic copy of the dissertation will be available only at the Library of the University of Tuscia.

Sincerely,

Chiara Evangelistella

Antoine Harfouche

TABLE OF CONTENT

SHORT ABSTRACT

Arundo donax has attracted renewed interest as a potential candidate bioenergy crop for use in biomass-to-liquid fuel-conversion processes and biorefineries, due to its high productivity, adaptability to marginal lands (e.g. drought-prone environments), and suitability for biofuel and biomaterial production. However, currently genomic resources publicly available for supporting *A. donax* improvement and information about its molecular response to drought in field are still limited.

RNA-Seq was used to *de novo* assemble and characterize the *A. donax* leaf transcriptome. The sequencing generated 1,249 million clean reads that were assembled into 62,596 unique sequences (unitranscripts). TransDecoder and Trinotate softwares were used to obtain putative coding sequences and to annotate them by mapping to UniProtKB/Swiss-Prot and UniRef90 databases, searching for known transcripts, proteins, protein domains and signal peptides. Furthermore, the unitranscripts were annotated by mapping them to the NCBI non-redundant, GO and KEGG databases using Blast2GO. The transcriptome was also characterized by customized BLAST searches to investigate homologous transcripts of genes coding for stress-associated proteins (SAPs), and key genes involved in important metabolic pathways, such as lignin, cellulose, purine and thiamine biosynthesis, stomatal development and carbon fixation. Additionally, 8,364 SSR markers were identified, representing the first genetic marker catalog of *A. donax*. 53 SSRs were then predicted to be polymorphic between ecotype-specific assemblies, suggesting genetic variability among the *A. donax* ecotypes (EcoA, EcoB and EcoC). Differentially expressed transcripts (DETs) between well-watered (WW) and natural moderate drought stress (mDr) conditions were also identified in three *A. donax* ecotypes and three different timepoints. All ecotypes over-expressed transcripts encoding PFP and JIPs, and employed the ABA (abscisic acid)-dependent pathways to cope with drought stress, differentially regulating transcripts belonging to different transcription factors (TFs) families and encoding proteins involved in ABA perception. In addition, differential expression (DE) analysis in EcoA and EcoB could suggest a greater control of stomata size, thereby water loss, and thus a greater ability to tolerate drought stress, compared to EcoC. In addition, EcoA and EcoB showed a down-regulation of genes involved in the lignin biosynthetic pathway, and also the down-regulation of genes involved in the drought-induced leaf senescence progression suggests a delay in leaf senescence, and a possible

increase in drought tolerance. Furthermore, DE results indicate that EcoA could be able to control stomatal density and size to fine-regulate water flow and gas exchange under drought stress. Moreover, all ecotypes down-regulated *AQP* genes for water flow control across cells. Finally, in EcoC transcripts encoding TNKS1-like were up-regulated, suggesting a role of this gene, involved in DNA damage repair, cell death and transcription and chromatin regulation, in its drought response.

This study provides the first publicly available transcriptome for *A. donax* bioenergy crop. The functional annotation and characterization of the leaf transcriptome will be highly useful for providing insight into the molecular mechanisms underlying the extreme adaptability of *A. donax*. The identified SSRs will facilitate the harnessing of untapped genetic diversity. Finally, DE analysis allowed the identification of potential candidate genes for molecular breeding of drought tolerant *A. donax* ecotypes.

Keywords

Arundo donax, drougth, biofuel, *De novo* leaf transcriptome, RNA-Seq, Genic-SSRs, DETs.

EXTENDED ABSTRACT

Introduction

Arundo donax has attracted renewed interest as a potential candidate bioenergy crop for use in biomass-to-liquid fuel-conversion processes and biorefineries. This is due to its high productivity, adaptability to marginal lands, among which drought-prone environments, and suitability for biofuel and biomaterial production. Despite its importance, currently the genomic resources publicly available for supporting the improvement of this species are still limited and also there is little information about its molecular response to drought in field.

Results

Illumina next-generation mRNA-sequencing (mRNA-Seq) was used to *de novo* assemble and characterize the *A. donax* leaf transcriptome. The sequencing generated 1,249 million clean reads that were assembled using single-*k*-*mer* (SK) and multi-*k*-*mer* (MK) approaches into 62,596 unique sequences (unitranscripts) with an N50 of 1,134 bp. TransDecoder and Trinotate software suites were used to obtain putative coding sequences and to annotate them by mapping to UniProtKB/Swiss-Prot and UniRef90 databases, searching for known transcripts, proteins, protein domains and signal peptides. Furthermore, the unitranscripts were annotated by mapping them to the NCBI non-redundant, GO and KEGG pathway databases using Blast2GO software. The transcriptome was also characterized by customized BLAST searches to investigate homologous transcripts of key genes involved in important metabolic pathways, such as lignin, cellulose, purine and thiamine biosynthesis and carbon fixation. Moreover, a set of homologous transcripts of key genes involved in stomatal development and of genes coding for stress-associated proteins (SAPs) was identified. Additionally, 8,364 simple sequence repeat (SSR) markers were identified that represent the first genetic marker catalog of *A. donax*. Furthermore, 53 SSRs (PolySSRs) were then predicted to be polymorphic between ecotype-specific assemblies, suggesting a certain degree of genetic variability in the studied *A. donax* ecotypes (called EcoA, EcoB and EcoC). Furthermore, differentially expressed transcripts (DETs) between well-watered (WW) and natural moderate drought stress (mDr) conditions were identified in three *A. donax* ecotypes and three different timepoints (referred as T1, T2 and T3). All the three ecotypes over-expressed transcripts encoding for PFP and JIPs proteins, and employed the ABA (abscisic acid)-dependent pathways to cope with drought stress, several differentially regulated transcripts belonging to different transcription factors (TFs) families (e.g. NAC, WRKY, MYC, MYB, AP2/ERF, bZIP, FAR, FRS, bHLH, PIF) and encoding for proteins involved in

ABA perception (e.g. PP2C, PYL4-like and PYR1-like). The differential regulation of *FAR1* and *FRS* genes in EcoA and EcoB could be associated with a greater control of stomata size, thereby of water loss, and thus suggesting a greater ability to tolerate drought stress compared to EcoC. In addition, in EcoA the down-regulation of *SDD1* gene, together with the up-regulation of *FAR1* transcript, could be associated with a molecular adaptive response of this ecotype which controls stomatal density and size to fine-regulate water flow and gas exchange under drought stress. Furthermore, all the three ecotypes down-regulated different *AQP* genes (e.g. *NIP2;2* and *NIP 1;1-like*; *TIP1;1* and *TIP4;2*; *SIP2;1* and *PIP1;*1 and *PIP2;7*) in order to regulate water flow across cells and subcellular compartments. Finally, for EcoC transcripts encoding for a TNKS1-like proteins were up-regulated, suggesting a role of this gene, which is involved in DNA damage repair, cell death pathways, transcription regulation and chromatin modification/remodeling, in the drought stress response of this ecotype. In addition, for EcoA and EcoB the down-regulation of genes involved in the lignin biosynthetic pathway was observed, which represents an important result in the context of *A. donax* as bioenergy crop for biofuel production. Finally, in EcoA and EcoB the down-regulation of *EIL3* gene could be associated with a delay in the progression of drought-induced leaf senescence and, a possible increase in water stress tolerance in these ecotypes.

Conclusions

This study provides the first publicly available transcriptome resource for *A. donax* bioenergy crop. The functional annotation and characterization of the leaf transcriptome will be highly useful for providing insight into the molecular mechanisms underlying the extreme adaptability of *A. donax*. The identification of homologous transcripts involved in key metabolic pathways offers a platform that will direct future efforts in genetic improvement of this species. In addition, the identified SSRs will facilitate the harnessing of untapped genetic diversity. This transcriptome should be of value to ongoing functional genomics and genetic studies in this crop of paramount economic importance. Finally, differential expression (DE) analysis allowed the identification of potential candidate genes useful for accelerating future molecular breeding program aiming to obtain increasingly drought tolerant *A. donax* ecotypes.

Keywords

Arundo donax, drougth, biofuel, *De novo* leaf transcriptome, RNA-Seq, Genic-SSRs, DETs.

AIM OF THE WORK

The aims of this PhD thesis were addressed to the *de novo* assembly, the characterization and the investigation of the leaf transcriptome generated using RNA-Seq data obtained from three *A. donax* ecotypes, grown in field under well-watered (WW) and natural moderate drought stress (mDr) conditions. Among the different objectives, there were providing insight into the molecular mechanisms underlying the extreme adaptability of *A. donax* and identifying homologous transcripts involved in key metabolic pathways that could be a platform for the genetic improvement of this species. In addition, another aim of the present thesis was the identification of simple sequence repeat (SSR) markers and then polymorphic SSRs between the three analyzed ecotypes, in order to study their putative genetic variability. Finally, the analysis of differentially expressed transcripts in the three ecotypes grown under WW and mDr conditions represents another objective of the work, with the purpose to identify candidate genes for drought tolerance in this species, and to provide the basis for furthering molecular breeding of *A. donax* ecotypes able to thrive and succeed in drought-prone marginal lands.

1. STATE OF THE ART

1.1 Origin, diffusion and reproduction of *A. donax*

Arundo donax L., which common name is "giant cane" or "giant reed", is a plant that grows spontaneously in different kinds of environmental conditions, mostly in temperate and hot zones all over the world. *A. donax* belongs to the Poaceae family, in the tribe of Arundinaceae together with other species such as *Arundo plinii*, *Arundo collina*, *Arundo mediterranea* (Mariani *et al*., 2010) and other ornamental species.

The plant originally developed in East Asia, successively spread into the Mediterranean area and then around the entire world due to human activity (Mariani *et al*., 2010; Hardion *et al*., 2012). In the Mediterranean region human domestication also occurred (Zeven and de Wet, 1982). Another hypothesis about *A. donax* origin suggests that *A. donax* and its related species (*A. plinii*, *A. collina* and *A. mediterranea* or *Arundo micrantha*) originated in the Mediterranean area (Zeven and de Wet, 1982). However, to date this plant is widespread in all continents and in very different kinds of environments.

A. donax is a hydrophyte plant that prefers growing in soil rich in water, mainly near channels, rivers, lakes, and in marshy environments, in which it reaches the maximum biomass yields (Corno *et al*., 2014).

The phylogenetic origin is still unknown. The high number of chromosomes and their small size have made very difficult the identification of *A. donax* chromosome number, so that different authors have reported different chromosome numbers; in fact, some authors reported 110 chromosomes (Hunter, 1934; Pizzolongo, 1962; Bucci *et al*., 2013); others counted instead 108 chromosomes (Christopher and Abraham, 1971), whereas Haddadchi *et al*. (2013) reported 84 chromosome (2n = 7x = 84). An hypothesis, that could explain the formation of the 2n=108 or 110 chromosomes number, is based on the fusion of reduced (n = 36) and unreduced (n = 72) gametes from fertile *A. plinii* progenitors (2n = 72) (Hardion *et al*., 2012; Bucci *et al*., 2013).

As expected for an agamic-reproducing plant species, little genetic diversity has been reported, with few exceptions in Australia (Haddadchi *et al*., 2013) and in USA (Khudamrongsawat *et al*., 2004).

Genetic improvement of *A. donax*, which aims to enhance its performance as bioenergy crop plant, is mainly based on clonal selection. In fact, experimental evidences reported heritable phenotypic differences among clones, that could be useful to improve some plant

traits (e.g. number of culms, culm diameter and height) (Cosentino *et al*., 2006; Pilu *et al*., 2014). Other methods for improving this species are based on physical and chemical mutagenesis, that aim to modify the genome of this species (Dhir *et al*., 2010; Takahashi *et al*., 2010). However, these techniques are hampered by gene redundancy due to the ploidy level, and thereby the mojaority of recessive mutations would be masked by the dominant wild-type alleles. Consequently, the rare events of phenotypic modifications could be due to rare gain of function mutations and also genomic rearrangement (Comai, 2005). Experimental evidence demonstrated that it is possible to obtain a transient expression of green fluorescent protein (GFP) and β-glucuronidase (GUS) genes in *A. donax* using micro-projectile bombardment-mediated transformation, even if the high frequency plant regeneration from embryogenic callus still remains to be optimized, and the micro-projectile bombardment-mediated transformation will be possible when a reliable and reproducible regeneration system become available (Dhir *et al*., 2010).

The lack of division of the megaspore cell mother (Bhanwra *et al*., 1982) makes *A. donax* a sterile plant, unable to produce viable seeds (Wijte *et al*., 2005). Therefore, *A. donax* grows and disperses via asexual vegetative reproduction, that has allowed the rapid spread all over the world. Propagation of *A. donax* can occur every year directly from rhizomes, the underground structure of the plant, that explores soil laterally during spring and summer, allowing the growth and the germination of new buds. Another method of propagation is the rooting at the nodes of stem fragments or lateral canes that fall to the soil (Corno *et al*., 2014). Under suitable environmental conditions (e.g. moisture), these fragments can produce roots from nodes and new shoots can grow (Boland, 2006). So, the main mean for cane dispersion is water, and during floods stem fragments are transported and they can generate new canes; in fact *A. donax* usually grows along water bodies.

Hydroponic (Ceotto and Di Candilo, 2010; Pilu *et al*., 2014) and micropropagation (Takahashi *et al*., 2010) methods can be used also to develop plants with both root and active photosynthetic apparatus.

1.2 *A. donax* as a biomass crop in different environmental and climatic conditions

The extreme adaptability of *A. donax* to a wide range of environments, soils and growing conditions, together with high biomass yield and also low input requirements confer to this important species many advantages compared to other energy crops.

Firstly, *A. donax* is characterized by high amount of biomass production per surface unit. However, several factors (e.g. plant age, pedo-climatic conditions, plant density and agronomics) impact on biomass yield, so that different values were reported (Corno *et al*., 2014). Particularly, the average aboveground biomass production was observed to be around 15.5 kg dry matter (DM) m^{-2} (Giessow *et al*., 2011). The underground part of the plant has been less studied, and it was estimated to be around 22.5% compared to the aboveground biomass (Sharma *et al*., 1998); so the total biomass production in *A. donax* was estimated around 20 kg DM m^{-2}.

The ability of *A. donax* to grown in different environmental conditions has been previously reported (Lewandowski *et al*., 2003). Especially, sality seems not to influence the growth of this plant, so that it has been reported to grow easly near beach and estuaries with brackish water conditions (Else, 1996), and to be a good solution for saline soils, having also the potential to treat saline wastewaters (Williams *et al*., 2008). In addition, also water availability seems not to influence plant growth; experimental evidence demonstrated also that *A. donax* could tolerate both soil in which there was lack of water and also water-satured soil (Lewandowski *et al*., 2003). Moreover, Pompeiano *et al*., 2016 demonstrated that water scarcity can impact on *A. donax* physiological responses to long-term stress, but also they reported that the species had the ability to cope with prolonged water stress, as it lost limited quantity of aboveground biomass; after drought stress, a rapid restoration of physiological functions were reported upon re-watering, demonstrated the extreme environmental plasticity of *A. donax*. In addition, recent experimental evidences suggest that the yield of irrigated and drought-stressed *A. donax* plants was greater compared to other bioenergy crops (*Cynara cardunculus* and *Miscanthus x giganteus*) and that the stress physiology of *A. donax* permits the adaptation to grow in semi-arid and hot Mediterranean climates, allowing the use of this species as a viable crop for biomass production in drought-prone marginal lands (Haworth *et al*., 2016).

Furthermore, *A. donax* cultivation is practically possible in all climatic regions. Besides, researches has been trying to develop clones resistant to cold stress, in order to grow *A. donax* also in cold climates (Pompeiano *et al*., 2013).

Moreover, the extreme adaptability of *A. donax* makes this plant suitable also for marginal lands (i.e. soils that are not suitable for traditional agriculture and food production) (Lewandowski *et al*., 2003). In fact, experimental evidence demonstrated that *A. donax* can

grow in low fertility sandy soil, showing in any case competitive yields (Nassi o Di Nasso *et al*., 2013).

In addition, wildfire can promote *A. donax* expansion and dominance, and it has been observed that *A. donax* is able also to re-sprout after a fire, starting to produce again additional biomass (Coffman *et al*., 2010). So, to end up, the ability of *A. donax* to grow in different ecological conditions and also to withstand the competition with other plant species has furthered the colonization of this species in a broad range of areas (Quinn et al., 2007; Virtue *et al*., 2010), and especially in riparian habitats (Quinn and Holt, 2008).

1.3 *A. donax* as a biomass feedstock for bioenergy and biofuel production

The relevance of *A. donax* as a non-food crop species is due to different reasons: as previously said, *A. donax* is characterized not only by the high biomass yield, low input requirements and low production costs (Corno *et al*., 2014), but also because of its applications in the production of bioenergy/biofuels using biological fermentation, i.e. biogas (Ragaglini *et al*., 2014) and bio-ethanol (Jaradat, 2010), and also for biomass combustion (Dahl and Obernberger, 2004). In addition, also industrial utilizes and the extraction of chemical compounds are largely confirmed (Tracy and DeLoach, 1998).

Concerning biogas production, *A. donax* can be efficiently used for biogas production in substitution or also in partial integration with the traditional energy crops in co-digestion with animal slurries and other kind of biomass (Corno *et al*., 2014). However, it has been reported that *A. donax* shows a lower anaerobic biogasification potential than other energy crops such as corn, sorghum, rye and triticale (Schievano *et al*., 2015). This lower productivity is the consequence of the fact that *A. donax* is a sterile plant without viable seeds, that contain starch reserve; thereby, the chemical composition of *A. donax* aboveground biomass mainly encompassed structural polymers (hemicelluloses, cellulose and lignin). These polymers are less degradable than starch, for its natural recalcitrance (Himmel *et al*., 2007) due to the cell wall ultra-structure (Adani *et al*., 2011). Therefore, biomass pre-treatment could be a suitable method to enhance biomass biodegradability during anaerobic digestion by reducing natural recalcitrante (Di Girolamo *et al*., 2013). Nevertheless, when *A. donax* biogas productivity is compared with other crops by referring to the surface area unit (ha), it can be observed that bio-methane production is higher than those of others species, because of its high biomass yield (Corno *et al*., 2014).

As regards second generation bio-ethanol, it is generated from lingocellulosic biomass containing C_6 and C_5 sugar polymers (Chandel *et al*., 2007; Alvira *et al*., 2010; Menon and Rao, 2012), and its main advantages are linked to the high biomass yield per ha, negative greenhouse emissions and positive energetic balance (Kheshgi *et al*., 2000). In *A. donax* the bio-ethanol production is very high (around 11,000 L ha^{-1}) when it is compared to the other energy species (50% more than sugar cane and sugar beet, and 20% more than *Miscanthus*), due to the large amount of produced biomass (Qin *et al*., 2011). Also in the case of bio-ethanol production, biomass pre-treatment is necessary to allow the degradation of lignocellulosic material and to remove recalcitrance (Adani *et al*., 2011), that means removing henicellulose, rearranging lignin structure and preparing cellulose for hydrolysis into monomers (Scordia *et al*., 2011).

In addition, *A. donax* is also a promising energy crop for the combustion process, because of the high biomass yield and heating value, which is similar to those of the other species. In fact, it has been observed that the high heating value (HHV) is not statistically different from those of other energy crops. However, it can be deduced that, in a temperate climate region, *A. donax* potentially can be able to produce more energy per ha compared to miscanthus, switchgrass and poplar (about 40.3%, 52.6% and 59.9%, respectively) (Corso *et al*., 2014), considering *A. donax* biomass production (37.7 Mg DM ha^{-1}) (Angelini *et al*., 2009).

Finally, *A. donax* is used also for other purposes, i.e. in the direct use of canes and/or in the extraction of chemical compounds for industrial use. In the past, it was used for the production of musical instruments (Perdue, 1958; Obataya and Norimoto, 1999), and it is still used for reeds production for the mouthpieces of this sector. The chemical composition of *A. donax*, i.e. water-soluble extractives (e.g. glucose, fructose and sucrose) can modify the reed properties changing acoustic characteristics (Obataya and Norimoto, 1999). In the past, *A. donax* canes, which are very robust and strong, were used to reinforce buildings (Barreca, 2012). Nowadays, *A. donax* is still employed for prefabricated wall panel and chipboards (Flores *et al*., 2011; Ferrandéz-Garcia *et al*., 2012; Flores-Yepes *et al*., 2012). Furthermore, new efficient methods for the extraction, isolation and characterization of cellulose and lignin (Neto *et al*., 1997; Seca *et al*., 2000) that are important raw materials for several industrial products, have been developing. The chemical structure of lignin suggests that it is a source of chemical compounds, such as *p*-hydroxyphenylpropane (Seca *et al*., 2000) for biopolymer and plastic production. Moreover, the production of bio-oil by the simultaneous de-polymerization of polymers (hemicellulose, cellulose and lignin) at high temperature has been reported. The

bio-oil is a mixture of complex organic compounds that is considered as wood protector, for its hydrophobic properties, conferring wood resistance against biotic agents, fungi and termite attack (Aysu and Küçük, 2013; Temiz *et al*., 2013). Also, chemical and thermal treatments of *A. donax* biomass allow the extraction of compounds that can be used in the pharmaceutical, cosmetic and food sectors (Corso *et al*., 2014).

1.4 A glimpse into Next-Generation Sequencing (NGS) technologies

After Sanger sequencing, which is considered as a 'first-generation' technology, the newer methods are referred to as Next-Generation Sequencing (NGS). NGS techniques became available around 2005 (Metzker, 2010), and the first was the Solexa sequencing technology. Since then, several different sequencing methods have been developed, and they are constantly improved at an astounding rate. These sequencing technologies can be divided into three main methods: sequencing by synthesis and sequencing by ligation (referred as second-generation sequencing), and single-molecule sequencing (termed as third-generation sequencing).

Similarly to Sanger method, in the sequencing by synthesis technique the bases composition is determined through the chemiluminescence detection after nucleotide incorporation during synthesis of the complementary DNA strand by DNA polymerase. In this method, DNA is fragmented to the correct length size, ligated to adaptor sequences, and then clonally amplified to increase the fluorescent or chemical signal. Templates are separated and immobilized in preparation for flow cell cycles. Three technologies, different in read length, amplification and immobilization of templates, are grouped under the sequencing by synthesis method: the Roche 454 Pyrosequencing (http://www.my454.com/), the Ion Torrent System (https://www.iontorrent.com/) and the Illumina Genome Analyzer (http://www.illumina.com/). These three technologies use different chemistry, but all employ sequential washes of nucleotides along with diverse chemistries for fluorescence or chemical detection of nucleotide incorporation (Egan *et al*., 2012).

Unlike sequencing by synthesis in which DNA polymerase is used for the sequence elongation during bases determination, massively parallel sequencing by ligation technique employs the mismatch sensitivity of DNA ligase to determine nucleotides sequence (Landegren *et al*., 1988). Oligonucleotide probes of different length sizes are used with different fluorescent tags, depending on the nucleotide to be incorporated in the sequence. The

DNA fragment templates are primed with a kwon anchor sequence, to allow probes to hybridize. In the flow cell, DNA ligase joins the fluorescently labeled probes to the primer and template. To determine which probe is incorporated, florescent imaging is performed and the process is repeated with several different set of probes to determine nucleotides sequence. Two methods, varying for probes usage and read length, are classified under the sequencing by ligation technique: the ABI/SOLiD System (http://www.appliedbiosystems.com) and Polonator G.007 System (http://www.azcobiotech.com/instruments/polonator.php) (Egan *et al*., 2012).

Finally, the single-molecule sequencing (SMS) uses chemiluminescence as signal of nucleotide incorporation during sequencing of a single nucleic acid molecule, eliminating the need of DNA template amplification. Consequently, SMS provides several advantages over the second-generation sequencing methods: elimination of PCR errors and biases during clonally amplification of templates, low concentration of starting material (< 1 µg) and a simplified sample preparation (Metzker, 2010). The SMS technologies, differing in the detection of light emission, the reducing of background florescence, the chemistry used and the immobilization of molecules during the flow cell cycles (Egan *et al*., 2012), include the Helicos Genetic Analysis System (http://seqll.com/) and the Pacific BioSciences technique (http://www.pacb.com/).

1.5 Uncovering transcriptomes in non-model plant species using RNA-Sequencing

A transcriptome is the complete set of all RNA molecules in a given cell or tissue of an organism, and their quantity for a specific developmental and physiological condition. So, it represents a snapshot of all active genes, but also the combination of all isoform sequences due to alternative splicing or variant alleles. To study and interpreting the functional elements of the genome of a species, the molecular constituents of cells and tissues and the different developmental stages, it is essential understanding the transcriptome (Wang *et al*., 2009).

Microarray has been widely employed to study patterns of expression of thousand of genes in a single experiment, but its main disadvantage is represented by the lack of the capacity to detect novel genes and also the sensitivity to gene expression levels. With the advent of NGS technologies, the scope and the extent of trascriptome characterization and gene expression analysis were drastically changed.

Next-generation RNA-Sequencing (RNA-Esq.) represents the first sequencing based technology for the survey of the whole transcriptome of a species in a very high-throughput and quantitative manner. It also allows both single-base resolution for annotation and 'digital' gene expression levels at genome scale, at a much lower cost than both array methods and large-scale Sanger EST sequencing (Wang *et al*., 2009).

Due to the many advantages, RNA-Seq has experienced an astonishing development and the rapid generation of transcript datasets for gene discovery and expression analysis has allowed the study of transcriptomes also in non-model species (Marioni *et al*., 2008; Li *et al*., 2012). So, RNA-Seq technology enables the analysis of the complexity of transcriptomes in any organisms, allowing a comprehensive view of gene expression variation under environmental changes (e.g. under biotic or abiotic stress), in different developmental stages (e.g vegetative growth and flowering) and tissues (e.g. roots and leaves) (Ward *et al*., 2012).

RNA-Seq can be used also for the *de novo* assembly of transcriptomes without the need to mapping to a reference genome sequence, an aspect that makes this technology an invaluable tool for the identification of novel genes in non-model plant species (Zhou *et al*., 2012). *De novo* RNA-Seq enables also the detection of genetic polymorphisms, which are useful for molecular markers-assisted selection (MAS), comparative genetic and genomics studies, and to identify variation in traits of interest between cultivars and close-related species (Martin *et al*., 2013). These genetic markers developed using RNA-Seq, and also with the help of other NGS technologies, represent also an important resource for association mapping studies in non-model plant species. Association mapping (linkage disequilibrium mapping) identifies quantitative trait loci (QTLs) that account for the phenotypic variation among individuals in a species (Unamba *et al*., 2015). RNA-Seq offers also the possibility to perform expression QTLs (eQTLs) studies, in order to uncover the genetic variants that explain variation in gene expression levels (Majewski and Pastinen, 2011). Therefore it helps, together with other NGS technologies, in the dissection of complex genetic traits and accelerates plant breeding for traits as disease resistance, salinity and drought tolerance, in both model and non-model plant species.

An other advantage of RNA-seq is its application to advance the field of proteomics; recent experimental evidences demonstrated that RNA-Seq transcriptome profiling can provide an effective dataset for proteomic analysis of non-model organisms by *de novo* assembly (Lopez-Casado *et al*., 2012).

Finally, RNA-Seq technology can be considered also a potentially valuable tool to advance studies of plant evolution and polyploidy in both model and non-model species. Besides, a comparison of the leaf transcriptome of an allopolyploid relative of soybean with its diploid progenitors allowed the study of the contribution of the different genomes to the transcriptome (Ilut *et al*., 2012). Another study examined the transcriptome of nine tissues of three species of the Poaceae family to determine whether orthologous genes from these three species show the same expression patterns, and they observed a genomic collinearity and a conservation of expression between species. Albeit this study employed model species, these results could be very useful also in non-model species of Poaceae family, as functional genomic information from model species is transferred to agronomically important non-model species (e.g. wheat, barley, and sugarcane) (Davidson *et al*., 2012).

2. INTRODUCTION

The increased world petroleum demand, together with the dwindling fossil fuel reserves, and the environmental impact of carbon dioxide emission on global warming have demanded the scientific community to cope with the need for finding sustainable alternative renewable energy resources (Ohlrogge *et al*., 2009). As the systems for lignocellulosic bioethanol production become more efficient and cost effective, a continuous and reliable supply of plant biomass that can be produced at a low cost and with minimal use of water, fertilizer and arable land can be considered as an attractive solution for biofuel production (Byrt *et al*., 2011). These benefits may also be exploited by green chemistry and biorefinery. Furthermpre, it is widely accepted that global warming will intensify the frequency of drought periods over the 21st century (Dai, 2012). Drought, the condition in which there is a limited water availability for normal plant growth and development, represents the main environmental constrain that limits plant growth, resulting in a reduced biomass yield (Bruce *et al*., 2002). In many countries it has already began to develop methods in order to avoid this major threat, and potentially compensate the productivity gains expected from advances in both agricultural and crop breeding strategies (Shanker *et al*., 2014).

Giant reed (*Arundo donax* L.) is a polyploid perennial rhizomatous grass, belonging to the Poaceae family, with C_3 photosynthesis. Although the debate about its origin is still outstanding, recently a Middle-East provenance of *A. donax* has been suggested (Hardion *et al*., 2014). *A. donax* is recognized as one of the most promising lignocellulosic crops for the

Mediterranean area, where its potential as a biomass feedstock is acknowledged for its high yields, low irrigation and nitrogen input requirements, high resistance to drought, and the necessity of a single annual harvest during late fall to early spring, whereas delayed harvest is tolerated (Lewandowski *et al*., 2003). *A. donax* biomass feedstock can be readily converted to heat and/or electricity and can also be processed to produce liquid transport fuels and biomaterials (Corno *et al*., 2014). Because of these advantages, *A. donax* is expected to play a major role in the provision of lignocellulosic biomass across much of Europe.

Even though *A. donax* can produce panicle-like flowers, no viable seed production for Mediterranean ecotypes have been reported up to now (Balogh *et al*., 2012). Consequently, there has been little agronomic improvement in *A. donax* either through traditional breeding or through genetic engineering. The increased access and affordability of next-generation sequencing (NGS) resources (i.e. genomics and transcriptomics) provides new opportunities for rapidly employing new plant species such as *A. donax* for use as biofuel feedstock. Besides, recent advances in genomic technologies allow the development of genomic tools to enable rapid improvements of crop plants (Pérez-de-Castro *et al*., 2012). Particularly, gene expression analysis using mRNA-sequencing (RNA-Seq) technology, has provided a key source of biological information for many commercially grown crops. This allowed breeders to understand the molecular basis of complex plant mechanisms and to lead to the identification of new targets for manipulating essential processes. RNA-Seq also holds great potential as a platform for molecular breeding, the effective study of plant responses and adaptations to abiotic and biotic stresses, and the generation of molecular markers (Martin *et al*., 2013). However, the lack of genomic and transcriptomic sequence information for *A. donax* limits molecular investigation of its adaptations to the environmental stresses. When a reference genome is not available as for *A. donax*, RNA-Seq reads can be assembled *de novo* into a transcriptome using bioinformatics tools such as Trans-ABySS (Grabherr *et al*., 2011) or Trinity (Haas *et al*., 2013). *De novo* RNA-Seq assembly facilitates the study of transcriptomes for non-model plant species without sequenced genomes by enabling an almost exhaustive survey of their transcriptomes and allowing the discovery of virtually all expressed genes in a plant tissue. It also helps to reconstruct transcript sequences from RNA-Seq reads without the aid of genome sequence information. Such a powerful approach presents a considerable challenge, and so to study plant transcriptomes a set of transcript sequences in plant RNA-Seq data must be first determined via *de novo* transcriptome assembly.

Molecular markers have proven to be an important and effective tool for germplasm evaluation, genetic diversity analysis, gene mapping and marker-assisted selection (MAS) for crop improvement. However, genomic studies of *A. donax* have lagged behind that of other biofuel crops, such as switchgrass and miscanthus, partly due to the lack of polymorphic DNA molecular markers. To date, there are no molecular markers reported for *A. donax*. Microsatellites (simple sequence repeats, SSRs) are an ideal choice for developing markers for *A. donax* because of their abundance, high polymorphism, codominance, reproducibility, and cross-species transferability. Exploiting transcriptome data for the development and characterization of gene-based SSR markers is of paramount importance. In contrast to genomic SSRs, genic-SSRs are located in the coding region of the genome, are developed in a relatively easy and inexpensive way, and are highly transferable to related taxa (Zhao *et al*., 2013). Therefore, they can directly influence phenotype and also be in close proximity to genetic variation in coding or regulatory regions corresponding to traits of interest. SSRs located in coding and untranslated regions can be efficient functional markers in genic regions (Li *et al*., 2004). Despite their advantages, no SSR markers for *A. donax* are currently available.

In the present PhD thesis, the sequencing, *de novo* assembly and annotation of the leaf transcriptome of *A. donax*, as well as the development of polymorphic and functional genic-SSR markers were reported. RNA was extracted from leaf tissues of three ecotypes of *A. donax* grown under two water regimes in field conditions. Illumina sequencing produced 1,252 million RNA-Seq reads and 1,249 million of clean reads were used for the transcriptome *de novo* assembly. The leaf transcriptome assembly consists of 62,596 unitranscripts with a mean length of 842 bp, and a total nucleotide count of about 52.7 megabasepair (Mbp). A global comparison of homology between the transcriptomes of *A. donax* and four other species of the Poaceae family revealed a high level of global sequence similarity within this family. The assembly was functionally annotated using several public databases: the Swiss-Prot section of the UniProt KnowledgeBase (UniProtKB/Swiss-Prot), UniProt Reference Clusters (UniRef90), NCBI (National Center for Biotechnology Information) non-redundant (NR) database, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG). In addition, comparative analyses with several phylogenetically related species with more complete genomic information were performed, allowing the identification of putative genes controlling important agronomic or domestication traits. Particularly, genes encoding for stress-associated proteins (SAPs), for

lignin and cellulose biosynthesis, purine and thiamine metabolism, and stomatal development and distribution were analyzed in the leaf transcriptome. The identification of homologs of these genes lead to supportable hypotheses of their conserved functions and to reasonable strategies for their use in *A. donax* genetic improvement. Additionally, simple sequence repeats (SSRs) were identified in the leaf transcriptome, thereby obtaining the first genetic marker catalog for *A. donax*, which could be used for population genetics studies. This new transcript dataset provides the most comprehensive resource currently publicly available for gene expression and gene discovery in *A. donax*, and the SSR markers developed in this study will facilitate further genetic and genomic research in *A. donax* and related plant species. Finally, differentially expressed transcripts (DETs) between well-watered (WW) and natural moderate drought stress (mDr) conditions were detected for each ecotype and timepoint, and also potential candidate genes were identified and will be useful for accelerating molecular breeding of drought tolerant *A. donax* ecotypes.

3. MATERIAL AND METHODS

3.1 Plant material and experimental design

Three ecotypes of *A. donax*, named EcoA, EcoB and EcoC, were selected from a panel of 81 ecotypes collected along the Mediterranean basin for this study because of their great biomass yield potential and contrasting responses to drought stress. EcoA originates from a coastal habitat of Greece that is characterized by hot and dry climatic conditions, whereas EcoB from a coastal habitat of Croatia that is considered a transitional zone between subtropical and semiarid climates, and EcoC from an hilly habitat in northern Portugal that is characterized by a transitional zone between Mediterranean subtropical and European oceanic climates.

In 2014, six replicate plots of each selected ecotype were planted in Savigliano (SAV, 44°35'N, 07°37'E, 349 m above sea level), northern Italy, in a plot-scale experimental design using homogeneous rhizome cuttings. Each plot measured 10 m^2 (2.5 m × 4.0 m) and contained 30 rhizomes planted at a distance of 0.5 m × 1.0 m. During this first year of growth, no serious insect infestation or diseases were observed, and no pesticide or fungicide was used. Plants were irrigated until field capacity and then grown under two watering regimes during the dry season: well-watered (WW) and natural moderate drought stress (mDr). The mDr was gradually imposed by decreasing the water availability from 80% to approximately 40% of field capacity, whereas WW plants, used as control, were maintained to

approximately 80% of field capacity with drip-irrigation to replenish evapotranspiration losses. Soil volumetric water content was continuously checked in the two treatments, and was maintained near the field capacity in WW. Additionally, daily water deficit was calculated as the difference between the water availability (i.e. rainfall) and the crop water demand (i.e. specific crop evapotranspiration). The whole experiment lasted 73 days, during which three timepoints, designated as T1–T3 and corresponding to 23 (93.7 mm of water deficit), 43 (147.1 mm of water deficit), and 73 (270 mm of water deficit) days after water stress imposition, were performed for leaves.

3.2 RNA isolation, cDNA library construction and Illumina sequencing

Fully expanded, non-senescing leaves (the 5th from the top) were harvested at the three time points (T1, T2 and T3), immediately frozen in liquid nitrogen, and stored at -80°C until further processing. A total of fifty-four leaf samples (3 ecotypes × 2 treatments × 3 biological replicates × 3 time points), were then ground in liquid nitrogen with precooled mortars and pestles. These samples are envisaged to represent a large proportion of the total leaf transcriptome in *A. donax* during the exponential growth phase in different environmental conditions. Total RNA was extracted from 50 mg of grinded tissue using a Spectrum^{TM} Plant Total RNA Extraction Kit (Sigma-Aldrich, St. Louis, MO) including DNase I treatment with an On-Column DNase I Digestion Set (Sigma-Aldrich, St. Louis, MO) according to the manufacturer's instructions. Purity of total RNA was determined by spectrophotometric analysis using a T60 UV/VIS spectrophotometer (PG Instruments, Leicestershire, United Kingdom). RNA samples with 260/280 and 260/230 ratios ranging from 1.8 to 2.2 were accepted. The integrity of extracted RNA was visually checked by 1.2% (w/v) agarose gel electrophoresis. The ethidium bromide-stained gels were photographed using a Kodak-DC290 zoom camera and analyzed with the Kodak 1D Image Imaging Analysis Software (Eastman Kodak Company, Rochester, NY). The quantity of total RNA was determined with a Qubit® 2.0 Fluorometer using the Qubit® RNA BR assay kit (Invitrogen^{TM}).

The fifty-four total RNA samples from the three *A. donax* ecotypes were used for the construction of cDNA libraries and for Illumina sequencing reactions. Total RNA integrity was further checked using a Caliper GX (PerkinElmer, Waltham, MA). 1.5 µg of good quality RNA was used as input of the 'TruSeq Stranded mRNA Sample Prep kit' (Illumina, San Diego, CA) for library preparation following the manufacturer's instructions. Briefly, the poly-A containing mRNA in each of the samples was purified and enriched using magnetic oligo(dT)-rich beads, and then fragmented using divalent cations under high temperature.

Next, cDNA was synthesized by reverse transcription and standard blunt-ending plus add 'A'. Illumina TruSeq adapters with indexes were ligated to the ends of the cDNA fragments. After ligation reaction and separation of non-ligated adapters, samples were amplified by PCR to selectively enrich the cDNA fragments in the library that have adapters at both ends. To ensure that the quality of the libraries was appropriate for sequencing, the concentration and insert size of final libraries were determined using a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) and Agilent 2100 Bioanalyzer High Sensitivity and DNA 1000 assay (Agilent Technologies, Santa Clara, CA). Finally, cDNA libraries were processed with Illumina cBot for cluster generation on the flowcell following the manufacturer's instructions and sequenced in single-end mode using a HiSeq2500 sequencing platform (Illumina, San Diego, CA). The CASAVA v1.8.2 of the Illumina pipeline was used to process raw data for format conversion and de-multiplexing. On average, 23 million 50 bp reads were produced for each sample.

3.3 Quality control and filtering of read dataset

Reads were processed for quality assessment and removal of low-quality bases before *de novo* assembly of a leaf transcript catalogue. Subsequently, quality control of raw reads was performed using FastQC v0.11.3 (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and a quality score above Q30 was maintained in all fastq files for downstream analysis. Reads with at least one undefined (N) base were further excluded using an in-house built script in Python v2.7. Finally, FASTX_collapser (downloaded from http://hannonlab.cshl.edu/fastx_toolkit/index.html) was used to collapse identical sequences from fastq files of each biological replicate in the two conditions and in each time point into a single file, while maintaining read counts. Hence, a comprehensive data set, representing a large proportion of *A. donax* leaf transcriptome, was generated into a single fasta file, which was used for the subsequent assembly step.

3.4 *De novo* assembly strategy for leaf transcriptome reconstruction

In the assembly pipeline, the *de novo* leaf transcriptome reconstruction was performed using a two-step procedure. The first step (hereafter termed pre-assembly step) was carried out using two different approaches in order to capture a maximum number of transcripts. These two approaches employed different sequences of a fixed length of *k* nucleotides (*k-mers*) and were defined as single-*k-mer* (SK) and multi-*k-mer* (MK). In the SK method, the

pre-assembly was generated using a de Bruijn graph based on the *de novo* transcriptome Trinity Assembler v2.0.4 (Hass *et al*., 2013), with a default *k-mer* parameter (K = 25). In parallel, a MK approach using Trans-ABySS v1.5.1 (Robertson *et al*., 2010) and rnaSPAdes v0.1.1 (available at http://bioinf.spbau.ru/en/rnaspades) was carried out. In the pre-assembly step with Trans-ABySS, *k-mer* values of K = 17, 21 and 25 were used. The three raw transcriptomes assembled with different *k-mers* were then merged using the 'transabyss-merge' option. Whereas the pre-assembly step performed with rnaSPAdes used two *k-mer* values (k = 21 and 33). For both SK and MK methods, the minimum transcript length was set to 200 bp. The different assemblies were performed using a 16-core Intel® Xeon(R) CPU E5-2650 v2 workstation with 64 GB of RAM.

In the second step of the assembly procedure, the pre-assemblies from SK and MK approaches were clustered using CD-HIT v4.6.4 (Fu *et al*., 2012). The identity and the word size settings were 0.95 and 10, respectively, merging clusters with an alignment overlap above 95% identity to generate an overall leaf transcriptome assembly. The processed sequences were further subjected to the EvidentialGene tr2aacds pipeline (available at http://arthropods.eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html) to generate a final assembly containing a NR unique sequence (unitranscript) dataset for *A. donax* leaf transcriptome. The tr2aacds pipeline aims to select an 'optimal' set of *de novo* assembled transcripts, based on coding potential, from a pool of RNA sequences. First, the algorithm generates coding sequences (CDS) and amino acid sequences for each assembled transcript, and then removes redundant sequences using the amino acid information in order to select the best coding sequences amongst identical sequences. The pipeline also implements self-on-self searches to identify highly similar sequences using the Basic Local Alignment Search Tool (BLAST). The alignment results and CDS/protein identities are subsequently used to select and output transcripts organized as primary or alternate. The transcripts classified as dropped did not pass the internal filters and were rejected. The primary and alternate transcripts were used for further assessments.

3.5 Reads Mapping Back to Transcriptome (RMBT)

Two aligners were used for the mapping of reads back to the leaf transcriptome catalogue: Bowtie2 v2.1.0 (Langmead and Salzberg, 2012) and Burrows-Wheeler Alignment (BWA) tool v0.7.12 (Li and Durbin, 2009). The 'end-to-end' and the 'sensitive' options were used for

724 the alignment using Bowtie2. For BWA, the 'bwtsw' algorithm was used to index the leaf
725 transcriptome database, and the 'aln' and 'samse' options to align reads. SAMtools v0.1.19-
726 96b5f2294a (Li *et al*., 2009) was used to calculate mapping statistics from BAM files.

727

728 **3.6 Full-length transcript analysis**

729 The *de novo* leaf transcriptome assembly was compared to five databases from the
730 Ensembl Plants FTP server (release-29, ftp://ftp.ensemblgenomes.org/pub/plants/release-29/):
731 barley (*Hordeum vulgare*) coding sequence v1.29 (August 2014), rice (*Oryza sativa*) coding
732 sequence IRGSPv1.0.29, sorghum (*Sorghum bicolor*) coding sequence v1.29, common wheat
733 (*Triticum aestivum)* coding sequence IWGSC1.0+popseq.29, and foxtail millet (*Setaria*
734 *italica*) coding sequence JGIv2.0.29, using NCBI-BLAST v2.2.28+. BLASTn searches were
735 performed with the following parameters: -task dc-megablst, -evalue 1e-20, -max_target_seqs
736 1, and -outfmt 6. In addition, the 'analyze_blastPlus_topHit_coverage' script from the Trinity
737 software package (available at https://github.com/trinityrnaseq/trinityrnaseq/wiki) was used to
738 calculate the target database alignment coverage and to assess the number of queries that were
739 assembled to full-length and specific length thresholds, respectively, in the *de novo* assembly.

740 The quality and completeness of the *de novo* assembly were further evaluated using
741 BUSCO (Benchmarking Universal Single-Copy Orthologs) software v2.0 [30]. This quality
742 assessment tool provides high-resolution quantifications for genomes, gene sets and
743 transcriptomes and checks whether each of the BUSCO group is complete, duplicated,
744 fragmented or missing in the genome or transcriptome assembly. The leaf unitranscripts were
745 compared to the set of Embryophyta genes, which contains 1,440 BUSCO groups from a total
746 of 31 species in order to obtain a quantitative measure of the transcriptome completeness,
747 based on evolutionarily-informed expectations of gene content from near-universal single-
748 copy orthologs (Simão *et al*., 2015).

749

750 **3.7 Functional annotation of the leaf transcriptome catalogue**

751 TransDecoder v2.0.1 (available at http://transdecoder.github.io) was used to identify Open
752 Reading Frames (ORFs) and to predict potential coding sequences using the assembled
753 unitranscripts as input. After ORFs were extracted from the assembly, the Trinotate v2.0.2
754 (available at http://trinotate.github.io) pipeline was used to annotate the leaf transcript ORF

dataset. Both nucleotide transcripts and protein sequences were used to search against the UniProtKB/Swiss-Prot (uniprot_sprot.trinotate_v2.0.pep.gz) and UniRef90 (uniprot_uniref90.trinotate_v2.0.pep.gz) databases (downloaded at https://data.broadinstitute.org/Trinity/Trinotate_v2.0_RESOURCES/), using NCBI-BLASTx and BLASTp v2.2.28+ (-evalue 1e-10 -max_target_seqs 1 -outfmt 6), respectively. The UniProtKB is a collection of functional information on proteins, with accurate, consistent and rich annotation; the section Swiss-Prot contains manually annotated records. The UniRef databases provide NR clustered sets of sequences from the UniProtKB (including isoforms) and the UniProt Archive (UniParc) records (a comprehensive and NR database that contains most of the publicly available protein sequences). Functional domains were identified using the Pfam domain database (Pfam-A.hmm.gz, available at https://data.broadinstitute.org/Trinity/Trinotate_v2.0_RESOURCES/) using HMMER v3.1b2 (Finn *et al*., 2011). Potential signal peptides were identified using SignalP v4.1 (Petersen *et al*., 2011). All the leaf transcriptome annotations were loaded into the SQLite database (Trinotate.sprot_uniref90.20150131.boilerplate.sqlite.gz, downloaded at https://data.broadinstitute.org/Trinity/Trinotate_v2.0_RESOURCES/). The maximum e-value for reporting the best hit and associated annotation was 1e-5.

Blast2GO v3.1 (Conesa *et al*., 2005) was used to retrieve the leaf unitranscripts related to GO and to reveal biological pathways. Briefly, the assembled transcripts were aligned using the CloudBlast (Matsunaga *et al*., 2008) service targeted against the NCBI NR protein database with default parameters. The matching transcripts were further processed to detect associated GO terms, describing biological processes (BPs), molecular functions (MFs) and cellular components (CCs). UniProtKB, Gramene Proteins Database (GR_protein), Protein Data Bank (PDB) and The Arabidopsis Information Resource (TAIR) mapping databases were used to identify GO categories. The InterProScan (IPS) function in the Blast2GO software was further used to retrieve protein domains and motif information. Blast2GO also produces the enzyme code (EC) numbers for transcripts with an e-value less than 1e-5. Subsequently, transcripts with EC numbers were identified and mapped to the KEGG database.

Custom databases of *A. thaliana* and *S. italica* genes involved in the biosynthesis of particular cell wall components (lignin and cellulose), which should aid in the optimization of biofuel production from lignocellulosic biomass, the development of bioenergy crops, and in improving drought tolerance through stomatal development and distribution, were further

created. A database for SAP genes, which are potential candidates for biotechnological approaches aiming to improve abiotic stress tolerance in plants, was also created. BLASTn searches were performed against the custom databases with an e-value cut-off of 1e-5.

3.8 Identification of genic-SSRs and development of polymorphic SSR markers

The *A. donax* leaf unitranscripts were scanned for single sequence repeat (SSR) markers using the MISA (MIcroSAtellite identification tool) program downloaded from the Leibniz Institute of Plant Genetics and Crop Plant Research website (http://pgrc.ipk-gatersleben.de/misa/). Only perfect SSRs including mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with numbers of uninterrupted repeat units more than 10, 6, 5, 5, and 5, respectively, were targeted. The maximum interruption distance between two SSRs in a compound microsatellite was set to 100 bases.

In order to retrieve candidate polymorphic SSRs (PolySSRs) among EcoA, EcoB and EcoC, the leaf transcriptome of each ecotype was *de novo* assembled using Trinity software v2.0.4 (Haas *et al*., 2013) with the default parameters (K = 25 and a minimum transcript length of 200 bp). Clustering of redundant transcripts was performed with 95% identity and a word size of 10 using CD-HIT v4.6.4 (Fu *et al*., 2012). The CandiSSR pipeline (available at http://www.plantkingdomgdb.com/CandiSSR/index.html) (Xia *et al*., 2015) was used to identify microsatellite polymorphisms between *A. donax* unitranscripts and the ecotype-specific leaf transcriptome assemblies. This procedure (i) employs MISA software to identify repeat sequences in the reference transcriptome (using the criteria for SSR identification described above); (ii) retrieves the flanking regions (100 bp both upstream and downstream) of the identified SSRs in the reference sequences; (iii) aligns them using BLAST to the non-reference transcriptome (e.g. among different individuals); (iv) filters out the low-quality hits; (v) extracts the non-reference sequences with the flanking regions; (vi) searches for the specific reference SSRs within them; (vii) removes the invalid search items; (viii) filters out the low-quality PolySSRs; (ix) calculates the sequence similarity of the flanking regions of the identified PolySSRs; and finally (x) designs primer pairs for each PolySSR using Primer3 (http://primer3.sourceforge.net/) (Koressaar and Remm, 2007; Untergasser *et al*., 2012), assessing also the global similarity of the primer binding regions. Parameters for the primer design were as follows: minimum, maximum, and optimal sizes were 18, 24, and 20 nt; minimum and maximum GC content were 40 and 60%; and minimum and maximum T_m were

52 and 63°C, respectively. Finally, Blast2GO v3.1 (Conesa *et al*., 2005) was used to functional annotate the transcripts containing the PolySSRs, with the aim to analyze the GO categories and KEGG metabolic pathways of these unitranscripts.

3.9 Validation of PolySSR markers by PCR and Sanger sequencing

In order to validate microsatellites identified in the *A. donax* leaf transcriptome, a total of five primer pairs were randomly selected from those obtained from the CandiSSR pipeline (CPSSR_1|AAG_F 'TGCGTGAGGAACAAGATGCT', CPSSR_1|AAG_R 'TTGCCTCCTCCTTGCAAGAG', CPSSR_3|AGA_F 'TGGAAGAAAAGGAAGCGGGAA', CPSSR_3|AGA_R 'TCCAGAACATGAGGCACCAA', CPSSR_4|AGC_F 'AGCTGCAACTTGCCCTTTTG', CPSSR_4|AGC_R 'AGGCTGTTGCATCTGGTGAA', CPSSR_5|AGC_F 'AAATCCCCGCTACATGGCAA', CPSSR_5|AGC_R 'TCCTTGCCCAATCCCGAATC', CPSSR_9|CAG_F 'TGCTGTTCTTGGTGCTGTGA' and CPSSR_9|CAG_R 'ACTTTGCAACAGGTGGTCGA') to be synthesized (Eurofins Genomics, Milan, Italy) and used for screening polymorphisms among the three ecotypes. The primer pairs were of sizes ranging from 20 to 21 bp and melting temperatures from 59°C to 60°C. Genomic DNA was extracted using the 'DNeasy Plant Mini kit' (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Polymerase chain reactions (PCRs) were conducted in a total volume of 25 µL containing 5 µl of 5× colorless GoTaq® reaction buffer, 0.80 µl of 25 mM $MgCl_2$, 0.50 µl of 10 mM dNTP mix, 0.10 µl of 5u/µl GoTaq® DNA polymerase (Promega™, Madison, USA), 0.65 µl of 10µM (0.25 µM) sense and antisense primers and 20 ng of template DNA. The PCR mixture was subjected to 94°C for 3 min, following by 35 cycles of 45 s at 94°C, 30 s annealing at 55-60°C (based on the melting temperature of primers), 25 s at 72°C, and a final step at 72°C for 2 min. PCR products were visualized via 3% MetaPhor gel and ethidium bromide staining and purified using ExoSAP-IT® treatment (Amersham

Biosciences, Buckinghamshire, UK), according to the manufacturer's instructions. Purified PCR products were sequenced on a ABI 3730XL at Eurofins Genomics (Milan, Italy). Finally, the presence of the polymorphic SSR fragment between the three *A. donax* ecotypes was assessed using the sequence alignment editor BioEdit v7.2.5 (http://www.mbio.ncsu.edu/bioedit/bioedit.html).

3.10 Identification of differentially expressed transcripts in three *A. donax* ecotypes under drought stress

3.10.1 Alignment and abundance estimation of RNA-Seq reads

In order to analyze gene expression variation under moderate drought stress condition, the 'align_and_estimate_abundance.pl' Perl script from the Trinity software package (available at https://github.com/trinityrnaseq/trinityrnaseq/wiki) was first used to align high-quality reads of the three biological replicates of each ecotype under the two water regimes against the *A. donax* leaf transcriptome using Bowtie v0.12.9 (Langmead *et al*., 2009) and then to estimate abundance of reads using RSEM (RNA-Seq by Expectation Maximization) v1.2.29 (Li and Dewey, 2011).

3.10.2 Pairwise statistical analysis to reveal differentially expressed transcripts

In order to identify transcripts whose expression was modulated under moderate drought stress condition, the function that performs pairwise differential expression (DE) analysis in the Blast2GO v4 software (Conesa *et al*., 2005) was used to identify differentially expressed transcripts in each ecotype at each timepoint between treatments (WW and mDr). This function uses the R (http://www.r-project.org/) Bioconductor package EdgeR (Robinson *et al*., 2010) and the TMM (Trimmed mean of M values) (Robinson and Oshlack, 2010) as the scaling normalization method to retrieve differentially expressed transcripts (DETs). The fold change ($\geq$ 2 fold) and the *p*-value (< 0.05) were used to assess the significance of differential gene expression.

3.10.3 Functional characterization and analysis of differentially expressed transcripts

The identified DETs were functional characterized using the Blast2GO v4 software to reveal transcripts related to GO and biological pathways (KEGG pathways). First, the Viridiplantae subset (taxonomy ID: 33090) of the NCBI NR database was interrogated using the CloudBlast service (Matsunaga *et al*., 2008) and then GO categories and KEGG pathways present in the DETs were investigated.

The GO database is extensively adopted to regulate genes representation across species and also to provide a controlled and structured dictionary of terms to analyze and illustrate gene annotations and products (Harris *et al*., 2008). Conversely, KEGG pathway is a database resource of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes and also human diseases and drug development; especially, it is used for deciphering high-level systemic functions and utilities of the biological system (e.g. cell, organism and ecosystem) and molecular-level information, particularly for the biological interpretation of large-scale datasets of sequences from high-throughput technologies (Kanehisa *et al*., 2008).

4. RESULTS

4.1 Sequencing of the *A. donax* leaf transcriptome

A flowchart overview of the steps followed in the assembly process is outlined in Fig. 1. To obtain a broad sample and accurate estimate of the *A. donax* leaf transcriptome, 54 independent cDNA libraries were constructed from leaves sampled from three biological replicates of three ecotypes at three time points, and under two water stress treatments. All 54 samples were subjected to Illumina RNA-Seq. A total number of 1,252 million 50 bp single-end (SE) reads from the 54 fastq files accounting for 187.1 Gb of raw data were generated from the Illumina sequencing with a GC content ranging from approximately 47 to 50%. Prior to the *de novo* assembly step, raw reads were analyzed to assess quality metrics using the FASTQC software. All files contained high-quality reads with Phred scores ranging from 33 to 37 (Phred quality score, Q, is a measure of base-calling quality, and is inversely related to the probability that the corresponding base-call is incorrect; a Q30 threshold is considered of high quality). Using in-house developed scripts in Python, 2.85 million reads containing

uncalled bases (N) were excluded. Using Fastx_collapser, identical sequences were further collapsed into a single sequence and then the resulting files were concatenated into a unique file which was used for the subsequent assembly step (Fig. 1). This provided a data set of 1,249 million of high-quality unique reads representing the leaf transcriptome at three different developmental stages of three different ecotypes grown in two contrasting conditions and as such encompasses a large proportion of the leaf transcriptome in the species *A. donax*.

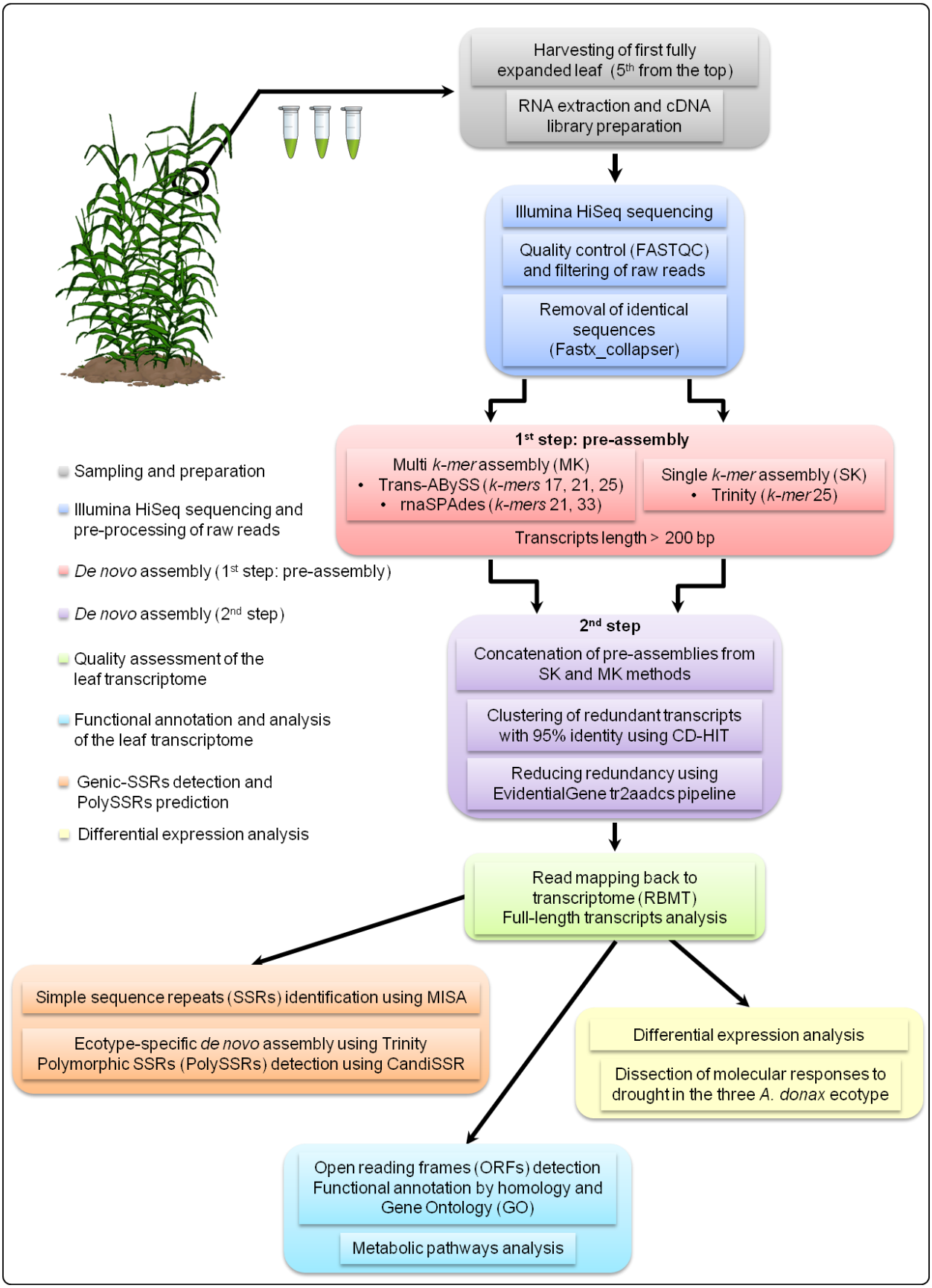

Fig. 1 Flowchart of the pipeline for the *A. donax* leaf transcriptome sequencing, *de novo* assembly, annotation and differential gene expression analysis. The pipeline performs multiple operations from sampling and preparation to sequencing and *de novo* assembly to functional annotation and differential gene expression

analysis. First, mRNA extraction from the first fully expanded leaf (5th from the top) was carried out followed by cDNA preparation and library construction (grey). Sequencing was performed using an Illumina HiSeq platform. The sequenced reads were then subjected to quality control and filtering, and identical sequences were removed (blue). Next, *de novo* pre-assemblies of transcripts were generated using a two-step approach: first, multi *k-mer* (TransABySS and rnaSPAdes) and single *k-mer* (Trinity) methods were used to generate the pre-assemblies (pink); second, pre-assemblies were then concatenated and redundant transcripts were removed using CD-HIT and the EvidentialGene tr2aadcs pipeline (purple). The quality of the *de novo* assembled leaf transcriptome was then evaluated (green). The non-redundant (NR) transcript dataset was functionally annotated by homology and Gene Ontology (GO), and metabolic pathways were analysed (mint blue). Simple sequence repeats (SSRs) and polymorphic SSRs (PolySSRs) were also identified (orange). Finally, differential gene expression analysis was performed in order to dissect the specific molecular responses to drought in the three ecotypes (yellow).

4.2 Leaf transcriptome *de novo* assembly of *A. donax*

Since no genome-wide DNA sequence is available for *A. donax* for alignment purposes, a specialized bioinformatics workflow for *de novo* transcriptome assembly were developed (Fig. 1). Accurate reconstruction and distinction between highly similar transcripts of homologous and paralogous genes as well as transcript isoforms in polyploid species present substantial challenges for *de novo* transcriptome assembly. For this reason, to optimize transcriptome assembly and obtain a high quality transcriptome, it is useful to combine the output of different *de novo* assemblers using a range of different *k-mer* sizes, while minimizing sequence redundancy (Nakasugi *et al*., 2014). Hence, the pre-assembly step was performed using two parallel approaches: a SK method using Trinity software and a MK method using Trans-ABySS and rnaSPAdes. The Trinity pre-assembly generated 136,294 transcripts with a total length of 74,660,373 bp. The median length of these transcripts (N50 value) was 659 bp and the mean and maximum transcript lengths were 547 and 10,795 bp, respectively (Table 1). N50 is defined as the length N for which 50% of all bases in the assembly are in a contig of length smaller than N (Miller *et al*., 2010). The MK approach using rnaSPAdes resulted in 138,896 raw transcripts with a total length of 66,613,178 bp, a mean and maximum transcript length of 479 bp and 8,173 bp, respectively. The N50 value was 538 bp (Table 1). Finally, the MK approach using Trans-ABySS produced 158,095 transcript sequences with a total length of 66,613,178 bp, a mean and maximum transcript lengths of 465 bp and 10,360 bp, respectively. The N50 value was 520 bp (Table 1). To reduce redundancy and possible artifacts due to the ploidy level of *A. donax* the pre-assemblies from the SK and MK methods were subsequently merged and clustered using the CD-HIT program. The processed sequences were further subjected to the EvidentialGene tr2aacds pipeline in order to select the best biological dataset of primary and alternative transcripts. The resulting final leaf unitranscript catalogue consisted of 62,596 NR unique

sequences with a considerably increased N50 value of 1,134 bp and a mean transcript length of 842 bp, covering a total nucleotide count of 52,719,740 bp with an average GC content of 49% (Table 1).

Table 1 Statistics of the leaf transcriptome pre-assemblies and the final *de novo* assembly of *A. donax*

Metric	Assembler			Final Transcriptome
	Trinity[a]	rnaSPAdes[b]	Trans-ABySS[c]	
Number of sequences	136,294	138,896	158,095	62,596
Total nucleotide count (bp)	74,660,373	66,613,178	73,637,461	52,719,740
Max. transcript length (bp)	10,795	8,173	10,360	10,360
Mean transcript length (bp)	547	479	465	842
N50 (bp)	659	538	520	1,134

Summary statistics of the *A. donax* pre-assemblies using Trinity, rnaSPAdes and Trans-AbySS, and of the final *A. donax* leaf transcriptome after clustering and redundancy removal.
[a] Trinity: *K-mer* 25; [b] rnaSPAdes: *K-mer* 21, 33; [c] Trans-ABySS: *K-mer* 17, 21, 25; The minimum transcript length was set to 200 bp.

The assembled sequence lengths ranged from the 200 bp cut-off value to a maximum transcript length of 10,360 bp. The majority of the assembled sequences were in the ranges of 200-500 and 501-1000 bp, while the frequency of longer transcripts gradually decreased and only a minor proportion reached lengths above 3,000 bp (Fig. 2).

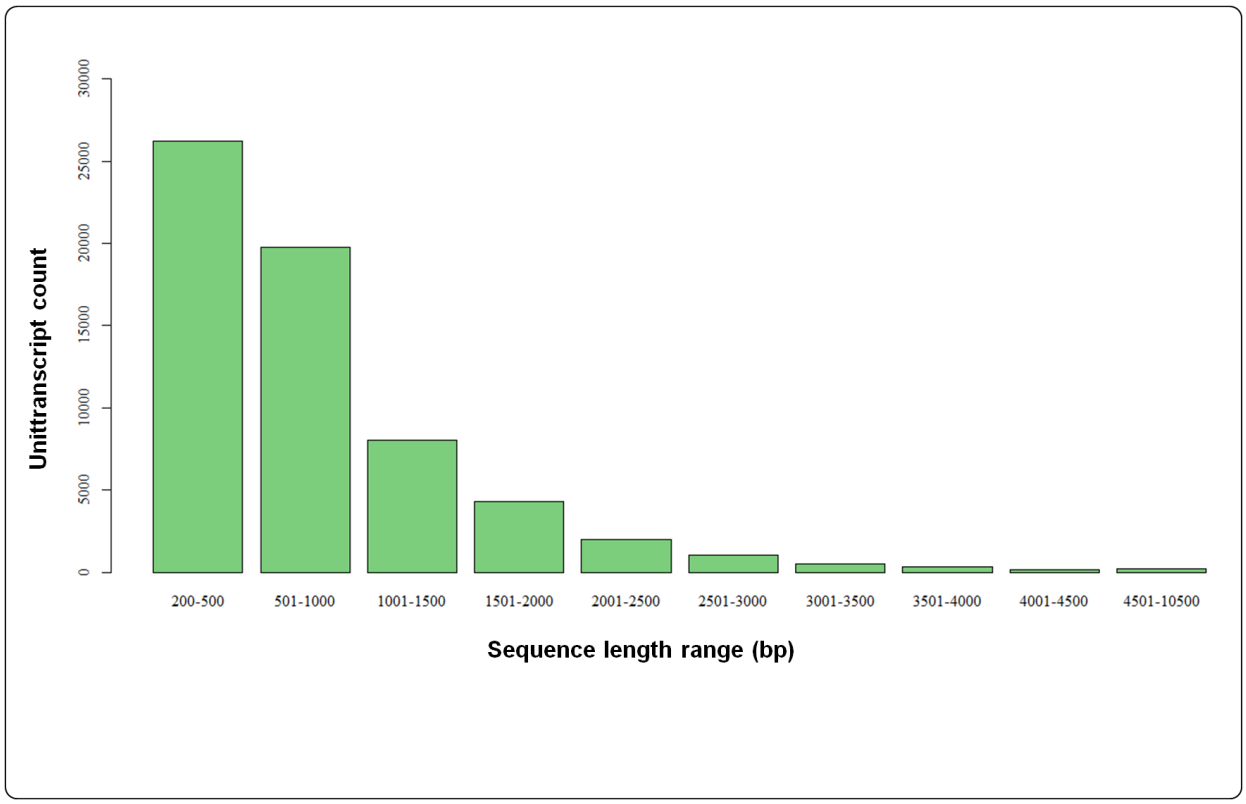

Fig. 2 Sequencing and *de novo* assembly of *A. donax* leaf transcriptome. Sequence length distribution of *A. donax* non-redundant (NR) unique unitranscript sequences. The X-axis represents the length range bins in bp. The Y-axis represents the frequency of transcripts in each bin.

4.3 Quality assessment of the leaf transcriptome assembly

To evaluate the assembly consistency, the filtered unique reads were mapped back to the final assembled leaf transcriptome. The overall alignment rate using the alignment softwares Bowtie2 and BWA was 67.62% and 62.74%, respectively. About 40% of the reads uniquely mapped to a single assembled transcript using either aligner, whereas 27.48% and 22.36% of the reads aligned to more than one transcript (Table 2).

Table 2 Percentage of reads mapped back to the *A. donax* leaf transcriptome

	Aligner	
	Bowtie2	**BWA**
Reads aligned 1 time (%)	40.14	40.38
Reads aligned > 1 times (%)	27.48	22.36
Overall alignment rate (%)	67.62	62.74

Percentage of reads uniquely mapped, aligned more than one time to the unitranscripts, and overall alignment rate using Bowtie2 and BWA aligners.

The level of homology between the leaf transcriptome dataset and those from related species of the Poaceae family was further assessed in order to evaluate how many transcripts were reconstructed at full-length and those partially reconstructed. A considerable proportion of the *A. donax* sequences aligned completely (100% of coverage) or virtually (70% of coverage) to the transcripts of related species. The highest number of hits was retrieved when *A. donax* sequences were aligned to the *H. vulgare* and *T. aestivum* transcriptomes. However, only a slightly smaller number of hits was detected for alignment against the *S. bicolor, O. sativa* and *S. italica* transcriptomes (Table 3).

Table 3 Full-length transcripts analysis

Template transcripts dataset	100% coverage	> 70% coverage	> 20% coverage
H. vulgare	6,526	11,443	17,430
T. aestivum	6,503	12,385	21,661
S. bicolor	6,281	10,851	16,521
O. sativa	6,176	11,283	16,631
S. italica	6,066	11,039	17,378

Percentage A. donax leaf transcripts aligned completely (100%), virtually (> 70%) or partially (> 20%) to the transcripts of related species.

We also checked the quality of our assembled unitranscripts by comparing them to the set of Embryophyta genes using BUSCO quality assessment tool [30]. Out of the 1,440 BUSCO groups searched, 70.07% (1,009 BUSCOs) were ‘complete’ (i.e. 813 single-copy and 196 duplicated), 13.19% (190 BUSCOs) were ‘fragmented’ and the remaining 16.74% (241 BUSCOs) were ‘missing’.

4.4 Functional annotation of the leaf transcriptome by homology

In order to determine putative functions of the identified *A. donax* unitranscripts, the Trinotate pipeline were adopted to detect homology between predicted *A. donax* ORFs/potential proteins and sequences deposited in universal databases. ORFs and potential coding sequences were first predicted using the TransDecoder software. From the initial leaf unitranscript catalogue of 62,596 NR unique sequences, 98,781 ORFs and 83,758 potential proteins could be predicted. The retrieved nucleotide sequences and putative protein sequences were then functionally annotated using the Trinotate pipeline, searching for nucleotide (BLASTx) and protein (BLASTp) homology (e-value < 1e-10) against the UniProtKB/Swiss-Prot and UniRef90 databases. In addition, presence of known functional protein domains and potential signal peptides was analyzed using the Pfam protein domain database and SignalP, respectively. In total, 34,177 nucleotide sequences (54.59%) and 33,597 protein sequences (34.01%) displayed significant homology when aligned against the UniProtKB/Swiss-Prot database using BLASTx and BLASTp searches, respectively. When aligned against the UniRef90 database the number of homologous nucleotide and protein sequences increased to 50,850 (81.23%) and 54,660 (55.33%), respectively. Furthermore, 31,513 (31.90%) unique Pfam protein motifs could be assigned and 2,729 (2.76%) protein sequences were predicted to have signal peptides. Among the Pfam domains, the most abundant classes were leucine rich repeat (LRR), pentatricopeptide repeat (PPR) and AAA domains. These protein domains can be found in proteins involved in various functions such as protein-protein interactions, transcription regulation, signal transduction and protein degradation. The Trinotate pipeline reported also the number of transcripts with unique GO assignments obtained from BLAST and Pfam searches. Consequently, 7,494 GO assignments were obtained from BLAST, whereas 1,961 unique GO terms were obtained from Pfam (Table 4).

Table 4 Overview of functional annotation by homology

Category	No. of transcripts
Predicted ORFs	98,781
Predicted proteins	83,758
Tr_EMBL_Top_BLASTX_hit	50,850
sprot_Top_BLASTX_hit	34,177

Tr_EMBL_Top_BLASTP_hit	54,660
sprot_Top_BLASTP_hit	33,597
Gene_ontology_blast	7,494
Pfam	31,513
Gene_ontology_pfam	1,961
SignalP	2,729

Summary of the functional annotation by homology.
ORFs = open reading frames; Tr_EMBL_Top_BLASTX_hit = top blastx hits against UniRef90database; sprot_Top_BLASTX_hit = top blastx hits against UniProtKB/Swiss-Prot database; Tr_EMBL_Top_BLASTP_hit = top blastp hits against UniRef90database; sprot_Top_BLASTP_hit = top blastp hits against UniProtKB/Swiss-Prot database; Gene_ontology_blast = gene ontology using blast; Gene_ontology_pfam = gene ontology using pfam.

4.5 Functional annotation of the leaf transcriptome by GO

To further functionally characterize the *A. donax* leaf trancriptome the detected transcripts were related to GO and biological pathways using the Blast2GO software. First, the blast-hit distribution of the 62,596 unitranscripts blasted against the NCBI NR protein database were assessed. The transcript dataset displayed 26,323 hits with the NR database and the highest sequence similarity (the lowest e-value) to *S. italica* (629 top-hits), followed by *S. bicolor* (205 top-hits), *Zea mays* (135 top-hits), and *O. sativa* ssp. *japonica* (123 top-hits) (Fig. 3a). The number of transcripts with InterPro (IP) sequence motifs and associated GO terms were further assessed. About 64% (40,116) of *A. donax* unitranscripts contained IP motifs, of which about 46% (18,293) were associated to GOs terms (Fig. 3b). The mapping databases used in the Blast2GO suite were UniProtKB, GR_protein, PDB and TAIR (Fig. 3c).

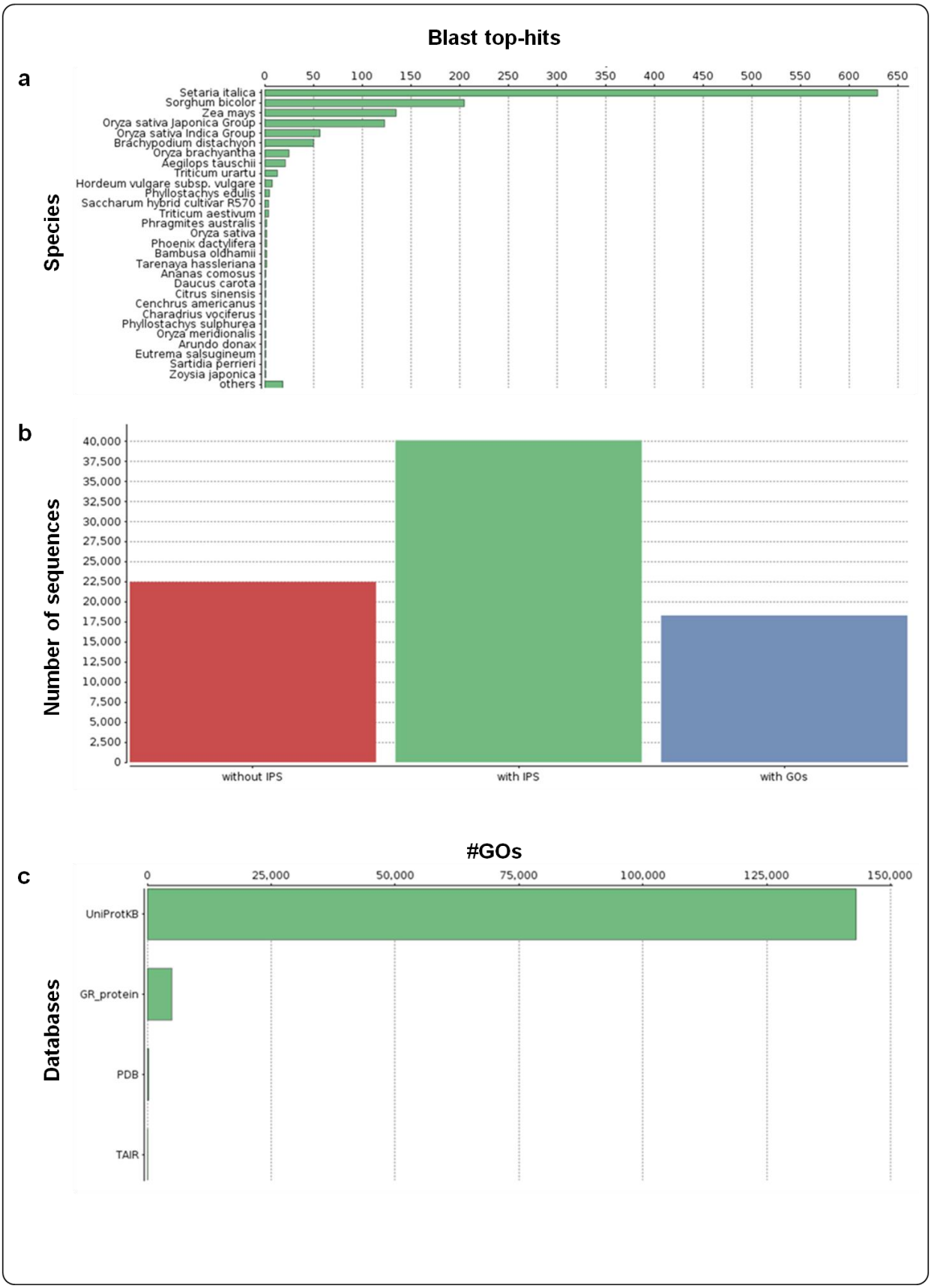

Fig. 3 Graphical representations of functional annotations in *A. donax* leaf transcriptome. **a** BLAST top-hits species distribution of *A. donax* unitranscripts against the non-redundant (NR) protein database. **b** Histogram of leaf transcriptome sequences with InterPro domains and Gene Ontology (GO) terms. The X-axis represents

transcripts without InterProScan (IPS), with IPS and with GO. The Y-axis shows the frequency of transcripts in each bin. **c** Representation of mapping databases (UniprotKB, GR_protein, PDB, TAIR) sources.

GO analysis revealed 9,778 sequences associated to 820 GO terms. Among the three main categories, BP category was the most abundant (6,299 sequences, 520 GOs), followed by MF (2,441 sequences, 151 GOs) and CC (1,038 sequences, 104 GOs) categories. Within the BP category, metabolic processes (28.75%), biosynthetic processes (25.37%), and cellular processes (17.38%) were most represented. Likewise, genes encoding binding proteins (24.66%) and genes encoding proteins related to catalytic activities (22.45%) were most abundant in the MF category. In the CC category, cell (63.20%), cell part (63%), organelle (54.14%) and membrane (37.18%) were the most abundantly represented GO terms (Fig. 4). These numbers indicate the importance of essential metabolic and biosynthetic processes in *A. donax*. However, a large proportion of transcripts showed no similarity to any known putative protein in the public databases, probably due to the presence of high numbers of long non-coding RNAs (Liu *et al*., 2015) and/or novel genes in *A. donax* that require more experimentation to be characterized in detail.

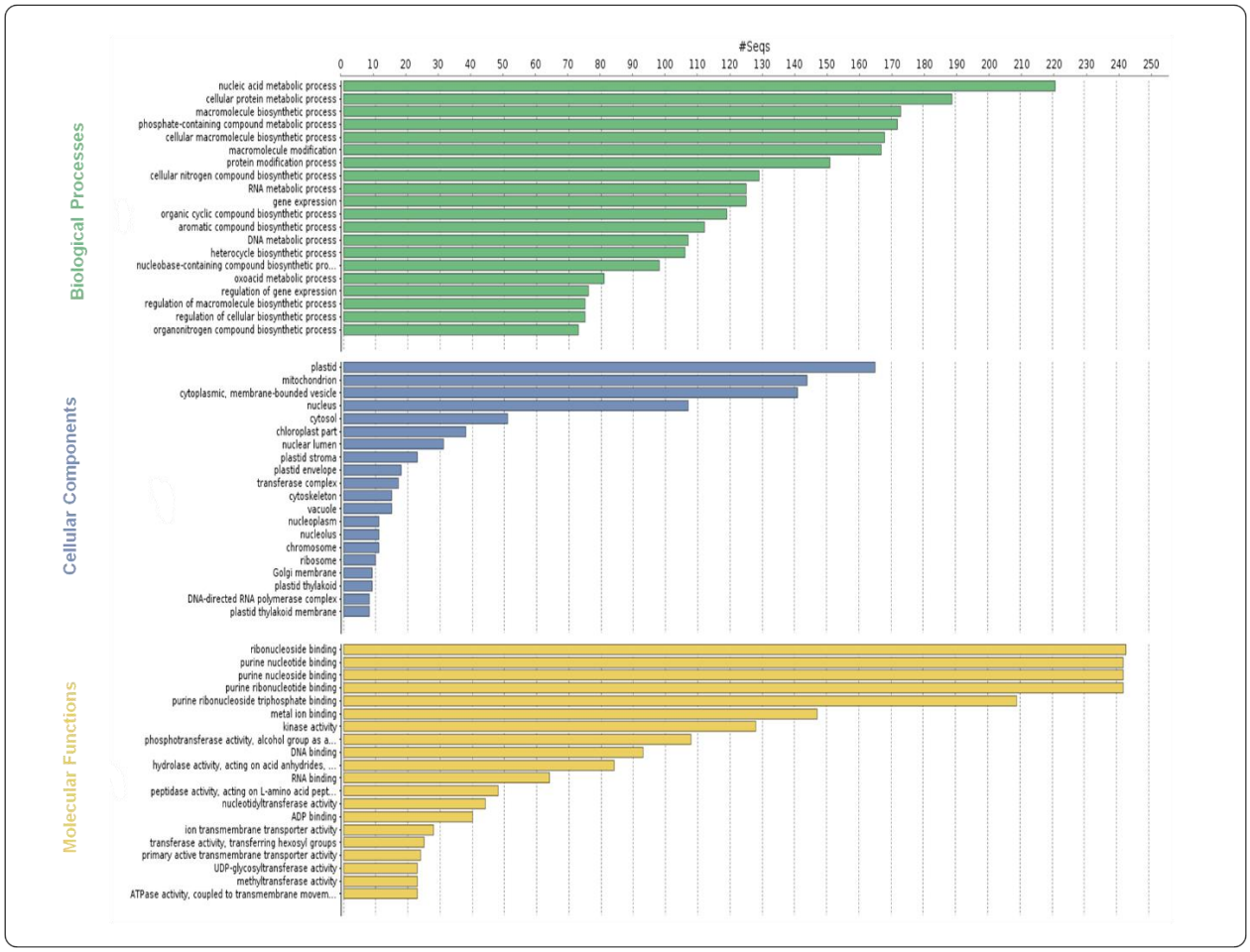

Fig. 4 Gene Ontology (GO) functional classification using Blast2GO. Histograms of the frequency of transcripts annotated to specific GO categories; biological process, cellular components and molecular functions are represented by green, blue and yellow bars, respectively.

4.6 Biological pathway analyses in *A. donax*

In order to investigate functional biological pathways in *A. donax*, the Blast2GO function that assigns EC numbers to transcripts were exploited with a blast hit e-value ≤ 1e-5. The largest proportion of such retrieved transcripts encoded for proteins with EC numbers related to transferase activity (33.53%), followed by hydrolases (29.48%) and oxidoreductases (13.30%). Next, transcripts with assigned EC numbers were mapped against the KEGG database to identify the putative presence of curated pathway components in *A. donax*. Among the 128 KEGG pathways analyzed, nucleotide metabolism, specifically purine metabolism (929 sequences), was the most represented pathway in terms of the number of homologous leaf transcripts. Furthermore, thiamine metabolism (749 sequences), metabolism of terpenoids and polyketides and other secondary metabolites (370 sequences), aminobenzoate degradation (252 sequences) and starch and sucrose metabolism (240 sequences) were highly represented.

Because of its high representation and the putative role of adenine in triggering abiotic stress tolerance, leading to increased plant biomass (Sukrong *et al*., 2012), purine metabolism were further analysed. In the *A. donax* leaf transcriptome, 926 homologs to 36 purine metabolism genes were detected, suggesting high numbers of paralogs, possibly due to the high level of polyploidization in *A. donax*. Particularly, proteins with nucleosidase activity involved in adenine, hypoxanthine, xanthine and guanine biosynthesis (EC 3.2.2.1), proteins with xanthine dehydrogenase activity (EC 1.17.1.4), and the enzyme involved in the first stage of conversion of 5-hydroxyisourate to S-allantoin (EC 3.5.2.17) were highly abundant in the transcripts catalogue. The *A. donax* leaf transcriptome dataset further included several additional enzymes involved in the biosynthesis of various compounds of purine metabolism, allowing the reconstructing of large segments of the pathway (Fig. 5a).

The second most represented pathway encompassed the metabolism of thiamine, a universal cofactor involved in common metabolic pathways, including biotic and abiotic stress responses in plants. In the *A. donax* leaf transcript dataset, 749 sequences homologous to 5 thiamine metabolism genes were identified, highlighting the importance of this pathway in *A. donax*. Particularly, enzymes involved in the production of thiamine phosphates, such as thiamine diphosphokinase (EC 2.7.6.2) converting thiamine in thiamine diphosphate, and thiamine phosphatase (EC 3.6.1.15) converting thiamine diphosphate in thiamine phosphate, were retrieved (Fig. 5b).

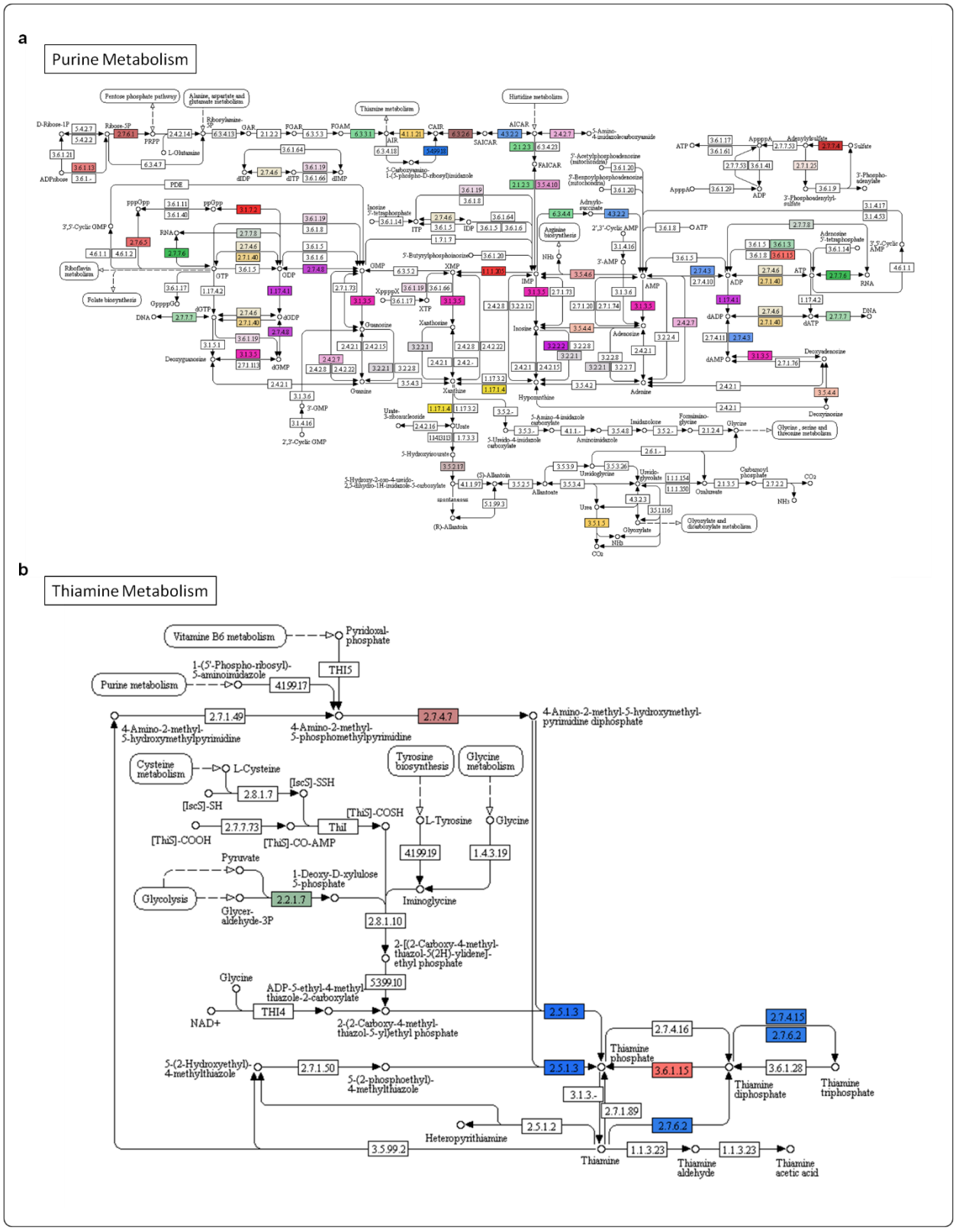

Fig. 5 Purine metabolism (**a**) and thiamine metabolism (**b**) pathways study by Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis showing the different identified enzymes in *A. donax* leaf transcriptome (one color for each Enzyme Code or EC).

Considering the substantial economic relevance of lignin metabolism in bioenergy crops, the *A. donax* leaf unitranscripts were also analyzed for coding sequences of the

phenylpropanoid biosynthesis pathway. Although this pathway was not among the most represented ones, 44 *A. donax* transcripts mapped to a peroxidase (EC 1.11.1.7), which is involved in the last step of guaiacyl, hydroxyl-guaiacyl, syringyl, and hydroxyl-phenyl lignin biosynthesis, and also to a beta-glucosidase (EC 3.2.1.21), which is involved in the synthesis of coumarinate (Fig. 6a). This indicates that *A. donax* is capable of synthesizing various types of lignin, provided that precursors are produced as well. However, despite its importance, lignin is not considered as the only target for improving the yield and quality of lignocellulosic biomass in bioenergy crops. A pivotal role towards this goal is assigned to polysaccharides. Therefore, the presence of transcripts coding for starch and sucrose metabolism was also verified in the *A. donax* leaf transcripts catalogue, identifying 30 different enzymes involved in this pathway (Fig. 6b). Amongst others, cellulose synthase (EC 2.4.1.12) involved in the synthesis of cellulose from UDP-glucose and cellulase (EC 3.2.1.4) involved in cellulose catabolism were identified. In addition, sucrose-phosphate synthase (EC 2.4.1.14) responsible for the phosphorylation of sucrose into sucrose-6-phosphate, could also be retrieved. Moreover, the *A. donax* leaf transcriptome contained several enzymes involved in the biosynthesis of amylopectin and glycogen, such as the branching enzyme (EC 2.4.1.18) (Fig. 6b). The identification of these crucial genes in biomass production might provide important targets for bioengineering approaches.

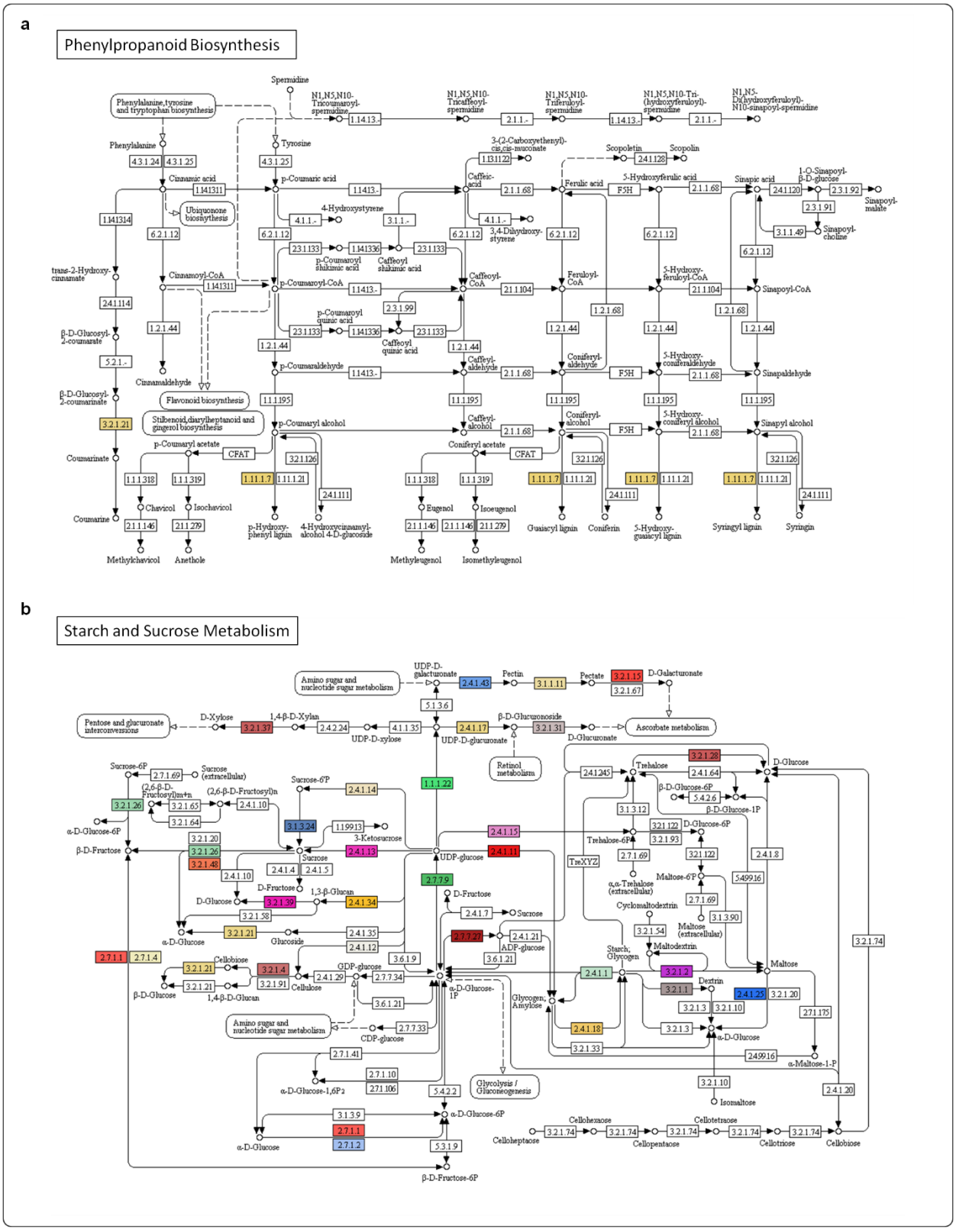

Fig. 6 Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis showing genes involved in phenylpropanoid biosynthesis (**a**) and starch and sucrose biosynthesis (**b**) representing each colored EC in *A. donax* leaf transcriptome.

To complete the analysis of carbohydrate metabolism in *A. donax*, the occurrence of components of the carbon fixation pathway in the leaf transcriptome was also explored and 15 different enzymes could be retrieved (Fig. 7). A large proportion of the carbon fixation pathway could thus be reconstructed, including, amongst others, phosphoglycerate kinase (EC 2.7.2.3) involved in the synthesis of 1,3-bisphopsho-glycerate, and aldolase (EC 4.1.2.13) required for the conversion of glyceraldehyde-3-phospahte into fructose-1,6-diphosphate and of erythrose-4-phospate into sedoheptulose-1,7- bisphosphate. Furthermore, the phosphoenolpyruvate carboxylase enzyme (EC 4.1.1.31) could be identified in the leaf transcripts. Finally, the malate dehydrogenase enzyme (EC 1.1.1.37) involved in malate biosynthesis and pyruvic-malic carboxylase (EC 1.1.1.39) involved in the conversion of malate into pyruvate were detected in the *A. donax* transcriptome (Fig. 7).

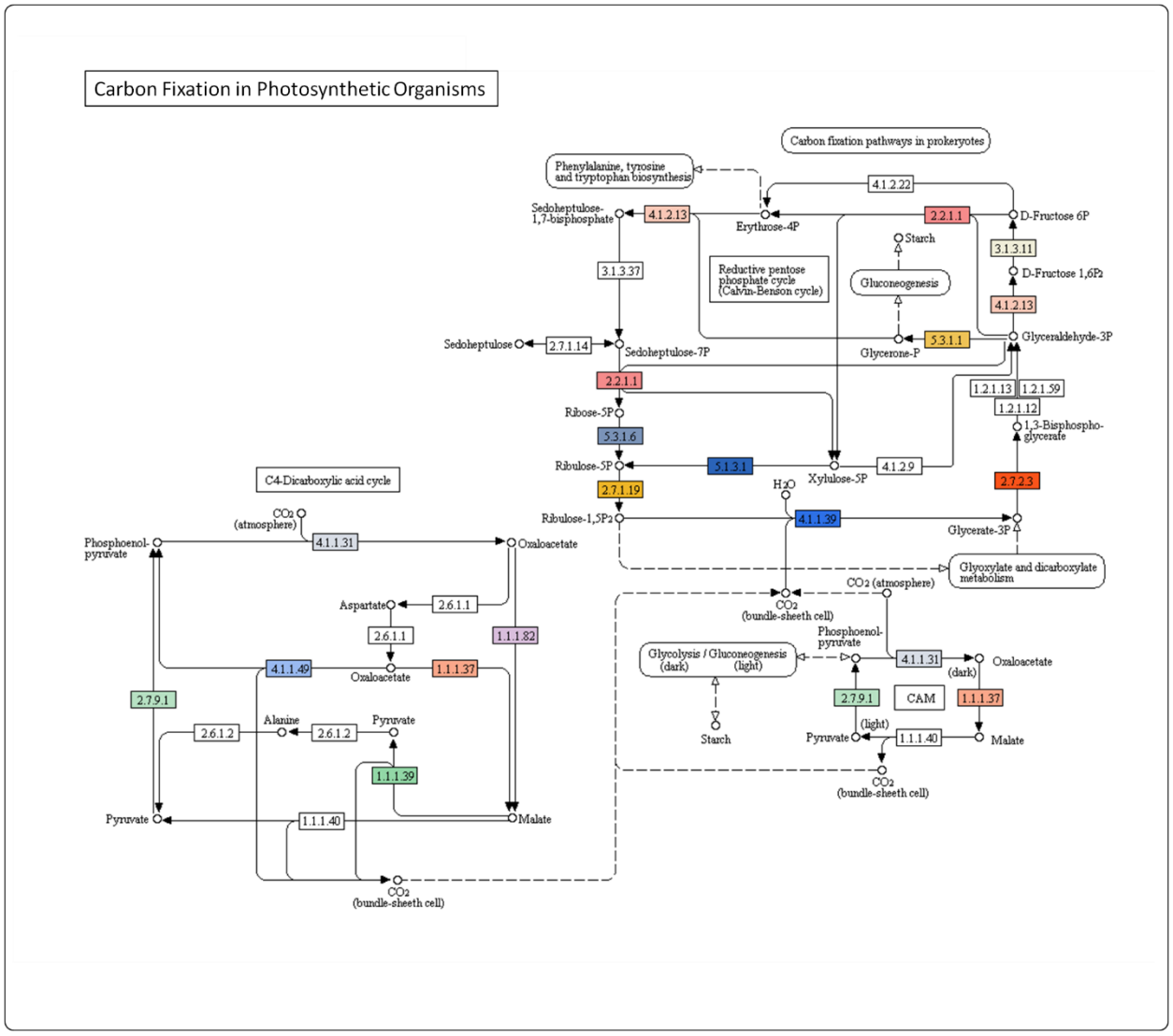

Fig. 7 Carbon fixation pathway genes found in *A. donax* leaf transcriptome are depicted by the different colored ECs (one color for each EC).

4.7 Interspecies comparisons of *A. donax* pathways

Since the KEGG pathways accumulate enzymes and possible conversions of many species it is difficult to estimate how this complexity relates to the presence of pathway components in single species. Therefore, sequence similarity of the *A. donax* transcripts to its most related species *S. italica*, *O. sativa*, *S. bicolor*, and the plant model species *A. thaliana* was analyzed. Transcripts involved in the biosynthesis of certain cell wall compounds (cellulose and lignin), in stomatal development, and transcripts encoding for SAPs homologous to those species were identified in the *A. donax* leaf transcriptome.

The *cellulose synthase* superfamily encompasses thirteen *cellulose synthase A* (*CesA*) genes and thirty six *cellulose synthase-like* (*Csl*) genes in *S. italica* (Muthamilarasan *et al*., 2015). In *A. thaliana*, the *cellulose synthase* superfamily includes ten *CesA* genes and thirty *Csl* genes (Richmond and Somerville, 2000). Among these genes involved in the biosynthesis of cellulose, *A. donax* transcripts homologous to nine out of thirteen *CesA* genes and two to *Csl* genes of *S. italica* were detected, as well as three genes of the *A. thaliana Ces* superfamily (Table 5), confirming the high homology of *A. donax* to *S. italica*.

Table 5 Comparison between *A. donax* transcriptome and *S. italica* and *A. thaliana* transcripts for cellulose biosynthesis

Gene symbol	Gene name	Accession no.[a]	Species	No. of *A. donax* transcripts
CesA1	*cellulose synthase A1*	Seita.5G122700.1	*S. italica*	7
CesA2	*cellulose synthase A2*	AT4G39350.1	*A. thaliana*	6
CesA2	*cellulose synthase A2*	Seita.4G211600.1	*S. italica*	4
CesA3	*cellulose synthase A3*	Seita.2G115400.1	*S. italica*	6
CesA4	*cellulose synthase A4*	Seita.9G227400.1	*S. italica*	1
CesA5	*cellulose synthase A5*	Seita.9G020600.1	*S. italica*	6
CesA6	*cellulose synthase A6*	AT5G64740.1	*A. thaliana*	7
CesA6	*cellulose synthase A6*	Seita.3G332300.1	*S. italica*	4
CesA8	*cellulose synthase A8*	Seita.5G319100.1	*S. italica*	3
CslB1	*cellulose synthase-like B1*	Seita.1G268900.1	*S. italica*	3
CslD1	*cellulose synthase-like D1*	AT2G33100.2	*A. thaliana*	2
CslE6	*cellulose synthase-like E6*	Seita.2G243900.1	*S. italica*	2

A total of nine *S. italica* transcripts and three *A. thaliana* transcripts involved in cellulose biosynthesis showed homology with *A. donax* transcripts.[a] Phytozome (v11.0) accession numbers.

For most genes, of both *S. italica* and *A. thaliana*, multiple homologous leaf transcripts of *A. donax* were retrieved, with a maximum of seven copies for *CesA1*, homologous to *S. italica*, and *CesA6*, homologous to *A. thaliana*. In line with the high homology of *A. donax* to the *S. italica* cellulose biosynthesis pathway, all ten lignin biosynthesis genes known for *S. italica* were also detected in the *A. donax* transcriptome (Table 6). Again, high copy numbers were detected for key genes in the lignin biosynthesis pathway, such as for *phenylalanine ammonia-lyase* (*PAL*) (13 copies) and *trans-cinnamate 4-hydroxylase* (*C4H*) (8 copies). These high copy numbers further stress the importance of key genes in biomass composition in *A. donax*.

Table 6 Comparison between *A. donax* transcriptome and *S. italica* transcripts for lignin biosynthesis

Gene symbol	Gene name	Accession no.[a]	Species	No. of *A. donax* transcripts
PAL	*phenylalanine ammonia-lyase*	Seita.1G240400.1	*S. italica*	13
CCoAOMT	*caffeoyl-CoA 3-O-methyl-transferase*	Seita.6G197700.1	*S. italica*	1
CCR	*cinnamoyl-CoA reductase*	Seita.2G256200.1	*S. italica*	1
CAD	*hydroxycinnamyl alcohol dehydrogenase*	Seita.1G057300.1	*S. italica*	1
COMT	*caffeate O-methyltransferase*	Seita.6G059400.1	*S. italica*	3
F5H	*ferulate 5-hydroxylase*	Seita.9G193900.1	*S. italica*	5
HCT	*hydroxycinnamoyl transferase*	Seita.7G155700.1	*S. italica*	7
C3H	*trans-cinnamate 3-hydroxylase*	Seita.3G194300.1	*S. italica*	6
C4H	*trans-cinnamate 4-hydroxylase*	Seita.5G361200.1	*S. italica*	8
4CL	*4 coumarate CoA-ligase*	Seita.6G093400.1	*S. italica*	3

A total of ten *S. italica* transcripts involved in lignin biosynthesis showed homology with *A. donax* transcripts.
[a] Phytozome (v11.0) accession numbers.

The presence of *A. donax* leaf transcripts homologous to *S. italica* and *A. thaliana* genes involved in stomatal development and distribution (Table 7), key genes in the response to

drought stress, was further assessed. Five of the seven genes [*erecta* (*ER*), *epidermal patterning factor-like protein 9* (*EPLF9/STOMAGEN*), *erecta-like1* (*ERL1*), *GT2-like1* (*GTL1*), *FMA/bHLH097* (*FAMA*)] known to be involved in stomatal development in *S. italica* showed sequence similarity with *A. donax* leaf transcripts, whereas no sequence similarity was detected between *A. donax* transcripts and *A. thaliana* genes. These results further highlight the phylogenetic distances between the investigated species.

Table 7 Comparison between *A. donax* transcriptome and *S. italica* transcripts for stomatal development and distribution

Gene symbol	Gene name	Accession no.[a]	Species	No. of *A. donax* transcripts
ER	*erecta*	Seita.4G086700.1	*S. italica*	5
EPLF9/ STOMAGEN	*epidermal patterning factor-like protein 9*	Seita.5G425700.1	*S. italica*	1
GTL1	*GT2-like1*	Seita.7G184400.1	*S. italica*	1
ERL1	*erecta-like1*	Seita.4G019700.1	*S. italica*	2
FAMA	*FMA/bHLH097*	XM_004961002.3	*S. italica*	5

A total of six *S. italica* transcripts involved in stomatal development and distribution showed homology with *A. donax* transcripts.

[a] Phytozome (v11.0) and NCBI accession numbers.

SAPs are fast emerging as potential gene family candidates for systems and synthetic biology approaches to improve multiple abiotic stress tolerance in plants. Thus, they would be invaluable in the improvement of bioenergy crops such as *A. donax*. So far, however, until relatively recently, such genes had only been identified in few plant species (for example, in rice and sorghum). Therefore, the homology of *A. donax* leaf transcripts to genes encoding SAPs not only in the *A. thaliana* and *S. italica*, but also in *O. sativa*, and *S. bicolor* transcript catalogs was investigated (Table 8).

Table 8 Comparison between *A. donax* transcriptome and *O. sativa* and *S. bicolor* transcripts for SAPs

Gene symbol	Gene name	Accession no.[a]	Species	No. of *A. donax* transcripts

SAP1	*stress associated protein 1*	LOC_Os09g31200.1	*O. sativa*	1
SAP4	*stress associated protein 4*	LOC_Os02g10200.1	*O. sativa*	1
SAP4	*stress associated protein 4*	Sobic.004G079100.2	*S. bicolor*	3
SAP5	*stress associated protein 5*	Sobic.002G245800.2	*S. bicolor*	2
SAP8	*stress associated protein 8*	LOC_Os06g41010.1	*O. sativa*	3
SAP9	*stress associated protein 9*	Sobic.002G046100.1	*S. bicolor*	1
SAP9	*stress associated protein 9*	LOC_Os07g07350.1	*O. sativa*	1
SAP11	*stress associated protein 11*	Sobic.002G345300.2	*S. bicolor*	3
SAP11	*stress associated protein 11*	LOC_Os08g39450.1	*O. sativa*	1
SAP15	*stress associated protein 15*	LOC_Os05g23470.1	*O. sativa*	1
SAP16	*stress associated protein 16*	LOC_Os07g38240.1	*O. sativa*	3

A total of seven *O. sativa* transcripts and four *S. bicolor* transcripts encoding for SAPs showed homology with *A. donax* transcripts.

[a] Phytozome (v11.0) accession numbers.

A. donax transcripts homologous to seven of the 18 genes encoding for *O. sativa* SAPs, and to four of the *S. bicolor* SAPs, were also identified, indicating a stronger phylogenetic relationship between *A. donax* and *S. bicolor* than between *A. donax* and *O. sativa*. No sequence similarity was detected between *A. donax* transcripts and those of *S. italica* and *A. thaliana*. Although multiple copies of several *A. donax* transcripts were identified again, this number was lower than for most processes described above, and might well be explained by the estimated high ploidy level of *A. donax*. Given the size of the SAP family and the low copy numbers detected this suggests that paralogous genes acquired substantial sequence diversity and neofunctionalization.

4.8 Genic-SSR marker mining and SSR polymorphism detection

Many of the sequenced mRNA-transcripts contained untranslated regions (UTRs) and occasional remaining introns. Such noncoding regions are prone to sequence diversity, which might be a source for potential polymorphic markers. Currently, no genetic markers are available for *A. donax*. Hence, the retrieved transcripts were mined for SSR markers, which are commonly used as a tool for assessing genetic diversity, the development of genetic maps, marker assisted selection (MAS) breeding, and comparative genomics. The 62,596 unitranscripts, accounting for 52.7 Mbp, were screened for repeat motifs to explore the SSR profile in the *A. donax* leaf transcriptome. A total of 8,364 SSRs were obtained from 7,491 unitranscript sequences (11.97%) so 775 sequences contained more than one SSR. In

addition, 294 sequences revealed compound SSR formation. A strong enrichment of mono- and tri-nucleotide repeat motifs was further detected among each of the SSR classes, while longer repeat motifs were clearly underrepresented (Table 9).

Table 9 Frequency of SSR classes in the *A. donax* leaf transcriptome

Search item	**Number**
Mono-nucleotide	4,170
Di-nucleotide	1,060
Tri-nucleotide	3,034
Tetra-nucleotide	57
Penta-nucleotide	29
Hexa-nucleotide	14
Total number of identified SSRs	8,364
Number of SSR containing sequences	7,491
Number of sequences containing more than 1 SSR	775
Number of SSRs present in compound formation	294

Summary statistics of repeat motifs identified in the *A. donax* leaf unitranscripts and frequency of different repeat type classes.

Within each SSR class strong nucleotide prevalence was observed. Thymine (2,326 SSRs; 55.78%) and adenine (1,413 SSRs; 33.88%) SSRs were most abundant in the mono-nucleotide class, followed by guanine (253 SSRs; 6.07%) and cytosine (178 SSRs; 4.27%) SSRs. The number of repeats within mono-nucleotide motifs ranged from 10 to a maximum of 31 repeats, and the highest frequencies observed were $(A)_{10}$, $(T)_{10}$, $(C)_{12}$, and $(G)_{11}$. Of the two possible types of mono-nucleotide motifs, considering sequence complementary, A/T repeats were considerably more abundant (3,739 SSRs; 44.70%), than C/G repeats (431 SSRs; 5.15%). However, the $(A/T)_{10}$ and $(C/G)_{11}$ repeats, accounting for 36.32% and 24.60%, respectively, were most abundant within mono-nucleotide motifs. With respect to di-nucleotide motifs, the GA (243 SSRs; 22.92%) and CT (172 SSRs; 16.22%) repeats were most represented within the 1,060 di-nucleotide motifs, followed by AG (164 SSRs; 15.47%) and TC (137 SSRs; 12.92%) repeats. Within the GA repeat type, the $(GA)_6$ motif was the most common (56 SSRs; 23.04%). Considering sequence complementary, AG/CT repeats were the most abundant within di-nucleotide motif (716 SSRs; 67.54%) with a maximum frequency of the $(AG/CT)_6$ repeat type (160 SSRs; 22.35%). Within the 3034 tri-nucleotide

repeat motifs, the CGC and CGG repeats were the most common, accounting for 213 (7.02%) and 211 (6.95%) SSRs, respectively. Considering the CGC repeat type, the tri-nucleotide $(CGC)_5$ was the most abundant (155; 72.77%) and, as per sequence complementary, the CCG/CGG motif accounted for 1,197 SSRs (39.45%) with $(CCG/CGG)_5$ occurring in 69.51% of the cases. Among the tetra-, penta- and hexa-nucleotide repeats, the highest frequencies were observed for the repeat types $(AAGG/CCTT)_5$, $(AAGGG/CCCTT)_5$ and $(AATCC/ATTGG)_5$. The frequencies of each of the different SSR repeats as compared to the total identified SSRs are reported in Table 10.

Table 10 SSR repeat motifs in *A. donax* unitranscripts

Motif	SSR number	Frequency (%)[a]
A/T	3,739	44.70
C/G	431	5.15
AC/GT	207	2.47
AG/CT	716	8.56
AT/AT	80	0.96
CG/CG	57	0.68
AAC/GTT	79	0.94
AAG/CTT	190	2.27
AAT/ATT	9	0.11
ACC/GGT	203	2.43
ACG/CGT	146	1.75
ACT/AGT	23	0.27
AGC/CTG	535	6.40
AGG/CCT	525	6.28
ATC/ATG	127	1.52
CCG/CGG	1,197	14.31
Others	100	1.20
Total	8,364	100

Amount of different SSR repeat motifs identified in *A. donax* leaf unitranscripts.

[a] Frequency of each repeat motif with respect to the total number of the identified SSRs.

Finally, in order to detect PolySSRs, ecotype-specific assemblies were reconstructed using the Trinity software. Using CandiSSR pipeline, a total of 53 PolySSRs were identified between *A. donax* unitranscripts (hereafter referred as reference transcriptome) and the three ecotype-specific transcriptomes, suggesting the existence of a certain degree of genetic diversity among the three ecotypes analysed in this study. The specific primer pairs for the 53 PolySSRs were designed using Primer3, which is implemented in the CandiSSR Perl script. The polymorphism encompassed di-nucleotide motifs with a minimum of six repeats and tri-nucleotide motifs with a minimum of five repeats in the reference transcriptome. The total number of PolySSR motifs ranged from 1 to 4 and the minimum and maximum number of repeats were 2 for AGC, GTT and TCC motifs, and 11 for AG, respectively (Table 11).

Table 11 PolySSR motifs in the reference and ecotype-specific transcriptomes

Motif	**SSR_number**[a]	**Min_rep**[b]	**Max_rep**[c]
AAG	1	3	5
AG	1	8	11
AGA	1	4	5
AGC	3	2	5
AGG	1	6	7
AT	2	3	6
CAG	2	4	5
CAT	4	3	5
CCA	1	4	5
CCG	1	3	5
CCT	2	5	6
CGA	1	4	5
CGC	2	3	5
CTC	2	4	6
CTG	4	3	6
CTT	1	3	5
GAA	1	4	5
GAG	2	3	6
GAT	1	4	5
GC	1	6	7
GCA	1	5	8
GCC	1	4	5
GCG	1	4	5
GCT	3	4	5

GGA	1	4	5
GGC	3	3	9
GGT	1	3	5
GTT	1	2	5
TC	2	3	6
TCC	1	2	5
TCT	1	4	5
TGC	2	4	5
TGG	1	4	5

Total count of each PolySSR motif with the minimum and the maximum number of repeats in the reference and ecotype-specific transcriptomes.

[a] Total number of PolySSR motifs: 53. [b] Minimum number of repeats. [c] Maximum number of repeats

4.9 Gene functions of the unitranscript sequences containing PolySSRs

To da To date, no gene-related markers have been identified in *A. donax*. Functional categorization of unitranscripts containing SSRs were therefore conducted in order to reveal homology with known proteins and determine the possible functions of the 53 polymorphic genic-SSRs by BLAST2GO analysis with the NR database of protein sequences. A total of 31 GO terms were allocated to the transcripts containing the PolySSRs. They were assigned to the BP (157 sequences, 20 GOs), CC (132 sequences, 8 GOs), and MF (49 sequences, 3 GOs) categories. The most represented subcategories in BP were metabolic process (20.38%), cellular process (19.11%), single-organism process (14.01%), regulation of biological process (9.55%) and response to stimuli (5.10%). Among the CC category, the subcategories cell and cell-part (26.51%), organelle (20.45%) and membrane (9.85%) were the most abundantly represented GO terms. Among the MF category, transcripts with binding (53.06%) and catalytic (40.81%) activities were retrieved. In addition, nineteen KEGG metabolic pathways were represented by at least one transcripts with the corresponding EC numbers. Particularly, the phenylpropanoid biosynthesis, starch and sucrose metabolism, purine and thiamine metabolisms were overrepresented.

Interestingly, functional annotation revealed that unitranscripts containing PolySSRs belonged to key genes involved in the hydrolysis of ABA glucose ester (ABA-GE) to free ABA (i.e. *β-glucosidase 10*) (Wang *et al*., 2011), in the positive regulation of ABA signaling pathway (i.e. *3-ketoacyl-CoA thiolase*) (Jiang *et al*., 2011), and to several transcription factors (TFs) involved in enhancing drought tolerance in plants, such as myeloblastosis (MYB) TFs (i.e. MYB1R1) (Baldoni *et al*., 2015) and heat stress TFs (i.e. HSFA3) (Guo *et al*., 2016). Furthermore, among the positive hits were genes known to be involved in photoperiod

perception, circadian regulation and regulation of flowering, such as *early flowering 4-like* genes (i.e. *ELF4- like 4*) (Kolmos *et al*., 2009), and in the *Agrobacterium tumefaciens*-mediated genetic transformation, such as *VirE2 interacting protein2* (i.e. *VIP2*) (Anand *et al*., 2007). Furthermore, homologous sequences to *glutathione S-transferase* (*GST*) genes, whose detoxification and antioxidant activities can be considered important factors in drought tolerance in plants, such as in *H. vulgare* (Rezaei *et al*., 2013) and in transgenic *A. thaliana* (Xu *et al*., 2015) were also found among the unitranscripts containing PolySSRs. These findings make these functional PolySSRs particularly useful for research into accelerated genetic studies and enhanced breeding programs in *A. donax*.

4.10 Experimental validation of PolySSR markers

To test whether the potential SSR loci that were mined were true-to-type ones for use in genetic diversity studies, a total of five primer pairs were randomly selected from those obtained from the CandiSSR pipeline and synthesized (see Material and Methods section). The results showed that all tested primer pairs amplified successfully. Out of the five pairs of primer, four produced amplicons with expected sizes, while one primer pair gave an amplicon larger than the expected size (i.e. CPSSR_1|AAG). Consequently, the former pairs of primer were then used to assess the genetic diversity among the three *A. donax* ecotypes. Out of the four primer pairs, one (i.e. CPSSR_4|AGC) successfully amplified polymorphic SSR fragment; particularly, $(AGC/GCT)_5$ repeats were detected in EcoA, whereas $(AGC/GCT)_4$ repeats were identified in EcoB and EcoC.

4.11 Differential gene expression analysis in three *A. donax* ecotypes in response to drought

In total, 2,641 transcripts were differentially expressed ($\geq$ 2 fold change and *p*-value 0.05) between treatments, regardless of ecotype. An higher number of DETs was retrieved in the first timepoint (T1; 2,136 DETs), followed by T2 (296 DETs) and T3 (209 DETs) timepoints.

Gene expression experienced an ecotype-specific variation under moderate drought stress condition. The total number of DETs for each ecotype in the three timepoints is reported in Table 12 and the up- and down- regulation is shown in Fig.8 and Fig. 9. Particularly, in EcoA at T1 the numbers of transcripts with decreased and increased expression were 1,725 and 79, respectively (Fig. 8a). In the two following timepoints, the numbers of DETs considerably

decreased to 32 (12 up-regulated and 20 down-regulated) and 90 (60 up-regulated, 30 down-regulated) in T2 and in T3, respectively (Table 12; Fig. 8a; Fig. 9). Likewise, in EcoB at T1 202 and 99 transcripts were down-regulated and up-regulated, respectively; at T2 the number of DETs slightly dropped off to 159 (73 up-regulated and 86 down-regulated), and at T3 only 37 transcripts (17 up-regulated, 14 down-regulated) were differentially expressed (Table 12; Fig. 8b; Fig. 10). On the contrary, EcoC showed a different trend in gene expression change, with only 31 DETs (17 up-regulated and 14 down-regulated) at T1, whereas 105 (90 up-regulated and 15 down-regulated) transcripts at T2 and 82 (19 up-regulated and 63 down-regulated) transcripts at T3 were differentially expressed (Table 12; Fig. 8c; Fig. 11).

Table 12 Total number of DETs in the three *A. donax* ecotypes at different timepoints

Ecotype	Timepoint[a]	DETs[b]
EcoA	T1	1804
	T2	32
	T3	90
EcoB	T1	301
	T2	159
	T3	37
EcoC	T1	31
	T2	105
	T3	82

[a] Timepoint: T1, 1st timepoint; T2, 2nd timepoint; T3, 3th timepoint.
[b] DETs: differentially expressed transcripts.

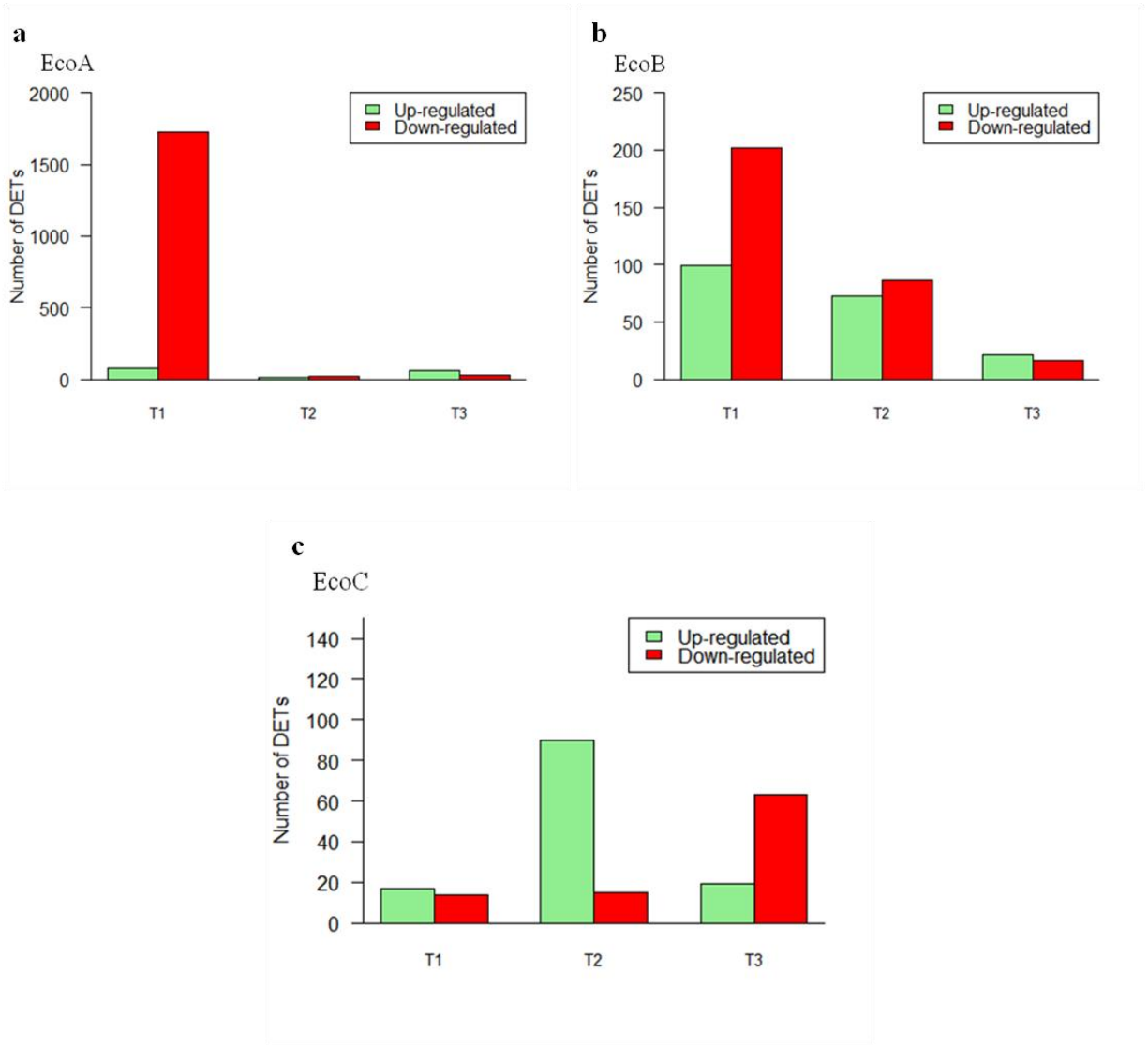

Fig. 8 Differentially expressed transcripts in the three *A. donax* ecotypes (EcoA, EcoB and EcoC) grown under drought stress at three timepoints (T1, T2 and T3). Number of up- and down-regulated transcripts in EcoA (a), EcoB (b) and EcoC (c). The x-axis and y-axis show the number of DETs and the three timepoint, respectively. Green and red bars showed the up-regulation and the down-regulation (≥ 2 fold and *p*-value > 0.05), respectively.

In order to assess the magnitude of gene expression variation under drought stress condition, five log(fold change) (logFC) ranges were used to partition DETs(Table 13). At T1, the highest numbers of over- and under-expressed transcripts were retrieved in the logFC ranges from 1 to 2 (i.e. 2- to 4-fold) and from 2 to 4 (i.e. 4- to 16-fold) in all the ecotypes. Further, a slight number of differentially expressed transcripts was observed only for EcoA in the logFC ranges from 4 to 6 (16- to 64-fold) and 7 to 10 (128- to 1024-fold). Similarly, at T2 and T3 the logFC from 1 to 2 and 2 to 4 showed the highest number of over- and under-

regulated transcripts in all the three ecotypes. Finally, only for EcoB at T2 three transcripts showed differential expression with fold change greater than 64-fold (Table 13).

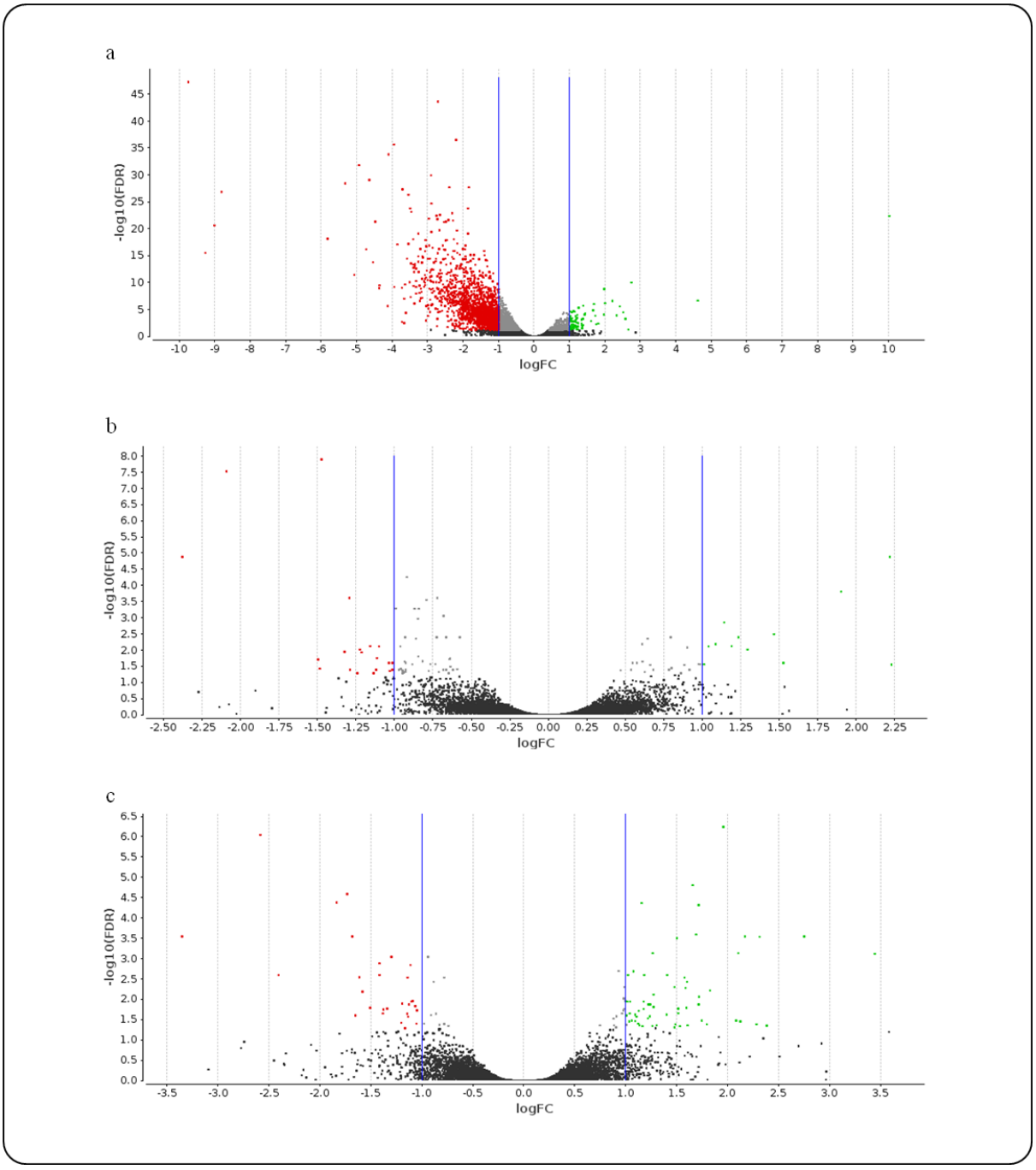

Fig. 9 Volcano plots showing DETs in EcoA at T1 (a), at T2 (b), and at T3 (c). The blue lines show logFC < -1 and logFC > 1. The red and green dots represent DETs with logFC < -1 and logFC > 1, respectively, and both with FDR < 0.05. The grey and black dots represent transcripts with FDR < 0.05 and logFC > -1, and FDR > 0.05 and logFC > 1, respectively.

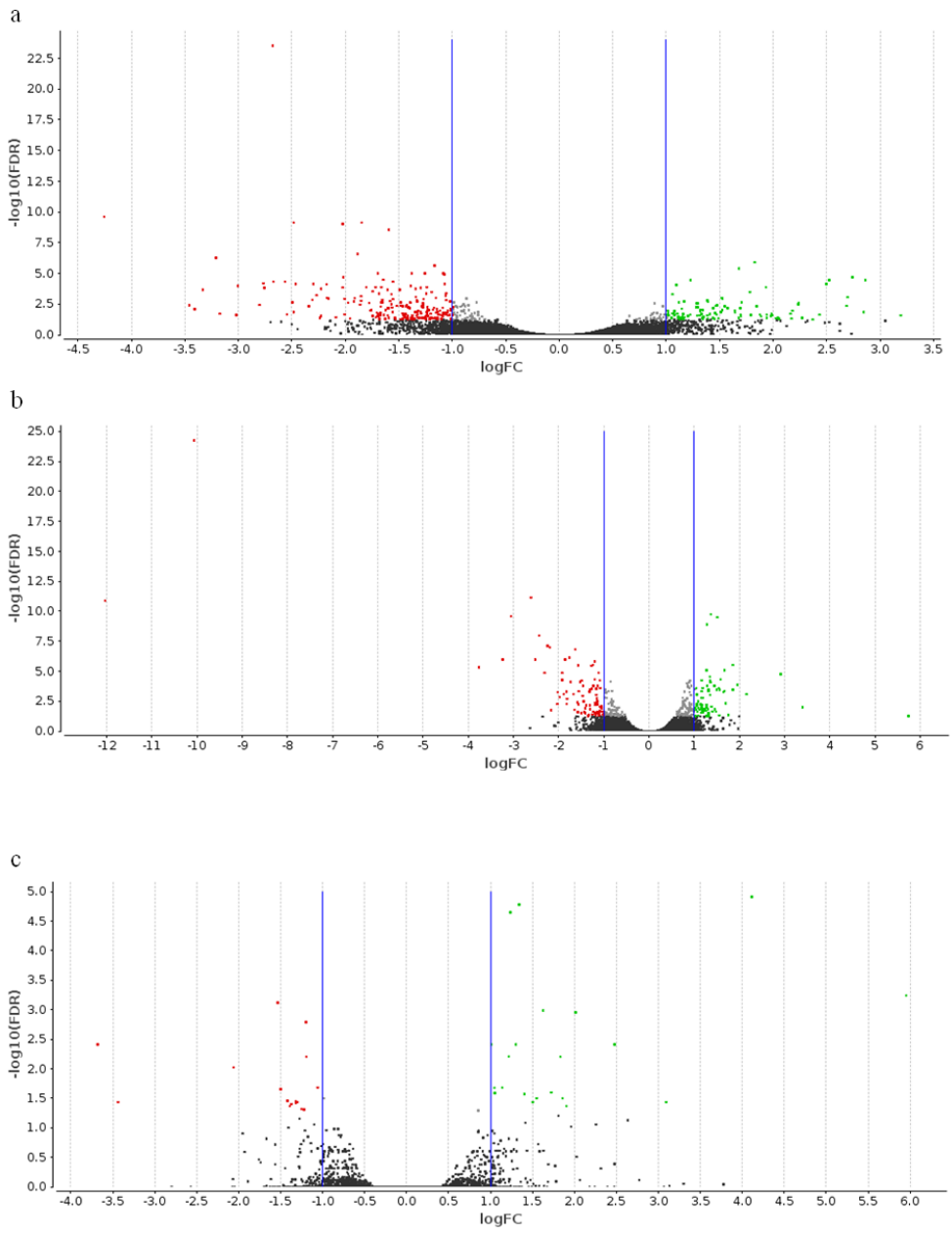

Fig. 10 Volcano plots showing DETs in EcoB at T1 (a), at T2 (b), and at T3 (c). The blue lines show logFC < -1 and logFC > 1. The red and green dots represent DETs with logFC < -1 and logFC > 1, respectively, and both with FDR < 0.05. The grey and black dots represent transcripts with FDR < 0.05 and logFC > -1, and FDR > 0.05 and logFC > 1, respectively.

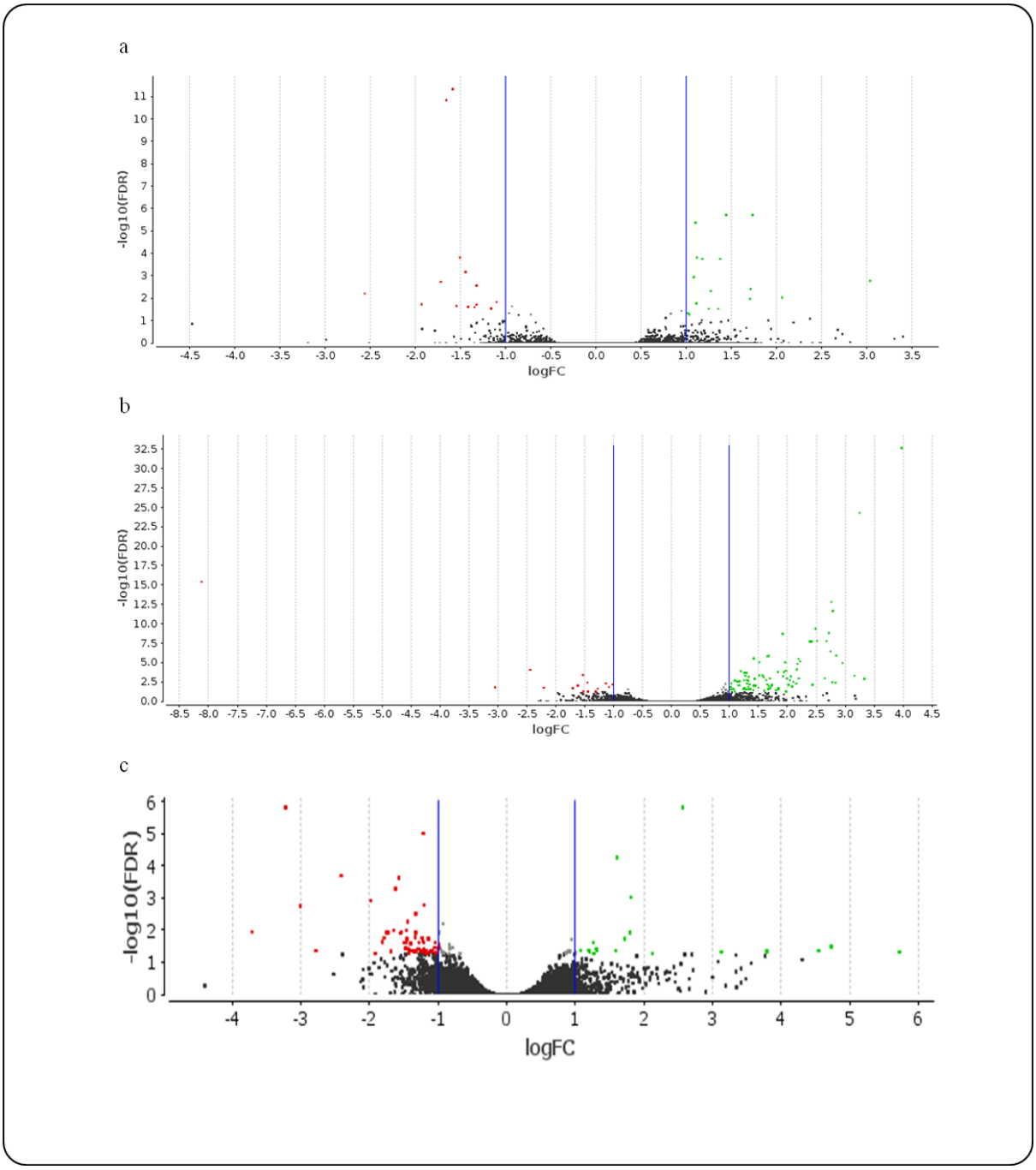

Fig. 11 Volcano plots showing DETs in EcoC at T1 (a), at T2 (b), and at T3 (c). The blue lines show logFC < -1 and logFC > 1. The red and green dots represent DETs with logFC < -1 and logFC > 1, respectively, and both with FDR < 0.05. The grey and black dots represent transcripts with FDR < 0.05 and logFC > -1, and FDR > 0.05 and logFC > 1, respectively.

Table 13 Number of over-expressed and under-expressed transcripts partitioned in their specific logFC ranges in each ecotype at the three timepoints

		T1[b]		
Expression change	**logFC range[a]**	**EcoA**	**EcoB**	**EcoC**
Over-expressed	10;12	0	0	0
	7;10	1	0	0
	4;6	1	0	0
	2;4	7	19	2
	1;2	70	80	15
Under-expressed	-10;-12	0	0	0
	-7;-10	4	0	0
	-4;-6	12	1	0
	-2;-4	384	34	1
	-1;-2	1325	167	13
		T2[b]		
Expression change	**logFC range[a]**	**EcoA**	**EcoB**	**EcoC**
Over-expressed	10;12	0	0	0
	7;10	0	0	0
	4;6	0	1	0
	2;4	2	3	30
	1;2	10	69	60
Under-expressed	-10;-12	0	2	0
	-7;-10	0	0	1
	-4;-6	0	0	0
	-2;-4	2	12	3
	-1;-2	18	72	11
		T3[b]		
Expression change	**logFC range[a]**	**EcoA**	**EcoB**	**EcoC**
Over-expressed	10;12	0	0	0
	7;10	0	0	0
	4;6	0	2	3
	2;4	9	3	4
	1;2	51	16	12
Under-expressed	-10;-12	0	0	0
	-7;-10	0	0	0
	-4;-6	0	0	0
	-2;-4	3	3	5
	-1;-2	27	13	58

Results were partitioned between over- and under- expressed transcripts using the logFC as parameter to measure the magnitude of expression difference in each ecotype and timepoint.

[a] log(fold change); The logFC is reported as a range between two values, which are separated by the semicolon.

[b] T1: 1st timepoint; T2: 2nd timepoint; T3: 3rd timepoint.

In order to decipher putative common molecular strategies involved in the drought stress response in the three *A. donax* ecotypes, the common DETs among them were investigated. Only a slight number of DETs (3 transcripts) was common between ecotypes, whereas the largest number of DETs are unique in each ecotype (Fig. 10). Among the three common DETs, two transcripts encode for pyrophosphate:fructose 6-phosphate 1-phosphotransferase (PFP) and for jasmonate-induced proteins (JIPs). The transcript encoding for JIPs were over-expressed in EcoA (3-fold) and EcoC (3-fold) at T1; in EcoB (3-fold) and EcoC (10-fold) at T2, and only in EcoC at T3 (52-fold). The transcript encoding for PFP were up-regulated in all the ecotypes at T1with a fold change value of 2. In the other two timepoints this transcript showed no significant differential expression between treatments. The other transcript showed homology with uncharacterized sequence in the Viridiplantae subset of NCBI NR database.

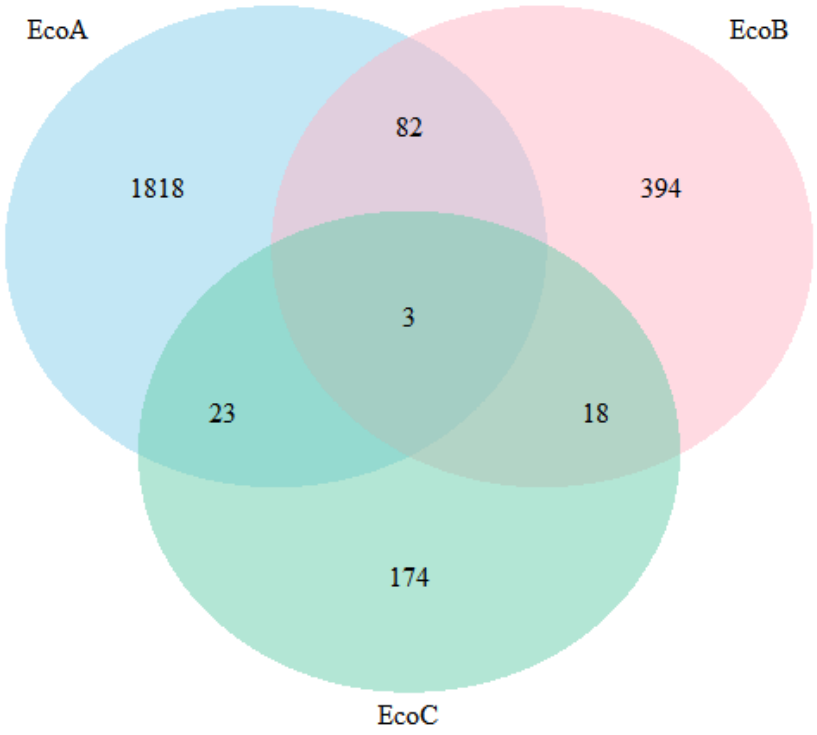

Fig. 12 Venn diagram of total DETs in the three *A. donax* ecotypes. Common and unique transcripts between the ecotypes are shown. Blue, pink and aquamarine circles represent EcoA, EcoB and EcoC, respectively.

In order to investigate transcriptional mechanisms and biological pathways involved in each ecotype-specific drought stress response, the DETs of each ecotype at each timepoint were further characterized by the analysis of KEGG pathways.

Among the KEGG pathways altered under drought stress, for EcoA at T1 the metabolism of terpenoids and polyketides and other secondary metabolites encompassed 27 enzymes, of which 25 and 2 were under- and over-expressed, respectively. Also the second most altered pathway was the purine metabolism, in which 9 and 3 transcripts were down- and up-regulated, respectively. Furthermore, another pathway affected by drought was the starch and sucrose metabolism, in which 11 enzymes (9 down-regulated and 2 up-regulated) were differentially expressed between treatments. Conversely, at T2 only one enzyme of the aminobenzoate degradation pathway was down-regulated under drought stress for EcoA (Table 14). At T3, glutathione, terpenoids and polyketides, and diterpenoid metabolisms were the most affected in EcoA, with 3 enzymes that were down-regulated under drought condition (Table 14). The remaining altered pathways for EcoA at T1 and T3 are shown in Table 14.

Table 14 The most represented KEGG pathways in EcoA at different timepoints

Timepoint	Pathway	Under-expressed transcripts/enzymes in pathway	Over-expressed transcripts/enzymes in pathway
T1	Metabolism of terpenoids and polyketides and other secondary metabolites	25	2
	Purine metabolism	9	3
	Starch and sucrose metabolism	9	2
	Sulfur metabolism	9	0
	Fructose and mannose metabolism	6	2
	Glycine, serine and threonine metabolism	7	1
	Phenylpropanoid biosynthesis	5	1
	Porphirin and chlorophyll metabolism	6	0
	Pyruvate metabolism	6	0
	Cysteine and methionine metabolism	6	0
T2	Aminobenzoate degradation	1	0
T3	Glutathione metabolism	3	0
	Metabolism of terpenoids and polyketides and other secondary	3	0

metabolites		
Diterpenoid biosynthesis	3	0
Aminobenzoate degradation	1	1
Purine metabolism	2	0
Pentose phosphate pathway	2	0
Porphryn and chlorophyll metabolism	2	0
Matabolism of xenobiotics by cytochrome P450	1	0
Tryptophan metabolism	1	0
Lisyne degradation	1	0

The top KEGG pathways in EcoA at each timepoint (T1, T2 and T3).

The metabolism of terpenoids and polyketides and other secondary metabolites experienced also an alteration in the expression of 11 transcripts (9 under-expressed and 2 over-expressed) for EcoB at T1 and T2, as well (Table 15). In addition, at T1 glycolysis/gluconeogenesis was the most second altered pathway for EcoB, in which 6 and 2 transcripts showed a down- and up-regulation, respectively (Table 15). In the starch and sucrose metabolism 7 transcripts (5 down-regulated and 2 up-regulated) showed differential expression for EcoB at T1 (Table 15). Again for EcoB in T2, the phenylpropanoid biosynthesis pathway showed 3 down-regulated transcripts. In the same timepoint and ecotype, both the aminobenzoate degradation and the metabolism of terpenoids and polyketides encompassed 2 down-regulated transcripts. Moreover, at T3 the starch and sucrose metabolism and the amino sugar and nucleotide sugar metabolism were the most represented in terms of number of DETs (i.e. enzymes) in pathways, with 2 up-regulated and 2 down-regulated transcripts, respectively (Table 15). The remaining altered pathways for EcoB at T1, T2 and T3 are shown in Table 15.

1442 **Table 15** The most represented KEGG pathways in EcoB at different timepoints

Timepoint	Pathway	Under-expressed transcripts/enzymes in pathway	Over-expressed transcripts/enzymes in pathway
T1	Metabolism of terpenoids and polyketides and other secondary metabolites	9	2
	Glycolysis/Gluconeogenesis	6	2
	Starch and sucrose metabolism	5	2
	Purine metabolism	4	2
	Pyruvate metabolism	5	0
	Amino sugar and nucleotide sugar metabolism	5	0
	Fructose and mannose metabolism	3	2
	Cysteine and methionine metababolism	3	1
	Sulfur metabolism	4	0
	Carbon fixation in photosynthetic organism	3	0
T2	Phenylpropanoid biosynthesis	3	0
	Aminobenzoate degradation	2	0
	Metabolism of terpenoids and polyketides and other secondary metabolites	2	0
	Purine metabolism	0	2
	Phenylalanine metabolism	2	0
	Ubiquinone and other terpenoid-quinone biosynthesis	2	0
	Drug metabolism - cytochrome P450	2	0
	Thiamine metabolism	0	1
	Phenylalanine, tyrosine and tryptophan biosynthesis	1	0
	Glycerophospholipid metabolism	1	0
T3	Starch and sucrose metabolism	0	2
	Amino sugar and nucleotide sugar metabolism	2	
	Phenylpropanoid biosynthesis	0	1

	Purine metabolism	0	1
	Steroid biosynthesis	1	
	Cyanoamino acid metabolism	0	1
	Glycerophospholipid metabolism	0	1
	Fructose and mannose metabolism	1	
	Thiamine metabolism	0	1

1443 The top KEGG pathways in EcoB at each timepoint (T1, T2 and T3).

1444

1445 Finally, for EcoC at T1 and T2 the metabolism of terpenoids and polyketides and other
1446 secondary metabolites included 3 and 6 over-expressed transcripts, respectively, followed by
1447 the glycolysis/gluconeogenesis and pentose phosphate pathways, in which 2 transcripts were
1448 up-regulated. In T2, the phenylalanine, tyrosine and tryptophan biosynthesis and the alpha-
1449 linolenic acid metabolism contained 5 and 3 over-expressed transcripts under drought
1450 condition (Table 16). The remaining altered pathways for EcoC at T1, T2 and T3 are shown
1451 in Table 16.

1452

1453 **Table 16** The most represented KEGG pathways in EcoC at different timepoints

Timepoint	**Pathway**	**Under-expressed transcripts/enzymes in pathway**	**Over-expressed transcripts/enzymes in pathway**
T1	Metabolism of terpenoids and polyketides and other secondary metabolites	0	3
	Glycolysis/Gluconeogenesis	0	2
	Pentose phosphate pathway	0	2
	Fructose and mannose metabolism	0	2
	Nitrogen metabolism	0	1
	Alanine, aspartate and glutamate metabolism	0	1
	Starch and sucrose metabolism	1	
	Aminobenzoate degradation	0	1
	Methane metabolism	0	1
	Galactose metabolism	0	1

T2	Metabolism of terpenoids and polyketides and other secondary metabolites	0	6
	Phenylalanine, tyrosine and tryptophan biosynthesis	0	5
	alpha-Linolenic acid metabolism	0	3
	Inositol phosphate metabolism	0	3
	Phosphatidylinositol signaling system	0	2
	Purine metabolism	2	0
	Cysteine and methionine metabolism	0	2
	Thiamine metabolism	1	0
	Carbon fixation in photosynthetic organisms	0	1
	Starch and sucrose metabolism	0	1
T3	Metabolism of terpenoids and polyketides and other secondary metabolites	2	
	Citrate cycle (TCA cycle)	2	
	Nitrogen metabolism	0	2
	Purine metabolism	2	
	Glycerophospholipid metabolism	2	
	Phenylpropanoid biosynthesis	0	1
	Thiamine metabolism	1	
	Photosynthesis	1	
	mTOR signaling pathway	0	1

The top KEGG pathways in EcoC at each timepoint (T1, T2 and T3).

The DETs of each ecotype at each timepoint were also characterized by mapping them against GO database using Blast2GO software.

Among the GO categories in EcoA, at T1 the CC was the most abundant category in terms of GO assignments (45.27%, 4151 sequences), followed by BP (37.44%, 3433 sequences) MF CC (17.29%, 1585 sequences). Particularly, metabolic process (GO:0008152, 798 sequences), cellular process (GO:0009987, 721 sequences), single-organism process (GO:0044699, 529 sequences), biological regulation (GO:0065007, 268 sequences), regulation of biological

process (GO:0050789, 237 sequences) and response to stimulus (GO:0050896, 224 sequences) were the mostly represented in the BP category. In the MF category, catalytic activity (GO:0003824, 680 sequences), binding (GO:0005488, 634 sequences), transporter activity (GO:0005215, 109 sequences) and nucleic acid binding transcription factor activity (GO:0001071, 48 sequences) were the most represented GO terms. Besides, cell (GO:0005623, 977 sequences), cell part (GO:0044464, 969 sequences), organelle (GO:0043226, 816 sequences) and membrane (GO:0016020, 541 sequences) were the mostly abundant in the CC category. Likewise, at T2 the BP was the most abundant GO category (41.56%, 32 sequences), followed by CC (38.96%, 30 sequences) and MF (19.48%, 15 sequences). In the BP category, metabolic process (GO:0008152, 10 sequences), cellular process (GO:0009987, 9 sequences) and response to stimulus (GO:0050896, 4 sequences) were the mostly represented GO terms. The MF category were mainly comprised of transcripts encoding fir proteins with binding activity (GO:0005488, 7 sequences), catalytic activity (GO: 0003824, 6 sequences) and transporter activity (GO:0005215, 2 sequences). Moreover, cell part (GO:0044464, 6 sequences), membrane (GO:0016020, 6 sequences), cell (GO:0005623, 6 sequences) and organelle (GO:0043226, 5 sequences) were the mostly predominant in the CC category. Finally, at T3 the BP was the most abundant GO category (42.32%, 204 sequences), followed by CC (36.93%, 178 sequences) and MF (20.75%, 100 sequences). Among the BP category, the mainly represented GO terms were cellular process (GO:0009987, 40 sequences), metabolic process (GO:0008152, 40 sequences), single-organism process (GO:0044699, 35 sequences), response to stimulus (GO:0050896, 20 sequences) and biological regulation (GO:0065007, 14 sequences). Binding (GO:0005488, 50 sequences), catalytic activity (GO:0003824, 40 sequences) and signal transducer activity (GO:0004871, 5 sequences) were the mostly represented GO terms in the MF category, whereas cell part (GO:0044464, 44 sequences), cell (GO:0005623, 44 sequences), organelle (GO:0043226, 33 sequences) and membrane (GO:0016020, 23 sequences) were the most abundant in the CC category.

Among the GO categories in EcoB, at T1 the CC category encompassed the majority of the GO assignments (41.15%, 614 sequences), followed by BP (38.14%, 569 sequences) and MF (20.71%, 309 sequences). Metabolic process (GO:0008152, 130 sequences), cellular process (GO:0009987, 125 sequences), single-organism process (GO:0044699, 95 sequences), localization (GO:0051179, 50 sequences), and biological regulation (GO:0065007, 49 sequences) were the main GO terms among the BP category. The MF category is mainly

made up of transcripts encoding for proteins with binding activity (GO:0005488, 131 sequences), catalytic activity (GO:0003824, 125 sequences) and transporter activity (GO:0005215, 31 sequences). The CC category largely encompassed cell (GO:0005623, 144 sequences), cell part (GO:0044464, 143 sequences), organelle (GO:0043226, 110 sequences), membrane (GO:0016020, 90 sequences) and membrane part (GO:0044425, 76 sequences). Moreover, at T2 the CC category encompassed the majority of the GO assignments (45.95%, 454 sequences), followed by BP (396%, 40.08 sequences) and MF (13.97%, 138 sequences). Among the BP category, cellular process (GO:0009987, 87 sequences), metabolic process (GO:0008152, 84 sequences), single-organism process (GO:0044699, 51 sequences), and response to stimulus (GO:0050896, 36 sequences) were the mostly represented terms. In the MF category, binding (GO:0005488, 63 sequences), catalytic activity (GO:0003824, 58 sequences) and transporter activity (GO:0005215, 10 sequences) were the main GO terms, whereas cell part (GO:0044464, 94 sequences), cell (GO:0005623, 94 sequences), organelle (GO:0043226, 83 sequences), membrane (GO:0016020, 67 sequences). Finally at T3, the CC category encompassed the majority of the GO assignments (41.36%, 79 sequences), followed by BP (37.70%, 72 sequences) and MF (20.94%, 40 sequences). Among the BP category, metabolic process (GO:0008152, 18 sequences), cellular process (GO:0009987, 14 sequences), single-organism process (GO:0044699, 14 sequences), localization (GO:0051179, 7 sequences) and biological regulation (GO:0065007, 5 sequences) were the mostly abundant GO assignments. In the MF category, catalytic activity (GO:0003824, 16 sequences), binding (GO:0005488, 16 sequences), and transporter activity (GO:0005215, 5 sequences) were the mostly represented GO terms. Finally cell part (GO:0044464, 20 sequences), cell (GO:0005623, 20 sequences), organelle (GO:0043226, 14 sequences), and membrane (GO:0016020, 10 sequences).

Finally, among the GO categories in EcoC the CC category covered the majority of the GO assignments (42.27%, 52 sequences), followed by BP (38.21%, 47 sequences) and MF (19.51%, 24 sequences) at T1. The BP category largely included cellular process (GO:0009987, 13 sequences), metabolic process (GO:0008152, 13 sequences) and single-organism process (GO:0044699, 8 sequences), whereas catalytic activity (GO:0003824, 12 sequences), binding (GO:0005488, 10 sequences) and transporter activity (GO:0005215, 2 sequences) were the mostly abundant among the MF category. The CC category encompassed mostly cell part (GO:0044464, 15 sequences), cell (GO:0005623, 15 sequences) and organelle (GO:0043226, 13 sequences). In T2, the BP category comprised the majority of the GO

assignments (43.11%, 241 sequences), followed by CC (37.39%, 209 sequences) and MF (19.50%, 109 sequences). Metabolic process (GO:0008152, 62 sequences), cellular process (GO:0009987, 50 sequences), single-organism process (GO:0044699, 46 sequences), and biological regulation (GO:0065007, 18 sequences) were largely represented in the BP category. In the MF category, catalytic activity (GO:0003824, 53 sequences), binding (GO:0005488, 44 sequences) and transporter activity (GO:0005215, 5 sequences) were the mostly abundant, whereas the CC category was mainly represented by cell part (GO:0044464, 51 sequences), cell (GO:0005623, 51 sequences), organelle (GO:0043226, 43 sequences), membrane (GO:0016020, 24 sequences) and membrane part (GO:0044425, 18 sequences) GO terms. Finally, in T3 the CC category encompassed the majority of the GO assignments (44.89%, 180 sequences), followed by BP (36.16%, 145 sequences) and MF (18.95%, 76 sequences). In the BP category, cellular process (GO:0009987, 31 sequences), metabolic process (GO:0008152, 31 sequences), single-organism process (GO:0044699, 25 sequences), localization (GO:0051179, 15 sequences), and response to stimulus (GO:0050896, 13 sequences) were the mostly abundant. Binding (GO:0005488, 33 sequences), catalytic activity (GO:0003824, 27 sequences), transporter activity (GO:0005215, 10 sequences) largely represented the MF category. The CC category mainly comprised cell part (GO:0044464, 39 sequences), cell (GO:00056203, 39 sequences), organelle (GO:0043226, 28 sequences), and membrane (GO:0016020, 27 sequences).

4.12 Dissection of the different molecular responses to drought and selection of candidate genes in the three *A. donax* ecotypes

To cope with drought stress, plants activate a complex gene network, which encompasses two groups of genes; the first group contains genes involved in the cell protection from damage induced by water-deficit and the second one includes genes encoding for regulatory proteins that control the stress signal and gene expression (Lata *et al*., 2011).

Under water stress, several transcription factors (TFs) families are involved in regulating the expression of drought-related genes. TFs are proteins acting jointly with other transcriptional regulators (e.g. chromatin remodeling proteins) to allow or obstruct RNA polymerase to bind transcriptional start sites (TSSs) on the DNA molecule. Approximately, the 7% of the coding sequence in plant genomes was assigned to TFs, confirming the importance and the complexity of transcriptional regulation (Udvardi *et al*., 2007). TFs interact with *cis*-elements in the promoter regions of stress-related genes and regulate the expression of several downstream genes resulting in complex regulatory transcriptional

networks that impart abiotic stress tolerance in plants (Agarwal and Jha, 2010), such as during cellular dehydration (Joshi *et al*., 2016). Among the DETs in EcoA, several TFs belonging to different protein families were retrieved, suggesting their important role in the drought stress response in this ecotype. These families included NAM (no apical meristem)-ATAF (*Arabidopsis* transcription activation factor)-CUC (cup-shaped cotyledon) (NAC; e.g. NAC48, 68, 78, 83, 94 and 25-like), WRKY (containing Trp, Arg, Lys and Tyr amino acid residues; e.g. WRKY11 and 53), myeloblastosis (MYB; e.g. MYB44-like and early flowering MYB, EFM), myelocytomatosis (MYC; e.g. MYC2), apetala2/ethylene response element binding factors (AP2/ERF; e.g. apetala2/ethylene-responsive element binding protein, AP2/EREBP; dehydration responsive element-binding factor 2, DREB2; ethylene responsive factor2, ERF2), phytochrome interacting factor (PIF; e.g. PIF4 and 1), basic leucine zipper (bZIP; ABI5), which were significantly down-regulated, and far-red-impaired response (FAR; e.g. FAR1) and FAR1-related sequence (FRS; e.g. FRS5) TFs, which showed an up-regulation. Conversely, few TF families were discovered in EcoB and EcoC; transcripts encoding for WRKY (e.g. WRKY19), FRS (e.g. FRS5) and PIF (e.g. PIF5) were up-regulated in EcoB, whereas transcripts encoding for bHLH (e.g. bHLH92), MYC (e.g. MYC2), PIF (e.g. PIF4 and 5) and MYB (e.g. SANT/MYB) were over-expressed in EcoC, except for PYF4 and SANT/MYB which were under-expressed. The up-regulation of the negative regulator of DREB2 protein was observed for EcoA; this protein was the E3 ubiquitin ligase DRIP2 (DREB2 interacting protein 2)-like which is involved in the DREB2 targeted degradation by 26S proteasome mediated proteolysis (Qin *et al*., 2008).

The plant ‘stress hormone’ abscisic acid (ABA) is crucial for plant growth and development and plays a pivotal role in integrating abiotic stress signals and downstream responses (Tuteja, 2007). It is perceived by an ABA-bound pyrabactin resistance1 (PYR1)/pyr1-like (PYL)/regulatory components of ABA receptors (RCARs) receptors complex combined with protein phosphatase 2C (PP2C), and consequently SNF1-related protein kinase 2 (SnRK2) is activated by phosphorylation. (Miyakawa *et al*., 2013; Nakashima and Yamaguchi-Shinozaki, 2013). Then, the activated SnRK proteins phosphorylate downstream targets like ABA-responsive element (ABRE)-binding proteins/ABRE-binding factors (AREB/ABF) TFs, which bind to the ABRE *cis*-element (PyACGTGG/TC) (Umezawa *et al*., 2013). In the absence of ABA, PP2C inhibits ABA signaling pathway by dephosphorylating SnRK2 proteins. In the present study, the down-regulation of several transcripts encoding for proteins involved in the ABA perception (PP2C,

PYL4-like and PYR1-like), and encoding for an intermediate in the ABA biosynthesis pathway (i.e. zeaxanthin) was detected in EcoA, whereas transcripts encoding for PP2C and zeaxanthin proteins were over-expressed in EcoB and EcoC.

In addition, stomata are specialized cellular structures in leaf epidermis tissue of plants, and consist of two guard cells that function as turgor-operated valves that regulate gas exchange (H_2O release and CO_2 uptake) between leaves and atmosphere. Stomata are non-randomly distributed, and their density and spacing pattern is finely regulated by genetic and environmental factors (Pillitteri and Dong, 2013). In the present study, for EcoA transcripts encoding for a subtilisin-like protease were strongly down-regulated. This protease, referred as stomatal density and distribution 1 (SDD1), is involved in the control of cell lineages development that leads to guard cell formation, negatively regulating stomatal development. Loss of SDD1 function results in the correct spaced but higher density stomata (Berger and Altmann, 2000; Von Groll *et al*., 2002). Conversely, for EcoB and EcoC no DETs involved in the stomatal density and development were detected.

Additionally, calcium (Ca^{2+}) is an important plant macro-nutrient element that is absorbed by roots and delivered to the shoot via the xylem where it plays a crucial role in regulating many processes, from plant growth and development, to the perception of extracellular signals and environmental cues (Tuteja and Mahajan, 2007). Besides, Ca^{2+} signaling is involved in the delay of cell cycle progression in response to abiotic stress (Sano *et al*., 2006). The regulation of gene expression by cellular Ca^{2+} is also crucial for plant defense against various abiotic stresses (Knight *et al*., 1997). In this study, the down-regulation of different transcripts encoding for different Ca^{2+} binding proteins was detected for EcoA. Among them, transcripts encoding for calcineurin B-like (CBL) interacting protein kinases (CIPK1, 6, 7, and 10), kinesin-like calmodulin binding protein (KCBP)-interacting Ca^{2+} binding protein (KIC), calmodulin-like protein (CML27 and 18), calmodulin-binding protein 60 D (CBP60D), and finally for calcineurin B1 (CNB1) were identified among DETs. For EcoB, several transcripts encoding for calcium-transporting ATPase plasma membrane-type (ACA), for calreticulin (CRT)-like proteins (CRT-3-like) were down-regulated, whereas transcripts encoding for CIPK29 and annexin P35 (ANN P35) were up-regulated. For EcoC, few transcripts were related to Ca^{2+} signaling; particularly, transcripts encoding for calmodulin-binding transcription activator 4 (CAMTA4) and CIPK1 were over-expressed, while transcripts for mitochondrial calcium uniporter (MCU) were under-expressed.

Moreover, aquaporins (AQPs) are membrane channel proteins, initially discovered as water channel, but their role as transporter of small neutral solutes, gasses and metal ions is now well accepted (Maurel *et al*., 2008). Water flow through plants can occur in three different ways: the apoplastic path along cell wall structure, the symplastic path from cell to cell through the plasmodesmata, and the transcellular path across the cellular membranes (Steudle and Peterson, 1998). Particularly, AQPs are involved in the symplastic path, facilitating the water flow across cells and subcellular compartments; as integral regulator of plant-water relation, they are assumed to play an important role in plant defense against abiotic stress (Afzal *et al*., 2016). Plant AQPs encompass seven subfamilies, divided based on their intercellular locations and sequence similarities. These include the plasma membrane intrinsic proteins (PIPs), tonoplast intrinsic proteins (TIPs), NOD26-like intrinsic proteins (NIPs) and small, basic intrinsic proteins (SIPs), the GlpF-like intrinsic proteins (GIPs), hybrid intrinsic protein (HIP), and the uncategorized X intrinsic protein (XIP) (Forrest and Bhave, 2007; Afzal *et al*., 2016). In this study, DETs belonged to four AQPs subfamilies were detected for EcoA; particularly, the down-regulation of transcripts encoding for TIPs (TIP1;1 and 4;2), NIPs (NIP2;2 and 1;1-like), SIP2;1 and PIP2;7 was observed for EcoA and PIP1;1 for EcoB, whereas no DETs belonged to AQPs subfamilies were identified for EcoC.

Under abiotic stress, another factor heat shock proteins (HSPs) play a fundamental role in protecting plants against different abiotic stress, acting in thermotolerance reactions, and also as molecular chaperones to avoid denaturation or aggregation of proteins, as well as facilitating protein refolding (Ahuja *et al*., 2010). In the present thesis, transcripts encoding for HSP70-17, heat shock proteins class III and V were down-regulated for EcoA and transcripts for HSP70-2 were under-expressed in EcoB. No transcripts encoding for HAPs were detected in EcoC.

Under drought condition, and in general abiotic stress, oxidative damage occurs and causes serious secondary negative effects on cells. Oxidative stress is associated with the production of reactive oxygen species (ROS) such as superoxide radical ($O_2^{\cdot -}$), singlet oxygen ($^1O_2$), hydrogen peroxide (H_2O_2), hydroxyl free radical ($^{\cdot}OH$), and causes lipid peroxidation (LPD), protein denaturation, DNA mutations and many others cellular oxidative damages, and ultimately cell death. Despite their negative role, they also act as second messengers in different cellular processes, including resistance to abiotic stresses, and the positive and the negative activities of ROS depend on the balance among ROS production and scavenging (Mattos and Moretti, 2015) Therefore, to survive under drought stress many plants produce

and accumulate compatible solute compounds, refereed as osmoprotectants or osmolytes. These molecules are involved in the maintaining of osmotic balance and stabilization of proteins and membranes under abiotic stress conditions (Yancey, 2005). In the present study, transcripts encoding for β-amylase (BAM1), proline (Pro) oxidase (POX) and Rieske-type iron-sulfur choline monooxygenase (CMO) enzymes were under-expressed in EcoA; conversely the over-expression of transcripts encoding for alpha,alpha-trehalose-phosphate synthase 7 (TPS7) was observed in EcoC. No transcripts encoding for osmolytes were detected in EcoB. In addition, several transcripts encoding for glutathione S-transferase (GST), manganese superoxide dismutase (MnSOD), thioredoxin-like 6 (TXNL6), glutaredoxin 2 (GLRX2), thioredoxin reductase (NTRA), and different lipoxygenases (LOX1 and LOX5) proteins were down-regulated in EcoA, whereas for EcoB transcripts for glutathione S-transferase TAU 1 (GSTU1) and thioredoxin (TXN) were down- and up-regulated, respectively. For EcoC, transcripts encoding for LOX and 2 -deoxymugineic-acid 2 -dioxygenase-like (IDS3-like) proteins were over- and under-expressed, respectively. In addition, poly(ADP-Ribose) polymerase (PARP) proteins play also a crucial role under oxidative stress, and the up-regulation of transcripts encoding for tankyrase-1-like (TNKS1-like) proteins were observed in EcoC.

Moreover, differential expression analysis revealed that transcripts related to secondary metabolism exhibited significant expression changes under drought stress. Particularly, phenolic compounds, such as flavonoids, anthocyanins, lignins, play different roles in response to abiotic stresses, particularly in the adaptation and the overcoming of the stress condition (Akula and Ravishankar, 2011). In this study, several transcripts encoding for phenolics showed a differential expression; particularly, transcripts encoding for proteins involved in the lignin biosynthesis, such as trans-cinnamate 4-monooxygenase (referred also as cinnamate 4-hydroxylase, C4H), caffeoylshikimate esterase-like (CSE-like), phenylalanine ammonia-lyase (PAL), 4-coumarate ligase (4CL), cinnamoyl-CoA reductase 2-like (CCR2-like) were down-regulated in EcoA. For EcoB, transcripts encoding for cinnamyl-alcohol dehydrogenase (CAD) and CSE-like proteins were under-expressed, whereas transcripts for CCR1-like proteins were strongly over-expressed in EcoC. In addition, transcripts encoding for anthocyanidin 3-O-glucosyltransferase 2-like (3'-GT 2-like), flavonoid 3,5-hydroxylase 1-like (F3'5'H 1-like), naringenin,2-oxoglutarate 3-dioxygenase (F3H) and terpene synthase 7 (TPS 7) proteins were down-regulated in EcoA, whereas transcripts encoding for flavonol synthase flavanone 3-hydroxylase-like (FLS1-like) and leucoanthocyanidin dioxygenase

(LDOX) proteins in EcoC and monoterpene synthase (MTS) proteins in EcoB were strongly up-regulated.

Finally, leaf senescence is a kind of programmed cell death (PCD), and so it is the ultimate stage of leaf development, which is highly coordinated at molecular, cellular, biochemical and physiological levels. Plant aging and environmental stresses (e.g. drought stress) can induce the onset of senescence, as a means to reduce canopy size and to avoid large water losses through transpiration, thereby contributing to maintain the suitable water balance in the whole plant under adverse environmental condition (Munné-Bosch and Alegre, 2004). In this study, both up-regulated and down-regulated DETs encoding for proteins associated with senescence process (senescence-associated genes, SAGs and ethylene-insensitive 3-like 3, EIL3) were identified in EcoA, whereas down-regulated transcripts encoding for EIL3 were detected in EcoB. For EcoC, no DETs encoding for senescence associated proteins were identified.

In conclusion, the analysis of leaf transcriptome profile in the three *A. donax* ecotypes grown under moderate drought stress indicated that transcriptional regulation via several transcription factors families, calcium and abscisic acid perception and signaling, and the regulation of stomata development and spacing play a central role in the drought stress response in *A. donax*. Furthermore, the transcriptional regulation of HSPs, aquaporins, and osmolyte genes and also the regulation of oxidative damage and senescence represent other important elements in the drought response in giant reed.

5. DISCUSSION

Although *A. donax* is one of the most promising and emerging non-food bioenergy crops in regions with Mediterranean climates, until very recently available genomics resources were scarce. The knowledge about the molecular mechanisms of *A. donax*, underlying its adaptability in marginal lands with low input requirements, its high biomass yield and tolerance to abiotic stresses (e.g. drought stress), is still in its infancy due to the little information in public databases about *Arundo* transcriptome profiles. Currently, only a Blast web-server is available for the scientific community to search for homologs of known genes in a transcriptome of bud, leaf, culm and root tissues (Sablok *et al*., 2014), and in a transcriptome of shoot and root tissues of young *A. donax* plants subjected to osmotic/water stress and grown in a growth chamber (Fu *et al*., 2016). In addition, the *A. donax* shoot transcriptome of an invasive ecotype collected along the Rio Grande River, bordering Texas

and Mexico, was recently published (Barrero *et al*., 2015), and the assembled and annotated transcripts are freely accessible at DDBJ/EMBL/GenBank database. Hence, actually only the transcriptome of this tissue is accessible for the scientific community, and no transcriptome data for *A. donax* leaf tissue are freely available in public database up to now. Together with this recent sequencing of the transcriptome, the transcriptomic data generated in this study provide invaluable resources to understand the biology of *A. donax*, offering new tools and methodologies to help develop agriculturally productive and economically important genotypes, and also DE analysis allowed the dissection of molecular mechanisms involved in the drought stress response in giant reed, thereby identifying candidate genes for drought tolerance in *A. donax*.

It is worth mentioning that genomes at species level are dynamic and are composed by genes which are present in each individual of a species (core genes) and genes in a set of individuals (dispensable genes); collectively, these genes constitute the so-called pan-genome. For example, transposable elements are largely responsible for the variation in both intergenic and genic regions, not only in closely related species but also between individuals within a species. These suggest that a single genome sequence might not reflect the entire genomic information of a species (Morgante *et al*., 2007). The presence of pan-genome and pan-transcriptome was observed also in many plant species. For instance, a study of the pan-genome was conducted by resequencing of six elite *Z. mais* inbred lines for commercial relevance in China and several hundreds of complete dispensable genes with presence/absence variation among these lines were identified (Lai *et al*., 2010). More recently, a transcriptome sequencing of 503 maize inbred lines revealed the dynamic nature of the maize pan-genome and identified representative transcript assemblies (RTAs) with 16.4% expressed in all lines and 82.7% expressed in subsets of the lines, demonstrating a considerable portion of variation may be outside the single reference genome for a species (Hirsch *et al*., 2014). Also DNA and RNA sequencing of seven accessions of *Glycine soya* (wild soybean) revealed that approximately 80% of the pan-genome was present in all the accessions, representing the core genome, and the remaining were dispensable genes that exhibited greater variation reflecting a putative role in adaptation to different environments (Li *et al*., 2014). This reflects the importance of sequencing and assembling numerous *A. donax* transcriptomes in order to capture all the possible variable dispensable genes/transcripts and decipher their role in adaptation to different environments. Further analysis of the pan-transcriptome rather than a single reference circumvents single sample

bias and ensures that the majority of genetic diversity within *A. donax* is fully captured. Furthermore, *de novo* construction of *A. donax* pan-transcriptome is also necessary to unravel variation in gene regulation and expression and provide additional candidate genes for the development of elite genotypes.

De novo RNA-Seq assembly facilitates the study of transcriptomes for plant species without sequenced genomes, but it is challenging to select the most accurate assembly in this context. Here, a specialized bioinformatics workflow for *de novo* transcriptome assembly and DE analysis was developed (Fig. 1). Illumina next-generation mRNA-Seq was successfully used to assemble and analyze the leaf transcriptome profile of three *A. donax* ecotypes representing a vast array of geographical adaptation across the species and possessing traits of agronomic importance grown under two different field conditions in order to increase the understanding of biological processes and shed a new light on the molecular mechanisms underlying the outstanding adaptability of this relatively unknown species. DE analysis was used to reveal the molecular mechanisms involved in drought stress response and to identify candidate genes for drought tolerance in *A. donax*.

More than 1,252 million of single-end transcript reads were generated by Illumina sequencing. Using a two-step procedure along with single and multiple *k-mer* methods, 62,596 unitranscripts, based on 1,249 million of filtered reads, were *de novo* assembled. The number of retrieved leaf transcripts was lower than the number of transcripts obtained in previous transcriptome reconstructions in *A. donax* (Sablok *et al*., 2014; Fu *et al*., 2016; Barrero *et al*., 2015), but it was similar to those recently obtained in *Lolium* species (*L. multiforum*: 55,570 sequences; *L. multiforum var. westerwoldicum*: 52,166 sequences; *L. temulentum*: 59,309 sequences) (Czaban *et al*., 2015). This was most probably due to the clustering of the different pre-assemblies by CD-HIT and the redundancy removal using the tr2aacds pipeline to obtain the final assembly.

The N50 length of the unitranscripts was 1,134 bp, and the average length was 842 bp (Table 1); these results are comparable to recently published plant transcriptomes, such as common reed (*Phragmites australis*; N50 = 1,187 bp, average length = 642 bp) (He *et al*., 2012), common wild rice (*O. rufipogon* Griff.; N50 = 1,147 bp, average length = 715 bp) (Tian *et al*., 2015) and sugarcane (*Saccharum officinarum*; N50 = 1,177 bp) (Li *et al*., 2016). Additionally, the GC content of the assembled unitranscripts (49%) was slightly higher compared to those obtained previously in *A. donax* [45.60% (Sablok *et al*., 2014) and 47.70%

(Fu *et al*., 2016)], but it was similar to that observed in *P. australis* using Illumina sequencing (49.9%) (He *et al*., 2012).

To evaluate assembly consistency and completeness, the percentage of reads mapping back to the assembled leaf transcriptome were assessed (Table 2). The observed percentage of mapped reads was about 65% and this result is similar with that retrieved in *T. aestivum* (Li *et al*., 2013).

We further assessed how much overlap there was between *A. donax* unitranscripts and those from related species (*H. vulgare*, *T. aestivum*, *S. bicolor*, *O. sativa* and *S. italica*). Similarly to *Lolium* species (Czaban *et al*., 2015), the highest number of hits was obtained for *H. vulgare* (6,526 transcripts with 100% coverage, and 11,443 transcripts with > 70% coverage) and *T. aestivum* (6,503 transcripts with 100% coverage, and 12,385 transcripts with > 70% coverage) (Table 3). We also checked the completeness of our transcriptome assembly using BUSCO quality assessment tool. The quality of the *A. donax* leaf transcriptome was comparable to or better than those of the majority of transcriptome assemblies listed in (Simão *et al*., 2015) and the *Spinacia oleracea* transcriptome assembly (Xu *et al*., 2015).

Although the study is based only on one vegetative tissue (leaf), it has broadened the knowledge of the transcriptome profile in *A. donax* grown under natural conditions. About 81% and 55% of the unitranscripts and predicted proteins were successfully assigned to genes by BLASTx and BLASTp searches against the UniRef90 database (Table 4). Moreover, a lower number of BLASTx (54%) and BLASTp (34%) matches were obtained with the UniProtKB/Swiss-Prot database, but this result was similar to that observed in *S. officinarum* (Li *et al*., 2016). Pfam domains and signal peptides were assigned to 32% and 2.76% of the identified proteins, respectively; these numbers are lower compared to those obtained in previous studies in perennial ryegrass (*Lolium perenne*) (Farrell *et al*., 2014), *Lolium* and *Festuca* species (i.e. *F. pratensis*; *L. multiflorum*; *L. m. westerwoldicum*; *L. temulentum*) (Czaban *et al*., 2015) and sugarcane (*S. officinarum*) (Li *et al*., 2016).

The Blast2GO suite using the NCBI NR protein database was also adopted to annotate the *A. donax* leaf transcriptome. A lower number of BLAST hits (26,323) against this database was obtained compared to the UniRef90 and UniProtKB/Swiss-Prot database. As expected, a considerable sequence similarity was observed with the related species *S. italica*, *S. bicolor*, *Z. mays*, and *O. sativa japonica* (Fig. 3). 820 GO terms were retrieved and of the three main GO categories, 63% of GO terms were assigned to BPs, about 18% were assigned to MFs,

and 13% to CCs. Moreover, three subcategories (metabolic processes, biosynthetic processes and cellular processes) were enriched in the BP category, two subcategories (binding proteins and proteins with catalytic activities) in the MF category and four subcategories (cell, cell part, organelle and membrane) in the CC category (Fig. 4). A high proportion of unitranscripts showed no similarity to any known putative protein in public databases. These transcripts could probably be present due to a combination of factors, such as the presence of long non-coding RNAs, untranslated regions, genes without any annotation information or lack of known protein domains, and finally novel genes in *A. donax* that require more experimentation to be characterized in details. Likewise, transcripts without BLAST matches were observed previously in plant transcriptome reconstruction in celery (*Apium graveolens*) (Fu *et al*., 2013), in perennial ryegrass (*L. perenne*) [61] and in sugarcane (*S. officinarum*) (Cardoso-Silva *et al*., 2014).

Biological pathway studies play a pivotal role in gaining insight into functional analysis and interpretation of transcriptomic data. KEGG is a highly integrated database that allows the comprehensive interrogation of an organism's genome content (Kanehisa *et al*., 2008). Here, purine (Fig. 5a) and thiamine (Fig. 5b) metabolism, phenylpropanoid biosynthesis (Fig. 6a, Table 6), starch and sucrose metabolism (Fig. 6b), cellulose biosynthesis (Table 5), carbon fixation (Fig. 7), stomatal development and distribution pathways (Table 7), and finally, the presence of homologous transcripts encoding for SAPs (Table 8) were analyzed in the *de novo* leaf transcriptome.

Purine metabolism, particularly adenine metabolism, has been reported for *Arabidopsis oxt1* mutants to be involved in activating plant abiotic defenses. The cellular adenine level plays a crucial role in signals that control abiotic stress responses and plant growth, indicating a connection between purine metabolism, plant growth responses and stress acclimation (Sukrong *et al*., 2012). Moreover, allantoin, a metabolic intermediate of purine catabolism, often accumulates in plants subjected to abiotic stress. Particularly, the role of allantoin in the activation of jasmonic acid responses via ABA has been recently reported for *Arabidopsis* knockout mutants (*aln*), suggesting a possible link of purine catabolism with stress phytohormone homeostasis and signaling (Takagi *et al*., 2016). Furthermore, thiamine (vitamin B_1) and its phosphate esters act as cofactors in response to abiotic and biotic stress in plants (Goyer, 2010). Interestingly, its metabolism can be altered in plants under environmental stress, *e.g.* in *Z. mays* under abiotic stress conditions (Rapala-Kozik *et al*., 2008).

Additionally, several putative targets for lignocellulosic biomass production and improvement were evaluated in the *de novo* transcripts catalogue. A prevalent proportion of plants biomass is composed of cell wall polymers; in secondary walls cellulose is the load-bearing unit that cross-links with hemicelluloses (xylan and glucomannan) and together with lignin, this complex network forms the secondary cell wall, an important source for biofuel production. Lignins are complex racemic aromatic heteropolymers that are products of the phenylpropanoid metabolism, whereas cellulose is made of β-1,4-linked glucosyl residues and its biosynthesis is mediated by cellulose synthase complexes, called rosettes, located in the plasma membrane. The proportion of these components differs between different plant species and environmental stimuli (Zhong and Ye, 2015). In the *A. donax* leaf transcriptome, many homologous transcripts encoding for lignins and cellulose were retrieved.

The analysis of the carbon fixation pathway revealed, as expected, that the majority of the metabolic genes of this pathway are expressed in *A. donax* leaves. It has long been known that the C fixed during photosynthesis is transformed into sugar or sugar derivatives in photosynthetic source cells. These compounds are then transported by the phloem to sink tissues where it is consumed by cells to obtain energy and building blocks for growth or stored in vacuoles, as polymers (starch or plastids) and as structural molecules (cellulose, hemicellulose, and lignin) (Poorter and Villar, 1997). In sugarcane, sucrose synthesized in leaves is deposited both inside and outside of stalk (culm) parenchyma cells and in secondary cell wall components, particularly as cellulose (Wang *et al*., 2013). Consequently, improving the production of photosynthates, which are subsequently accumulated in polymers and structural molecules, could be a strategy for the success of *A. donax* as a bioenergy crop.

Stomatal development, distribution and aperture mechanisms are finely regulated in plants in order to control transpiration and water use efficiency (WUE) under different environmental conditions. In the *A. donax* leaf transcriptome, five homologous transcripts were retrieved. Particularly, the detection of the *GTL1* protein as a transcriptional repressor of *SDD1*, negatively regulating stomatal density in order to regulate WUE (Yoo *et al*., 2011) is an important finding. The presence of *GTL1* in the assembled unitranscripts might be associated to an adaptive response of *A. donax* that prevents water loss in dry conditions.

Finally, the presence of homologs encoding for SAPs was investigated in the *A. donax* leaf unitranscripts. These A20/AN1 zinc-finger proteins were reported as an important component of stress responses in rice, where 18 SAP encoding genes were inducible by multiple abiotic stresses such as cold, salt and dehydration stresses (Vij and Tyagi, 2006).

The presence of homologous transcripts to genes involved in the above mentioned pathways gives new insights into molecular mechanisms underlying the extreme adaptability and robustness of *A. donax* to a wide range of environmental stresses, thereby contributing to its importance as a bioenergy crop.

Moreover, for the first time SSRs markers were identified in the *A. donax* leaf transcriptome. Using the MISA software, a total of 8,364 genic-SSRs were obtained from 7,491 sequences. In this SSRs catalog, the most frequent repeat type was the mono-nucleotide motif with 4,170 SSRs (49.86%), followed by tri-nucleotide repeats with 3,034 SSRs (36.27%) and di-nucleotide repeats with 1,060 SSRs (12.67%). This distribution frequency differs from those previously reported in sugarcane (Cardoso-Silva *et al*., 2014) in which the most frequent SSR motif is the tri-nucleotide repeat. Genome-wide SSRs analyses in nine species of the Poaceae family (Wang *et al*., 2015), revealed that the most abundant repeat type was a hexamer (58.82%). Conversely, our results were consistent to those obtained in locorice (*Glycyrrhiza uralensis*) in which the most frequent repeat motif was a mono-nucleotide, followed by tri-nucleotides and di-nucleotide motifs (Liu *et al*., 2015) and partially to those obtained in bhumyaamalaki (*Phyllanthus amarus*), in which the most abundant repeat type was a mono-nucleotide, followed by di- and tri-nucleotides (Bose Mazumdar and Chattopadhyay, 2015). Although natural genetic variation in *A. donax* was reported to be low for the Mediterranean basin and slightly higher for the Asian area (Mariani *et al*., 2010), as well as genetic uniformity was also observed in ecotypes in the United States (Ahmad *et al*., 2008), here 53 Candidate PolySSRs were predicted using CandiSSR software, suggesting a likely high level of genetic diversity among the analyzed ecotypes. To test whether the identified PolySSRs were true-to-type ones for use in genetic diversity studies, we evaluated a total of five primer pairs, and one pair resulted to amplify a polymorphic fragment among the three ecotypes (see Results section), suggesting that the PolySSRs that were mined using the CandiSSR pipeline represent a valuable resource for assessing the genetic variability in *A. donax*. The use of identified SSR markers will help in the exploitation of variation at genomic and expression level, in the analysis of genetic diversity in future *A. donax* genomic studies, and in the identification of ecotypes for commercial utilization. Particularly, the set of PolySSRs identified in this study will facilitate the assessment of the polymorphic status in other *A. donax* ecotypes, by easing the costly and time-consuming traditional laboratory assessment of SSR development.

In addition, the majority of genic-SSRs was found to be present in functional genes, indicating these markers could possibly be associated with phenotypes of agricultural value. SSRs located in coding and untranslated regions (transcribed SSRs) can be efficient functional markers in genic regions (Li *et al*., 2004). The functional categorization of *A. donax* unitranscripts containing SSRs revealed that these transcripts represent several key transcribed genes with biology, cellular and molecular function. Therefore, while allowing studies on genetic variation, SSR markers derived from transcripts also provides information on gene function related to possible phenotypic differences between the *A. donax* ecotypes. Newly designed SSR primers for *A. donax* can be used to detect SSRs in other species in Poaceae, and explore their transferability to other closely related species that also lack genomic resources.

Differential gene expression analysis allowed the identification of potential candidate genes useful for accelerating molecular breeding of drought tolerant *A. donax* ecotypes. Particularly, TFs enable plants to withstand under adverse environmental conditions (e.g. drought) and are considered promising genomic candidate for their application in crop improvement. In this study, a down-regulation of different TFs families (NAC, WRKY, MYC, MYB, AP2/ERF, bZIP) was observed in EcoA, except those encoding for *FAR1* and *FRS5* which were over-expressed. Conversely, for EcoB and EcoC an up-regulation of several TFs families (WRKY, FRS, PIF, bHLH, MYC, MYB) was detected, except for *PYF4* and *SANT/MYB* transcripts which were under-expressed. It has been proved that *FAR1* is involved in the crosstalk between light and ABA signaling in *Arabidopsis*, and it is considered as a positive regulator of ABA signaling, conferring adaptation to environmental stresses; particularly, in *Arabidopsis far1* knockout mutants showed wider stomata, faster water loss, and they are more sensitive to drought than wild type plants (Tang *et al*., 2013). Therefore, the up-regulation of *FAR1* and *FRS* in EcoA and EcoB could be associated with a greater control of stomata size, of water loss, and thus a greater ability to tolerate drought stress, compared to EcoC. Further, the down-regulation of transcripts encoding for proteins involved in the ABA perception (PP2C and ABA receptors) and for an intermediate in the ABA biosynthesis pathway was observed for EcoA, whereas for EcoB and EcoC these transcripts were up-regulated. In addition, EcoC up-regulated also the bHLH/MYC TFs, which are also transcriptional activators ABA signaling (Singh and Laxmi, 2015). Taken together these results suggest that all the three ecotype use the ABA-dependent pathways, but in a specific way and with different intermediates.

Furthermore, under water stress the regulation of stomatal density and distribution represents another important aspect for improving drought tolerance in plants. In this study, the down-regulation of transcripts encoding for SDD1, a subtilisin-like protease involved in the negative regulation of stomatal development, was retrieved in EcoA. Experimental evidence demonstrated that stomatal density is negatively correlated with stomatal size (Hetherington and Woodward, 2003); in addition, in a perennial grass (*Leymus chinensis*) it has been reported that moderate drought stress influenced positively the number of stomata and negatively the stomatal size (Xu and Zhou, 2008). This result, together with the up-regulation of *FAR1* transcript, could be associated with an adaptive response of this ecotype which controls stomatal density and size to fine-regulate water flow and gas exchange under drought stress.

Under drought stress, Ca^{2+} signaling plays an important role and it has been reported that higher level of cytoplasmic free Ca^{2+} concentration functions as second messenger to mediate defense response in plants (White and Broadley, 2003). In this study, several transcripts encoding several Ca^{2+} binding and associated proteins were differentially regulated under drought stress. For EcoA, several transcripts encoding for different CIPK, KIC, CML, CBP, and CNB proteins were under-expressed, whereas in EcoB several transcripts encoding for ACA and CRT-3-like proteins were down-regulated and others transcripts were up-regulated (*CIPK29* and *ANN P35*). EcoC showed only few transcripts encoding for Ca^{2+} binding and associated proteins; transcripts for CAMTA4 and CIPK1 proteins were up-regulated, whereas transcripts for MCU protein was down-regulated. Thus, these outcomes indicate that potentially EcoA doesn't use Ca^{2+}-sensors (Ca^{2+} binding proteins) or Ca^{2+} dependent protein kinases to regulate its phenotypic response to drought, whereas EcoB and EcoC could employ some of these proteins to control the expression of many genes and also stress responsive genes, thus resulting in the phenotypic response of water stress tolerance.

Transcriptome studies revealed differential expression patterns of multiple *AQP* genes in response to drought, suggesting their central role in water stress responses in plants (Afzal *et al*., 2016). In this study, the under-expression of different transcripts encoding for NIPs (NIP2;2 and 1;1-like), TIPs (TIP1;1 and 4;2), SIPs (SIP2;1) and PIPs (PIP2;7) was observed in EcoA; likewise, the down-regulation of transcripts for PIP1;1 aquaporin protein was detected in EcoB. Under drought stress, contrasting and heterogeneous roles for aquaporin genes have been reported (Afzal *et al*., 2016). However, AQPs have been generally associated with drought sensitivity and the down-regulation of these genes has also been reported in

grape and tobacco plants. In grapevine, leaf expression of *AQP* genes (*PIP1;1*, *PIP1;2*, *PIP1;3*,*PIP2;1*, *PIP2;2*, *TIP1* and *TIP2*) showed an initial significant decrease in plants subjected to moderate drought stress, followed by an increase to control expression values after a period of constant soil water deficit (Galmés *et al*. 2007). In tobacco plants, the root transcript levels of two AQP genes (*NtPIP1;1* and *NtPIP2;1*) were significantly decreased after drought stress, but at the same time the transcription level of *NtAQP1*gene was significantly increased (Mahdieh *et al*. 2008). Furthermore, root transcriptome analysis of water stressed chickpea (*Cicer arietinum*) plants showed both up- and down-regulation of different AQP genes (e.g. *PIPs*, *TIPs*, and *NIPs*) (Molina *et al*., 2008). Thus, AQPs show a variable drought response based on stress level, AQP isoform, species, and also tissue. Consequently, their complex and integrated roles in the regulation of water balance under drought stress make them potential candidates for developing drought stress resistant crop plants.

In addition, HSPs play an important role under drought stress. Generally, HSPs are induced under different abiotic stresses, acting as chaperons that can assist in protein refolding, in preventing aggregation, and also in protein import and translocation under stress conditions (Wang *et al*., 2004). In this study, a down-regulation of different transcripts encoding for HSPs were observed in EcoA and EcoB. Furthermore, also osmoprotection and ROS scavenging processes are fundamental to cope with drought stress in plants. In this study, the down-regulation of different transcripts involved in the osmoprotection process (BAM1, POX and CMO) were observed for EcoA, whereas the up-regulation of transcripts involved in the trehalose synthesis (TPS7) was detected in EcoC. In addition, also different transcripts encoding for ROS scavenging enzymes were down-regulated in EcoA, whereas in EcoB and EcoC both down- and up-regulation of these enzymes were observed. These findings seem to be in contrast with those generally observed under abiotic stresses (Al-Whaibi, 2011, Dedemo *et al*., 2013). However, it is noteworthy that under natural field condition plants often encounter a combination of stresses, and consequently gene expression regulation could be different from that observed in controlled growth conditions. For example, it has been previously demonstrated that plants response to the combination of drought and heat stresses, that is similar to many natural environments, is different when these stresses are studied separately in laboratory (Rizhsky *et al*., 2002; Rizhsky *et al*., 2004). For example, in plants subjected to drought stress an high content of proline was reported, but when the same plant species were subjected to combined drought and heat stresses, an high

content of sucrose and other sugars, instead of proline, was detected (Rizhsky *et al*., 2004). Further, also PARP proteins are involved in oxidative stress response, but also in other processes, such as including DNA damage repair, cell death pathways, transcription and chromatin modification and remodeling (Lamb *et al*., 2012). In this study, the up-regulation of the PARP protein TNKS1-like was observed in EcoC. In vertebrate, tankyrase 1 is a poly (ADP-ribose) polymerase (PARP) that modifies a shelterin subunit, called TRF1, decreasing its ability to bind telomeric DNA. Since TRF1 is a negative regulator of telomere length in vertebrates, over-expression of tankyrase leads to increase in telomere length (Smith and de Lange, 2000). However, analysis of AtPARP1/AtPARP2 orthologs in *Arabidopsis* revealed that AtPARPs play a minor role in telomere biology, even if their involvement in telomere length cannot be still excluded (Boltz *et al*., 2014).

Under drought stress the concentration of phenolic compounds (e.g. flavonoids, anthocyanins, lignins) were drastically enhanced (Selmar and Kleinwächter, 2013). In this study, EcoA and EcoB showed a down-regulation of different transcripts involved in the lignin biosynthesis. In addition, for EcoA transcripts encoding for proteins involved in the metabolism of flavonoids and anthocyanins were also down-regulated, whereas transcripts involved in the monoterpene biosynthesis were up-regulated in EcoB. Conversely, transcripts involved in lignin, flavonoids and anthocyanins metabolisms were significantly up-regulated in EcoC, as previously reported for in *A. donax* under osmotic/water stress in growth chamber (Fu *et al*., 2016). The down-regulation of genes involved in lignin biosynthetic pathway, observed in EcoA and EcoB, has been previously detected in maize, in which the lignin content in leaves of plants subjected to drought stress was lower compared to control plants, suggesting an adaptation to drought, since the maintenance of high levels of lignin in the region of leaf elongation region might negatively affect growth after rehydration (Vincent *et al*., 2005). These results are also relevant in the context of *A. donax* employment as bioenergy crop for biofuels production. It has been reported that reduced lignin content in plant biomass could enhance saccharification efficiency through enzyme hydrolysis, and consequently could reduce the cost of biofuel production (Fu *et al*., 2011).

In addition, leaf senescence is a kind of PCD process that can be activated under drought stress. Drought-induced leaf senescence, especially when accompanied with leaf abscission, avoids large water losses through transpiration, in order to maintain water balance in plant, and it is accompanied by macroscopic, cellular, biochemical and molecular changes (Munné-Bosch and Alegre, 2004). It is also characterized by an increased oxidative stress. Besides, in

sage (*Salvia officinalis*) a loss of antioxidant defences activity resulting in oxidative stress, mediated drought-induced leaf senescence (Munné-Bosch *et al*., 2001). In this study, senescence-associated genes were both down- and up-regulated in EcoA, whereas EcoB showed down-regulation of transcripts encoding for EIL3. As previously said, leaf senescence could be considered a process that aims to maintain a favorable water balance in plants under drought condition (Munné-Bosch and Alegre, 2004). However, it has also been reported that the suppression of drought-induced leaf senescence resulted in an outstanding drought tolerance in transgenic tobacco plants expressing P_{SARK}::*IPT* (i.e. plants in which the expression of *isopentenyltransferase*, *IPT*, gene were under the control of the *senescence associated receptor protein kinase*, *SARK*, promoter) (Rivero *et al*., 2007). *IPT* gene encoding for an enzyme that catalyzes the rate-limiting step in cytokinin (CK) synthesis, whereas *SARK* is gene encoding for a senescence-dependent receptor protein kinase, which has been shown to be induced during late maturation and decreased during the development of senescence, suggesting its role in signaling (Hajouj *et al*., 2000). Particularly, it has been reported that *EIL* genes (e.g. *EIL1*), together with *EIN3* gene, are involved in the developmental cues and environmental signals to induce the progression of leaf senescence (Li *et al*., 2013). Therefore, the down-regulation of EIL3 in EcoA and EcoB could be associated with a delay in the progression of drought-induced leaf senescence and an increase in water stress tolerance, albeit some other transcripts for SAGs in EcoA were up-regulated.

Finally, transcripts encoding for PFP and for JIPs, were strongly up-regulated in all ecotypes. Experimetal evidences demonstrated that PFP, which is a regulatory enzyme involved in glycolysis and gluconeogenesis, plays a role in the adaptation of *Arabidopsis* seedlings to salt and osmotic/water stresses (Lim *et al*., 2014). In addition, transgenic carrot plants with a mutated mammalian *phosphofructo-2-kinase/fructose 2,6-bisphosphatase gene* (*Fru 2,6-P_2 ase*) gene showed a marked increase in the fuctose 2,6-bisphosphate (Fru-2,6-P_2) level and in PFP activity; particularly, Fru-2,6-P_2 stimulated PFP which operates in the glyconeogenic direction in the taproots of drought stressed plants, whereas the glycolytic direction dominates in the non-stressed controls. Thereby, the metabolic status determining the PFP activity depends on the physiological stress situations and consequently, PFP could be considered an important sensor of the environmental changes (Kovács *et al*., 2006). Moreover, there is also a steadily increasing of evidences for the involvement of jasmonates in drought stress (Riemann *et al*., 2015). Particularly, under drought induction with sorbitol or mannitol an increased in endogenous contents of jasmonates, followed by the transcription of

JIPs were reported in barley leaves (Lehmann *et al*., 1995). Moreover, endogenous contents of jasmonates in maize root cells increased under drought stress (Xin *et al*., 1997), and also this molecule can elicit the osmolytes accumulation (e.g. betaine) in pear leaves (Gao *et al*., 2004).

6. CONCLUSION

In the current study we employed next-generation mRNA sequencing (RNA-Seq) technology to generate, investigate and characterize the *A. donax* leaf transcriptomic profile under moderate drought stress in the field. Currently, the data reported here provide the first publicly available comprehensive transcriptome resource for *A. donax*. To facilitate molecular research in *A. donax,* 62,596 leaf unitranscripts were functionally annotated and characterized. The occurrence of genes involved in crucial metabolic pathways were further assessed, thereby giving new insight into the molecular mechanisms underlying the extreme adaptability of *A. donax* to different environmental conditions and its importance as a bioenergy crop. Further, this study provides the first SSR markers catalog and a set of candidate PolySSRs for *A. dona*x, which will be a useful resource for future genetic and genomic studies in this still poorly investigated species. The genomic data herein generated for *A. donax* ecotypes represent a counterpart to the genomic resources already available for this species. In addition, candidate genes for drought tolerance in *A. donax* were identified and provided to the scientific community in order to further the development of drought tolerant ecotypes and, consequently, the use of this species in drought-prone marginal lands. Altogether, this study will facilitate future genetic and molecular studies in *A. donax* and related species, and also will help in accelerating and implementing breeding strategies in this important fast-growing bioenergy crop.

7. ACKNOWLEDGEMENTS

Undertaking this Ph.D has been a truly life-changing experience for me and it would not have been possible to achieve without the support and guidance that I received from many people.

A special thank you is addressed to Prof. Antoine Harfouche for giving me the opportunity to undertake my PhD thesis under his careful supervision. Working with him, I did not only

enriched my knowledge in science but also in bioentrepreneurship. Without his guidance and constant feedback this PhD would not have been achievable.

Also a resounding thank you to the members of my internal evaluation committee, Prof. Giuseppe Scarascia Mugnozza, Prof. Alessio Valentini and Prof. Maurizio Sabatti, for their support and critical evaluation during the three years of the PhD.

I would also like to thank my colleagues, Francesco Fabbrini, Riccardo Ludovisi, Andrea Firrincieli and Paolo Latini for their support for field and laboratory activities. Special appreciation to Ms. Gabriela Porcari for providing me with laboratory assistance.

A particularly special thank you is addressed to my family for the help and the support given to me during these three years.

I gratefully acknowledge the funding received towards my PhD partly from the University of Tuscia PhD fellowship. I am also grateful to the funding received through the European Commission's Seventh Framework Programme (FP7/2012-2017) under the grant agreement n° FP7-311929 (WATBIO).

8. REFERENCES

Adani F, Papa G, Schievano A, Cardinale G, D'Imporzano G, Tambone F (2011) Nanoscale structure of the cell wall protecting cellulose from enzyme attack. *Environ Sci Technol*, **45**: 1107-1113.

Afzal Z, Howton TC, Sun Y, Mukhtar MS (2016) The Roles of aquaporins in plant stress responses. *J Dev Biol*, **4**: 9.

Agarwal PK, Jha B (2010) Transcription factors in plants and ABA dependent and independent abiotic stress signaling. *Biol Plantarum*, **54**: 201-212.

2143 Ahmad R, Liow PS, Spencer DF, Jasieniuk M (2008) Molecular evidence for a single genetic
2144 clone of invasive *Arundo donax* in the United States. *Aquat. Bot*, **88**: 113-120.

2145 Ahuja I, de Vos RC, Bones AM, Hall RD (2010) Plant molecular stress responses face
2146 climate change. *Trends Plant Sci*, **15**(12): 664-674.

2147 Akula R, Ravishankar GA (2011) Influence of abiotic stress signals on secondary metabolites
2148 in plants. *Plant Signal Behav*, **6**(11): 1720-1731.

2149 Alvira P, Tomás-Pejó E, Ballesteros M, Negro MJ (2010) Pretreatment technologies for an
2150 efficient bioethanol production process based on enzymatic hydrolysis: a review.
2151 *Bioresour Technol*, **101**: 4851-4861.

2152 Al-Whaibi MH (2011) Plant heat-shock proteins: a mini review. *J King Saud Univ Sci*, **23**:
2153 139-150.

2154 Anand A, Krichevsky A, Schornack S, Lahaye T, Tzfira T, Tang Y, *et al* (2007) *Arabidopsis*
2155 VIRE2 INTERACTING PROTEIN2 is required for *Agrobacterium* T-DNA integration in
2156 plants. *Plant Cell*, **9**(5): 1695-1708.

2157 Angelini LG, Ceccarini L, Nassi o di Nasso N, Bonari E (2009) Comparison of *Arundo donax*
2158 L. and *Miscanthus* × *giganteus* in a long-term field experiment in Central Italy: analysis of
2159 productive characteristics and energy balance. *Biomass Bioenerg*, **33**: 635-643.

2160 Aysu T, Küçük MM (2013) Liquefaction of giant reed (*Arundo donax* L.) by supercritical
2161 fluid extraction. *Fuel*, **103**: 758-763.

2162 Baldoni E, Genga A, Cominelli E (2015) Plant MYB transcription factors: their role in
2163 drought response mechanisms. *Int J M Sci*, **16**(7): 15811-15851.

2164 Balogh E, Herr JM, Czakó M, Márton L (2012) Defective development of male and female
2165 gametophytes in *Arundo donax* L (Poaceae). *Biomass Bioenerg*, **45**: 265-269.

2166 Barreca F (2012) Use of giant reed *Arundo donax* L. in rural constructions. *Agric Eng Int*:
2167 *CIGR J*, **14**: 46-52.

2168 Barrero RA, Guerrero FD, Moolhuijzen P, Goolsby JA, Tidwell J, Bellgard SE, *et al* (2015)
2169 Shoot transcriptome of the giant reed, *Arundo donax*. *Data Brief*, **3**:1-6.

2170 Berger D, Altmann T (2000) A subtilisin-like serine protease involved in the regulation of
2171 stomatal density and distribution in *Arabidopsis thaliana*. *Genes Dev*, **14**: 1119-1131.

2172 Bhanwra R, Choda SP, Kumar S (1982) Comparative embryology of some grasses. *Proc*
2173 *Indian Acad Sci B*, **48**: 152-162.

2174 Boland JM (2008) The role of floods and bulldozers break-up and dispersal of *Arundo donax*
2175 (giant reed). *Madroño*, **53**: 216-222.

2176 Boltz KA, Jasti M, Townley JM, Shippen DE (2014) Analysis of poly(ADP-ribose)
2177 polymerases in *Arabidopsis* telomere biology. *PLoS ONE*, **9**(2): e88872.

2178 Bose Mazumdar A, Chattopadhyay S (2015) Sequencing, *de novo* assembly, functional
2179 annotation and analysis of *Phyllanthus amarus* leaf transcriptome using the Illumina
2180 platform. *Front Plant Sci*, **6**: 1199.

2181 Bruce WB, Edmeades GO, Barker TC (2002) Molecular and physiological approaches to
2182 maize improvement for drought tolerance. *J Exp Bot*, **53**: 13-25.

2183 Bucci A, Cassani E, Landoni M, Cantaluppi E, Pilu R (2013) Analysis of chromosome
2184 number and speculations on the origin of *Arundo donax* L. (giant reed). *Cytol Genet*, **47**:
2185 237-41.

2186 Byrt CS, Grof CPL, Furbank RT (2011) C4 plants as biofuel feedstocks: optimising biomass
2187 production and feedstock quality from a lignocellulosic perspective. *J Integr Plant Biol*,
2188 **53**: 120-135.

2189 Cardoso-Silva CB, Costa EA, Mancini MC, Balsalobre TWA, Canesin LEC, Rossini Pinto L,
2190 *et al* (2014) *De novo* assembly and transcriptome analysis of contrasting sugarcane
2191 varieties. *PLoS ONE*, **9**(2): e88462.

2192 Ceotto E, Di Candilo M (2010) Shoot cutting propagation of giant reed (*Arundo donax* L.) in
2193 water and moist soil: the path forward? *Biomass Bioenergy*, **11**: 1614-1623.

2194 Chandel AK, Kapoor RK, Singh A, Kuhad RC (2007) Detoxification of sugarcane bagasse
2195 hydrolysate improves ethanol production by *Candida shehate* NCIM 3501. *Bioresour*
2196 *Technol*, **98**: 1947-1950.

2197 Christopher I, Abraham A (1971) Studies on the cytology and phylogeny of South Indian
2198 Bambusoideae, Oryzoideae, Arundinoideae and Festucoideae. *Cytologia*, **36**: 579-594.

2199 Coffman GC, Ambrose RF, Rundel PW (2010) Wildfire promotes dominance of invasive
2200 giant reed (*Arundo donax*) in riparian ecosystems. *Biol Invasions*, **12**: 2723-2734.

2201 Comai L (2005) The advantages and disadvantages of being polyploid. *Nature*, **6**: 836-846.

2202 Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a
2203 universal tool for annotation, visualization and analysis in functional genomics research.
2204 *Bioinformatics*, **21**(18): 3674-3676.

2205 Corno L, Pilu R, Adani F (2014) *Arundo donax* L. A non-food crop for bioenergy and bio-
2206 compound production. *Biotechnol Adv*, **32**: 1535-1549.

2207 Cosentino SL, Copani V, D'Agosta GM, Sanzone E, Mantineo M (2006) First results on
2208 evaluation of Arundo donax L. clones collected in Southern Italy. *Ind Crop Prod*, **23**: 212-
2209 222.

2210 Czaban A, Sharma S, Byrne SL, Spannagl M, Mayer KF, Asp T (2015) Comparative
2211 transcriptome analysis within the Lolium/Festuca species complex reveals high sequence
2212 conservation. *BMC Genomics*, **16**: 249.

2213 Dahl J, Obernberger I (2004) Evaluation of the combustion characteristics of four perennial
2214 energy crops (*Arundo donax, Cynara cardunculus*, *Miscanthus giganteus* and *Panicum*
2215 *virgatum*). 2nd world conference on biomass energy, industry and climate protection, pp.
2216 1265-1270.

2217 Dai A (2012) Increasing drought under global warming in observations and models. *Nat Clim*
2218 *Chang*, **3**: 52-58.

2219 Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu SH, *et al* (2012)
2220 Comparative transcriptomics of three Poaceae species reveals patterns of gene expression
2221 evolution. *Plant J*, **71**: 492-502.

2222 Dedemo GC, Rodrigues FA, Roberto PG, Neto CB, França SC, Zingaretti SM (2013)
2223 Osmoprotection in sugarcane under water deficit conditions. *Plant Stress*, **7**(1): 1-7.

2224 Dhir S, Knowles K, Livia Pagan C, Mann J, Dhir S (2010) Optimization and transformation
2225 of *Arundo donax* L. using particle bombardment. *Afr J Biotechnol*, **9**: 6460-6469.

2226 Di Girolamo G, Grigatti M, Barbanti L, Angelidaki I (2013) Effects of hydrothermal pre-
2227 treatments on giant reed (*Arundo donax*) methane yield. *Bioresour Technol*, **147**: 152-159.

2228 Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in
2229 plant biology. *Am J Bot*, **99**(2): 175-185.

2230 Else JA (1996) Post-flood establishment of native woody species and an exotic, *Arundo*
2231 *donax*, in a Southern California riparian system. San Diego University; [Master Thesis].

Farrell JD, Byrne S, Paina C, Asp T (2014) *De novo* assembly of the perennial ryegrass transcriptome using an RNA-Seq strategy. *PLoS ONE*, **9**(8): e103567.

Ferrandéz-Garcia CE, Andreu-Rodrìguez J, Ferràndez-Garcìa M, Ferràndez-Villena M, Garcìa-Ortuño T (2012) Panel made from giant reed bonded with non-modified starches. *Bioresources*, **7**: 5904-5916.

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*, **39**: 29-37.

Flores JA, Pastor JJ, Martinez-Gabarron A, Gimeno-Blanes FJ, Rodrìguez-Guisado I, Frutos MJ (2011) Arundo donax chipboard based on urea-formaldehyde resin using under 4 mm particle size meets the standard criteria for indoor use. *Ind Crop Prod*, **34**: 1538-1542.

Flores-Yepes JA, Pastor-Perez JJ, Gimeno-Blanes FJ, Rodriguez-Guisado I, Frutos-Fernandez MJ (2012) Full recovery of Arundo donax particleboards from swelling test without waterproof additives. *Bioresources*, **7**: 5222-5235.

Forrest KL, Bhave M (2007) Major intrinsic proteins (MIPs) in plants: a complex gene family with major impacts on plant phenotype. *Funct Integr Genomics*, **7**: 263-289.

Fu C, Mielenz JR, Xiao X, Ge Y, Hamilton CY, Rodriguez M, *et al*. 2011. Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass. *Proc Natl Acad Sci*, **108**: 3803-3808.

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

Fu N, Wang Q, Shen HL (2013) *De novo* assembly, gene annotation and marker development using Illumina paired-end transcriptome sequences in celery (*Apium graveolens* L.). *PLoS ONE*, **8**(2): e57686.

Fu Y, Poli M, Sablok G, Wang B, Liang Y, Porta N, *et al* (2016) Dissection of early transcriptional responses to water stress in *Arundo donax* L. by unigene-based RNA-seq. *Biotechnol Biofuels*, **9**:54.

Galmés J, Pou A, Alsina MM, Tomàs M, Medrano H, Flexas J (2007) Aquaporin expression in response to different water stress intensities and recovery in Richter-110 (*Vitis* sp.): relationship with ecophysiological status. *Planta*, **226**: 671-681.

Gao XP, Wang XF, Lu YF, Zhang LY, Shen YY, Liang Z, *et al* (2004) Jasmonic acid is involved in the water-stress-induced betaine accumulation in pear leaves plant. Plant Cell Environ, **27**: 497-507.

Giessow J, Casanova J, Lecler R, MacArthur R, Fleming G (2011) *Arundo donax* - distribution and impact report. *California Invasive Plant Council*.

Goyer A (2010) Thiamine in plants: aspects of its metabolism and functions. *Phytochemistry*. **71**(14): 1615-1624.

Grabherr M, Haas B, Yassour M, Levin JZ, Thompson DA, Amit I, *et al* (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol*, **29**: 644-652.

Guo M, Liu JH, Ma X, Luo DX, Gong ZH, Lu MH (2016) The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. *Front Plant Sci*, **7**: 114.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, *et al* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*, **8**: 1494-1512.

Haddadchi A, Gross CL, Fatemi M (2013) The expansion of sterile *Arundo donax* (Poaceae) in southeastern Australia is accompanied by genotypic variation. *Aquat Bot*,**104**: 153-161.

Hajouj T, Michelis R, Gepstein S (2000) Cloning and characterization of a receptor-like protein kinase gene associated with senescence. *Plant Physiol*, **124**: 1305-1314.

Hardion L, Verlaque R, Baumel A, Juin M, Vila B (2012) Revised systematics of Mediterranean *Arundo* (Poaceae) based on AFLP fingerprints and morphology. *Taxon*, **61**(6):1217-1226.

Hardion L, Verlaque R, Saltonstall K, Leriche A, Vila B (2014) Origin of the invasive *Arundo donax* (Poaceae): a trans-Asian expedition in herbaria. *Ann Bot*, **114**: 455-462.

Harris MA, Deegan JI, Lomax J, Ashburner M, Tweedie S, Carbon S, *et al* (2008) The gene ontology project in 2008. *Nucleic Acids Res*, **36**: D440-D444.

Haworth M, Cosentino SL, Marino G, Brunetti C, Scordia D, Testa G *et al* (2016) Physiological responses of *Arundo donax* ecotypes to drought: a common garden study. *Glob Change Biol Bioenergy*, doi: 10.1111/gcbb.12348.

He R, Kim MJ, Nelson W, Balbuena TS, Kim R, Kramer R, *et al* (2012) Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, *Phragmites australis* (Poaceae), reveals genes involved in invasiveness and rhizome specificity. *Am. J. Bot*, **99**: 232-247.

Hetherington AM, Woodward FI (2003) The role of stomata in sensing and driving environmental change. *Nature*, **424**: 901-908.

Himmel EM, Ding S, Johnson DK, Adney WS, Nimlos MR, Brady JW, *et al* (2007) Biomass recalcitrance: engineering plant and enzymes for biofuel production. *Science*, **315**: 804-807.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B *et al* (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, **26**(1): 121-135.

Hunter AWS (1934) A karyosystematic investigation in Gramineae. *Can J Res*, **11**: 213-241.

Ilut DC, Coate JE, Luciano AK, Owens TG, May GD, Farmer A, Doyle JJ (2012) A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. *Am J Bot*, **99**(2): 383-396.

Jaradat AA (2010) Genetic resources of energy crops: biological system to combat climate change. *Aust J Crop Sci*, **4**: 309-323.

Jiang T, Zhang XF, Wang XF, Zhang DP (2011) *Arabidopsis* 3-ketoacyl-CoA thiolase-2 (KAT2), an enzyme of fatty acid β-oxidation, is involved in ABA signal transduction. *Plant Cell Physiol*, **52**(3): 528-538.

Joshi R, Wani SH, Singh B, Bohra A, Dar ZA, Lone AA *et al*. (2016). Transcription factors and plants response to drought stress: current understanding and future directions. *Front Plant Sci*, **7**: 1029.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, *et al* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, **36**(1): D480-D484.

Kheshgi HS, Prince RC, Marland G (2000) The potential of biomass fuels in the context of global climate change: focus on transportation fuels. *Annu Rev Environ*, 25: 199-244.

Khudamrongsawat J, Tayyar R, Holt JS (2004) Genetic diversity of giant reed (*Arundo donax*) in the Santa Ana River, California. *Weed Sci*, **52**: 395-405.

Knight H, Trewavas AJ, Knight MR (1997) Calcium signalling in *Arabidopsis thaliana* responding to drought and salinity. *Plant J*, **12**(5): 1067-1078.

Kolmos E, Nowak M, Werner M, Fischer K, Schwarz G, Mathews S, *et al* (2009) Integrating *ELF4* into the circadian system through combined structural and functional studies. *HFSP J*, **3**(5): 350-366.

Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**(10): 1289-1291.

Kovács G, Sorvari S, Scott P, Toldi O (2006) Pyrophosphate: fructose 6-phosphate 1-phosphotransferase operates in net gluconeogenic direction in taproots of cold and drought stressed carrot plants. *Acta Biol Szeged*, **50**(1-2): 25-30.

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H *et al* (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet*, **42**: 1027-1030.

Lamb RS, Citarelli M, Teotia S (2012) Functions of the poly(ADP-ribose) polymerase superfamily in plants. *Cell Mol Life Sci*; **69**: 175-189.

Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. *Science*, **241**: 1077-1080.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth*, **9**: 357-359.

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

Lata C, Yadav A, Prasad M (2011) Role of plant transcription factors in abiotic stress tolerance. In AK Shanker, B. Venkateswarlu (Eds) *Abiotic Stress Response in Plants - Physiological, Biochemical and Genetic Perspectives*, InTech, **10**, pp. 269-296.

Lehmann J, Atzorn R, Brückner C, Reinbothe S, Leopold J, Wasternack C, *et al* (1995) Accumulation of jasmonate, abscisic acid, specific transcripts and proteins in osmotically stressed barley leaf segments. *Planta*, **197:** 156-162.

Lewandowski I, Scurlock JMO, Lindvall E, Christou M (2003) The development and current status of perennial rhizomatous grasses as energy crops in the US and Europe. *Biomass Bioenerg*, **25**: 335-361.

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**:323.

Li D, Deng Z, Qin B, Liu X, Men Z (2012) *De novo* assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics*, **13**: 192.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**: 2078-2079.

Li HZ, Gao X, Li XY, Chen QJ, Dong J, Zhao WC (2013) Evaluation of assembly strategies using RNA-Seq data associated with grain development of wheat (*Triticum aestivum* L.). *PLoS ONE*, **8**(12): e83530.

Li M, Liang Z, Zeng Y, Jing Y, Wu K, Liang J, *et al* (2016) *De novo* analysis of transcriptome reveals genes associated with leaf abscission in sugarcane (*Saccharum officinarum* L.). *BMC Genomics*. **17**: 195.

Li YC, Korol AB, Fahima T, Nevo E: (2004) Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol*, 21: 991-1007.

Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, *et al* (2014) *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature biotechnol*, **32**(10): 1045-1052.

Li Z, Peng J, Wen X, Guo H (2013) *Ethylene-insensitive3* is a senescence-associated gene that accelerates age-dependent leaf senescence by directly repressing *miR164* transcription in *Arabidopsis*. *Plant Cell*, **25**: 3311-3328.

Lim H, Cho MH, Bhoo SH, Hahn TR (2014) Pyrophosphate: fructose-6-phosphate 1-phosphotransferase is involved in the tolerance of Arabidopsis seedlings to salt and osmotic stresses. *In Vitro Cell Dev Biol Plant*, **50**: 84-91.

Liu J, Wang H, Chua NH (2015) Long noncoding RNA transcriptome of plants. Plant *Biotechnol J*, **13**: 319-328.

Liu Y, Zhang P, Song M, Hou J, Qing M, Wang W, *et al* (2015) Transcriptome analysis and development of SSR molecular markers in *Glycyrrhiza uralensis* Fisch. *PloS ONE*, **10**(11): e0143017.

Lopez-Casado G, Covey PA, Bedinger PA, Mueller LA, Thannhauser TW, Zhang S, *et al* (2012) Enabling proteomic studies with RNA-seq: the proteome of tomato pollen as a test case. *Proteomics*, **12**: 761-774.

Mahdieh M, Mostajeran A, Horie T, Katsuhara M (2008) Drought stress alters water relations and expression of PIP-type aquaporin genes in *Nicotiana tabacum* plants. *Plant Cell Physiol*, **49**: 801-813.

Majewski J, Pastinen T (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet*, **27**(2): 72-79.

Mariani C, Cabrini R, Danin A, Piffanelli P, Fricano A, Gomarasca S, *et al* (2010). Origin, diffusion and reproduction of the giant reed (*Arundo donax* L.): a promising weedy energy crop. *Ann Appl Biol*. **157**: 191-202.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: anassessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*,**18**: 1509-1517.

Martin LB, Fei Z, Giovannoni JJ, Rose JK (2013) Catalyzing plant science research with RNA-seq. *Front Plant Sci*, **4**:66.

Matsunaga A, Tsugawa M, Fortes J (2008) Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications. IEEE Fourth International Conference on eScience.

Mattos LM, Moretti CL (2015) Oxidative stress in plants under drought conditions and the role of different enzymes. *Enz Eng*, **5**(1): 136.

Maurel C, Verdoucq L, Luu DT, Santoni V (2008) Plant aquaporins: membrane channels with multiple integrated functions. *Annu. Rev. Plant Biol.*, **59**: 595-624.

Menon V, Rao M (2012) Trends in bioconversion of lignocellulose: biofuels, platforms chemicals & biorefinery concept. *Prog Energy Combust*, **38**: 522-550.

Metzker ML (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet*, **11**: 31-46.

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**: 315-327.

Miyakawa T, Fujita Y, Yamaguchi-Shinozaki K, Tanokura M (2013) Structure and function of abscisic acid receptors. *Trends Plant Sci*, **18**(5): 259-266.

Molina C, Rotter B, Horres R, Udupa SM, Besser B, Bellarmino L, *et al* (2008) SuperSAGE: the drought stress-responsive transcriptome of chickpea roots. *BMC Genomics*, **9**: 553.

Morgante M, De Paoli E, Radovic S (2007) Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol*, **10**: 149-155.

Munné-Bosch S, Alegre L (2004) Die and let live: leaf senescence contributes to plant survival under drought stress. *Funct Plant Biol*, **31**: 203-216.

Munné‐Bosch S, Jubany‐Marí T, Alegre L (2001) Drought‐induced senescence is characterized by a loss of antioxidant defences in chloroplasts. *Plant, Cell Environ*, **24**: 1319-1327.

Muthamilarasan M, Khan Y, Jaishankar J, Shweta S, Lata C, Prasad M (2015) Integrative analysis and expression profiling of secondary cell wall genes in C_4 biofuel model *Setaria italica* reveals targets for lignocellulose bioengineering. *Front. Plant Sci*, **6**: 965.

Nakashima K, Yamaguchi-Shinozaki K (2013) ABA signaling in stress-response and seed development. *Plant Cell Rep*, **32**(7): 959-970.

Nakasugi K, Crowhurst R, Bally J, Waterhouse P (2014) Combining transcriptome assemblies from multiple *de novo* assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLoS ONE*, **9**(3): e91776.

Nassi o Di Nasso N, Roncucci N, Bonari E (2013) Seasonal dynamics of aboveground and belowground biomass and nutrient accumulation and remobilization in giant reed (Arundo donax L.): a three-year study on marginal land. *Bioenergy Res*, **6**: 725-736.

Neto CP, Seca A, Nunes AM, Coimbra MA, Domingues F, Evtuguin D, *et al* (1997) Variation in chemical composition and structure ofmacromolecular components in different morphological regions and maturity stages of *Arundo donax*. *Ind Crop Prod*, **6**: 51-57.

Obataya E, Norimoto M (1999) Mechanical relaxation processes due to sugars in cane (*Arundo donax* L.). *J Wood Sci*, **45**: 378-383.

2438 Ohlrogge J, Allen D, Berguson B, DellaPenna D, Shachar-Hill Y, Stymne S (2009) Driving
2439 on biomass. *Science*, **324**: 1019-1020.

2440 Perdue RE (1958) *Arundo donax* - source of musical reeds and industrial cellulose. *Econ Bot*,
2441 **12**: 368-404.

2442 Pérez-de-Castro AM, Vilanova S, Cañizares J, Pascual LM, Blanca JJ, Diez M *et al* (2012)
2443 Application of genomic tools in plant breeding. *Curr genomics*, **13**(3):179-195.

2444 Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) Signalp 4.0: discriminating signal
2445 peptides from transmembrane regions. *Nat Methods*, **8**(10): 785-786.

2446 Pillitteri LJ, Dong J (2013) Stomatal development in Arabidopsis. *Arabidopsis Book*, e0162.

2447 Pilu R, Cassani E, Landoni M, Cerino Badone F, Passera A, Cantaluppi E, *et al* (2014)
2448 Genetic characterization of Italian giant reed (*Arundo donax* L.) clones collection:
2449 exploiting clonal selection. Euphytica,**196**: 169-181.

2450 Pizzolongo P (1962) Osservazioni cariologiche su *Arundo donax* e *Arundo plinii*. Annu Bot,
2451 **27**:173-187.

2452 Pompeiano A, Remorini D, Vita F, Guglielminetti L, Miele S, Morini S (2016) Growth and
2453 physiological response of *Arundo donax* L. to controlled drought stress and recovery. *Plant*
2454 *Biosyst*, http://dx.doi.org/10.1080/11263504.2016.1249427.

2455 Pompeiano A, Vita F, Miele S, Guglielmetti L (2013) Freeze tolerance and physiological
2456 changes during cold acclimation of giant reed [*Arundo donax* (L.)]. *Grass Forage Sci*, doi:
2457 10.1111/gfs.12097.

2458 Poorter H, Villar R (1997) The fate of acquired carbon in plants: chemical composition and
2459 construction costs. In Bazzaz FA, Grace J editors. Resource Allocation in Plants. San
2460 Diego: *Academic Press*, pp. 39-72.

2461 Qin F, Sakuma Y, Tran LSP, Maruyama K, Kidokoro S, Fujita Y *et al*. (2008). *Arabidopsis*
2462 DREB2A-interacting proteins function as RING E3 ligases and negatively regulate plant
2463 drought stress-responsive gene expression. *Plant Cell*, **20**(6): 1693-1707.

2464 Qin Z, Zhuang Q, Zhu X, Cai X, Zhang X (2011) Carbon consequences and agricultural
2465 implications of growing biofuels crops on marginal agricultural lands in China. *Environ*
2466 *Sci Technol*, **45**: 10765-10772.

2467 Quinn LD, Holt JS (2008) Ecological correlates of invasion by Arundo donax in three
2468 southern California riparian habitats. *Biol Invasions*, **10**: 591-601.

2469 Quinn LD, Rauterkus MA, Holt JS (2007) Effects of nitrogen enrichment and competition on
2470 growth and spread of giant reed (*Arundo donax*). *Weed Sci*, **55**: 319-326.

2471 Ragaglini G, Dragoni F, Simone M, Bonari E (2014) Suitability of giant reed (*Arundo donax*)
2472 for anaerobic digestion: effect of harvest time and frequency on biomethane yield
2473 potential. *Bioresour Technol*, **152**: 107-115.

2474 Rapala-Kozik M, Kowalska E, Ostrowska K (2008) Modulation of thiamine metabolism in
2475 *Zea mays* seedlings under conditions of abiotic stress. *J Exp Bot*, **59**(15): 4133-4143.

2476 Rezaei MK, Shobbar ZS, Shahbazi M, Abedini R, Zare S (2013) Glutathione S-transferase
2477 (GST) family in barley: identification of members, enzyme activity, and gene expression
2478 pattern. *J Plant Physiol*, **170**(14): 1277-1284.

2479 Richmond TA, Somerville CR (2000) The cellulose synthase superfamily. *Plant Physiol*, **124**:
2480 495-498.

2481 Riemann M, Dhakarey R, Hazman M, Miro B, Kohli A, Nick P (2015) Exploring jasmonates
2482 in the hormonal network of drought and salinity responses. *Front Plant Sci*, **6**: 1077.

2483 Rivero RM, Kojima M, Gepstein A, Sakakibara H, Mittler R, Gepstein S, Blumwald E (2007)
2484 Delayed leaf senescence induces extreme drought tolerance in a flowering plant. *Proc Natl*
2485 *Acad Sci*, **104**(49): 19631-19636.

2486 Rizhsky L, Liang H, Mittler R (2002) The combined effect of drought stress and heat shock
2487 on gene expression in tobacco. *Plant Physiol*, **130**(3): 1143-1151.

2488 Rizhsky L, Liang H, Shuman J, ShulaevV, Davletova S, Mittler R (2004) When defense
2489 pathways collide: the response of Arabidopsis to a combination of drought and heat stress.
2490 *Plant Physiol*, **134**: 1683-1696.

2491 Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, *et al* (2010) *De novo*
2492 assembly and analysis of RNA-seq data. *Nat Methods*, **7**(11): 909-912.

2493 Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for
2494 differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1): 139-
2495 140.

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**: R25.

Sablok G, Fu Y, Bobbio V, Laura M, Rotino GL, Bagnaresi P, *et al* (2014) Fuelling genetic and metabolic exploration of C_3 bioenergy crops through the first reference transcriptome of *Arundo donax* L. *Plant Biotechnol J*, doi:10.1111/pbi.12159.

Sano T, Higaki T, Handa K, Kadota Y, Kuchitsu K, Hasezawa S *et al* (2006). Calcium ions are involved in the delay of plant cell cycle progression by abiotic stresses. *FEBS Lett*, ***580***(2), 597-602.

Schievano A, D'Imporzano G, Orzi V, Colombo G, Maggiore T, Adani F (2015) Biogas from dedicated energy crops in Northern Italy: electric energy generation costs. *Glob Change Biol Bioenergy*, **7**: 899-908.

Scordia D, Cosentino SL, Lee J, Jeffries TW (2011) Dilute oxalic acid pretreatment for biorefining giant reed (*Arundo donax* L.). *Biomass Bioenerg*, **35**: 3018-3024.

Seca AML, Cavaleiro JAS, Domingues FMJ, Silvestre AJD, Evtuguin D, Neto CP (2000) Structural characterization of lignin from the nodes and internodes of *Arundo donax* reed. *J Agric Food Chem*, **48**: 817-824.

Selmar D, Kleinwächter M (2013) Stress enhances the synthesis of secondary plant products: the impact of the stress-related over-reduction on the accumulation of natural products. *Plant Cell Physiol*, **54**(6): 817-826.

Shanker AK, Maheswari M, Yadav SK, Desai S, Bhanu D, Attal NB, Venkateswarlu B (2014) Drought stress responses in crops. *Funct Integr Genomics*, **14**: 11-22.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**: 3210-3212.

Singh D, Laxmi A (2015) Transcriptional regulation of drought response: a tortuous network of transcriptional factors. *Front. Plant Sci*, **6**:895.

Smith S and de Lange T (2000) Tankyrase promotes telomere elongation in human cells. *Curr Biol*, **10**(20): 1299-1302.

Steudle E; Peterson CA (1998) How does water get through roots? *J Exp Bot*, **49**(322): 775-788.

Sukrong S, Yun KY, Stadler P, Kumar C, Facciuolo T, Moffatt BA *et al* (2012) Improved growth and stress tolerance in the *Arabidopsis oxt1* mutant triggered by altered adenine metabolism. *Mol Plant*. **5**(6): 1310-1332.

Takagi H, Ishiga Y, Watanabe S, Konishi T, Egusa M, Akiyoshi N, *et al* (2016) Allantoin, a stress-related purine metabolite, can activate jasmonate signaling in a MYC2-regulated and abscisic acid-dependent manner. *J Exp Bot*, **67**(8): 2519-2532.

Takahashi W, Takamizo T, Kobayashi M, Ebina M (2010) Plant regeneration from calli in giant reed (*Arundo donax* L.). *Grassl Sci*, **4**: 56224-56229.

Tang W, Ji Q, Huang Y, Jiang Z, Bao M, Wang H, Lin R (2013) FAR-RED ELONGATED HYPOCOTYL3 and FAR-RED IMPAIRED RESPONSE1 transcription factors integrate light and abscisic acid signaling in Arabidopsis. *Plant Physiol*, **163**(2): 857-866.

Temiz A, Akbas S, Panov D, Terziev N, Alma MH, Parlak S, *et al* (2013) Chemical composition and efficiency of bio-oil obtained from giant cane (*Arundo donax* L.) as a wood preservative. *Bioresources*, **8**: 2084-2098.

Tian XJ, Long Y, Wang J, Zhang JW, Wang YY, Li WM, *et al* (2015) *De novo* transcriptome assembly of common wild rice (*Oryza rufipogon* Griff.) and discovery of drought-response genes in root tissue based on transcriptomic data. *PLoS ONE*, **10**(7): e0131455.

Tracy JL, DeLoach CJ (1998) Suitability of classical biological control for giant reed (*Arundo donax*) in the United States. *Proceedings of the Arundo and saltcedar management workshop*, Ontario.

Tuteja N (2007). Abscisic acid and abiotic stress signaling. *Plant Signal Behav*, **2**(3): 135-138.

Tuteja N, Mahajan S (2007) Calcium signaling network in plants: an overview. *Plant Signal Behav*, **2**(2): 79-85.

Udvardi MK, Kakar K, Wandrey M, Montanri O, Murray J, Andraiankaja A *et al* (2007) Legume transcription factors: global regulators of plant development and response to the environment. *Plant Physiol*, **144**: 538-549.

Umezawa T, Sugiyama N, Takahashi F, Anderson JC, IshihamaY, Peck SC, *et al* (2013) Genetics and phosphoproteomics reveal a protein phosphorylation network in the abscisic acid signaling pathway in *Arabidopsis thaliana*. *Sci Signal*, 6(270): rs8.

Unamba CI, Nag A, Sharma RK (2015) Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. *Front Plant Sci*, **6**: 1074.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, *et al* (2012) Primer3- new capabilities and interfaces. *Nucleic Acids Res*, 40(15): e115.

Vij S, Tyagi AK (2006) Genome-wide analysis of the stress associated protein (SAP) gene family containing A20/AN1 zinc-finger (s) in rice and their phylogenetic relationship with *Arabidopsis*. *Mol Gen Genomics*, **276**: 565-575.

Vincent D, Lapierre C, Pollet B, Cornic G, Negroni L, Zivy M (2005) Water deficits affect caffeate O-methyltransferase, lignification, and related enzymes in maize leaves. A proteomic investigation. *Plant Physiol*, **137**: 949-960.

Virtue JG, Reynolds T,Malone J, Preston C,Williams C (2010) Managing the weed risk of cultivated *Arundo donax* L. *Proceedings of 17th Australasian weeds conference*, Christchurch.

Von Groll U, Berger D, Altmann T (2002) The subtilisin-like serine protease SDD1 mediates cell-to-cell signaling during Arabidopsis stomatal development. *Plant Cell*, **14**: 1527-1539.

Wang J, Nayak S, Koch K, Ming R (2013) Carbon partitioning in sugarcane (*Saccharum* species). *Front Plant Sci*, **4**: 201.

Wang P, Liu H, Hua H, Wang L, Song CP (2011) A vacuole localized β-glucosidase contributes to drought tolerance in *Arabidopsis*. *Chinese Sci Bull*, **56**(33): 3538-3546.

Wang W, Vinocur B, Shoseyov O and Altman A (2004) Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci*, **9**(5): 244-252.

Wang Y, Yang C, Jin Q, Zhou D, Wang S, Yu Y, *et al* (2015) Genome-wide distribution comparative and composition analysis of the SSRs in Poaceae. *BMC Genet*, **16**:18.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat.Rev.Genet*, **10**: 57-63.

Ward JA, Ponnala L, Weber CA (2012) Strategies for transcriptome analysis in nonmodel plants. *Am J Bot*, **99**(2): 267-276.

White PJ, Broadley MR (2003) Calcium in plants. *Annal Bot*, **92**: 1-25.

Wijte AH, Mizutani T, Motamed ER, Merryfield ML, Miller DE (2005) Temperature and endogenous factors cause seasonal patterns in rooting by stem fragments of the invasive giant reed, *Arundo donax* (Poaceae). *Int J Plant Sci*, **166**: 507-517.

Williams CMJ, Biswas TK, Schrale G, Virtue JG, Heading S (2008) Use of saline land and wastewater for growing a potential biofuel crop (*Arundo donax* L.). In *Irrigation Australia 2008 Conference*.

Xia EH, Yao QY, Zhang HB, Jiang JJ, Zhang LP, Gao LZ (2015) CandiSSR: an efficient pipeline used for identifying candidate polymorphic SSRs based on multiple assembled sequences. *Front. Plant Sci*, **6**: 1171.

Xin ZY, Zhou X, Pilet PE (1997) Level changes of jasmonic, abscisic and indole-3yl acetic acids in maize under desiccation stress. *J Plant Physiol*, **151**: 120-124.

Xu C, Jiao C, Zheng Y, Sun H, Liu W, Cai X, Dai S (2015) *De novo* and comparative transcriptome analysis of cultivated and wild spinach. *Sci Rep*, **5**:17706.

Xu J, Xing XJ, Tian YS, Peng RH, Xue Y, Zhao W, Yao QH (2015) Transgenic *Arabidopsis* plants expressing tomato glutathione S-transferase showed enhanced resistance to salt and drought stress. *PloS ONE*, **10**(9): e0136960.

Xu Z, Zhou G (2008) Responses of leaf stomatal density to water status and its relationship with photosynthesis in a grass. *J Exp Bot*, **59**(12): 3317-3325.

Yancey PH (2005) Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses. *J Exp Biol*, **208**: 2819-2830.

Yoo CY, Hasegawa PM, Mickelbart MV (2011) Regulation of stomatal density by the GTL1 transcription factor for improving water use efficiency. *Plant Signal Behav*, **6**(7): 1069-1071.

Zeven AC, de Wet JMJ (1982) Dictionary of cultivated plants and their regions of diversity: excluding most ornamentals, forest trees and lower plants. 2nd rev ed, Centre Agricultural Pub & Document, Wageningen.

Zhao Y, Williams R, Prakash CS, He G (2013) Identification and characterization of gene-based SSR markers in date palm (*Phoenix dactylifera* L). *BMC Plant Biol*, **12**: 237-244.

Zhong R, Ye ZH (2015) Secondary cell walls: biosynthesis, patterned deposition and transcriptional regulation. *Plant Cell Physiol*, **56**(2): 195-214.

Zhou Y, Gao F, Liu R, Feng J, Li H (2012) *De novo* sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in *Ammopiptanthus mongolicus*. *BMC Genomics*, **13**: 266.

9. LIST OF PAPERS

Evangelistella C, Valentini A, Ludovisi R, Firrincieli A, Fabbrini F, Scalabrin S, Cattonaro F, Morgante M, Scarascia Mugnozza G, Keurentjes JJB, Harfouche A. *De novo* assembly, functional annotation and analysis of the giant reed (*Arundo donax* L.) leaf transcriptome provide tools for the development of a biofuel feedstock. (Submitted to *Biotechnology for Biofuels*).

Ludovisi R[*], Evangelistella C[*], Fabbrini F, Douthe C, Puertolas J, Latini P, Firrincieli A, Scarascia-Mugnozza G, Dodd IC, Davies B, Flexas J, Keurentjes JJB, Harfouche A. Morpho-physiological and molecular responses of *Arundo donax* L. to moderate drought stress.([*]: These authors contributed equally to this work; in preparation to submission in *Plant Cell & Environment*).

10. LIST OF TABLES

11. LIST OF FIGURES