Universidade de Lisboa

Faculdade de Ciências

Departamento de Informática

Transcriptomic analysis of maritime pine response to infection with *Bursaphelenchus xylophilus*, the causing agent of pine wilt disease

Mestrado em Bioinformática e Biologia Computacional

Especialização em Bioinformática

Dissertação orientada por:
António Marcos Costa do Amaral Ramos
Cátia Luísa Santana Calisto Pesquita

Daniel Filipe Branco Gaspar

2016

Universidade de Lisboa

Faculdade de Ciências

Departamento de Informática

Transcriptomic analysis of maritime pine response to infection with *Bursaphelenchus xylophilus*, the causing agent of pine wilt disease

Mestrado em Bioinformática e Biologia Computacional

Especialização em Bioinformática

Daniel Filipe Branco Gaspar

Dissertação orientada por:

António Marcos Costa do Amaral Ramos

Cátia Luísa Santana Calisto Pesquita

2016

“A ciência serve para nos dar uma ideia de quão extensa é a nossa ignorância”

Félicité Robert de Lamennais

ACKNOWLEGMENTS

Esta dissertação é o culminar de mais uma etapa de apredizagem e crescimento a nível pessoal e profissional. Embora, pela sua finalidade académica, uma dissertação seja um trabalho individual, há contributos que devem ser realçados. Nesse sentido, desejo expressar o meu sincero reconhecimento e gratidão pela inestimável ajuda:

Ao meu orientador, Doutor António Marcos Ramos, pela oportunidade que me concedeu, pelo apoio diário, pelos conselhos e rigor científico das orientações, indispensáveis na realização deste trabalho, e em especial pela amizade e confiança.

À minha orientadora, Doutora Cátia Pesquita, pela disponibilidade permanente, pelo acompanhamento e interesse demonstrado nas várias etapas deste trabalho. As suas sugestões e partilha de conhecimento foram indispensáveis.

À Doutora Anabel Chimenos, pela disponibilidade e apoio prestado em todas as tarefas realizadas, pela indispensável partilha de conhecimentos. Acima de tudo, pela amizade e convivência.

Aos Mestres Brígida Meireles e Pedro Barbosa, por toda a ajuda, pela integração e convivência no CEBAL, amizade e companheirismo.

A todas as pessoas do CEBAL que de alguma forma me ajudaram.

Por último, mas nunca em último, aos meus Pais e à minha Avó. Porque mesmo quando a distância nos separa, não me deixam caminhar sozinho.

A todos vós, o meu muito sincero, Obrigado.

RESUMO

A bioinformática é uma área mutidisciplinar que envolve a aplicação de técnicas computacionais para analisar informação biológica em larga escala. Este conjunto de ferramentas e técnicas computacionais foi desenvolvido para dar suporte à análise da crescente quantidade de dados gerados neste domínio, e em particular por técnicas de *next-generation sequencing*. Uma das áreas da biologia largamente dependente das ferramentas bioinformáticas é a análise do perfil do transcriptoma. Atualmente, a técnica de *RNA-sequencing* tem sido a abordagem predominante em estudos transcriptómicos de dados de sequenciação. Esta técnica tem sido bastante usada em estudos de resistência, especialmente em espécies florestais ameaçadas.

A Floresta é um recurso natural essencial em termos globais, não apenas pela sua importância a nível ecológico, mas também a nível económico e paisagístico. Ela representa um suporte de Vida na Terra, fornecendo inúmeros benefícios fundamentais para o equilibrio de diversos ecossistemas. No entanto, recentemente tem vindo a verificar-se um preocupante declínio de várias espécies florestais, sendo o Pinheiro Bravo (*Pinus pinaster* Ait.) uma das mais afectadas. Este declínio tem causado um impacto negativo no equilibrio dos ecossistemas e na manutenção da biodiversidade. Um dos organismos com maior potencial destrutivo para a área florestal de Pinheiro Bravo é o nemátodo da madeira do pinheiro (*Bursaphelenchus xylophilus*), um verme microscópico responsável pela doença da murchidão do pinheiro. Numa tentativa de reduzir as perdas resultantes da doença, surgiram vários estudos de resistência do hospedeiro para a identificação de árvores com menor susceptibilidade à infecção. No entanto, parte desses estudos apresenta uma abordagem mais tradicional, sem recurso às novas tecnologias de sequenciação. Nesse sentido, o presente trabalho, baseado no estudo de dados de *RNA-sequencing* produzidos pela plataforma de sequenciação Ion Proton, tem como principal objectivo a caracterização da resposta do Pinheiro Bravo à infecção com o nemátodo da madeira do pinheiro entre três diferentes estágios após inoculação. Para isso, foram

identificados genes diferencialmente expressos, vias metabólicas e marcadores moleculares potencialmente associados à resistência à doença.

Um total de 355,287 unigenes foram obtidos a partir de um conjunto de 176,282,168 *reads* sequenciadas para todas as bibliotecas, pela técnica de *de novo assembly*. A baixa percentagem de genes predictos (23.5%) a partir do conjunto de unigenes assemblados e o elevado número de genes sem anotação ou com anotação desconhecida, evidenciam as limitações existentes num estudo de RNA-Seq em espécies não-modelo, sem o genoma sequenciado, como é o caso do Pinheiro Bravo. Apesar disso, foram obtidos 17,533 genes diferencialmente expressos entre todas as comparações. No seguimento desta análise, há a evidência de duas fases de resposta à infecção. Em primeiro lugar, é desencadeada uma resposta imediata, logo após a infecção. Posteriormente, uma segunda fase de resposta parece acontecer aos 7 dias após a infecção. Foi ainda identificado um conjunto de genes candidatos envolvidos na resistência à doença nos vários estágios em estudo. Desse conjunto, é possível identificar genes envolvidos no metabolismo secundário, stress oxidativo e defesa contra infeção de agentes patogénicos. Este estudo representa uma nova abordagem ao nivel dos mecanismos moleculares e vias metabólicas envolvidas na defesa contra a infeção do nemátodo da madeira do pinheiro. Podendo assim ser um recurso útil para estudos ulteriores e também para programas de melhoramento com vista à seleção de plantas com menos susceptibilidade à doença.

Palavras-chave: Bioinformática; *Next-generation sequencing*; *RNA-Sequencing; Pinus pinaster*; *Bursaphelenchus xylophilus*; doença da murchidão do pinheiro.

ABSTRACT

Bioinformatics is a multidisciplinary field that involves the application of computational tools to analyze biological information on a large-scale. This set of computational techniques were developed to support the analysis of the increasing amount of data generated in this area, and in particular by next-generation sequencing (NGS). One of the main fields of biology that is largely dependent on bioinformatics tools is the transcriptome profile analysis. Currently, RNA-Sequencing (RNA-Seq) is the dominant transcriptomics approach for NGS data. RNA-Seq has been highly used in disease pathogenesis studies, especially in endangered forest species.

Forests are essential resources on a global scale, not only for the ecological benefits, but also for economical and landscape purposes. They represent one of the Life support systems on Earth, providing essential resources for a range of ecosystems. However, in recent years there has been a worrying decline of a large number of forest species around the world, with maritime pine (*Pinus pinaster* Ait.) being one of the most affected. This alarming decay is caused by abiotic and biotic factors. Within this last group of factors we must highlight the pine wood nematode (PWN), *Bursaphelenchus xylophilus* as one of the main responsible. PWN is a microscopic organism reported for the first time in Portugal in 1999, being the causal agent of pine wilt disease (PWD). In an attempt to reduce losses arising by PWD, the study of maritime pine resistance is one of the research programs that recently started in Portugal, aiming to improve their resistance and select trees with lower susceptibility to infection. However, just a few of these studies were based on next-generation sequencing data. Taking this into account, this study is an approach to pine wilt disease, using RNA-Sequencing data produced by Ion Proton platform. The aims of this study was to analyze RNA-Seq data to characterize the maritime pine transcriptome in the response to infection with *Bursaphelenchus xylophilus,* over three different time stages after inoculation of the PWN, by determining the differentially expressed genes,

regulatory networks and pathways, with the purpose of identifying potential genes involved in resistance against PWD.

A total of 355,287 unigenes were obtained by *de novo* assembly from the 176,282,168 sequenced reads for all libraries. Moreover, we obtained 17,533 differentially expressed genes (up and down regulated) between all comparatives. The low rate of predicted genes (23.5%) from the set of assembled contigs and the high number of genes without annotation or with "Unknown" annotation, evidences the existing limitations when working in RNA-Seq studies with non-model species like *Pinus pinaster.* Despite this, further analysis suggest an early response that may occur immediately after inoculation and a late response that may occur 7 days after inoculation.

A set of candidate genes involved in resistance against PWN infection were identified over different time points. These genes were related to secondary metabolism, oxidative stress and defense against pathogen infection. Our results provide new insights about the molecular mechanism and metabolic pathways involved in resistance of *Pinus pinaster* against PWN infection. It may be a useful resource in future studies and for future breeding programs to select plants with lower susceptibility to PWD.

Keywords: Bioinformatics; Next-generation sequencing; RNA-Sequencing; *Pinus pinaster*; *Bursaphelenchus xylophilus*; Pine wilt disease.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BP	Biological Process
CC	Cellular Component
cDNA	Complementary Deoxyribonucleic Acid
DEG	Differentially Expressed Genes
FDR	False Discovery Rate
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
MF	Molecular Function
MR	Mapped Reads
mRNA	Messenger Ribonucleic Acid
NGS	Next-Generation Sequencing
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PDA	Potato Dextrose Medium
PP	*Pinus pinaster*
PWD	Pine Wood Disease
PWN	Pine Wood Nematode
QC	Quality Control
RNA	Ribonucleic Acid
SAM	Sequence Alignment Map
SNP	Single Nuclear Polymorphism
VCF	Variant Call Format

1. - INTRODUCTION

Maritime pine (*Pinus pinaster* Ait) is one of the main forest species in southwestern Europe, having a high economic impact due to the value of the wood and resin. However, recently a serious decline of maritime pine populations has been observed, with pine wood nematode (*Bursaphelenchus xylophilus*) being one of the main agents responsible for the decline. Over the last years, some studies in this area were executed using different approaches. However, just a few of these studies were based on next-generation sequencing (NGS) data. This study is an approach to study pine wilt disease (PWD), using RNA-Seq data produced by the Ion Proton platform. RNA-Seq is largely used in resistance studies, being especially useful to characterize transcriptome profile over different time points. This technique involves a set of steps to process NGS data, allowing the identification of candidate genes and molecular markers associated to the resistance against PWD.

1.1 – Motivation

RNA-Seq is a revolutionary technology widely used to characterize transcriptome profile over different time points, using deep-sequencing technologies. However, in terms of bioinformatics analysis, these type of approaches require to take into account some aspects that can limit the appropriate approaches to use. The most important limitation is to work with non-model organisms like *Pinus pinaster*, which there is no genome sequence available in public databases. In this sense, this study pretends to contribute to the bioinformatics field, providing a RNA-Seq analysis workflow for a non-model species that could, in future, be applied and adapted to similar studies.

1.2 – Objectives

The purpose of this study was to analyze RNA-Seq data to characterize the maritime pine transcriptome in the response to infection with *Bursaphelenchus xylophilus*. In order to carry out this work, four libraries of RNA-Seq data were sequenced by the Ion Proton platform. The four libraries corresponds to three different time stages after inoculation of the PWN plus the control sample. So, by determining the differentially expressed genes over those libraries, and the regulatory networks and pathways involved, we were able to identify potential candidate genes associated with resistance against PWD. In this context, a RNA-Seq analyses workflow was established and a several bioinformatics tools were used to achieve these aims. In figure 1 are represented all capital stages followed in this study. This dissertation focuses only in the bioinformatics analysis of the sequenced libraries.

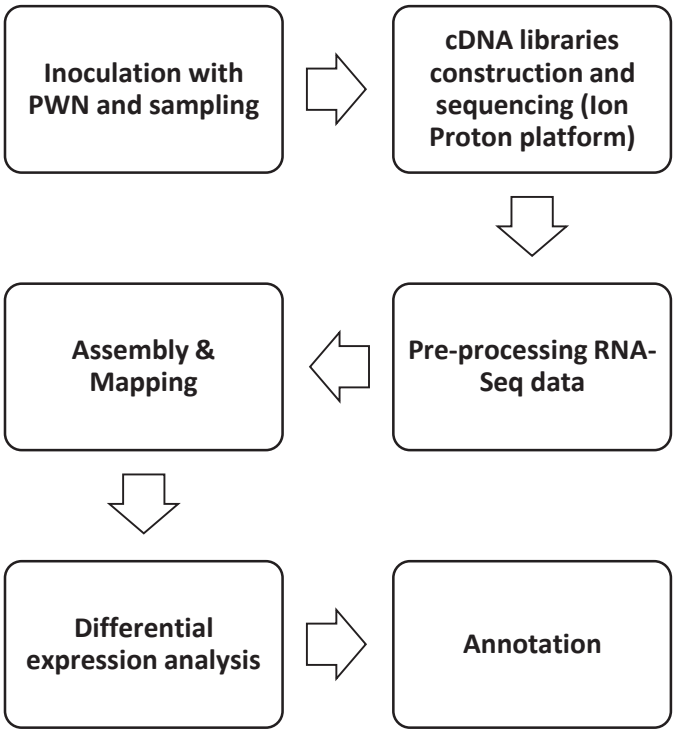

Figure 1 - Representation of workflow applied in this study.

1.3 – Maritime Pine and Pine wilt disease

Forests are much more than a large area of land covered with trees. They represent one of the life support systems on Earth, providing essential resources for a range of ecosystems. Furthermore, forests supply various products and services, generating a wide range of economic and social benefits. Due to the commercial value of wood products, maritime pine (*Pinus pinaster* Ait.) is one of the main conifer species in southwestern of Europe, covering approximately 4 million hectares in this area (Plomion *et al*., 2000). In Portugal, maritime pine has been considered by many as one of the predominant tree species, and by far the most widespread, mainly in the regions of Atlantic influence, covering over than 700 thousand hectares, that corresponds to 23% of the total forest surface (ICNF – IFN, 2013).

In recent years there has been a worrying decline of a large number of forest species around the world, with maritime pine being one of the most affected. This alarming decay is caused by abiotic and biotic factors, and within this last group of factors we must highlight the pine wood nematode, *Bursaphelenchus xylophilus* (Steiner & Buhrer, 1934) (Nickle, 1970) as one of the main culprits (Futai et al., 2008).

PWN is a quarantine organism in the European Union (Directive 77/93 EEC), being the causal agent of the pine wilt disease (PWD), that may kill a host tree within a short period of time after infection (Mota et al., 1999). Mostly due to this pathogen, the total area occupied by *P. pinaster* suffered an abrupt decline in Portugal, accounting for losses of 263,000 hectares between 1995 and 2010 (AFN, 2010). As a result, *P. pinaster* went from being the main forest species, in terms of distribution and area, to the third, behind eucalyptus and cork oak. Recently, it has been identified as an endangered species by the IUCN red list of threatened species (Farjon, 2010).

PWN was reported for the first time in Portugal in 1999 (Mota et al., 1999), and in less than 10 years the whole *P. pinaster* area has been affected. PWN is transported between host trees by an insect vector, a longhorn cerambycid beetle (*Monochamus*

galloprovincialis Oliv.) (Sousa E. et al, 2001). The transmission may occur in two forms: by oviposition, whereby the female beetles laying their eggs under the bark of stressed or recently killed trees by the PWN, and the nematodes migrate to pupae just before adult beetles emerge, ensuring successful survival of the parasite; or via transmission by feeding that occurs through beetle feeding wounds (primary transmission). Nematodes carried by beetles move into wounds and breed in the xylem, nonetheless, the survival of nematodes is not guaranteed (Edwards & Linit, 1992; Fielding & Evans, 1996). This is a close relationship between PWN and its vector beetle, resulting in the epidemiological cycle of PWD (Futai et al., 2008).

PWD expression depends not only on the pathogenicity of PWN and susceptibility of host trees but also on environmental conditions such as high temperature and large soil moisture, the optimal conditions for PWN proliferation (Fielding & Evans, 1996). The symptoms caused by PWD are common to other diseases, and therefore can easily be confused. A typical early symptom is needle discoloration. Needles turn grayish green, then tan, and finally brown. Then, resin flow ceases and the wood is dry when cut (Futai et al., 2008).

The defensive mechanisms of host trees can be divided into early and advanced stage (Fukuda, 1997). In the first stage, defensive response occurs in both susceptible and resistant trees, nonetheless, late response is found only in susceptible trees (Fukuda, 1997). In the same species, it has been verified the existence of trees with different levels of susceptibility, some of which survive the infection, thus, constituting an opportunity for selective breeding. This has been the approach in breeding programs developed in China and Japan over the last years (FAO, 1985).

1.4 - Next-generation sequencing

Before discussing the applications and impact of next-generation sequencing (NGS), also known as massively parallel sequencing, on genomics research, it is necessary to look back on the history of sequencing development, to review basic concepts and the evolution of NGS systems. The NGS term describes a set of platforms that represent the evolution of sequencing technology from the Sanger system, and has provided unprecedented opportunities. Their use has changed scientific approaches, enabling whole genome or individual genes sequencing, having applications in various fields, including plant biology (Liu et al., 2012).

In 2005 the first NGS platform was launched by 454 Life Sciences (www.454.com). This system is based in the principle of pyrosequencing or sequencing by synthesis. In brief, this process starts with an emulsion PCR in which single-stranded DNA binding beads are encapsulated. During the pyrosequencing mechanism, a successful incorporation of a nucleotide is converted to light emission from the release of pyrophosphate molecules (Liu et al., 2012) (Mardis, 2013). Initially, the 454 system had a read length of 100-150bp, however, it was upgraded to 600-700bp with a 99.9% accuracy after filtering and with an output of 0.7Gb data per run (Liu et al., 2012). The 454 platform was a revolutionary technology that represented an important progress in terms of speed, throughput and allowed reducing the per-base cost over Sanger technology (Van Dijk, Auger, Jaszczyszyn, & Thermes, 2014).

The second platform launched, and presently the most widely used, was the Illumina system from Solexa (www.illumina.com). Briefly, in this system, libraries are loaded into a flow cell and each bound fragment is amplified into a clonal cluster through bridge amplification. Four kinds of fluorescently labeled nucleotides are added and as they are incorporated a characteristic signal is emitted. This emission wavelength is recorded and used to identify the base (Mardis, 2013). Illumina have shorter read lengths (150-300bp) when compared with the 454 system, but produce more reads and have higher

throughput per run (~1,500Gb). The Illumina sequencing system also has a lower cost per base than older platforms.

In 2006, Applied Biosystems released the SOLiD platform (Sequencing by Oligo Detection), a system that requires an emulsion PCR approach with small magnetic beads for DNA fragment amplification. The technology of two-base sequencing is used during the sequencing mechanism, where the libraries are sequenced by 8 base-probe ligation with a specific fluorescent marker, which identifies a two-base combination. The probes light signal is recorded and after five cycles the sequence of an entire fragment can be deduced (Liu et al., 2012; Zhang et al., 2011). The SOLiD system has a high accuracy (99.99%) after filtering, producing reads with an average length of 85bp. Nonetheless, is slightly more expensive than Illumina system and may take a few more days per run (Liu et al., 2012).

One of the most recent NGS platform is the Ion Proton™ system, developed by Ion Torrent in 2010. This technology differs from other existing platforms in base detection. It measures slight variations in pH levels, which is caused by the releasing of Hydrogen ions during base incorporation into a strand of DNA by a polymerase (Ion Proton™ system guidelines), instead of measuring light released from fluorescent or chemiluminescent reagents as other platforms do. This sequencer machines use only an ion sensor, therefore it does not require camera scanning or light. For this approach, libraries are amplified by emulsion PCR and each fragment is attached to one bead. These beads are placed into the wells of Ion Chips (Mardis, 2013). The Ion Proton platform has a higher sequencing speed and lower cost per base comparing to oldest platforms. Moreover, it produces up to 10Gb throughput per run with a read length of up to 200bp (Ion Proton™ system Documentation).

Over the last years, some new NGS platforms emerged on the market. These technologies, also called third generation sequencing, promise to deliver entire genomes in less than a day, increasing the applicability of sequencing technologies (Schadt et al., 2010). An example of third generation platform is the single-molecule real-time, launched

by Pacific Biosciences, enabling real-time observation of DNA synthesis (Schadt et al., 2010).

Since the introduction of the first NGS platform, there was a revolution in the biological research field, which allowed a fast progress in terms of reducing costs, increasing throughput and accuracy. Every day, more organisms are being sequenced, and lots of new raw data are constantly becoming available to be analyzed. This fast paced evolution provides new opportunities and enables additional studies and projects in genomics, metagenomics, epigenomics, exomics and also in transcriptomics which has contributed to the decline of microarrays technology. Furthermore, NGS is being used in forensic genetics and in clinical diagnostics for genetic diseases (Van Dijk et al., 2014; Mardis, 2013; Liu et al., 2012).

Taking this into account and despite some hurdles to be considered, in a near future NGS tools will provide us new applications in research fields and clinical diagnostics that would have been unthinkable some years ago.

1.4.1 - RNA-Sequencing

Transcriptome analysis provides information about all transcriptional activity in a cell or organism, and it has recently gained popularity and been applied to disease pathogenesis studies and identification of biomarkers (Wang et al., 2009).

Initially, the most commonly used technique in transcriptome analysis was microarrays. However this technique has several limitations, like reliance upon existing knowledge of gene sequences or high background levels (Wang et al., 2009). Due to this and to NGS evolution, RNA-sequencing is nowadays the dominant transcriptomics approach for gene expression analysis, identifying differentially expressed genes under different conditions and allowing new insights in various fields such as plant biology (Wang et al., 2009). Unlike microarrays, RNA-Seq does not need probes or reference sequences, produces low

background noise and can identify novel transcripts and splicing events, among other advantages. RNA-Seq revolutionized the scientific approaches in transcriptome analysis, offering a number of advantages compared to microarrays.

For non-model organisms like *P. pinaster*, for which there is no genome sequence available, RNA-Seq is an efficient means to generate functional genomic data (Parchman et al., 2010). Once RNA-Seq raw reads have been obtained, the first step of data analysis is to trim raw reads with low quality bases and adapters. Then, processed reads are assembled into contigs before aligning them to the genomic sequence to reveal transcription structure and finally predict candidate coding regions and annotate them against a database (Wang et al., 2009). An example of RNA-Seq data analysis workflow can be observed in figure 2.

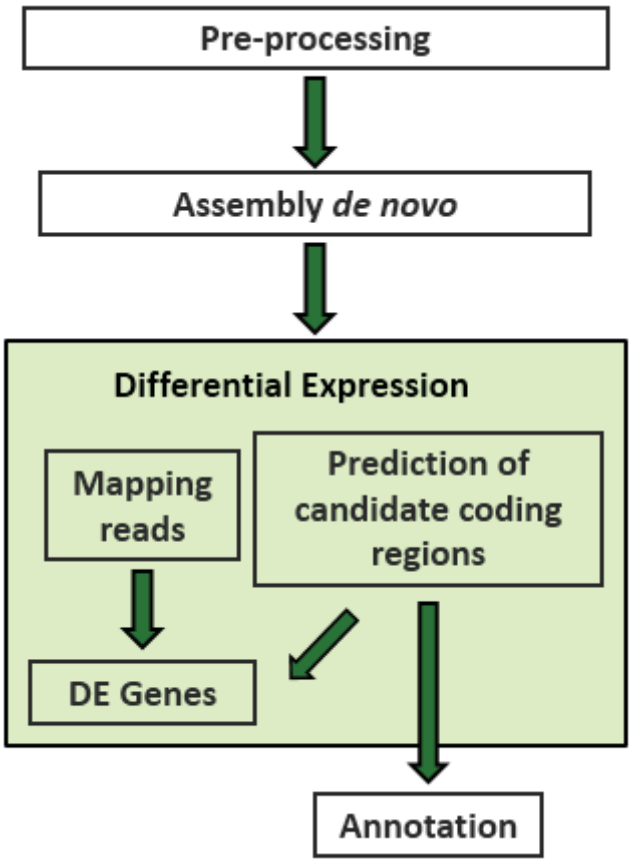

Figure 2 - Example of RNA-Seq data analysis workflow

1.5 - Bioinformatics tools for RNA-Seq data analysis

Next-generation sequencing of RNA libraries has become progressively used in a huge variety of transcriptomic studies. However, this information needs to be processed appropriately, thus, a set of freely bioinformatics tools has been developed for multiple genomic features analysis (Kalari et al., 2014). In this context, some bioinformatics tools commonly used in each stage of RNA-Seq analysis are presented below.

1.5.1 - Pre-processing data tools

Once raw reads have been produced by NGS systems, the first challenge of data analysis is to check the quality of the reads and trim adaptors sequences and low quality bases. For this step, a set of tools like FastQC software (Andrews, 2010) and PRINSEQ (Schmieder & Edwards, 2011) are frequently used, these tools apply a set of control tests on a raw sequence data and provides statistics which reports an overview of quality scores in our RNA-Seq data. FastQC tool outputs a set of graphics where potential problematic areas (low quality) are identified, and it also provides graphical information about GC content, N content, sequence length distribution and overrepresented sequences (Andrews, 2010). Based on FastQC results, it is necessary to trim low quality bases. For this step, is necessary to establish a threshold value for quality and for read length. A commonly software used for trimming low quality reads are Sickle (Joshi & Fass, 2011) and FASTX Toolkit[1]. They take an input file in fastq format and outputs a trimmed version with a specific thresholds defined by user. Briefly, Sickle uses a sliding windows along approach, either to trim the 3'-end or the 5'-end of reads, when quality is sufficiently low or high respectively. Moreover, Sickle also discard reads based upon the length threshold defined previously (Joshi & Fass, 2011).

[1] Website: http://hannonlab.cshl.edu/fastx_toolkit/

1.5.2 - De novo assembly tools

The genomes of a large number of organisms have been sequenced by NGS over the last years, but it still represents a small percentage of all known organisms. Working with non-model species without a reference genome is challenging because it is imperative to determine the transcript sequences from RNA-Seq data *de novo,* a process known as *de novo* transcriptome assembly. A set of software packages like TransAByss (Robertson et al., 2010), Velvet/Oases (Zerbino & Birney, 2008) (Schulz et al., 2012) or Trinity (Grabherr et al., 2011) has been developed to perform *de novo* assemblies. Most of the available assemblers implement algorithms based on de Bruijn graph (De Bruijn, 1946). In brief, in de Bruijn graph a node is defined by a substring of a fixed length k, denoted as k-mer, usually shorter than the read length. The nodes are connected by edges only if their overlap is exactly k-1 nucleotides. This representation enumerates all possible solutions by which linear sequences can be reconstructed given overlaps of k-1. However, the adaptation of de Bruijn graph to *de novo* assembly may have some issues, such as working with large amounts of data sets or providing robustness in cases of sequencing errors that can introduce false nodes (Grabherr et al., 2011).

One of the assemblers implementing the de Bruijn graph is the Trinity assembler (Grabherr et al., 2011). It is a widely used tool for de novo transcriptome assembly. Trinity includes three modules: Inchworm, that uses a greedy k-mer-based approach, assembles the RNA-Seq data into the unique sequences of transcripts; Chrysalis, which constructs a de Bruijn graph for each cluster of related contigs; Butterfly, which analyzes the paths taken by reads and reports all plausible transcript sequences (Grabherr et al., 2011).

1.5.3 - Mapping tools

Mapping, also called alignment of reads to a reference genome or transcriptome, is an essential step in NGS data analysis. This challenge consists in aligning a set of sequenced reads against a reference genome. Numerous tools have been developed to perform this process, including BWA (H. Li & Durbin, 2009) RapMap (Srivastava et al., 2015), Bowtie (Langmead et al., 2009) or SOAP (R. Li et al., 2008), among others. Due to the use of different algorithms, each tool provides different trade-off between speed and quality of the mapping (Hatem et al., 2013). Thus, algorithms must follow some assumptions, like aligning single reads across splice junctions *de novo,* or handle paired-end reads and run in a reasonable amount of time (Grant et al., 2011). To evaluate RNA-Seq alignments, a set of metrics needs to be checked. For example, each tool provides a score relative to mapping quality (MAPQ), the possibility of limiting the number of allowed mismatches or the gap length. The mapping process is a crucial step to perform differential expression analysis because the latter is performed over the unique mapped reads. From the large set of mapping tools referred before, RapMap is one of the most recent publicly available. This tool is based on the algorithm called quasi-mapping that uses a combination of data structures, a hash table, suffix array and efficient rank data structure, taking advantage of the transcriptome structure and providing read mapping information for each query that is useful for downstream analysis (Srivastava et al., 2015).

1.5.4 - Differential expression for RNA-Seq data analysis

High-throughput sequencing technologies led to a massive increase in transcriptomic data represented by counts. Analysis of such data is often concerned with detecting differential expression between different stages. The discovery of the differential expression data between different stages is, particularly but not exclusively, done by using biological replicates samples among each stage. Briefly, this type of analysis consists of normalizing the raw input counts and performing statistical tests to accept or reject

the null hypothesis of no differential expression between two or more groups of samples under different experimental conditions (Rashi Gupta, 2012; Soneson & Delorenzi, 2013). Several tools have been developed for inferring differential expression for RNA-Seq data, however, in this thesis we will focus only on EdgeR, an R package from Bioconductor.

1.5.4.1 - EdgeR

R is a programming language for statistical computing, providing a wide variety of statistical and graphical techniques (R Core Team, 2015). R can be extended via *packages*. The two biggest repositories for R packages are the CRAN (https://cran.r-project.org/) and Bioconductor (Huber et al., 2015), but only Bioconductor is relevant in this context.

Bioconductor is an open source, open development software project to provide tools for the analysis and comprehension of high-throughput genomic data, based primarily on the R programming language (Gentleman et al., 2004; Huber et al., 2015).

The EdgeR package provides a set of tools to identify differential gene expression in sequence count data from high-throughput sequencing technologies, allowing the analysis from different groups of data (Robinson et al.,2010). We can describe EdgeR's model as a statistical software based on the negative binomial distributions, which includes empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. The input data is summarized into a table of counts, with rows corresponding to genes and columns to samples. To assess differential expression, EdgeR uses an exact test analogous to Fisher's, but adapted to over dispersed data (Robinson et al., 2010).

1.5.4.2 - Prediction of candidate coding regions (TransDecoder)

Open reading frames (ORFs) are regions of nucleotide sequences between a start and a stop codon, and may indicate candidate protein coding regions in a DNA sequence. In

computational biology, identification of candidate coding regions represents a challenge in RNA-Seq studies conducted in species without a reference transcriptome. For this purpose, some software packages like TransDecoder (B. J. Haas et al., 2013) have been developed.

TransDecoder was integrated into the Trinity package, being useful for the identification of potential protein-coding regions within reconstructed transcripts generated by *de novo* assembly using Trinity. However, it can also be used as standalone tool. TransDecoder is executed in several steps. Initially, it processes a FASTA file containing transcript sequences and extracts the long ORFs. By default it considers as long ORFs the ones that are at least 100 amino acids in length. Additionally, an extra step to identify ORFs by homology via BlastP and/or Pfam against SwissProt and/or UniRef90 databases, respectively, can also be executed. To finalize the prediction, TransDecoder integrates the results obtained in the previous steps and outputs the final set of candidate coding regions (B. Haas, 2014).

1.5.5 - Transcriptome annotation

Transcriptome annotation provides information related to the function and biological process of assembled transcripts and the proteins they encode. The first step to perform transcriptome annotation involves *de novo* transcriptome assembly to infer transcripts from RNA-Seq data, or mappping reads onto a reference genome, when it is available. Then, annotation can be performed using one of the available tools implemented for this purpose, such as InterProScan (Jones et al., 2014; Quevillon et al., 2005).

InterProScan is one of the bioinformatics tools available for transcriptome annotation. This software searches protein sequences over non-redundant public domain databases, such as Pfam, Gene3D and Panther, providing information related to protein domains and important sites and classifying them into families (Jones et al., 2014).

Once the InterProScan results are obtained, it is possible to filter them and identify Gene Ontology (GO) terms or KEGG pathways. GO terms are used to describe gene function, classifying them into three categories: molecular function (MF), cellular component (CC) and biological process (BP) (Gene Ontology Consortium). There are several tools that analyse and organize GO terms data sets, including CateGOrizer (Na et al., 2014). This tool takes GOs IDs as input and performs a step-wise classification against a GO_slim database (Zhi-Liang, Jie, & James, 2008).

KEGG database (Kanehisa et al., 2015) is an integrated database which includes genomic, chemical and systemic functional information. Therefore, KEGG is widely used as a knowledge base for interpretation of large-scale datasets generated by high-throughput sequencing, being a reference resource for gene and protein annotation (Kanehisa et al., 2015).

1.6 - SNP calling

Single nucleotide polymorphisms (SNPs) are one of the most common type of genetic variation among individuals. They can be used as biological markers, helping in a set of research studies, which include the susceptibility and response to pathogens, such as maritime pine susceptibility to PWN.

Advances in NGS technologies provided new guidelines for identification of genetic variants such as SNP calling, but an accurate SNP calling can be difficult if NGS data suffer from high error rates or low-coverage (Nielsen et al., 2011). Moreover, assembly and alignment processes have a crucial role in a successful SNP detection (Nielsen et al., 2011).

The identification and filtering of SNPs from the raw data requires utilization of many processing steps and the application of a set of tools. Probably, the most widespread package for SNP calling is the genome analysis toolkit (GATK) (McKenna et al., 2010).

GATK package provides a wide variety of tools for variant discovery and genotyping, which include the Haplotype Caller and the UnifiedGenotyper, the tool used in this study. This tool uses a Bayesian genotype likelihood model to estimate the most likely genotypes and allele frequency for each sample in a population of N samples. UnifiedGenotyper generates an unfiltered, highly sensitive callset in variant call format (VCF). VCF is a text file format, containing meta-information lines about position and quality of each variant in genome. To filter the generated data, SelectVariants is a tool that has been widely used. It provides a new VCF file containing the selected subset of variants, following specific thresholds for quality defined by user. An useful tool for variant annotation is SnpEff (Cingolani et al., 2012). This software provides an annotation for variants and predicts the effects they produce on predicted genes.

2. - MATERIAL AND METHODS

This section describes the workflow applied in this study to perform the analysis of the RNA-Sequencing data produced to investigate the maritime pine response to infection with PWN. For this purpose, a set of maritime pine trees was inoculated with PWN and four sampling time points were established. RNA extraction was performed and four cDNA libraries were constructed. All libraries were sequenced as single-end reads on the Ion Proton platform. All specifications for these steps are presented in appendix. Data analysis workflow included the pre-processing of the raw reads and the *de novo* transcriptome assembly, mapping, differential gene expression analysis between all conditions, transcriptome annotation and finally the SNP calling.

2.1 - Pre-processing RNA-Sequencing data and assembly

The quality of the RNA-Seq reads from the four sequenced libraries was checked using FastQC software Version 0.11.3, a quality control tool for high throughput sequence data. Based on the FastQC results, a quality threshold of 12 and a read length of 80bp were defined. These parameters were used to run Sickle tool Version 1.33, trimming poor quality bases and adapters sequences from the raw data, which produced a set of processed reads to proceed with the RNA-Seq analysis. Discarding low quality bases from the raw data allows to reduce errors in subsequent procedures, therefore, pre-processing of raw reads is an important step, contributing to the reliability of the final results.

Due to the fact that there was no reference genome available for *P. pinaster*, it was necessary to perform a *de novo* transcriptome assembly. The processed reads from all libraries were assembled into contigs using Trinity 2.1.1 with the default parameters. In order to improve assembly by reducing gaps between contigs, clipping 5' and 3' low quality regions and obtain larger contigs, the CAP3 software (Huang & Madan, 1999) was

used. The resultant assembly was the basis for the next procedures, being used as the reference transcriptome assembly.

2.2 - Prediction of candidate coding regions

The sequences from the reference transcriptome were analyzed with TransDecoder-2.0.1 software to identify the open reading frames (ORF). This software is even it is able to predict ORFs by itself, allowing the improvement of such predictions performing homology searches. Thus, the ORF transcripts identified were further scanned for homology to known proteins against SWISS-PROT (Boeckmann et al., 2003) and Pfam (Finn et al., 2015) databases by running BlastP (Altschul et al., 1990) and HmmScan (Eddy, 1995), respectively. At the end, TransDecoder provides a final set of candidate coding regions, namely, predicted genes representing the basis for their annotation.

2.3 - Mapping and differential expression analysis

Mapping the reads against the transcriptome assembly was performed using RapMap, a new fast sensitive and accurate mapping tool. In brief, it consisted in building the index over the reference transcriptome, which was subsequently used along with a set of reads as input, to report the alignments in SAM format. This mapping output provided a report for each read (mapped or unmapped), which included the position of the mapped reads in the reference sequence, a quality mapping score, and was useful to infer gene expression information.

Before performing a differential gene expression analysis, it is imperative to determine the number of unique mapped reads, which was accomplished with SAMtools -1.3 (H. Li et al., 2009). SAMtools provides a set of utilities for manipulating alignments in the SAM format. The unique mapped reads were identified from the SAM files by using the flag

"NH:i:1", which is produced by RapMap and indicates solely the reads that mapped only once in the reference transcriptome.

The EdgeR package of Bioconductor was used to identify transcripts that were differentially expressed between the conditions. To adjust for library sizes and skewed expression of transcripts, the estimated abundance values were normalized using the Trimmed Mean of M-values normalization method (Robinson & Oshlack, 2010) included in the EdgeR package. As our experiment did not have biological replicates, it was necessary to determine the biological variability. Thus, in accordance with the EdgeR guidelines, a BCV (biological coefficient variation) of 0.1 was assigned (McCarthy, Chen, & Smyth, 2012). This procedure has been successfully used previously in other studies, for which biological replicates were also not available (Sebastiana et al., 2014). After the identification of the differentially expressed (DE) genes a multiplicity correction was performed by applying the Benjamini-Hochberg method (Yoav & Yosef, 1995) on the p-values, to control the false discovery rate (FDR). Finally, in order to obtain the most significant DE genes, the results were filtered using a FDR value ≤ 0.01.

2.4 - Transcriptome annotation

The ORFs transcripts identified by TransDecoder were used for transcriptome annotation. This procedure was performed using InterProScan. The protein domains, GO terms and KEGG pathways associated with the genes annotated which are encoding enzymes were identified. A python script was run to filter GO's and KEGGs from the InterProScan output. Categorizer was used for the analysis of the GOs. From a list of GOs IDs belonging to one of the GO category (BP, CC, MF), it classifies them by their corresponding subcategories against the GO Slim plant, counting the number of GOs within each subcategory, and reporting its percentage over the total set of GO IDs provided.

In relation to functional annotation for differential expressed genes, the contigs were annotated against the non-redundant NCBI plants database (version of August, 2015) using BlastX (e-value 1e-5).

2.5 SNP calling

Variant calling was performed with the GATK toolkit, which offers a variety of tools for variant discovery. Similarly to differential expression analysis, the unique mapped reads were used for SNP calling. The first step was to create a dictionary as a reference from the assembly, which was done using Picard tools (Broad institute, n.d.). Once the dictionary was created, the next step was to produce an unfiltered highly sensitive call set of variants in VCF format, using the "UnifiedGenotyper" tool available in the GATK toolkit. This initial set of variants was then filtered, using the "SelectVariants" option with the parameters SNP quality (QUAL >= 60), individual coverage (DP >= 25) and genotype quality (GQ - phred quality >= 40), in order to produce the final set of high-confidence SNPs. Finally, SnpEff was used to annotate and predict effects of the filtered SNPs.

3. - RESULTS

The key results of the RNA-Sequencing analyses from maritime pine response to infection with PWN are shown in this chapter. These results include the most relevant metrics and statistics obtained in each bioinformatics analysis step described previously.

3.1 - Pre-processing of RNA-sequencing data and assembly

A total of 176,282,168 raw reads were generated for all libraries. After checking quality control using FastQC, low-quality bases were trimmed by Sickle and 144,422,207 high quality reads were obtained with an average range length between 119bp and 122bp (table 1). A total of 81.9% of the original number of reads were retained after applying the quality control procedures.

The *de novo* assembly performed with Trinity 2.1.1 produced 483,428 contigs. Additional clustering of these contigs was performed with CAP3 (Huang & Madan, 1999), which resulted in an improved assembly comprising 355,287 contigs with a total length of 147,022,102 base pairs. Moreover, the largest contig had 7,285 base pairs. QUAST software (Gurevich et al, 2013) was used to obtain a set of different assembly metrics such as the N50, percentage of GC content, and the distribution of number of contigs above different length ranges. This allows a general view of the assembly status. Key results from the QUAST are presented in table 2.

Regarding gene prediction for the transcriptome assembly, we used TransDecoder software to identify protein coding regions within the unigenes, which yielded a total number of 83,468 predicted genes from the 355,287 assembled contigs.

Table 1 - Number of sequenced reads and its average read length for each library. Number and percentage of processed reads after control quality.

Sample	Number of sequenced reads	Average read length (bp)	Number of reads after QC	% reads after QC
Pp01 – Control	47,903,109	122	39,091,399	81.6
Pp02 – 6h+24h	38,483,969	119	30,863,177	80.2
Pp03 – 48h	44,943,925	122	37,186,370	82.7
Pp04 – 7 days	44,951,165	121	37,281,261	82.9
Total	176,282,168	121	144,422,207	81.9

Table 2: Key results from QUAST software

Metric	Value
Total number of contigs	355,287
Nº of contigs >=200 bp	355,287
Nº of contigs >=500 bp	66,262
Nº of contigs >=1000 bp	15,997
Nº of contigs >=2000 bp	2,583
Nº of contigs >=4000 bp	74
Nº of contigs >= 6000 bp	3
Total length of contigs	147,022,102 bp
Largest contig	7,285 bp
GC %	44.2%
N50	408

3.2 - Mapping and differential expression analysis

The mapped reads (MR) report for each library is presented in table 3. A total of 102,863,100 pre-processed reads were mapped by RapMap against the transcriptome assembly for all libraries, which corresponded to an average of 71.3% of the total number of pre-processed reads. The lowest percentage of mapped reads were obtained for the control library (Pp01) and for the last sampling time point Pp04, with values of 70.6% and 70.5%, respectively. On the other hand, for the Pp02 library, which corresponds to the 6h+24h sampling time points after inoculation, the highest percentage of MR (73%) was obtained (Table 3).

For all downstream analyses it was essential to filter the unique mapped reads (UMR) from this set of mapped reads. A total of 54,497,857 UMR were retained, which corresponded to approximately 37.8% of total processed reads (Table 3). Similarly to MR, the lower percentage of UMR was detected in Pp01 and Pp04 (36.9% and 37%, respectively), while the Pp02 library had the highest percentage of UMR (39.4%) (Table 3).

Table 3: Number of mapped reads, unique mapped reads and their percentages for each library

Sample	Number of reads mapped	Number of unique mapped reads	% of mapped reads	% of unique mapped reads
Pp01 – Control	27,578,068	14,439,253	70.55%	36.94%
Pp02 – 6h+24h	22,536,600	12,167,028	73.02%	39.42%
Pp03 – 48h	26,465,242	14,086,581	71.17%	37.88%
Pp04 – 7 days	26,283,190	13,804,995	70.50%	37.03%
Total	102,863,100	54,497,857	71.3%	37.8%

Statistical analysis in EdgeR software identified a total of 17,533 differentially expressed genes (DEG) (adjusted P-value ≤ 0.05 and FDR value= 0.01) within the 42,606 significant tests. The number of tests (up and down regulated) for each comparison between two different stages are summarized in table 4. The highest number of tests were identified between the control sample and the Pp02 where 4,969 genes were up regulated and 5,104 genes were down regulated. Moreover, 85 genes were always differentially expressed (up or down) in all comparisons.

Table 5 shows the number of genes differentially expressed (up and down) uniquely for each comparison. These results are in agreement with the total number of significant tests, since the highest number of DEG were present between Pp01 and Pp02 libraries.

Table 4: Total number of differentially expressed tests (up and down) between each comparison

	Pp01 vs Pp02	**Pp01 vs Pp03**	**Pp01 vs Pp04**	**Pp02 vs Pp03**	**Pp02 vs Pp04**	**Pp03 vs Pp04**
UP	4969	3354	3001	2874	4746	3235
DOWN	5104	3549	2637	2964	3957	2216

Table 5: Number of differentially expressed genes (up and down) uniquely for each comparison

	Pp01 vs Pp02	**Pp01 vs Pp03**	**Pp01 vs Pp04**	**Pp02 vs Pp03**	**Pp02 vs Pp04**	**Pp03 vs Pp04**
UP	630	384	264	362	716	312
DOWN	675	334	254	222	539	253

3.3 - Transcriptome annotation

Functional annotation over the 83,468 predicted genes by TransDecoder was performed using BlastP against the NCBI NR-plants database, with results showing a total of 70,646 annotated genes. However, 25,545 annotated genes had "Unknown" description, predominantly being associated to *Picea sitchensis*, a conifer of the *Pinaceae* family. From this set of annotated genes, the subset containing only the DE genes also contained 8,996 with an "Unknown" description or no description available. Also for the DE genes, most of the "Unknown" descriptions were related to *Picea sitchensis*.

We also carried out analysis about protein domains using InterProScan, which provided information related to the Gene Ontology annotations and KEGG pathways in the set of all predicted genes.

Gene Ontology (GO) analysis was performed by running queries against the CateGOrizer plant database, providing information related to three ontologies, which include biological process, cellular component and molecular function. First, a GO analysis was performed for all predicted genes, for which the results are shown in figures 3, 4 and 5. A total of 38,762 (46.4%) genes were associated with at least one GO term and a total of 1,810 different GO terms were found over the whole gene set.

With respect to the biological process branch, we found 1,737 hits assigned to 30 GO terms. The most significant were cellular process (GO:0009987) (32.8%), metabolic process (GO:0008152) (26.9%) and biosynthetic process (GO:0009058) (11.1%) (Figure 3). In the case of cellular component, 690 hits were assigned to 26 terms. The largest proportion GOs were assigned to cell (GO:0005623) (28.84%), intracellular (GO:0005622) (26.96%) and cytoplasm (GO:0005737) (11.45%) (Figure 4). In the molecular function category, we detected 1,437 hits corresponding to 24 GO terms. In this category, the most representative terms were catalytic activity (GO:0003824) (44.2%), transferase activity (GO:0016740) (13.9%) and hydrolase activity (GO:0016787) (11.7%) (Figure 5).

Subsequently, to further investigate the biological response associated to PWN infection, we performed a GO's analysis for DE genes between all conditions, the results are shown in figures 6, 7 and 8. In this analysis, we identified a total of 9,119 DE genes (52.0%) associated with at least one GO term and a total of 1,292 different GO terms were found. For the biological process term we identified 36 GO subcategories with a total of 1,477 hits. The most representative subcategories were cellular process (GO:0009987) (27.4%), metabolic process (GO:0008152) (22.55%) and biosynthetic process (GO:0009058) (9.1%) (Figure 6). Regarding cellular component terms, 24 subcategories were found with a total of 486 hits. The subcategories with more hits were cell (GO:0005623) (28.6%), intracellular (GO:0005622) (27.4%) and cytoplasm (GO:0005737) (12.4%) (Figure 7). Lastly, for molecular function term we identified 24 subcategories with a total of 1,039 hits. The most relevant subcategories were catalytic activity (GO:0003824) (42.9%), transferase activity (GO:0016740) (12.9%) and hydrolase activity (GO:0016787) (11.6%) (Figure 8).

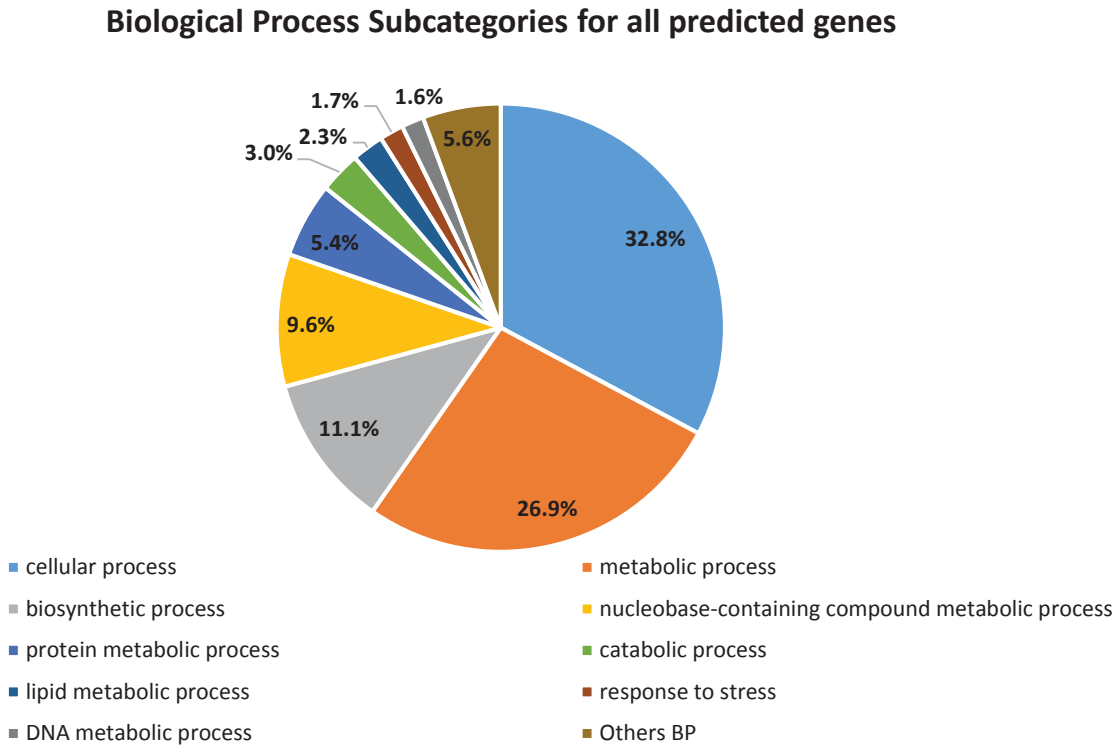

Figure 3 - Gene Ontology analysis of RNA-Seq data. Distribution of biological process subcategories for all predicted genes

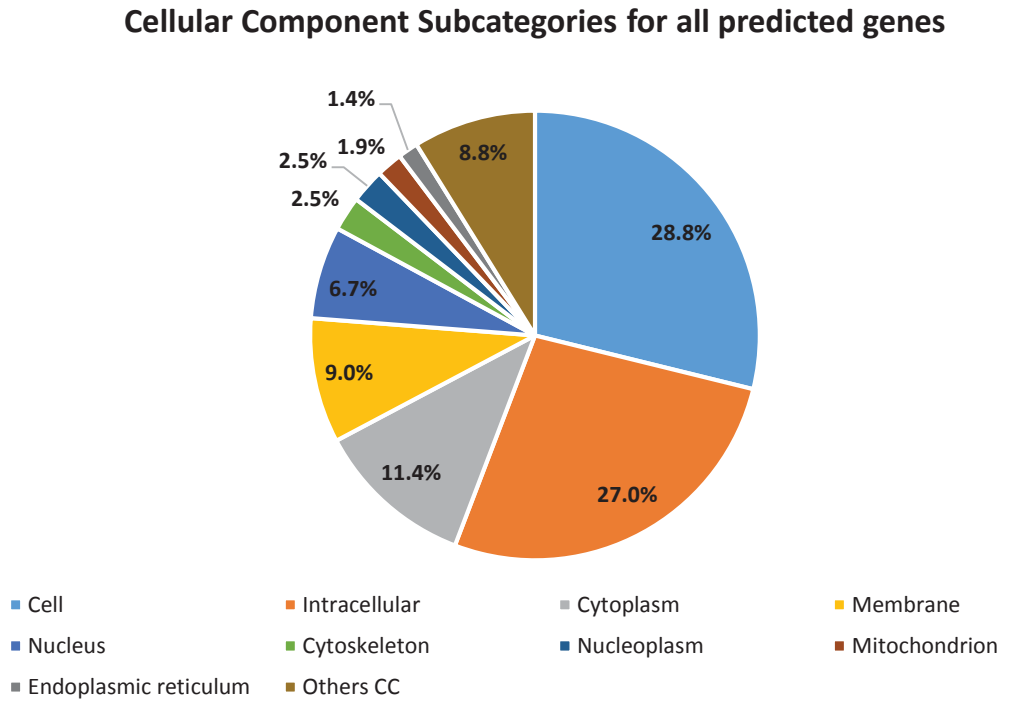

Figure 4 - Gene Ontology analysis of RNA-Seq data. Distribution of cellular component subcategories for all predicted genes

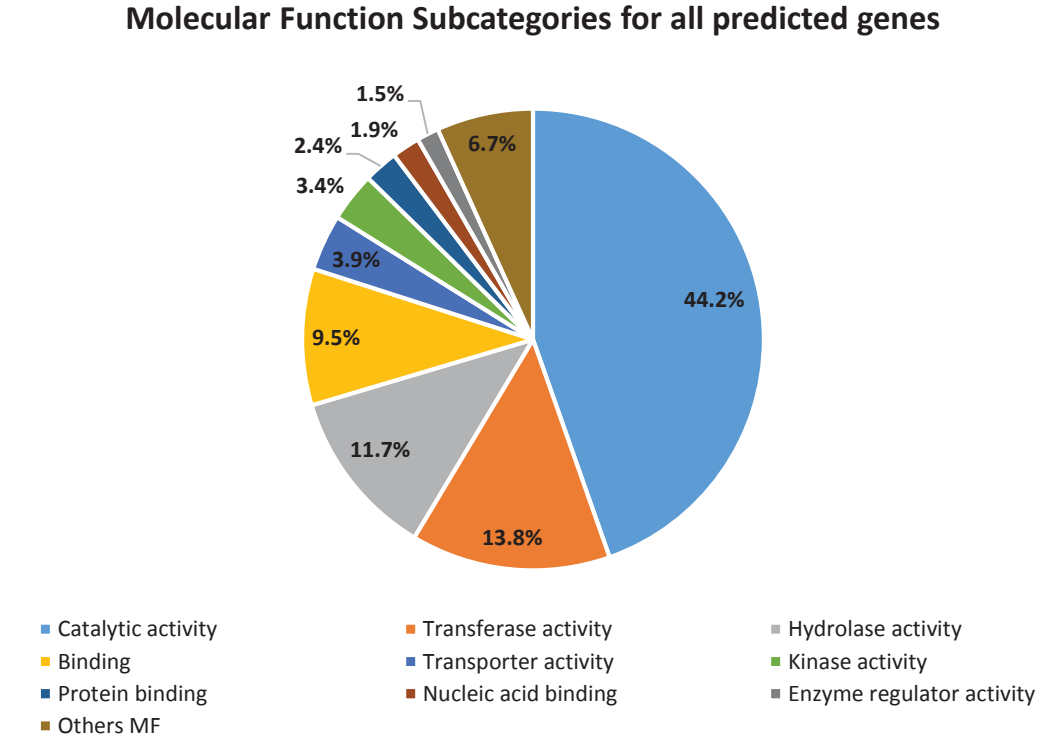

Figure 5 - Gene Ontology analysis of RNA-Seq data. Distribution of molecular function subcategories for all predicted genes

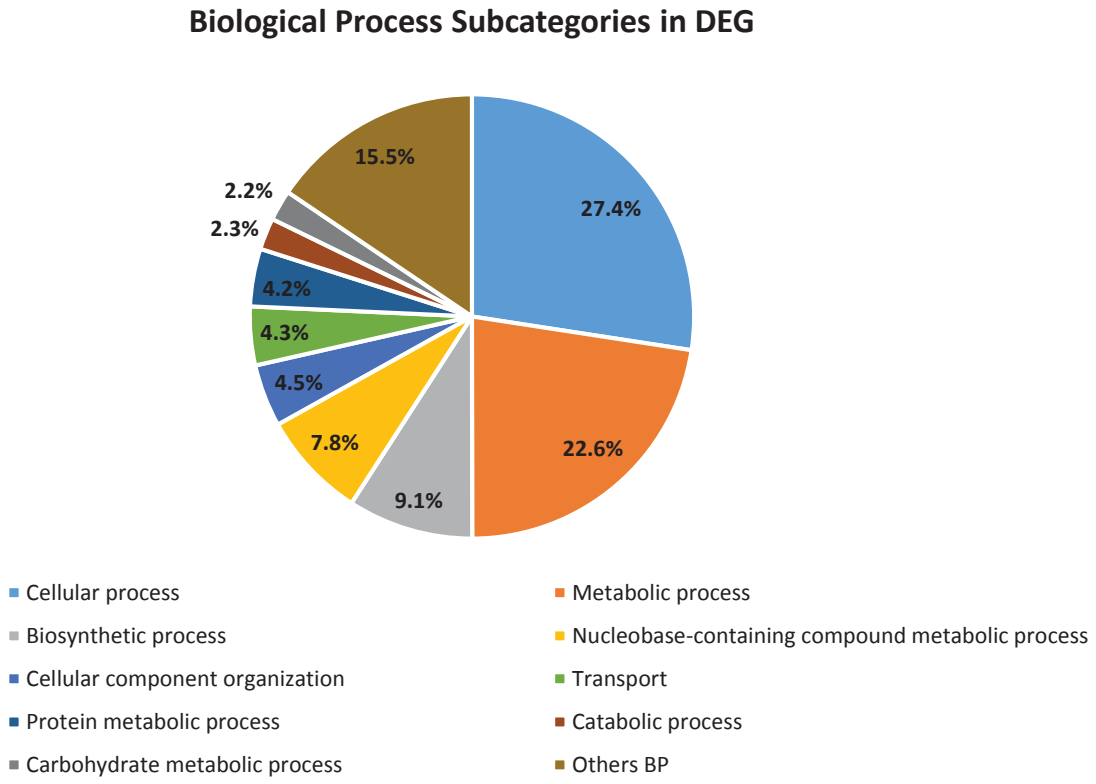

Figure 6 - Gene Ontology analysis of RNA-Seq data. Distribution of biological process subcategories in DEG

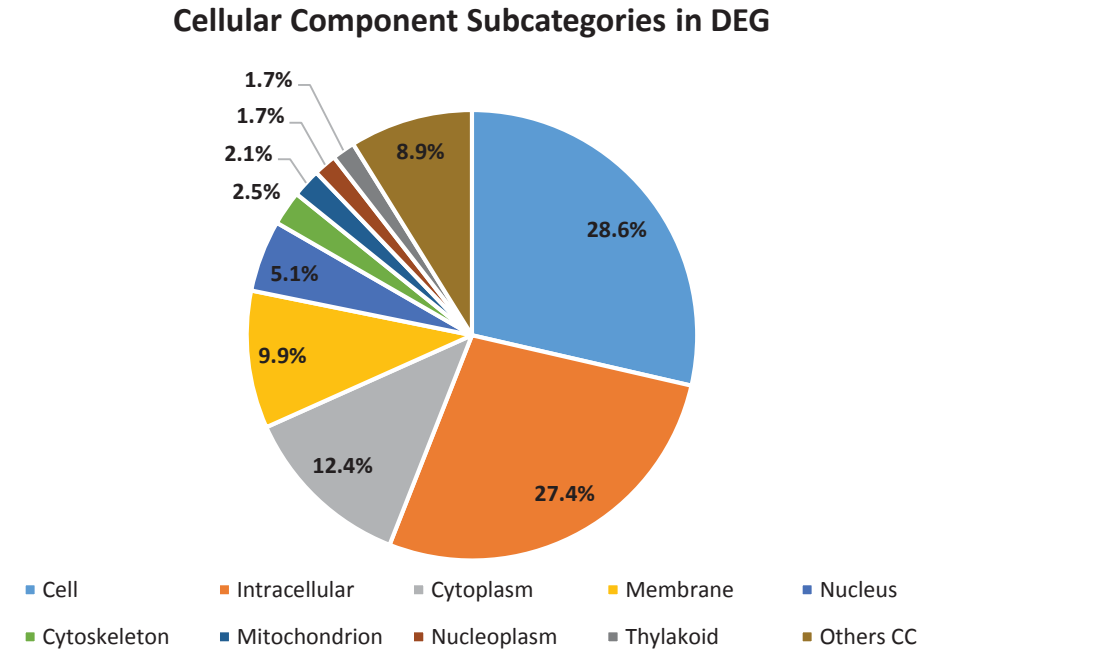

Figure 7 - Gene Ontology analysis of RNA-Seq data. Distribution of cellular component subcategories in DEG

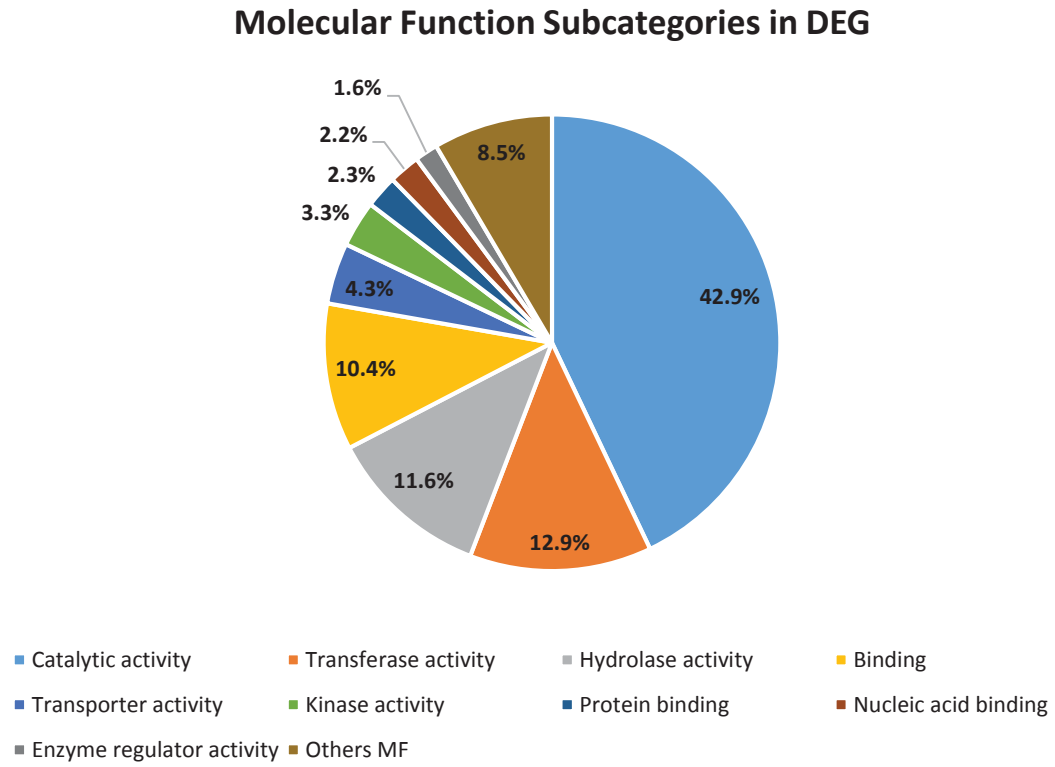

Figure 8 - Gene Ontology analysis of RNA-Seq data. Distribution of molecular function subcategories in DEG

Similarly to the GO analysis, we performed the KEGG pathways analysis for all predicted genes and for the DE genes. In the predicted genes set, we identified 4,904 genes associated with at least one KEGG pathway and a total of 111 KEGG pathways were found. KEGG analysis of DE genes between stages revealed that 1,154 were associated with at least one KEGG pathway and a total of 102 different KEGG pathways were found over this set of genes.

The ten most representative pathways for predicted genes and for DE genes with the number of enzymes associated are shown in table 6.

Table 6 – Summary of most representative KEGG pathways detected in predicted genes and in DEG

Pathways	Enzymes
Purine metabolism	35
Pyrimidine metabolism	26
Cysteine and methionine metabolism	20
Aminoacyl-tRNA biosynthesis	20
Starch and sucrose metabolism	19
Phenylalanine, tyrosine and tryptophan biosynthesis	18
Terpenoid backbone biosynthesis	17
Pyruvate metabolism	17
Porphyrin and chlorophyll metabolism	17
Glycolysis/ Gluconeogenesis	17

3.4 - SNP calling analysis

For SNP discovery and filtering, GATK was used with stringent parameters. Variants were called using the UnifiedGenotyper and further filtering was performed using the SelectVariants option. In total, 36,295 different SNPs were detected. Among these SNP's, 32.0% were found in exons, while 30.6% were detected in an intergenic region, a portion of DNA sequences located between genes (Table 7). Moreover, with respect to the SNPs found in each functional class, we identified 48.5% associated to missense mutations, 50.7% associated to silent mutations and less than 1% associated to nonsense mutations (Table 8).

Table 7 - SNP calling analysis. Number and percentage of effects by region

Region	Count	Percent
Exon	15,232	31.9%
Intergenic	14,600	30.6%
Splice site region	1	<0.1%
Transcript	31	0.1%
UTR 3 Prime	9,072	19.0%
UTR 5 Prime	8,718	18.3%

Table 8 - SNP calling analysis. Number and percentage of effects by functional class

Type	Count	Percent
MISSENSE	7,410	48,5%
NONSENSE	121	0,8%
SILENT	7,732	50,7%

Table 9 - SNP calling analysis. Number and percentage of effects by type

Type	Counts	Percent
3 prime UTR variant	9,072	19.0%
5 prime UTR premature start codon gain variant	1,245	2.6%
5 prime UTR variant	7,473	15.9%
Initiator codon variant	8	<0.1%
Intergenic region	14,600	30.6%
Missense variant	7,350	15.4%
Missense variant + splice region variant	1	<0.1%
Splice region variant	1	<0.1%
Start lost	23	<0.1%
Stop gained	121	0.3%
Stop lost	28	0.1%
Stop retained variant	14	<0.1%
Synonymous variant	7,718	16.2%

4. - DISCUSSION

In this study, we used an approach based in RNA-Sequencing technology to generate the transcriptome profile of maritime pine in different stages after inoculation with PWN, identifying candidate genes associated to resistance mechanism.

One of the main challenges in RNA-Seq studies for non-model organisms like maritime pine is the *de novo* transcriptome assembly. This is a crucial step, which can yield some undetected errors by the error-prone nature of high-throughput sequencing reads. The error rate of Ion Proton sequencing is between 1% and 3% affecting the accuracy of the de novo transcriptome assembly, since the de Bruijn graph can introduce false nodes, which may have important implications for gene prediction, differential expression analysis and SNP calling. This fact is evidenced in this study due to the low rate of predicted genes from the set of assembled contigs. Only 83,468 genes were predicted from 355,287 assembled contigs. These results can be explained, in part, either by sequencing errors or by assembly errors. In particular, the Ion Proton sequencing error rate is larger than other sequencing platforms, which increases the probability of the errors mentioned previously. Another relevant factor that contributes to the low rate of predicted genes is the unavailability of a reference genome for *P. pinaster*. In RNA-Seq approaches, the availability of reference genome is important because it provides a full description of genetic sequences and other useful biological knowledge stored in genome. In addition, with the usage of a genome reference it is easier to analyze and compare regions that could be less probably achieved with the *de novo* assembly.

When a reference genome is not available, the genetic description contained in the assembled transcripts can be successfully identified by homology only if the protein products have homologies in different protein databases, giving a set of predicted genes. From the total genes predicted in these study, 70,646 of them were annotated, providing a genomic resource to further deepen the study of candidate genes associated to pine

wood disease resistance. However, 25,545 annotated genes had “Unknown” description, mainly associated to *Picea sitchensis*. This high number of “Unknown” annotations can be explained again due to the unavailability of a reference genome for *P. pinaster*.

Despite the limitations mentioned above related with the RNA-Seq approaches, this study provides new advances in the comprehension of maritime pine resistance to PWN, by identifying a set of candidate genes potentially involved in defensive mechanism. However, additional studies are required to identify the real role of each gene in this complex defensive system.

Functional annotation with GO terms for predicted genes resulted in 38,762 (46,4%) unigenes with at least one assignment into one of the three categories of GO terms (BP,MF and CC). In one of the GO categories, the GO terms fell mainly into two or three subcategories. The GO subcategories identified with more evidences are in accordance with other reports (Santos et al., 2012), and may represent a typical gene expression profile for *P.pinaster* after infection with PWN.

Most plant defensive responses to pathogens have evolved into a complex system, simultaneously combining a number of mechanisms and pathways. To identify pathways involved in defense against PWN, we performed KEGG analysis for our set of predicted genes. The different KEGG pathways associated with the predicted genes are in agreement with Physiome Project Models for *P. pinaster* (http://nsr.bioeng.washington.edu/jsim/models/kegg/organism.html?eppi) except pyrimidine metabolism. The most prevalent pathways were purine and pyrimidine metabolism. These subunits of nucleic acids are major energy carriers and precursors for the synthesis of nucleotide cofactors such and NAD and SAM (Moffatt & Ashihara, 2002).

The comparison of sequence data from all libraries revealed a total of 17,533 DEG. Note that this high number of genes were obtained using a FDR value of 0.01. Usually this kind of studies make use of a FDR value equal to 0.05. Due to the huge number of DEG found

with that FDR value, we were forced to decrease it in order to reduce the complexity of the set DEG to analyze.

The highest number of DEG were identified in the comparison between control sample (Pp01) and the first time point (PP02 – 6h + 24h), suggesting an immediate response to PWN after inoculation. This observation is in accordance with previous results obtained in *Pinus thunbergii* Parl., that propose an early response to PWN in susceptible and in resistant trees (Shin et al., 2009). Within this early stage of response and comparing with the control sample, several genes potentially involved in the defensive response were detected. The "TMV resistance protein N-like" gene was down-regulated. This gene produces a resistance protein that guards the plant against pathogens, trigging a defense system, which restricts the pathogen growth (http://www.uniprot.org/uniprot/Q40392). We also highlighted "putative TIR-NBS-LRR protein" that belongs to disease resistance proteins family (http://www.uniprot.org/uniprot/Q9ZVX6). These proteins have been referenced as commonly involving in defensive mechanisms in various diseases. Several up-regulated genes for this comparison were also identified, including the "mildew resistance locus 6 calmodulin binding protein" gene, which triggers a response in the occurrence of an infection caused by a foreign body (http://www.uniprot.org/uniprot/B2KZI2). The processes used by the PWN to invade the *Pinus pinaster* tissues are likely to represent a very similar mechanism, hence this results provides further support for the involvement of the mildew resistance locus 6 calmodulin binding protein gene in the initial response of plants to infections with parasites or other agents. Also "sucrose synthase" was identified, an enzyme that provides the substrate for cellulose synthase, playing an important role in secondary cell wall synthesis (Nairn et al., 2008). The over expression of this enzyme as a response to infection, gives insights that not just proteins related to defensive mechanism are used to fight the infection. Thus, some mechanisms are activated to reconstruct the cell damage originated by the PWN.

In the (Shin et al., 2009) report it was also suggested that there is a late response in susceptible trees. This was observed in our data, indicating that this response may occur

approximately one week after inoculation, due to the large amount of DEG between Pp02 – 6h +24h and Pp04 – 7 days after inoculation identified. Measuring differences between early and late responses can elucidate the different mechanisms activated. As down-regulated genes between Pp02 and Pp04 we identified a "dehydrin 2 partial", which has been associated to plant response and adaptation to abiotic stress, such as water stress, being involved in a commonly mechanism developed in these stages (Hanin et al., 2011). This make sense, once the PWN attack the conducting vessels of the plant, affecting the water transportation, resulting in a water stress state. A "putative intracellular pathogenesis related type 10 protein" was identified as down-regulated. This protein was already found in conifers, displaying a transient accumulation in needles of drought-stressed trees (Dubos, 2001).As a consequence of the water stress, the needles became drought stressed, which is one of the most characteristic symptoms of PWD. As up-regulated between Pp02 and Pp04 a "heat shock protein 81-1-like" was found. Heat shock proteins, also known as stress proteins, are highly conserved among different organism. Under stressful conditions they protect cells by stabilizing unfolded proteins, giving the cell time to repair damage proteins (http://www.enzolifesciences.com/). It is unclear the precise role that this protein is playing in the *Pinus pinaster* response to the PWD. A"light harvesting complex a protein" was also found, which is involved in light energy transfer to one chlorophyll a molecule at the reaction center of a photosystem. This protein is not directly related with defensive mechanism, but it plays an important role, trying to maximize the production of energy, which could be essential in helping the resistance system. Furthermore, due to the high number of DEG among all conditions, a set of genes without expression in control sample (Pp01) and highly expressed in the others conditions (Pp02, Pp03, Pp04) were identified. This means that these genes were induced only after inoculation with PWN. Within those genes we highlighted "GDSL esterase/lipase At5g03610" which belongs to an important lipases gene family, where most of these contain a signal peptide, and are potentially involved in defensive reactions (Ling, 2008; Oh et al., 2005). The role of this proteins is to trigger systemic resistance signaling.

Moreover, we identified “translationally-controlled tumor protein homolog”, a highly conserved protein among many eukaryotic organisms that has been referenced as participant in important cellular processes like the protection of cells against various stress and apoptosis (Bommer & Thiele, 2004). Additionally, “jacalin-related lectin 3” protein was identified, which belongs to a subgroup of proteins often associated with biotic and abiotic stimuli. This subgroup of proteins has been referenced as a component of the plant defense system (Xiang et al., 2011). In this study, the identification of several DEG related to biotic and abiotic stresses further validates the hypothesis that these mechanisms may play a crucial role in the plant defense system.

Another interesting analysis is to monitor the evolution of defensive mechanism, thus, between Pp02-6h+24h and Pp03-48h, we identified as down-regulated “Cytochrome P720B1” that is involved in the biosynthesis of diterpene resin acids, a major component of the conifer oleoresin defense system (Geisler et al., 2016). It makes sense, once one of the main symptoms associated to PWN infection is the “tracheid cavitation” caused by destruction of cells surrounding the resin ducts. We also identified “auxin-induced protein 1”, auxins hormones regulate and control vital mechanisms, being involved in growth, development and in defense via signaling involving different interactions of molecules (Carna et al., 2014). This protein seems to have an important role in the first stage of the response against the infection. Finally, as down-regulated in this comparison, we also found “putative NBS-LRR protein G6207” that has been widely referenced in plants disease resistance mechanism (McHale et al., 2006). On the other hand, as up-regulated genes between Pp02 and Pp03, we identified a “laccase” protein. This kind of proteins are involved in lignin biosynthesis and plant pathogenesis (Christopher, Yao, & Ji, 2014). Lignin forms important structural materials in the support tissues of vascular plants. It make sense that one of the mechanisms activated is to reinforce the cell walls, especially in wood and bark.

Lastly, in Pp03 and Pp04 comparison, we highlighted “phospholipase D alpha 1-like” and “tau class glutathione S-transferase” being over expressed in Pp04. The first plays an

important role in various cellular processes, including response to stress (http://www.uniprot.org/uniprot/Q38882), while the second, has been associated to oxidative stress response mechanism (Kilili et al., 2004). One of the basal defense mechanism used by plants to combat pathogenic invasion is to generate oxidative stress, which has been already identified in the PWD as a response to the infection.

The SNP calling analysis performed in this study confirmed that the RNA-Seq approach is an efficient way to identify SNPs without complete sequencing of the whole genome. However, in our study, SNP calling was done over pools of sequenced individuals. This approach has the limitation of not allowing the determination of the genotypes for each individual. If the SNP calling was done without pools, this could permits to relate the expression profiles for each individual, because may be the SNPs identified in a gene could provoke the over or under expression of it. GATK package with stringent parameters yielded a total of 36,295 SNPs. In relation to the genomic regions where SNPs were identified, we not only identify SNPs in exons (31.9%), but also SNPs located in intergenic regions (30,6%), which have been recognized as playing important roles in gene regulation and disease response mechanism. Related with the effects by functional class, over than 50% has a silent effect, which means that SNP does not change the protein sequence. However, about 48.5% has a missense effect. In this situations, these changes are responsible for coding a different amino acids. When a new amino acid is coded, the sequence of the protein coded by a particular gene is also changed. These changes may occur between amino acids with markedly different properties, which in turn can affect the enzyme catalytic activity, or affect the secondary and tertiary structure of the protein, among others. Hence, these are very important SNPs. Moreover, about the 0.8% of the SNPs identified are nonsense, which provokes an unexpected stop codon truncating the protein function.

Additionally, we identified 4,061 SNPs over 17,533 DEG. From this set of genes, 1,452 have at least one SNP. These results could be promising to provide molecular markers for analyzing genome and identifying genomic regions that are expressed in different stages

of PWD resistance phenotype. It has been demonstrated that the use of molecular genetic markers to detect the presence of genetic loci controlling quantitative genetic variation, well known as quantitative trait loci (QTL), would seem to be particularly beneficial for improving disease resistance (Gibson & Bishop, 2005). Thus, the identification of QTLs would be useful for marker-assisted selection in PWN resistant breeding programs in *Pinus pinaster* because resistance tests are time consuming and laborious.

5. - CONCLUSIONS

Currently, PWD, caused by *Bursaphelenchus xylophilus*, is the most deadly maritime pine disease. Several studies have been performed but only a few of them were based in NGS data.

This study establishes a new approach for the understanding of the molecular response of maritime pine, which is susceptible to PWN, over different time points after inoculation with PWN. This was done using RNA-Seq data that is becoming widely used in resistance studies at transcriptome level.

The low rate of predicted genes from the set of assembled contigs and the high number of genes without annotation or with "Unknown" annotation, evidences the existing limitations when working in RNA-Seq studies with non-model species like *Pinus pinaster*. Despite these limitations, we were able to find some insights related with the defensive mechanism of *Pinus pinaster* against PWN.

The functional annotation of the predicted genes reveals the complexity of the system involved in the defensive mechanism against PWN, combining a number of mechanisms and pathways, simultaneously.

As pointed out in previous studies, the occurrence of two phases of response against PWN was identified from the results of the differential expression analysis: an early response which may occur immediately after infection, and a late response which may occur approximately seven days after infection (Shin et al., 2009). Additionally, we were able to get a set of candidate genes involved in response to PWD related to secondary metabolism, oxidative stress and defense against pathogen infection, among others. Some of those candidate genes highlighted in this study are “TMV resistance protein N-like”, “Putative TIR-NBS-LRR protein”, “Mildew resistance locus 6 calmodulin binding protein”, “Dehydrin 2 partial”, “Putative intracellular pathogenesis related type 10 protein”, “Heat shock protein 81-1-like”, “Light harvesting complex a protein”, “GDSL

esterase/lipase At5g03610", "Translationally-controlled tumor protein homolog", "Jacalin-related lectin 3", "Cytochrome P720B1", "Tracheid cavitation", "Auxin-induced protein 1", "Putative NBS-LRR protein G6207", "Phospholipase D alpha 1-like", "Tau class glutathione S-transferase".

Taking all these together, our results indicate that the workflow was successfully applied and it can be used as a guideline for similar studies with non-model species. Furthermore, the results provide new insights about the molecular mechanisms and metabolic pathways involved in resistance of *Pinus pinaster* against PWN infection.

The set of candidate genes identified over the different time points after inoculation may be a useful resource in future studies and for future breeding programs to select plants with lower susceptibility to PWD. Moreover the SNP calling results could be promising to provide molecular markers for identifying genomic regions that are expressed in different stages of PWD resistance phenotype. However, these markers need to be validated in large populations. Another future work opportunity based in this study is to identify gene clusters that share the same pattern of behavior through time after inoculation. Last but not least, it could be interesting to compare these results with the molecular response of a conifer species, which are referred as tolerant to PWN. In this sense, the workflow carried out in this study could be applied and adjusted to these non-model conifer species.

6. - REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. http://doi.org/10.1016/S0022-2836(05)80360-2

Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baermann, G. (1917). Ein einfache Methode zur Auffindung von Anklyostomum (Nematoden) Larven in Erdproben. *Ned Tijdschr Geneeskd*, *57*, 131–137.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., … Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*(1), 365–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12520024

Bommer, U. A., & Thiele, B. J. (2004). The translationally controlled tumour protein (TCTP). *International Journal of Biochemistry and Cell Biology*, *36*(3), 379–385. http://doi.org/10.1016/S1357-2725(03)00213-9

Broad institute. (n.d.). Picard. Retrieved from http://broadinstitute.github.io/picard/

Carna, M., Repka, V., Skupa, P., & Sturdik, E. (2014). Auxins in defense strategies. *Biologia*, *69*(10), 1255–1263. http://doi.org/10.2478/s11756-014-0431

Christopher, L. P., Yao, B., & Ji, Y. (2014). Lignin Biodegradation with Laccase-Mediator Systems. *Frontiers in Energy Research*, *2*, 12. http://doi.org/10.3389/fenrg.2014.00012

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. http://doi.org/10.4161/fly.19695

De Bruijn, N. G. (1946). A Combinatorial Problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, *49*, 758–764.

Dubos, C. (2001). Drought differentially affects expression of a PR-10 protein, in needles of maritime pine (Pinus pinaster Ait.) seedlings. *Journal of Experimental Botany*, *52*(358), 1143–1144. http://doi.org/10.1093/jexbot/52.358.1143

Eddy, S. R. (1995). Multiple alignment using hidden Markov models. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *3*, 114–20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7584426

Edwards, O. R., & Linit, M. J. (1992). Transmission of Bursaphelenchus xylophilus through Oviposition Wounds of Monochamm carolinensis (Coleoptera: Cerambycidae). *Journal of Nematology*, *24*(1), 133–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2619244&tool=pmcentrez&rendertype=abstract

Farjon, A. (2010). *A handbook of the world's conifers*. Leiden: Brill. Retrieved from http://www.worldcat.org/isbn/9789004177185

Fielding, N. J., & Evans, H. F. (1996). The pine wood nematode Bursaphelenchus xylophilus (Steiner and Buhrer) Nickle (= B . lignicolus Mamiya and Kiyohara): an assessment of the current position. *Forestry*, *69*(1), 35–46.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. http://doi.org/10.1093/nar/gkv1344

Fukuda, K. (1997). Physiological process of the symptom development and resistance mechanism in pine wilt disease. *Journal of Forest Research*, *2*(3), 171–181. http://doi.org/10.1007/BF02348216

Futai, K., Sutherland, J. R., & Takeuchi, Y. (2008). *Pine wilt disease*. Tokyo: Springer.

Geisler, K., Jensen, N. B., Yuen, M. M. S., Madilao, L., & Bohlmann, J. (2016). Modularity of Conifer Diterpene Resin Acid Biosynthesis: P450 Enzymes of Different CYP720B Clades Use Alternative Substrates and Converge on the Same Products. *Plant Physiology*, *171*(May), pp.00180.2016. http://doi.org/10.1104/pp.16.00180

Gibson, J. P., & Bishop, S. C. (2005). Use of molecular markers to enhance resistance of livestock to disease: a global approach. *Revue Scientifique et Technique (International Office of Epizootics)*, *24*(1), 343–53. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16110901

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–52. http://doi.org/10.1038/nbt.1883

Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., … Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, *27*(18), 2518–2528. http://doi.org/10.1093/bioinformatics/btr427

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. http://doi.org/10.1093/bioinformatics/btt086

Haas, B. (2014). TransDecoder (Find Coding Regions Within Transcripts). Retrieved May 16, 2016, from http://transdecoder.github.io

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–512. http://doi.org/10.1038/nprot.2013.084

Hanin, M., Brini, F., Ebel, C., Toda, Y., Takeda, S., & Masmoudi, K. (2011). Plant dehydrins and stress tolerance: versatile proteins for complex mechanisms. *Plant Signaling & Behavior*, *6*(10), 1503–9. http://doi.org/10.4161/psb.6.10.17088

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(1), 184. http://doi.org/10.1186/1471-2105-14-184

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*(9), 868–877. http://doi.org/10.1101/gr.9.9.868

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. http://doi.org/10.1038/nmeth.3252

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, *30*(9), 1236–40. http://doi.org/10.1093/bioinformatics/btu031

Joshi, N., & Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Retrieved from https://github.com/najoshi/sickle

Jusheng, H. (1985). A brief account of forest tree improvment in China. *Genetic Resources Information (FAO)*, *14*, 2–6.

Kalari, K. R., Nair, A. A., Bhavsar, J. D., O'Brien, D. R., Davila, J. I., Bockol, M. A., … Kocher, J.-P. A. (2014). MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. *BMC Bioinformatics*, *15*(1), 224. http://doi.org/10.1186/1471-2105-15-224

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–62. http://doi.org/10.1093/nar/gkv1070

Kilili, K. G., Atanassova, N., Vardanyan, A., Clatot, N., Al-Sabarna, K., Kanellopoulos, P. N., … Kampranis, S. C. (2004). Differential roles of tau class glutathione S-transferases in oxidative stress. *The Journal of Biological Chemistry*, *279*(23), 24540–51. http://doi.org/10.1074/jbc.M309882200

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. http://doi.org/10.1186/gb-2009-10-3-r25

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–60. http://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9. http://doi.org/10.1093/bioinformatics/btp352

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, *24*(5), 713–4. http://doi.org/10.1093/bioinformatics/btn025

Ling, H. (2008). Sequence analysis of GDSL lipase gene family in Arabidopsis thaliana. *Pakistan Journal of Biological Sciences*, *11*(5), 763–767. http://doi.org/10.3923/pjbs.2008.763.767

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, *2012*. http://doi.org/10.1155/2012/251364

Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. http://doi.org/10.1146/annurev-anchem-062012-092628

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–97. http://doi.org/10.1093/nar/gks042

McHale, L., Tan, X., Koehl, P., Michelmore, R. W., Jones, D., Jones, J., … Delarue, M. (2006). Plant NBS-LRR proteins: adaptable guards. *Genome Biology*, *7*(4), 212. http://doi.org/10.1186/gb-2006-7-4-212

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–303. http://doi.org/10.1101/gr.107524.110

Moffatt, B. A., & Ashihara, H. (2002). Purine and pyrimidine nucleotide synthesis and metabolism. *The Arabidopsis Book / American Society of Plant Biologists*, *1*, e0018. http://doi.org/10.1199/tab.0018

Mota, M., Braasch, H., Bravo, M. A., Penas, A. C., Burgermeister, W., Metge, K., & Sousa, E. (1999). First report of Bursaphelenchus xylophilus in Portugal and in Europe. *Nematology*, *1*(February 2016), 727–734. http://doi.org/10.1163/156854199508757

Na, D., Son, H., Gsponer, J., Huang, D., Sherman, B., Lempicki, R., … Liebman, M. (2014). Categorizer: a tool to categorize genes into user-defined biological groups based on semantic similarity. *BMC Genomics*, *15*(1), 1091. http://doi.org/10.1186/1471-2164-15-1091

Nairn, C. J., Lennon, D. M., Wood-Jones, A., Nairn, A. V, & Dean, J. F. D. (2008). Carbohydrate-related genes and cell wall biosynthesis in vascular tissues of loblolly pine (Pinus taeda). *Tree Physiology*, *28*(7), 1099–110. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18450574

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–51. http://doi.org/10.1038/nrg2986

Oh, I. S. et al. (2005). Secretome Analysis Reveals an Arabidopsis Lipase Involved in Defense against Alternaria brassicicola. *The Plant Cell*, *17*(10), 2832–2847. http://doi.org/10.1105/tpc.105.034819

Parchman, T. L., Geist, K. S., Grahnen, J. a, Benkman, C. W., & Buerkle, C. A. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, *11*, 180. http://doi.org/10.1186/1471-2164-11-180

Plomion, C., Pionneau, C., Brach, J., Costa, P., & Baillères, H. (2000). Compression wood-responsive proteins in developing xylem of maritime pine (Pinus pinaster ait.). *Plant Physiology*, *123*(3), 959–969. http://doi.org/10.1104/pp.123.3.959

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research*, *33*(Web Server issue), W116–20. http://doi.org/10.1093/nar/gki442

Rashi Gupta, I. D. B. A. B. (2012). Differential Expression Analysis for RNA-Seq Data. *ISRN Bioinformatics*, *2012*. http://doi.org/10.5402/2012/817508

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., … Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–12. http://doi.org/10.1038/nmeth.1517

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–40. http://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. http://doi.org/10.1186/gb-2010-11-3-r25

Santos, C. S., Pinheiro, M., Silva, A. I., Egas, C., & Vasconcelos, M. W. (2012). Searching for resistance genes to Bursaphelenchus xylophilus using high throughput screening. *BMC Genomics*, *13*, 599. http://doi.org/10.1186/1471-2164-13-599

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227–40. http://doi.org/10.1093/hmg/ddq416

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, *27*(6), 863–4. http://doi.org/10.1093/bioinformatics/btr026

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, *28*(8), 1086–92. http://doi.org/10.1093/bioinformatics/bts094

Sebastiana, M., Vieira, B., Lino-Neto, T., Monteiro, F., Figueiredo, A., Sousa, L., … Schmittgen, T. (2014). Oak Root Response to Ectomycorrhizal Symbiosis Establishment: RNA-Seq Derived Transcript Identification and Expression Profiling. *PLoS ONE*, *9*(5), e98376. http://doi.org/10.1371/journal.pone.0098376

Shin, H., Lee, H., Woo, K. S., Noh, E. W., Koo, Y. B., & Lee, K. J. (2009). Identification of genes upregulated by pinewood nematode inoculation in Japanese red pine. *Tree Physiology*, *29*(3), 411–421. http://doi.org/10.1093/treephys/tpn034

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*(1), 91. http://doi.org/10.1186/1471-2105-14-91

Sousa E, Bravo MA, Pires J, Naves P, Penas AC, Bonifácio L, M. M. (2001). Bursaphelenchus xylophilus (Nematoda: Aphelenchoididae) associated with Monochamus galloprovincialis (Coleoptera: Cerambycidae) in Portugal. *Nematology*, *3*, 89–91.

Srivastava, A., Sarkar, H., & Patro, R. (2015). RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-seq Reads to Transcriptomes. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2015/10/22/029652.abstract

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9). http://doi.org/10.1016/j.tig.2014.07.001

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. http://doi.org/10.1038/nrg2484

Xiang, Y., Song, M., Wei, Z., Tong, J., Zhang, L., Xiao, L., … Wang, Y. (2011). A jacalin-related lectin-like gene in wheat is a component of the plant defence system. *Journal of Experimental Botany*, *62*(15), 5471–83. http://doi.org/10.1093/jxb/err226

Yoav, B., & Yosef, H. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, *57*, 289–300.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–9. http://doi.org/10.1101/gr.074492.107

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, *38*(3), 95–109. http://doi.org/10.1016/j.jgg.2011.02.003.

Zhi-Liang, H., Jie, B., & James, M. R. (2008). CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. *Online Journal of Bioinformatics*, *9*(2), 108–112.

7. - APPENDIX

In this section are presented the methodologies used for PWN inoculation, sampling procedures, RNA extraction and cDNA synthesis. All this steps were done in INIAV I.P. – Instituto Nacional de Investigação Agrária e Veterinária. Moreover, methodologies for libraries preparation and sequencing are also showed. This procedures were carried out in Biocant. These tasks were not my responsibility.

7.1 - Biological Material, pine wood nematode inoculation and sampling

A total of seventeen potted 3-year old *Pinus pinaster* trees were used in this study. These plants were derived from seeds and maintained in natural environmental conditions during the assay. *Bursaphelenchus xylophilus* culture was grown in PDA (*Potato Dextrose Medium*) with *Botrytis cinerea*. After a significant growth, a suspension of nematodes was transferred to test tubes with 5ml of water and barley grains previously autoclaved. Later they were incubated for a week at 25ºC and relative humidity of 70%, (optimal conditions for nematodes growth). Before inoculation, nematodes were extracted from test tubes using the Baermann funnel technique (Baermann, 1917). Then, the culture was placed at 4ºC to stop multiplication and passing from juvenile stage to adult stage.

Inoculation with PWN was conducted following the method of Futai and Furuno (1979). Shortly, a suspension with 2,000 nematodes was pipetted into a small vertical wound (1cm) made on the upper part of the main pine stem with a sterile scalpel. A sterilized piece of gauze was placed around the wound site and fixed with parafilm to maintain the optimal humidity level. This procedure was done in fifteen *P. pinaster* plants, while the two remaining plants were used as control (inoculation with water).

Four sampling time points were established, including 6h, 24h, 48h and 7 days after inoculation. For each time point, a set of three *P. pinaster* plants were collected. Briefly, a small piece of stem tree above inoculation point was cut and flash frozen at -80ºC for further RNA extraction.

7.2 - RNA extraction, cDNA synthesis, library preparation and sequencing

All collected samples were ground in liquid nitrogen and a total RNA extraction was performed from 2g of plant material, according to an optimized method from Provost *et al*, (2007). Then, a DNase treatment was carried out following the instructions of the manufacturer (Kit TURBO DNA-free by life technologies).

An amount of approximately 1 microgram of total RNA was used for cDNA synthesis, following the ImProm-II™ Reverse Transcription System protocol kit (Promega). Before sequencing, four pools of cDNA were constructed (pool 1- control; pool 2-6+24h; pool 3-48h; pool 4- 7 days).

cDNA libraries were constructed with the Ion Total RNA-Seq Kit v2 (Life Technologies). Briefly, mRNA was fragmented with RNAse III. After short fragment removal, RNA adapters were ligated and the cDNA first and second strands synthesized. cDNA was then amplified with specific barcoded primers by PCR amplification and the resulting fragments selected for the correct size with magnetic beads.

Finally, the positive spheres from the four libraries were loaded into an Ion PI chip v2 and the transcriptomes were sequenced as single-end reads in the Ion Proton System (Life Technologies). All procedures were carried out according to manufacturer's instructions.