UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

RNA-Seq for the detection of differential expressed genes under several experimental conditions

Isabel Cristina Moniz Fernandes

Mestrado em Bioinformática e Biologia Computacional

Versão Provisória

Dissertação orientada por:
Professor Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo

2021

Agradecimentos

Em primeiro lugar, gostaria de agradecer ao meu orientador, Professor Doutor Octávio Paulo, pela mentoria deste projeto, pela tranquilidade e entusiasmo com que acompanhou o meu trabalho, pela confiança que depositou em mim e pelo constante reforço positivo que me motivou a dar sempre o meu melhor.

Quero agradecer também a parceria com a equipa do ISA que permitiu a execução deste projeto, nomeadamente o Doutor José C. Ramalho, a Doutora Ana I. Ribeiro-Barros e a Doutora Ana D. Caperta. Deixo ainda um agradecimento especial à Doutora Isabel Marques pelos conselhos, críticas e sugestões que me desafiaram a ser mais exigente comigo mesma e contribuíram para enriquecer esta tese e o meu rigor científico. Gostaria de agradecer também o financiamento parcial deste trabalho no âmbito do programa de investigação e inovação da União Europeia Horizon 2020, através do projeto BreedCAFS, e à Fundação para a Ciência e a Tecnologia, através do projeto PTDC/ASPAGR/31257/2017 e das unidades de investigação UIDB/00239/2020 (CEF) e UIDP/04035/2020 (GeoBioTec).

A toda a equipa $CoBiG_2$ que, mesmo em anos tão atípicos de isolamento e distanciamento, sempre me fez sentir acolhida. Obrigada por todas as partilhas científicas, técnicas, zoológicas, gastronómicas, bem-dispostas, sarcásticas, divertidas e, acima de tudo, pelo constante sentimento de entreajuda e disponibilidade. Vocês são os maiores!

A todos os meus amigos que, de perto ou de longe, me ajudaram a manter à tona. Um agradecimento especial para: a Carlota, que me acompanhou nesta jornada desde o primeiro dia e a quem devo tantos desabafos e gargalhadas (*sticky man*!); a Carina e a Bea, que me enchem sempre de mimos, admiração e orgulho; a Lála e a Rita, os meus maiores exemplos de inteligência, dedicação e perseverança, que são verdadeiras companheiras de vida; a Joana, a minha *partner in crime* desde sempre e para sempre; o *emCANTUS*, por trazer alegria à minha vida com a música que fazemos juntos; e os *AnimaVoce*, por, além de tudo o resto, serem uma segunda família.

Aos meus primos, Sílvia e Luís, por me abrirem sempre as portas de vossa casa e por todos os conselhos para ultrapassar este desafio. Aos meus cunhados, Manuel e Andreia, pelo carinho, pelas palhaçadas e por serem como irmãos. Aos meus pais, irmãos e sobrinhos, por me apoiarem incondicionalmente e por alinharem comigo de alma e coração em todas as aventuras em que me meto ("e não eram muitas, bastantes!"). E a ti, Nuno, por seres o meu maior pilar, pelas rugas que me vais dar de tanto rir, por seres o ombro onde chorar, pela leveza, pela alegria, pela compreensão, pelo amor e por seres o meu lar. Sem ti não faria sentido.

Para as minhas avós, Merita e Manuela.

Abstract

Bioinformatics aims to analyze and store complex biological datasets, which, due to the multidisciplinarity of the field, can be essential for finding meaning in biological systems, contributing to the modern life sciences knowledge. Transcriptomics is currently one of the areas if bioinformatics in greater expansion, namely through RNA sequencing (RNA-seq), which is an efficient transcriptome profiling approach. Its main application is the analysis of differentially expressed genes (DEGs), to assign biological meaning to specific tissues, environmental conditions, and other aspects. Reproductive strategies, resistance and stress responses can be evaluated through this technique, leading to a better understanding of the species fitness and survival.

This thesis intended to detect and functionally annotate DEGs through the application of RNA-seq pipelines. Moreover, since there's still no gold standard for its best practices, this work mostly aimed to find the best suited tools and methods for each data type, such as length, depth and replicates, according to the research goals. Furthermore, it established a better understanding of the different expression profiles of species from three different genera, namely *Casuarina*, *Coffea* and *Limonium*. In general, the RNA-seq workflow was performed as follows: quality analysis, assembly (for non-model species), alignment, quantification, differential expression, and functional annotation. Since this project was developed as four separated analyses, each step and respective tools were evaluated according to each dataset features. The results of these analyses break the path for further studies and integration with other omics, which can help unravel relevant mechanism and pathways of the studied species.

During the work of this thesis, a large set of scripts were developed to speed up and automatize the analysis, using Python and R languages, which have been made publicly available and can be applied by other users that work on similar studies.

Keywords: RNA-seq, *Casuarina*, *Coffea*, *Limonium*, transcriptomics.

Resumo

A bioinformática é uma área relevante e atual das ciências da vida que visa desenvolver ferramentas para analisar e armazenar conjuntos grandes e complexos de dados biológicos. Graças à sua multidisciplinaridade, englobando áreas como a biologia, a matemática, a estatística e as ciências computacionais, esta pode ser essencial na descoberta do significado de sistemas biológicos, contribuindo para o desenvolvimento do conhecimento moderno da biologia e da medicina. A transcriptómica é uma das áreas da bioinformática em maior expansão e desenvolvimento na atualidade, nomeadamente através das tecnologias de sequenciamento de RNA (RNA-seq). Este método caracteriza-se por ser uma abordagem eficiente na criação de perfis transcriptómicos, cuja principal aplicação é a análise de genes diferencialmente expressos (DEGs), com o objetivo de atribuir significado biológico a tecidos específicos, genótipos distintos, diferentes condições ambientais, e outras propriedades. As estratégias reprodutivas, os mecanismos de resistência e as respostas ao stresse de uma determinada espécie, sistema, órgão ou tecido, podem ser avaliadas através desta técnica, através da pesquisa de diferenças significativas entre o objeto de estudo e o seu controlo, conduzindo potencialmente a um melhor conhecimento acerca da adaptação e sobrevivência das espécies.

O trabalho desenvolvido nesta tese pretendeu detetar e anotar funcionalmente os genes com expressão significativamente diferencial, recorrendo à aplicação de *pipelines* de RNA-seq. Adicionalmente, uma vez que ainda não existe um *gold standard* para as melhores práticas a aplicar na utilização deste método, este projeto teve ainda como objetivo principal a seleção das ferramentas mais adequadas para cada tipo de dados, de acordo com os objetivos da investigação. Uma vez que o tipo de ferramentas e parâmetros usados têm uma grande influência sobre o sucesso da análise, os critérios de escolha da abordagem a utilizar reveste-se de grande importância numa boa análise bioinformática. Não obstante, os métodos de preparação das bibliotecas e processos laboratoriais que a envolvem podem influenciar os resultados, pelo que a sua otimização contribui também para a qualidade dos mesmos. Concretamente, este trabalho permitiu conhecer e compreender melhor os diferentes perfis de expressão de espécies de três géneros distintos, nomeadamente *Casuarina*, *Coffea* e *Limonium*, tendo em conta as condições de crescimento a que foram submetidos antes da colheita das amostras. Na primeiro análise, pretendeu-se estudar a expressão dos genes envolvidos na resistência à salinidade dos solos em *Casuarina glauca* para promover a proteção dos biomas nativos nos locais onde estas são invasoras. No segundo caso, na tentativa de estudar o impacto das alterações climáticas nas produções de café, analisou-se o efeito da elevação de [CO_2] nos perfis de expressão de *Coffea canephora* e *Coffea arabica*, e ainda o efeito combinado da elevação de [CO_2] e da temperatura. Finalmente, pretendeu-se estudar quais os mecanismos genéticos responsáveis pela substituição das estratégias de reprodução sexuadas pelas apomíticas, obrigatórias e facultativas, em diversas espécies de *Limonium*.

Em geral, o *workflow* de RNA-seq foi aplicado da seguinte forma: controlo de qualidade com corte e filtragem das *reads*; assemblagem *de novo* do transcriptoma, para as espécies não modelo; alinhamento das *reads* com o transcriptoma assemblado ou o genoma de referência, nos casos em que este está disponível; quantificação da expressão dos genes em cada amostra; análise estatística para deteção de DEGs; e anotação funcional dos DEGs identificados para identificação dos processos biológicos mais relevantes. Uma vez que este projeto foi desenvolvido em quatro análises separadas, cada etapa da *pipeline* e as suas respetivas ferramentas e métodos foram avaliados de acordo com as características de cada conjunto de dados, tendo em consideração especificidades como tipo, tamanho e profundidade das bibliotecas de RNA-seq, e o número de amostras e seus replicados. O controlo de qualidade foi maioritariamente conseguido através de filtragem e corte das *reads* de baixa qualidade, pequeno tamanho ou presença de contaminantes, nomeadamente adaptadores, usando os *softwares* FASTQC e Trimmomatic. A assemblagem *de novo* dos organismos não modelo, no caso *Casuarina glauca* e *Limonium* spp., foi realizada com o programa Trinity e permitiu não apenas a deteção e anotação funcional de DEGs, como a criação de transcriptomas que poderão ser usados em análises futuras. Uma das vantagens deste *software* é a integração com ferramentas importantes para a análise subsequente, nomeadamente para o alinhamento das *reads* e posterior quantificação dos perfis de expressão génica, as quais foram executadas com as ferramentas Bowtie2 e RSEM, respetivamente. No caso das amostras de café, para as quais existe um genoma de referência, o alinhamento das *reads* foi conseguido através da aplicação do *software* STAR, tendo sido os perfis de expressão genética traçados com a utilização do HTSeq. A análise de expressão diferencial propriamente dita foi determinada pela utilização individual ou combinada de um conjunto de pacotes de *R* especialmente adequados para o tipo de dados em questão, nomeadamente DESeq/DESeq2, edgeR. NOISeq. Estas ferramentas integram a possibilidade de normalização das contagens utilizando métodos otimizados para expressão diferencial, que têm em consideração fatores como os valores de contagem das *reads*, a profundidade de sequenciação e a composição de RNA, o que permite minimizar o viés intrinsecamente ligado ao tamanho relativo dos transcriptomas. Posteriormente, a anotação funcional e restantes análises foram realizadas de acordo com os objetivos específicos de cada investigação, recorrendo sempre a bases de dados como Gene Ontology, UniProtKB e KEGG. De modo geral, as ferramentas mais referenciadas na literatura estão otimizadas para DNA e apesar de já se terem desenvolvido algumas ferramentas especificamente focadas em RNA, ainda são necessários estudos de revisão para comparar a qualidades dos seus resultados com os métodos de referência, a fim de assegurar a manutenção ou melhoria da qualidade das análises. Os resultados deste trabalho permitiram a coautoria e publicação de dois artigos científicos sobre as estratégias de resistência ao stress ambiental das duas espécies de café estudadas, e a primeira autoria de dois artigos em vias de publicação sobre os mecanismos já referidos em *Casuarina glauca* e *Limonium* spp.

No decorrer do trabalho desenvolvido nesta tese, além da investigação para a aplicação das ferramentas mais adequadas e atuais de acordo com o estado da arte, foi ainda criado um conjunto de

scripts com a finalidade de agilizar e automatizar as análises, utilizando as linguagens *Python* e *R*. Na sua globalidade, estes *scripts* têm a intenção de acelerar os processos mais morosos e, sobretudo, facilitar a aplicação recorrente de tarefas nas análises a diferentes conjuntos de dados, representado um investimento de tempo para análises futuras. Entre as referidas tarefas encontram-se a aplicação automatizada de ferramentas de controlo de qualidade, assemblagem, alinhamento, expressão diferencial e anotação funcional a múltiplos dados, a recolha automatizada de informações em bases de dados recorrendo a *APIs*, o rápido processamento de resultados de expressão diferencial e sua consequente graficagem, a manipulação de genomas de referência para uma facilitada anotação funcional e a automatização da análise de enriquecimento. Os *scripts* estão disponibilizados publicamente e poderão ser aplicados ou adaptados por utilizadores que trabalhem com dados e objetivos semelhantes.

De modo global, os resultados deste projeto permitiram abrir caminho para futuros estudos e, especialmente, para a integração com outras ómicas, que poderão ajudar a desvendar os mecanismos e vias metabólicas mais relevantes das espécies estudadas, ou espécies relacionadas, tendo em conta os objetivos da investigação. Considerando a célebre evolução das tecnologias de sequenciação e dos *softwares* de análise, prevê-se que durante a próxima década o conhecimento acerca dos mecanismos genéticos que permitem a funcionalidade das células e seus os processos biológicos mais relevantes tenha um crescimento exponencial, permitindo um desenvolvimento sem precedentes da área das ciências da vida.

Palavras-chave: RNA-seq, *Casuarina*, *Coffea*, *Limonium*, transcriptómica.

Contributions

This thesis led to the co-authorship of two original articles and contributed largely to the first authorship of two papers, soon to be published:

- Marques, I., Fernandes, I., David, P., Paulo, O. S., Goulao, L. F., Fortunato, A. S., Lidon, F. C., DaMatta, F. M., Ramalho, J. C., & Ribeiro-Barros, A. I. (2020). Transcriptomic Leaf Profiling Reveals Differential Responses of the Two Most Traded Coffee Species to Elevated [CO2]. *International journal of molecular sciences*, *21*(23), 9211. https://doi.org/10.3390/ijms21239211
- Marques, I., Fernandes, I., Paulo, O. S., Lidon, F. C., DaMatta, F. M., Ramalho, J. C., & Ribeiro-Barros, A. I. (2021). A Transcriptomic Approach to Understanding the Combined Impacts of Supra-Optimal Temperatures and CO_2 Revealed Different Responses in the Polyploid and Its Diploid Progenitor *C. canephora*. *International journal of molecular sciences*, *22*(6), 3125. https://doi.org/10.3390/ijms22063125
- Fernandes, I., Sarjkar, I., Sem, A., Graça, I., Marques, I., Lidon, F. J., Pawlowski, K., Paulo, O. S., Ramalho, J. C. & Ribeiro-Barros, A. I. Salt stress tolerance in *Casuarina glauca*: insights from branchlets transcriptome analysis. (Provisory title)
- Fernandes, I., Conceição, S. I. R., Marques, I., Róis, A. S., Paulo, O. S., Caperta, A. D. Transcriptomic analysis provides new insights into the mechanisms of sexual and apomictic regulation in *Limonium spp.* (Provisory title)

Moreover, this project made possible the availability and accessibility of four raw datasets in the largest publicly available repository of high throughput sequencing data, Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI), under the accession BioProjects: PRJNA706159 (available online at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA706159), PRJNA606444 (available online at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA606444), PRJNA630692 (available online at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA630692) and PRJNA752506 (available online at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA752506). Also, a set of scripts which were developed during this work are publicly available at https://github.com/ziisabel/CoBiG2/tree/cobig2.

The work developed in this thesis was conducted in the Computational Biology and Population Genomics ($CoBiG^2$) research group from the Centre for Ecology, Evolution and Environmental Changes (CE3C) of Faculdade de Ciências da Universidade de Lisboa (FCUL), in collaboration with Instituto Superior de Agronomia (ISA), through:

- Dr. José C. Ramalho, Dr. Ana I. Ribeiro-Barros and Dr. Isabel Marques from the Plant-Environment Interactions and Biodiversity Lab (PlantStress & Biodiversity) and Forest Research Centre (CEF) departments, who provided the:
  - *Casuarina glauca* datasets, under the project PTDC/AGR-FOR/4218/2012

    - *Coffea canephora* and *Coffea arabica* datasets, under the project PTDC/ASPAGR/31257/2017
  - Dr. Ana D. Caperta from the Linking Landscape, Environment, Agriculture and Food (LEAF), who provided the *Limonium* datasets, under the project PTDC/AGRPRO/4285/2014.

The development of this thesis was made possible by the funding support from the European Union's Horizon 2020 research and innovation program (grant agreement no. 727934, project BreedCAFS), and from national funds from Fundação para a Ciência e a Tecnologia, I.P. (FCT), Portugal, through the project PTDC/ASPAGR/31257/2017 (CoffeeOmicsClimate), through which I received an 18-month research grant, and the research units UIDB/00239/2020 (CEF) and UIDP/04035/2020 (GeoBioTec).

Table of Contents

List of Figures

List of Tables

Acronyms and Initialisms

DNA – Deoxyribonucleic Acid

cDNA – Complementary Deoxyribonucleic Acid

RNA – Ribonucleic Acid

mRNA – Messenger Ribonucleic Acid

RNA-seq – Ribonucleic Acid Sequencing

ChIP-seq – Chromatin Immunoprecipitation Sequencing

Hi-C – High Throughput Chromosome Conformation Capture

SNP – Single-nucleotide Polymorphism

PCR – Polymerase Chain Reaction

GC – Guanine-cytosine

HTS – High-throughput Sequencing

SRA – Sequence Read Archive

NCBI – National Center for Biotechnology Information

UniProtKB – UniProt Knowledgebase

DEG – Differentially Expressed Gene

SE – Single-end

PE – Paired-end

GO – Gene Ontology

QC – Quality Control

DBG – *De Bruijn graphs*

BUSCO – Benchmarking Universal Single-Copy Orthologs

STAR – Spliced Transcripts Alignment to a Reference

RSEM – RNA-seq by Expectation-Maximization

FC – Fold change

CPM – Counts Per Million

RPKM – Reads Per Kilobase Million

FPKM – Fragments Per Kilobase Million

TPM – Transcripts Per Kilobase Million

TMM – Trimmed Mean of Means

RLE – Relative Log Expression

MRN – Median Ratio Normalization

LRT – Likelihood Ratio Test

FDR – False Discovery Rate

FNR – False Negative Rate

FPR – False Positive Rate

API – Application Programming Interfaces

PMC – PubMed Central

NIH/NLM – National Institute of Health's/National Library of Medicine

KEGG – Kyoto Encyclopedia of Genes and Genomes

STRING – Search Tool for the Retrieval of Interacting Genes/Proteins

PPI – Protein-Protein Interaction

BLAST – Basic Local Alignment Search Tool

ORA - Over-Represented Analysis

GSEA – Gene Set Enrichment Analysis

WebGestalt – WEB-based GEne SeT AnaLysis Toolkit

GGI – Gene-Gene Interaction

NOD^+ - *Casuarina glauca* plants nodulated by *Frankia* strain Thr

KNO_3^+ – *Casuarina glauca* plants supplemented with mineral nitrogen

PCA – Principal Component Analysis

PcoA – Principal Coordinate Analysis

CC – Cellular Component

MF – Molecular Function

BP – Biological Process

aCO_2 – Ambiente CO_2

eCO_2 – Elevated CO_2

FAD - Flavin Adenine Dinucleotide

LOX – Lipoxygenase

BH – Benjamini & Hochberg

REVIGO – Reduce and Visualize Gene Ontology

iTAK – Plant Transcription factor & Protein Kinase Identifier and Classifier

TF – Transcription Factor

KO – Knocked Out

ORF – Open Reading Frame

Icatu – *Coffea arabica* cv. Icatu

CL153 – *Coffea canephora* cv. Conilon Clone 153

RuBisCO – Ribulose-1,5-bisphosphate carboxylase-oxygenase

Chapter 1

1. Introduction

Bioinformatics is a multidisciplinary field that combines biology, computational science, mathematics, statistics, and information engineering (Moore, 2007). Its main goal is to develop methods and software tools to analyze, interpret and find meaning for large and complex sets of biological data. Acting as the linguistics of genetics, it comprises the compilation, storage, retrieval, manipulation and modelling of data, through the development of algorithms and software. Specifically, bioinformatics can deal with the analysis of sequences of biological molecules, namely nucleotides (*i.e.,* DNA and RNA) and amino acids, being particularly useful to compare different sequences within an organism or between organisms (Austin, 2014). Using *in silico* analyses of biological queries, it looks for patterns to explore and predict molecular functions, study evolutionary relationships, discover stress response mechanisms and unravel metabolic pathways. It is also used to develop databases that store relevant biological information, which can either be used to achieve new findings and to guarantee the reproducibility of work (Kulkarni et al., 2018). Data obtained by genome or transcriptome[1] sequencing can be used to gain insight into biological processes and phylogenetic relationships, complementing proteome[2] and metabolome[3] research (Ulfenborg, 2019). Therefore, bioinformatics is the key to making the most out of these studies as an invaluable tool for attaining knowledge and import meaning to living systems, being essential in modern biology and medicine.

1.1. RNA Sequencing

The individuality of an organism is partially defined by its gene expression profile and its analysis, also known as transcriptomics, is one of the most commonly conducted procedures in molecular biology research (Chatterjee et al., 2018). Ribonucleic acid sequencing (RNA-seq) has revolutionized transcriptomics, since it offers remarkable opportunities to life sciences by shaping and refining knowledge to better comprehend cellular mechanisms (Ari, 2016). For many years, Sanger sequencing (Sanger) was the most widely used sequencing method and, although it's still in use to

[1] Set of all RNA transcripts, including coding and non-coding, in an individual or a population of cells.
[2] Entire set of proteins that can be expressed by a genome, cell, tissue, or organism.
[3] Complete set of small-molecule chemicals found within a biological sample.

validate results and to cover areas not amendable to newer methods, it has been largely replaced by deep-sequencing technologies, especially for large-scale automated genome analyses (Shendure et al., 2017).

RNA sequencing is a major quantitative transcriptome profiling system, which enables the whole transcriptome to be surveyed in a remarkably high-throughput manner (Wang et al., 2009). Through the sequencing of complementary DNA (cDNA), it can produce millions of short nucleotide sequences (30-400 nucleotides in size) also known as reads. Due to its efficiency, it became one of the most chosen approaches to study gene expression profiles, aiming to assign biological meaning to the differences found between different development stages, tissues, genotypes, physiological or environmental conditions (Kukurba & Montgomery, 2015). The interpretation of these changes can contribute to the understanding of development, physiology, stress tolerance, chemical signals, metabolic regulation, diseases and other key aspects that are essential to species fitness and survival. RNA-seq is particularly helpful in the analyses of poorly characterized species, since it doesn't require prior knowledge of the genome or genomic features under investigation, nor a reference genome to attain useful transcriptomic information (Strickler et al., 2012). Also, it has very low background signal, which can be entirely absent, since DNA sequences can be unequivocally mapped to unique regions of the genome (Wang et al., 2009). Therefore, the development of this method represents a good opportunity to analyze not only the expression of protein-coding regions, but also noncoding RNA and *de novo* transcriptome assembly of new species or organisms (Chatterjee et al., 2018).

Sequence-based methods such as RNA-seq directly determine the cDNA sequence, both mapping and quantifying transcriptomes, without the need of predefined transcripts/genes, and thus allowing the detection of novel transcripts and full sequencing of the whole transcriptome (Wang et al., 2009). However, like any other technology, RNA-seq also faces some challenges. Among its disadvantages, it lacks optimized and standardized analysis protocols and involves working with considerably larger files, requiring a more extensive and complex bioinformatics analysis, with longer analysis times and expensive computation infrastructures (Rao et al., 2019). However, there are multiple computational tools already available and under development, gradually improving these limitations. Also, since RNA-seq relies on reverse transcription and PCR amplification before sequencing, it can consequently induce some types of biases, including random hexamer priming bias, GC content bias and depletion of 3' and 5' ends of the transcripts, which impacts read nucleotide content and annotation, biasing the quantification of gene expression (Hansen et al., 2010; Roberts et al., 2011). Moreover, since RNA-seq involve cDNA synthesis, which requires several additional steps, there's an increased signal degradation and chances of sample contamination (Ozsolak & Milos, 2011).

Nevertheless, this methodology shows high levels of reproducibility for both technical and biological replicates, requiring less RNA sample since there's no amplification step (Wang et al., 2009). Furthermore, RNA-seq has better accuracy and broader dynamic range, which helps to identify more

differentially expressed genes (DEGs) with higher fold-changes[4] (Zhao et al., 2014). Due to its strengths, it is particularly useful in studies with splice variants and non-coding transcripts, enhancing stress related predictions, biological processes, phenomena comprehension and biomarker discovery (Rodríguez-Garcia et al. 2017). Also, since researchers avoid needing any preconceived notions about what to detect (via probes or primers), the overall bias is decreased (Rao et. Al, 2019). Furthermore, it has been demonstrated that RNA-seq is superior in low abundance transcripts detection, since it's possible to increase sequence coverage depth to detect rare transcripts, single transcripts per cell or weakly expressed genes, also allowing the differentiation and identification of genetic variants, such as gene fusions, single nucleotide variants and indels[5] (Zhao et al., 2014).

The protocol of RNA-seq starts by isolating total RNA from the sample of interest, which is then purified to enrich for the specific type of RNA to be profiled (i.e., mRNA, microRNA) (Kukurba & Montgomery, 2015). Afterwards, the RNA sequencing library is prepared through the randomly fragmentation by hydrolysis or nebulization of long RNAs and reverse-transcription into multiple cDNA fragments via random hexamer or oligo-dT[6] priming (Hrdlickova et al., 2017). Alternatively, fragmentation can be done after the creation of the cDNA library. RNA fragmentation produces a globally even read coverage with a decrease towards the transcript ends, whereas cDNA fragmentation results in an overall lower read coverage with an increase of read coverage on both ends (Kukurba & Montgomery, 2015). Sequencing adaptors are then added to each fragment, with or without an amplification step, generating short-reads from one (single-end) or both (paired-end) ends using one of the high-throughput (HTS) platforms, which size selects fragments suitable for sequencing and produces up to hundreds of millions of reads. Lastly, the resulting reads are aligned to a reference genome or transcriptome and categorized as exonic, junction or poly(A)[7] end-reads, generating a base-resolution expression profile for each gene (Figure 2) (Wang, 2009). RNA-seq has a number of different applications, namely in studies related to alternative gene spliced transcripts, post-transcriptional modifications, gene fusions, mutations, SNPs, small RNAs (*i.e.*, snoRNA, miRNA, rRNA), ribosomal profiling and differential gene expression over time in one culture or between samples under control and experimental conditions.

[4] Ratio between expression levels of a gene
[5] Insertion or deletion of bases
[6] Synthetic single-stranded 18-mer oligonucleotide
[7] Polyadenylic acid tails present at 3′

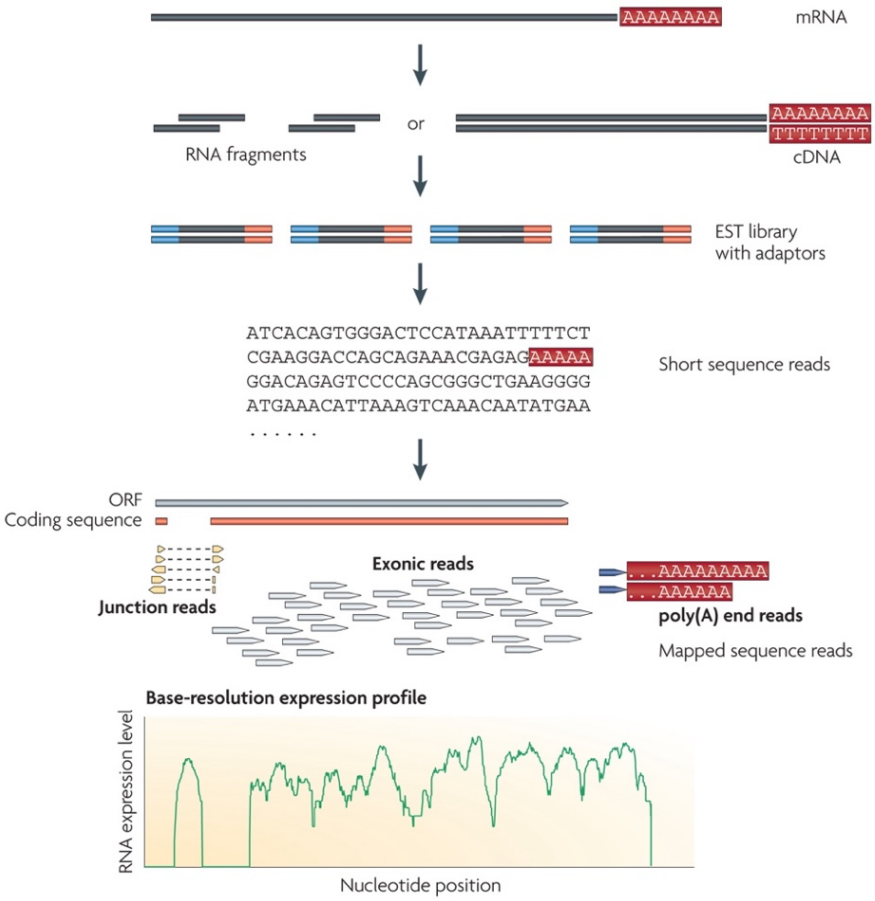

Figure 1.1. Standard RNA-seq protocol (Wang et al., 2009).

When designing RNA-seq experiments there are a few aspects to consider, namely library construction, sequencing depth and number of replicates. Library preparation methods vary depending on the type of RNA, strand specificity and type of reads. Using total RNA allows detection of non-coding as well as mRNA, but may require additional enrichment steps (e.g., ribosomal RNA depletion) to allow the detection of low abundance transcripts (Kim et al., 2019). Poly(A) RNA enrichment can be used to purify mRNA and is useful in studies of eukaryotic organisms (Chung et al., 2018). Non-stranded RNA-seq leads to the loss of orientation of the original RNA transcript, and without that intel it's challenging to accurately determine gene expression from genes that have at least partially overlapping genomic *loci* but are transcribed from opposite strands (Zhao et al., 2015). Conversely, with strand-specific RNA-seq, also known as stranded RNA-seq, it's possible to retain the information pertaining to strand origin which can be useful to accurately quantify gene expression levels for overlapping genes and to identify antisense or non-coding RNA.

Sequencing can also involve two types of reads: single-end (SE), or paired-end (PE). SE sequencing involves sequencing cDNA fragments from only one end to the other, generating the sequence of base pairs (Corley et al., 2017). This solution delivers large volumes of high-quality data, rapidly and economically and it can be a good choice for some methods, such as small RNA-seq or chromatin immunoprecipitation sequencing (ChIP-seq), especially on a low budget (Illumina,

2017). PE sequencing allows users to sequence both ends of the cDNA fragments, aligning the forward and reverse reads as read pairs and producing twice the number of reads for the same time and effort in library preparation (Corley et al., 2017). Overall, PE reads are more useful for alignment than for differential expression (DE) detection (Chhangawala et al., 2015). Since both ends of cDNA fragments are expected to map nearby on the transcriptome, this method deals more efficiently with multi-mapping, solving most of these ambiguities. In addition, sequences aligned as read pairs enable more accurate read alignment and the detection of genomic rearrangements, providing high-quality alignments, even across DNA regions with repetitive sequences, gene fusions and novel transcripts (Rogers et al., 2014). As such, PE sequencing represents a better choice for *de novo* transcriptome assembly, indels discovery and isoform expression analysis, as well as to characterize poorly annotated transcriptomes, since it produces longer contigs by filling gaps in the consensus sequence (Deng et al., 2014). Experiments designed to study splice variants, epigenetic modifications (methylation) and SNPs identification benefit from paired-end runs (Au et al., 2010).

Sequencing depth, also known as library size, represents the number of reads sequenced for a given sample (Conesa et al., 2016). As the sample is sequenced to a deeper level, the reads are likely to cover a larger proportion of the genome/transcriptome, allowing more transcripts to be detected with more precise quantification. Ideal sequencing depth varies with the goals of the research and the complexity of the target transcriptome. Gene expression profiling experiments, that are looking for a quick snapshot of highly expressed genes may only need 5 to 25 million reads per sample. In these cases, researchers can pool multiple RNA-seq samples into one lane of a sequencing run, which allows for high multiplexing of samples (Kukurba & Montgomery, 2015). Experiments looking for a more global view of gene expression, and some information on alternative splicing, typically require 30 to 60 million reads per sample. This range encompasses most published RNA-seq experiments for mRNA and whole transcriptome sequencing. However, when working with complex transcripts libraries, up to 500 million reads may be required to cover full sequence diversity (Fu et al., 2014). Studies have shown that increasing depth can reduce quantification errors, within a limited range. An appropriately stable detection of coding genes can be reached at ~30 M reads per sample, since increases in already considerably deep depths may not contribute significantly to error minimization (Fonseca et al., 2014). Also, when held above 10 M reads per sample, increasing read depth seems to have a small effect on workflow performance, which is much more greatly impacted by the number of biological replicates, giving diminishing returns on power to detect DEG when this number is not increased (Baccarella et al., 2018).

Read length depends on libraries' application and final size. Simple gene expression studies can obtain decent results from short SE reads (e.g., 1 x 50 bp to 1 x 75 bp), while novel transcriptome assemblies, splice junction detection and annotation projects tend to benefit from longer, PE reads (e.g., 2 x 75 bp or 2 x 100 bp) (Chhangawala et al., 2015). However, increasing both read length and library size can have larger impact in the detection of low-expressed genes on DE studies (Lamarre, 2018).

Also, since shorter reads (< 50 bp) have lower mapping quality due to large percentages of multiple alignments, it's important to guarantee sufficient read length and sequencing depth to control technical noise and enable accurate gene identification and expression profiling (Rizzetto et al., 2017).

The number of biological replicates in an RNA-seq experiment, which consists of different biological samples that are processed separately, depends on biological variability of the organisms under study (Conesa et al., 2016). These replicates are required to make inferences on the population, since without the estimation of variability within a group it's impossible to estimate significant differences between groups and conclusions from such results should not be generalized (Manga et al., 2016). As such, two biological replicates must be the minimum for inferential analysis. However, since statistical power increases with increasing numbers of replicates, it's advisable to have at least three clean replicates, considering the chance that one or more replicates within a condition should be rejected (Schurch et al., 2016). Moreover, studies have shown a fairly greater proportional increase in the number of DEGs when moving from two to three replicates compared to going from three to four, suggesting three as a minimal ideal number of replicates (Manga et al., 2016). Furthermore, increases in the replicate number and library size seem to improve sensitivity and specificity, respectively, of Gene Ontology (GO) enrichment analysis, enhancing the underlying biological inferences (Lamarre et al., 2018). Overall, while PE reads and high coverage help to reconstruct lowly expressed transcripts, replicates are essential to resolve false-positive calls, as mapping artifacts or contaminations, at the low end of signal detection (Conesa et al., 2016). Nevertheless, due to budget limitations, transcriptomic studies aim for a suitable trade-off between the number of replicates, reads size and depth, to provide sufficient statistical confidence for efficient, powerful, yet cost-effective analysis (Manga et al., 2016).

Being a HTS sequencing technique, RNA-seq poses a great demand for bioinformatics-based analysis of the generated data, since the analysis of the massive amount of data generated by this large-scale sequencing still faces many obstacles (Zhao et al., 2016). As tools have to balance between sensitivity, specificity and speed, no software can be best suited for all applications. As such, due to the huge number of tools available and the lack of a gold standard method to perform this kind of analysis, one of the main challenges is to make an informed choice of the tools to apply for each data type, organism and project goals (Chatterjee et al., 2018). Also, raw data can require terabytes of storage, which represents many challenges from simply moving the data off the machine, to the outmatched of common desktop computers by the volume of data from a single run (Costa et al., 2010). As such, the use of small clusters of computers is highly advantageous to reduce computational bottleneck.

To obtain meaningful biological knowledge from raw sequencing data is essential to apply *in silico* modular bioinformatics pipelines, using specific software and biological references (Simoneau et al., 2021). The choice of each software and its respective parameters are typically made according to the sequencing protocol and the biological questions. Overall, to achieve such knowledge, the bioinformatic analysis pipeline of RNA-seq consists of five or six fundamental steps, depending on the

existence of a reference genome or transcriptome available for the reviewed organism: 1. raw data processing and quality control; 2. *de novo* transcriptome assembly (optional for non-model species without a reference genome, or when the available reference has low quality and is poorly annotated); 3. mapping the reads to the transcriptome or reference genome; 4. gene, or isoform level quantification; 5. statistic modeling for DE detection; 6. pathway and/or network level analyses to gain biological insight through systems biology approaches (Mutz et al. 2013; Conesa et al., 2016).

1.2. Quality Analysis

Millions to billions of raw short reads are the starting point of RNA-seq computational data analyses, which are submitted to quality control (QC), before alignment and downstream analysis (Zhao et al., 2016). Quality control consists in the analysis of sequence quality, sequence content, sequence length distribution, duplication levels, overrepresented sequences and adaptor content. This control aims at detecting sequencing errors, contaminations, and PCR artifacts (Sathyanarayanan et al., 2019). In addition, QC also includes the analysis of read alignment, namely read uniformity and GC content, quantification, considering 3' bias[8], biotypes, and low-counts, and reproducibility, including correlation, principal component analysis, and batch effects (Conesa et al., 2016; Tarazona et al., 2015). RNA-seq is usually considered unbiased. However, fragmentation and library construction can introduce some biases into the analysis. The number of reads from each transcript is proportional to the number of cDNA fragments rather than the number of transcripts. Since longer transcripts are usually more fragmented, more reads will be assigned to them compared to shorter transcripts. Therefore, when performing DE analysis, DEGs are more likely to be enriched for longer than shorter transcripts, since the statistical power is higher for longer transcripts due to its larger counts (Ma et al., 2019).

First, raw reads can be converted to FASTQ files, which contains not only sequence data, but also per base quality information about that same data, with four lines dedicated to each single sequence read, according to Table 1. Due to its simplicity, fastq format became widely used as a simple interchange file format (Cock et al., 2010).

Table 1.1. Per base quality information about each single raw sequence read in FASTQ format.

Line	Description
1	Character '@' and information about the read
2	The actual DNA/RNA sequence
3	Character '+' and sometimes the same info in line 1
4	Quality scores encoded by a string of characters (same number of characters as line 2)

[8] Enrichment in the 3' end of polyadenylated mRNAs.

In order to quantify quality, encoding scales provide the mapping of quality scores to the quality encoding characters. Although there are different quality encoding scales, which differ by offset in the ASCII[9] table, the most commonly used one is the Sanger format. There are several different ways to encode phred scores with ascii characters. The two most common are called Phred+33 and Phred+64, that take the phred quality score and add 33 or 64, respectively, then using the ascii character corresponding to the sum (figure A1). Since nowadays the Phred+33 is widely used, the Phred+64 is only found on older data that was sequenced several years ago.

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0........10........20........30........40
```

Figure 1.2. Mapping of quality scores (Phred+33) to the quality encoding characters, according to Sanger encoding scale.

Each quality score represents the logarithmically based probability (Pe) that the corresponding nucleotide call is incorrect, which depends on how much signal was captured for the base incorporation. The scores generally range from 2 to 40 with higher scores indicating greater confidence in the call. Phred score is calculated as follows:

$$Q_{PHRED} = -10 * \log_{10}(P_e)$$

It's common to filter out bases with Phred scores below 20, which represents a probability of 1/100 incorrect base calling, although individual preferences according to the specificities of the dataset and the research goals may lead to the choice of other thresholds. Interpretation of Phred scores in terms of base calling probabilities and accuracy are summarized in table 2.

Table 1.2. Interpretation of quality score values in probabilities of erroneous and accuracy base calling.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

To improve the quality of reads, there are a number of QC tools available. While some perform just a simple FASTQ quality assessment, others can automatically filter and trim low-quality data. Among the many available QC tools for HTS data, we can find FastQC (Andrews, 2010) and FastQ Screen (Wingett & Andrews, 2018), both from the same bioinformatics group at the Babraham Institute. FastQC provides a modular set of analyses that can be used to give an overview of whether data has any problems needed to be approached before doing further analysis (Babraham bioinformatics, 2010).

[9] American Standard Code for Information Interchange.

Its output includes an html file with easy-to-read graphical modules, flagging each one as passed (green), warning (yellow), or failure (red), divided in 11 to 12 modules: basic statistics, per base sequence quality, base tile sequence quality, per sequence quality score, per base sequence content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, adapter content and, if relevant, k-mer[10] content. Although FastQC was designed for DNA studies, it can also be used to evaluate RNA data, taking into account that, in this case, some of the modules are expected to present warnings or failures. Since this tool doesn't have the ability to change data, it can be necessary to subsequently submit it to another software capable of filtering and trimming as needed. FastQ Screen is a tool to validate the origin of genomic data, independently of the laboratory protocol followed (i.e., DNA, RNA-Seq, ChIP-Seq or Hi-C[11]), by quantifying the proportion of reads that map to a set of reference genomes (Wingett & Andrews, 2018). It can be used to determine the origin of samples from uncertain or multiple sources, to identify regions rich in low-complexity sequences, or as a quality measure to determine samples' contamination. Furthermore, this tool is able to filter reads mapped or unmapped to specific genomes, allowing to remove contaminants, or to select specific reads according to the research goals.

After analyzing reads integrity, data can be filtered and trimmed to improve quality, since it often needs to be mapped to a reference genome or transcriptome before downstream analysis. This process can be done by trimming low quality ends or entirely removing low quality reads. Overall, trimming has been widely used in HTS analyses, specifically prior to genome or transcriptome assembly, metagenome reconstruction, RNA-seq, epigenetic studies and comparative genomics (Del Fabbro et al., 2013). One of the most popular trimming tools is Trimmomatic, which is a fast, multithreaded command line tool for Illumina SE and PE data (in fastq file format), developed to crop and filter reads. The available trimming steps are: ILLUMINACLIP, to cut adapters and other Illumina-specific sequences; SLIDINGWINDOW, to clip reads once the average quality within a selected window falls below a threshold; MAXINFO, which balances read length and error rate to maximize the value of each read; LEADING and TRAILING, to cut low-quality bases off the start or the end of the reads, respectively; CROP and HEADCROP, to cut reads to a specified length by removing bases from the end or the start of the reads, respectively; AVGQUAL, to drop reads with average quality below a threshold; TOPHRED33 and TOPHRED64, to convert quality scores to Phred-33 or Phred-64, respectively (Bolger et al., 2014).

Although the main goal of this approach is to improve the quality of reads, aggressive trimming may lead to inconsistences across different genes, resulting in differential bias, which can have a large impact in RNA-seq-based gene expression estimates, especially with short-reads (Williams et al., 2016). Furthermore, since short-read sequence aligners take quality information into account and can

[10] Subsequences of length *k* contained within a biological sequence.
[11] Chromosome conformation capture method that performs high-throughput PE sequencing of fragments' nucleotides.

effectively remove adapters via *soft-clipping*[12], conservative trimming may be unnecessary (Liao & Shi, 2020). As such, despite its popularity, trimming may not be required for both mapping and quantification of RNA-seq reads, when performed at gene level, or SNP-calling. Recent studies show that the accuracy of quantification or variant calling from untrimmed reads can be comparable or even slightly better than that from trimmed reads (Liao & Shi, 2020; Bush, 2020). Therefore, trimming is characterized as redundant and, according to the researchers, it can unnecessarily increase data time analysis and costs. Also, this method seems to have little impact on assembly completeness for coding genes, although these results depend strongly on experiment goals (Yang et al., 2019). Nevertheless, if reads have very low-quality ends, gentle trimming may contribute to improve analysis (Williams et al., 2016). When extremely large numbers of reads are available, modest trimming may offer advantages, namely in older low-quality datasets, or in library preparation protocols that are susceptible to adapter contamination, allowing the recovery of reads without deteriorating expression estimation. One possible improvement may be the use of longer reads (e.g., 100-150 bp), so that reads remain long enough after trimming, or the addition of a minimum read length filter to shorter reads trimming, in order to minimize the introduction of unpredictable changes in expression estimates (Williams et al., 2016). Thus, to determine the best strategy, researchers should consider a trade-off between different trimming approaches according to downstream applications and to the available computational time (Yang et al., 2019).

1.3. Assembly

After quality control, cleaned reads can be either aligned to a reference genome or transcriptome, possibly requiring and additional assembly step. Transcriptome reconstruction can be classified as reference-based, when a reference genome is available to guide de assembly, and *de novo* assembly, when a reference isn't available or is incomplete (Marchant et al., 2016). Reference-based approaches are less computationally demanding than *de novo* methods, being suitable for detection of low abundance transcripts in organisms with reliable reference genomes (Benjamin, et al., 2014). Due to its high efficiency and sensitivity, it is also sometimes possible to use the reference from a closely related species. Conversely, *de novo* assemblers have the advantage of not requiring a reference, which allows the discovery of novel transcripts, although this approach need deeper sequencing (Hansen et al., 2012). This method can be used to reconstruct transcriptomes from a large number of cDNA fragments, without *a priori* knowledge of their correct sequence or order (Hölzer et al., 2019). Overall, *de novo* assemblies can generate accurate reference sequences, even for species with complex or polyploid genomes (Gutierrez-Gonzalez & Garvin, 2017). These assemblies may provide useful information about unknown or poorly annotated genomes, clarifying highly similar or repetitive regions and

[12] Masking of portions of reads that do not align to the genome from end to end.

identifying structural variants and complex rearrangements (e.g., deletions, inversions, or translocations). Nevertheless, due to the nature of transcriptomes, technical challenges and bioinformatics tasks, it can be difficult to correctly assemble sequences from some organisms. Features like alternatively spliced variants of the same gene, genomes with high degree of polymorphism, or high dynamic range of expression can make the reconstruction of transcripts a major challenge (Strickler et al. 2012; Kukurba and Montgomery 2015). Some studies suggest that combining reference-based and *de novo* strategies, merging short-read and long-read technologies, provide better results to assemble fragmented transcripts and can improve genome reconstruction by integrating information of a related genome (Marchant et al., 2016; Lischer & Shimizu 2017). The ideal assembly would require high coverage, high read length, and very good read quality. Since no sequencing platform provides all these features, the assembly method should be chosen based on the particular assets of each dataset and project goals. Through the combination of different approaches, it's possible to enhance the detection of a broader range of structural variant types and the accuracy of identification of complex rearrangements, yielding higher-quality assemblies. However, since there isn't a gold standard method for all analysis, the best strategy highly depends on the research design, the available genomic data and computational resources (Benjamin, et al., 2014).

Over the last years, different assemblers were developed to fit particular needs such as the assembly of smaller and simpler prokaryotic transcripts, or large and more complex eukaryotic transcriptomes, where alternative splicing has to be considered to reconstruct different isoforms (Hölzer & Marz, 2019). To perform *de novo* assembly, it's fundamental to have a graph representation of relationships between reads, sharing common prefixes or suffixes. The most widely used assemblers are based either in *de Bruijn graphs* (DBG). Although less informative than other algorithms, it's less computationally intensive, allowing the application of more sophisticated arguments, due to its simpler structure (Rizzi et al., 2019). Through *de Bruijn* graph assemblers, reads are broken into k-mers[13], which are then used as nodes in the graph assembly. Later, nodes that overlap by some amount (generally, k=1) are connected by an edge and the assembler construct sequences based on the *de Bruijn* graph. There are a number of different good quality assemblers available based in DBG, among which is Trinity. This is a simple and intuitive software package for conducting efficient and robust *de novo* transcriptome assembly from Illumina RNA-seq data, which also supports genome-guided assembly, useful in non-model organisms (Grabherr et al., 2013). This package requires little to no parameter tuning and includes scripts for generating statistics to assess assembly quality, and for wrapping external tools to perform downstream analyses (Hölzer & Marz, 2019). Partitioning data into several independent DBG, preferably one per gene, it uses parallel computing to reconstruct transcripts, including alternatively spliced isoforms (Haas et al., 2013). Studies have demonstrated that Trinity is one of the highest effective assembly methods, generating consistently good assemblies and producing

[13] Nucleotide sequence (substring) of length k, contained within a biological sequence (read).

longer contigs (Hölzer & Marz, 2019). Since Trinity was developed for RNA-seq data, unlike assemblers written for genomic DNA, it produces one contig per isoform rather than per locus, and different transcripts are expected to have different coverage, due to their distinct expression levels (Haas et al., 2013). Trinity can process either strand-specific Illumina PE and non-strand-specific and single-end SE libraries, and it also supports tools that take its output transcripts and test for DE, while accounting for both technical and biological sources of variation and correcting for multiple hypothesis testing (Figure 1.3). Despite Trinity's effectiveness, without a reference genome, it may not be possible to completely understand the structural context for the transcript variations (e.g., number of skipped exons, or retained introns), if the experimental goals involve profiling expression at the isoform level.

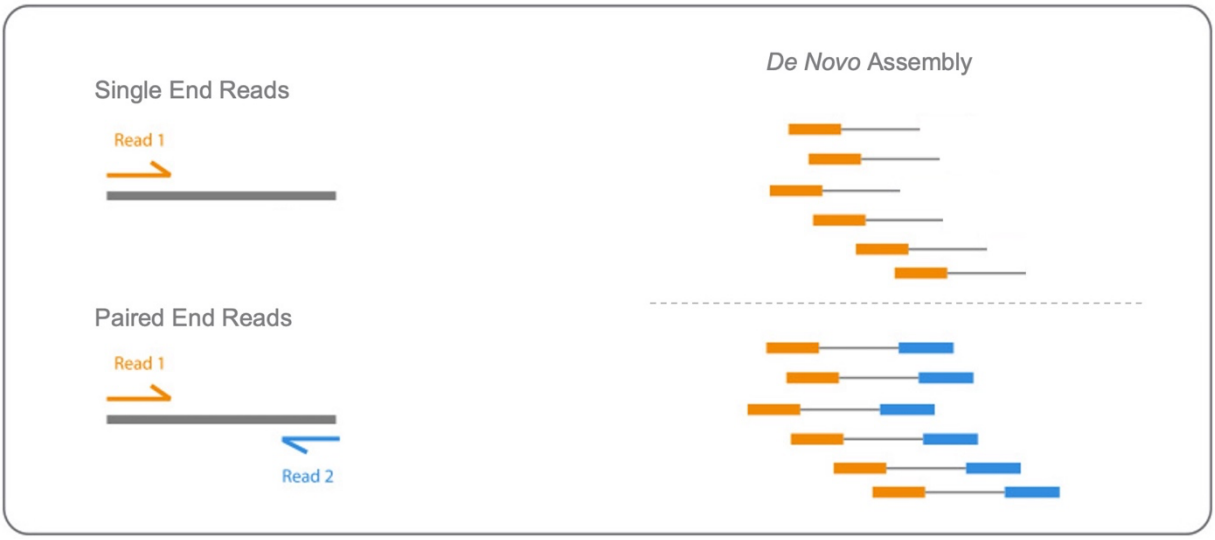

Figure 1.3. *De novo* transcriptome assembly of single-end (SE) and paired-end (PE) reads.

After assembly, it's important to assess its quality to ensure accurate mapping and quantification results for downstream analysis. Although Trinity can generate N50 statistics and counts of the number of contigs, these metrics should not be treated as good indicators by themselves. However, the more transcripts Trinity assembles, especially with just a few hundred bases long, the more likely contigs represent subsequences of actual genes (Adam & Weeks, 2020). Assemblies based upon the same read data can be evaluated with respect to the numbers of genes that are complete, fragmented, or missing from the assembly. One of the available tools to assess completeness is Benchmarking Universal Single-Copy Orthologs (BUSCO). This tool wraps HMMER, which performs biosequence analysis using profile hidden Markov models, searching databases for sequence homology, and determining whether assembly contigs are orthologs[14] with a particular BUSCO dataset entry (Simão et al., 2015).

[14] Genes in different species evolved from a common ancestral gene.

1.4. Alignment

One of the fundamental tasks in RNA-seq data analysis is mapping large sets of high-throughput sequenced reads. Pairwise sequence alignment is a technique that intends to find identical or similar regions between two sequences, regardless of whether they are related by common descent (divergent) or not (convergent), to help give biological meaning to sequenced data and predict gene functions, e.g., through the detection of protein family, or repetitive elements across genomes (Baichoo & Ouzounis, 2017). Using an appropriate aligner, reads can be mapped to either a genome or a transcriptome. Transcriptome-based methods employ unspliced aligners to map RNA-seq reads directly to the transcriptome and do not require considering splice junctions between exons. Conversely, spliced aligners that map reads to the genome can be used to help identify novel exons, isoforms, or transcribed regions of the genome, as well as cope with events like intron retention, which are more difficult to account for when using methods that align reads to the transcriptome (Srivastava et al., 2020). Frequently, sequenced reads are mapped and aligned to a reference genome. However, this technique is particularly prone to errors for RNA-seq data due to the presence of output reads spanning exon-exon splice junctions (Raplee et al., 2019). Moreover, RNA-seq datasets typically consist of millions to billions of relatively short sequence fragments of the original RNA transcripts, which are usually generated from large genome of many important species, making this task very computationally intensive (Dobin et al., 2013). Furthermore, transcripts are often spliced, requiring mapping to noncontiguous regions of the genome and thus creating a unique challenge for the RNA-seq mapping, both in terms of speed and accuracy. Since different aligners may have distinct mapping performances and the selection of the right tool is strongly related to each project goal.

Overall, when conducting gene expression estimation, studies suggest applying aligners that produce high percentages of total aligned reads and aligned reads with at most one mismatch (Yang et al., 2015). Using this quality performance criterion, and although a single aligner program cannot universally be applied to all RNA-seq datasets, Spliced Transcripts Alignment to a Reference (STAR) is among the best tools available. STAR is a feature-rich software, with support for annotated and novel splice-junctions, chimeric and circular RNA, and fusion read detection. It is capable of running parallel threads on multicore systems with high productivity, making it fast (Dobin et al., 2013). Due to its moderate error rate, it allows highly accurate spliced reads alignment at ultrafast speed, providing scalability for emerging sequencing technologies (Dobin & Gingeras, 2015). STAR can generate various data files useful for downstream analyses such as transcript/gene expression quantification, differential gene expression, novel isoform reconstruction, and signal visualization. Moreover, STAR shows better alignment precision and sensitivity for both experimental and simulated data in comparison to other aligners (Dobin et al., 2013).

Another suitable aligner is Bowtie2, which is an ultrafast and memory-efficient tool that aligns reads to reference sequences, claiming to combinate high speed, sensitivity and accuracy (Langmead &

Salzberg, 2012). It generates multiple alignments for each read, reporting the single best alignment per read. Bowtie2 supports gapped, local and paired-end alignment modes, indexing genomes with an FM Index[15] to reduce its memory footprint. In local mode, this aligner can trim low-quality bases from either end of the reads to increase alignment score. Bowtie2 is particularly useful to align reads from 50-100 bp to relatively long genomes (*e.g.*, mammalian), being optimized for the read lengths and error modes yielded by typical Illumina sequencers. It can be used by itself, or as a resource for other software, as is the case of FastQ Screen.

After mapping, it's important to measure the quality of the alignment to ensure accurate downstream analysis. The percentage of mapped reads is a global indicator of the overall sequencing accuracy and of the presence of contaminating DNA (Conesa et al., 2016). Depending on the aligner and the species under investigation, 70 to 90 % of regular RNA-seq reads are expected to map onto the reference genome in a good quality alignment, with a significant fraction of reads mapping equally well to a limited number of identical regions, known as multi-mapping reads. If the same reads are mapped against the transcriptome, total mapping percentages are expected to be slightly lower due to the loss of reads correspondent to unannotated transcripts. The number of multi-mapping reads is expected to be significantly higher since several reads will map onto exons shared by different isoforms of the same gene (Conesa et al., 2016). Since alignment is the first step in RNA-seq cleaned data analysis, all subsequent analysis relies profoundly upon this initial step and can be positively or negatively influenced by it, especially to detect and identify differential gene expression (Raplee et al., 2019). However, some studies show that mapping methods may have minimal impact on the final DEGs analysis, if a good quality annotated reference genome is available. Although there are multiple genome annotations available, some are incomplete or inaccurate, which suggests that the choice of the genome annotation may have a relevant effect on downstream RNA-seq data analysis (Costa-Silva et al., 2017; Zhao & Zhang, 2016).

1.5. Quantification

After alignment, reads mapped to each gene are counted to determine the expression profile of the samples. Since transcript quantification is proportional to gene expression, in addition to many other factors, it is a prerequisite and a key step in the RNA-seq data analysis pipeline, and the accuracy of expression quantification can profoundly affect downstream analysis (Srivastava et al., 2020). There are two main approaches for transcript quantification, namely alignment-based and alignment-free tools. The first maps all the reads to a genome or transcriptome, and then counts the number of reads that map to an individual transcript or gene, while the second quantifies expression counting unique *k-mers* in a sequencing library, without mapping to a reference (Babarinde et al., 2019). Unlike alignment-

[15] Compressed self-index, that compresses data and indexes at the same time.

base, alignment-free approaches don't require the alignment of every read before quantification, which makes them much faster. However, they are unsuitable to the quantification of repeat-derived RNA, since they only exploit unique splicing patterns to collect unique k-mers, being most suitable to transcript-level quantification and in genomes with decent transcript annotations. Alignment-free methods tend to perform poorly with lowly expressed transcripts or short RNA and can be misleading at gene-level quantifications, but in some cases, the loss of sensitive can be an acceptable trade-off for speed improvement (Babarinde et al., 2019).

There are a number of quantification methods to choose from, among which are two of the most widely used alignment-based tools, namely RNA-seq by Expectation-Maximization (RSEM) and HTSeq. RSEM iteratively assigns reads to each transcript based on the probabilities of the reads being derived from each transcript and taking into account positional biases produced by library-generating protocols (Haas et al., 2013). This method can estimate expression at both gene and isoform-level and can be applied to single-end or paired-end reads, with or without a reference genome, being particularly useful for quantification with *de novo* transcriptome assemblies (Li & Dewey, 2011). Also, studies show that higher sequencing depth can increase the accuracy of RSEM quantification (Babarinde et al., 2019). By default, RSEM used parameters specifically chosen for RNA-seq quantification, and can be performed using Bowtie (default), Bowtie2, or STAR aligners. HTSEq is a quantification tool increasingly in used due to its speed, comparable to alignment-free methods, but with improved sensitivity. HTSeq uses a naive count-based approach for expression estimation and can be used either for strand-specific or non-stranded samples (Chandramohan et al., 2013). Unlike RSEM, HTSeq outputs only counts of reads aligned to genes but not the counts of reads involved in a particular isoform, i.e., the gene is considered to be a union of all exons. However, HTSeq is faster, easier to handle, less dependent on the choice of the mapper and, because of its computational efficiency, is great for preliminary data analysis or for quick assessment of relative expression estimates.

In transcriptomic data analysis, quantification can be done either at transcript-level or gene-level, where the number of reads of different isoforms are counted individually, or grouped by gene, respectively. Gene-level estimation is considerably simpler and more stable than transcript-level, being better in terms of robustness, accuracy, statistical performance and interpretation (Soneson et al., 2015). Furthermore, it removes a lot of confounding information related to minor transcript isoforms (Babarinde et al., 2019). Overall, for RNA-seq analysis, gene-level quantification should be preferable unless there is a particular reason to consider isoforms, such as splicing factor mutation, which will specifically impact expression of particular transcripts rather than genes (Conesa et al., 2016).

One setback of transcript-level quantification is that the measurement of DE can often overemphasize changes in the several minor transcripts of a gene, whilst its major transcript is relatively unchanged, making interpretation a challenge. On the other hand, analysis at the gene-level loses much of the complexity of transcript expression and is not easily suited to the analysis of particular types of non-coding genes, such as anti-sense or sense intronic transcripts, which are difficult to interpret in

gene-level quantification (Babarinde et al., 2019). Although the transitioning from transcripts to genes is substantially complex, some studies suggest that taking advantage of transcript-level abundance estimates when defining or analyzing gene-level abundances can lead to improved differential gene expression results compared to simple counting. Through an adequate combination of both approaches, it can be possible to increase sensitivity and accuracy of gene-level quantification (Yi et al., 2018).

A typical quantification begins with the number of reads or *k-mers* that mapped to a transcript or gene. This number depends on the expression level, library size, percentage of aligned reads, transcript length, GC content, and other confounding parameters (e.g., batch effect[16], or operator bias). To surpass these issues, quantification is typically followed by expression normalization (Babarinde et al., 2019). Accurate gene expression quantification requires not only accurate sequence read alignment, but also an adequate normalization method. Several metrics have been proposed for measuring transcript abundance levels based on RNA-seq data, normalizing for depth of sequencing and length of transcripts. Normalization makes the expression levels more comparable between and/or within samples and converts gene expression data from counts to a continuous scale. While normalization is essential for DE analyses, it is also necessary for exploratory data analysis, data visualization, and to explore or compare counts between or within samples (Khetani & Mistry, 2017). The main factors often considered during normalization are sequencing depth, to compare expression between samples, gene length, to compare expression between different genes within the same sample, and RNA composition, recommended for accurate comparison of expression between samples and particularly important when performing DE analyses (Evans et al., 2018; Robinson & Oshlack, 2010). RNA-seq can present some statistical issues for gene expression analysis, like large numbers of genes and few replicates; discrete, positive and skewed data; large dynamic range with presence of null counts and variable library size. Also, the detection of DEGs is inherently biased, since there is more power to detect DE of longer genes (Soneson & Delorenzi, 2013). As such, since DE analysis is mainly concerned with relative changes in expression levels between conditions rather than estimating absolute expression levels, normalization has a great impact on its results, assuming that DE and non-DE genes behave the same, balanced expression (Evans et al., 2018).

Gene expression estimates can be evaluated by the number of genes falsely quantified and the number of genes with falsely estimated $\log_2$FC (Yang et al., 2015). Studies have shown that bellow an absolute $\log_2$FC, normalization performances tend to suffer, generating bias and higher variance values. As such, literature isn't clear when reporting relative expression level units in RNA-seq data and the most fitting normalization method to use depends on which assumptions are valid for the biological experiment (Conesa et al., 2016). In the Counts Per Million (CPM) method, counts are scaled by the total number of reads, taking into account the sequencing depth. Although it can be used for gene count comparisons between replicates of the same sample group, having several advantages for samples where

[16] Data variability that is *not* due to the variable of interest.

RNA quality is low, it is unsuitable for within sample comparisons or DE analysis. Among the most frequently reported unit of expressions for RNA-seq data are RPKM, FPKM and TPM. These methods perform counts per length of transcript (kb) per million reads mapped, accounting for sequencing depth and gene length. The RPKM and FPKM methods are only recommend for gene count comparisons within samples, while TPM can also implement comparisons between samples of the same sample group. However, the order in which these methods normalize the read counts causes differences within samples that can add bias to the results, making them not very suitable for DE analysis (Evans et al., 2018). Also, studies have shown that RPKM normalization is not able to control false positives in data having a small number of genes with very high read counts (Lin et al., 2016). Unlike these approaches, the DESeq, the Relative Log Expression (RLE) and the Trimmed Mean of Means (TMM) are designed to account for extreme differences in read count number, being able to compensate effectively for RNA-seq data with a large dynamic range. Also, they can deal efficiently with the intrinsic bias resulting from the relative size of transcriptomes (Lin et al., 2016). The TMM uses a weighted trimmed mean of the log expression ratios between samples that accounts for sequencing depth, RNA composition and gene length. It is recommended for gene count comparisons between and within samples and for DE analysis. Although TMM seems to be very restrictive, showing low sensitivity, *i.e.*, generating less numbers of DEGs, it has high specificity. Having the ability to detect DEGs while controlling false positives, it shows great classification performance (Li et al., 2020). DESeq normalization uses the ratio of each read count to the geometric mean of all read counts for a gene across all samples. The median of these ratios, called the *size factor*, is then used to scale samples (Evans et al., 2018). Doing so, these method accounts for sequencing depth and RNA composition, being suitable for gene count comparisons between samples and for DE analysis, but not for counts within samples. The Median Ratio normalization (MRN) is an alternative and similar method to the TMM, with the goal of being more robust. Read counts are divided by the total count of their sample and averaged across all samples in a condition for a given gene, producing an average count-normalized value for each gene and each condition. The original counts are then normalized by the median of the ratios of these values between conditions and their library size.

1.6. Differential Expression

Count tables resulting from RNA-seq data quantification are typically analyzed through the application of statistical methods to detect DEGs. Numerous methods have been developed specifically for RNA-seq, attempting to accurately quantify the abundance of transcripts or genes within different conditions and time points, and to correlate deviations in its abundance to genetic and environmental changes in order to comprehend genome function and adaptation (Fang et al., 2012). Parametric and nonparametric are two broad classifications of statistical procedures. They differ from nonparametric in data distribution assumption, biological replicates handling, ability to perform multi-group

comparisons and computational intensity. The use of nonparametric methods is a way to reduce the difficulty of modelling counts in some experiments, since without depending on underlying distributional assumptions, they can give reliable results on a vast variety of data sets. However, since parametric methods use the most powerful test statistic, being more efficient when the distributional assumption holds, even when the sample size is small, they are the most used in RNA-seq studies. Nevertheless, if the distribution is poorly approximated, the results may not be reliable (Li & Tibshirani, 2013). Many of these methods are implemented in *R/Bioconductor*[17] packages, namely the widely used and best performing parametric DESeq, DESeq2 and edgeR, and nonparametric NOISeq.

DESeq fits a generalized linear model to estimate variance-mean dependence in count data, testing for DE based on the negative binomial distribution and a Likelihood Ratio Test (LRT). The core assumption of this method is that the mean is a good predictor of the variance, i.e., that genes with similar expression levels also share similar variance across replicates, and thus it estimates a function for each condition that allows to predict the variance from the mean. DESeq incorporates information from all the genes in a set of samples to circumvent small sample sizes and it normalizes data through its intrinsic method, already discussed above. Assuming that most genes behave the same within replicates as across conditions, it allows testing on samples without replicates, estimating variance by treating all samples as if they were replicates of the same condition. Although statistical significance can't be truly credited without replicates, it is anticipated that the estimated variance should not be affected too much by the influence of DEGs. In that case, it's assumed that the dispersion estimated by DESeq will be too high, so that the test will err to the side of being too conservative, i.e., high false negative rate (FNR) and low false positive rate (FPR) (Anders and Huber, 2010). Thus, the results from such analysis are expected to be more incomplete than incorrect, which can be particularly useful for exploratory analysis.

DESeq2 also uses a negative binomial distribution to model read counts (Ren & Kuan, 2020). First, it normalizes the counts of each gene employing a generalized linear model that uses the variance of all the genes to improve the variance estimated for each individual gene (MRN). Then it applies an empirical Bayes shrinkage to detect and correct for dispersion and high variance $\log_2$ fold change ($\log_2$FC) estimates, which represents the ratio of the expression of the two samples. Unlike DESeq, it doesn't allow the absence of replicates and it can find the value of the parameter that makes the likelihood largest, also known as maximum likelihood estimation (MLE) (Love et al., 2014).

The edgeR method is a *Poisson* super dispersion model that uses weighted likelihood methods to implement a flexible empirical Bayes approach to allow gene-specific variation estimates. Its algorithm is also based on a negative binomial distribution, with variance and mean bound by local regression, and it uses as default method of normalization the TMM. However, there are alternative normalization methods available in edgeR to account for data that fail to conform to a negative binomial distribution,

[17] Free, open source and open development software project for the analysis and comprehension of genomic data, based primarily on statistical R programming language.

which is assumed with TMM. According to studies, although poisson distributions can correctly characterize data from technical replicates, negative binomial models are more appropriate to describe data from biological replicates, which have much larger variance, being a suitable choice for RNA-seq data analysis (Li & Tibshirani, 2013). Like DESeq, edgeR can be applied even with very few or no replicates (Chen et al., 2014), in which case the dispersion value has to be fixed manually. Although both DESeq and edgeR incorporate information sharing in the dispersion estimation, the way that this information sharing is done accounts for the main difference between the two methods. Overall, DESeq tends to be more conservative than edgeR, showing better control of false positive rate, but edgeR is more suitable for experiments with fewer than 12 replicates (Schurch et al., 2015).

Using a very different approach, NOISeq is an exploratory analysis tool that tests for DE between two experimental conditions with no parametric assumptions, that can simulate technical replicates. This method relies on the premise that read counts follows a multinomial distribution, where probabilities for each feature are the probability of a read to map to it (Tarazona et al., 2011). Adaptative to the data, NOISeq empirically models the noise in the counting data and allows data analysis without replication. In this method, genes are differentially expressed if the ratio of log_2 between two conditions and the value of the difference between the two corresponding conditions are likely to be higher than a noise. The noise distribution is obtained by comparing all pairs of repetitions within the same condition.

Both edgeR and DESeq2 require raw read counts in a data matrix, which perform the same kind of normalization to account for differences in sequencing depth, and low count variability. Also, both tools assume that RNA-seq data display overdispersion with variance greater than expected for random sampling and to a negative binomial distribution, employing Bayesian methods to fit gene expression counts into it. On the other hand, NOIseq and DESeq2 presents the highest true positive rate (TPR), accuracy and specificity and are suitable for experiments with a large number of samples and an annotated genome. However, NOIseq can obtain a low amount of unidentified DEGs, independently of the mapper previously used, since DE analysis is more influenced by the methodology of DEG identification than the adopted methodology of mapping or quantification of reads. In general, although some methods perform better than others and the results also depend on the nature of data and the experiment goals, the combination of different methods seems to produce more suitable results, presenting better balance than individual solutions (Costa-Silva et al., 2017).

The output of RNA-seq DE analysis is a list of significant DEGs. The significance of these DEGs can be determined through a filter of corrected p-value and/or log_2FC, depending on the type of data and the experiment goals. The False Discovery Rate (FDR) is a statistical approach that is typically applied in RNA-seq studies, to correct for multiple comparisons in multiple hypothesis testing. It is defined as the expected proportion of false discoveries, *i.e.*, incorrectly rejected null hypothesis (Li et al., 2012). To filter significant DEGs, FDR is usually ranged between 0.01 to 0.1. The log_2FC filter should be used more carefully, since it only accounts for the size of the effect of DE. For some genes, subtle differences in expression can have a substantial impact in biological pathways, while for others,

only great changes can effectively cause real impacts. Moreover, log_2FC is sensitive to lowly-expressed genes, where the variability is high and thus the value might not be accurate. However, for cases where a statistical significance can be attributed and/or when in the presence of a huge number of DEGs, a log_2FC threshold can be useful to narrow down the search to genes of interest (Chen et al., 2016).

1.7. Functional Analysis

After DE analysis, it's important to interpret the results using functional analysis tools to gain biological insights on the DEGs lists. These tools span a wide variety of approaches and can be roughly categorized in three main classes: over-representation analysis, functional class scoring and pathway topology. These analyses are highly dependent on biological information from multiple studies, like gene and protein annotations, or pathway networks. Since the goal of functional analysis is to offer biological knowledge, it's necessary to analyze results in the context of experimental hypothesis, using tools to validate experimental results and to make hypotheses, suggesting genes/pathways that may be involved with the condition of interest. However, these results shouldn't be used to make definite conclusions about the pathways, since those conclusions require experimental validation.

1.7.1. Databases

Recent advances in HTS technologies coupled with decreases in sequencing costs have resulted in the collection of large volumes of RNA-seq data from several organisms. These data are usually deposited in online repositories in formats that are text and/or table-based (Robinson et al., 2018). To store more complex data, there are also several biological databases available. These are typically huge, organized bodies of exception-ridden, vast and incomplete biological information, generally associated with software developed to quickly update, query, and retrieve components of the stored data (Rhee & Crosby, 2005). Moreover, databases are also useful to provide web application programming interfaces (APIs) to automatically exchange and integrate data from multiple databases (Zou et al., 2015). Several databases have been designed and interpreted to ensure unambiguous results (Rhee & Crosby, 2005). A range of information can be retrieved by using biological databases, namely genomic sequences, metabolic interactions, functional relationships, protein families and homologous. Databases can be classified as primary, containing only sequences or structural information, or secondary, which content derives from the analysis or treatment of primary data (Zou et al., 2015).

The National Center for Biotechnology Information (NCBI) provides access to most biomedical and genomic information through its powerful and large servers, contributing to advances in science and health. Among the numerous resources of NCBI, the Sequence Read Archive (SRA) constitutes the largest publicly available repository of high throughput sequencing data, from all branches of life as well as metagenomic and environmental surveys. This stores raw sequencing data and alignment

information to enhance reproducibility and facilitate new discoveries through data analysis (Leinonen et al., 2011).

The UniProt Knowledgebase (UniProtKB) is one of the many available databases and represent the main hub for the compilation of accurate, consistent and rich proteins' functional annotations. Mainly, this database captures amino acid sequences, protein names or descriptions, taxonomic data and citation info, adding as much annotation information as possible to each entry, like biological ontologies, classifications, cross-references and annotations' quality (i.e., experimental or computational evidence). The UniProtKB consists of two sections namely *UniProtKB/Swiss-Prot* and *UniProtKB/TrEMBL*. The first contains high-quality, non-redundant, manually-annotated records extracted from literature and curator-evaluated computational analysis. The second comprises computationally analyzed records that provides high annotation coverage of the proteome, awaiting full manual annotation (The UniProt Consortium, 2021).

The PubMed Central (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institute of Health's/National Library of Medicine (NIH/NLM) that comprises over 32 million citations. Since it's beyond the ability of anyone to comprehend information in such amounts without computational help, there is an increasing need of relying on the creation of controlled structured vocabularies such as ontologies (Hill et al., 2008). These systems allow experimental data to constitute a formal, structured representation of the reality, captured by the underlying biological science. Gene Ontology (GO) is a bio-ontology that describes gene products with three independent categories: biological process, cellular component and molecular function, which may produce multiple GO terms assigned to one query sequence (Ashburner et al., 2000). As such, a GO annotation represents an association between a gene product type and that product's function, what biological processes it contributes to, and where in the cell it is capable of functioning. The Gene Ontology Annotation Database (GOA) stores the corpus of all GO terms' annotations to UniProtKB entries, being accessible through the QuickGO interface, which is a web-based tool for searching and view data from the GOA database (Binns et al. 2009).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a prominent reference knowledge base resource that integrates and interprets genomic, chemical, health and systemic functional information for understanding high-level functions and utilities of the biological system. Large-scale molecular datasets generated by sequencing and other high-throughput experimental technologies are integrated on molecular wiring diagrams of interaction, reaction and relation networks, representing systemic functions of the cell and the organism (Kanehisa & Goto, 2000). Wikipathways, another database of biological pathways, is known for its collaborative nature and open science approaches, maintained by and for the scientific community. It presents a model for pathway databases that enhances and complements other platforms, such as KEGG, paving the way towards more sustainable, community-driven biology databases, yielding to provide intuitive views of the myriad of interactions underlying biological processes (Martens et al., 2021).

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) is a database of known and predicted protein-protein interactions (PPIs), that includes direct (physical) and indirect (functional) associations, using computational prediction, knowledge transfer between organisms, and combining interactions from other primary databases (Jensen et al., 2009). PPI networks are a vital element for system-level comprehension of cellular machinery and can perform tangible, practical purposes such as filtering and assessing high-throughput functional genomics data to provide intuitive visual scaffolds for annotating the structural, functional, and evolutionary properties of proteins.

1.7.2. Functional Annotation

The Basic Local Alignment Search Tool (BLAST) is a software specialized in finding regions of similarity between biological sequences. Through the comparison of nucleotide or amino acid sequences to sequence databases it calculates the statistical significance of the resemblance, above a certain threshold (Boratyn et al., 2013). The BLAST algorithm is divided in three main functions: *blastn*, which searches nucleotide databases using a nucleotide query; *blastp*, which searches protein databases using protein queries; and *blastx*, which searched protein databases using a translated nucleotide query. This program also includes *tblastn* and *tblastx* tools, which searches translated nucleotide databases using protein and translated nucleotide queries, respectively. The *blastx* tool is particularly useful in RNA-seq studies, since it allows the identification of potential protein products encoded by transcripts, allowing transcriptome annotation of non-model organisms.

When working with a model organism, annotation can be made by running a BLAST search against the organism's reference protein dataset. For non-model organisms, because of the absence of a specific reference for comparison, the analysis has to be done against the proteins of a closely related species (when available), or a dataset containing all known proteins. Lastly, the percentage of best aligning targets can be recovered to select the best hit for each transcript. To do so, *blastx* can search multiple databases, which can be select according to the research goals.

Table 1.3. Blast databases (*excluding those in PAT, TSA and env_nr).

Protein Database	Content
Non-redundant protein sequences (nr)	Non-redundant sequences from GenBank + RefSeq + PDB + SwissProt + PIR + PRF*
RefSeq Select Proteins (refseq_select)	Representative transcripts for protein-coding genes from RefSeq
Reference Proteins (refseq_protein)	All protein sequences from RefSeq
Model Organisms (landmark)	Proteomes from representative genomes spanning a wide taxonomic range
UniProtKB/Swiss-Prot (swissprot)	UniprotKB/SwissProt sequences (last major release)
Patented protein sequences (pataa)	Patented division of GenBank
Protein DataBank proteins (pdb)	Protein sequences with 3D structure records
Metagenomic proteins (env_nr)	Protein sequences translated from the CDS annotation of the metagenomic nucleotide sequences
Transcriptome Shotgun Assembly Proteins (tsa_nr)	Protein sequences translated from CDS's annotated on transcriptome shotgun assemblies

Although the default *nr* database is comprehensive and frequently updated, considering its huge size, it requires a significantly high computational power and storage, and a complete search can be overly time-consuming. In most cases, there are better choices of database, such as a subset of *GenBank* for the organism of interest or a UniProtKB complete proteome.

1.7.3. Enrichment analysis

A common feature of HTS technologies, alongside other bioinformatics analysis, is the output of lists with hundreds to thousands of genes. However, to gain greater biological insight on DEGs, the interpretation of each gene individually is, evidently, not practical. As such, several tools have been developed to search for sets of genes with particular interest, which typically interact in a common biological pathway. Consistent perturbations over such gene sets frequently suggest mechanistic changes. One of the ways to identify these interactions is through enrichment analysis, which compare a given set of genes to a background, that can be a whole reference genome, or the only genes expressed by the organism in study. In practice, these background gene sets are compiled from gene and pathway annotation databases such as GO, KEGG, or Wikipathways (Simillion et al., 2017).

Enrichment analysis approaches search for sets of genes that are significantly over-represented in a given list, compared to a background set of genes. Many tools can perform Over-Representation Analysis by querying biological databases that typically categorize genes into groups (gene sets) based on shared functions, involvement in pathways, presence in specific cellular locations, or other common categorizations. Known genes are organized into categories based on its functional annotation. The proportion of DEGs associated with a specific category is compared to the proportion of total genes associated with the same category in the background set. Over-represented categories are, then, determined based on the probability of having a significantly higher proportion of DEGs in a specific category than was expected for the organism in investigation.

A Gene Set Enrichment Analysis (GSEA) is a particular type of Over-Represented Analysis (ORA) that takes into account the DE values obtained from the DE analysis, ranking genes with decreasing $\log_2$FC. An ORA simply uses a *hypergeometric test* to find enriched categories among all DEGs, regardless its expression regulation, although it can accept while either flat or ordered gene lists by decreasing or Wikipathways. This tool includes an ordered query option, which is useful for RNA-seq data when ranking DEGs by $\log_2$FC. Using this option, g:Profiler performs incremental enrichment analysis with increasingly larger numbers of genes starting from the top of the list. This approach identifies functional terms that are associate to the most significant changes, as well as broader terms that characterize the gene set as a whole. importance. Conversely, a GSEA uses a permutations algorithm, with the possibility to rank categories based on FDR and then selects the top N most significant genes from positive and negative related categories, separately. One of the tools that can perform GSEA is the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt), which is a popular suite

of comprehensive, powerful, flexible and interactive tools for functional enrichment analysis in various biological contexts. Currently, it supports 12 organisms, 342 gene identifiers and 155 175 functional categories, as well as user-uploaded functional databases. Also, to facilitate comprehension of the enrichment results, it includes methods to reduce redundancy between enriched gene sets, improving the visualization of results. ORA can be performed in several tools, among which is g:Profiler. This is a public web server that has a simple user-friendly web interface with powerful visualizations, currently available for more than 400 species, including mammals, plants, fungi, insects from Ensembl and Ensembl Genomes. This software contains a function specifically developed for ORA, the g:GOSt, which performs functional profiling of gene lists using various kinds of biological evidence, such as GO terms, KEGG pathways.

The selection of an inappropriate background set can heavily influence an enrichment protocol, resulting in concepts and genes appearing to be more significant than they actually are, or appearing significant (*i.e.,* biased) when the bias is actually due to methodology rather than biology. As such, it is imperative to think carefully about the set from which an interesting subset of genes was taken. A good rule of thumb for background selection is only to include those genes or proteins that have a chance of making it into the *interesting* set and exclude all others. Most use all genes present in the input dataset or even all genes annotated on the genome as the background. Doing so, however, introduces a particular type of bias into the results, which we refer to as sample source bias. Sample source bias occurs when the gene sets returned by GSEA describe the sample rather than the condition being tested. Carefully selecting the background set can eliminate this bias. Although it is arguably an important consideration, surprisingly very few authors have addressed the issue of background selection.

1.7.4. Other analysis

There are numerous other analysis that can be performed using RNA-seq data, according to the type of data, organism, experiments' goals, computational power and time available. Gene-gene and protein-protein interactions (GGIs/PPIs) are two kind of analysis that can help unveil biological processes of the cells. GGI is the modification of the effect of one gene caused by another gene or several other genes (Jiang et al., 2013). GGIs software can report on how and which genes work together, providing a powerful tool for systematically defining gene function and pathways. Transcription of any given eukaryotic gene can be regulated by as many as over a hundred different proteins that act through protein-protein and protein-DNA[18] interactions (Cole et al., 2017). On the other hand, PPIs are high specificity physical contacts established between two or more proteins as a result of biochemical events involving electrostatic forces, hydrogen bonding and the hydrophobic effect (Tripathi et al., 2019). These contacts induce a variety of interactions and associations among the

[18] Essential components of all biological systems, fundamental to almost all biological processes.

proteins, but it's important to keep in mind that proteins that share a functional contact do not necessarily interact directly with each other (Rao et al., 2014; De Las Rivas & Fontanillo, 2010). PPIs are fundamental to the formation of macromolecular structures and enzymatic complexes that are the basis of nearly every biological process ranging from signal transduction and cellular transport to catalyzing metabolic reactions, developmental control, activating or inhibiting other proteins and biomolecular synthesis (Tripathi et al., 2019). For these reasons, finding PPIs is becoming one major objective of system biology. PPIs can be classified in many different ways, according to their structural and functional features, namely interaction surface (homo or heterooligomeric), stability (obligate or nonobligate) and persistence (transient or permanent). A single PPI may be classified by a combination of multiple of these features (Rao et al., 2014).

The ShinyGO (Xijin et al. 2019) is an intuitive, graphical web application that can contribute to gain insights from gene sets, such as a list of DEGs. This method is based on a hypergeometric distribution, followed by FDR correction, and can perform a set of functional analysis, including enrichment analysis, hierarchical clustering tree and networks, and retrieving PPIs from the STRING database (Ge et al., 2020).

Another interesting approach is the analysis of clusters of potentially co-expressed genes. Co-expression clustering aims to group genes together based on similar expression profiles, which may reflect functional similarity, and is typically used to detect genes in novel pathways or networks. By taking an entire expression matrix and computing pairwise co-expression values, it implements comparisons across conditions or time-points, making possible the identification of biologically relevant pathways and networks. Exploring the topology of the generated networks, it can be possible to make inferences on gene co-regulation. One useful tool for study gene co-expression is Clust, which is a fully automated command line tool that performs optimized consensus clustering of well-correlated genes in heterogeneous datasets, helping to understand expression patterns of the DEGs. This tool can simultaneously cluster multiple datasets, enabling the combination of large quantities of public expression data for novel comparative analysis (Abu-Jamous & Kelly, 2018).

1.8. Organisms

1.8.1. Casuarina glauca

Casuarina glauca, commonly known as scaly oak, is a fast-growing multipurpose actinorhizal tree of the Casuarinaceae family in the order Fagales. Native to Australia, *C. glauca* is highly resilient to extreme environmental conditions, such as salinity and drought, having the ability of grow in difficult sites and colonize eroded lands, improving their fertility. As a result, this tree is increasingly used for reforestation and renovation of degraded lands in tropical and subtropical areas, such as China and Egypt (Zhong et al., 2013). This plant can establish root-nodule symbiosis with N_2-fixing bacteria of

the genus Frankia, where N_2 fixation occurs through the action of prokaryotic nitrogenase, which helps improve phosphorous and water uptake by the root system, limiting de necessity of chemical fertilizers. The increased capability of actinorhizal plants to cope with extreme environmental conditions has been attributed to their ability to establish this symbiose. However, studies suggests that increased N nutrition, photosynthesis potential and machinery, proline accumulation as well as the enhancement of the antioxidant status that maintains cellular homeostasis are important factors responsible for salt tolerance (Ngom et al., 2016; Graça et al., 2020). As such, the analysis of the transcriptome of different samples of *C. glauca* under various salt concentrations, can unravel the molecular mechanisms underlying stress tolerance and minimize the impact of its invasive behavior in native biomes, contributing, in complement to previous studies in this species, to halophyte research (Duro et al., 2016; Graça et al., 2019).

1.8.2. Coffea canephora and Coffea arabica

The coffee tree is a shrub that belongs to the Rubiaceae family. Coffee is produced in about 80 countries of the tropical region, where it plays a crucial economic and social role (Ramalho et al., 2014). With an annual provision of around 9M tons of green beans and the involvement of 100-125M people in its extensive chain of value, coffee is one of the most important agricultural products worldwide. Although Coffea genus enclosed at least 125 species, only two dominate coffee market: *Coffea arabica* L. (Arabica coffee) and *Coffea canephora* Pierre ex A. Froehner (Robusta coffee). Together, they are responsible for roughly 99% of the world coffee production, with the former generating around 60-65% (DaMatta et al., 2006). The origin of *C. arabica* is apparently related to a single natural polyploidization event, occurred between the diploids *C. canephora* and *Coffea eugenioides* in a very recent evolutionary time (<50,000 years ago), on the plateaus of Central Ethiopia. Since the two parental species are closely related, the genetic variation between the allotetraploid *C. arabica* and its diploid progenitors is quite small, which is a concern for the sustainability of the crop in the climate change's context (Scalabrin et al., 2020).

Despite those genetic similarities, the two species present different behaviors and adaptations. Arabica coffee grows best at high altitude and is a more expensive bean to grow, due to its longer maturation period and its selective harvesting, resulting in a relatively low yield. On the other hand, Robusta coffee grows well at lower altitudes, which suggests that it is better suited for extreme conditions, such as higher temperatures and consequent potential fungal contamination. Additionally, Robusta has a much higher caffeine content than Arabica, which is said to offer the plant natural insecticidal properties, adding to the sturdy nature of the variety and helping make it better able to withstand the environmental stresses present at low altitudes. Studies suggest that Arabica has evolved to contain lower levels of caffeine, since its bitterness as a defense against insects is not imperative on higher ground. Understanding the effect of extreme temperatures and elevated air CO_2 is crucial for

mitigating the impacts of the coffee industry (Simpson, 2017). As such, the analysis of significant changes in the transcriptomic responses of *C. arabica* and *C. canephora* caused by climate changes can reveal the main biological processes involved in these plant's resistance to abiotic stress responses, improving yields and its underlying economy.

1.8.3. Limonium spp.

The cosmopolitan species-rich genus *Limonium* from the Plumbaginaceae family comprises annual and perennial herbs, shrubs and lianas, often adapted to extreme coastal environments, which has both sexual and apomictic (asexual seed formation) modes of reproduction (Kubitzki, 1993). The *Plumbaginaceae* family is distributed across several parts of the world having preference for cold, arid, saline costal habitats and saline steppes (halophytes). In mainland Portugal there are ca. 17 species, among which are *L. multiflorum*, *L. ovalifolium*, *L. nydeggeri* and *L. dodartii* (Costa, 1998). *L. multiflorum* is an apomict tetraploid species, while *L. ovalifolium* is a related putative sexual diploid species with morphological affinities with *L. nydeggeri*. Studies in *L. ovalifolium* and *L. multiflorum* have showed differences in these species' reproductive strategies (Róis et al., 2012, 2015). *L. dodartii* is a facultative apomictic tetraploid species, which is a trait that gives perennial plants the capability of obtaining stability for the colonization of large areas (Hojsgaard et al., 2014). Due to its ecological and ornamental importance, the analysis of *Limonium* transcriptomes can provide useful insights to the characterization of genetic factors specific to autonomous apomixis and to disclose the molecular mechanisms in controlling the switch from sex to autonomous apomixis.

1.9. Objective

Overall, the work presented in this thesis aimed to detect and functionally annotate significant DEGs, through the application of an RNA-seq pipeline, using suitable tools for each data type and research goals. Moreover, this project intends to contribute to a better understanding of the different expression profiles of the species in investigation, according to each growth conditions, allowing for further studies and integration with other omics. Specifically, the main purposes of this work are:

- Assemble *Casuarina glauca* and *Limonium spp.* transcriptomes to use as reference for the subsequent expression analysis and functional annotation;
- Analyze differential gene expression to characterize relevant mechanisms and adaptations of studied species of *Casuarina*, *Limonium* and *Coffea* genera;
- Improve and optimize data analysis to achieve higher quality results regarding transcriptome assembly, differential gene expression and functional annotation, by choosing the best parameters, tools and software for each dataset.

Chapter 2

2. Methodology

A basic RNA-seq analysis pipeline consists of a few fundamental steps, depending on the existence of a reference genome or transcriptome available for the reviewed organism: raw data processing; *de novo* transcriptome assembly (for non-model species without a reference genome); mapping the reads to the transcriptome or reference genome; gene expression quantification; statistic testing for differential expression (Mutz et al. 2013). This project was conducted as four separate analyses, from three different genera (i.e., *Casuarina*, *Coffea* and *Limonium*), which were analyzed through the application of well documented, efficient and effective bioinformatic methods and tools. The main methods applied to each dataset were the following:

- Integrity and quality control of raw data;
- *De novo* transcriptome assembly of non-model organisms (i.e., *Casuarina* and *Limonium*);
- Alignment/mapping of reads against the:
  - *de novo* transcriptome assembly for *Casuarina* and *Limonium*;
  - a reference genome for *Coffea*;
- Gene expression quantification;
- DEGs detection;
- Functional analysis.

As mentioned before, Illumina RNA-seq raw datasets were obtained from ongoing studies made in collaboration with the CoBiG^2 research group. Each of the datasets was processed equivalently by applying a set of tasks in three separate and individual analyzes, to detect DEGs between plants grown under different environmental conditions or different genotypes of the same plant. Furthermore, DEGs identified in the analysis were functionally analyzed to better understand the mechanisms and pathways involved in the regulation of the most relevant adaptations of each species.

The analyses were started on a MacBook Air with a 1.8 GHz Dual-Core Intel Core i5 process and 4GB 1600 MHz DDR3 and completed on a MacBook Pro with a 2.3 GHz Dual-Core Intel Core i5 processor and 8 GB 2133 MHz LPDD3. Also, some of the most complex and time-consuming analysis were performed using three different FCUL servers (one 8-core server with 173 GB of RAM and two 16-core servers with 62 GB of RAM) running Ubuntu. Most of the analysis on the servers were implemented using external access through VPN. Analysis on FCUL servers were performed using multiple packages of the 4.7.11. version of Miniconda3 for Linux-64, which is a free minimal installer

for conda, with Python version 3.7. The remaining analysis were performed using the fast, free, open-source Visual Studio Code version 1.42.0 code editor and R Studio version 1.2 (RStudio Team, 2019) for macOS, with R version 3.6.0. Figures were edited using the powerful, open-source, free design tool Inkscape version 1.0.1 (Inkscape Project., 2020) for macOs.

2.1. *Casuarina glauca*

RNA-seq specimens of *C. glauca* from plants either nodulated by *Frankia* strain Thr (NOD^+) or supplemented with mineral nitrogen (KNO_3^+) were grown under environmental controlled conditions. Salt stress was gradually imposed by the addition of 200, 400 and 600 mM NaCl concentrations to specimens of each plant. Control plants didn't receive any addition of NaCl, and no replicates were sampled for this analysis. Libraries were prepared with the TruSeq RNA Sample Prep Kit v2 (Illumina, USA) and sequenced in the platform Illumina NovaSeq 6000 (2x125 bp pair-end reads; 30 million reads per sample) at Macrogen (Korea). RNA-seq datasets were acquired through collaboration with Dr. Ana I. Ribeiro-Barros from the Plant-Environment Interactions and Biodiversity Lab (PlantStress & Biodiversity) and Forest Research Centre (CEF) of Instituto Superior de Agronomia (ISA).

The sequencing process generated a total of 8 libraries, whose respective raw fastq files were deposited in the NCBI Sequence Read Archive (SRA), under BioProject SymbSaltStress with accession PRJNA706159 (accessible at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA706159). The raw reads were assessed for integrity, quality and contamination through the application of the following tools: FastQC version 0.11.9 (Andrews, 2010) to analyze reads quality and FastQ Screen version 0.13 (Wingett & Andrews, 2018) to survey putative contaminants. FastQ Screen was instructed to map raw reads against the genomes of the 14 default pre-indexed species and adaptors available on Babraham Bioinformatics website, namely: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Arabidopsis thaliana*, *PhiX174*, Enterobacteria Phage Lambda, UniVec vectors database, Mitochondria RNA, FastQ Screen rRNA custom database and FastQ Screen sequencing adapters database. To perform the screen, 6 threads[1] were allocated to the process, with Bowtie 2 selected as aligner. Due to the large number of datasets, both FastQC and FastQ Screen reports were compiled using MultiQC version 1.10.1 for easier visualization and quicker interpretation of results. Trimmomatic version 0.39.1 (Bolger et al. 2014) was then used to eliminate the remaining adaptors and low-quality or too small reads, using ILLUMINACLIP with *keepBothReads* option, SLIDINGWINDOW:4:15, MAXINFO:36:0.5 and MINLEN:36. After filtering and trimming, Trinity version 2.8.5 (Haas et al. 2013) was used to perform *de novo* transcriptome assembly, combining all *C. glauca* samples to generate one single assembly. This

[1] Number of CPU cores.

software was developed specifically for short reads and is advantageous for non-model plant sequence assemblies (Grabherr et al. 2011). The assembled transcriptome was assessed for completeness through gVolante (Nishimura et al., 2017) online interface with BUSCO v2.0.1 option (Simão et al. 2015). To align the reads against the assembled transcriptome, the sequences were processed with Trinity tool Bowtie2 version 2.3.5 (Langmead and Salzberg 2012) and the aligned reads for each condition were quantified at gene-level expression with RSEM version 1.3.2. (Li and Dewey 2011). The normalized expression of all samples was estimated using the Trimmed Mean of M values (TMM). A Principal Component Analysis (PCA) was performed to survey the relatedness of all samples using the function plotPCA in R studio using R version 3.6.0 (R Core Team, 2019).

To identify DEGs, a combined set of R packages were applied, namely edgeR version 3.26.8, DESeq version 1.36.0 and NOISeq version 2.28.0. The dispersion value in edgeR, which had to be fixed manually due to the absence of replicates, was set to 0.1. In NOISeq, the DE was computed using a stringent threshold of q=0.9, along with the following parameters: pnr=0.2, nss=5, v=0.02. DESeq was used with *method="blind"* and *sharingMode="fit-only"* to allow the analysis without any replicates. To study the effect of salinity, these three tools were used individually to identify DEGs at each salinity level (200 mM, 400 mM and 600 mM NaCl) against the control (0 mM NaCl), for KNO_3^+ and NOD^+ plants, separately. The results were adjusted with the Benjamini and Hochberg's approach for controlling the false discovery rate (FDR) (Benjamini and Hochberg 2000). A filter of FDR < 0.05 and a normalized non-zero $|\log_2 FC| > 2$ were set to define DEGs. The results from edgeR, DESeq and NOISeq were combined to increase the accuracy of the analysis, and only the genes detected as differentially expressed by the three tools were used on downstream analysis. To visualize the resulting expression profiles, volcano plots, heatmaps, barplots and Venn diagrams were plotted for each comparison, using the stats and graphics core R packages in R studio and Python's Matplotlib 3.2.1 library (Caswell et al. 2020) in Visual Studio Code.

The BLAST version 2.9.0 command line application from the NCBI C++ Toolkit was used for functional annotation of DEGs. Through the application of the blastx tool, DEGs were mapped against a local UniProtKB/Swiss-Prot database, filtering the results by maximum E-Value of 1.0E-3 and minimum Identities of 40% (Chen et al. 2017). The resultant annotated proteins were then characterized by Cellular Component (CC), Molecular Function (MF) and Biological Process (BP) Gene Ontology (GO) terms, using the Uniprot and QuickGO APIs to retrieve direct terms and GO term ancestors. Then, Clust version 1.8.10 command line tool was applied to visualize the expression patterns of the detected DEGs and to find co-expressed genes, based on a selected cluster tightness of 5. Then, the results were filtered to keep only the DEGs that were found uniquely in one of the clusters. The resultant lists were searched to find GO terms related to salt stress response. Later, an ORA was implemented by g:GOSt functional profiling tool from gProfiler website, which was applied using g:SCS tailored algorithm that uses a minimum hypergeometric test (Fisher's exact test). ShinyGO version 0.61 webtool, which is also based on hypergeometric distribution followed by FDR correction, was used to retrieve PPIs from the

STRING database. In both tools, *Arabidopsis thaliana* was selected as the organism of interest and separate log_2FC ranked lists of DEGs for each salinity-stress comparison were used as inputs.

2.2. *Coffea canephora* and *Coffea arabica*

Plants from the two main producing species *C. canephora* Pierre ex A. Froehner cv. Conilon Clone 153 (CL153) and *C. arabica* L. cv. Icatu Vermelho (Icatu), were grown in 12-L pots for 1.5 years. Afterwards, plants were transferred into walk-in growth chambers (EHHF 10000, ARALAB, Portugal), and grown for another 10 months in 28-L pots under environmental controlled conditions of temperature (25/20 °C, day/night), relative humidity (70–75%), irradiance (*ca*. 700 µmol m^{-2} s^{-1}), photoperiod (12 h), and air [CO_2] of 380 µmol mol^{-1}. During the whole experiment, the plants were grown in an optimized substrate consisting of a mixture of soil, peat, and sand (3:1:3, v/v/v) and fed on a monthly basis with 5 g of the following fertilizer mixture: 7% $Ca(NO_3)_2$, 5% KNO_3, 7.8% P_2O_5, 17% K_2O, 1.6% MgO, 20% $MgSO_4$, 0,02% H_3BO_3 and 0.01% $ZnSO_4$. To reinforce the N and Ca availability, a complementary fertilization of 2 g was conducted every 3 months with a mixture of 27% NH_4NO_3 and 6% CaO. Both fertilizers were provided as solid spheres that slowly dissolved over successive watering, allowing a gradual release of minerals to the soil/plant. To complement the availability of micronutrients, 500 mL of a solution containing 0.02% Fe-EDTA, 0.01% $CuSO_4$, 0.01% $MnCl_2$, and 0.005% H_2MoO_2, were added on a monthly basis (Ramalho et al., 2013). All RNA-seq datasets were acquired through collaboration with Dr. José C. Ramalho from the Plant-Environment Interactions and Biodiversity Lab (PlantStress & Biodiversity) and Forest Research Centre (CEF) of Instituto Superior de Agronomia (ISA).

2.2.1. CO_2

After the acclimation period, plants from each genotype were grown under either control ambient air [CO_2] of 380 µmol mol^{-1} (aCO_2) or elevated air [CO_2] of 700 µmol mol^{-1} (eCO_2). Total RNA from 3 replicates of the two [CO_2] levels for each genotype was isolated using the RNeasy Plant Mini Kit (Qiagen, Germany). The 12 mRNA libraries were constructed with the Illumina "TruSeq Stranded mRNA Sample Preparation kit" (Illumina, USA) and sequenced separately on the Illumina Hiseq 2000 platform (1x50 bp single-end reads; 28 million reads per sample) at the MGX (Montpellier GenomiX, France, www.mgx.cnrs.fr/). Raw reads were quality-checked using FastQC version 0.11.8 and screened for contaminants using FastQ Screen version 0.13 against the genome of the 14 default pre-indexed FastQ putative contaminant species and adapters. After trimming with Trimmomatic version 0.38, cleaned reads were mapped to the reference genome of *C. canephora* downloaded from the Coffee Genome Hub (http://coffee-genome.org) (Denoed et al. 2014) using STAR version 2.6.1 with default settings. HTSeq-

count version 0.11.0 was used to quantify uniquely mapped reads to each gene, discarding reads in multiple alignments, to avoid the increase of false positives. Relevant parameters used included the default mode *union* and the option *stranded=reverse*. Samtools version 1.9 and gffread version 0.9.9 were used throughout the analysis to convert files and obtain general statistics of the genome mapping. Afterwards, $\log_2$ transformation of Fragments Per Kilobase Of Exon Per Million Fragments Mapped (FPKM+1) and quantile normalization were performed using Cufflinks version 2.2.1. For exploratory analysis, Principal Coordinate Analysis (PCoA) was conducted to verify the relatedness of every normalized replicate, using the function prcomp in R software version 3.5.1.

DESeq2 version 3.8 was employed to identify differentially expressed genes (DEGs) reflecting the effect of eCO_2 (eCO_2 vs aCO_2; aCO_2 as control) and differences between the two genotypes (CL153 vs. Icatu; Icatu as control). The resulting values were adjusted using the Benjamini and Hochberg's approach for controlling the FDR. Genes with a normalized non-zero $\log_2$ FC and an FDR < 0.01 were defined as differentially expressed. Functional annotations for the protein coding genes in the *C. canephora* genome were downloaded from the Coffee Genome Hub. Because a very high number of reads were mapped to the *C. canephora* genome it was further used as the reference genome for the analyses. BLAST2GO version 1.4.4 was used for mapping and functional annotation of DEGs (parameters: E-Value-Hit-Filter 1.0E−6, Annotation Cutoff 55, GO Weight 5, Hsp-Hit Coverage Cutoff 20). After mapping, DEGs were filtered to exclude multiple hits to the same gene, keeping only the one that showed the highest identity percentage. The genes were characterized using GO terms of Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). A local BLAST database was built to map DEGs to the highest identity Uniprot gene hits. GO mapping and GO annotation was then performed with Blast2GO command line interface using the Uniprot genes and the local database.

The raw reads had already been subjected to QC, alignment to the reference genome, DE analysis and functional annotation, which results were obtained in the form of CSV files containing full lists of annotated DEGs. First, Venn diagrams were plotted using VennDiagram version 1.6.20 R package to show the overlap of common DEGs between different comparisons and highlight DEGs specific to the effect of eCO_2 and to different responses of each genotype. In order to obtain a full set of GO terms associated with each DEG, all GO ancestors of each identified term were retrieved using QuickGO API. To relate the transcriptomic answer of the two coffee genotypes with their physiological and biochemical responses, DEGs were searched for the GO terms referenced in Scalabrin et al. (2020) (i.e., photosynthesis, chlorophyll metabolic process, ribulose-1,5-bisphosphate carboxylase/oxygenase, antioxidant activity, cellular respiration, malate dehydrogenase activity, and pyruvate kinase activity) and FAD and LOX-related proteins, which have been reported as the most important for lipid profile dynamics related to stress acclimation in coffee plants. This search was performed using the QuickGO API, identifying all direct and descendant GO terms.

Then, a GSEA was performed using WebGestalt webtool, with a range of 5 to 2000 genes for category, a Benjamini & Hochberg (BH) multiple test adjustment and a "TOP" significance level of 10,

which ranks categories based on FDR and them selects the top 10 most significant ones. Ranked lists of DEGs, sorted by descending log_2FC were used as input and mapped against the *Arabidopsis thaliana* functional database. GO terms, KEGG and Wikipathways with an FDR < 0.05 were considered enriched. Results were plotted using the R ggplot2 version 3.3.2 library.

Additionally, the libraries (raw fastq files) generated by the sequencing process were deposited in the NCBI SRA, under BioProject CoffeeOmics Climate (PTDC/ASP-AGR/31257/2017), with accession PRJNA606444 (accessible at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA606444).

2.2.2. CO_2 + Temperature

After acclimation, temperature was raised at a rate of 0.5 °C day^{-1} (diurnal temperature) from 25/20 °C to 42/30 °C, with 7 days of stabilization at 31/25, 37/28, and 42/30 °C. The individual and combined effects of increased temperature and [CO2] were studied for a moderate supra-optimal temperature of 37/28 °C (37 °C) and an extreme supra-optimal temperature of 42/30 °C (42 °C), in comparison to the control temperature of 25/20 °C (25 °C), at either aCO_2 or eCO_2 conditions, for each genotype.

Total RNA from 3 replicates of the three temperatures (25ºC, 37ºC and 42ºC), each under one of the two [CO_2] levels (aCO_2 and eCO_2) for each genotype was isolated using the RNeasy Plant Mini Kit (Qiagen, Germany) according to the manufacturer's instructions (two genotypes × two CO_2 treatments × three temperatures × three biological replicates). The 36 messenger RNA (mRNA) libraries were constructed with the Illumina "TruSeq Stranded mRNA Sample Preparation kit" (Illumina, San Diego, CA) and sequenced separately on Illumina Hiseq 2000 platform (1x50 bp single-end reads; 30 million reads per sample) at the MGX platform (Montpellier GenomiX, France, www.mgx.cnrs.fr/).

High-quality reads were obtained after several steps of quality checks which included trimming, removal of adaptor/primer and low-quality reads using FastQC version 0.11.8 and Trimmomatic version 0.38 through the trimming steps: ILLUMINACLIP to cut adaptors, SLIDINGWINDOW:4:15 to trim low-quality reads and MINLEN:38 to drop small reads. FastQ Screen version 0.13 was used to check for contaminants against the genome of the most common model organisms (*e.g.*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Escherichia coli*) and adapter databases (*e.g.*, Mitochondria RNA, PhiX, Vector from UniVec database, FastQ Screen rRNA custom database and FastQ Screen Adapters database).

The filtered high-quality reads were mapped to the reference genome of *C. canephora* genome downloaded from the Coffee Genome Hub using STAR version 2.6.1, with default settings. Htseq-count version 0.11.0 was used with default mode *union* and the option *stranded=reverse* to count only uniquely mapped reads to each gene, discarding reads in multiple alignments and thus avoiding the increase of false positives. Samtools version 1.9 and gffread version 0.9.9 were used throughout the analysis to obtain general statistics of the genome mapping. Principal Coordinate Analysis (PCoA) was

performed on the expression data of genes, FKPM normalized and log_{10}-transformed, using the function prcomp in R software version 3.5.1.

The raw reads had already been subjected to QC, alignment to a reference genome, quantification and normalization, which results were obtained in the form of Tab-delimited files containing gene-level quantifications ready for DE analysis. Gene expression normalization of all the samples was estimated in FPKM. The changes in the relative abundance of the genes between the different genotypes/CO_2-treatments/temperature-treatments were estimated using DESeq2 v1.28.1 and edgeR v3.30.3. Only the DEGs identified by both tools as differentially expressed significantly were used in subsequent analyses. The resulting values were adjusted using the Benjamini and Hochberg's approach for controlling the FDR. Genes with a normalized non-zero log_2 fold change expression and an FDR <0.01 were defined as differentially expressed. Python's matplotlib library was used to plot Venn diagrams and barplots.

DEGs were annotated following the functional annotation of the reference genome, *C. canephora* downloaded from the Coffee Genome Hub. GO enrichment analyses were applied to understand the functional classification of temperature-responsive DEGs through an ORA, using gProfiler under FDR < 0.01. Results were summarized using REVIGO by removing redundant GO terms with allowed similarity=0.5. Enrichment non-redundant results were plotted using the R ggplot2 version 3.3.2 library. This same package was used to plot a heatmap with dendrograms to visualize DEGs based on the differential expression patters between the different treatments. To prevent highly differentially expressed genes from clustering together without considering their expression pattern, log_2 fold change was scaled by gene across treatments (row Z-score).

Additionally, the libraries (raw fastq files) generated by the sequencing process were deposited in the NCBI SRA, under BioProject CoffeeOmics Climate (PTDC/ASP-AGR/31257/2017) II, under accession PRJNA630692 (accessible at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA630692).

2.3. *Limonium* spp.

All RNA-seq datasets (18 libraries) were obtained through collaboration with Dr. Ana D. Caperta from the Linking Landscape, Environment, Agriculture and Food (LEAF) of Instituto Superior de Agronomia (ISA).

Quality control of the raw reads, including contaminants survey, was performed using FastQC version 0.11.9 (Andrews 2010) and FastQ Screen version 0.14.0 (Wingett and Andrews 2018), ran against the genome of its default pre-indexed species and adaptors. Then, since all raw reads presented a quality base score over 36, Trimmomatic version 0.39 (Bolger et al. 2014) was only used to eliminate adaptors and filter reads of length bellow 36 base pairs (bp). A *de novo* transcriptome assembly was performed using Trinity version 2.11.0 (Grabherr et al. 2011), in which cleaned reads from all samples were combined to generate one global assembly, since this software has shown a consistent performance

and has a high read alignment rate (Wang and Gribskov 2016). The assembly was assessed for completeness using BUSCO version 5 (Waterhouse et al. 2017), through gVolante2 (Nishimura et al., 2017). After alignment against the transcriptome using Bowtie2 aligner version 2.3.5 (Langmead and Salzberg 2012), sequences were quantified at gene-level expression with RSEM version 1.3.3 (Li and Dewey 2011), through Trinity pipeline. A Principal Component Analysis (PCA) was performed to survey the relatedness of normalized gene counts using the function plotPCA in R studio version 4.0.2 (R Core Team, 2020). Then, a heatmap with dendograms was plotted with R ggplot2 version 3.3.2 library (Wickham, 2016) using the same counts to cluster samples based on their expression profiles similarity.

To study significant differences between sexual and apomictic plants, differential expression analysis was performed with edgeR version 3.30.3 (Robinson et al., 2010), which is a flexible empirical Bayes approach that uses weighted likelihood methods to estimate gene-specific variation even with very few or no replicates (Chen et al., 2014). DEGs were searched in apomictic plants (*L. multiflorum*) in S1 relative to sexual plants (*L. nydeggeri* and *L. ovalifolium*) in both S1 and S2. Also, differential expression analysis was performed in apomictic plants in S2 relative to sexual plants in S3/S4 and to facultative apomictic plants (*L. dodartii*) in S4. Apomictic plants were set as the samples to test, while sexual and facultative apomictic plants were set as the controls, according to each comparison. Genes with a normalized $|\log_2$ fold change ($\log_2$ FC)$| > 2$ were defined as differentially expressed and used in the downstream analysis. In the comparison between apomictic and facultative plants, in which all samples have at least 3 replicates, DEGs were also filtered by $p < 0.01$. Venn diagrams were used to plot DEGs overlapping between different comparisons, through matplotlib version 3.3.3 (Caswell et al. 2020) in Python version 3.9.0 (Python Software Foundation 2020).

Functional annotation of DEGs was performed with Basic Local Alignment Search Tool (BLAST) version 2.10.1 command line tool from the NCBI C++ Toolkit (National Center for Biotechnology Information 2020). Blastx was used to map DEGs to *Arabidopsis thaliana* homologs, against a local Swissprot database, filtering gene hits by maximum E-Value of 1.0E-3 and minimum Identities of 40% (Chen et al., 2017). Then, to avoid duplicated results, DEGs annotated to the same *A. thaliana* homolog were filtered by identities and sequence length, keeping the transcripts with the highest values. Uniquely annotated DEGs were characterized with GO terms, using UniprotKB website REST API (The Uniprot Consortium 2019). GO enrichment analyses were applied to $\log_2$ FC ordered lists of DEGs through an Over-representation analysis (ORA), using the g:GOSt *functional profiling* tool from gProfiler website (Raudvere et al. 2019), with the g:SCS tailored algorithm under FDR<0.01. Enrichment results were summarized using REVIGO (Supek et al., 2011) through the remotion of redundant GO terms with allowed similarity=0.5 and then plotted with the R ggplot2 version 3.3.2 library.

After DE analysis, gene expression of DEGs was analyzed to find DEGs that were knocked out either in the test sample or the control, selecting the genes that had no expression in only one of those

samples. Also, DEGs shared by more than one comparison were searched for opposite regulation. KEGG and WikiPathways enrichment analysis was then performed with gProfiler in the results from both methods to find relevant metabolic pathways.

According to literature, TFs can be involved in plants fertility. Also, studies have suggested that the following genes can be responsible for male sterility: ROS1, DMC1, MS2, pop1 and 4CLL1. As such, these genes, along with a list of *A. thaliana* TFs retrieved from the Plant Transcription factor & Protein Kinase Identifier and Classifier (iTAK) (Zheng et al., 2016), were searched among DEGs to find if they were downregulated or repressed in apomictic plants. Furthermore, annotated lists of DEGs were searched for GO terms related to pollen tube, such as the biological processes pollen tube reception (GO:0010483), pollen tube development (GO:0048868), pollen tube growth (GO:0009860), regulation of pollen tube growth (GO:0080092) and pollen tube adhesion (GO:0009865), and the cellular components pollen tube (GO:0090406) and pollen tube tip (GO:0090404).

Additionally, the libraries (raw fastq files) generated by the sequencing process were deposited in the NCBI SRA, under BioProject Sexual and apomictic regulation in Limonium spp., under accession and PRJNA752506 (accessible at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA752506).

Chapter 3

3. Results

The main goal of my master's thesis is to apply a state-of-the-art workflow for the analysis of transcriptomic datasets from different plants samples, with special focus on the observation of samples subjected to abiotic stress-inducing factors, compared to control conditions. After revising the current literature, a set of tools was chosen to perform the analysis, aiming to obtain high quality results with useful biological meaning. Additionally, a set of scripts were developed to speed up and automatize the analysis, using Python and R languages. These scripts are publicly available at https://github.com/ziisabel/CoBiG2/tree/cobig2 and can be used to perform similar analysis. The `mainFunctions.py` script was developed to be called by other scripts, since it includes a set of functions that are needed in multiple of them, contributing to reduce code redundancy. The applied workflow initially involves a preprocessing step to remove adaptors, contaminants, low quality bases and sequencing errors. For this step, I developed 2 Python scripts, namely `fastqAnalysis.py`, which allows users to apply FastQ Screen, FastQC and/or MultiQC to assess and visualize the quality of multiple sequencing files and filter contaminants, and `runTrimmomatic.py`, which runs Trimmomatic for multiple files simultaneously, from either single-end or paired-end reads.

Afterwards, for the organisms that don't have a quality reference genome available, I developed the Python script `trinityMultiple.py`, which performs *the novo* assembly of multiple sequencing files with Trinity, according to the users input parameters. To prepare the reference genome for alignment, sometimes it's necessary to make some amendments to the respective file, such as convert a multiple-line into a single-line per gene FASTA file. To achieve this, I developed the script `clean_fasta.py`, which was projected to work with the *Coffee arabica* reference genome from RefSeq, GCF_003713225.1_Cara_1.0_cds_from_genomic.fna (available online at https://www.ncbi.nlm.nih.gov/assembly/GCF_003713225.1). This same file can be used to extract gene and protein names using the Python script `get_Carabica_annotations.py`. Additionally, I developed the Python script `write_annotation_genes.py`, which extracts ORF and protein names from GFF3 annotation files, searching UniprotKB to complete that information with gene name and related GO terms.

The next step is to align the reads with a reference genome or assembled transcriptome, for which I developed the Python script `STARmultiple.py` to perform multiple sequencing files with STAR, also from either single-end or paired-end reads. To visualize samples and inspect for outlier, I developed the R script `pca_DESeq2.R`, which uses the mean of ratios normalization method from DESeq2 package, with either *rlog* or *vst* data transformation, to normalize the gene counts raw data and plot the respective Principal Component Analysis (PCA). Subsequently, the differential expression analysis was performed with DESeq2, edgeR and NOISeq, through the development of the following scripts: `DE_analysis_DESeq2.R`, `DE_analysis_edgeR.R` and `DE_analysis_NOISeq.R`. Also, to overlap the results from DESeq2 and edgeR for the same data, I developed the `DE_analysis_overlap_DEGs.py`.

After differential analysis, the resultant lists of DEGs can be used for multiple purposes. To visualize DEGs based on their differential expression patterns between different treatments relative to the control, I developed the R script `heatmap.R`. This script includes the scale of the $\log_2$ fold change by gene across treatments (row Z-score) to prevent highly differentially expressed genes from clustering together without considering their expression pattern. To easily organize the generated information in table format, I develop the Python script `get_DEGs_tables.py`, which writes DEGs annotations in a tab-delimited text file (.TAB), including gene name, protein name and GO terms. Additionally, to filter redundant genes, with the exact same annotations, I developed the Python scripts `get_unique_DEGs.py`, which filters DEGs based on the most similar homologs identified through blastx (higher identities percentage) and `get_TAIRs_longer_transcript.py`, which filters DEGs based on the longer transcript among genes blasted to the same homolog gene. Another way to annotate DEGs, especially when there isn't an annotated reference genome available, is to use the BLAST tool to find homologs genes from related species. To perform this task, I developed the Python scripts `runBlastx.py`, which finds DEGs homologs through blastx, and `write_annotations_DEGs.py`, which writes tab-delimited text files (.TAB) with UniProtKB annotations searched with the respective API, using the blastx results. This script is specially optimized to work with *Arabidopsis thaliana* homologs. Other useful way to interpret the biological meaning of DEGs is to study the transcription factors (TFs) among them. As such, I developed the Python script `get_TairsTFfromiTAK.py`, which generates a tab-delimited text file (.TAB) associating DEGs and their regulation type to the respective TF family names from their *A. thaliana* homologs, retrieved from the iTAK (Plant Transcription factor & Protein Kinase Identifier and Classifier) database. Moreover, it can be useful to known which DEGs are totally knocked out (KO) in one of the samples (control or test), for which I developed the Python script `get_KOgenes.py`, which creates multiple tab-delimited text files (.TAB) with lists of KO genes in a sample, relative to its test or control. Additionally, I developed the Python script `get_oppositeRegulationDegs.py`, which creates

two tab-delimited text files (.TAB), one for up and another for down-regulated DEGs, with lists of DEGs shared between two comparisons presenting opposite regulation.

The GO enrichment analysis is another useful resource to assign biological meaning to this type of data. To optimize and speed-up that process, I developed 5 scripts (4 Python + 1 R) dedicated to the over-representation analysis (ORA) with gProfiler and to the GO terms redundancy reduction with REVIGO. The `generate_GMT.py` script allows the user to generate personalized GMT files to use as input background in gProfiler, which is very important to obtain meaningful results, that accounts for technologic, detection and biological bias. The `get_gProfiler_input.py` script creates multiple TAB files that can be used as input in gProfiler to perform ORA. After running the analysis, the CSV files generated from gProfiler can be used to feed the `clean_gProfiler.py` script, which reformats them for easier readability and data extraction, either for plotting or further manipulation with REVIGO. This tool can be used when the list of GO enriched terms is very long to summarize the results by eliminating redundant terms within a defined range of similarity between terms. Then, the CSV files generated in that process, can be edited with the `clean_revigo.py` script for easier visualization and direct plotting with the `plot_enrichment_gProfiler.R` script. The results are documented via graphical output and tables.

3.1. *Casuarina glauca*

Quality control generated an average of ca. 25 M (71%) clean reads from ca. 35 M raw reads in NOD^+ plants and an average of ca. 26 M clean reads from ca. 36 M raw reads in KNO_3^+ plants (Figure 3.1; Table 3.1). The minimum per base sequence quality was improved from 2 in raw reads to 31 in clean reads (Figure 3.2). The GC content was in average 46% pre-processing and 45% post-processing. According to the set parameters, minimum sequence length was reduced to 36 bp after trimming, maintaining the maximum of 125 bp per clean read. As expected in this type of RNA-seq data, the first 12-13 bp of all samples failed the per base sequence content report due to library preparation bias. The percentage of duplicated reads ranged from 60% to 70% and the adapter content reached a local maximum of 20% in the 3' end of reads.

Table 3.1. Sequencing data from *C. glauca* NOD^+ and KNO_3^+ samples, grown in different salinity stresses (200 mM, 400 mM and 600 mM NaCl), plus the control (0 mM NaCl). Raw reads, obtained after sequencing, generated clean reads after submission to quality control with FastQC and Trimmomatic software [Min. quality: minimum sequence per base quality].

Plant-type	[NaCl] mM	Strand	Raw Reads			Clean Reads			
			Total Sequences	% GC	Min. quality	Total Sequences	%	% GC	Min. quality
NOD^+	0	forward	37010538	46	28	26559076	72	46	33
		reverse	37010538	46	15	26559076	72	46	31
	200	forward	33417448	45	28	25496114	76	45	33
		reverse	33417448	45	15	25496114	76	45	31
	400	forward	33716971	45	28	22905738	68	45	33
		reverse	33716971	45	15	22905738	68	45	31
	600	forward	33881260	46	27	24012917	71	45	33
		reverse	33881260	46	15	24012917	71	45	31
	Average		**34506554**	**46**	**21**	**24743461**	**72**	**45**	**32**
KNO_3^+	0	forward	37089848	46	28	27361402	74	45	33
		reverse	37089848	46	2	27361402	74	46	31
	200	forward	38696059	46	28	27277745	71	45	33
		reverse	38696059	46	15	27277745	71	46	31
	400	forward	33546743	45	28	24471108	73	45	33
		reverse	33546743	45	15	24471108	73	45	31
	600	forward	35472156	46	28	26000474	73	45	33
		reverse	35472156	46	15	26000474	73	46	31
	Average		**36201202**	**46**	**20**	**26277682**	**73**	**45**	**32**

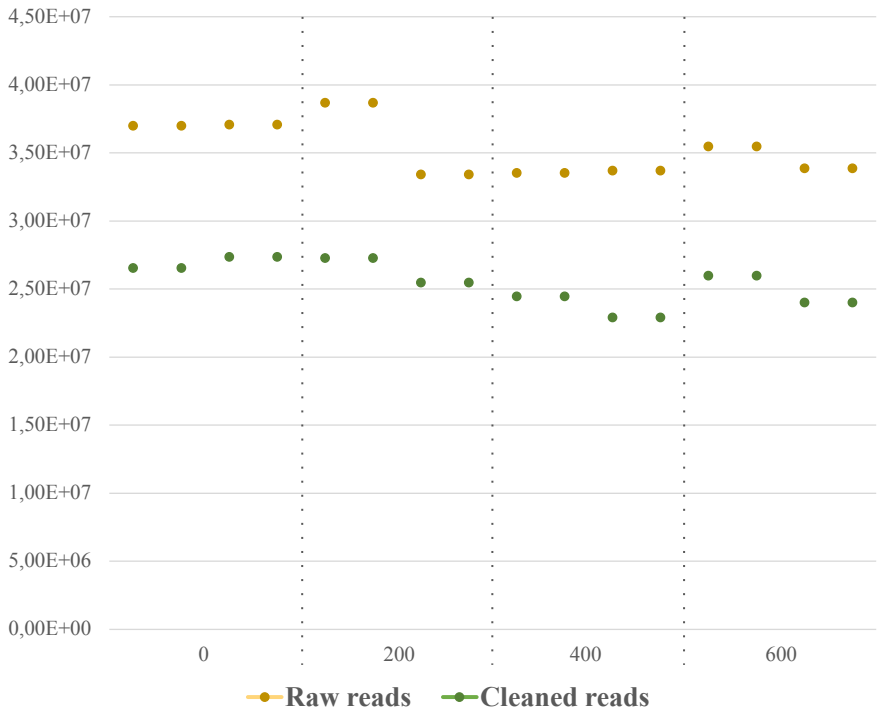

Figure 3.1. Total number of reads per sample, before and after processing with Trimmomatic, according to FastQC. Plot created with Excel and edited in Inkscape.

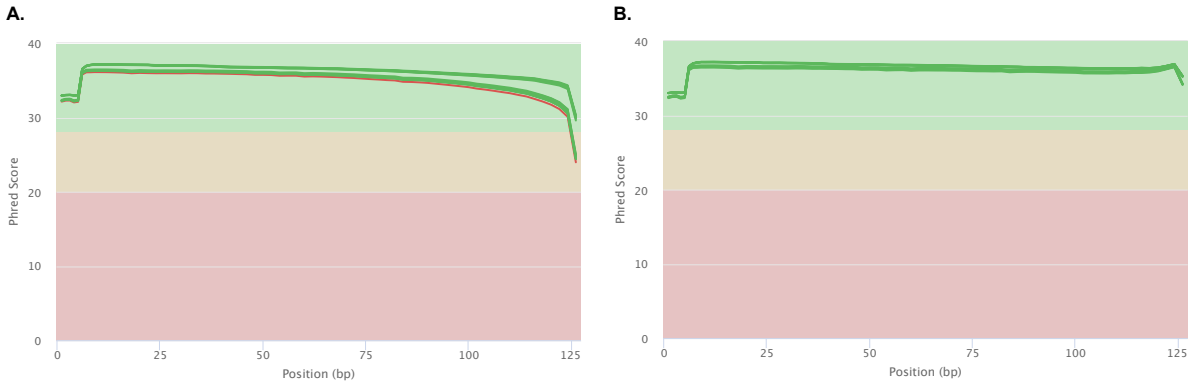

Figure 3.2. Per base sequence quality of reads, according to FastQC. Plot created with MultiQC and edited in Inkscape. (A) Raw reads. (B) Clean reads post-processing with Trimmomatic.

Only a small number of raw reads of each sample mapped to the surveyed sequences in FastQ Screen, with < 3% mapping uniquely to *Arabidopsis thaliana*, < 2% to adaptors and < 0.1% to other sequences (Figure 3.3). Since these values were particularly low and this analysis is focused on short reads of a non-model species, this result was expected, and all trimmed reads were used in the assembly to avoid losing important information.

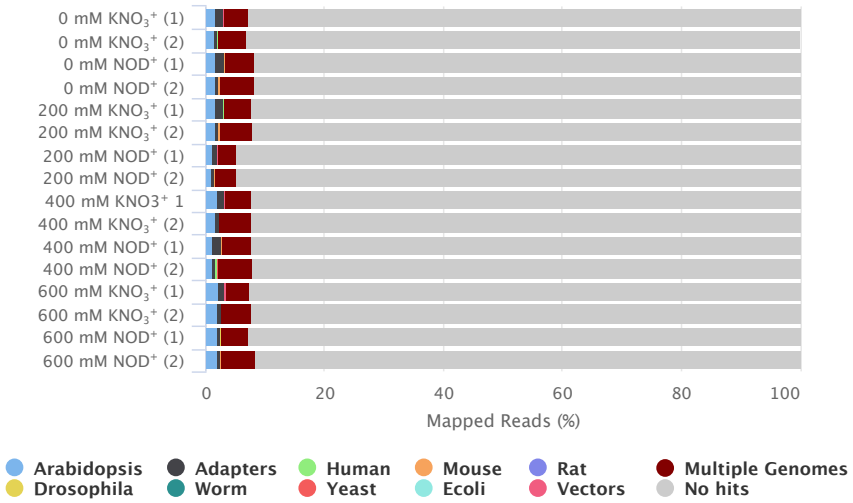

Figure 3.3. Percentage of mapped raw reads of *C. glauca* KNO_3^+ and NOD^+ plants to main contaminants, according to FastQ Screen. Plot created by MultiQC and edited on Inkscape [(1) forward read; (2) reverse read].

The *de novo* assembly showed 41% GC content within a total of 181,484 contigs, 86,202 unigenes and a contig N50 size of 2792. More than 96% of the reads were mapped back to the transcriptome, which had almost 95% completeness, indicating that a high-quality transcriptome assembly has been generated for downstream analyses. The detailed list of basic metrics approaching composition, alignment and completeness quality of the assembly can be seen in Figure 3.4 and in Table 3.2.

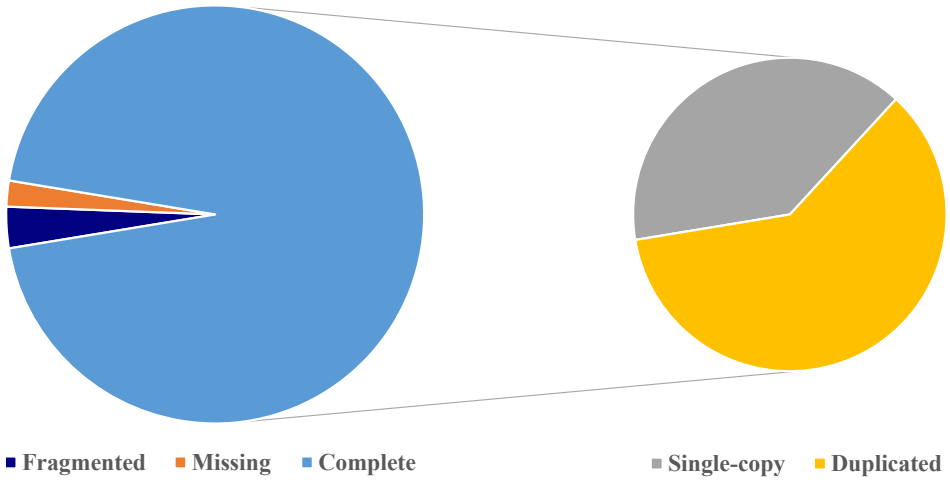

Figure 3.4. Completeness quality of the *novo* assembly of *Limonium* transcriptome, according to BUSCO database, using gVolante.

Table 3.2. Basic metrics of composition, alignment and completeness quality of the *de novo* transcriptome assembly of *C. glauca* NOD^+ and KNO_3^+, grown at 0 mM, 200 mM, 400 mM and 600 mM NaCl.

Completeness assessment	**Values**
Total number of core genes queried	1440
Number of complete core genes detected	1366 **(94.86%)**
Number of missing core genes	28 (1.94%)
Average number of orthologs per core genes	1.92
% of detected core genes that have more than 1 ortholog	60.54
Scores in BUSCO format	C: 94.8% [S: 37.4%, D: 57.4%], F: 3.2%, M: 2.0%
Length statistics and composition	
Total assembled contigs	**181484**
Total assembled genes	**86202**
Total length (nt)	289780263
Longest contig (nt)	17867
Shortest contig (nt)	183
Mean contig length (nt)	1597
Median contig length (nt)	1036
# Contig sequences with ORF	76648
% Mean ORF	40
N90	727
N50 sequence length (nt)	**2792**
L50 sequence count	34704
N10	5729
Number of sequences > 1K (nt)	92205 (50.8%)
Number of sequences > 10K (nt)	170 (0.1%)
Base composition (%)	A: 30.98; T: 27.94; G: 19.24; C:21.85
N	0
GC-content (%)	**41.09**
Number of non-ACGTN (nt)	0
Alignment statistics	
Total reads (PE)	204084574
Total aligned reads	**196275145 (96.2%)**
Concordantly unaligned and discordantly aligned reads	435225 (0.2%)
Total aligned mates	**11388138 (2.8%)**
Overall alignment rate	**99.2%**

Principal Component Analyses (PCA) showed a distinct separation of samples in four different quadrants: (i) 0 mM and 200 mM NaCl KNO_3^+; (ii) 0 mM and 200 mM NaCl NOD^+; (iii) 400 mM and 600 mM NaCl KNO_3^+; and (iv) 400 mM and 600 mM NaCl NOD^+. PC1 accounted for 86% of the total variance, with a clear division between the samples with lower (0 and 200 mM) and higher (400 and 600 mM) [NaCl], while PC2, comprising 9% of the variance, distinguished KNO_3^+ from NOD^+ plants (Figure 3.5).

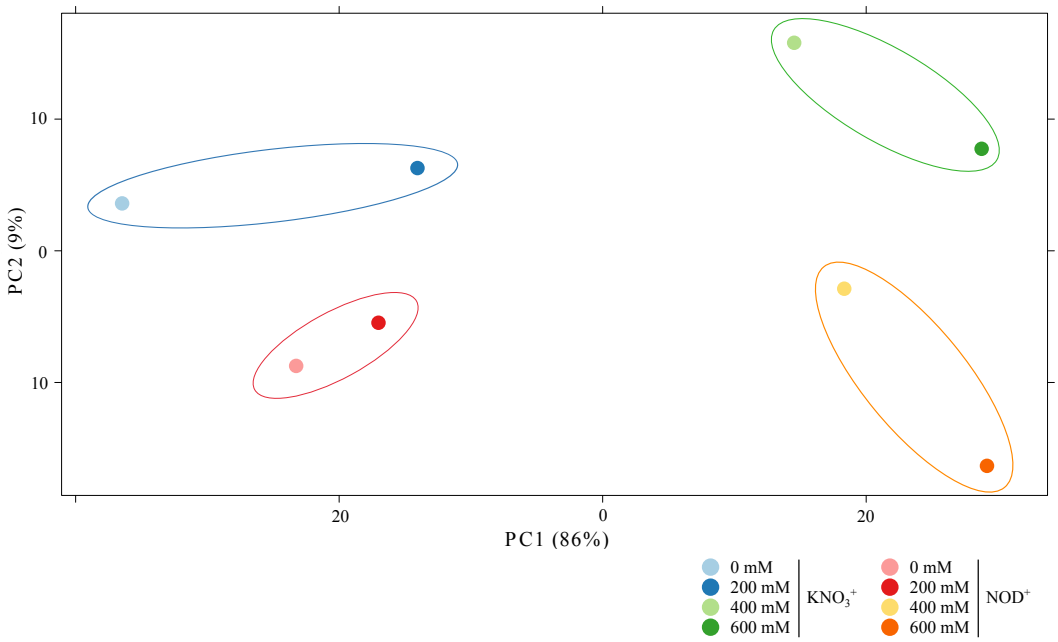

Figure 3.5. PCA of gene expression counts from *C. glauca* NOD^+ and KNO_3^+ samples, grown at control (0 mM NaCl) and salinity-stressed conditions (200 mM, 400 mM and 600 mM NaCl).

The KNO_3^+ plants expressed a total of 19,913 genes at 0 mM NaCl and an average of 19,645 genes among salinity-stressed samples (Table 3.3.). Overall, the number of DEGs increased with increasing salinity, although the percentage of significant DEGs between each salinity condition and the control was extremely low, ranging from 0.04% to 2% with increasing salinity (Figure 3.6). In the same plants, the totality of DEGs were downregulated at 200 mM NaCl and 88% were down-regulated at 400 mM NaCl and 600 mM NaCl. The NOD^+ plants expressed a slightly higher number of genes compared to KNO_3^+, with a total of 20,278 at 0 mM NaCl and an average of 19,780 among salinity-stressed samples. Similar to KNO_3^+ plants, a decreasing number of common genes and an increasing number of DEGs were observed with increasing salinity. However, the percentage of significant DEGs at all salinity conditions was slightly higher in KNO_3^+ plants, varying only from 0.03% to 1%, with increasing salinity. In these plants, the majority of DEGs were also down-regulated, ranging from 60% to 85% with increasing salinity, with the exception of 200 mM NaCl, where DEGs were equally distributed between the two types of regulation.

Table 3.3. DE quantification in *C. glauca* NOD^+ and KNO_3^+, at salinity-stressed conditions relative to control (0 mM NaCl). [200: 200 mM vs. 0 mM, 400: 400 mM vs 0 mM and 600: 600 mM vs. 0 mM].

Plant-type	[NaCl] mM	Expressed Genes			Detected DEGs			Overlapping DEGs		
		Stressed	Control	Common (%)	DESeq	NOISeq	edgeR	Total (%)	Up	Down
KNO_3^+	200	20765	19913	15928 (78%)	58	330	23	**9 (0.04%)**	0	9
	400	19078	19913	15079 (77%)	524	1213	381	**238 (1%)**	28	210
	600	19092	19913	14976 (77%)	650	1450	612	**359 (2%)**	45	314
	Average	**19645**	**19913**	**15328 (77%)**						
NOD^+	200	19397	20278	16386 (83%)	44	204	20	**6 (0.03%)**	3	3
	400	20470	20278	16137 (79%)	359	962	184	**104 (0.5%)**	42	62
	600	19472	20278	15547 (78%)	548	1261	373	**254 (1%)**	37	217
	Average	**19789**	**20278**	**16023 (80%)**						

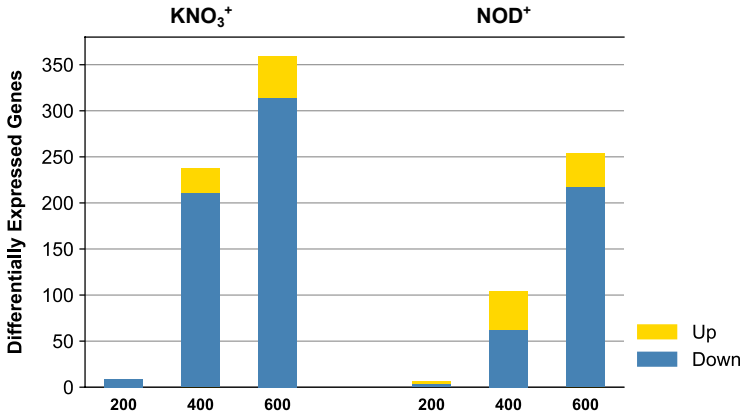

Figure 3.6. Total number of DEGs in *C. glauca* plants NOD^+ and KNO_3^+. [200: 200 mM vs. 0 mM NaCl, 400: 400 mM vs 0 mM NaCl and 600: 600 mM vs. 0 mM NaCl].

The DEGs were searched for salt-treatment specificity in order to find which genes were differentially expressed in only one of the treatment conditions (200mM, 400mM, or 600 mM NaCl) relative to control. Not surprisingly, the number of treatment-specific DEGs increased with the salt concentration in both plant groups (Figure 3.7). Also, the combined number of DEGs specific to 400 mM or 600 mM NaCl, relative to control, accounted for the majority of all treatment-specific DEGs in both plants, being around 66% in KNO_3^+ and 85% in NOD^+ plants. Moreover, the number of DEGs shared by 200 mM and 400 mM NaCl was always smaller than the ones shared by 400 mM and 600 mM NaCl.

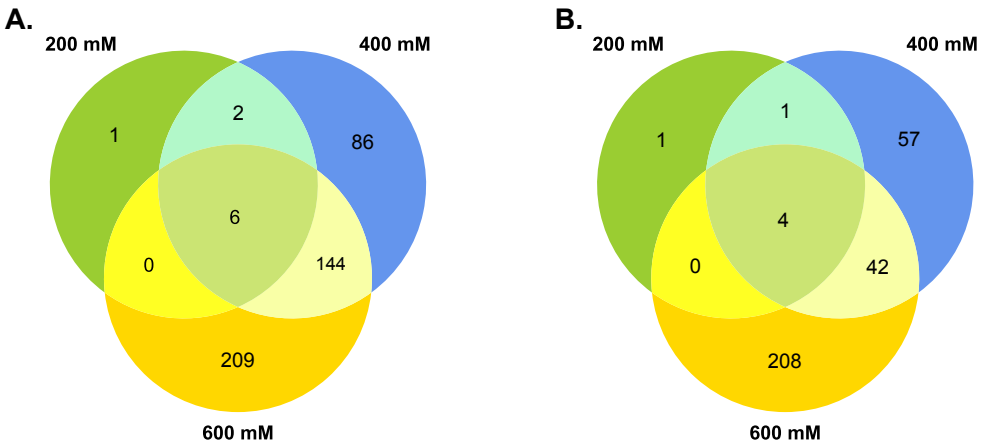

Figure 3.7. Treatment-specific and overlapping DEGs in *C. glauca* KNO_3^+ and NOD^+ plants, at 200 mM, 400 mM and 600 mM NaCl, relative to control. (A) KNO3$^+$ (B) NOD^+. [200 mM: 200 mM vs. 0 mM, 400 mM: 400 mM vs 0 mM and 600 mM: 600 mM vs. 0 mM].

Comparing KNO_3^+ and NOD^+ plants, no DEGs were detected in both plants at 200 mM NaCl, while 32 (10%) and 135 (28%) were shared at 400 and 600 mM NaCl, respectively (Figure 3.8). Regardless salinity concentration, the number of plant-specific DEGs was always higher in KNO_3^+ than in NOD^+.

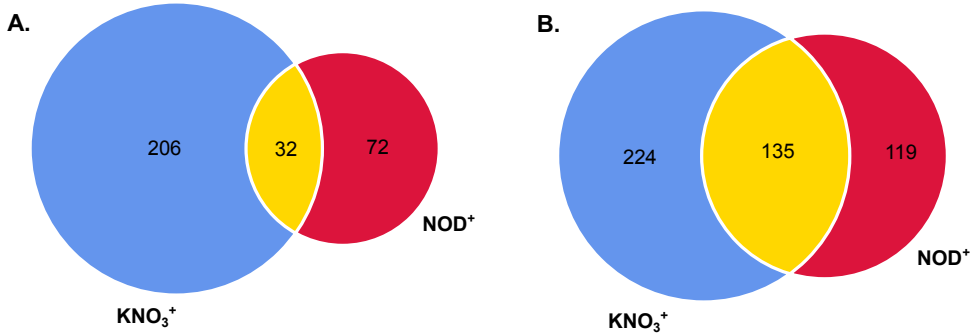

Figure 3.8. Treatment-specific and overlapping DEGs of *C. glauca* of either KNO_3^+ or NOD^+ plant-type, under different stress-salinity conditions. (A) 400mM vs 0 mM NaCl; (B) 600mM NaCl vs 0 mM.

After functional annotation, less than half of all DEGs were uniquely mapped to UniProtKB/Swiss-Prot database proteins. The differential expression values of annotated DEGs were only slightly higher in NOD^+ plants, in which $\log_2$ FC ranged from -7.8 to 5.1, comparatively to KNO_3^+ which varied from -6.9 to 4.9. In KNO_3^+ plants, the only annotated DEGs at 200 mM NaCl were ERF020 and GT-3B, which are related to stress signal and response to salt, respectively.

When comparing differential expression across conditions, the dendogram of the heatmap grouped samples into two main clusters, one with all samples of NOD^+ plus KNO_3^+ at 200 mM, and another with the remainder KNO_3^+ samples. In the first one, the two samples grown at 200 mM were closer together, followed by the 400 mM sample and then by the 600 mM sample (Figure 3.9).

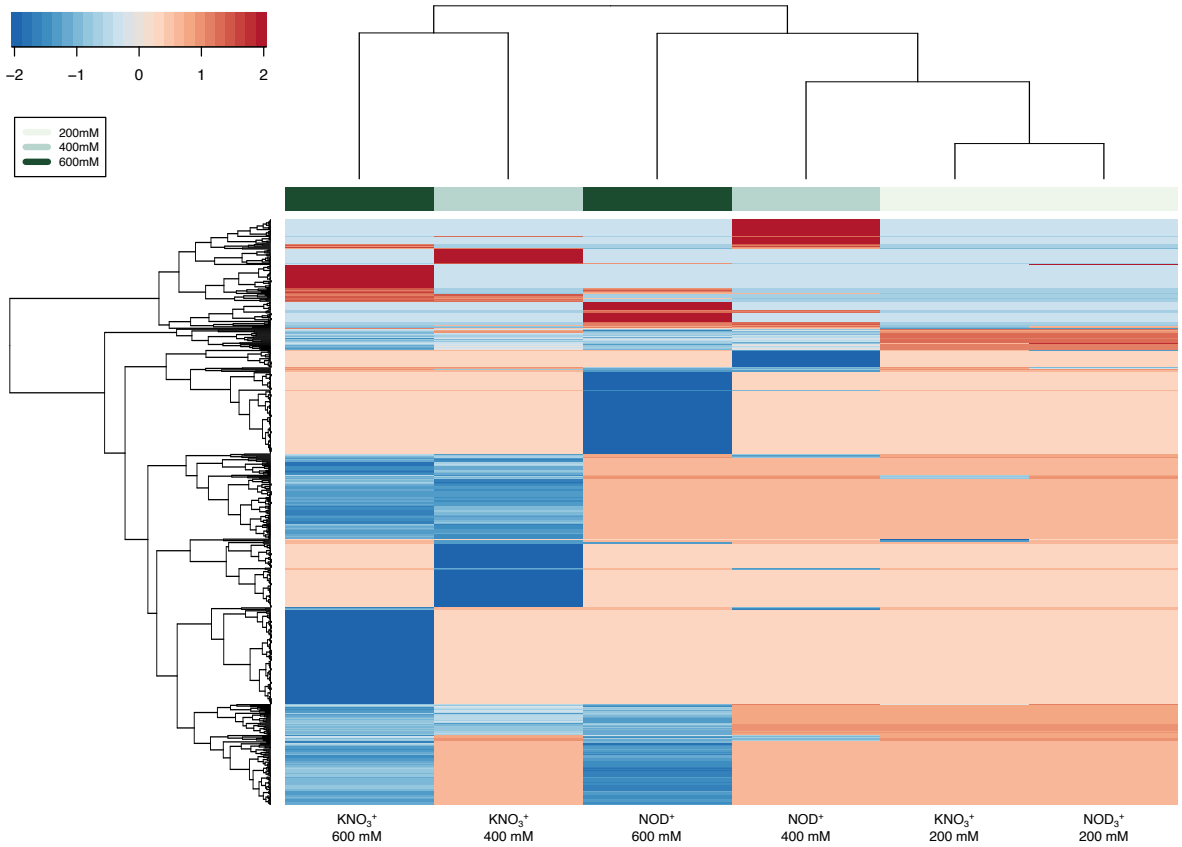

Figure 3.9. Heatmap and dendograms of the normalized $\log_2$ FC of DEGs in *C. glauca* KNO_3^+ and NOD^+, grown at 200 mM, 400 mM and 600 mM NaCl, relative to control (0 mM NaCl). The plotted values are scaled by row to improve visualization. Column color labels groups comparisons by salinity concentration (darker: highest salinity; lightest: lowest salinity).

In KNO_3^+ plants, co-expression analysis grouped DEGs in 10 clusters (Figure 3.10). Clusters C0 to C2, C8 and C9 consistently presented a clear downregulation of DEGs at 600 mM NaCl relative to the control, with differences in the regulation pattern at the intermediate salinities. Cluster C0 showed a progressive downregulation with increasing salinity from 200mM to 600mM NaCl, while C1 displayed an accentuated downregulation only after 400mM NaCl. Although clusters C2 and C3 presented a sharp downregulation between 400mM and 600mM, they also showed an increasing upregulation from control to 400mM NaCl, more evidently in C3. The downregulation pattern in clusters C8 and C9 was marked at 400mM NaCl relative to the control, exhibiting a more progressive decrease in C8, and showing only a slight change at the highest salinity condition in both clusters. Oppositely, clusters C4 and C5 consistently presented a strong upregulation of DEGs from 0 to 600mM NaCl. Overall, in clusters C6 and C7 a downregulation was observed at 400mM NaCl, which was more accentuated in C7 and reverted at 600mM NaCl in both cases.

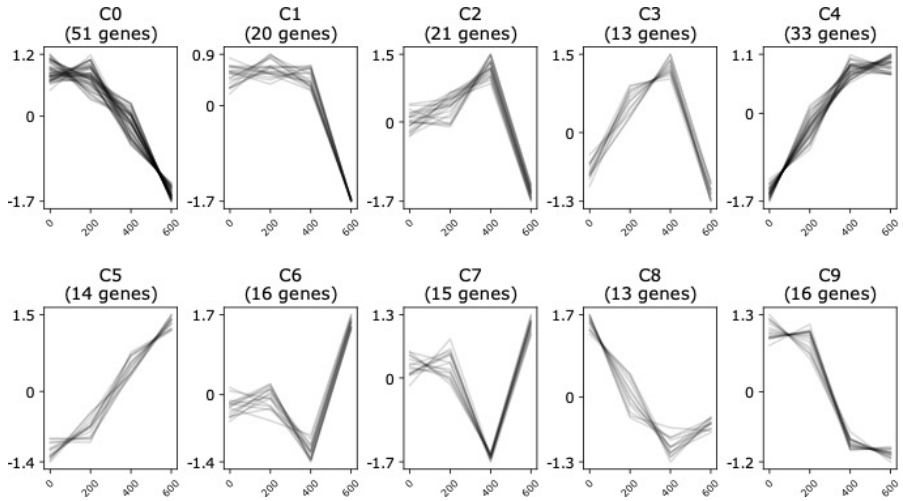

Figure 3.10. Expression pattern of DEGs in *C. glauca* KNO_3^+ plants, clustered by potential co-expression.

In NOD^+ plants, co-expression analysis grouped DEGs in 11 clusters (Figure 3.11). Clusters C0 to C5 and C10 presented a noticeable downregulation of DEGs at 600 mM NaCl relative to control, with different patterns of regulation at the intermediate salinities. Clusters C0 and C1 showed a similar behavior than the same clusters in $KNO3^+$ plants, with the first having a progressive downregulation with increasing salinity and the second only showed an accentuated downregulation at 600mM NaCl. Clusters C2 and C3 showed an upregulation at 200mM NaCl, which overall was reverted at higher salinity levels. In clusters C4 and C5 the major downregulation happened rapidly between 200mM and 400mM NaCl, being moderately reverted in C5 at 600 mM NaCl. In cluster C10, the downregulation occurred in two marked steps, between control and 200 mM NaCl and between 400 mM and 600 mM NaCl. Clusters C6 and C7 presented a strong downregulation at 400mM NaCl, which was inverted at 600 mM NaCl. However, the upregulation at 600 mM was much more evident in C7, showing an upregulation relative to the control.

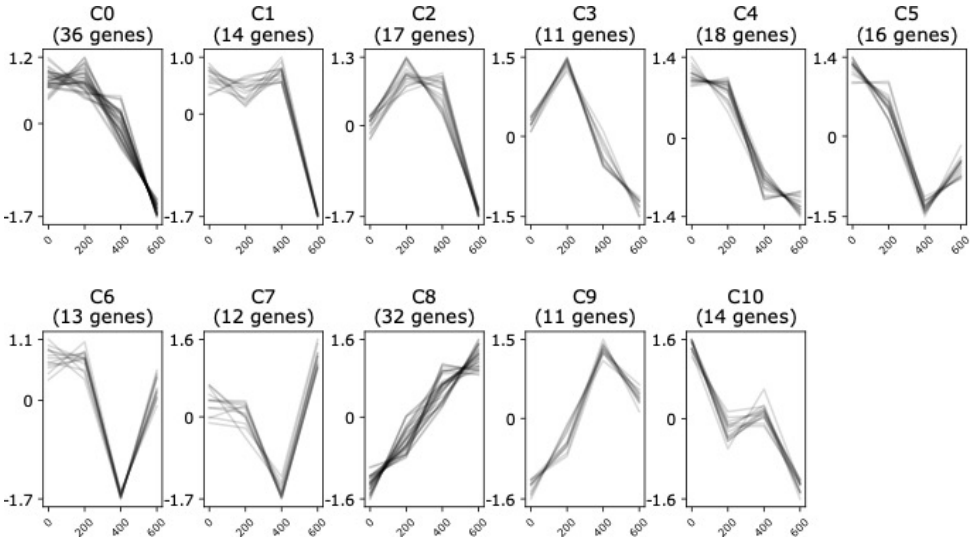

Figure 3.11. Expression pattern of DEGs in *C. glauca* NOD^+ plants, clustered by potential co-expression.

DEGs in almost all clusters, from both plants, were associated with binding, catalytic activity, transport, signaling. Also, there was a predominance of terms related to defense responses and defense to stresses and stimulus, namely hypoxia, water deprivation, oxidative, osmotic, cold and salt. Moreover, in NOD^+ plants, it was also found a relation with responses to wounding, auxin, carbon dioxide, ozone, jasmonic and abscisic acid and biotic stimulus.

A few sets of enriched GO terms were found in the ORA, which included DEGs at 400 mM and 600 mM NaCl (Figure 3.12). Due to its low number, no enriched GO terms were found for DEGs at 200 mM NaCl. In KNO_3^+ plants, DEGs at 400 mM NaCl were found to be enriched for UDP-glycosyltransferase activity, cell wall organization, external encapsulating structure organization, defense response to bacterium, polysaccharide biosynthetic process, leaf abscission and anchored component of plasma membrane. In NOD+ plants DEGs were only enriched for glutamate dehydrogenase (NAD^+) activity, terpenoid catabolic process and plasma membrane. Furthermore, DEGs from both plants were also enriched for cell periphery. At 600 mM NaCl, DEGs from both plants were enriched for multiple binding functions, protein serine/threonine kinase activity, protein phosphorylation, defense responses and abscission. Moreover, in NOD^+ plants, DEGs were also enriched for oxidoreductase and phospholipase activities, multi-organism process, lipid catabolic process, response to hypoxia and cell death.

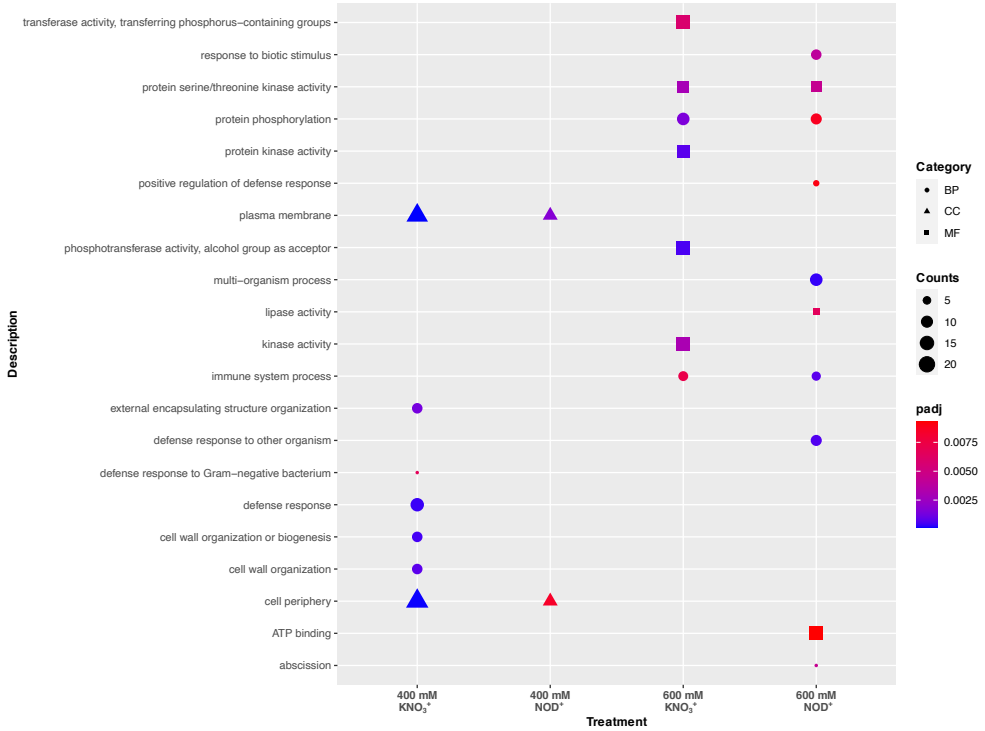

Figure 3.12. Enriched GO terms among down-regulated DEGs, considering the effect of salt-stress at 400 mM and 600 mM in *C. glauca* KNO_3^+ and NOD^+ plants. GO terms are grouped (shape) by the main category – Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Counts (size) indicate the number of DEGs annotated with each GO term and dots are colored by adjust p-value (padj).

Using ShinyGO, enriched KEGG pathways were only observed at 600 mM NaCl in NOD^+ plants, namely cyanoamino acid metabolism, amino sugar and nucleotide sugar metabolism and starch and sucrose metabolism (Figure 3.13). With the same software, PPI networks from STRING database were found at 600 mM NaCl in both plant groups. In KNO_3^+ plants, three DEGs at 600 mM NaCl were mapped in the network, namely SAG101, CRT3 and TAO1. The first two, which were related to regulation of response to stress and positive regulation of response to stimulus, were directly linked in the network. Also, these two genes were indirectly linked to RBOHD, which was associated to cell death in NOD^+ plants. In NOD^+ plants, four DEGs were mapped in the network, namely SAG101, CRT3, TAO1 and PUB17. Again, the first two were directly linked in the network. However, in this case, while CRT3 was associated to cell death, SAG101 was related to immune system process, response to bacterium, regulation of response to stress and positive regulation of response to stimulus. Also, PUB17, which was associated with immune system process, was indirectly linked to both of these genes.

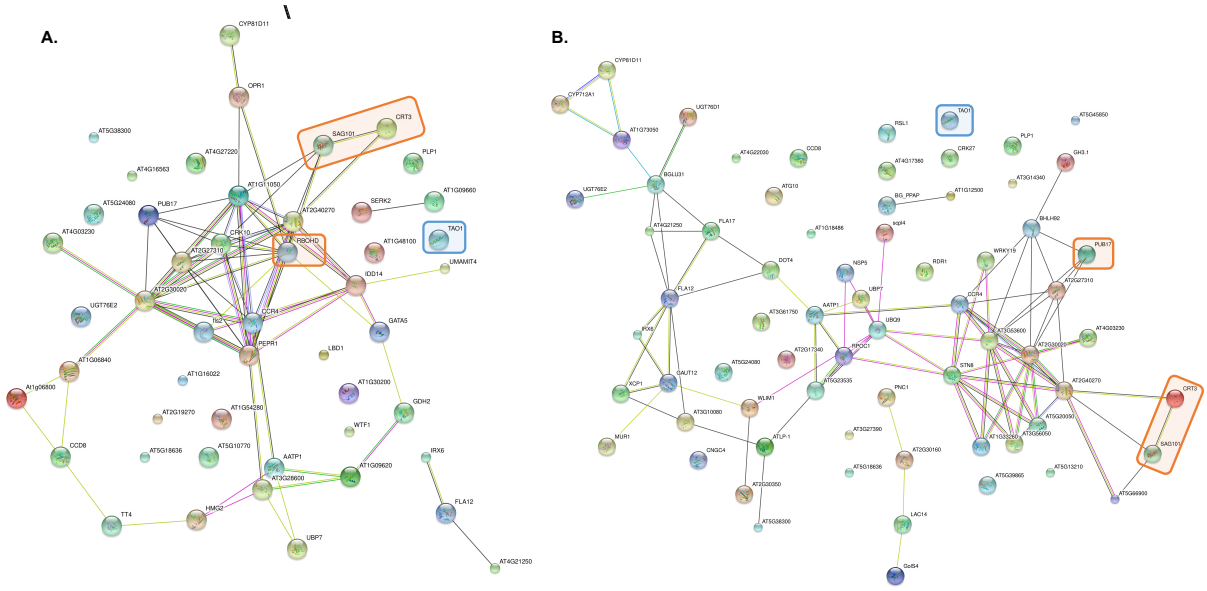

Figure 3.13. PPI networks of DEGs in *C. glauca*, grown at 600 mM NaCl relative to control, retrieved through ShinyGO, based on STRING database. (A) KNO_3^+ plants. (B) NOD^+ plants. [Orange: directly or indirectly linked; Blue: unlinked].

3.2. *Coffea canephora* and *Coffea arabica*

3.2.1. CO_2

Marques, I., Fernandes, I., David, P., Paulo, O. S., Goulao, L. F., Fortunato, A. S., Lidon, F. C., DaMatta, F. M., Ramalho, J. C., & Ribeiro-Barros, A. I. (2020). Transcriptomic Leaf Profiling Reveals Differential Responses of the Two Most Traded Coffee Species to Elevated [CO_2]. *International journal of molecular sciences*, *21*(23), 9211. https://doi.org/10.3390/ijms21239211

In analysis previous to this work, quality assessment and data filtering generated an average of 26 M (93%) clean reads from 28 M raw reads, with a high proportion (81%) of uniquely reads aligned to the reference genome of *C. canephora*, which demonstrated a very good coverage over the species transcriptome (Table 3.4). Through visual inspection of the PCA, replicate 1C was considered an outlier and thus removed from the downstream analysis (Figure 3.14).

Table 3.4. Genome mapping showing the alignment and reads counting results of the transcriptome of Icatu and CL153 against the genome of *Coffee canephora*. A, B, C correspond to the individual biological replicates. RAW READS: number of reads obtained after sequencing. CLEAN READS: number of reads passing the Illumina quality filters and downstream filters. % CLEAN: percentage of reads passing filters compared to the number of raw reads. % MULTIPLE MAP: proportion of reads aligned to exons of several overlapping genes compared to the number of clean reads. % UNMAPPED: proportion of non-aligning reads compared to the number of clean reads.

GENOTYPES	[CO_2] (µL L^{-1})	REPLICATES	RAW READS	CLEAN READS	% CLEAN	% MULTIPLE MAP	% UNMAPPED
ICATU	380	1A	28702752	26442162	92.1	9.10	3.2
		1B	28603251	26372236	92.2	12.7	3.1
		1C	27795986	25107797	90.3	23.2	3.2
		Average	28367329	25974065	91.5	15.0	3.2
	700	3A	30895839	29009249	93.8	13.3	3.1
		3B	25630485	23784221	92.8	17.6	3.2
		3C	31962251	29719153	92.9	18.1	3.3
		Average	29496191	27504207	93.2	16.3	3.2
CL153	380	5A	24532884	22853193	93.1	18.8	2.8
		5B	28922635	26926317	93.1	19.8	2.8
		5C	25702571	23940567	93.1	19.5	2.7
		Average	26386030	24573359	93.1	19.4	2.8
	700	7A	29807104	27910922	93.6	12.9	2.5
		7B	26162690	24625490	94.1	12.8	2.5
		7C	25883732	24238764	93.6	14.2	2.8
		Average	27284508	25591725	93.8	13.3	2.7
Total average			27883515	25910839	92.9	16.1	2.9

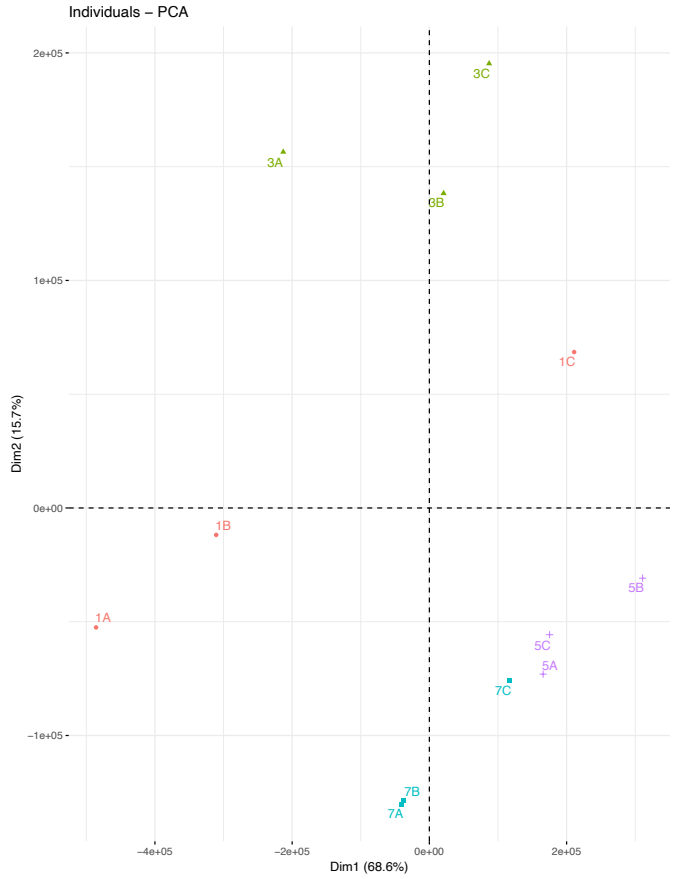

Figure 3.14. PCA of rlog transformed gene expression data in Icatu and CL153, grown either in aCO_2 or eCO_2.

Icatu expressed 21,714 genes under aCO_2 and 21,659 genes under eCO_2, which was more than the genes expressed by CL153 at the same conditions: 20,728 and 21,186, respectively. Conversely, as a response to eCO_2, CL153 presented 6,486 (31%) DEGs, while Icatu only had 4,895 (23%). In both genotypes, the proportion of up- and down-regulated DEGs was close to 50%. Overall, the log_2 FC values of DEGs ranged from -8.08 to 6.11 in Icatu and from -5.31 to 9.88 in CL153. A high number of DEGs (2799) was found to share the same response to eCO_2 in the two genotypes, corresponding to 57% of all DEGs in Icatu, and to 43% of CL153 DEGs (Figure 3.15).

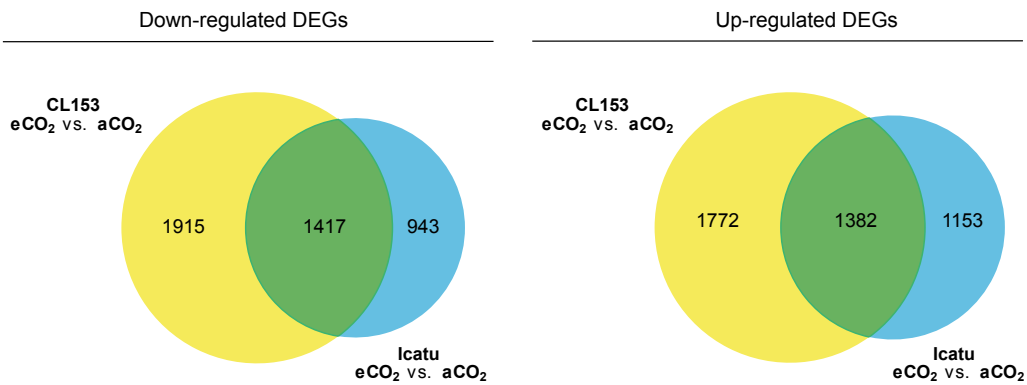

Figure 3.15. Down- and up-regulated DEGs at eCO_2 relative to control, shared by the two genotypes (green), specific of CL153 (yellow) and specific of Icatu (blue).

In the DE analysis between genotypes, 6,764 genes were found to be differentially expressed under eCO_2 and 5,914 under aCO_2. Of the totality of DEGs in CL153 relative to Icatu, 37% were present regardless [CO_2], while 34% were only present at eCO_2 and 27% only at aCO_2, showing that the differences between genotypes are accentuated with increased [CO_2] (Figure 3.16). The log_2 FC values of DEGs between genotypes varied from -13.36 to 14.23 under aCO_2 and from -10.99 to 12.17 under eCO_2.

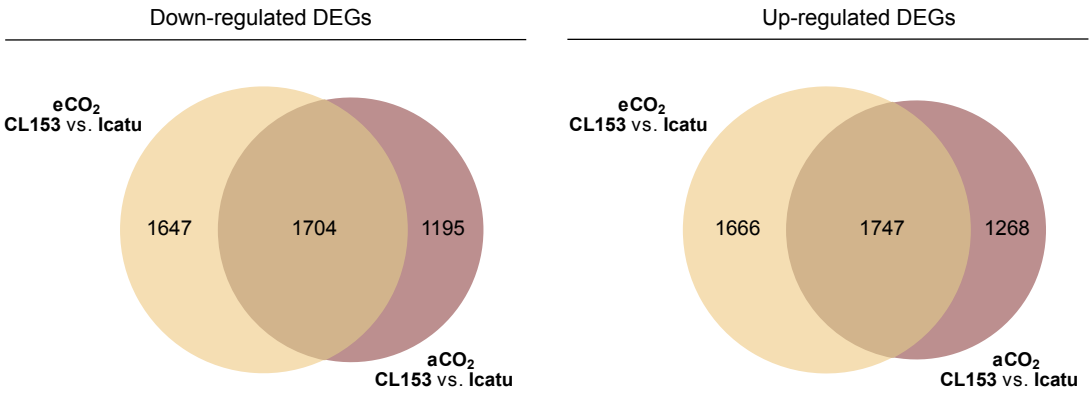

Figure 3.16. Down- and up-regulated DEGs of CL153 relative to Icatu, present at both [CO_2] (brown), specific at eCO_2 (yellow) and specific at aCO_2 (pink).

In addition to the 2799 DEGs at eCO_2 shared by the two genotypes with the same regulation pattern, 566 (17%) DEGs showed opposite patterns. From those, 311 DEGs were down-regulated in CL153 but up-regulated in Icatu, while 255 exhibited the opposite regulation (Figure 3.17).

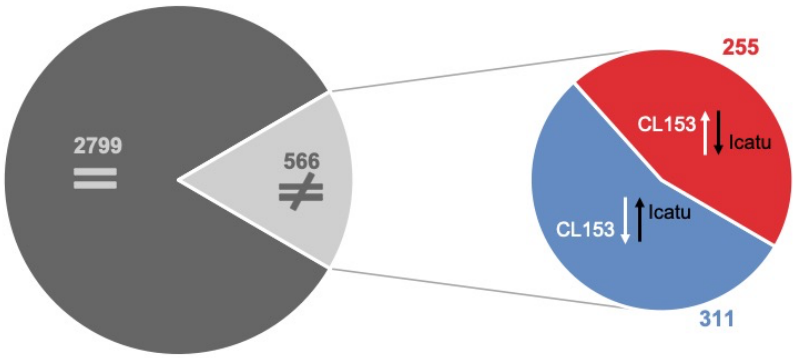

Figure 3.17. Expression of up- (↑) and down-regulated (↓) DEGs at eCO_2 relative to aCO_2 in both genotypes, either exhibiting similar (=) or opposite (≠) patterns.

The functional characterization of eCO_2-responsive DEGs from Icatu revealed an association with 3,923 (32%) GO terms in the BP category, 4,097 (34%) in the MF and 4,168 (34%) in the CC.

However, only 3 of the GO terms from the BP category were found to be significantly enriched, all related to up-regulated DEGs. In comparison, the functional characterization of eCO_2-responsive DEGs from CL153 established a relation with 5,205 (32%) BP's, 5,373 (33%) MF's and 5,590 CC's (35%). Although the proportion of annotations in relation to the number of DEGs was similar in both genotypes, there was a greater number of enriched GO terms in CL153, more specifically 8 terms from the BP category, 4 terms from the MF and 1 from the CC. From these, 2 BP and the CC terms were associated to down-regulated DEGs, while the remainder are part of the annotation of up-regulated DEGs. The full set of enriched GO terms of both genotypes is presented in figure 3.18.

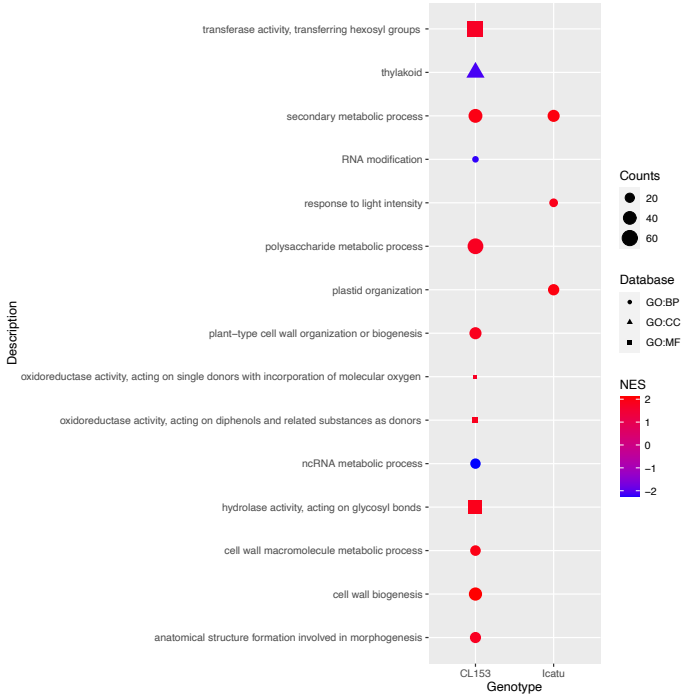

Figure 3.18. Significantly enriched GO terms, according to GSEA, among down- (blue) and up-regulated (red) DEGs, considering the effect of eCO_2 in CL153 (left) and Icatu (right). GO terms are grouped by main category – Biological Process (GO:BP), Molecular Function (GO:MF) and Cellular Component (GO:CC). Counts indicate the number of DEGs annotated with each term and dots are colored by ascending normalized enrichment score (NES).

Moreover, among DEGs in CL153 relative to Icatu, 2 GO terms were significantly enriched, under eCO_2, namely plastid membrane and oxidoreductase activity acting on diphenols and related substances as donors, associated with down- and up-regulated DEGs, respectively (Figure 3.19).

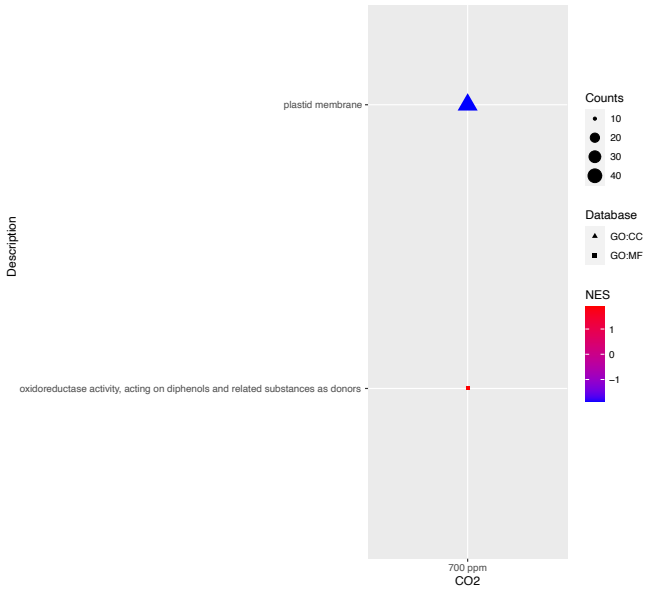

Figure 3.19. Significantly enriched GO terms, according to GSEA, among down- (blue) and up-regulated (red) DEGs, comparing CL153 to Icatu, under the effect of eCO_2. GO terms are grouped by main category – Molecular Function (GO:MF) and Cellular Component (GO:CC). Counts indicate the number of DEGs annotated with each term and dots are colored by ascending normalized enrichment score (NES).

When searching for pathway enrichment, 2 KEGG pathways were found to be significantly enriched among up-regulated DEGs at eCO_2 relative to aCO_2, in CL153. Also, 1 enriched pathway from WikiPathways was found in up-regulated DEGs between genotypes, under eCO_2 (Table 3.5).

Table 3.5. GSEA of differentially expressed genes DEGs from KEGG and WikiPathways databases. Counts indicate the number of DEGs annotated with each pathway and normalized enrichment scores (NES).

Database	ID	Description	Counts	NES
		eCO_2 vs. aCO_2 | CL153		
KEGG	map00906	carotenoid biosynthesis	12	1.99
	map00073	cutin, suberine and wax biosynthesis	5	1.76
		CL153 vs. Icatu | eCO_2		
WikiPathways	WP3661	genetic interactions between sugar and hormone signaling	12	1.67

Genes associated with photosynthesis and some related biochemical components were found among eCO_2-responsive DEGs in both genotypes, although with different responses between them. The terms ‘photosynthesis’ and ‘chlorophyll metabolic process’ were mostly (78%/70%) related to down-regulated DEGs under eCO_2 in CL153, while Icatu only showed a slight (51%/53%) down-regulation of these DEGs (Figure 3.20). Conversely, GO terms associated with RuBisCO were largely (67%) associated with up-regulated DEGs in both genotypes under eCO_2.

Antioxidant components were slightly more related to up-regulated DEGs under eCO_2, more evidently in CL153 (57%) than Icatu (51%). Terms involving lipid metabolic FAD and LOX showed a similar parallelism between genotypes, being majorly associated with up-regulated DEGs (CL153: 67%; Icatu: 63%). GO terms comprising cellular respiration and pyruvate kinase activity were even more markedly related to up-regulated DEGs, especially in Icatu (CL153: 69% and 75%; Icatu: 82% and 100%, respectively). However, while in CL153 malate dehydrogenase related-DEGs were mostly up-regulated (83%), in Icatu they were evenly distributed between the two types of regulation (50%).

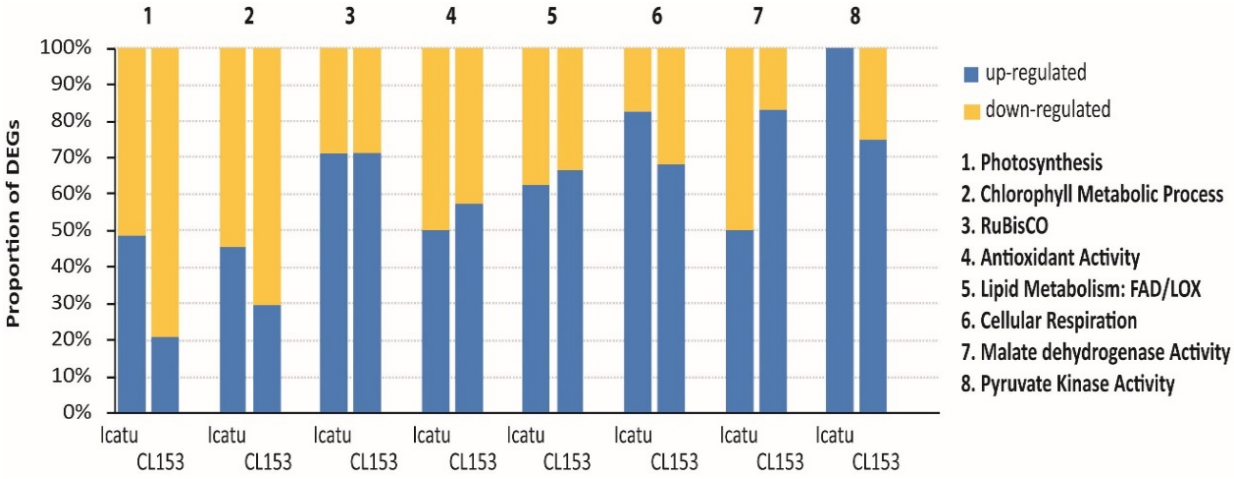

Figure 3.20. Proportion of up-(blue) and down-regulated (yellow) DEGs at eCO_2 vs. aCO_2, associated to specific physiological and biochemical responses in Icatu and CL153.

3.2.2. CO_2 + Temperature

Marques, I., Fernandes, I., Paulo, O. S., Lidon, F. C., DaMatta, F. M., Ramalho, J. C., & Ribeiro-Barros, A. I. (2021). A Transcriptomic Approach to Understanding the Combined Impacts of Supra-Optimal Temperatures and CO_2 Revealed Different Responses in the Polyploid *Coffea arabica* and Its Diploid Progenitor *C. canephora*. *International journal of molecular sciences*, *22*(6), 3125. https://doi.org/10.3390/ijms22063125

Previously to this work, quality control analysis generated an average of 26 M clean reads, from an average 28 M raw reads (Table 3.6). Overall, a high proportion of reads were aligned to the reference genome, since only an average of ca. 3% of reads were not mapped to the reference genome of *C. canephora*. The number of genes expressed by each sample varied in a range of 2% to 6% between the control temperature (25 °C) and the two different supra-optimal temperatures of 37 °C and 42 °C, more evidently under eCO_2. Overall, fewer genes were expressed as temperatures increased, especially in combination with eCO_2 and in CL153, where the lowest number of expressed genes was observed (Figure 3.21A). The PCoA revealed a stronger transcriptomic response as a result of the effect of the highest temperature, since PC2 allowed a clear separation of all samples under 42ºC from samples under 25ºC and 37ºC, regardless of [CO_2] and genotype. PC1 was able to cluster CL153 samples, except for the plants grown at the highest temperature under aCO_2, whereas Icatu samples presented a wider variation. Moreover, PC1 also separated samples at aCO_2 from eCO_2 under 42ºC in both genotypes (Figure 3.21B).

Table 3.6. Summary of sequencing data and mapped reads for the samples of *Coffee arabica* cv. Icatu (Icatu) and *C. canephora* cv. CL153 (CL153). A, B, C correspond to biological replications. Plants were grown in two different stress temperatures, 42/30°C and 37/28°C and the control temperature at 25/20ºC, in either 380 µmol mol^{-1} (aCO_2) or 700 µmol mol^{-1} (eCO_2). RAW READS: number of reads obtained after sequencing. CLEAN READS: number of reads passing the Illumina quality filters and downstream filters. % CLEAN: percentage of reads passing filters compared to the number of raw reads. % MULTIPLE MAP: proportion of reads aligned to exons of several overlapping genes compared to the number of clean reads. % UNMAPPED: proportion of non-aligning reads compared to the number of clean reads.

GENOTYPE	[CO2] (µL L^{-1})	TEMP (ºC)	REPLICATES	RAW READS	CLEAN READS	% CLEAN	% MULTIPLE MAP	% UNMAPPED
Icatu	380	25	1A	28702752	26442162	92.12	9.06	3.21
			1B	28603251	26372236	92.20	12.78	3.10
			1C	27795986	25107797	90.33	23.23	3.21
			Average	28367330	25974065	91.60	15.02	3.18
		37	9A	28029335	26158146	93.32	9.39	3.77
			9B	26467175	24778725	93.62	9.45	3.74
			9C	26399997	24568206	93.06	9.58	3.76
			Average	26965502	25168359	93.34	9.47	3.75
		42	2A	35347710	32590153	92.20	10.02	3.71
			2B	36899936	34008889	92.17	8.77	3.33
			2C	31079752	28583341	91.97	8.98	3.32

			Average	34442466	31727461	92.11	9.25	3.45
Icatu	700	25	3A	30895839	29009249	93.89	13.35	3.19
			3B	25630485	23784221	92.80	17.66	3.26
			3C	31962251	29719153	92.98	18.10	3.31
			Average	29496192	27504208	93.22	16.37	3.25
		37	10A	29875514	27911427	93.43	9.60	3.78
			10B	35512604	32842431	92.48	15.58	3.67
			10C	24494246	22612119	92.32	15.02	3.61
			Average	29960788	27788659	92.74	13.40	3.68
		42	4A	25150070	23507568	93.47	19.66	3.70
			4B	26381269	24732239	93.75	15.80	3.62
			4C	23576551	22009513	93.35	19.21	3.68
			Average	25035963	23416440	93.52	18.22	3.66
CL153	380	25	5A	24532884	22853193	93.15	18.82	2.83
			5B	28922635	26926317	93.10	19.89	2.86
			5C	25702571	23940567	93.14	19.53	2.70
			Average	26386030	24573359	93.13	19.41	2.79
		37	11A	28288493	25823242	91.29	19.20	3.10
			11B	27548373	25230205	91.59	21.19	3.07
			11C	27339032	25133850	91.93	16.89	2.79
			Average	27725299	25395766	91.60	19.09	2.98
		42	6A	32771150	30503930	93.08	18.96	2.98
			6B	29703647	27717207	93.31	13.25	2.58
			6C	25487141	23755944	93.21	14.62	2.68
			Average	29320646	27325694	93.20	15.61	2.74
CL153	700	25	7A	29807104	27910922	93.64	12.96	2.53
			7B	26162690	24625490	94.12	12.87	2.56
			7C	25883732	24238764	93.64	14.27	2.88
			Average	27284509	25591725	93.80	13.36	2.65
		37	12A	33425021	30805515	92.16	15.54	2.97
			12B	29484379	27221963	92.33	15.72	2.90
			12C	28295180	26140159	92.38	14.71	2.96
			Average	30401527	28055879	92.29	15.32	2.94
		42	8A	21937614	20616287	93.98	17.78	2.98
			8B	23329397	21918759	93.95	18.00	3.05
			8C	29033721	27027236	93.09	21.94	3.18
			Average	24766911	23187427	93.67	19.24	3.07
Average				**28346097**	**26309087**	**92.85**	**15.31**	**3.17**

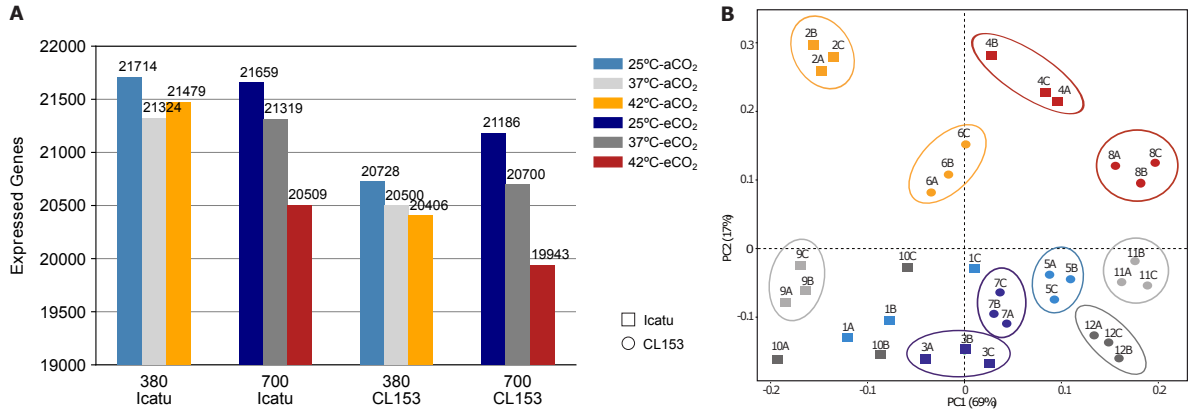

Figure 3.21. Gene expression profiles across samples. (**A**) Number of expressed genes in Icatu and CL153, grown either in aCO_2 or eCO_2, at control temperature conditions (25 °C) and the two supra-optimal temperatures (37 °C and 42 ºC). (**B**) PCoA of rlog transformed gene expression data.

The highest number of DEGs was consistently found at 42ºC under eCO_2 in both genotypes (Icatu: 13,134; CL153: 12,115). However, although the lowest number of DEGs was found at 37ºC in both genotypes, it was lower under eCO_2 in Icatu (9,545), but under aCO_2 in CL153 (8,240).

In both genotypes, under the same [CO_2], the majority of DEGs were shared by the two supra-optimal temperatures, ranging from 62% to 85% in Icatu and from 65% to 86% in CL153 (Figure 3.22). The number of specific DEGs was higher at 42ºC than 37ºC in both genotypes. Moreover, under eCO_2 specific DEGs in each genotype and at each temperature almost always decreased in relation to aCO_2, with the exception for Icatu at 42ºC under eCO_2, which reported the highest number of specific DEGs (Figure 3.22A).

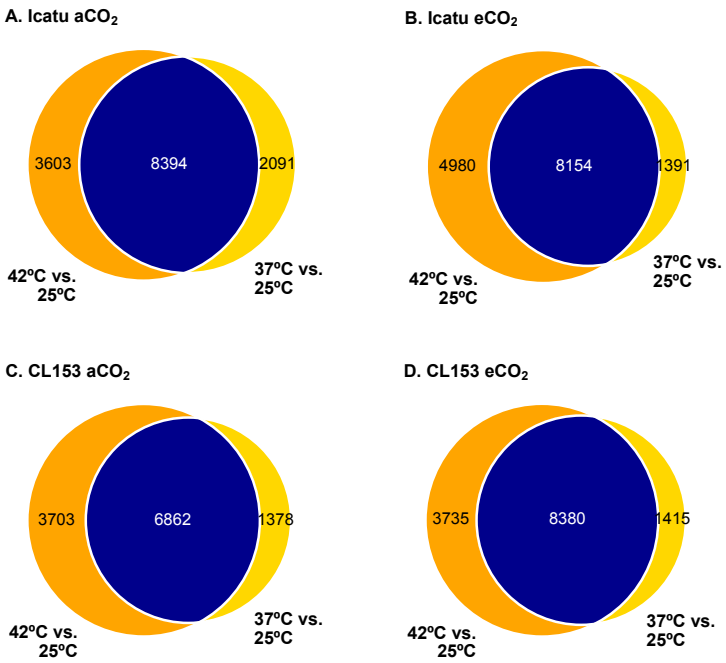

Figure 3.22. Treatment-specific and shared transcriptional patterns among DEGs at the two supra-optimal temperatures of 37°C and 42 °C, relative to control, found in plants of Icatu (A, B) and CL153 (C, D), under aCO_2 or eCO_2.

Overall, the distribution of treatment-specific DEGs between up- and down-regulation was similar in both genotypes, with slightly more up- than down-regulated DEGs across conditions, especially under eCO_2 at either supra-optimal temperature (Figure 3.23).

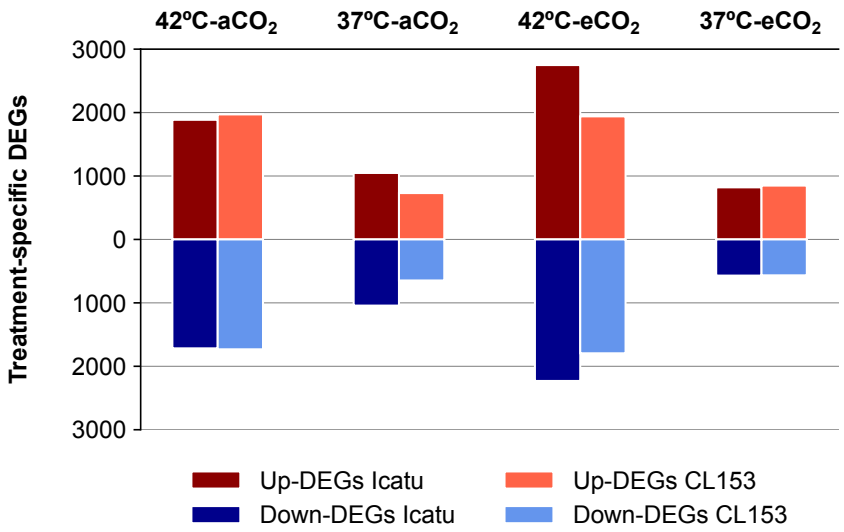

Figure 3.23. The effect of the supra-optimal temperatures of 37ºC and 42 °C on the number of up- and down-regulated treatment-specific DEGs in Icatu and CL153, under aCO_2 or eCO_2.

When clustering treatment-specific DEGs by log_2 FC, there is a clear separation of DEGs in Icatu at 42ºC, under either [CO_2] from the remainder samples (Figure 3.24). On a separate cluster, Icatu samples at 37ºC are closer together, followed by CL153 under aCO_2 and then under eCO_2. The CL153 samples at 42ºC appear closer to the 37ºC cluster than to Icatu samples at 42ºC. Moreover, the differences between samples of the two supra-optimal temperatures in CL153 seems to be increased by the effect of eCO_2.

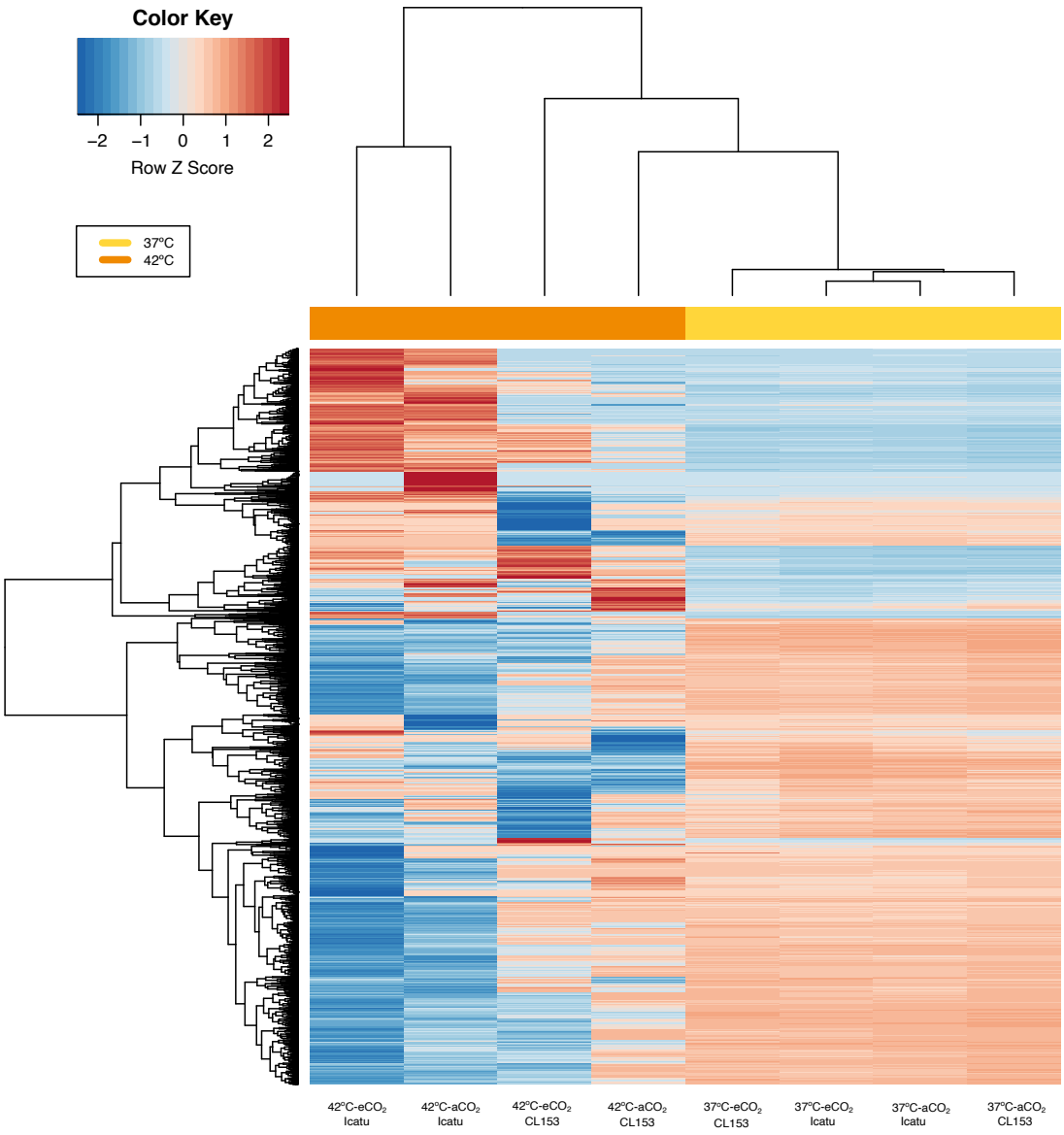

Figure 3.24. Heatmaps and dendrograms of the normalized $\log_2$ FC of treatment-specific DEGs in Icatu and CL153 as a response to 37 °C and 42 °C supra-optimal temperatures under aCO_2 or eCO_2. The plotted values are scaled by row for improved visualization. Hot colors represent up-regulated DEGs and cold colors represent down-regulated DEGs.

In both species, an average of 74% of DEGs were annotated with GO terms, according to the functional annotation of the reference genome of *C. canephora*. Among up-regulated DEGs, a set of different GO terms were found to be enriched at either or both supra optimal temperatures and [CO_2] conditions, specifically 16 in Icatu and 11 CL153 (Figure 3.25A). Up-regulated DEGs at 37°C showed more enriched GO terms than at 42°C, under both [CO_2] conditions (Figure 3.25A). Moreover, within the same temperature treatment, eCO_2 triggered more enriched GO terms than aCO_2. The enriched GO term associated with more DEGs, RNA binding, was found at 42 °C in eCO_2. Among the remaining terms there was a predominance of GO terms related to photosynthesis, including thylakoid, photosystem, photosynthesis, light reaction and chlorophyll binding, which were found only at 37 °C,

under both [CO_2]. Also, in Icatu, up-regulated DEGs at 42°C under aCO_2 were enriched for heat shock protein binding, while under eCO_2 there was an enrichment of terms related to proteins folding and binding.

In down-regulated DEGs there was an enrichment of 15 different GO terms in Icatu and 18 in CL153 (Figure 3.25B). At 37ºC, down-regulated DEGs showed no to very few enriched terms under aCO_2, while under eCO_2 there was an enrichment in a set of general molecular functions, such as oxidoreductase, transferase, binding and catalytic activities, which differed between genotypes. At 42°C in aCO_2, microtubule motor activity and binding were highly enriched in down-regulated DEGs in Icatu, while CL153 showed an enrichment in the molecular functions related to transport. In this highest temperature, eCO_2 triggered less enriched terms, with Icatu plants being mostly enriched in secondary metabolic process and lignin catabolic process, while CL153 plants showed an enrichment in molecular functions linked to calcium ion binding, and transferase and oxidoreductase activities.

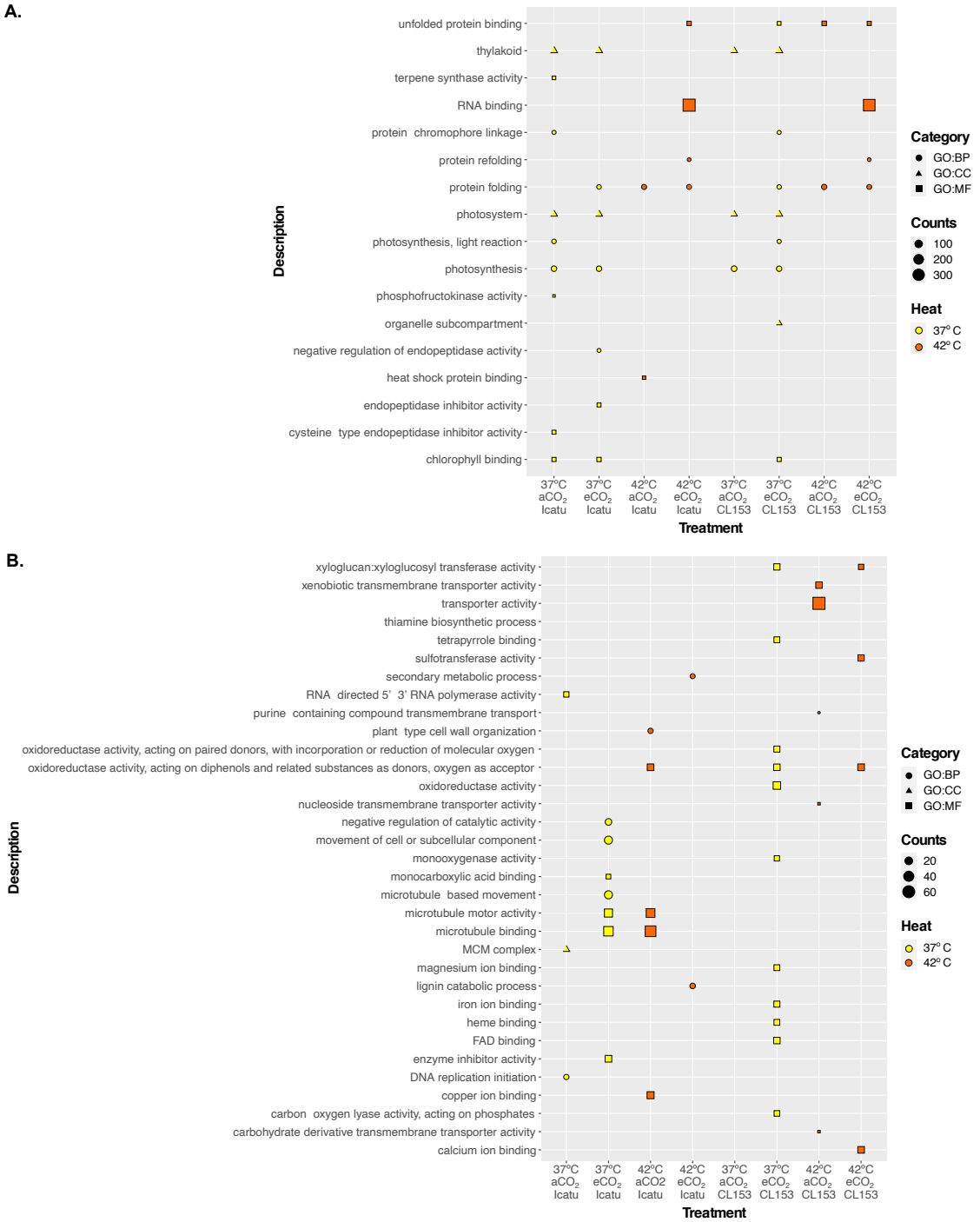

Figure 3.25. ORA of GO terms performed against the functional annotation of the *C. canephora* genome. Enriched GO terms among up-regulated (**A**) and down-regulated (**B**) DEGs in Icatu and CL153, considering the effect of supra-optimal temperatures at 37 °C and 42°C, under either aCO_2 or eCO_2. GO terms are grouped by the main categories: biological process (GO:BP), molecular function (GO:MF), and cellular component (GO:CC). Counts indicate the number of DEGs annotated with each GO term and dots are colored by temperature treatment.

In the two genotypes, a total of 667 DEGs were found to be related to photosynthesis and other essential metabolic functions in coffee, such as antioxidant activity, lipid metabolism and respiratory process. From these, 180 were directly associated with the photosynthetic pathway, particularly 122 to photosynthesis, 49 to chlorophyll metabolism and 9 to RuBisCO activities. Moreover, 73 DEGs were

linked to antioxidant activities and 123 to lipid metabolism. The remainder 291 DEGs were found related to the respiratory pathway, specifically 170 to cellular respiration, 30 to malate dehydrogenase activity, and 91 to pyruvate kinase activity. The log_2 FC of these DEGs varied widely but the most extreme values were always found at 42°C, regardless [CO_2] conditions.

Comparing responses between genotypes, overall, an equivalent number of DEGs related to photosynthesis, chlorophyll metabolic process and RuBisCO were found as a response to supra-optimal temperatures (Figure 3.26). However, the regulation of such genes was not always concordant between them. Although these DEGs were mostly up-regulated at 37ºC in both genotypes, DEGs at 42ºC were majorly up-regulated in CL153, but down-regulated in Icatu. The remaining terms were mostly associated with down-regulated DEGs, especially at 42°C under eCO_2, although differences were still found between the two genotypes. The proportion of down-regulated DEGs involved in antioxidant activities and lipid metabolism were always higher in Icatu than CL153, except at 37 °C in eCO_2, while in CL153 those DEGs were mostly up-regulated at 37°C, independently of [CO_2]. Moreover, while the down-regulation of such DEGs increased with eCO_2, regardless of temperature, in Icatu there was a decrease with eCO_2 at 37ºC. Overall, in Icatu plants under both [CO_2] conditions, DEGs involved in cellular respiration were mostly up-regulated at 37°C, but down-regulated at 42°C. Conversely, in CL153, these DEGs were mostly down-regulated, with the exception of DEGs at 37ºC under aCO_2. Furthermore, DEGs involved in pyruvate kinase (PK) and malate dehydrogenase (MDH) activity, which are involved in glycolysis, were mostly down-regulated in all treatments, except in the MDH of CL153 plants at 37 °C under aCO_2.

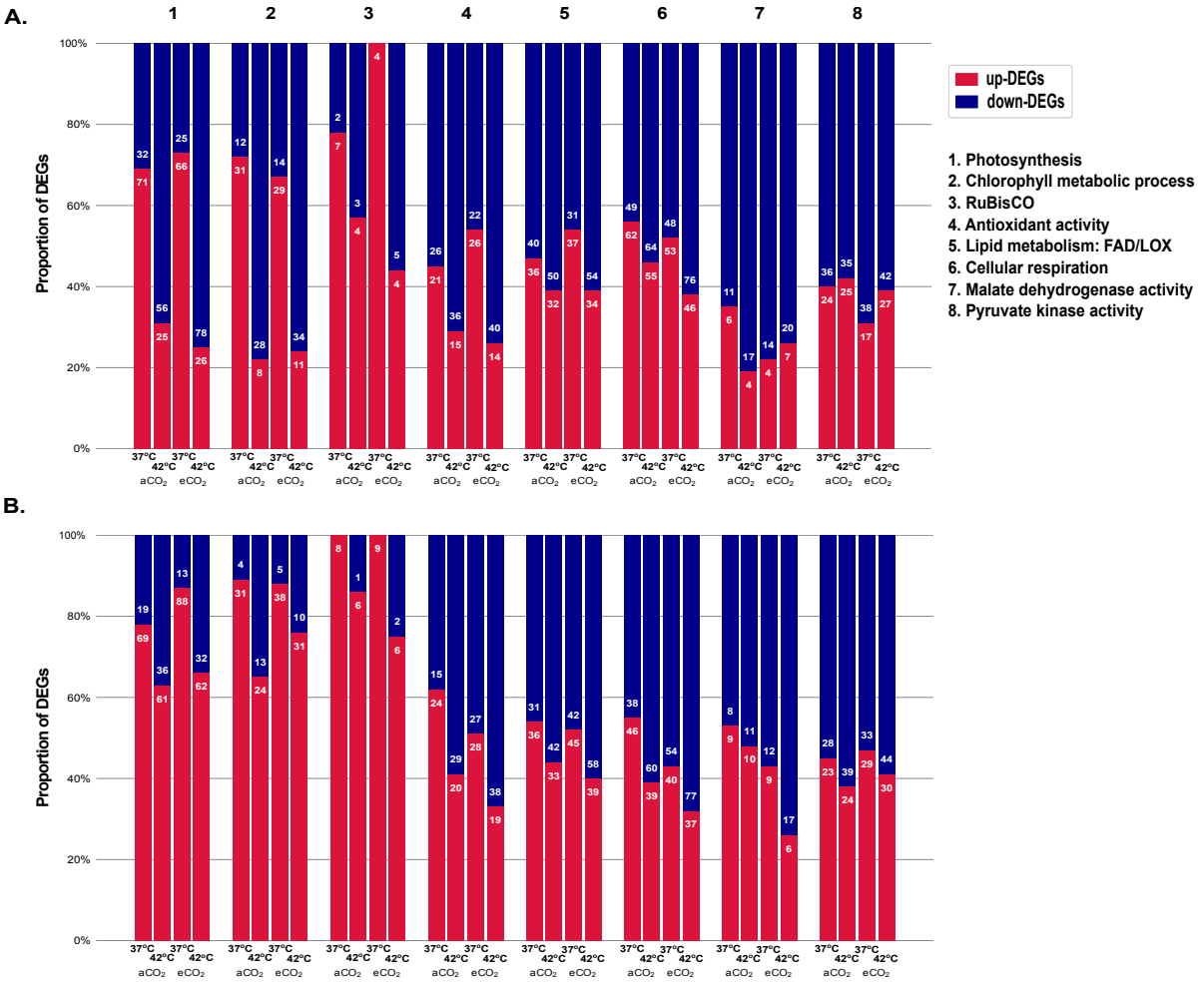

Figure 3.26. Proportion (%) of the regulation of DEGs, related to photosynthesis and biochemical processes in Icatu (**A**) and CL153 (**B**) plants, as a response to 37°C and 42°C, under aCO_2 or eCO_2. Indicated in each bar are the number of DEGs associated with each term.

Looking at more specific functions, it was found that more than half of the photosynthetic DEGs were involved in binding activities. Furthermore, DEGs related to the reaction centers of photosystems (PSs) I and II were down-regulated in Icatu and up-regulated in CL153 at 42°C and independently of the [CO_2]. Under these extreme conditions, the same regulation pattern was observed in DEGs involved in chlorophyll a-b binding and most PsbQ and PsbP genes.

Analyzing the heatmap of the DEGs related to the same photosynthesis and relevant metabolic activities, there was a clear separation in two clusters, with the DEGs of Icatu at 42ºC, under both [CO_2] separated from the remaining treatments. Among these, DEGs were clustered primarily by temperature-treatment, and then by genotype, regardless of [CO_2] (Figure 3.27).

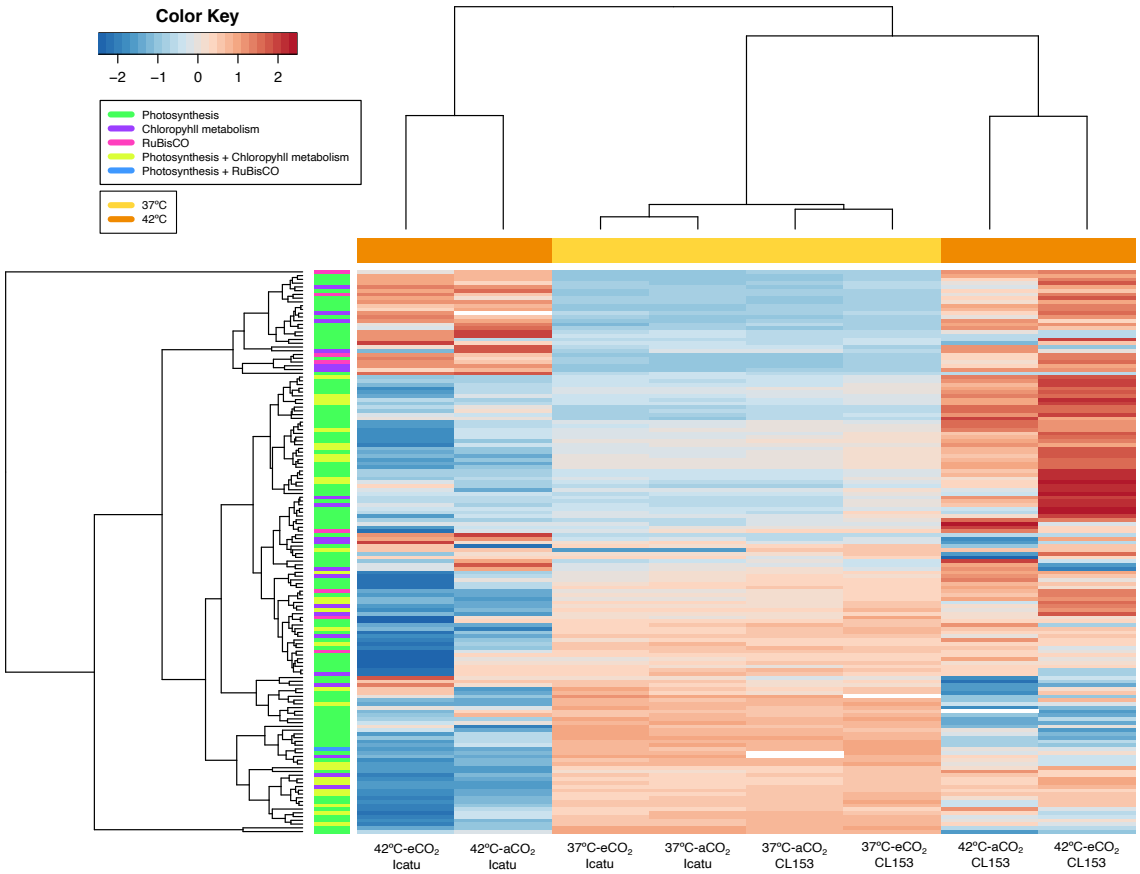

Figure 3.27. Heatmaps and dendrograms of the normalized log_2 FC of photosynthesis-related DEGs as a response to the supra-optimal temperatures of 37ºC and 42ºC, under aCO2 or eCO2, in Icatu and CL153 plants. The plotted values are scaled by row for improved visualization. Hot colors represent up-regulated DEGs and cold colors represent down-regulated DEGs. Column color labels group comparisons by temperature treatment, while row color labels group genes by GO annotation.

3.3. *Limonium* spp.

Limonium raw sequences ranged from 15378236 to 29800668 in apomictic plants and from 9372951 to 22701315 in sexual plants, being higher in the first group (Table 3.7).

Table 3.7. Sequencing data from apomictic, facultative apomictic and sexual *Limonium* spp. samples. Raw reads, obtained after sequencing, generated clean reads after submission to quality control with FastQC and Trimmomatic software. [A-E indicate different biological replicates].

Reproductive strategy	Species	Stage	Replicates	Raw reads	Clean Reads
Apomicitc	*L. multiflorum*	S1	A	15378236	15219989
			B	26741297	26455783
			C	21263703	21069237
			D	19483758	19286282
		S2	A	16866816	16652632
			B	29800668	29525079
			C	18981960	18799302
			D	20651158	20501631
			E	24619068	24417862
Facultative apomicitc	*L. dodartii*	S4	A	21100986	20800610
			B	16353613	16134223
			C	23992940	23473601
Sexual	*L. ovalifolium*	S1	A	18792168	18610595
		S2	A	20711554	20512815
		S3/S4	A	9372951	9193765
	L. nydeggeri	S1	A	19245651	19060161
		S2	A	22701315	22458578
		S3/S4	A	18317853	18141561

The assembled transcriptome generated a total of 162520 trinity unigenes, with a 43% GC content and a contig N50 of 2128 (Table 3.8). According to BUSCO, a 90% completeness was achieved in *de novo* assembled transcriptome indicating that we have generated a high-quality transcriptome assembly that could be used for further downstream analysis.

Table 3.8. Quantification of basic quality and completeness metrics of *Limonium de novo* transcriptome assembly. Samples of apomictic (*L. multiflorum*) plants in stages S1 and S2, facultative apomictic (*L. dodartii*) plants in stage S4 and sexual (*L. ovalifolium*, *L. nydeggeri*) plants in stages S1, S2 and S3/S4 were combined to perform a *de novo* transcriptome assembly using Trinity software.

Basic Metrics	Values
Total trinity unigenes	162520
Total trinity transcripts (isoforms)	315983
Percent GC (%)	42.94
Contig N10	4592
Contig N20	3604

Contig N30	2998
Contig N40	2528
Contig N50	2128
Completness	
Total number of core genes queried	2326
Number of core genes detected	
Complete	2090 (89.85%)
Complete + Partial	2159 (92.82%)
Number of missing core genes	167 (7.18%)
Average number of orthologs per core gene	2.93
% of detected core genes that have more than 1 ortholog	70.81
Scores in BUSCO format	C:89.8% [S:26.2%, D:63.6%], F:3.0%, M:7.2%

The total number of expressed unigenes among *Limonium* samples was highest in apomictic plants in S2, followed by facultative apomictic plants in S4 (Figure 3.28). Among sexual plants, the number of expressed unigenes was higher in *L. nydeggeri* than in *L. ovalifolium* in S3/S4, but lower in the other stages.

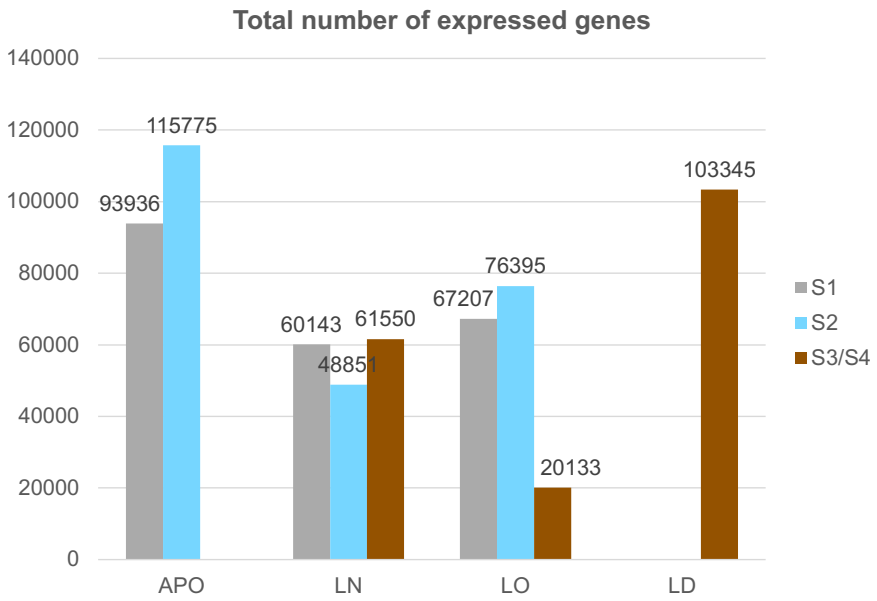

Figure 3.28. Total number of genes expressed by *Limonium* samples from apomictic *L. multiflorum* (APO), facultative apomictic *Limonium dodartii* (LD), and sexual *L. nydeggeri* (LN) and *L. ovalifolium* (LO) ovules in stages S1, S2, S3/S4.

In the PCA analysis, PC1, which accounted for 71% of the variance, revealed a clear cluster of sexual plants in the right side of the graph (Figure 3.12). Also, PC2 separated the sexual plants from S1 and S2 (upper-right quadrant) from S3/S4 plants (lower-right quadrant). Moreover, PC1 grouped all samples from apomictic and facultative apomictic plants from S1 in the left side of the graph, showing a higher dispersion for the remaining stages of these plants (Figure 3.29).

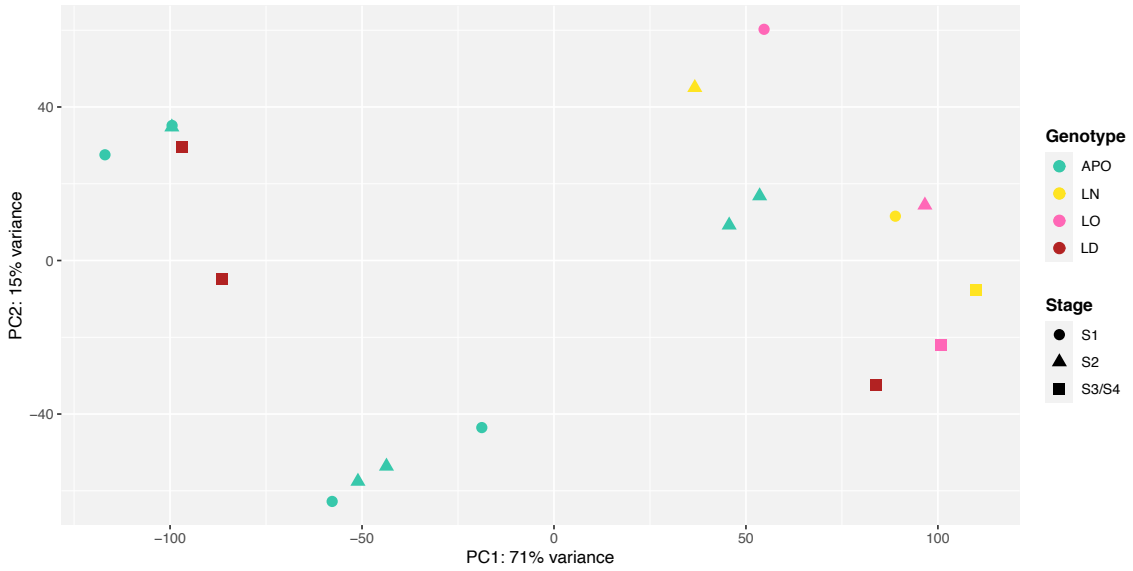

Figure 3.29. PCA of gene expression counts from *Limonium nydeggeri* (LN), *L. ovalifolium* (LO), *L. multiflorum* (APO) *and L. dodartii* (LD). *Limonium* samples were collected at multiple stages (S1, S2 and S3/S4).

The heatmap clustered all apomictic plants in the same main branch, along with the facultative apomictic samples, which were scattered among them (Figure 3.30). Also, almost all sexual plants were clustered in the same main branch, except for *L. ovalifolium* in S3/S4, which was presented in a separate one.

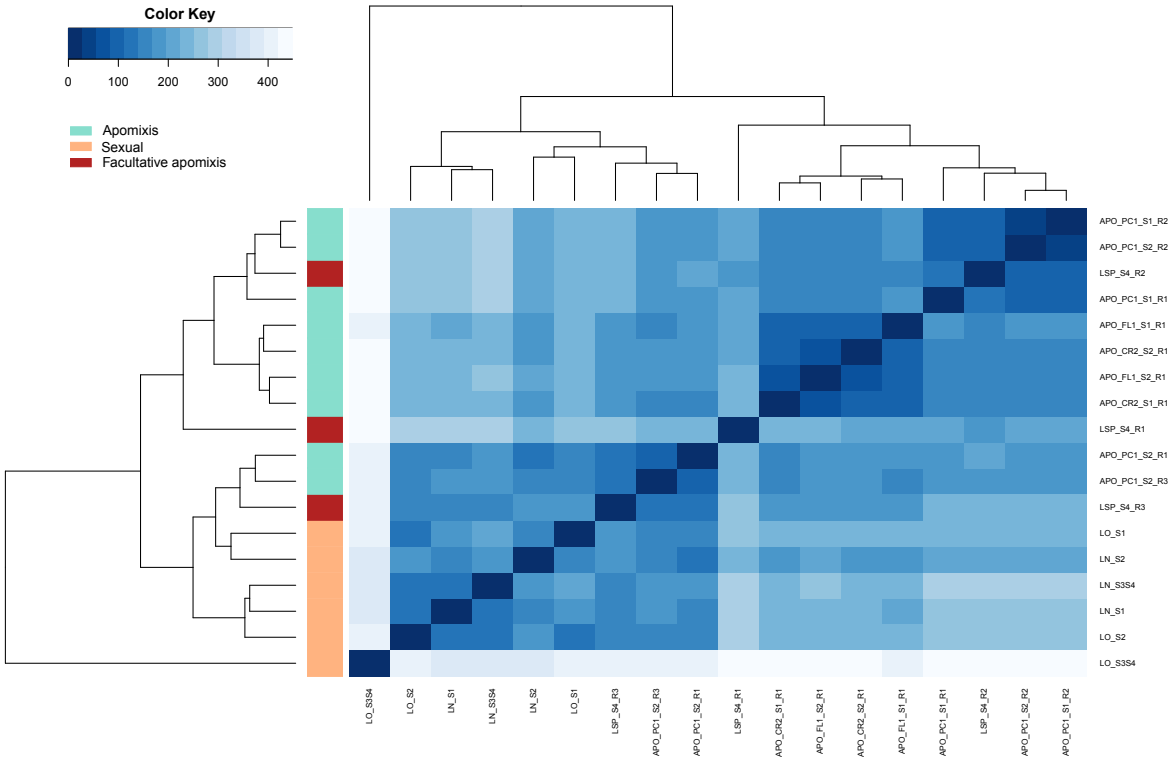

Figure 3.30. Heatmap and dendogram of the normalized log_2 gene counts of apomictic (APO), facultative apomictic (LSP) and sexual (L. nydeggeri – LN, L. ovalifolium - LO) plants in S1, S1/S2, S3/S4 and S4 stages. Column color labels groups comparisons by type of reproduction (orange: sexual; light green: apomixis; dark red: facultative apomixis).

The highest number of DEGs was found in apomictic plants in S2 relative to sexual plants, especially relative to *L. ovalifolium* in S3/S4 (12837 vs. 4401; 11% vs. 4% of expressed genes in apomictic S2) (Figure 3.31). In apomictic S1 there were considerably less DEGs relative to sexual plants in S1 and S2, with just slightly more DEGs relative to *L. nydeggeri* than to *L. ovalifolium*. (average of 3151 vs. 3061; 3% of expressed genes in apomictic S1). When comparing the two sexual plants, there was an increasing number of DEGs with the progression of stages, which varied from 616 to 1611. Among *L. ovalifolium* plants, the lowest number of DEGs was found between S2 and S1 (693), followed by the comparison between S3/S4 vs. S2 (1650) and finally S3/S4 vs. S1 (2067). However, in *L. nydeggeri*, S3/S4 vs. S1 plants generated less DEGs (351) than S2 vs. S1 (796) and S3/S4 vs. S2 (1346). Apomictic plants presented 1096 DEGs between its S1 and S2 stages and 806 DEGs in S2 plants relative to the facultative apomictic *Limonium dodartii* plants.

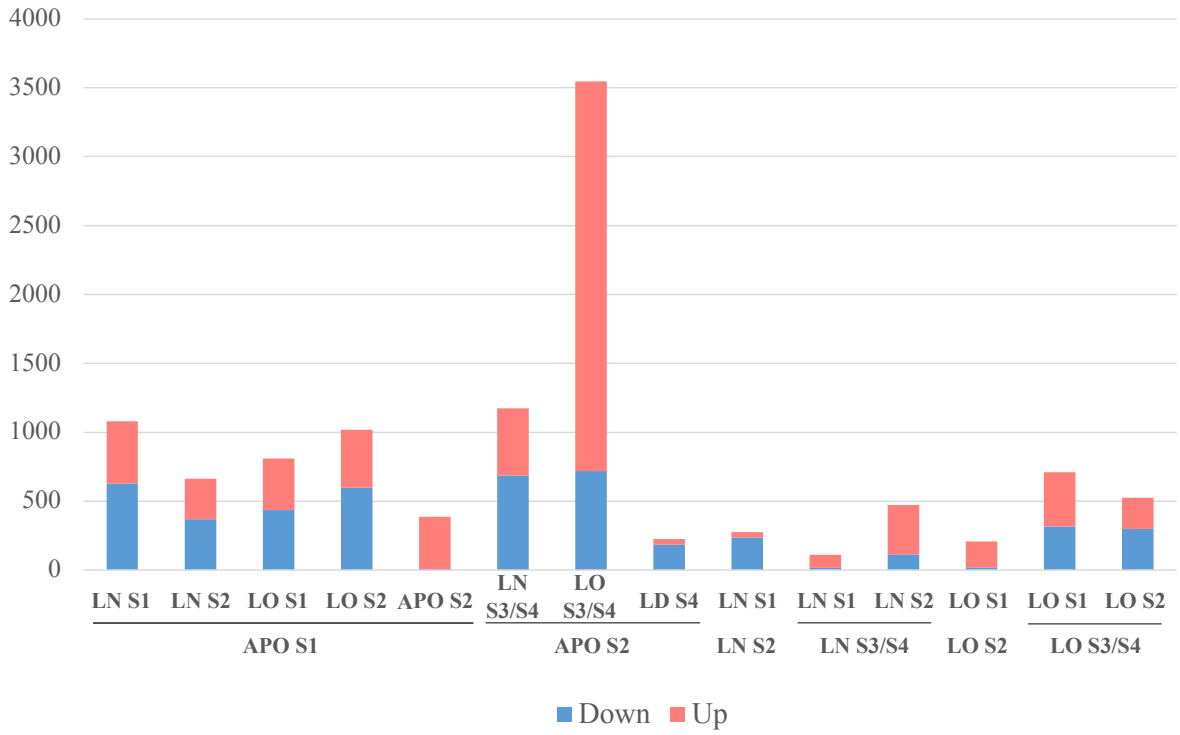

Figure 3.31. Number of uniquely annotated differentially expressed genes (DEGs) in *Limonium* samples from apomictic *L. multiflorum* (APO), facultative apomictic *L. dodartii* (LD), and sexual *L. nydeggeri* (LN) and *L. ovalifolium* (LO) plants in stages S1, S2 and S3/S4. DEGs represent the number of significant genes found to be differently expressed in each comparison (namely: APO S1 vs. **LN S1**; APO S1 vs. **LN S2**; APO S1 vs. **LO S1**; APO S1 vs. **LO S2**; APO S2 vs. **APO S1**; APO S2 vs. **LN S3/S4**; APO S2 vs. **LO S3/S4**; APO S2 vs. **LD S4**; LN S2 vs. **LN S1**; LN S3/S4 vs. **LN S1**; LN S3/S4 vs. **LN S2**; LO S2 vs. **LO S1**; LO S3/S4 vs. **LO S1**; LO S3/S4 vs. **LO S2**).

In sexual plants, the number of DEGs overlapped between different comparisons, differed between the two genotypes (Figure 3.32). While in *L. nydeggeri* the highest overlap was found between S3/S4 vs. S2 and S2 vs. S1, in *L. ovalifolium* the overlap was higher between S3/S4 vs. S2 and S3/S4 vs. S1 (Figure 3.15A, 3.15B). Furthermore, the number of DEGs specific to a comparison was higher in S3/S4 vs. S2 (210; 36% of total DEGs) and in S3/S4 vs. S1 (242; 25% of total DEGs) in *L. nydeggeri* and *L. ovalifolium*, respectively. When comparing the two genotypes in the same stages, the highest overlap was found between S1 and S2. Also, the number of DEGs specific to a stage increased with the progressing of the stages.

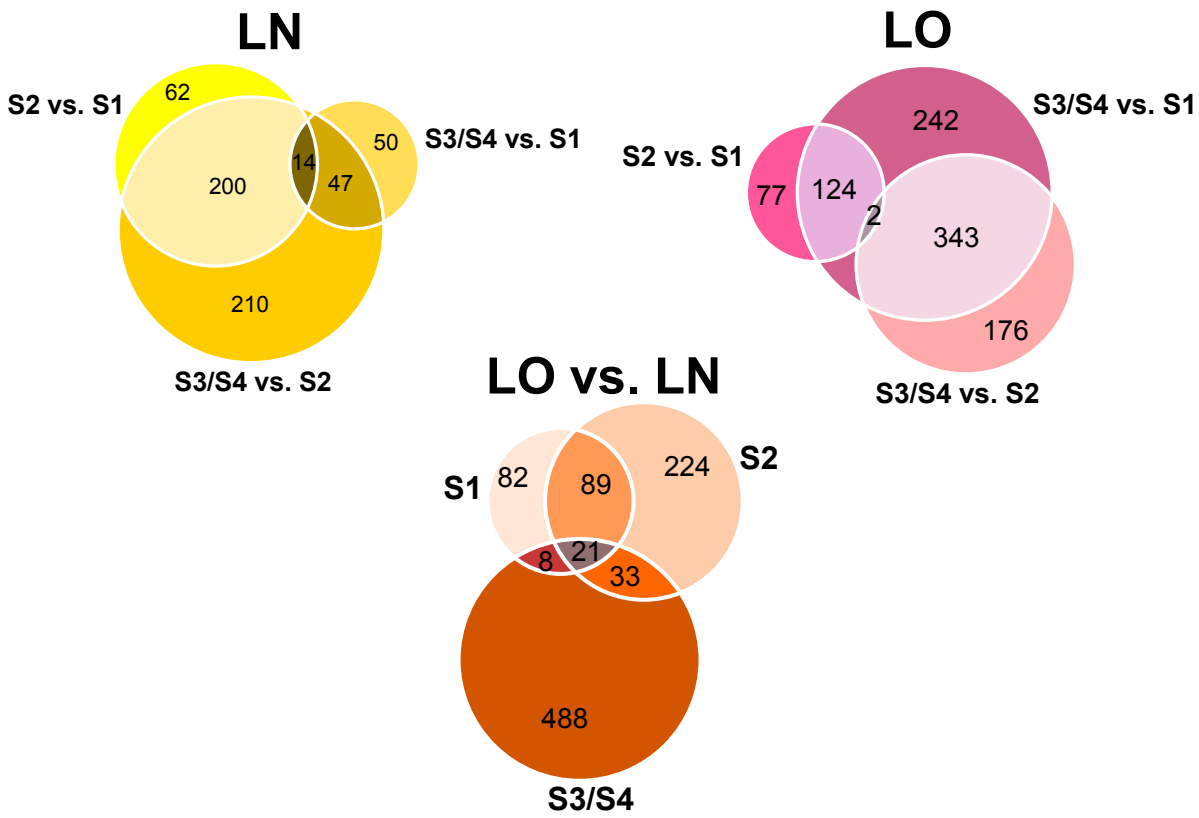

Figure 3.32. Weighted Venn diagrams of specific and overlapping differentially expressed genes (DEGs) found in *Limonium* plants, namely *L. nydeggeri* and *L. ovalifolium* sexual plants. DEGs were filtered by |log2 fold-change (Log_2FC)| >2. Number of overlapping and specific DEGs in: **[A]** *L. nydeggeri* in S2 relative S1, S3/S4 relative to S1 and S3/S4 relative to S2; **[B]** *L. ovalifolium* in S2 relative S1, S3/S4 relative to S1 and S3/S4 relative to S2; **[C]** *L. ovalifolium* relative to *L. nydeggeri* in S1, S2 and S3/S4.

When comparing the initial and final stages of both sexual and apomictic plants, there was a higher overlap between DEGs at S3/S4 vs. S2 in *L. nydeggeri* and DEGs at *L. multiflorum* S2 vs. S1, than in *L. ovalifolium* at the same stages (133 vs. 13), but the opposite in DEGs at S3/S4 vs. S1 in sexual plants (40 vs. 74) (Figure 3.33). In both cases, the number of DEGs specific to a comparison was the highest in *L. ovalifolium*.

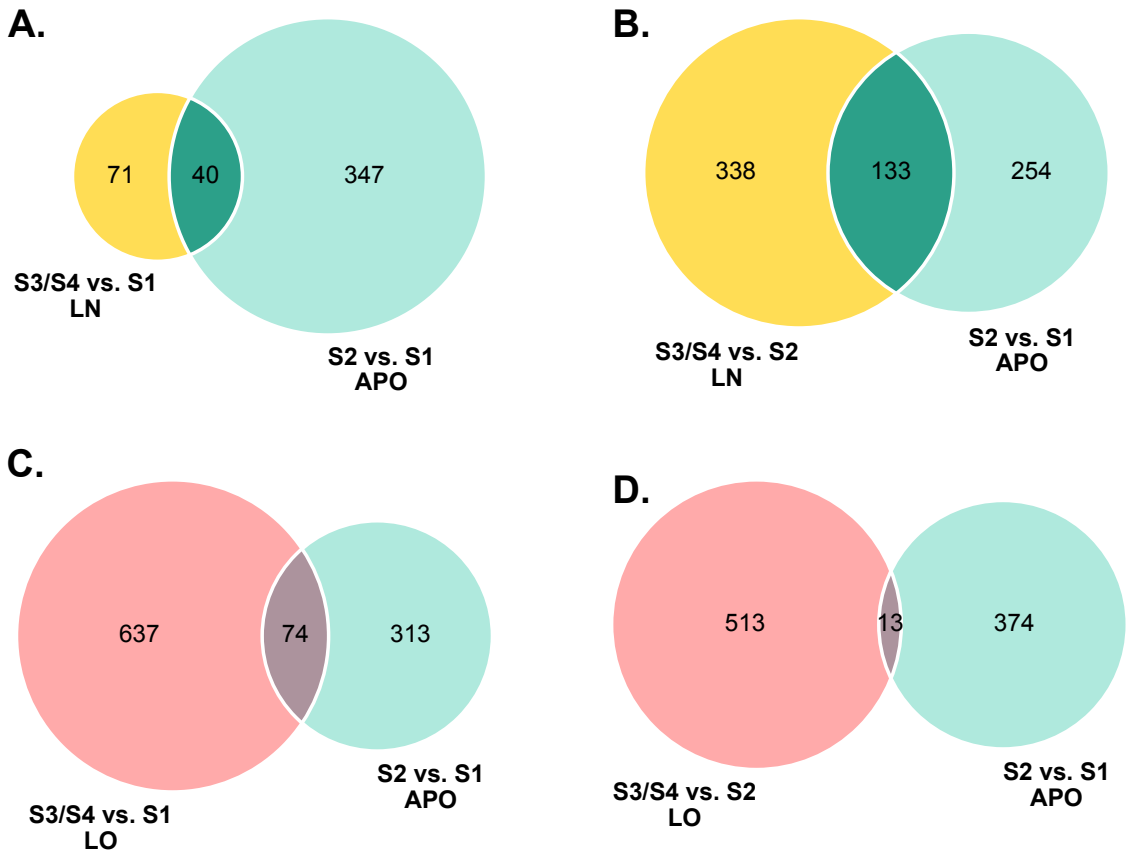

Figure 3.33. Weighted Venn diagrams of specific and overlapping differentially expressed genes (DEGs) found in *Limmonium* plants, namely apomictic *L. multiflorum* and sexual *L. nydeggeri* and *L. ovalifolium*. DEGs were filtered by |log2 fold-change (Log_2FC)| >2. Number of overlapping and specific DEGs in: **[A]** *L. nydeggeri* in S3/S4 relative to S1 (yellow) and *L. ovalifolium* in S3/S4 relative to S1 (pink); **[B]** *L. nydeggeri* in S3/S4 relative to S2 (yellow) and *L. ovalifolium* in S3/S4 relative to S2 (pink).

Analyzing DEGs between apomictic and sexual plants at the same stages, there was almost the same number of DEGs in comparisons between S1 and S2 (1287 and 1274) (Figure 3.34A-B). Among these, the number of overlapping DEGs was higher in S1 (602 vs. 407), but the number of specific DEGs was higher in S2 (611 vs. 207). In the comparisons with apomictic plants in S2, the number of overlapping DEGs was much higher with sexual plants than with facultative apomictic plants (838 vs. 102) (Figure 3.34C). Moreover, the percentage of specific DEGs was the highest in *L. ovalifolium* (2562; 66% of total DEGs), followed by *Limonium nydeggeri* (272; 7% of total DEGs) and finally by *L. dodartii* (71; 2% of total DEGs) (Figure 3.34C).

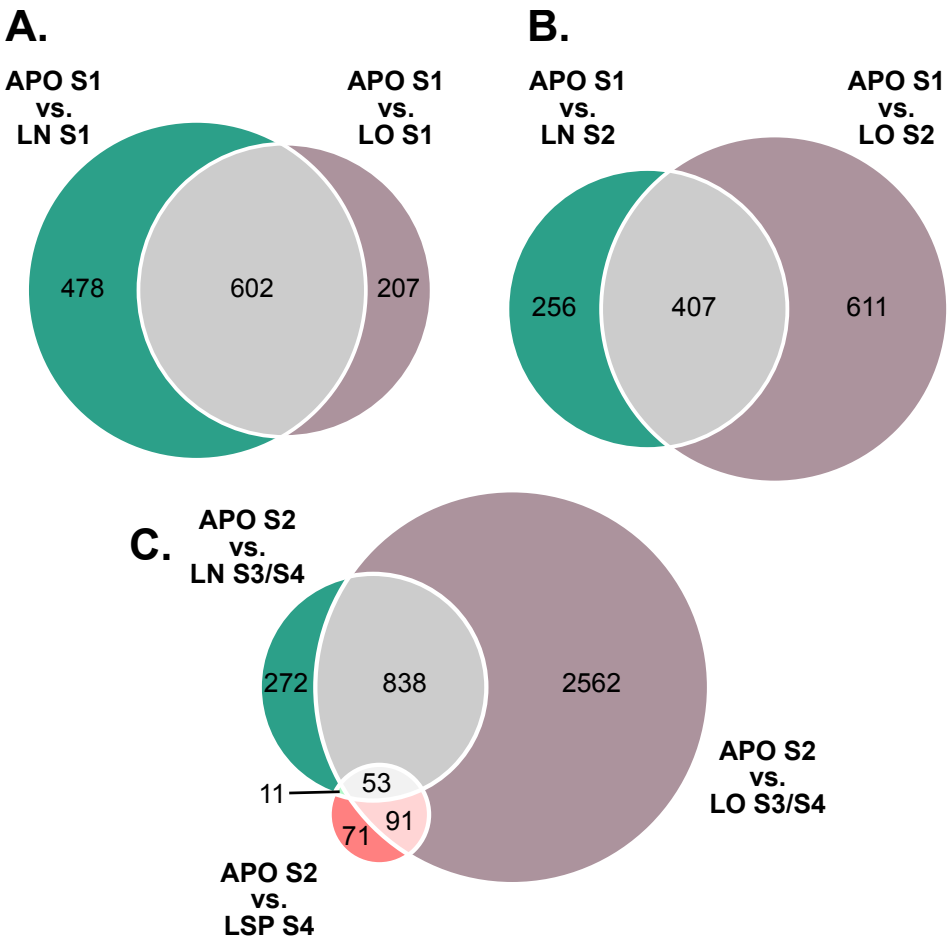

Figure 3.34. Weighted Venn diagrams of specific and overlapping differentially expressed genes (DEGs) found in *Limonium* plants, apomictic *L. multiflorum*, facultative apomictic *Limonium dodartii* and sexual *L. nydeggeri* and *L. ovalifolium*. DEGs were filtered by |log2 fold-change (Log_2FC)| >2. Number of overlapping and specific DEGs in: **[A]** apomictic in S1 relative to *L. nydeggeri* (green) and relative to *L. ovalifolium* (purple) **[B]** apomictic in S1 relative to *L. nydeggeri* in S2 (green) and relative to *L. ovalifolium* in S2 (purple); **[C]** apomictic in S2 relative to *L. nydeggeri* in S3/S4 (green), to *L. ovalifolium* in S3/S4 (purple) and to *Limonium dodartii* in S4 (red).

In general, TF potentially related to male sterility were mostly associated with up-regulated DEGs in apomictic plants in both S1 and S2 relative to all stages of sexual and facultative apomictic plants (Figure 3.35). These TF were classified into 10 major families: AP2/ERF, bHLH, bZIP, C2C2, C2H2, HB, MADS, MYB, NAC and WRKY. The most representative families in all comparisons were WRKY, which play a role in plant disease resistance, abiotic stress responses, nutrient deprivation, senescence, seed and trichome development, embryogenesis, as well as additional developmental and hormone-controlled processes, MYB, which are essential in regulatory networks controlling development, metabolism and responses to biotic and abiotic stresses, and AP2/ERF, which are also key regulators of several abiotic stresses and respond to multiple hormones.

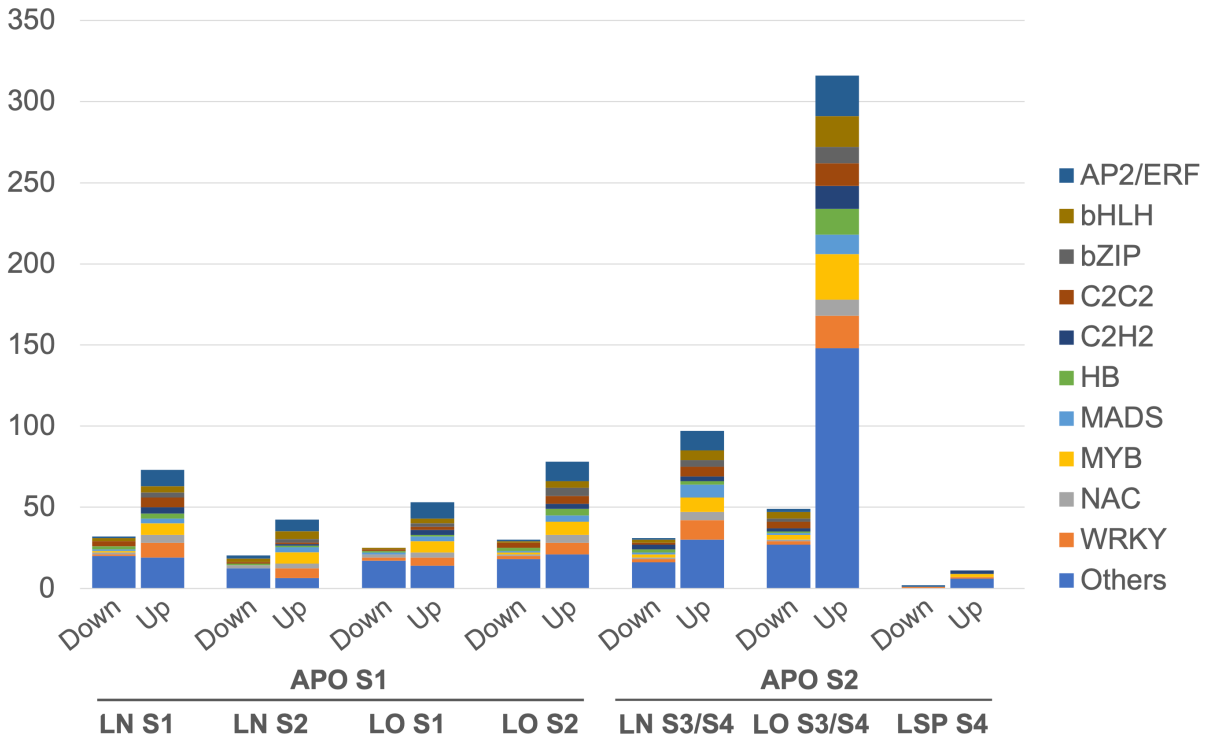

Figure 3.35. Distribution of differentially expressed transcription factors potentially related to male sterility classified into the 10 families with the highest number of differentially expressed genes (DEGs) in *Limonium* plants, namely apomictic *L. multiflorum*, sexual *L. nydeggeri* and *L. ovalifolium*, and facultative apomictic *Limonium dodartii*: AP2/ERF, bHLH, bZIP, C2C2, C2H2, HB, MADS, MYB, NAC and WRKY.

Overall, among DEGs annotated with pollen tube related GO terms, the majority were down-regulated in *L. multiflorum*, related to either sexual and facultative apomictic plants, in both S1/S2 and S3/S4. However, DEGs of *L. multiflorum* in S2 relative to *L. ovalifolium* in S3/S4 were slightly up-regulated (35 vs. 31) (Figure 3.36).

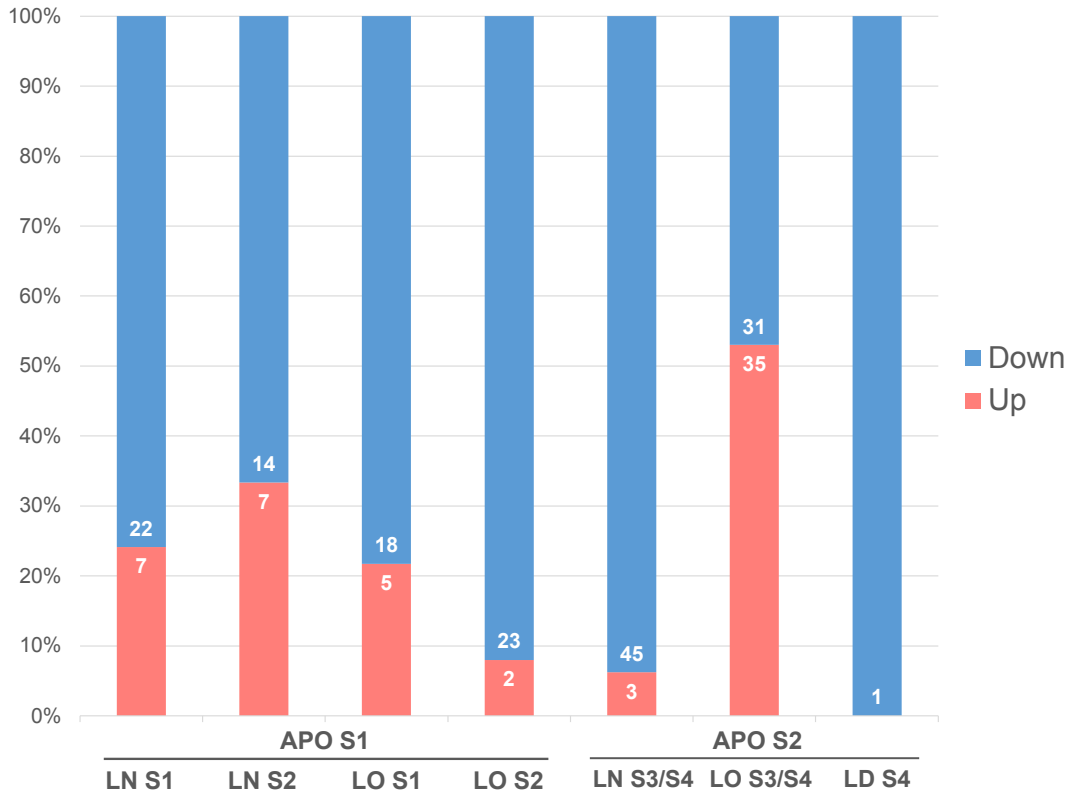

Figure 3.36. Regulation of differentially expressed genes (DEGs) in *Limonium* plants in stages S1, S2, S3/S4 and S4, namely apomictic *L. multiflorum*, sexual *L. nydeggeri* and *L. ovalifolium*, and facultative apomictic *Limonium dodartii*, annotated with Gene Ontology (GO) terms related to pollen tube. DEGs were searched for the following terms: the biological processes pollen tube reception (GO:0010483), pollen tube development (GO:0048868), pollen tube growth (GO:0009860), regulation of pollen tube growth (GO:0080092) and pollen tube adhesion (GO:0009865), and the cellular components pollen tube (GO:0090406) and pollen tube tip (GO:0090404). DEGs represent the number of significant genes found to be differently expressed in each comparison (namely: APO S2 vs. APO S1; LN S2 vs. LN S1; LN S3/S4 vs. LN S1; LN S3/S4 vs. LN S2; LO S2 vs. LO S1; LO S3/S4 vs. LO S1; LO S3/S4 vs. LO S2; APO S1 vs. LN S1; APO S1 vs. LN S2; APO S1 vs. LO S1; APO S1 vs. LO S2; APO S2 vs. LN S3/S4; APO S2 vs. LO S3/S4; APO S2 vs. LSP S4).

Some DEGs were searched between multiple comparisons, although presenting opposite regulation. Among DEGs of *L. multiflorum* in S1, only 24 were down-regulated in relative to *L. nydeggeri* but up-regulated in *L. ovalifolium* (Figure 3.37A). However, among DEGs of *L. multiflorum* in S2 there were more opposite regulations across comparisons at S3/S4. While 124 DEGs were down-regulated relative to *L. nydeggeri*, but up-regulated relative to *L. ovalifolium*, 21 were up-regulated in *L. nydeggeri*, but down-regulated in *L. ovalifolium* (Figure 3. 37B). Also, while 20 DEGs were up-regulated relative to *L. nydeggeri*, but down-regulated relative to *L. dodartii*, 3 DEGs were down-regulated relative to *L. nydeggeri*, but up-regulated relative to *L. dodartii* (Figure 3. 37C). Furthermore, while 65 DEGs were up-regulated relative to *L. nydeggeri*, but down-regulated relative to *L. dodartii*, 3 DEGs were down-regulated relative to *L. nydeggeri*, but up-regulated relative to *L. dodartii* (Figure 3. 37D). Among the 124 DEGs of *L. multiflorum* in S2 which were down-regulated relative to *L. nydeggeri* in S3/S4, but up-regulated relative to *L. ovalifolium* in S3/S4, it was found an enrichment of the WikiPathways Flavonoid biosynthesis.

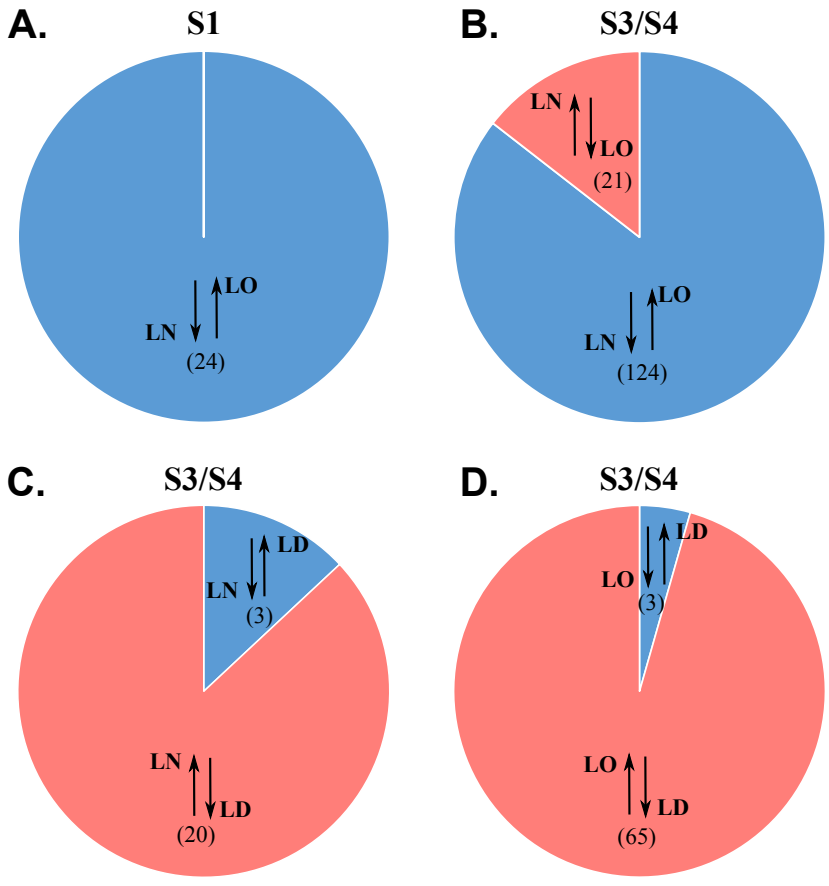

Figure 3.37. Total number of differentially expressed genes (DEGs) shared between two comparisons with opposite regulation, namely apomictic plants (*Limonium multiflorum*) in S1 relative to sexual (*L. nydeggeri* and *L. ovalifolium*) in S1 and apomictic plants (*Limonium multiflorum*) in S2 relative to sexual (*L. nydeggeri* and *L. ovalifolium*) in S3/S4 and relative to facultative apomictic plants (*L. dodartii*) in S3/S4. Annotations of the transcripts *Arabidopsis th*aliana homologs were retrieved from UniprotKB. [↑: Up-regulated; ↓: Down-regulated].

A set of the DEGs found in apomictic plants, either in S1 and S2, were found to be totally knocked out (KO) in both sexual plants (*L. nydeggeri* and *L. ovalifolium*), in S1, S2 and S3/S4, and facultative apomictic plants (*L. dodartii*) in S4. In apomictic DEGs in S1, 16 and 34 DEGs were KO in *L. nydeggeri* in S1 and S2, respectively, and 22 and 5 DEGs were KO in *L. ovalifolium* in S1 and S2, respectively. In apomictic DEGs in S2, 30 and 2170 DEGs were KO in *L. nydeggeri* and *L. ovalifolium*, respectively, both in S3/S4. Also, 1 DEG was KO in *L. dodartii* in S4. Among the DEGs in *L. multiflorum* in S1 which were KO in *L. ovalifolium* in S1 it was found and enrichment in Tryptophan metabolism according to KEGG database. Furthermore, DEGs in *L. multiflorum* in S2 which were KO in *L. nydeggeri* in S3/S4 were enriched in Ethylene signalling pathway, according to WikiPathways, and in Aminoacyl-tRNA biosynthesis and Lysine degradation, according to KEGG.

Chapter 4

4. Discussion & Conclusion

The growing volume of transcriptomic data can open doors to major biological breakthroughs. However, RNA-seq for differential expression analysis is still in development and, although several software has been developed for this technique, there is no ideal method or tool for all the analyses and no clear consensus about its best practices. This difficulty is aggravated when a reference genome is lacking, due to the several issues related with transcriptome assembly, such as redundancy and the need of high computer power. Generally, it's up to the user to choose the better and more up-to-date software according to each dataset and main objectives of the research, which can greatly influence the outcome of the experiment. As such, the use of appropriate parameters and tools is of the utmost importance in a proper bioinformatics analysis. Nevertheless, it's also important to ensure that the laboratory procedures that precede it are suitable, as the quality of the data also has a great impact on the analysis results. Datasets containing an appropriate number of biological replicates are important for an unbiased analysis of the data, an improved quality of libraries is key to reduce duplicates through the reduction of the number of PCR cycles during preparation and an increased sequencing depth can help mitigate the false discovery rate by reducing the genetic background and increasing sequencing sampling. The main purpose of this thesis was the detection and functional annotation of significant DEGs, using the best parameters, tools and software for each dataset, according to the research goals. Thus, the analysis presented here sets up a starting point for future transcriptome studies in *Casuarina glauca*, *Coffea arabica*, *C. canephora* and *Limonium spp*.

After sequencing, quality control is an essential step to ensure the quality of the samples to correctly detect DEGs. The FastQC tool, which was used in this project, is a reference and useful tool to get an overview of samples quality. In general, reads presented good quality, leading only to moderate filtering and trimming, which aimed to maximize general quality, remove adapters and to avoid excessively multiple alignments in the alignment step, allowing libraries to keep their length and depth at adequate levels, according to the state of the art. However, although this step is important to eliminate very short and low-quality reads, most recent studies suggests that the trimming of adapters should be omitted in good quality samples, since they can be eliminated through *soft-clipped* in the assembly step

with the most common assemblers (*e.g.*, STAR) (Liao & Shi, 2020). By doing so, adapters can be eliminated from the samples, reducing computational time and the risk of losing reads with relevant biological meaning. Thus, in future analyses, this step can be improved by choosing to eliminate only low-quality bases or other contaminants. Also, since most reference tools used in quality control are optimized for DNA, it would be interesting to develop more software specifically designed for RNA, which would allow the user to confidently perform filtering, managing to keep the greatest number of reads, with the highest possible quality.

RNA-seq robustness can also be affected by the presence of duplicated reads due to the amplification step. However, the most used tools for PCR duplicate removal rely only on the mapping coordinates of sequencing reads, which does not allow distinguishing between PCR duplicates and valid biological read duplicates. As such, since most identical reads reflect biological reality, the removal of duplicates can erroneously eliminate usable reads, particularly from short transcripts and small RNAs (Fu et al., 2018). Nevertheless, while duplication removal can produce a high proportion of false negatives, the absence of this step is biased towards false positive results (Klepikova et al., 2017). Currently, there are tools that allows the estimation of PCR duplication rate, assessing the fraction of duplicates that correspond to natural read duplicates. However, studies show that only a small fraction of read duplicates in RNA-seq data are due to PCR amplification when the library complexity is adequate, reinforcing the need to adjust the library preparation to improve sequencing library complexity, especially in very low input or extremely deep RNA-sequencing projects (Parekh et al. 2016; Bansal, 2017). Plant RNA-seq data frequently contains large amounts of duplicated reads, due to the fact that gene expression is controlled by only a small number of transcripts in many plant tissues, such as leaves. Also, studies have found that a large fraction of computationally identified read duplicates are not from PCR amplification and can be explained by sampling and fragmentation bias, having been observed that their elimination not only does not improve accuracy nor precision, but can additionally worsen the power and the FDR for differential gene expression (Parekh et al. 2016). As such, although the duplication rate can still be used as quality control, in RNA-seq projects PCR duplicates are normally kept in. Nevertheless, today it's possible to use a laboratory technique that allows the identification of reads unquestionably, giving bioinformatics the possibility of differentiating the natural duplication from the duplication generated by amplification, through Unique Molecular Identifiers (UMI). This approach is mainly recommended for very low input samples and very deep sequencing of RNA-seq libraries and can be useful for future studies (Fu et al., 2018). Furthermore, when dealing with non-model organisms, it is often useful to filter out duplicate annotations from multiple genes that map to the same homolog when blasting against reference genomes from related species. This approach often leads to a large reduction in the number of DEGs and, while this process can help to reduce the duplication rate introduced by library preparation, it can also exclude some

important information, and, thus, the choice of filtering these genes should be made based on the duplication rate of the assembly and the number of repeated annotations among DEGs.

The detection and elimination of outliers is another important step to achieve good results in the analysis of differential expression, since the presence of these elements can substantially alter the differences found between the compared samples. The first step in detecting outliers is to use a sufficiently large number of replicates that allow the user to identify those that do not fit into the remaining group of samples. The most adequate number of replicates can be statistically measured considering several factors, especially intra-specific variation. As such, in this project, the detection of outliers was only possible in the *Coffea* datasets, which was performed through visual inspection of the PCA graphics. Very recent studies allow this procedure to be performed using software specially designed for RNA-seq data, which allows the detection of significant differences between samples of the same group, identifying those who, unequivocally, should not be part of it (Kumar et al., 2020; Chen et al., 2020). Thus, in future analyses, it would be interesting to use one of these tools, comparing its results with those of visual inspection to ensure greater quality control.

I chose to use two different approaches when dealing with datasets without replicates. In the analysis of the *Casuarina* dataset, I applied more conservative parameters, filtering the results based on the overlap between 3 different DEGs detection tools, namely *NOISeq*, *DESeq* and *edgeR*. By doing so, I intended to decrease the probability of detecting false positives derived from the inability to assign significance (*p-value*) to the results. However, this methodology also minimizes the number of true positives, allowing only a small number of DEGs to be identified. Thus, this type of procedure allows for a preliminary analysis where only some probable DEGs and their respective functions are highlighted. On the other hand, in the analysis of *Limonium* I decided to use only *edgeR*, which is referred to in the literature as being very efficient, even in cases where there are few or no replicates (Chen et. al., 2016). At the risk of having a higher number of false positives, this approach is more exploratory and can be useful to pinpoint the most important genes and functions that can potentially be differential expressed, broadening the results and allowing for a more general view of the transcriptome, which can be even more useful for future studies.

The *de novo* transcriptome assembly is an invaluable tool to study organisms, especially when a reference genome is missing, since it can be used to perform annotation using related species reference genomes. However, the study of non-model organisms is still one of the great challenges of RNA-seq analysis. Since most research involving non-model species are focused on protein-coding genes, many genes are considered as unrelated or uninformative in annotation databases, due to the fact that only a small proportion of the whole gene set are assigned to pathways (Sundaram et al., 2017). To better understand the genetic processes behind the non-model organisms, it would be important to assemble

good quality reference genomes that could be used in the future to improve transcriptomic studies and better unravel their biological functions and pathways.

Functional analysis of transcriptomic datasets is important to assign biological meaning to data. As such, and although there is a huge number of knowledge databases and annotations tools, the GO database represents one of the most popular in RNA-seq studies (Ashburner et al., 2000). However, whether through direct analysis of the transcriptome annotation, or through third-party software and methods, such as enrichment analysis, the conclusions drawn from GO terms should always consider the prior knowledge about the organism and/or mechanisms under study. Furthermore, many genes annotated with the same GO terms doesn't necessarily have the exact same functions, since many terms are too broad to draw conclusions about the specific functions and pathways in which each gene is involved. Nevertheless, this type of analysis can be very interesting to get a general idea of the main biological processes, molecular functions and cellular components associated with the genes under study, allowing the characterization of pathways according to the DEGs regulation. Another important aspect in this type of analysis is the identification of which genes are essential in each metabolic pathway under study, due to the ability of some genes to enable or unable an entire process, potentiating or disabling other genes. For the future, it would be interesting to develop more tools that consider the role of genes in metabolic pathways and the interactions and interdependencies between them.

In this thesis, I created a considerable set of scripts with the intention of automating processes that would otherwise become very time consuming and processes whose regularity of application justifies its automation, recovering the time invested in its development. These scripts have been documented and developed so that they can be used by other users with similar data and objectives. Given the interconnection of these scripts, it's possible to use them not only independently, but also sequentially. One of the projects for the future could be the creation of a pipeline that automatically uses subsets of these scripts to generate faster results. However, since there is still no single best practices and pipelines for all RNA-seq data analyzes, the individualized way in which the scripts were created allows them to be used more freely according to the data and objectives of each project, without creating one single model for all cases, which would limit its uses. Another promising important development for the future is the creation of a user-friendly web interface that would allow the application of these scripts, individually or sequentially, so that users could applied them easily without having to deal with the code behind them, which could be especially useful for inexperienced researchers, with the potential to significantly speed up the analysis. Reproducibility is a key element in science research, due to the ability of replicating experiments independently of location and users. As such, all data produced by this work is publicly available (or soon to be published), alongside the full computational analysis workflow, which is clearly described in this manuscript. However, different releases of the same tools and/or the system libraries used by such tools might lead to some

reproducibility issues, which should be accounted for and carefully addressed. Reproducibility is a key element in science research, due to the ability of replicating experiments independently of location and users. As such, all data produced by this work is publicly available (or soon to be published), alongside the full computational analysis workflow, which is clearly described in this manuscript. However, different releases of the same tools and/or the system libraries used by such tools might lead to some reproducibility issues, which should be accounted for and carefully addressed.

One of the biggest concerns of RNA-seq analysis is that the choice of tools and parameters depends heavily on the user. Thus, the quality of the analysis is not only reflected by the quality of the data or equipment, but also by the experience and competence of the users. As such, bearing in mind that this is an area in constant expansion, one of the most important steps is to always update knowledge according to the state of the art to guarantee the best quality of results. Since there is no gold standard for RNA-seq data analysis, it is very important that the user knows the current best practices and tools available for each data type, as well as how to interpret the results according to their context. Moreover, bioinformatics should always work with other scientists whose expertise can contribute to find the most accurate biological meaning of the results. Ideally, this partnership should start before the data collection, so that they can make decisions together about RNA-seq methods, namely sequencing platforms, read depth and length, and number of replicates of each sample, according to the available budget and to the type of libraries needed, to better define goals, minimize bias, optimize sequencing quality, manage expectations, and improve the overall analysis results.

This project contributes to a better understanding of the different expression profiles of the species in investigation, according to each growth conditions, allowing for further studies and integration with other omics, that undoubtedly expand the application of RNA-seq. More specifically, this work allowed to uncover some mechanisms of gene expression associated with the resistance of *Casuarina glauca*, which can be used in the future to develop comprehensive stress tolerance studies of these species and related species. Although there are already several studies with *Coffea*, the genetic implications of climate change in this genus are not yet fully known. With this project, it was possible to study some of the main genes of its two most commercially traded species related with stress tolerance and susceptibility, namely elevated CO_2 and high temperatures, which can greatly change with global warming. Moreover, since the mechanisms behind the genetic activation and inactivation of apomictic phenomena in *Limonium* are still unknown, this work can be seen as a first step in exploring this adaptation, which can be further developed in future studies.

Over the past decade, RNA-seq has become an essential method for analyzing transcriptomes and mRNA splicing. Nowadays, RNA-seq methods can be used to study many different aspects of RNA biology, including single-cell gene expression, translation, and RNA structure. New applications such

as spatial transcriptomics (spatialomics), new technologies for long, direct RNA-seq reading and better computational tools for data analysis are currently being explored and developed, contributing to a more complete understanding of RNA mechanisms, from the location of transcription to the folding and intermolecular interactions that govern its function (Stark et al., 2019). Single-cell sequencing is becoming standard in many laboratories and advances in *in situ* RNA-seq and imaging methods have already made it possible to generate transcriptome data similar to the amounts of data available from droplet-based single-cell methods (Hwang et al., 2018). Furthermore, the revolution from bulk RNA sequencing to single-molecular, single-cell and spatial transcriptome approaches has enabled increasingly accurate, individual cell resolution incorporated with spatial information, which is likely to be widely adopted in the future if its technical limitations can be overcome. It's also anticipated that the development of long-read sequencing methods will replace the Illumina short-read RNA-seq as the default method for a substantial proportion of users. Although some improvements are still needed, namely in throughput increase and error rates reduction, if it becomes as affordable and reliable as short-read, the advantages of long-read sequencing are such that it is likely to be the preferred choice in the future (Amarasinghe et al., 2020). Moreover, chromatin structure technologies, such as chromatin conformation capture analysis (3C) and its several derivatives including circular chromosome conformation capture (4C), carbon copy chromosome conformation capture (5C), ChIP-Loop, Hi-C and capture Hi-C, have been developed and improved to detect chromatin structure as well as unknown interacting regions, which can be combined with RNA-seq analysis to detect structure variation-related differentially expressed genes (Han et al., 2018). Due to foreseeable major advances in technology, general predictions about the development of RNA-seq over the next decade are likely to be too conservative. Nevertheless, RNA-seq is expected to develop fast and to open doors for an unprecedently knowledge of the architecture and functionality of the cell, unraveling multiple areas of biology.

References

1. Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P., & Coombes, K. (2019). A protocol to evaluate RNA sequencing normalization methods. *BMC bioinformatics, 20*(Suppl 24), 679. https://doi.org/10.1186/s12859-019-3247-x
2. Abu-Jamous, B., & Kelly, S. (2018). Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome biology, 19*(1), 172. https://doi.org/10.1186/s13059-018-1536-8
3. Adam, F., & Weeks, N. (2020, September 29). Best practices for de novo transcriptome assembly with Trinity. Retrieved March 28, 2021, from https://informatics.fas.harvard.edu/best-practices-for-de-novo-transcriptome-assembly-with-trinity.html
4. Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology, 21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5
5. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
6. Ansorge W. J. (2009). Next-generation DNA sequencing techniques. *New biotechnology*, *25*(4), 195–203. https://doi.org/10.1016/j.nbt.2008.12.009
7. Ari Ş., Arikan M. (2016) Next-Generation Sequencing: Advantages, Disadvantages, and Future. In: Hakeem K., Tombuloğlu H., Tombuloğlu G. (eds) Plant Omics: Trends and Applications. Springer, Cham. https://doi.org/10.1007/978-3-319-31703-8_5
8. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics, 25*(1), 25–29. https://doi.org/10.1038/75556
9. Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic acids research, 38*(14), 4570–4578. https://doi.org/10.1093/nar/gkq211
10. Austin, C. P. (2014). Bioinformatics. *National Human Genome Research Institute.* Retrieved February 19, 2021, from https://www.genome.gov/genetics-glossary/Bioinformatics
11. Babarinde, I. A., Li, Y., & Hutchins, A. P. (2019). Computational Methods for Mapping, Assembly and Quantification for Coding and Non-coding Transcripts. *Computational and structural biotechnology journal, 17*, 628–637. https://doi.org/10.1016/j.csbj.2019.04.012
12. Babraham bioinformatics. (2010, April 26). FastQC: a quality control tool for high throughput sequence data. Retrieved March 25, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
13. Baccarella, A., Williams, C. R., Parrish, J. Z., & Kim, C. C. (2018). Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. *BMC bioinformatics, 19*(1), 423. https://doi.org/10.1186/s12859-018-2445-2
14. Baichoo, S., & Ouzounis, C. A. (2017). Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. *Bio Systems, 156-157*, 72–85. https://doi.org/10.1016/j.biosystems.2017.03.003
15. Bansal V. (2017). A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC bioinformatics, 18*(Suppl 3), 43. https://doi.org/10.1186/s12859-017-1471-9

16. BBC Research. (2020, May). Global Next-Generation Sequencing (NGS) Services Market. Retrieved February 19, 2021, from https://www.bccresearch.com/partners/verified-market-research/next-generation-sequencing-services-market.html

17. Bohnert, R., Behr, J., & Rätsch, G. (2009). Transcript quantification with RNA-Seq data. *BMC Bioinformatics*, *10*(Suppl 13), P5. https://doi.org/10.1186/1471-2105-10-S13-P5

18. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

19. Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic acids research, 41(*Web Server issue), W29–W33. https://doi.org/10.1093/nar/gkt282

20. Bush S. J. (2020). Read trimming has minimal effect on bacterial SNP-calling accuracy. *Microbial genomics, 6*(12), 10.1099/mgen.0.000434. https://doi.org/10.1099/mgen.0.000434

21. Capobianco E. (2014). RNA-Seq Data: A Complexity Journey. *Computational and structural biotechnology journal, 11*(19), 123–130. https://doi.org/10.1016/j.csbj.2014.09.004

22. Chandramohan, R., Wu, P. Y., Phan, J. H., & Wang, M. D. (2013). Benchmarking RNA-Seq quantification tools. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2013, 647–650. https://doi.org/10.1109/EMBC.2013.6609583

23. Chatterjee, A., Ahn, A., Rodger, E. J., Stockwell, P. A., & Eccles, M. R. (2018). A Guide for Designing and Analyzing RNA-Seq Data. *Methods in molecular biology (Clifton, N.J.)*, *1783,* 35–80. https://doi.org/10.1007/978-1-4939-7834-2_3

24. Chen, F., Dong, M., Ge, M., Zhu, L., Ren, L., Liu, G., & Mu, R. (2013). The history and advances of reversible terminators used in new generations of sequencing technology. *Genomics, proteomics & bioinformatics*, *11*(1), 34–40. https://doi.org/10.1016/j.gpb.2013.01.003

25. Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., & Gu, J. (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC bioinformatics, 18*(Suppl 3), 80. https://doi.org/10.1186/s12859-017-1469-3

26. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England), 34(17)*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

27. Chen, X., Zhang, B., Wang, T., Bonni, A., & Zhao, G. (2020). Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC bioinformatics, 21*(1), 269. https://doi.org/10.1186/s12859-020-03608-0

28. Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research, 5, 1438. https://doi.org/10.12688/f1000research.8987.2

29. Chhangawala, S., Rudy, G., Mason, C. E., & Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome biology, 16*(1), 131. https://doi.org/10.1186/s13059-015-0697-y

30. Chung, M., Teigen, L., Liu, H., Libro, S., Shetty, A., Kumar, N., Zhao, X., Bromley, R. E., Tallon, L. J., Sadzewicz, L., Fraser, C. M., Rasko, D. A., Filler, S. G., Foster, J. M., Michalski, M. L., Bruno, V. M., & Dunning Hotopp, J. C. (2018). Targeted enrichment outperforms other enrichment techniques and enables more multi-species RNA-Seq analyses. *Scientific reports, 8*(1), 13377. https://doi.org/10.1038/s41598-018-31420-7

31. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research, 38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

32. Cole, B. S., Hall, M. A., Urbanowicz, R. J., Gilbert-Diamond, D., & Moore, J. H. (2017). Analysis of Gene-Gene Interactions. *Current protocols in human genetics, 95*, 1.14.1–1.14.10. https://doi.org/10.1002/cphg.45

33. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8

34. Corley, S. M., MacKenzie, K. L., Beverdam, A., Roddam, L. F., & Wilkins, M. R. (2017). Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. *BMC genomics, 18*(1), 399. https://doi.org/10.1186/s12864-017-3797-0

35. Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, *12*(12), e0190152. https://doi.org/10.1371/journal.pone.0190152

36. Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of biomedicine & biotechnology, 853916*. https://doi.org/10.1155/2010/853916

37. DaMatta, F. M., Ramalho, J. D. C. (2006) Impacts of drought and temperature stress on coffee physiology and production: A review. *Braz. J. Plant Physiol*, *18*:55–81. https://doi.org/10.1590/S1677-04202006000100006

38. De Las Rivas, J., & Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology, 6*(6), e1000807. https://doi.org/10.1371/journal.pcbi.1000807

39. Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS one, 8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

40. Deng, Y., Lei, Q., Tian, Q., Xie, S., Du, X., Li, J., Wang, L., & Xiong, Y. (2014). De novo assembly, gene annotation, and simple sequence repeat marker development using Illumina paired-end transcriptome sequences in the pearl oyster Pinctada maxima. *Bioscience, biotechnology, and biochemistry, 78*(10), 1685–1692. https://doi.org/10.1080/09168451.2014.936351

41. Di Resta, C., Galbiati, S., Carrera, P., & Ferrari, M. (2018). Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. *EJIFCC*, *29*(1), 4–14.

42. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England), 29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

43. Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current protocols in bioinformatics, 51*, 11.14.1–11.14.19. https://doi.org/10.1002/0471250953.bi1114s51

44. Duarte, G. T., Volkova, P. Y., & Geras'kin, S. A. (2021). A Pipeline for Non-model Organisms for *de novo* Transcriptome Assembly, Annotation, and Gene Ontology Analysis Using Open Tools: Case Study with Scots Pine. *Bio-protocol*, *11*(3), e3912. https://doi.org/10.21769/BioProtoc.3912

45. Duro N, Batista-Santos P, da Costa M, Castro IV, Ramos M, Ramalho JC, Pawlowski K, Máguas C, Ribeiro-Barros A (2016) The impact of salinity on the symbiosis between *Casuarina glauca* Sieb. ex Spreng. and N_2-fixing Frankia bacteria based on the analysis of Nitrogen and Carbon metabolism. *Plant Soil 398*: 327–337. https://doi.org/10.1007/s11104-015-2666-3

46. Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics, 19*(5), 776–792. https://doi.org/10.1093/bib/bbx008

47. Fang, Z., Martin, J., & Wang, Z. (2012). Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell & bioscience, 2(1), 26*. https://doi.org/10.1186/2045-3701-2-26

48. Fonseca, N. A., Marioni, J., & Brazma, A. (2014). RNA-Seq gene profiling--a systematic empirical comparison. *PloS one, 9*(9), e107026. https://doi.org/10.1371/journal.pone.0107026

49. Fu, G. K., Xu, W., Wilhelmy, J., Mindrinos, M. N., Davis, R. W., Xiao, W., & Fodor, S. P. (2014). Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proceedings of the National Academy of Sciences of the United States of America, 111*(5), 1891–1896. https://doi.org/10.1073/pnas.1323732111

50. Fu, Y., Wu, P. H., Beane, T., Zamore, P. D., & Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC genomics, 19(*1), 531. https://doi.org/10.1186/s12864-018-4933-1

51. Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics (Oxford, England), 36(*8), 2628–2629. https://doi.org/10.1093/bioinformatics/btz931

52. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. & Regev, A. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol, 29*(7), 644–652. https://doi.org/10.1038/nbt.1883

53. Graça, I., Mendes, V. M., Marques, I., Duro, N., da Costa, M., Ramalho, J. C., Pawlowski, K., Manadas, B., Pinto Ricardo, C. P., & Ribeiro-Barros, A. I. (2019). Comparative Proteomic Analysis of Nodulated and Non-Nodulated *Casuarina glauca* Sieb. ex Spreng. Grown under Salinity Conditions Using Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH-MS). *International Journal of Molecular Sciences*, *21*(1), 78. https://doi.org/10.3390/ijms21010078

54. Gurson, N. (2015, September 9). When Do I Use Sanger Sequencing vs NGS? *Retrieved January 28*, 2021, from https://www.thermofisher.com/blog/behindthebench/when-do-i-use-sanger-sequencing-vs-ngs-seq-it-out-7/

55. Gutierrez-Gonzalez, J. J., & Garvin, D. F. (2017). De Novo Transcriptome Assembly in Polyploid Species. *Methods in molecular biology (Clifton, N.J.), 1536*, 209–221. https://doi.org/10.1007/978-1-4939-6682-0_15

56. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols, 8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

57. Han, J., Zhang, Z., & Wang, K. (2018). 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular cytogenetics*, *11*, 21. https://doi.org/10.1186/s13039-018-0368-2

58. Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research, 38*(12), e131. https://doi.org/10.1093/nar/gkq224

59. Hansen, K. D., Irizarry, R. A., & Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics (Oxford, England), 13*(2), 204–216. https://doi.org/10.1093/biostatistics/kxr054

60. Hill, D. P., Smith, B., McAndrews-Hill, M. S., & Blake, J. A. (2008). Gene Ontology annotations: what they mean and where they come from. *BMC bioinformatics, 9* Suppl 5(Suppl 5), S2. https://doi.org/10.1186/1471-2105-9-S5-S2

61. Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. Wiley interdisciplinary reviews. *RNA, 8*(1), 10.1002/wrna.1364. https://doi.org/10.1002/wrna.1364

62. Hoeijmakers, W. A., Bártfai, R., & Stunnenberg, H. G. (2013). Transcriptome analysis using RNA-Seq. *Methods in molecular biology (Clifton, N.J.)*, *923*, 221–239. https://doi.org/10.1007/978-1-62703-026-7_15

63. Hölzer, M., & Marz, M. (2019). *De novo* transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience, 8*(5), giz039. https://doi.org/10.1093/gigascience/giz039

64. Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine, 50*(8), 1–14. https://doi.org/10.1038/s12276-018-0071-8

65. Illumina. (2017, May). Advantages of paired-end and single-read sequencing. Retrieved March 24, 2021, from https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

66. Inkscape Project. (2020). Inkscape. Retrieved from https://inkscape.org

67. Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., & von Mering, C. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research, 37*(Database issue), D412–D416. https://doi.org/10.1093/nar/gkn760

68. Jiang R. (2013) Gene-Gene Interaction. In: Gellman M.D., Turner J.R. (eds) Encyclopedia of Behavioral Medicine. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1005-9_690

69. Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research, 28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

70. Ke, R., Mignardi, M., Hauling, T., & Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Human mutation*, *37*(12), 1363–1367. https://doi.org/10.1002/humu.23051

71. Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N., & Shoaib, M. (2018). A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. Evolutionary bioinformatics online, 14, 1176934318758650. https://doi.org/10.1177/1176934318758650

72. Khetani, R., & Mistry, M. (2017). Count normalization with DESeq2. Retrieved March 28, 2021, from https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

73. Kim, I. V., Ross, E. J., Dietrich, S., Döring, K., Sánchez Alvarado, A., & Kuhn, C. D. (2019). Efficient depletion of ribosomal RNA for RNA sequencing in planarians. *BMC genomics, 20*(1), 909. https://doi.org/10.1186/s12864-019-6292-y

74. Klepikova, A. V., Kasianov, A. S., Chesnokov, M. S., Lazarevich, N. L., Penin, A. A., & Logacheva, M. (2017). Effect of method of deduplication on estimation of differential gene expression using RNA-seq. *PeerJ, 5*, e3091. https://doi.org/10.7717/peerj.3091

75. Kukurba K. R., Montgomery S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, *11*, 951–969. https://doi.org/10.1101/pdb.top084970

76. Kulkarni, N., Alessandrì, L., Panero, R., Arigoni, M., Olivero, M., Ferrero, G., Cordero, F., Beccuti, M., & Calogero, R. A. (2018). Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics*, *19* (Suppl 10), 349. https://doi.org/10.1186/s12859-018-2296-x

77. Kumar, G., Ertel, A., Feldman, G., Kupper, J., & Fortina, P. (2020). iSeqQC: a tool for expression-based quality control in RNA sequencing. *BMC bioinformatics, 21*(1), 56. https://doi.org/10.1186/s12859-020-3399-8

78. Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M., & Maza, E. (2018). Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size. *Frontiers in plant science, 9*, 108. https://doi.org/10.3389/fpls.2018.00108

79. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

80. Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic acids research, 39*(Database issue), D19–D21. https://doi.org/10.1093/nar/gkq1019

81. Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics, 12*, 323. https://doi.org/10.1186/1471-2105-12-323

82. Li, J., Witten, D. M., Johnstone, I. M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics (Oxford, England), 13*(3), 523–538. https://doi.org/10.1093/biostatistics/kxr031

83. Li, J., & Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical methods in medical research, 22*(5), 519–536. https://doi.org/10.1177/0962280211428386

84. Li, J. R., Liu, C. C., Sun, C. H., & Chen, Y. T. (2018). Plant stress RNA-seq Nexus: a stress-specific transcriptome database in plant cells. *BMC genomics*, *19*(1), 966. https://doi.org/10.1186/s12864-018-5367-5

85. Li, X., Cooper, N., O'Toole, T. E., & Rouchka, E. C. (2020). Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies. *BMC genomics, 21*(1), 75. https://doi.org/10.1186/s12864-020-6502-7

86. Lin, Y., Golovnina, K., Chen, Z. X., Lee, H. N., Negron, Y. L., Sultana, H., Oliver, B., & Harbison, S. T. (2016). Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. *BMC genomics, 17*, 28. https://doi.org/10.1186/s12864-015-2353-z

87. Liao, Y., & Shi, W. (2020). Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. *NAR genomics and bioinformatics, 2*(3), lqaa068. https://doi.org/10.1093/nargab/lqaa068

88. Lischer, H., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC bioinformatics, 18*(1), 474. https://doi.org/10.1186/s12859-017-1911-6

89. Liu, D., & Graber, J. H. (2006). Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. *BMC bioinformatics, 7*, 77. https://doi.org/10.1186/1471-2105-7-77

90. Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication?. *Bioinformatics (Oxford, England), 30*(3), 301–304. https://doi.org/10.1093/bioinformatics/btt688

91. Loraine, A. E., McCormick, S., Estrada, A., Patel, K., & Qin, P. (2013). RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. *Plant physiology*, *162*(2), 1092–1109. https://doi.org/10.1104/pp.112.211441

92. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology, 15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

93. Ma, F., Fuqua, B. K., Hasin, Y., Yukhtman, C., Vulpe, C. D., Lusis, A. J., & Pellegrini, M. (2019). A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC genomics, 20*(1), 9. https://doi.org/10.1186/s12864-018-5393-3

94. Manga, P., Klingeman, D. M., Lu, T. Y., Mehlhorn, T. L., Pelletier, D. A., Hauser, L. J., Wilson, C. M., & Brown, S. D. (2016). Replicates, Read Numbers, and Other Important Experimental Design Considerations for Microbial RNA-seq Identified Using Bacillus thuringiensis Datasets. *Frontiers in microbiology, 7*, 794. https://doi.org/10.3389/fmicb.2016.00794

95. Marchant, A., Mougel, F., Mendonça, V., Quartier, M., Jacquin-Joly, E., da Rosa, J. A., Petit, E., & Harry, M. (2016). Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus. *Insect biochemistry and molecular biology, 69*, 25–33. https://doi.org/10.1016/j.ibmb.2015.05.009

96. Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., A Miller, R., Digles, D., Lopes, E. N., Ehrhart, F., Dupuis, L. J., Winckers, L. A., Coort, S. L., Willighagen, E. L., Evelo, C. T., Pico, A. R., & Kutmon, M. (2021). WikiPathways: connecting communities. *Nucleic acids research, 49*(D1), D613–D621. https://doi.org/10.1093/nar/gkaa1024

97. Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature reviews. Genetics*, *12*(10), 671–682. https://doi.org/10.1038/nrg3068

98. Meera K. B., Khan, M. A., & Khan, S. T. (2019). Next-Generation Sequencing (NGS) Platforms: An Exciting Era of Genome Sequence Analysis. *Microbial Genomics in Sustainable Agroecosystems*, *89*–109. https://doi.org/10.1007/978-981-32-9860-6_6

99. Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

100. Moore J. H. (2007). Bioinformatics. *Journal of cellular physiology*, *213*(2), 365–369. https://doi.org/10.1002/jcp.21218

101. Naranpanawa, D., Chandrasekara, C., Bandaranayake, P., & Bandaranayake, A. U. (2020). Raw transcriptomics data to gene specific SSRs: a validated free bioinformatics workflow for biologists. *Scientific reports, 10*(1), 18236. https://doi.org/10.1038/s41598-020-75270-8

102. Ngom, M., Gray, K., Diagne, N., Oshone, R., Fardoux, J., Gherbi, H., Hocher, V., Svistoonoff, S., Laplaze, L., Tisa, L. S., Sy, M. O., & Champion, A. (2016). Symbiotic Performance of Diverse Frankia Strains on Salt-Stressed Casuarina glauca and Casuarina equisetifolia Plants. *Frontiers in plant science, 7*, 1331. https://doi.org/10.3389/fpls.2016.01331

103. Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews. *Genetics, 12*(2), 87–98. https://doi.org/10.1038/nrg2934

104. Patil, R. (2006). Comparative analysis of parametric, nonparametric and permutation methods for differential expression. *Theses. 433*. https://digitalcommons.njit.edu/theses/433

105. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Scientific reports, 6*, 25533. https://doi.org/10.1038/srep25533

106. Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of clinical medicine*, *9*(1), **132**. https://doi.org/10.3390/jcm9010132

107. Pertsemlidis, A., & Fondon, J. W., 3rd (2001). Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome biology, 2*(10), REVIEWS2002. https://doi.org/10.1186/gb-2001-2-10-reviews2002

108. Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, *93*(2), 105–111. https://doi.org/10.1016/j.ygeno.2008.10.003

109. Ramalho, J. C., Rodrigues, A. P., Semedo, J. N., Pais, I. P., Martins, L. D., Simões-Costa, M. C., Leitão, A. E., Fortunato, A. S., Batista-Santos, P., Palos, I. M., Tomaz, M. A., Scotti-Campos, P., Lidon, F. C., & DaMatta, F. M. (2013). Sustained photosynthetic performance of Coffea spp. under long-term enhanced [CO2]. *PloS one, 8*(12), e82712. https://doi.org/10.1371/journal.pone.0082712

110. Ramalho, J. C., DaMatta, F. M., Rodrigues, A. P., Scotti-Campos, P., Pais, I., Batista-Santos, P., Partelli, F. L., Ribeiro, A., Lidon, F. C., Leitão, A. E. (2014). Cold impact and acclimation response of Coffea spp. plants. Theor. Exp. *Plant Physiol, 26*. 5–18. https://doi.org/10.1007/s40626-014-0001-7

111. Ramalho, J. C., Rodrigues, A. P., Lidon, F. C., Marques, L., Leitão, A. E., Fortunato, A. S., Pais, I. P., Silva, M. J., Scotti-Campos, P., Lopes, A., Reboredo, F. H., & Ribeiro-Barros, A. I. (2018). Stress cross-response of the antioxidative system promoted by superimposed drought and cold conditions in Coffea spp. *PloS one, 13*(6), e0198694. https://doi.org/10.1371/journal.pone.0198694

112. Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E., & Liguori, M. J. (2019). Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Frontiers in genetics, 9*, 636. https://doi.org/10.3389/fgene.2018.00636

113. Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. *International journal of proteomics, 147648*. https://doi.org/10.1155/2014/147648

114. Raplee, I. D., Evsikov, A. V., & Marín de Evsikova, C. (2019). Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research. *Journal of personalized medicine, 9*(2), 18. https://doi.org/10.3390/jpm9020018

115. Ren, X., & Kuan, P. F. (2020). Negative binomial additive model for RNA-Seq data analysis. *BMC bioinformatics, 21*(1), 171. https://doi.org/10.1186/s12859-020-3506-x

116. Rhee, S. Y., & Crosby, B. (2005). Biological databases for plant research. *Plant physiology, 138*(1), 1–3. https://doi.org/10.1104/pp.104.900158

117. Rizzetto, S., Eltahla, A. A., Lin, P., Bull, R., Lloyd, A. R., Ho, J., Venturi, V., & Luciani, F. (2017). Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Scientific reports, 7*(1), 12781. https://doi.org/10.1038/s41598-017-12989-x

118. Rizzi, R., Beretta, S., Patterson, M., Pirola, Y., Previtali, M., Della Vedova, G. & Bonizzoni, P. (2019). Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. *Quantitative Biology, 7*(4): 278–292. https://doi.org/10.1007/s40484-019-0181-x

119. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology, 12*(3), R22. https://doi.org/10.1186/gb-2011-12-3-r22

120. Robinson, A. J., Tamiru, M., Salby, R., Bolitho, C., Williams, A., Huggard, S., Fisch, E., Unsworth, K., Whelan, J., & Lewsey, M. G. (2018). AgriSeqDB: an online RNA-Seq database for functional studies of agriculturally relevant plant species. *BMC plant biology, 18*(1), 200. https://doi.org/10.1186/s12870-018-1406-2

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

122. Rodríguez-García, A., Sola-Landa, A., & Barreiro, C. (2017). RNA-Seq-Based Comparative Transcriptomics: RNA Preparation and Bioinformatics. *Methods in molecular biology (Clifton, N.J.)*, *1645*, 59–72. https://doi.org/10.1007/978-1-4939-7183-1_5

123. Rogers, R. L., Cridland, J. M., Shao, L., Hu, T. T., Andolfatto, P., & Thornton, K. R. (2014). Landscape of standing variation for tandem duplications in Drosophila yakuba and Drosophila simulans. *Molecular biology and evolution, 31*(7), 1750–1766. https://doi.org/10.1093/molbev/msu124

124. RStudio Team (2019). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

125. Sathyanarayanan, A., Manda, S., Poojary, M., & Nagaraj, S. H. (2019). Exome Sequencing Data Analysis. In Encyclopedia of bioinformatics and computational biology (Vol. 3, pp. 164-175). Amsterdam: *Elsevier*. https://doi.org/10.1016/B978-0-12-809633-8.20094-0

126. Sayols, S., Scherzinger, D., & Klein, H. (2016). dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC bioinformatics, 17*(1), 428. https://doi.org/10.1186/s12859-016-1276-2

127. Scalabrin, S., Toniutti, L., Di Gaspero, G., Scaglione, D., Magris, G., Vidotto, M., Pinosio, S., Cattonaro, F., Magni, F., Jurman, I., Cerutti, M., Suggi Liverani, F., Navarini, L., Del Terra, L., Pellegrino, G., Ruosi, M. R., Vitulo, N., Valle, G., Pallavicini, A., Graziosi, G., Klein, P. E., Bentley, N., Murray, S., Solano, W., Al Hakimi, A., Schilling, T., Montagnon, C., Morgante, M., Bertrand, B. (2020). A single polyploidization event at the origin of the tetraploid genome of

Coffea arabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Scientific reports, 10*(1), 4642. https://doi.org/10.1038/s41598-020-61216-7

128. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England), 31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

129. Simillion, C., Liechti, R., Lischer, H. E., Ioannidis, V., & Bruggmann, R. (2017). Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC bioinformatics, 18*(1), 151. https://doi.org/10.1186/s12859-017-1571-6

130. Simoneau, J., Dumontier, S., Gosselin, R., & Scott, M. S. (2021). Current RNA-seq methodology reporting limits reproducibility. *Briefings in bioinformatics, 22*(1), 140–145. https://doi.org/10.1093/bib/bbz124

131. Simpson B.M. Preparing Smallholder Farm Families to Adapt to Climate Change. Pocket Guide 2: Managing Crop Resources. *Catholic Relief Services*; Baltimore, MD, USA: 2017

132. Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA (New York, N.Y.), 22*(6), 839–851. https://doi.org/10.1261/rna.053959.115

133. Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345–353. https://doi.org/10.1038/nature24286

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

135. Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research, 4*, 1521. https://doi.org/10.12688/f1000research.7563.2

136. Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M. I., Kingsford, C., & Patro, R. (2020). Alignment and mapping methodology influence transcript abundance estimation. *Genome biology, 21*(1), 239. https://doi.org/10.1186/s13059-020-02151-8

137. Strickler, S. R., Bombarely, A., & Mueller, L. A. (2012). Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American journal of botany*, *99*(2), 257–266. https://doi.org/10.3732/ajb.1100292

138. Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. Nature reviews. *Genetics, 20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

139. Sundaram, A., Tengs, T., & Grimholt, U. (2017). Issues with RNA-seq analysis in non-model organisms: A salmonid example. *Developmental and comparative immunology*, *75*, 38–47. https://doi.org/10.1016/j.dci.2017.02.006

140. Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic acids research, 43*(21), e140. https://doi.org/10.1093/nar/gkv711

141. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome research, 21*(12), 2213–2223. https://doi.org/10.1101/gr.124321.111

142. The UniProt Consortium. (2021, January 28). UniProt: The universal protein knowledgebase in 2021. Retrieved March 30, 2021, from https://www.uniprot.org/help/uniprotkb

143. Tripathi, L. P., Chen, Y., Mizuguchi, K. & Murakami, Y. (2019). Network-Based Analysis for Biological Discovery. In Encyclopedia of Bioinformatics and Computational Biology (Vol. 3, pp. 283-291). Amsterdam: *Elsevier*. https://doi.org/10.1016/B978-0-12-809633-8.20674-2

144. Ulfenborg B. (2019). Vertical and horizontal integration of multi-omics data with miodin. *BMC bioinformatics*, *20*(1), 649. https://doi.org/10.1186/s12859-019-3224-4

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

146. Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC bioinformatics, 17*, 103. https://doi.org/10.1186/s12859-016-0956-2

147. Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research, 7*, 1338. https://doi.org/10.12688/f1000research.15931.2

148. Xiao, T., & Zhou, W. (2020). The third generation sequencing: the advanced approach to genetic diseases. *Translational pediatrics*, *9*(2), 163–173. https://doi.org/10.21037/tp.2020.03.06

149. Yang, C., Wu, P. Y., Tong, L., Phan, J. H., & Wang, M. D. (2015). The impact of RNA-seq aligners on gene expression estimation. ACM-BCB: the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, *Computational Biology and Biomedicine*, 462–471. https://doi.org/10.1145/2808719.2808767

150. Yang, S. F., Lu, C. W., Yao, C. T., & Hung, C. M. (2019). To Trim or Not to Trim: Effects of Read Trimming on the De Novo Genome Assembly of a Widespread East Asian Passerine, the Rufous-Capped Babbler (Cyanoderma ruficeps Blyth). *Genes, 10*(10), 737. https://doi.org/10.3390/genes10100737

151. Yi, L., Pimentel, H., Bray, N. L., & Pachter, L. (2018). Gene-level differential analysis at transcript-level resolution. *Genome biology, 19*(1), 53. https://doi.org/10.1186/s13059-018-1419-z

152. Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one, 9*(1), e78644. https://doi.org/10.1371/journal.pone.0078644

153. Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D., & Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC genomics, 16*(1), 675. https://doi.org/10.1186/s12864-015-1876-7

154. Zhao, S., Zhang, B., Zhang, Y., Gordon, W., Du, S., Paradis, T., Vincent, M. & von Schack, D. (July 27th 2016). Bioinformatics for RNA-Seq Data Analysis, Bioinformatics - Updated Features and Applications, Ibrokhim Y. Abdurakhmonov, IntechOpen, https://doi.org/10.5772/63267. Available from: https://www.intechopen.com/books/bioinformatics-updated-features-and-applications/bioinformatics-for-rna-seq-data-analysis

155. Zhao, S., & Zhang, B. (2016, January 14). Impact of Gene annotation on rna-seq data analysis. Retrieved March 28, 2021, from https://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/impact-of-gene-annotation-on-rna-seq-data-analysis

156. Zhong, C., Mansour, S., Nambiar-Veetil, M., Bogusz, D., & Franche, C. (2013). Casuarina glauca: a model tree for basic research in actinorhizal symbiosis. *Journal of biosciences, 38*(4), 815–823. https://doi.org/10.1007/s12038-013-9370-3

157. Zhou, Q., Su, X., Jing, G., Chen, S., & Ning, K. (2018). RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. *BMC genomics, 19*(1), 144. https://doi.org/10.1186/s12864-018-4503-6

158. Zou, D., Ma, L., Yu, J., & Zhang, Z. (2015). Biological databases for human research. *Genomics, proteomics & bioinformatics, 13*(1), 55–63. https://doi.org/10.1016/j.gpb.2015.01.006