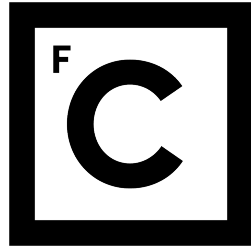


UNIVERSIDADE DE LISBOA  
FACULDADE DE CIÊNCIAS  
DEPARTAMENTO DE BIOLOGIA VEGETAL



**Ciências**  
**ULisboa**

**Alternative splicing detection across different  
tissues in cork oak**

**Pedro Miguel R. Barros**

MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL  
ESPECIALIZAÇÃO EM BIOINFORMÁTICA

Dissertação orientada por:  
Prof. Doutora Cátia Pesquita  
Dr. Marcos Ramos

2017



## Resumo

As florestas de sobreiro (*Quercus suber* L.) são recursos únicos e emblemáticos em Portugal, com elevado impacto económico, ecológico e social. A disponibilidade recente da sequência do genoma de sobreiro forneceu um importante contributo para revitalizar a pesquisa em temas como desenvolvimento de cortiça e melhoramento da planta, assim como promover a competitividade da indústria da cortiça. No entanto, é ainda necessário adicionar mais detalhe à anotação estrutural do genoma, nomeadamente ao nível dos transcritos, incluindo previsão de eventos de *splicing* alternativo. O *splicing* alternativo (AS) é um processo usado durante a expressão génica que origina diferentes variantes de transcritos (isoformas) e produtos proteicos a partir um único gene. No presente estudo, procedemos à análise de dezasseis bibliotecas de RNA-seq, preparadas a partir de quatro tecidos de sobreiro (folhas, felema, entrecasco e xilema), de modo a prever novas formas de AS para genes já previstos e melhorar a anotação estrutural do genoma.

Um protocolo bioinformático foi definido para testar o desempenho do *software* HISAT2 e STAR para mapeamento de *reads* de RNA-seq no genoma de referência, e do *software* Cufflinks e StringTie para (re)construção de transcritos. O alinhamento de *reads* no genoma efetuado com STAR resultou em taxas de mapeamento (de 84,22% a 86,86%) superiores aos resultados atingidos com HISAT2 (73,88% a 76,55%). Assim, os resultados de mapeamento com STAR foram utilizados para a (re)construção de transcritos. O uso do StringTie para este processo foi globalmente mais conservador do que com Cufflinks, gerando menos transcritos novos, mas com melhor cobertura de *reads* por pares de base. Para melhorar a precisão da anotação e reduzir falsos positivos, foi realizado um passo adicional de otimização

com StringTie. Desta otimização resultou uma anotação que prevê a ocorrência de 7 958 novos transcritos (8% dos transcritos totais), dos quais 5 453 são novas isoformas para genes previstos na anotação de referência. Esta nova anotação foi utilizada como referência para estimar a abundância dos transcritos em cada um dos tecidos estudados e efetuar a análise de expressão diferencial. Cerca de 16% de todos os genes expressos nos quatro tecidos e que contêm intrões apresentaram *splicing* alternativo, e os principais eventos de *splicing* foram *alternative acceptor site* e *intron retention*. Grupos de transcritos com expressão diferencial entre os quatro tecidos foram identificados e a análise de enriquecimento funcional confirmou os principais processos biológicos esperados para cada tecido: os transcritos mais expressos nas folhas e no xilema estavam relacionados com a fotossíntese e com transporte, respetivamente; transcritos mais expressos na periderme (felema e entrecasco) mostraram um enriquecimento em categorias funcionais relacionadas com a síntese de suberina e outros componentes de parede celular presentes nas células de cortiça. Estes grupos específicos mostraram também um enriquecimento em transcritos envolvidos na resposta ao stresse (biótico ou abiótico). Nos tecidos que compõem a periderme, este enriquecimento foi observado principalmente no entrecasco, enquanto que no felema foi detetado um enriquecimento em transcritos envolvidos no metabolismo secundário. A presente tese permitiu a definição de um protocolo padrão que poderá ser usado para estudar o *splicing* alternativo no sobreiro e para uma análise mais aprofundada na nova versão do genoma, que estará disponível em breve.

**Palavras Chave:** *Quercus suber*, anotação do genoma, transcrição, isoformas, diferenciação da periderme

## Abstract

Cork oak (*Quercus suber* L.) forests are unique and emblematic resources for Portugal, with high economical, ecological and social significance. The recent availability of the cork oak genome sequence provided an important contribution to reinvigorate research in fundamental topics such as cork development and plant improvement, and to promote the competitiveness of cork industry. Yet, further analysis is required to add detail to genome structure annotation, namely at the transcript level, also taking into account alternative splicing. Alternative splicing (AS) is a process used during gene expression to yield different transcript variants and protein products derived from a single gene. In the present study, we analyzed sixteen RNA-seq libraries prepared from four cork oak tissues (leaf, xylem, phellem and inner bark), in order to predict new AS forms for the already predicted genes and improve genome structural annotation.

A bioinformatics pipeline was defined in order to test the performance of HISAT2 and STAR for read mapping against the reference genome, and Cufflinks and StringTie for transcript assembly. STAR yielded higher mapping efficiencies (84.22% to 86.86%) for the cork oak datasets, as compared to HISAT2 (73.88% to 76.55%), and the corresponding mapping data was selected for transcript assembly. The use of StringTie for this step was globally more conservative than Cufflinks, generating less novel transcripts, but with better support by read per base coverage. A further optimization step was performed using StringTie in order to improve annotation precision. The final transcript annotation was selected from this optimization step, predicting 7,958 novel transcripts (8% of total transcripts in the new annotation), 5,453 of which were novel isoforms for genes in reference annotation. This new annotation was used as reference to estimate

transcript abundance in each tissue and differential expression analysis. Approximately 16% of all intron-containing genes expressed in the four tissues were alternatively spliced and the main event found in the four cork oak tissues was alternative acceptor site, followed by intron retention. Transcript clusters showing differential expression among the four tissues were identified and functional enrichment analysis confirmed the main biological processes expected for each tissue: transcripts highly expressed in leaves and xylem were mostly related to photosynthesis and transport, respectively; transcripts highly expressed in peridermis (phellem and inner bark) showed an enrichment in functional categories related to the synthesis of suberin and other component of cork cell walls. These tissue-specific clusters also showed an enrichment in transcripts involved in the response to stress (biotic or abiotic). Yet, in peridermis, this enrichment was mostly observed in inner bark samples, while phellem samples showed an enrichment in transcripts related to secondary metabolism.

This thesis allowed the definition of a standard workflow that can be used to study alternative splicing in cork oak and used for further analysis on the new improved genome version that will be available soon.

**Keywords:** *Quercus suber*, genome annotation, transcription, isoforms, peridermis differentiation

## Resumo Alargado

As florestas de sobreiro (*Quercus suber* L.) são recursos únicos e emblemáticos em Portugal, com elevado impacto económico, ecológico e social. A disponibilidade recente da sequência do genoma de sobreiro forneceu um importante contributo para revitalizar a pesquisa em temas como desenvolvimento de cortiça e melhoramento da planta, assim como promover a competitividade da indústria da cortiça. No entanto, é ainda necessário adicionar mais detalhe à anotação estrutural do genoma, nomeadamente ao nível dos transcritos, incluindo a previsão de eventos de *splicing* alternativo. O *splicing* alternativo (AS) é um processo usado durante a expressão génica, que resulta da remoção alternativa de exões ou inclusão de intrões nas regiões codificantes do RNA mensageiro, originando diferentes variantes de transcritos (isoformas) a partir um único gene. Um procedimento comum para reconstruir transcritos tendo por base dados de RNA-seq envolve o alinhamento de *reads* sequenciadas no genoma de referência (quando disponível para o organismo em estudo) e (re)construção de transcritos com base no agrupamento de *reads* sobrepostas num determinado *locus*. Esta abordagem permite não só identificar novas isoformas de genes previamente anotados, como também identificar novos genes ainda não anotados. Seguidamente é possível estimar a abundância dos transcritos e determinar a ocorrência de AS nos vários conjuntos de dados. No presente estudo, procedemos à análise de dezasseis bibliotecas de RNA-seq, preparadas a partir de quatro tecidos de sobreiro (folhas, felema, entrecasco e xilema), de modo a prever novas formas AS para genes já previstos e melhorar a anotação estrutural do genoma.

Um protocolo bioinformático foi definido para comparar o desempenho de diferentes algoritmos nos dois passos mais críticos da análise

de RNA-seq: mapeamento de *reads* no genoma de referência, usando HISAT2 e STAR; e construção do transcriptoma, com Cufflinks e StringTie. O alinhamento de *reads* no genoma efectuado com STAR resultou em taxas de mapeamento (de 84,22 % a 86,86 %) superiores aos resultados atingidos com HISAT2 (73,88 % a 76,55 %). Tendo em conta estes resultados, os alinhamentos efectuados com STAR foram utilizados para a (re)construção de transcritos. O uso do StringTie para este passo foi globalmente mais conservador do que com Cufflinks, gerando menos transcritos novos, mas com melhor cobertura de *reads* por pares de base. Para melhorar a precisão da anotação e reduzir falsos positivos, foi realizado um passo adicional de otimização usando StringTie, utilizando parâmetros mais restritivos relacionados com a cobertura mínima de fragmentos considerados e de conexões entre dois exões adjacentes (*splice junctions*). Desta otimização resultou uma anotação que prevê a ocorrência de 7 958 novos transcritos (8% dos transcritos totais), dos quais 5 453 são novas isoformas para genes previstos na anotação de referência. A percentagem de genes com mais de uma isoforma rondou os 10%, havendo um aumento de cerca de 5% relativamente à anotação de referência. Como a versão do genoma do sobreiro utilizado é ainda preliminar (não existindo modelos de genes completamente anotados para determinar a precisão e sensibilidade da nova anotação) a escolha desta nova anotação mais conservadora constituiu uma estratégia para reduzir o número de transcritos incorretamente construídos.

A nova anotação estrutural foi utilizada como referência para estimar a abundância dos transcritos em cada um dos tecidos estudados e efetuar a análise de expressão diferencial. Um total de 25 149 genes (29 296 transcritos) foram considerados expressos nos quatro tecidos analisados, 21 032 dos quais apresentaram uma estrutura de transcrito com mais do que um exão. Destes genes, apenas 3 279 (15.60%) apresentaram *splicing* alternativo, contendo mais do que uma isoforma anotada. Os principais eventos de *splicing* foram *alternative acceptor*



*site e intron retention*. A análise de expressão diferencial identificou um total de 22 449 transcritos diferencialmente expressos entre os quatro tecidos. Estes transcritos foram agrupados de acordo com o seu padrão de expressão, pelo método *k*-means, e foram selecionados cinco grupos de transcritos expressos maioritariamente num único tecido. Foi também selecionado um grupo cuja expressão era maioritariamente encontrada na periderme (camada composta por felema e entrecasco). Estes grupos foram submetidos a uma análise de enriquecimento de termos GO (*Gene Ontology*) de modo a averiguar as classes funcionais com maior representação em cada grupo. A análise de enriquecimento funcional confirmou os principais processos biológicos esperados para cada tecido. Os dois grupos de transcritos mais expressos nas folhas incluíam um total de 990 transcritos, apresentando um enriquecimento significativo de termos associados a fotossíntese e estruturas cloroplastidiais. O grupo que continha transcritos maioritariamente expressos no xilema incluía 488 transcritos, apresentando um enriquecimento em termos relacionados com transporte de metabolitos e iões. O grupo de transcritos mais expressos na periderme (692 transcritos) mostrou um enriquecimento em categorias funcionais relacionadas com a síntese de suberina e outros componentes de parede celular presentes nas células do felema (que originarão a cortiça). Uma vez que o entrecasco não apresenta suberificação das paredes celulares, este resultado sugere um eventual envolvimento deste tecido na produção de componentes de parede de células de felema. Um grupo de 340 transcritos apresentou expressão maioritariamente no felema e um enriquecimento em classes funcionais relacionadas com a síntese de taninos, que preenchem o conteúdo celular destas células durante os primeiros anos de desenvolvimento. O grupo de transcritos expressos maioritariamente no entrecasco (786 transcritos) apresentou um enriquecimento em termos associados com a resposta ao stresse (biótico ou abiótico). Embora termos relacionados com stress tenham sido também encontrados nos grupos específicos de folha e xilema,

nos tecidos que compõem a periderme este enriquecimento foi principalmente observado no entrecasco, sugerindo que este tecido está também envolvido na protecção da planta.

A presente tese seguiu uma abordagem conservadora para (re)construção do transcritoma com base em RNA-seq, de modo a reduzir erros na anotação, o que poderá também ter eliminado transcritos bem anotados, mas com baixa cobertura. No entanto o protocolo definido nesta tese poderá ser futuramente usado para uma análise mais detalhada sobre o *splicing* alternativo no sobreiro, usando a nova versão do genoma, que estará disponível em breve.

## Acknowledgements

Aos meus orientadores, Cátia Pesquita e Marcos Ramos um muito obrigado por terem aceite guiar-me nesta aventura que foi o ano que passou. Obrigado ao Marcos, por me ter recebido muito bem em Beja e me ter incluído na sua equipa.

Ao Pedro e Anabel, obrigado por me guiarem na fase inicial do trabalho, respondendo a todas as minhas dúvidas e ajudando a resolver problemas como a máxima entrega e paciência.

À Brígida, ao Daniel, à Ana Ferro, à Lia e a toda a equipa do CEBAL, obrigado por me receberem tão bem e terem facilitado a minha integração nos meses que passei em Beja. Foi um período intenso mas bastante marcante pela positividade, e sem vocês não seria a mesma coisa.

À Prof. Margarida Oliveira, um obrigado por me permitir entrar nesta aventura do mestrado, mesmo que isso tenha afetado o meu desempenho no laboratório durante um ano. Tenciono compensar o tempo fora do laboratório nos próximos tempos.

Aos meus amigos e colegas de trabalho do GPlantS, obrigado pelo apoio e motivação que me deram durante todo este período.

À Isabela e à Joana por serem as colegas de turma mais espetaculares, embora tenham nascido depois de 1990, o que me fez sentir muito velho nos primeiros dias mas depois passou-me. Não poderia pedir melhor companhia no meu regresso às aulas.

Obrigado ao Miguel por me aturar todos os dias e por me apoiar nas minhas decisões.

Obrigado Mãe e Pai por tudo.



*À minha avó Maria,  
que me aconselhou a estudar para um dia ser alguém...*



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Alternative splicing prediction based on RNA-seq . . . . .	2
1.2.1	Read mapping to a reference genome . . . . .	4
1.2.2	Transcript assembly and expression estimation . . . . .	7
1.3	Cork oak genomics . . . . .	9
1.4	Objectives . . . . .	11
<b>2</b>	<b>Methods</b>	<b>13</b>
2.1	Library construction and transcriptome sequencing . . . . .	13
2.2	Read processing and mapping on the reference genome . . . . .	13
2.3	Transcript assembly . . . . .	16
2.4	Identification of splicing events . . . . .	17
2.5	Differential expression analysis . . . . .	17
2.6	Functional annotation . . . . .	18
<b>3</b>	<b>Results and Discussion</b>	<b>19</b>
3.1	Study design . . . . .	19
3.2	Mapping analysis: STAR vs HISAT2 . . . . .	19
3.3	Transcriptome reconstruction: StringTie vs Cufflinks . . . . .	22
3.4	Tuning transcript assembly with StringTie . . . . .	26
3.5	Landscape of AS in cork oak . . . . .	28
3.6	Differential expression analysis across the four tissues . . . . .	31
<b>4</b>	<b>Conclusions</b>	<b>39</b>

## CONTENTS

---

<b>A Supplementary materials</b>	<b>41</b>
A.1 Transcriptome reconstruction: StringTie vs Cufflinks . . . . .	41
A.2 Tuning transcript assembly with StringTie . . . . .	43
A.3 Differential expression analysis across the four tissues . . . . .	44
<b>References</b>	<b>49</b>



# List of Figures

1.1	Schematic representation of the most common AS events found in eukaryotic species . . . . .	3
1.2	Overview of the main stages of RNA-seq analysis . . . . .	4
2.1	Cork oak tissues sampled for RNA extraction . . . . .	14
2.2	Bioinformatics pipeline used for prediction of alternative splicing events . . . . .	15
3.1	Percentage of reads mapped in proper pair on the cork oak genome in unique and multiple positions, using HISAT2 and STAR. . . . .	21
3.2	Categorization of Cufflinks and StringTie transcripts by estimated depth of read coverage. . . . .	23
3.3	Isoforms per gene frequency determined for reference annotation and further annotations obtained with StringTie assemblies . . . . .	28
3.4	Categorization of strt.cafj10-T10 annotated transcripts by estimated depth of read coverage. . . . .	29
3.5	Expression profile and enriched GO terms on xylem and leaf-specific transcript clusters . . . . .	33
3.6	Expression profile and enriched GO terms on peridermis-specific transcript clusters . . . . .	34
A.1	Categorization of Cufflinks and StringTie transcripts by estimated depth of read coverage (complementary plots) . . . . .	42
A.2	FPKM distribution across samples used for transcript expression analysis . . . . .	44

## LIST OF FIGURES

---

A.3 Hierarchical clustering and heatmap based on sample-to-sample euclidean distances . . . . .	45
A.4 Principal component analysis of the 16 samples used for transcript expression analysis . . . . .	45
A.5 $k$ -means clustering of differentially expressed transcripts . . . . .	46
A.6 Examples of differential splicing found between tissues for cork oak <i>Alpha-amylase-like</i> and <i>Topoisomerase II-like</i> genes . . . . .	47

# List of Tables

2.1	Combination of parameters used in StringTie runs to optimize transcript assembly . . . . .	17
3.1	Mean number of read pairs per tissue mapped on the cork oak genome in unique and multiple positions, using HISAT2 and STAR	20
3.2	Number of genes and transcripts predicted for cork oak using Cufflinks and StringTie . . . . .	24
3.3	Quantification of single- and multi-exon genes expressed in leaf, phellem, inner bark and xylem . . . . .	30
3.4	Alternative splicing events occurring in leaf, phellem, inner bark and xylem . . . . .	31
A.1	Characterization of genome annotation files, regarding number of exons and transcripts per genes, from reference and further annotations generated after the optimization rounds with StringTie. .	43



# Chapter 1

## Introduction

### 1.1 Motivation

Alternative splicing is a process used in gene expression to yield different transcript variants (or isoforms) derived from a single gene. It occurs during the maturation of the messenger RNA and it is based based on the differential use of splice sites, leading to different combinations of exons (coding regions). This process imposes an extra layer of complexity for transcript annotation of newly sequenced genomes. This is the case of cork oak (*Quercus suber*), a forest species with high economic and social significance, whose genome has been recently assembled (Ramos *et al.*, Submitted). Although a preliminary structural annotation based on gene models has already predicted gene boundaries and transcript variants, this needs to be validated on real datasets, using RNA-seq.

Transcript assembly using RNA-seq data is of high complexity since transcripts may be composed of many exons, and these may be shared between two or more isoforms. Many algorithms have been developed to predict transcript structures based on RNA-seq, dealing with read mapping to a reference genome, prediction of splicing sites and assembly of transcripts. The selection of the most suitable software for each analysis step mostly depends on the sequencing technology used and the availability of a fully annotated reference genome. Yet, other factors such as short error prone reads (present in RNA-seq datasets), alignment artifacts, low levels of gene expression or even lack of annotation detail for a

## 1. INTRODUCTION

---

given genome (as expected in the present cork oak assembly) may greatly impact the performance of these softwares. Therefore, we aimed to develop a workflow for transcriptome assembly and alternative splicing prediction based on RNA-seq in cork oak, by comparing the performance of two state-of-the-art softwares developed for read mapping and transcript assembly, respectively, and generate a more detailed genome annotation. After selection of the softwares with the best performance, we aimed to validate the new annotation and assess the extent of alternative splicing found in the datasets, also obtaining a set of high confidence transcripts (i.e. validated by RNA-seq).

### 1.2 Alternative splicing prediction based on RNA-seq

One of the key steps to improve the structural annotation of a genome is to look into transcribed sequences [messenger RNA (mRNA) or long non-coding RNA (lncRNA)], defining its intron and exon boundaries and uncover different variants that can occur due to alternative splicing. Precursor mRNAs commonly undergo a maturation step in which intronic regions are removed, based on nucleotide signatures defining its boundaries (splice sites) and adjacent exons are joined. Alternative splicing (AS) of precursor mRNA is a molecular phenomenon found in eukaryotic species that may generate different mature mRNAs (isoforms) based on the differential use of splice sites (Nilsen & Graveley, 2010). AS may lead to the translation of different proteins, regulating proteome diversity and/or protein function, or simply regulate transcript abundance by generating non-functional isoforms that are degraded through nonsense-mediated mRNA decay (Hug *et al.*, 2016). The most common AS events include (by order of mean frequency in plants): intron retention, when an intron is not spliced and integrates the mature mRNA; alternative acceptor (3') or donor site (5'), when an alternative splice site located within an exon is used, leading to the removal of the intron and part of the exon; exon skipping, when an exon is also spliced together with adjacent introns (Figure 1.1). Other types of splicing, which include mutually exclusive exons or introns, intraexonic deletions, or alternative acceptor and donor sites are also

## 1.2 Alternative splicing prediction based on RNA-seq

---

frequently reported, but usually occurring at a lower frequency [e.g. [Dubrovina et al. \(2013\)](#); [Huang et al. \(2015\)](#); [Marquez et al. \(2012\)](#); [Xie et al. \(2015\)](#); [Xu et al. \(2014\)](#); [Zhang et al. \(2017\)](#)].

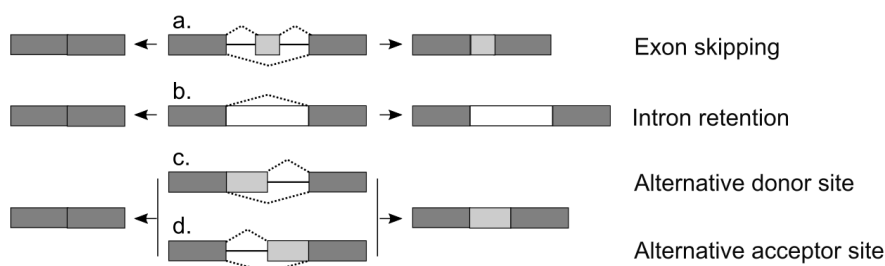


Figure 1.1: Schematic representation of the most common AS events found in eukaryotic species: (a) exon skipping; (b) intron retention; (c) alternative donor site; (d) alternative acceptor site. Gray rectangles represent exons; white rectangles and black lines represent introns; dotted lines represent splicing.

High throughput RNA-Sequencing (RNA-Seq) has become a standard technique to assess transcriptomic changes occurring in living organisms when comparing tissues or cell types, or in response to a given factor. Although third generation sequencing techniques, such as PacBio (Single Molecule Real-Time) or nanopore sequencing, generate long-reads and may sequence entire transcripts, the most widely used approach is still short-read sequencing, such as Illumina, due to its high accuracy and lower cost ([Heather & Chain, 2016](#)). A common workflow for RNA-seq studies start with read filtering based on quality, followed by read mapping against a reference genome (if available for the target organism) and assembly of the reads in order to build the transcripts (Figure 1.2). After this, estimation of gene/transcript abundance can be performed and differential expression can be assessed by comparing different datasets. Many algorithms have been designed to analyze RNA-seq data allowing transcriptome assembly with or without a reference annotation. Choosing the most suitable software for each analysis step mostly depends on the sequencing technology used and the availability of a fully annotated reference genome.

## 1. INTRODUCTION

---

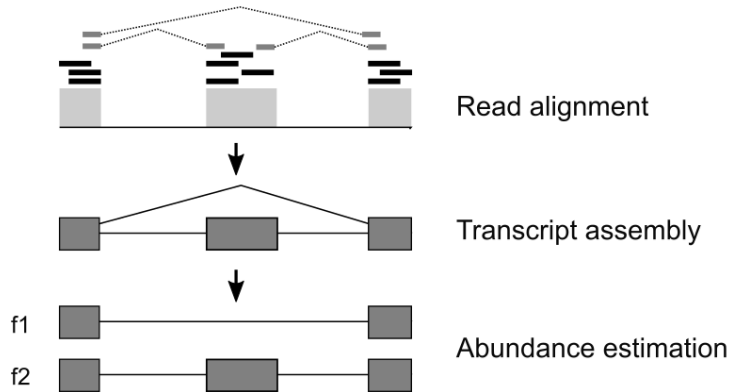


Figure 1.2: Overview of the main stages of RNA-seq analysis. Adapted from (Florea & Salzberg, 2013).

### 1.2.1 Read mapping to a reference genome

The mapping (or alignment) of large sets of reads to a reference genome is a central step in RNA-seq analysis. Programs designed to map short reads to a reference genome face the challenging task of aligning spliced reads and correctly determining exon-intron junctions, taking into account the variable size gaps generated by introns. Therefore, specific software need to be designed to properly handle intron-sized gaps, where common software (such as BWA and Bowtie) designed for DNA alignment tend to fail (Baruzzo *et al.*, 2016). In addition, a viable aligner also needs to handle paired-end, run fast/efficiently and have the ability to align reads across unannotated splice junctions (Baruzzo *et al.*, 2016; Engström *et al.*, 2013). Most splice-aware aligners are based on prediction of splice junctions using a reference genome and annotation, while only a few others are available for transcriptome assembly without a reference [e.g. TransABySS (Robertson *et al.*, 2010) and Trinity (Haas *et al.*, 2013)]. The aligners that require a reference genome can be divided according to two main strategies for read alignment, the exon-first approach and the seed-extend approach (Alamancos *et al.*, 2014; Reddy *et al.*, 2013). In the exon-first approach, unspliced reads are aligned first in order to find candidate exonic regions represented as read-clusters, and remaining reads are then mapped using specific algorithms to find connections between clusters. This strategy is followed, for example, by MapSplice (Wang



## 1.2 Alternative splicing prediction based on RNA-seq

---

*et al.*, 2010), SpliceMap (Au *et al.*, 2010), Tophat2 (Kim *et al.*, 2013). The seed-extend approach first uses part of the read (seed) to align to the reference and select candidate alignment sites; this seed is then extended using specific search algorithms and candidate splice-sites are located. This approach is the base of GSNAP (Wu & Nacu, 2010) and STAR (Dobin *et al.*, 2013), for example. Exon-first approach is generally faster, but strongly dependent on the coverage of unspliced read clusters to assign spliced reads, while seed-extend approach is less dependent on unspliced reads but may require more computer memory (Alamancos *et al.*, 2014). Regardless of the strategy used, these aligners also require a genome indexing stage for the use of improved searching algorithms based on hash tables or suffix arrays.

The performance of most of these aligners are frequently compared, by developers themselves upon the release of a new software or update, or through benchmarking analyses by independent authors (e.g. Baruzzo *et al.* (2016); Dobin *et al.* (2013); Engström *et al.* (2013); Kim *et al.* (2013, 2015). TopHat became a widely used tool (Baruzzo *et al.*, 2016) due to its improved performance at the time and inclusion in the well recognized Tuxedo pipeline for RNA-seq analyses (Trapnell *et al.*, 2012). However, one major disadvantage of this software was the increased runtime (Baruzzo *et al.*, 2016; Dobin *et al.*, 2013; Kim *et al.*, 2015). More recently, the same authors released HISAT, a faster aligner with reduced memory consumption, due to an hierarchical indexing algorithm that is based on the Burrows-Wheeler transform and the FM index (Kim *et al.*, 2015). HISAT algorithm employs a global FM index to represent the entire genome and numerous local and overlapping FM indexes that represent small parts of the genome. In addition, HISAT uses a mixed exon-first strategy, using Bowtie to handle low-level operations, with seed-extend approach to find genomic locations. Candidate locations for part of each read are first localized using the global index and the best matches are selected based on internal parameters. Reads are then extended until a mismatch is found, and at this stage the local FM index is retrieved and a local search is performed (Kim *et al.*, 2015). Besides the fast performance, HISAT accuracy is comparable to that of TopHat (Baruzzo *et al.*, 2016; Kim *et al.*, 2015).

Before the development of HISAT, STAR stood out among the available algorithms for its high mapping speed, accuracy and the first to efficiently deal with

## 1. INTRODUCTION

---

longer read lengths (from third generation sequencing) (Dobin *et al.*, 2013). After its release, STAR rapidly became the second most used open source software for RNA-seq analysis based on genome alignment, following TopHat (Baruzzo *et al.*, 2016). In a first stage, STAR applies a sequential search for a maximum mappable length of a seed, implemented as a speed-efficient suffix array search (MMP, maximum mappable prefix), starting from the first base. The MMP refers to the longest substring that matches exactly one or more substrings in the genome. The algorithm finds the MMP for the seed and if a splice junction is present, the first seed will be mapped to the donor splice site and a new search is repeated for the unmapped portion, until an acceptor splice site is found. This sequential application of MMP search is one of the key elements that makes the STAR extremely fast. In the second stage, STAR starts building the alignments for the entire read sequence by stitching together all the aligned seeds, first by clustering a set of proximal anchor seeds in limited genomic window, and then by stitching pairs of seeds, taking into account gap and mismatch penalties (Dobin *et al.*, 2013).

Both, HISAT and STAR recommend the use of gene annotations to identify and correctly map spliced alignments across known splice junctions. New candidate splice junctions are further identified based on mapping evidence and can be used in a two-pass mode to allow the detection of more spliced reads mapped to novel junctions. The two-pass mode is available for STAR, requiring a second run, while HISAT makes use of splice sites found during previous alignments when aligning further reads, using the same run (Dobin *et al.*, 2013; Kim *et al.*, 2015). The performance of both aligners is comparable regarding mapping yields, accuracy and splice junction call (Baruzzo *et al.*, 2016; Kim *et al.*, 2015). However, STAR showed to perform better than HISAT with complex datasets (simulated) containing increased rates of polymorphisms (Baruzzo *et al.*, 2016). One main disadvantage of STAR is the large memory requirements due to the use of the suffix-array method. HISAT uses the Burrows-Wheeler transform, improving its processing speed and requiring a lower amount of random access memory (Kim *et al.*, 2015).

### 1.2.2 Transcript assembly and expression estimation

One of the main goals of RNA-Seq is to accurately identify the full-length structure of the transcripts that are expressed in a given dataset, and estimate their relative abundances. This can be performed by using as input the data generated from read mapping to a reference, and assigning clusters of reads to transcriptionally active regions. However, this task is of high complexity even when a reference genome is available, since transcripts may be composed of many introns and occur in different splice forms. In addition, transcript assembly can be affected by precursor factors, such as short error prone reads, alignment artifacts or bias introduced by the library construction process, which may introduce more "noise".

Methods for estimating transcript expression and AS prediction based on a reference genome can be divided in two types: event-based models and isoform resolution models (Alamancos *et al.*, 2014; Liu *et al.*, 2014). The event-based approach is not an assembly approach per se, but mostly an expression estimation technique, based on counting the number of reads falling on a given locus (normalized for transcript length and the total number of mapped reads). Event-based models estimate differential splicing by counting reads located at exon level and test two possible splicing outcomes, inclusion and/or exclusion of an exon. However, these models are highly based on genome annotation and are not suitable for predicting novel splice forms, particularly when the reference annotation provided is incomplete (Liu *et al.*, 2014). Isoform resolution models mostly deal with estimating transcript expression based on isoform reconstruction, also predicting new isoforms. Most algorithms for transcript assembly based on a genome perform a clustering of overlapping reads for each locus and then build graphs that represent all isoform possibilities. This process is highly affected by sequence coverage due to the difficulty of unambiguously assembling multiple isoforms, particularly when considering isoforms with highly variable coverage. Since many theoretical splice variants can arise during graph construction, an analytical step is then applied to select the subset of transcripts most likely to be represented in the RNA-seq library (Liu *et al.*, 2014; Pertea *et al.*, 2015). Two representative programs used for transcript assembly and selection are Cufflinks, a widely used

## 1. INTRODUCTION

---

tool that is part of the Tuxedo package (Trapnell *et al.*, 2012), and StringTie, a recently developed software (Pertea *et al.*, 2015) that showed improved performance and high accuracy in recent benchmark analysis (Hayer *et al.*, 2015; Williams *et al.*, 2017).

Cufflinks predicts transcript structures using an overlap graph, with the nodes being the sequenced reads and edges representing the overlap between two reads that have also compatible alignments (similar splice patterns). This overlap graph is structured as a directed acyclic graph and each path represents a putative transcript. These graphs are then parsed using a parsimony-based algorithm, selecting the minimum number of transcripts that will explain all reads mapped the graph (Florea & Salzberg, 2013; Trapnell *et al.*, 2010), without taking into account transcript abundance. Estimation of transcript abundance can be further performed by assigning to each read a probability of belonging to any of the isoforms (a value that depends on the current abundance estimation of isoforms) and iteratively assigning reads to isoforms according to this probability to determine the maximum-likelihood expression levels for all isoforms (Florea & Salzberg, 2013; Trapnell *et al.*, 2010).

StringTie uses splice graphs to represent all possible isoforms for a locus under different paths, in which nodes represent exons or exons portions and edges represent introns connecting two exons. Exons and intron structure in the graphs are based on gene clusters grouped from read mappings. The selection step greatly differs from the approach followed by Cufflinks, since StringTie dynamically creates a separate flow network for each splice graph to estimate isoform abundance using a maximum flow algorithm (Pertea *et al.*, 2015). After building a splice graph, the algorithm iteratively searches for the heaviest path (an path-compatible isoform structure with the highest per-base read coverage throughout all nodes) and creates a flow network to estimate abundance. This is performed using a maximum flow algorithm that determines the maximum number of reads that can be associated with the selected isoforms. Afterwards, StringTie removes the reads that contributed to this estimation and updates the per-base coverage of the splice graph. Thus, StringTie uses coverage to constrain the algorithm, working as an optimization technique, which may improve the accuracy of the assembly (Pertea *et al.*, 2015, 2016).

Pertea *et al.* (2015) observed that both Cufflinks and StringTie could lead to excessive gene predictions when analyzing real and fully annotated human datasets, if new predictions were considered false positives. In that study, StringTie showed the best sensitivity in transcript assembly and abundance estimation, while Cufflinks was the next best assembler for real datasets, among the four that were tested (Pertea *et al.*, 2015). Another benchmarked analysis of seven genome-guided assembler algorithms, highlighted the strength of both Cufflinks and StringTie in isoform prediction, inclusively when genome annotation was not provided (Hayer *et al.*, 2015). In the presence of a more detailed annotation, Cufflinks predictions tended to have better precision, whereas StringTie showed better recall. Regarding isoform abundance estimation, Cufflinks showed higher correlation to true transcript FPKM values, although StringTie predicted a higher number of truly expressed genes (number of isoforms with true expression predicted by the algorithm as being expressed). The good performance showed by both algorithms makes them top candidates for genome-guided transcript assembly, particularly when none or incomplete annotations are available as it is the case of undergoing genome sequencing projects.

### 1.3 Cork oak genomics

Cork oak (*Quercus suber* L.) forests are unique and emblematic resources, with high economical, ecological and social significance. In the Iberian Peninsula, most cork oak woodlands ("montados") are savannah-type complex ecosystems maintained by human management (Bugalho *et al.*, 2011). Cork oak occupies 23% of the total Portuguese forest area and Portugal is the world leader in cork production (49,6%) and exportation (more than 60% of the world exported cork volume, APCOR 2013).

Cork is a senescent tissue, with unique physical properties and wide range of commercial applications, that protects the tree from adverse environmental conditions. Cork (or phellem) development starts with the formation of phellogen, a specialized lateral cambium that produces a phellem layer to the outside of the trunk and phelloderm to the inside. Phellem differentiation will involve cell expansion, extensive deposition of suberin and waxes in the cell walls and an

## 1. INTRODUCTION

---

irreversible program of senescence ending in cell death (Graça & Pereira, 2004). Contrastingly, phelloderm is composed of living and non-suberized cells that will accumulate below the phellogen, close to mature phloem cells used as storage tissue. Hereafter, the layer corresponding to phelloderm and mature phloem will be referred to as inner bark, while the layer composed of phellem and phelloderm will be referred to as peridermis (Pereira, 2007).

The few studies that have been made to understand cork development in *Q. suber* mostly focused the comparison between xylem and phellem tissues (Ricardo *et al.*, 2011; Soler *et al.*, 2007). These studies have found different genes or proteins over-represented in cork cells, with predicted involvement in suberin biosynthesis, response to stress, as well as meristem identity (Ricardo *et al.*, 2011; Soler *et al.*, 2007). More recently, Rains *et al.* (2017) compared the transcriptomes of outer bark and inner bark from poplar (*Populus tremula* x *P. alba* hybrid) and identified similar regulators and effector genes involved in suberin biosynthesis as previously identified in cork Soler *et al.* (2007), suggesting a conservation of this pathway in woody species. However, one exclusive feature found in cork oak is the ability to develop new cork layers after each harvest. This occurs by the regeneration of a new phellogen within inner bark, highlighting the importance of this layer in cork production (Pereira, 2007). Still, the metabolic pathways found in cork oak's inner bark remain unknown.

Given the unique ability to regenerate cork layers after harvest, cork production is exploited as a sustainable system through many production cycles (Oliveira & Costa, 2012). Cork producers/industry are currently focusing on improving agronomical practices to intensify the traditional savannah-type cork producing ecosystem and reduce at least the time for first cork harvest. To efficiently support decisions on intensifying strategies that could promote the competitiveness of cork industry, or even develop breeding programs, it is crucial to uncover new data on fundamental questions about cork oak biology.

Recently, a window of opportunity to develop fundamental and applied knowledge in this species was opened along with the Cork Oak Genome Sequencing initiative (GENOSUBER) (Ramos *et al.*, Submitted). A draft genome assembly with an estimated genome size of 953.3 Mb was already obtained using a combination of paired-end and mate-pair libraries sequenced using the Illumina

technology. The draft genome is organized in 23,347 scaffolds, with the majority of the assembly being represented in a considerable smaller number of larger scaffolds (longer than 10,000 bp). Structural annotation of the genome predicted 79,752 genes with complete open reading frames, and 83,814 transcripts, already indicating alternative splicing events in some genes (Ramos *et al.*, Submitted).

## 1.4 Objectives

The work described in the present thesis aimed to characterize the extent of alternative splicing events in cork oak and improve genome structural annotation using a RNA-seq dataset generated for four different cork oak tissues: leaves, phellem, inner bark and xylem. The workflow designed for this purpose included the comparison of two different algorithms for read mapping (STAR and HISAT2) and transcriptome assembly (Cufflinks and StringTie).

This work also aimed to shed a light into the metabolic pathways found in phellem and inner bark, through the identification of genes expressed predominantly in these tissues.





# Chapter 2

## Methods

### 2.1 Library construction and transcriptome sequencing

Developing phellem, inner bark and xylem were collected from adult branches (Figure 2.1) from the same cork oak genotype used for genome sequencing (HL8). To increase transcriptome diversity, fully expanded leaves from the same branch were also collected. RNA was extracted from each tissue and cDNA libraries were prepared and sequenced using paired-end protocol and 100 bp read length on Illumina HiSeq-4000 platform (performed at the Beijing Genomics Institute, China). Four technical replicates were sequenced for each tissue.

### 2.2 Read processing and mapping on the reference genome

Raw reads were filtered to remove adaptor sequences and reads containing undetermined nucleotides (N's) and further processed to trim/remove low quality reads (minimum quality  $\geq 20$ , minimum length  $> 80\%$  of total read length) using Sickle (Joshi, NA and Fass, JN, 2011).

Read mapping on the reference genome was performed using two spliced alignment software, HISAT2 v2.0.5 (Kim *et al.*, 2015) and STAR v2.5.0 (Dobin *et al.*,

## 2. METHODS

---



Figure 2.1: Cork oak tissues sampled for RNA extraction. Samples from developing phellem (a), inner bark (b) and xylem (c) were collected from adult branches (left panel). Leaves (d) from the same branches were also sampled (right panel).

2013), in order to compare their performance (Figure 2.2). The reference genome annotation was used in both approaches to guide the alignment and provide coordinates of splice junctions from annotated transcripts. A genome index for HISAT2 was created using `hisat2-build` command and options `"-ss"` and `"-e"`, which take as input two files containing splice junction and exon coordinates, respectively. These files were previously obtained using `hisat2_extract_splice_sites.py` and `hisat2_extract_exons.py` scripts (available in the HISAT2 package) and the reference genome annotation file. HISAT2 was then run for all processed paired libraries using default parameters (with `-asOUT` option to generate a list of novel splice junctions) and final mapping statistics were recorded from standard output. The genome index for STAR was built using `"-runMode genomeGenerate"` and `"-sjdbGTFfile"` options from STAR command, the latter taking as input the reference annotation file. Mapping was further performed for each read paired library independently using the option `"-runMode alignReads"`, in two rounds (multi-sample two-pass mapping). The first round of mappings was used to obtain a standard output file for each paired library, containing a set of validated splice junctions (`SJ.out.tab`). The second round of mappings was further performed using the additional option `"-sjdbFileChrStartEnd /path/to/sj1.tab /path/to/sj2.tab ..."` which takes a list of `SJ.out.tab` files for a given set of samples. This two-pass mapping method was only made for STAR alignments since a similar option in HISAT2 is already provided as default.

## 2.2 Read processing and mapping on the reference genome

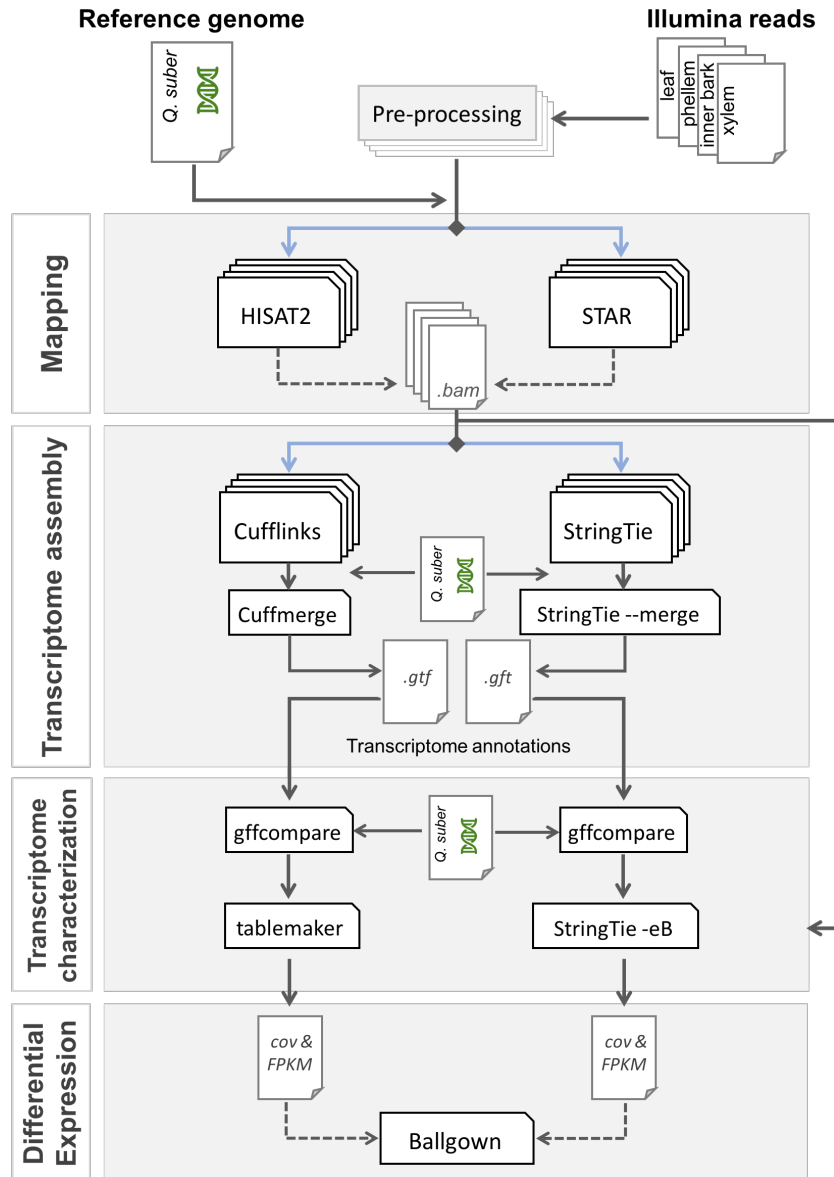


Figure 2.2: Bioinformatics pipeline used for prediction of alternative splicing events in cork oak. After a pre-processing stage where high quality reads were filtered, mapping was performed using HISAT2 and STAR. The BAM output files from one of the aligners was selected to transcript re-construction. This step was performed with Cufflinks and StringTie and generated two GTF genome annotation files that were compared to reference annotation using gffcompare. StringTie -eB and Tablemaker were used to generate transcript coverage estimations based on the BAM files and using the new annotation files as reference. The annotation generating the best estimation of transcript diversity was used for gene expression analysis using Ballgown.

## 2. METHODS

---

Final mapping statistics from both mapping tools were collected from standard output. Final alignment files (bam format) were filtered for high quality mappings in proper pair using samtools (Li *et al.*, 2009).

### 2.3 Transcript assembly

Preliminary transcript assembly, based on read mapping on cork oak genome, was performed using Cufflinks v2.2.1 (Trapnell *et al.*, 2010) and StringTie v1.3.3 (Pertea *et al.*, 2015), in order to assess their performance using default parameters and the option to include the reference genome annotation (Figure 2.2). Transcripts were assembled for each sample individually, using the corresponding mapping file as input, and then all assemblies were merged along with the reference annotation using cuffmerge or stringtie `-merge` modes. Each final GTF (General Transfer Format) file obtained with Cufflinks and StringTie was compared to original genome annotation using gffcompare utility (<https://github.com/gpertea/gffcompare>) to classify transcripts as they relate to reference transcripts and collapse contained transfrags (intron-redundant) using `-C` option. To further evaluate the coverage of the assembled transcripts, each new annotation file was used as reference to analyze read alignment files and generate coverage data, using Tablemaker (for Cufflinks annotation, <https://github.com/leekgroup/tablemaker>) and StringTie with `"-e"` and `"-B"` options (for StringTie annotation). Afterwards, histograms for transcript coverage were obtained using a custom built R script (<https://github.com/pedro-mb/RNA-seq-scRipts>).

Based on previous analysis, StringTie was selected for transcript assembly, and further assembly options were tested in order to improve transcript annotations. Thus, StringTie was repeated adjusting minimum read coverage for predicted transcripts [`-c` 2.5 (default) and 5], minimum anchor length on each side of a junction [`-a` 10 (default) and 15] and junction coverage [`-j` 1 (default), 2 and 10]. The `-merge` step was further performed for each independent StringTie optimization round and in this step the parameter setting a minimum input transcript per million (TPM) to include in the merge [`-T` 1 (default), 4 and 10] was also tested. All the tested combinations are indicated in Table 2.1. Final GTF files were compared to original genome annotation using gffcompare to collapse

## 2.4 Identification of splicing events

contained transfrags (intron-redundant) using -C option. Several parameters were used to evaluate all optimization runs, including total number of new isoforms and exons, new predicted genes and number isoforms per gene. The new genome annotation file providing the best prediction of the transcriptome was selected and used as reference to generate transcript coverage data among libraries, using StringTie with "-e" and "-B" options.

Table 2.1: Combination of parameters used in StringTie runs to optimize the transcript assembly during individual library assemblies and merge steps. The tested parameters (Param.) included: minimum read coverage for predicted transcripts (-c), minimum anchor length (-a), junction coverage (-j) and minimum input transcript expression in TPM to include in merge (-T). The default conditions are indicated as "str.t.def".

Step	Param.	str.t.cafj10				
		str.t.def	str.t.cafj	str.t.cafj10	-T4	-T10
Assembly	-c	2.5	5	5	5	5
	-a	10	15	15	15	15
	-j	1	2	10	10	10
Merge	-T	1	1	1	4	10

## 2.4 Identification of splicing events

To assess the alternative splicing transcriptional landscape among tissues, transcripts with mean FPKM values in the four replicates above 1 were selected in order to generate four transcriptome sub-sets, representing the transcript universe expressed in each tissue. The structural annotation for each sub-set was obtained by parsing the new genome annotation file and used as input in ASTALAVISTA v4.0 (Foissac & Sammeth, 2007).

## 2.5 Differential expression analysis

Expression analysis was performed based on the transcript coverage obtained in Section 2.3 after StringTie optimization. Differential expression was performed

## 2. METHODS

---

using Ballgown (Frazee *et al.*, 2015; Fu *et al.*, 2017) package for R environment and based on FPKM abundance estimates. Differentially expressed transcripts were identified using multigroup comparisons ( $q$ -value  $< 0.01$ ). To identify transcript clusters with similar expression profiles within the 4 tissues, a  $k$ -means clustering analysis was performed using Jensen-Shannon distance calculated based on the mean FPKM values determined for each tissue [number of clusters ( $k$ ) = 16]. This analysis was performed using functions from CummeRbund package for R (Goff *et al.*, 2013).

### 2.6 Functional annotation

Nucleotide sequences for each transcript were retrieved from cork oak draft genome using gffread (<https://github.com/gpertea/gffread>) based on the selected new genome annotation. Open reading frames (ORFs) were predicted using TransDecoder v3.0.1 (<https://github.com/TransDecoder/TransDecoder/>) using the default criteria for ORF retention, enriched with homology searches on Swiss-Prot database (April 2017) using BlastP [BLAST+ v2.6.0, Camacho *et al.* (2009)] and on Pfam database (March 2017) for protein domains using hmm-scan (HMMER 3.1, <http://hmmer.org/>). For functional annotation of cork oak proteome, the corresponding peptide sequences were used as query for BlastP homology search against the *Arabidopsis thaliana* protein database (TAIR10), with a cutoff e-value of  $1 \times 10^{-3}$  (Camacho *et al.*, 2009). Gene ontology (GO) enrichment analysis was performed using the *Arabidopsis* homologs of the cork oak differentially expressed transcripts identified in Section 2.5, using BiNGO plug-in (Maere *et al.*, 2005) for Cytoscape v3.2.1 (Cline *et al.*, 2007).

# Chapter 3

## Results and Discussion

### 3.1 Study design

The reference draft genome annotation obtained for cork oak predicted the occurrence of alternative splicing (AS) in multiple genes, although in a limited extent (Ramos *et al.*, Submitted). In the present study, we analyzed sixteen RNA-seq libraries prepared from four cork oak tissues, in order to predict new AS forms for the already predicted genes and improve genome structural annotation.

In the workflow designed for this study, we compared the performance of different software in two of the most critical stages in transcript assembly: STAR and HISAT2 for read mapping against the reference genome; StringTie and Cufflinks for transcriptome reconstruction (Figure 2.2). After that, an optimization step was performed using the selected transcriptome assembler and a new genome structure annotation was obtained. This new annotation was used as reference to estimate transcript abundance in each tissue and differential expression analysis was further performed using Ballgown comparing the datasets obtained for each tissue.

### 3.2 Mapping analysis: STAR vs HISAT2

An average of 622.86 million high quality reads per tissue (78.87 million reads per library) were obtained after pre-processing, which accounted for approximately

### 3. RESULTS AND DISCUSSION

---

93.5% of total raw reads. Globally, mapping analysis using STAR yielded a higher percentage of uniquely mapped reads in proper pair, ranging from 60.49 to 73.45 million mapped read pairs for leaf and xylem, respectively (Table 3.1). Read alignment using HISAT2 generated 53.11 to 64.77 million uniquely mapped reads in proper pair, on average for phellem and xylem libraries, respectively. Consequently, mapping rates were higher using STAR, ranging from 84.22% to 86.86% of total high quality reads (Figure 3.1), compared to HISAT2 (73.88% to 76.55%). The results obtained with STAR are relative to the two-pass mode, which is default in HISAT2. This two-pass mode is performed with a first run to report a list of splice junctions validated by reads with long anchors, followed by the second run, which takes this information to align reads with short anchors. In the present study, the percentage of uniquely mapped reads in proper pair was relatively higher in the first STAR run (85.23% to 87.90%, data not shown). The decrease observed after two-pass was caused by a slight increase in multimapped reads, which can be explained by the increase in reference annotation detail with the novel splice junctions, and deduced increase in mapping accuracy.

Table 3.1: Mean number of read pairs per tissue (n=4) mapped on the cork oak genome in unique and multiple positions, using HISAT2 and STAR. Total number of unmapped pairs is also represented.

Aligner	Tissue	Uniquely mapped	Multimapped	Unmapped
STAR	Leaf	60,487,184.00	7,441,070.50	2,231,786.00
	Phellem	70,677,430.75	8,942,502.00	4,304,146.75
	Inner bark	62,457,384.75	7,904,287.00	2,428,973.50
	Xylem	73,451,037.25	8,665,381.25	2,438,255.00
HISAT2	Leaf	53,106,302.25	11,711,617.50	5,342,120.75
	Phellem	62,006,482.50	14,030,328.50	7,887,268.50
	Inner bark	54,657,527.00	12,418,154.75	5,714,963.50
	Xylem	64,771,539.75	13,849,563.75	5,933,570.00

STAR and HISAT2 are two of the most recent splice aligners developed to align reads to a reference genome, sharing the advantage of being faster than other aligners available (Baruzzo *et al.*, 2016; Kim *et al.*, 2015). Kim *et al.* (2015), the developers of HISAT2, demonstrated that STAR showed higher memory usage



### 3.2 Mapping analysis: STAR vs HISAT2

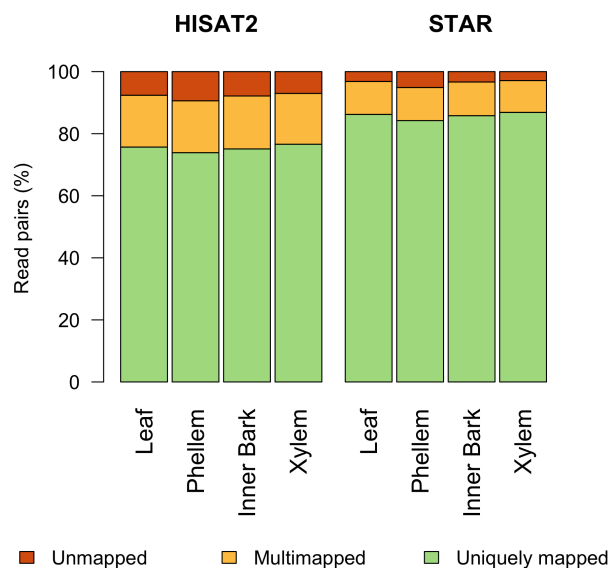


Figure 3.1: Percentage of reads mapped in proper pair on the cork oak genome in unique (green) and multiple (light orange) positions, using HISAT2 and STAR. Percentage of unmapped pairs is also represented (dark orange). Each column represents the sum of read pairs obtained for the technical replicates in each tissue.

(28 Gb) and run-time (50.6 min) compared to HISAT2 (4.7 Gb and 26.7 min, respectively), when analysing the same data set. The alignment accuracy of both aligners for 100-bp simulated reads were highly similar, particularly when comparing the two-pass modes (96.7% and 97.6% correctly and uniquely mapped reads with STAR and HISAT2, respectively). However, recent benchmarking analysis of multiple splice-aware aligners on simulated datasets highlighted STAR, but not HISAT2, for its accuracy in detecting splice events, being amongst the alignment software with best performance using default parameters (Baruzzo *et al.*, 2016). All the programs analyzed showed significantly low accuracy for detecting non-canonical junctions as compared to canonical junctions although HISAT2, STAR and two others perform best in this cases. Moreover, HISAT2 showed increased accuracy for aligning spliced reads with shortest anchors and without annotation (Baruzzo *et al.*, 2016). In addition, a pipeline including STAR as read aligner showed the best balance of precision and recall in analyzing real RNA-seq datasets derived human clinical samples (Williams *et al.*, 2017). In the

### 3. RESULTS AND DISCUSSION

---

present study, there was no previous knowledge on the true alignment location of reads, and alignment precision could not be evaluated. Still based previous benchmarking analyses and on the observed mapping efficiencies, we decided to continue the work with the alignments obtained with STAR since these can provide a broader view in transcriptome diversity.

### 3.3 Transcriptome reconstruction: StringTie vs Cufflinks

In order to predict new AS forms in cork oak, transcriptome assembly was performed first to test Cufflinks and StringTie. Both programs were run using default parameters and each run generated one GTF file per RNA-seq library, containing the corresponding transcript annotation predictions. These annotation files were then merged with the reference annotation, using the correspondent merge options, to generate a unified set of non-random transcripts (or isoforms) found across all samples. The performance of StringTie and Cufflinks was evaluated by assessing the total amount of new isoforms and new genes predicted and the corresponding coverage.

Globally, Cufflinks predicted a total of 120,988 novel transcripts, which corresponds to approximately 64.00% of total transcripts annotated on the GTF file (189 285). StringTie was more conservative in predicting new transcripts since from the final annotated transcripts (121,088) only 35.22% (42,652) were new isoforms or new candidate loci (Table 3.2). Consequently, the number of predicted new genes was also higher for Cufflinks assembly than for StringTie. The major contribution for new transcript predictions is from novel isoforms for genes already predicted in the reference annotation. These account for 77% (93,247) and 70% (30,116) of novel transcripts predicted by Cufflinks and StringTie, respectively. The remaining transcript classes follow the global trend observed, being more abundant in Cufflinks assembly, except for transcripts overlapping reference introns and other classes, not discriminated on the table. This group includes single exon transfrags overlapping a reference exon and at least 10 bp of a reference intron and transfrags predicted within 2K bases of a reference transcript, with

### 3.3 Transcriptome reconstruction: StringTie vs Cufflinks

no splice junction evidence between them. The higher number of cases found for StringTie may reflect a higher sensitivity to detect these cases, however, they only account for close to 3% of total novel transcripts.

The new transcript annotations were further evaluated by generating coverage data based on read mappings on reference genome. The distribution of coverage was assessed for the most abundant transcript classes in both annotations: transcript classes with a complete match with reference transcripts (confirmed by Cufflinks and StringTie), novel isoforms from reference genes, and unknown transcripts (intergenic transcripts, new candidate genes) (Figure 3.2 and Figure A.1, Appendix A). The high number of transcripts predicted in both annotations resulted in more than 50% of transcripts with coverage lower than 500 reads per base pair. This was more evident for Cufflinks annotation, where this percentage percentage rised to 75% (Figure 3.2). StringTie annotation allowed better support by read coverage, which is particularly evident for novel isoforms (Figure 3.2 and Figure A.1, Appendix A).

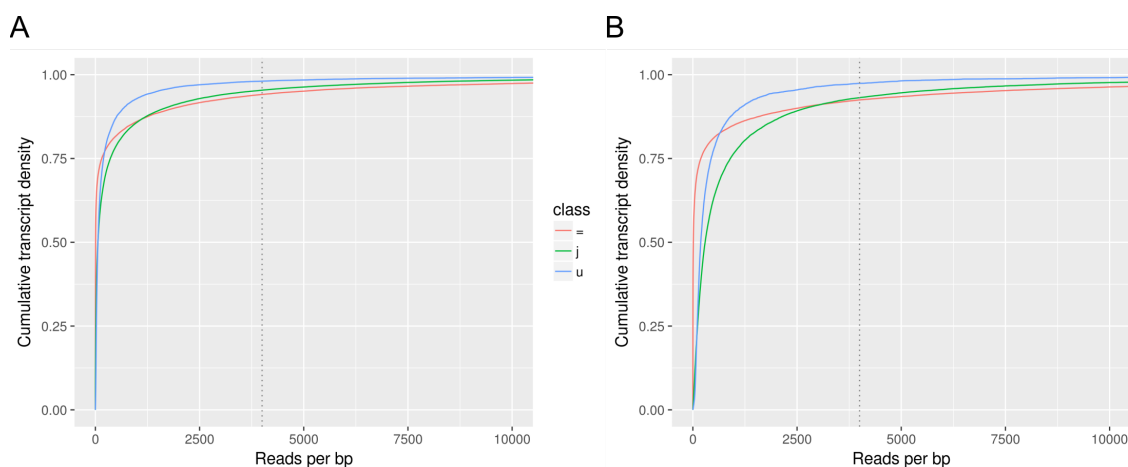


Figure 3.2: Categorization of Cufflinks and StringTie predicted transcripts by estimated depth of read coverage (reads per base pair). Transcripts were grouped according to gffcompare classification as: transcripts with complete match of intron-exon coordinates with reference transcripts (=, red), novel isoforms (j, green) and unknown intergenic transcripts (u, blue). Vertical grey dashed lines represent a threshold of 4,000 reads/bp used to create Figure A.1.

Table 3.2: Total number of genes and transcripts predicted in the original reference annotation (reference) and in the new annotations built using Cufflinks (cuff.def) and StringTie (strt.def) with default parameters. The results obtained for further optimizations with StringTie are also indicated (strt.cafj, strt.cafj10, strt.cafj10-T4 and strt.cafj10-T10, see Table 2.1 for specific details). Novel transcripts are new non-redundant mRNA sequences not predicted in the reference annotation.

	reference	cuff.def	strt.def	strt.cafj	strt.cafj10	strt.cafj10 -T4	strt.cafj10 -T10
Genes (total)	79,752	88,380	83,584	84,703	86,163	81,885	80,416
Transcripts (total)	83,813	189,285	121,088	117,547	112,217	96,601	89,801
Novel transcripts	-	120,988	42,652	38,501	32,631	15,707	7,958
<i>Novel isoforms</i>	-	93,247	30,116	26,100	20,522	10,417	5,453
<i>Unkown</i>	-	17,919	7,321	7,510	8,020	3,315	1,451
<i>Single exonic overlap</i>	-	5,845	2,990	2,194	1,195	747	441
<i>Opposite exonic overlap</i>	-	3,574	992	1,182	1,023	448	218
<i>Intronic overlap</i>	-	27	215	270	313	96	42
<i>Other</i>	-	376	1,018	1,245	1,558	684	353

Novel transcripts are classified by gffcompare, according to their location based on reference annotation, and summarized in the table as: novel isoforms (at least one splice junction is shared with a reference transcript), unknown (intergenic transcript), single exonic overlap (transfrag overlaps with one exon from a reference transcript), opposite exonic overlap (exonic overlap with reference on the opposite strand), intronic overlap (transfrag overlapping entirely within a reference intronic region). Other, less abundant classes are also indicated.

### 3.3 Transcriptome reconstruction: StringTie vs Cufflinks

---

The differences observed between Cufflinks and StringTie may arise from the different algorithms used to assemble transcripts. As mentioned in Section 1.2.2, Cufflinks creates overlap graphs, by connecting reads that overlap at a given genomic region and later applies a parsimony-based algorithm to generate the minimal number of transfrags necessary to explain all reads in the graph (Trapnell *et al.*, 2010). StringTie creates splice graphs for clusters of reads grouped at a given region to predict transfrags, and further creates a separate flow network for each transfrag to estimate its expression level applying a maximum flow algorithm (Pertea *et al.*, 2015). Thus, StringTie workflow dynamically considers transcript abundance during the assembly process, while Cufflinks relies on the parsimony principle. One option provided by both assemblers allows the user to define minimal abundance of transfrags to be reported in the final output annotation (Pertea *et al.*, 2015; Trapnell *et al.*, 2010). StringTie "-c" option sets the minimum read coverage (default 10), while Cufflinks "-min-frags-per-transfrag" sets the minimum number of reads aligned to a given transfrag (default 10). This may also influence the results obtained, as filtering solely based on the number of reads aligned may not be a proper estimation of abundance, since it does not account for the length of the transcript. Cufflinks was highlighted in a previous comparison of different methods to analyze AS from RNA-seq data in plants as more robust in detecting novel transcript isoforms, particularly in cases where an incomplete genome annotation was available (simulated RNA-seq datasets) (Liu *et al.*, 2014). StringTie was not included in this study (Liu *et al.*, 2014), but StringTie developers reported higher, but still close precision and sensitivity compared with Cufflinks in transcript prediction in simulated datasets and without providing reference annotation (Pertea *et al.*, 2015). In real datasets (RNA-seq data from human tissues), the number of transcripts predicted by both assemblers were similar. Yet, if only genes fully annotated were considered, StringTie showed to predict 44% more known transcripts (i.e. exactly matching known annotation) than Cufflinks (Pertea *et al.*, 2015). Other benchmarking analysis validated the strength of StringTie and Cufflinks under ideal conditions (perfect alignments and high transcript coverage), with StringTie showing better recall and Cufflinks better precision (Hayer *et al.*, 2015). However, these authors also highlighted that all algorithms designed to delineate transcript forms tend to

### 3. RESULTS AND DISCUSSION

---

make many false discoveries. Considering the quality of the inferred quantified transcript expression, StringTie showed to provide a better estimation, with the highest number of true positives (correctly called expressed) and lowest number of false positives and false negatives (Hayer *et al.*, 2015). Moreover, recent evidence by Williams *et al.* (2017) showed that a pipeline including STAR as read aligner and StringTie as assembler showed the best balance of precision and recall using real RNA-seq datasets derived human clinical samples.

Taking into account the results obtained for cork oak datasets, and the reported higher accuracy of StringTie, we decided to continue transcript prediction with this software. However, an additional optimization step was performed in order to improve transcript prediction.

#### 3.4 Tuning transcript assembly with StringTie

The number of isoforms per gene is a critical aspect for a robust estimation of isoform abundance, since isoform resolution models (applied by Cufflinks and StringTie for this purpose) may introduce a level of uncertainty in read assignments to exons, when these are shared between two or more isoforms from the same gene (Liu *et al.*, 2014; Pertea *et al.*, 2015). This adds to the already expected uncertainty related to ambiguous read mappings and low levels of gene expression in a given condition. Therefore transcriptome assembly with StringTie was repeated using different parameters in order to improve stringency in prediction of new transcripts based on coverage and transcript expression. The tested conditions are indicated in Table 2.1.

When increased stringency was just applied during the assembly step for each individual libraries, followed by a default merge step (str.t.def, str.t.cafj and str.t.cafj10), it was possible to observe a general decrease in total predicted genes and transcripts (Table 3.2). A greater decrease was observed when minimum junction coverage was increased to 10 (str.t.cafj10). A major contribution for this decrease was observed for novel isoforms and single exonic overlaps. Interestingly, the remaining classes followed the opposite trend, although in a smaller extent and not contributing significantly to the overall decrease observed in the number

### 3.4 Tuning transcript assembly with StringTie

---

of novel transcripts. The increase observed in classes such as the intergenic transcripts (unkown) or intronic overlaps could be related to the imposed increase in the minimum coverage for junctions. An increase in splice junction coverage may have increased fragmentation in transfrags, since the evidence for a link between two (or more) given exons was not considered when these were not supported by more than 2 (strt.cafj) or 10 (strt.cafj10) spliced reads. Two further `-merge` runs were performed using the assemblies obtained in `strt.cafj10`, testing a filtering parameter that discards assembled transfrags with estimated expression below 4 (-T4) or 10 (-T10) TPM. This resulted on a massive decrease in novel predicted transcripts (from 50% in `strt.cafj10-T4` to 80% in `strt.cafj10-T10`).

In addition to counting the total number of transcripts generated by all rounds of optimization, further characterization of the resulting annotations files was performed, accounting for number of exons, isoforms per gene and changes in gene structure. The use of default options during the merge step, with no filtering for low abundant genes (`strt.def`, `strt.cafj` and `strt.cafj10`), resulted in an increase in total number of exons and maximum number of exons in a gene, as well as single exon genes (Table A.1, Appendix A). This resulted in a wider distribution of the number of isoforms per genes, compared with the reference annotation (Figure 3.3). The filtering of low abundant transfrags to include in the merge steps (`strt.cafj10-T4` and `strt.cafj10-T10`) resulted in more conserved annotations with metrics closer to reference annotation, in what concerns structural features such as the number of exons and isoforms per gene (Table A.1, Figure 3.3). Taking into account all the new annotations it can be concluded that the number of genes with more than one isoform (and more likely to be alternatively spliced) only comprises about 10% to 20% of the all genes annotated (Figure 3.3). However, it must be noticed that this percentage may increase if this assessment is made for each tissue individually, since the expected number of genes actually being expressed would be less, thus reducing the transcript universe. It is also important to note that not all genes present in the reference genome annotation are expected to be expressed in the tissues used in the present analysis, and no new isoforms were determined for these cases.

The several annotations obtained in this study provide an estimation of total transcriptome diversity for the four tissues used. Nevertheless, the application of

### 3. RESULTS AND DISCUSSION

---

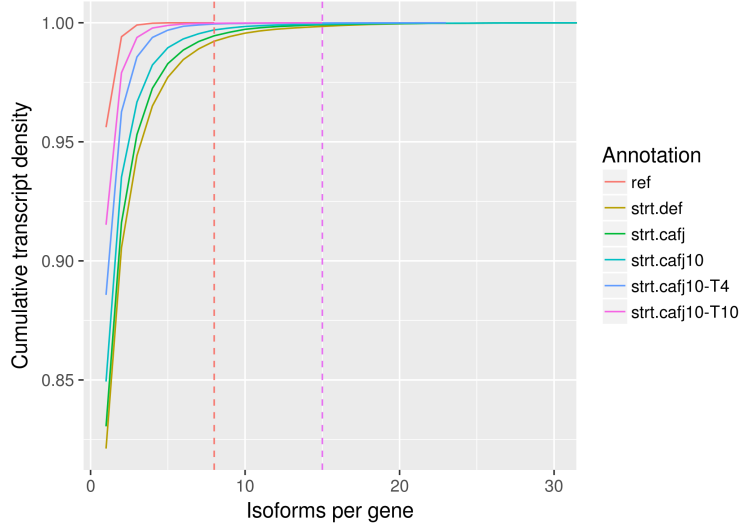


Figure 3.3: Isoforms per gene frequency determined for reference annotation (ref) and further annotations obtained after StringTie assemblies using specific optional parameters as described in Table 2.1. Vertical dashed lines indicate the maximum number of isoforms per gene found for reference (red) and most stringent StringTie run (str.t.cafj10-T10, pink).

stringent filtering for low abundant transcripts during the merging step severely decreased the total number of predicted transcripts/isoforms. Since this work is dealing with real datasets in a still poorly annotated genome, it is not possible to evaluate the precision of the assemblies. Therefore it was decided to continue the work with the most conservative assembly generated by str.t.cafj10-T10. The novel transcripts predicted in this annotation have a wider distribution of read coverage (Figure 3.4) than that obtained for default StringTie run (Figure 3.2).

### 3.5 Landscape of AS in cork oak

To further assess the occurrence of AS across the four different tissues, transcript coverage data was obtained from str.t.cafj10-T10 final annotation file. Transcripts expressed in each tissue (average FPKM  $> 1$ ) were identified based on this data and the corresponding structural annotations were analyzed. A total of 18,798 genes were expressed in leaves (22,046 transcripts), 20,611 genes in phellem (23,916 transcripts), 21,504 genes in inner bark (25,105 transcripts) and 21,190



### 3.5 Landscape of AS in cork oak

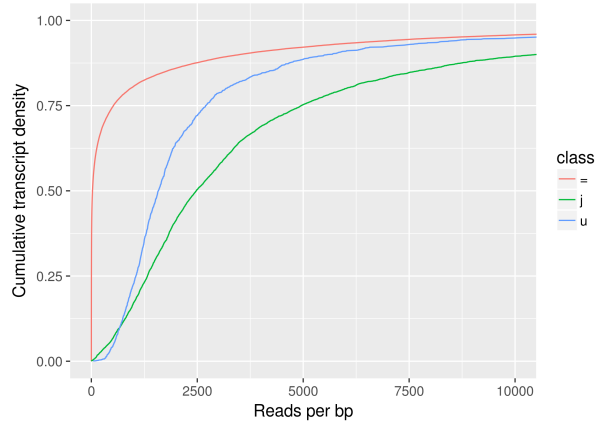


Figure 3.4: Categorization of *strf.cafj10-T10* annotated transcripts by estimated depth of read coverage (reads per base pair). Transcripts were grouped according to *gffcompare* classification as: transcripts with complete match of introns-exon coordinates with reference transcripts (=, red), novel isoforms (j, green) and unknown intergenic transcripts (u, blue).

genes in xylem (24,628 transcripts) (Table 3.3). Globally, 25,149 genes (29,296 transcripts) were expressed in at least one cork oak tissue, representing 31.27% of all genes present in the new transcriptome annotation (80,416 genes, Table 3.2). This result may be explained by the fact that only four tissues have been considered in the present analysis and only represent a portion of the whole cork oak transcriptome diversity. The remaining genes considered as not expressed were "inherited" from the original reference annotation, which contained a total of 79,752 gene predictions (Table 3.2).

A great proportion of the genes expressed in each of four tissues contained more than one exon, but only approximately 16% of these genes were alternatively spliced (i.e. present in each sample with more than one transcript isoform) (Table 3.3). In Poplar and Eucalyptus, analysis of AS occurring in xylem samples showed that only 28.3% and 20.7% of intron-containing genes, respectively, were alternatively spliced (Xu *et al.*, 2014). These results are closely related, yet above the estimation obtained for cork oak in the present study. However, when more comprehensive sets of libraries are used, for example including more tissues or full plants, AS events are detected in a higher proportion. In maize seedlings, AS was detected in 45.5% of multi-exonic genes (Huang *et al.*, 2015), while in Arabidopsis,

### 3. RESULTS AND DISCUSSION

Table 3.3: Quantification of single- and multi-exon genes expressed in leaf, phellem, inner bark and xylem. The extent of AS in each tissue was assessed through the quantification of multi-transcript genes composed of more than one exon.

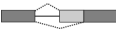





	Leaf	Phellem	Inner Bark	Xylem	All
Single-exon genes	3,011	3,370	3,374	3,403	4,117
Multi-exon genes	15,787	17,241	18,130	17,787	21,032
<i>Multi-transcript genes</i>	<i>2,532</i>	<i>2,646</i>	<i>2,848</i>	<i>2,758</i>	<i>3,279</i>
(%)	(16.04 %)	(15.35 %)	(15.71 %)	(15.05%)	(15.60 %)
Total	18,789	20,611	21,504	21,190	25,149

recent estimations ranged between 45% to 61% (Filichkin *et al.*, 2010; Marquez *et al.*, 2012). When considering the four tissues used in the present study, the proportion of AS events does not increase (Table 3.3). However, it should be stressed that the transcript annotation was obtained using stringent parameters, which may discard true, but low-abundant transcripts (false-negatives).

Intron retention (IR) has been previously proposed to be the most abundant AS event in Arabidopsis, representing close to 40% of total events (Reddy *et al.*, 2013). Chamala *et al.* (2015) also reported similar trends in other plant species. However, other studies have reported some variability namely in the proportion of alternative acceptor site (AA) events relative to intron retention (Huang *et al.*, 2015; Xu *et al.*, 2014; Zhang *et al.*, 2017), suggesting some differences in the regulation of splice site use in plants. In cork oak the most abundant event in all tissues was AA (32-34%), followed by IR (19-21%) (Table 3.4). Exon skipping (ES) was the third most abundant event followed by alternative donor site (AD). Other less abundant and more complex events included double exon skipping (ES1+2), alternate acceptor/donor sites (AA/AD) (Table 3.4), and double or alternate intron retention (not shown). The distribution of events was similar between all four tissues, and could represent a trend in this species, at least for this tissues under the conditions tested.

### 3.6 Differential expression analysis across the four tissues

Table 3.4: Top six of the most frequent AS events occurring in leaf, phellem, inner bark and xylem. An illustration of the intron-exon structure of each event is shown on the first column. The raw number of events (and percentage) is shown for each tissue

	Leaf		Phellem		Inner Bark		Xylem	
	Events	%	Events	%	Events	%	Events	%
AA 	683	31.87	719	32.77	814	33.68	781	33.66
IR 	448	20.91	450	20.51	457	18.91	444	19.14
ES 	413	19.27	398	18.14	441	18.25	440	18.97
AD 	358	16.71	357	16.27	388	16.05	370	15.95
ES1+2 	40	1.87	33	1.50	47	1.94	43	1.85
AA/AD 	36	1.68	39	1.78	36	1.49	35	1.51

AA: alternative acceptor (3' splice) site; IR: intron retention; ES: exon skipping; AD: alternative donor (5' splice) site; ES1+2: double exon skipping; AA/AD: AA or AD.

### 3.6 Differential expression analysis across the four tissues

To identify the regulatory pathways involved in tissue-specific development, and more particularly phellem and inner bark, we performed differential expression analysis at transcript level. The distribution of transcript abundances (estimated as FPKM values) across samples and a hierarchical clustering analysis was performed to check the consistency of all RNA-seq libraries. FPKM distribution was homogeneous within all replicates for each tissue (Figure A.2, Appendix A) and hierarchical clustering based on samples euclidean distance also showed great similarity within replicates (Figure A.3, Appendix A). Two major clusters were formed, one included leaf and xylem and the other including phellem and inner bark. Principal component analysis also confirmed the consistency among technical replicates and great variability between tissues (Figure A.4, Appendix A).

Since the RNA-seq datasets were obtained from four different tissues, none of them was considered a reference and differential expression was assessed using multi-group comparison. A total of 22,449 transcripts (out of 24,874 transcripts that passed the variance filter) were differentially expressed ( $q$ -value  $< 0.01$ ). The

### 3. RESULTS AND DISCUSSION

---

differentially expressed transcripts (DETs) were further clustered based on their pattern of expression into 16 clusters (Figure A.5, Appendix A). Out of these, 6 clusters were selected based on their expression profile: one representing xylem-enriched DETs (cluster 13, Figure 3.5A), two representing leaf-enriched DETs (cluster 5 and 16, Figure 3.5B) and three representing peridermis-specific DETs (cluster 3, 11, 15) (Figure 3.6). Due to time constraints, a complete functional annotation of the cork oak transcriptome was not possible, so it was decided to perform an homology search against the Arabidopsis transcriptome for an estimation of the functional categories enriched in each cluster.

The xylem-specific cluster (13) included 488 transcripts, 429 of which showed homology to Arabidopsis proteome. This cluster contained an enrichment of GO terms that relate to xylem function as a conduit for water and nutrient transport (Figure 3.5A). These include anion transport, transmembrane transporter activity and extracellular region and are annotated to transcripts encoding for sulfate, zinc, potassium and amino acid transporters. Other enriched GO terms in xylem-specific cluster related to response to disease or pathogens (response to virus, chitinase activity). These transcripts may contribute to the establishment of a response to specific pathogens that target the vascular system after being absorbed by the roots. Chitinase activity is important to hydrolyze chitin, a primary cell wall component in fungi (Yadeta & J Thomma, 2013). Cluster 5 and 16 included together 990 DETs (of which 811 had a match in the Arabidopsis proteome) showing higher expression in leaves, as compared to the other tissues (Figure 3.5B). A significant enrichment in photosynthesis-related terms was found (e.g. photosynthesis, photosynthetic membrane, thylakoid, chlorophyll binding), which included transcripts encoding structural units of Photosystems I and II (light-harvesting complex and reaction center sub-units), which are involved in light absorption and electron transfer and located in the thylakoid membranes from the chloroplasts (Nelson & Ben-Shem, 2004). Leaf-specific clusters also showed an over-representation of functional classes related to biotic and abiotic stimulus, as well as defense response. Other enriched GO terms, such as oxygen binding and monooxygenase activity were mostly related to a group of cytochrome P450 encoding transcripts (CYP81, CYP82, CYP87, CYP706, CYP71, CYP715,

### 3.6 Differential expression analysis across the four tissues

CYP76) involved in the metabolism (biosynthesis or catabolism) of hormones or secondary metabolites with a role in stress response (Bak *et al.*, 2011).

The peridermis-specific clusters included 340 transcripts enriched in phellem (Figure 3.6A), 786 transcripts enriched in inner bark (Figure 3.6B) and 692 transcripts highly expressed in both tissues but not in leaf or xylem (Figure 3.6C). Blastp search against the Arabidopsis proteome identified 318, 643 and 600 close homologs, respectively for each cluster. Similarly to xylem and leaf, GO terms

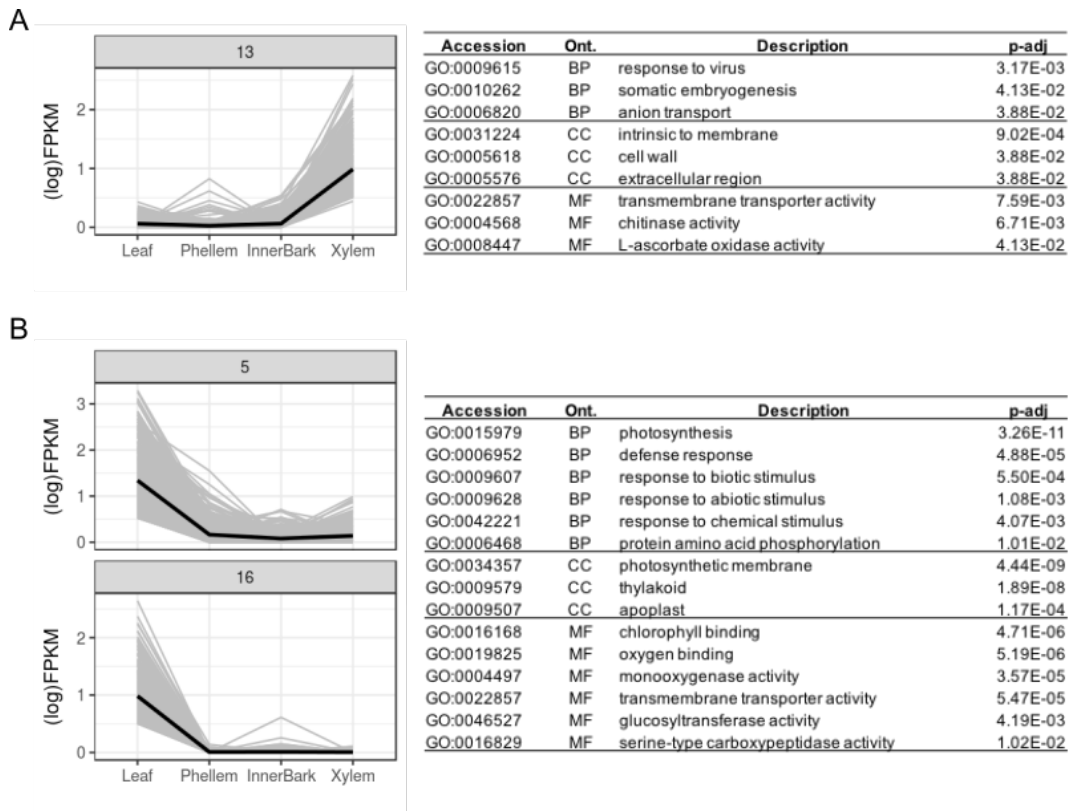


Figure 3.5: Expression profile and enriched GO terms of xylem and leaf-specific transcript clusters. Clusters of highly expressed transcripts in xylem (A), leaf (B) are represented. For each cluster (right panel), gray lines represent transcript expression profiles (log<sub>10</sub> FPKM) across the leaf, phellem, innerbark and xylem datasets, and the black line highlights the average expression for each cluster. Tables (left panel) contain representative GO terms shown to be enriched in each cluster [Hypergeometric test, with Benjamini and Hochberg false discovery rate correction, adjusted *p*-value (p-adj) < 0.05]. GO terms accessions (Acc.) include biological process (BP), cellular component (CC) and molecular function (MF) ontologies.

### 3. RESULTS AND DISCUSSION

related to response environmental cues were found to be enriched, but mostly in inner bark-enriched DETs (e.g. immune response, response to biotic stimulus, defense response) (Figure 3.6B). The overexpression of transcripts involved in plant

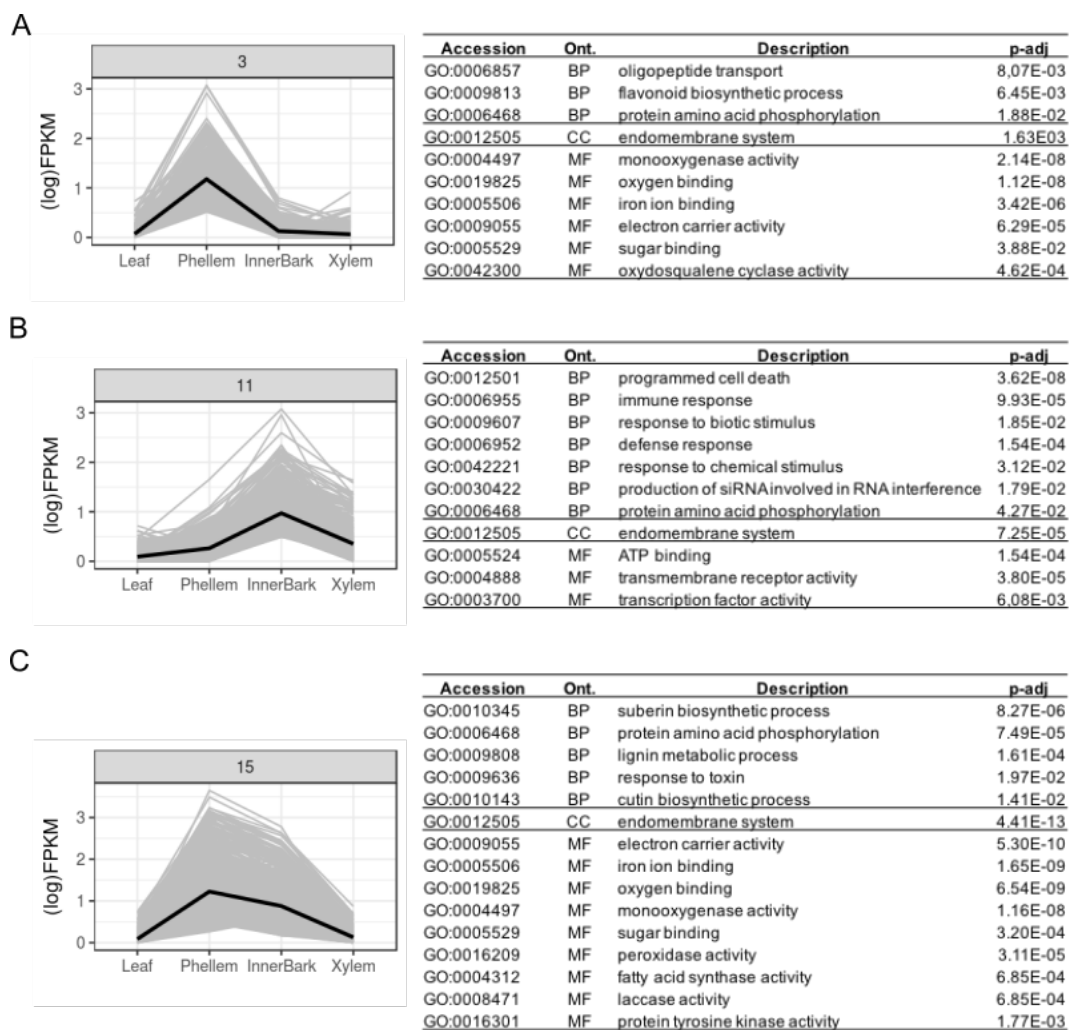


Figure 3.6: Expression profile and enriched GO terms of peridermis-specific transcript clusters. Clusters of highly expressed transcripts in phellem (A), inner bark (B) and both (C) are represented. For each cluster (right panel), gray lines represent transcript expression profiles (log<sub>10</sub> FPKM) across the leaf, phellem, innerbark and xylem datasets, and the black line highlights the average expression for each cluster. Tables (left panel) contain representative GO terms shown to be enriched in each cluster [Hypergeometric test, with Benjamini and Hochberg false discovery rate correction, adjusted  $p$ -value (p-adj) < 0.05]. GO terms accessions (Acc.) include biological process (BP), cellular component (CC) and and molecular function (MF) ontologies.

### 3.6 Differential expression analysis across the four tissues

---

response to environmental factors is not surprising since tissue samples were collected directly from a field grown tree, and may reflect to some extent endogenous contamination by pathogens and/or exposure to drought or heat, eliciting a response by the tree. However, this particular association with inner bark suggests a role of this layer in tree protection. As previously mentioned (Section 1.3) the peridermis results from the meristematic activity of the phellogen, that produces phellem to the outside and phelloderm (inner bark) to the inside (Graça & Pereira, 2004; Pereira, 2007). Unlike phellem cells, the fate and function of inner bark cells developed from the phellogen is largely unknown. Mature phellem cells are dead and their highly suberized cell walls impose a mechanical barrier against pathogens. Since inner bark is made of living cells and not suberized (Pereira, 2007), these can provide a second and more specialized layer of protection against pathogens that may cross the phellem barrier, through pores or cracks present in phellem.

During the first four years of development, the phellem cells are filled with tannins and already have suberized walls, and after the fifth year they become devoid of cellular content (Pereira, 2007). Cork physical properties (impermeability and flexibility) are mostly derived from cell wall chemical composition, which includes aliphatic suberin, aromatic suberin (also referred as cork lignin), waxes, tannins and polysaccharides (Graça, 2015; Pereira, 2007). In agreement with the cellular differentiation processes expected in cork cells, DETs with a role in the metabolic pathways involved in the synthesis of phenylpropanoid based polymers, such as tannins, lignin and suberin, were overrepresented in peridermis-specific clusters. Transcripts acting in the flavonoid biosynthetic pathway, necessary for tannins, were found exclusively in phellem-specific cluster 3 (Figure 3.6A), while transcripts related to suberin, lignin present in cluster 11 (Figure 3.6C). The latter cluster contains transcripts predominantly expressed in phellem, but also in inner bark at a lower level. This suggests that those pathways are also taking place in inner bark layer facing phellogen, even though the walls of these cells are not typically suberized (Pereira, 2007). Other transcripts abundant in these clusters belong to the super family of P450 cytochromes (monooxygenase activity, oxygen binding). Some of them belong to the same families as the ones found for leaves (CYP71, CYP81, CYP82, CYP87, CYP71, CYP76) while the remaining

### 3. RESULTS AND DISCUSSION

---

were related to specialized functions, such as: regulation of cell growth and differentiation through the brassinosteroid metabolism (CYP734 and CYP714 found in cluster 11) (Bak *et al.*, 2011; Vogt, 2010); suberin biosynthesis (CYP86 found in cluster 3 and 11); and phenylpropanoid pathway (CYP84 found in cluster 3).

GO enrichment analysis confirmed the representation of general metabolic processes and activities expected for each of the targeted tissues for RNA-seq. Although just providing a general overview of complex pathways acting in each tissue, it validated the new transcript annotation generated in this work. Still, a complete analysis of the transcripts overexpressed in each tissue, combined with a complete and more specific functional annotation, will be required highlight other important pathways not uncovered by GO term enrichment. Nevertheless, this analysis already shed some light into the metabolic pathways taking place in phellem and inner bark. Interestingly, transcripts involved in the synthesis of phellem cell wall components were enriched not only in phellem but also in inner bark (Figure 3.6C). This may suggest either, a contribution of this tissue in the synthesis of monomers that may be exported to the apoplast and be included in the assembly of phellem cell walls, or a certain level of cross contamination during sample collection, given to proximity of both tissues. Yet it should be highlighted that genes involved in tannin biosynthesis, which is the cellular content in young phellem cells (Graça & Pereira, 2004), were only found in the phellem-specific cluster (Figure 3.6A).

From the 22,449 DETs identified in this study, 15,139 (67.43%) were annotated as the unique isoforms for the corresponding gene entity. The remaining 7,310 DETs belong to multi-isoform loci, and 4,109 of these transcripts could be considered alternatively spliced, i.e. more than one isoform from the same gene was differentially expressed. This condition corresponded to 1,819 genes, which are candidates to investigate differential splicing between different tissues. Figure A.6 (Appendix A) shows two examples of differential splicing, detected for one cork oak *Alpha-amylase-like* gene (homolog to AT1G69830 gene from Arabidopsis) and one cork oak *Topoisomerase II-like* gene (homolog to AT3G23890 gene from Arabidopsis). The *Alpha-amylase-like* gene contained two isoforms that were only detected in inner bark and xylem, while all four annotated isoforms were expressed in leaves and phellem (Figure A.6 I). The *Topoisomerase-like*



### 3.6 Differential expression analysis across the four tissues

---

gene contained three isoforms (255, 254 and 253) that were highly expressed in inner bark and, to a lesser extent, in phellem, while a fourth isoform (252) was more abundant in xylem (Figure A.6 II). Further studies may be performed to evaluate the biological significance of these changes, and identify other cases of differential splicing.



# Chapter 4

## Conclusions

This project aimed to improve the detail of the genome annotation presently available for the draft genome sequence from cork oak, predicting new alternative splicing forms of genes expressed in four different tissues. The transcriptomes of leaf, phellem, inner bark and xylem were sequenced by RNA-seq and a sequence analysis workflow was defined in order to test the performance of HISAT2 and STAR for read mapping, and Cufflinks and StringTie for transcript assembly. STAR showed to be the aligner generating the highest mapping efficiencies for all the tissues and was selected for further analysis. StringTie was selected to assemble the transcriptome, since it was globally more conservative than Cufflinks, generating less novel transcripts, which could be better supported by coverage. Assembly with StringTie was further optimized in order to improve annotation precision, and a final and most conservative annotation was selected to assess transcript expression. Since the cork oak genome version is still a draft and no correct gene models can be used to test precision and recall, the selection of the most conservative annotation was a strategy to decrease the number of incorrectly assembled transcripts, although it could have discarded low abundant but correctly assembled transcripts.

The new transcript annotation was further used to estimate transcript expression and evaluate the extent of alternative splicing within the four tissues. Globally, about 16% of all intron-containing genes expressed in the four tissues were alternatively spliced. The analysis of AS events suggested that the main

## 4. CONCLUSIONS

---

event found in the four cork oak tissues is alternative acceptor (3' splice) site, followed by intron retention. Differentially expressed transcripts were identified and grouped according to their main expression in each tissue. The most enriched functional categories identified for each group were in agreement with the function of each tissue. In this way, transcripts highly expressed in leaves and xylem were mostly related to photosynthesis and transport, respectively, while transcripts highly expressed in peridermis (phellem and inner bark) showed an enrichment on functional categories related to the synthesis of suberin and other component of cork cell walls. All tissue-specific clusters showed an enrichment in transcripts involved in the response to stress (biotic or abiotic). However, this result was more striking in the inner bark-specific cluster, suggesting that this tissue is creating a second layer of protection in the trunk, after the physical barrier imposed by cork.

In conclusion, this thesis allowed the definition of a standard workflow that can be used to study alternative splicing in cork oak. Considering that this work was performed on a draft genome, it is likely that the final annotation generated still contains some assembly errors (e.g. due to genome fragmentation). Since a new improved genome version will be available soon, the workflow used in the present study will be a valuable contribution for transcript annotation.

# Appendix A

## Supplementary materials

### A.1 Transcriptome reconstruction: StringTie vs Cufflinks

## A. SUPPLEMENTARY MATERIALS

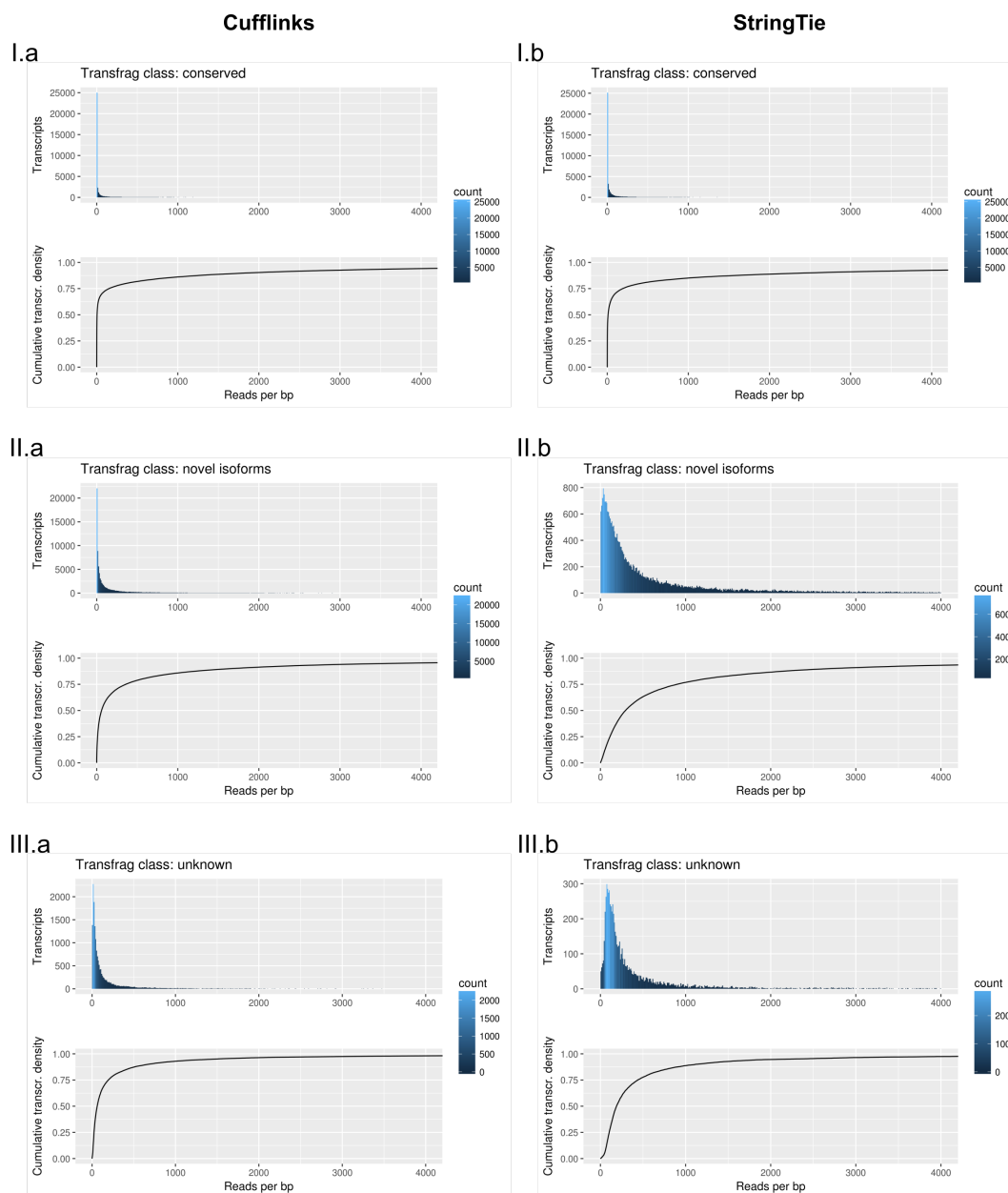


Figure A.1: Categorization of Cufflinks (a) and StringTie (b) predicted transcripts by estimated depth of read coverage (reads per base pair). Transcripts were grouped according to gffcompare classification as: transcripts with conserved intron-exon coordinates with reference transcripts (I), novel isoforms (II) and unknown intergenic transcripts (III). Each panel is composed of a histogram (up) and correspondent cumulative density plot (down) representing transcript number according to their predicted coverage, up to 4000 reads per bp.

## A.2 Tuning transcript assembly with StringTie

Table A.1: Characterization of genome annotation files, regarding number of exons and transcripts per genes, from reference and further annotations generated after the optimization rounds with StringTie.

	reference	strt.def	strt.cafj	strt.cafj10	strt.cafj10-T4	strt.cafj10-T10
Total number of exons	285193	372013	361602	346795	315905	300490
Mean exons in a gene	4	4	4	4	4	4
Max exons in a gene	73	122	96	160	73	73
Single exon genes	25536	28094	29157	30651	27170	26137
(%)	(32.01%)	(33.61%)	(34.42%)	(35.57%)	(33.18%)	(32.50%)
Genes with 1 isoform	76254	68638	70347	73181	72528	73595
(%)	(95.61%)	(82.12%)	(83.05%)	(84.93%)	(88.57%)	(91.52%)
Mean isoforms per gene:	1.05	1.45	1.39	1.30	1.18	1.12
Max isoforms per gene:	8	43	38	48	23	15

## A. SUPPLEMENTARY MATERIALS

---

### A.3 Differential expression analysis across the four tissues

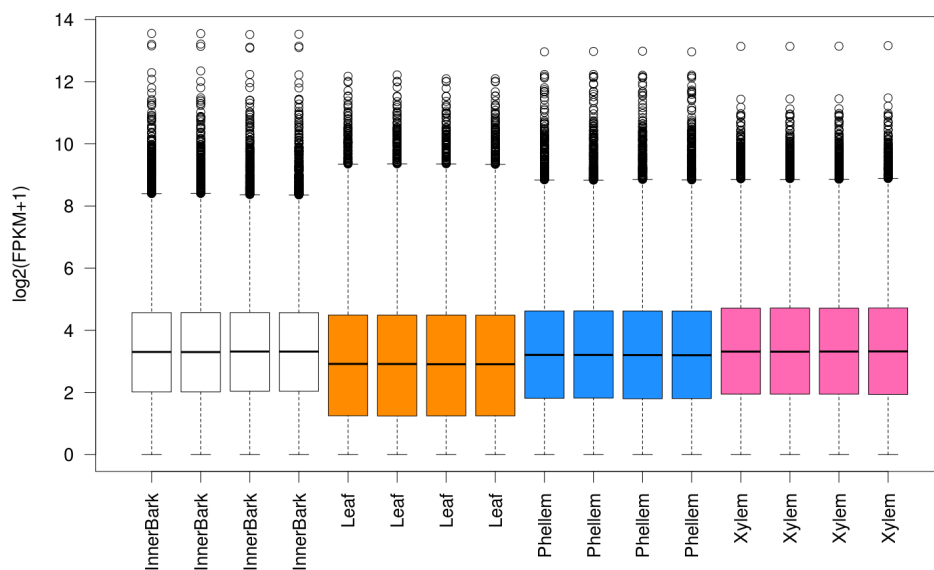


Figure A.2: Distribution of FPKM values determined for transcripts expressed in the 16 samples used for differential expression analysis. The four libraries obtained for each tissue are shown: white for inner bark, orange for leaf, blue for phellem and pink for xylem.



### A.3 Differential expression analysis across the four tissues

---

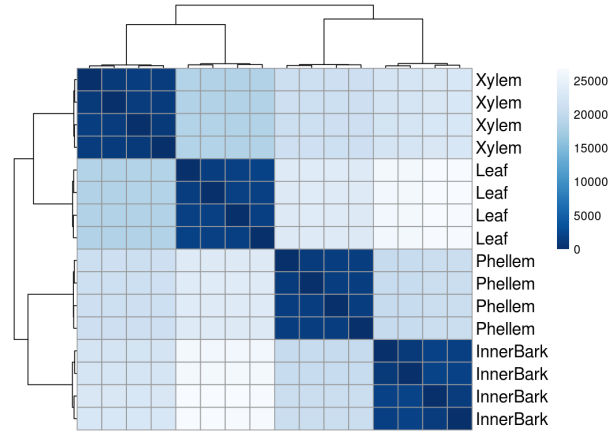


Figure A.3: Hierarchical clustering and heatmap based on sample-to-sample euclidean distances, computed from the FPKM values estimated for all expressed transcripts.

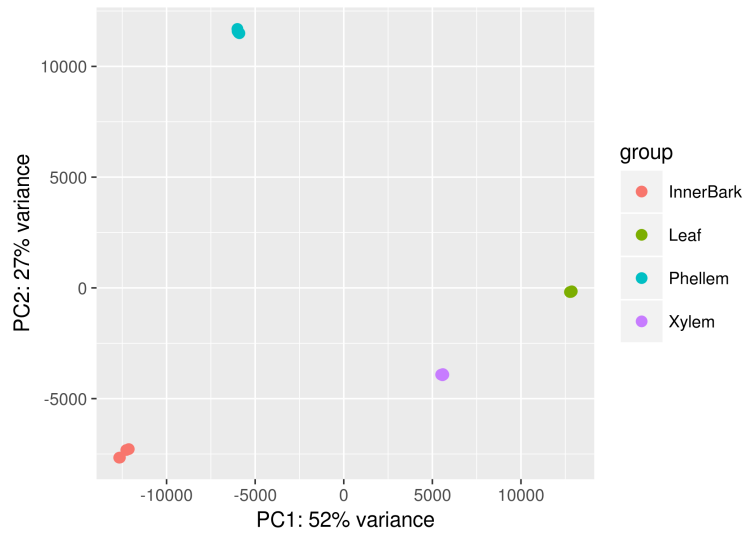


Figure A.4: Principal component analysis of the samples used for transcript expression analysis, based on FPKM values estimated for expressed transcripts. This analysis was performed using the four technical replicates obtained for each tissue, which clustered in close proximity: red for inner bark, green for leaf, blue for phellem and purple for xylem.

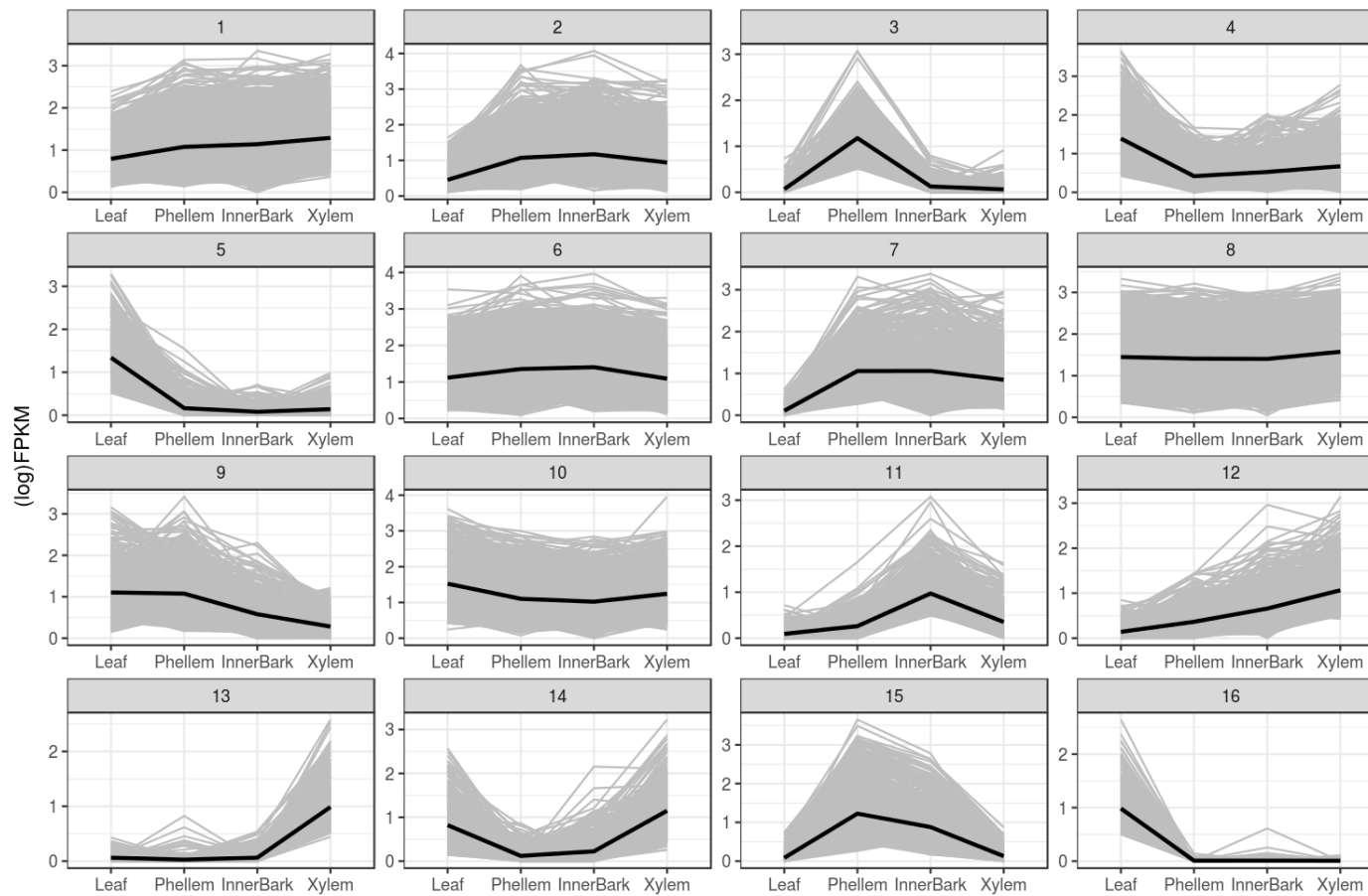
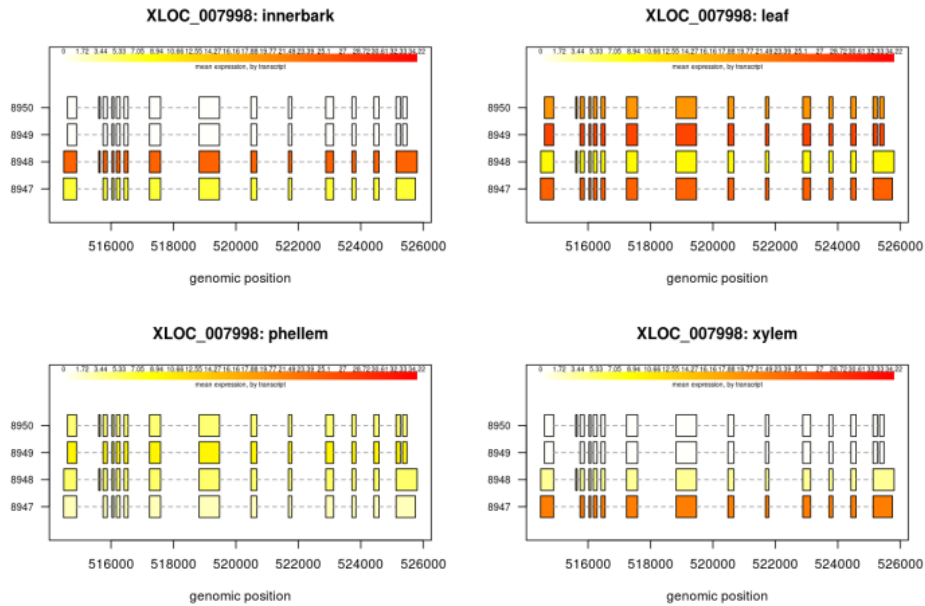


Figure A.5:  $k$ -means clustering analysis of differentially expressed transcripts ( $k=16$ ). The gray lines represent mean expression profile (log<sub>10</sub> FPKM) for each transcript across tissues. The black line represents the mean expression profile observed in each cluster.

### A.3 Differential expression analysis across the four tissues

I.



II.

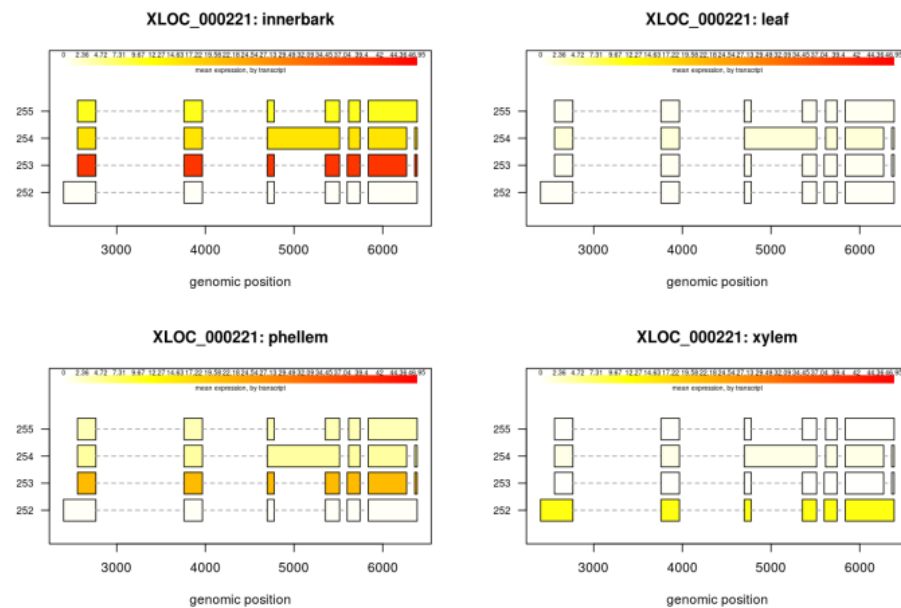


Figure A.6: Examples of differential splicing found between tissues for two cork oak loci: (I) *Alpha-amylase-like* and (II) *Topoisomerase II-like*. Exon-intron structure is shown for each annotated isoform (internal annotation IDs are shown on the left side of each box) and estimated mean FPKM expression determined for each tissue is represented by the color scale (white to red).



## References

- ALAMANCOS, G.P., AGIRRE, E. & EYRAS, E. (2014). Methods to study splicing from high-throughput RNA sequencing data. In *Spliceosomal Pre-mRNA Splicing*, 357–397, Humana Press, Totowa, NJ. [4](#), [5](#), [7](#)
- 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**, 4570–4578. [5](#)
- BAK, S., BEISSON, F., BISHOP, G., HAMBERGER, B., HÖFER, R., PAQUETTE, S. & WERCK-REICHHART, D. (2011). Cytochromes P450. *The Arabidopsis Book*, **9**, e0144. [33](#), [36](#)
- BARUZZO, G., HAYER, K.E., KIM, E.J., DI CAMILLO, B., FITZGERALD, G.A. & GRANT, G.R. (2016). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature methods*. [4](#), [5](#), [6](#), [20](#), [21](#)
- BUGALHO, M.N., CALDEIRA, M.C., PEREIRA, J.S., ARONSON, J. & PAUSAS, J.G. (2011). Mediterranean cork oak savannas require human use to sustain biodiversity and ecosystem services. *Frontiers in Ecology and the Environment*. [9](#)
- CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T.L. (2009). BLAST+: architecture and applications. *Bmc Bioinformatics*, **10**, 421. [18](#)
- CHAMALA, S., FENG, G., CHAVARRO, C. & BARBAZUK, W.B. (2015). Genome-Wide Identification of Evolutionarily Conserved Alternative Splicing

## REFERENCES

---

- Events in Flowering Plants. *Frontiers in Bioengineering and Biotechnology*, **3**, 30
- CLINE, M.S., SMOOT, M., CERAMI, E., KUCHINSKY, A., LANDYS, N., WORKMAN, C., CHRISTMAS, R., AVILA-CAMPILO, I., CREECH, M., GROSS, B., HANSPERS, K., ISSERLIN, R., KELLEY, R., KILLCOYNE, S., LOTIA, S., MAERE, S., MORRIS, J., ONO, K., PAVLOVIC, V., PICO, A.R., VAILAYA, A., WANG, P.L., ADLER, A., CONKLIN, B.R., HOOD, L., KUIPER, M., SANDER, C., SCHMULEVICH, I., SCHWIKOWSKI, B., WARNER, G.J., IDEKER, T. & BADER, G.D. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature protocols*, **2**, 2366–2382. 18
- 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**, 15–21. 5, 6, 13
- DUBROVINA, A.S., KISELEV, K.V. & ZHURAVLEV, Y.N. (2013). The role of canonical and noncanonical pre-mRNA splicing in plant stress responses. *BioMed research international*, **2013**, 264314. 3
- ENGSTRÖM, P.G., STEIJGER, T., SIPOS, B., GRANT, G.R., KAHLES, A., RÄTSCH, G., GOLDMAN, N., HUBBARD, T.J., HARROW, J., GUIGÓ, R., BERTONE, P. & RGASP CONSORTIUM (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, **10**, 1185–1191. 4, 5
- FILICHKIN, S.A., PRIEST, H.D., GIVAN, S.A., SHEN, R., BRYANT, D.W., FOX, S.E., WONG, W.K. & MOCKLER, T.C. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome research*, **20**, 45–58. 30
- FLOREA, L.D. & SALZBERG, S.L. (2013). Genome-guided transcriptome assembly in the age of next-generation sequencing. *IEEE/ACM transactions on computational biology and bioinformatics*, **10**, 1234–1240. 4, 8

## REFERENCES

---

- FOISSAC, S. & SAMMETH, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic acids research*, **35**, W297–9. 17
- FRAZEE, A.C., PERTEA, G., JAFFE, A.E., LANGMEAD, B., SALZBERG, S.L. & LEEK, J.T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature biotechnology*, **33**, 243–246. 18
- FU, J., FRAZEE, A., COLLADO-TORRES, L., JAFFE, A. & LEEK, J. (2017). ballgown: Flexible, isoform-level differential expression analysis. r package version 2.8.4. <https://github.com/alyssafrazee/ballgown>. 18
- GOFF, L., TRAPNELL, C. & KELLEY, D. (2013). cummerbund: Analysis, exploration, manipulation, and visualization of cufflinks high-throughput sequencing data. r package version 2.18.0. <https://github.com/vikas0633/R/blob/master/CummeRbund.Rmd>. 18
- GRAÇA, J. (2015). Suberin: the biopolyester at the frontier of plants. *Frontiers in chemistry*, **3**, 329. 35
- GRAÇA, J. & PEREIRA, H. (2004). The periderm development in *Quercus suber*. *IAWA Journal*, **25**, 325–335. 10, 35, 36
- 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., POCHE, N., STROZZI, F., WEEKS, N., WESTERMAN, R., WILLIAM, T., DEWEY, C.N., HENSCH, 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**, 1494–1512. 4
- HAYER, K.E., PIZARRO, A., LAHENS, N.F., HOGENESCH, J.B. & GRANT, G.R. (2015). Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics (Oxford, England)*, **31**, 488–495. 8, 9, 25, 26

## REFERENCES

---

- HEATHER, J.M. & CHAIN, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, **107**, 1–8. [3](#)
- HUANG, J., GAO, Y., JIA, H., LIU, L., ZHANG, D. & ZHANG, Z. (2015). Comparative transcriptomics uncovers alternative splicing changes and signatures of selection from maize improvement. *BMC genomics*, **16**, 363. [3](#), [29](#), [30](#)
- HUG, N., LONGMAN, D. & CÁCERES, J.F. (2016). Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic acids research*, **44**, 1483–1495. [2](#)
- JOSHI, NA AND FASS, JN (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for fastq files (version 1.33). <https://github.com/najoshi/sickle>. [13](#)
- KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, **14**, R36. [5](#)
- KIM, D., LANGMEAD, B. & SALZBERG, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, **12**, 357–360. [5](#), [6](#), [13](#), [20](#)
- LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, .G.P.D.P. (2009). The sequence alignment/map (sam) format and samtools. *Bioinformatics*, **25**, 2078–2079. [16](#)
- LIU, R., LORAINE, A.E. & DICKERSON, J.A. (2014). Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *Bmc Bioinformatics*, **15**, 364. [7](#), [25](#), [26](#)
- MAERE, S., HEYMANS, K. & KUIPER, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, **21**, 3448–3449. [18](#)



## REFERENCES

---

- MARQUEZ, Y., BROWN, J.W.S., SIMPSON, C., BARTA, A. & KALYNA, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome research*, **22**, 1184–1195. [3](#), [30](#)
- NELSON, N. & BEN-SHEM, A. (2004). The complex architecture of oxygenic photosynthesis. *Nature reviews. Molecular cell biology*, **5**, 971–982. [32](#)
- NILSEN, T.W. & GRAVELEY, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463. [2](#)
- OLIVEIRA, G. & COSTA, A. (2012). How resilient is *Quercus suber* L. to cork harvesting? A review and identification of knowledge gaps. *Forest Ecology and Management*, **270**, 257–272. [10](#)
- PEREIRA, H., ed. (2007). *Cork: Biology, Production and Uses*. Elsevier Science, Amsterdam, 1st edn. [10](#), [35](#)
- PERTEA, M., PERTEA, G.M., ANTONESCU, C.M., CHANG, T.C., MENDELL, J.T. & SALZBERG, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, **33**, 290–295. [7](#), [8](#), [9](#), [16](#), [25](#), [26](#)
- PERTEA, M., KIM, D., PERTEA, G.M., LEEK, J.T. & SALZBERG, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols*, **11**, 1650–1667. [8](#)
- RAINS, M.K., GARDIYEHWA DE SILVA, N.D. & MOLINA, I. (2017). Reconstructing the suberin pathway in poplar by chemical and transcriptomic analysis of bark tissues. *Tree physiology*, 1–22. [10](#)
- RAMOS, A.M., USIÉ, A., BARBOSA, P., BARROS, P.M., CAPOTE, T., CHAVES, I., SIMÕES, F., ABREU, I., CARRASQUINHO, I., FARO, C., GUIMARÃES, J., MENDONÇA, D., NÓBREGA, F., RODRIGUES, L., SAIBO, N.J.M., VARELA, C., EGAS, C., MATOS, J., CÉLIA, M., OLIVEIRA, M.M., RICARDO, C.P.P. & GONÇALVES, S. (Submitted). The draft genome sequence of cork oak. *Under Revision*. [1](#), [10](#), [11](#), [19](#)

## REFERENCES

---

- REDDY, A.S.N., MARQUEZ, Y., KALYNA, M. & BARTA, A. (2013). Complexity of the alternative splicing landscape in plants. *The Plant cell*, **25**, 3657–3683. [4](#), [30](#)
- RICARDO, C.P.P., MARTINS, I., FRANCISCO, R., SERGEANT, K., PINHEIRO, C., CAMPOS, A., RENAUT, J. & FEVEREIRO, P. (2011). Proteins associated with cork formation in *Quercus suber* L. stem tissues. *Journal of proteomics*, **74**, 1266–1278. [10](#)
- ROBERTSON, G., SCHEIN, J., CHIU, R., CORBETT, R., FIELD, M., JACKMAN, S.D., MUNGALL, K., LEE, S., OKADA, H.M., QIAN, J.Q., GRIFFITH, M., RAYMOND, A., THIESSEN, N., CEZARD, T., BUTTERFIELD, Y.S., NEWSOME, R., CHAN, S.K., SHE, R., VARHOL, R., KAMOH, B., PRABHU, A.L., TAM, A., ZHAO, Y., MOORE, R.A., HIRST, M., MARRA, M.A., JONES, S.J.M., HOODLESS, P.A. & BIROL, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature methods*, **7**, 909–912. [4](#)
- SOLER, M., SERRA, O., MOLINAS, M., HUGUET, G., FLUCH, S. & FIGUERAS, M. (2007). A genomic approach to suberin biosynthesis and cork differentiation. *Plant physiology*, **144**, 419–431. [10](#)
- TRAPNELL, C., WILLIAMS, B.A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M.J., SALZBERG, S.L., WOLD, B.J. & PACHTER, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**, 511–515. [8](#), [16](#), [25](#)
- TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D.R., PIMENTEL, H., SALZBERG, S.L., RINN, J.L. & PACHTER, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, **7**, 562–578. [5](#), [8](#)
- VOGT, T. (2010). Phenylpropanoid Biosynthesis. *Molecular plant*, **3**, 2–20. [36](#)
- WANG, K., SINGH, D., ZENG, Z., COLEMAN, S.J., HUANG, Y., SAVICH, G.L., HE, X., MIECZKOWSKI, P., GRIMM, S.A., PEROU, C.M., MACLEOD, J.N.,

## REFERENCES

---

- CHIANG, D.Y., PRINS, J.F. & LIU, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research*, **38**, e178–e178. [4](#)
- WILLIAMS, C.R., BACCARELLA, A., PARRISH, J.Z. & KIM, C.C. (2017). Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *Bmc Bioinformatics*, **18**, 38. [8](#), [21](#), [26](#)
- WU, T.D. & NACU, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, **26**, 873–881. [5](#)
- XIE, B.B., LI, D., SHI, W.L., QIN, Q.L., WANG, X.W., RONG, J.C., SUN, C.Y., HUANG, F., ZHANG, X.Y., DONG, X.W., CHEN, X.L., ZHOU, B.C., ZHANG, Y.Z. & SONG, X.Y. (2015). Deep RNA sequencing reveals a high frequency of alternative splicing events in the fungus *Trichoderma longibrachiatum*. *BMC genomics*, **16**, 54. [3](#)
- XU, P., KONG, Y., SONG, D., HUANG, C., LI, X. & LI, L. (2014). Conservation and functional influence of alternative splicing in wood formation of *Populus* and *Eucalyptus*. *BMC genomics*, **15**, 780. [3](#), [29](#), [30](#)
- YADETA, K.A. & J THOMMA, B.P.H. (2013). The xylem as battleground for plant hosts and vascular wilt pathogens. *Frontiers in plant science*, **4**, 97. [32](#)
- ZHANG, R., CALIXTO, C.P.G., MARQUEZ, Y., VENHUIZEN, P., TZIOUTZIOU, N.A., GUO, W., SPENSLEY, M., ENTIZNE, J.C., LEWANDOWSKA, D., TEN HAVE, S., FREI DIT FREY, N., HIRT, H., JAMES, A.B., NIMMO, H.G., BARTA, A., KALYNA, M. & BROWN, J.W.S. (2017). A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic acids research*, **45**, 5061–5073. [3](#), [30](#)