Universidade de Trás-os-Montes e Alto Douro

Variation detection in organelle genomes of *Quercus* species

Dissertação de Mestrado em Bioinformática e Aplicações às Ciências da Vida

Ana Rita Moreira Ferreira

Orientador: Doutor António Marcos Ramos

Coorientadora: Professora Doutora Irene Oliveira

Coorientadora: Doutora Ana Isabel Usié Chimenos

Vila Real, 2020

Universidade de Trás-os-Montes e Alto Douro

Variation detection in organelle genomes of *Quercus* species

Dissertação de Mestrado em Bioinformática e Aplicações às Ciências da Vida

Ana Rita Moreira Ferreira

Orientador: Doutor António Marcos Ramos

Coorientadora: Professora Doutora Irene Oliveira

Coorientadora: Doutora Ana Isabel Usié Chimenos

Composição do júri:

Presidente:

- Doutor Eduardo José Solteiro Pires

Vogais:

- Doutora Maria Teresa Correia Guedes Lino Neto

- Doutor António Marcos Costa do Amaral Ramos

Vila Real, 2020

Orientação Científica

Doutor António Marcos Ramos

Centro de Biotecnologia Agrícola e Agroalimentar do Alentejo, CEBAL

Doutora Ana Isabel Usié Chimenos

Centro de Biotecnologia Agrícola e Agroalimentar do Alentejo, CEBAL

Professora Doutora Irene Oliveira

Universidade de Trás-os-Montes e Alto Douro, UTAD

Statement of Integrity

It is stated on the honour commitment that this academic work was conducted with integrity and expressly prepared by the author, as an original dissertation, for the purpose of obtaining a Master's degree in Bioinformatics and Applications to the Life Sciences at the University of Trás-os-Montes and Alto Douro.

All non-original contributions were duly identified with an indication of the source and I confirm that I have not used any form of undue use of information or falsification of results throughout the process of preparing this dissertation.

Acknowledgments

Esta dissertação marca o fim de um capítulo importante e não podia começar sem antes dirigir alguns agradecimentos a quem me ajudou a chegar até aqui.

Ao Centro de Biotecnologia Agrícola e Agroalimentar do Alentejo (CEBAL), por me receberem, por todas as condições cedias essenciais para a realização do estágio e desta dissertação.

Ao meu orientador, Doutor Marcos Ramos, por ter aceitado a minha orientação e me ter dado esta oportunidade, pela ajuda e disposição.

À minha coorientadora, Doutora Ana Isabel Chimenos, pelo apoio diário prestado desde o primeiro dia, por tudo o que me ensinou, pelo rigor e dedicação indispensável para a realização desta dissertação. Muito obrigada também pela amizade e momentos descontraídos.

À minha coorientadora, Professora Doutora Irene Oliveira, por toda a ajuda e disponibilidade ao longo desta etapa.

À minha família do coração pelo companheirismo e apoio ao longo de todos estes anos, por nunca falharem em nenhum momento e mostrarem sempre o verdadeiro significado de amizade.

À Sara, por ser um pilar indispensável, por toda a força, pela paciência, por não me deixar desistir e me motivar, por acalmar as tempestades, pela firmeza e por ser sempre o meu porto de abrigo.

Por último, o mais importante de todos, à minha família. Por todo o apoio, por nunca deixarem que nada me falte e por continuarem a aliviar-me o peso do mundo. Sem vocês, nada disto seria feito.

Mais uma vez, o meu sincero obrigado!

Abstract

The use of next-generation sequencing (NGS) technologies has been revolutionizing the study of genetics. The big amount of generated data by these technologies allows the characterization of species genomes and genomic variation between species and within the same one. The present dissertation focused on the study of single nucleotide polymorphisms (SNPs) in organelle genomes of two well-known species of *Quercus*: *Quercus suber* and *Quercus ilex rotundifolia*, commonly named cork oak and holm oak respectively.

Chloroplasts and mitochondria are organelles present in plant cells and play a crucial role in photosynthesis and energy metabolism, respectively, among other important physiological functions. Each of these organelles has its own genome, distinct from the nuclear genome. Within the *Quercus* genus, the chloroplast genome sequences have been determined for over 20 species, including cork oak which was assembled by the Genosuber consortium, and only the cork oak mitochondrial genome has been determined to date. Moreover, the amount of information on genomic variation is very scarce in chloroplast genomes, or non-existent in mitochondrial genomes. Therefore, there is a clear need to increase our knowledge in this field, given the importance of these species in the ecosystem and their socio-economic impact especially in the south region of the Iberic Peninsula.

The pipeline of this study involves the use of high-throughput sequencing data of 47 individuals (39 cork oaks and 8 holm oaks) using NGS techniques and tools to perform quality control, preprocessing, read mapping, variant calling and annotation. Additionally, to achieve the best performance on preprocessing and variant calling the used tools were tested using different parameters on a smaller group of individuals.

As it was expected given the higher conservation of chloroplast genomes, the presented results show a higher variation on mitochondrial genomes, especially when comparing cork oak with holm oak trees. These variations suggest a different capacity in both species and some studies have been reporting that holm oak is more resistant than cork oak and these variations may be the reason for that. With this in mind, it is possible to say that holm oak trees have greater ability to withstand climate change and therefore be a good model for selection of important molecular markers.

Keywords: *Quercus* genus, Organelle genomes, Next-Generation Sequencing, SNPs, Bioinformatics.

Resumo

A utilização de tecnologias de *next-generation sequencing* (NGS) tem vindo a revolucionar os estudos genéticos. A grande quantidade de dados gerados por estas tecnologias permite a caracterização dos genomas das espécies e a variação genómica entre espécies e dentro da mesma. A presente dissertação focou-se no estudo de polimorfismos de nucleótidos únicos (SNPs) em genomas de organelos de duas espécies bem conhecidas de *Quercus*: *Quercus suber* e *Quercus ilex rotundifolia*, vulgarmente designadas por sobreiro e azinheira, respetivamente.

Os cloroplastos e mitocôndrias são organelos presentes nas células vegetais e desempenham um papel crucial na fotossíntese e no metabolismo energético, respetivamente, entre outras importantes funções fisiológicas. Cada um destes organelos tem o seu próprio genoma, distinto do genoma nuclear e pouco se sabe sobre eles em espécies de *Quercus*. No entanto, as sequências do genoma do cloroplasto foram determinadas em mais de 20 espécies, incluindo o sobreiro cujo *assembly* foi feito pelo consórcio Genosuber. Por outro lado, apenas o genoma mitocondrial do sobreiro foi determinado até à data. Além disso, a quantidade de informação sobre a variação genómica é muito escassa nos genomas dos cloroplastos, ou inexistente nos genomas mitocondriais. Portanto, existe uma clara necessidade de aumentar os nossos conhecimentos neste campo, dada a importância destas espécies no ecossistema e o seu impacto socioeconómico, especialmente na região sul da Península Ibérica.

A estrutura deste estudo envolve a utilização de dados de sequenciação de alto rendimento de 47 indivíduos (39 sobreiros e 8 azinheiras) utilizando técnicas e ferramentas de NGS para realizar o controlo de qualidade, pré-processamento, mapeamento, determinação de variantes e anotação. Para além disso, para alcançar o melhor desempenho no pré-processamento e na determinação de variantes, as ferramentas utilizadas foram testadas utilizando diferentes parâmetros num grupo mais pequeno de indivíduos.

Como era de esperar dada a maior conservação dos genomas dos cloroplastos, os resultados apresentados mostram uma maior variação nos genomas mitocondriais, especialmente quando se compara o sobreiro com a azinheira. Estas variações sugerem uma capacidade diferente em ambas as espécies e alguns estudos têm relatado que a azinheira é mais resistente do que o sobreiro e estas variações podem ser a razão para isso. Com isto em mente, é possível dizer que as azinheiras têm maior capacidade de resistir às alterações climáticas e, portanto, ser um bom modelo para a seleção de marcadores moleculares importantes.

Palavras-chave: Género *Quercus*, genomas de organelos, Sequenciação de Nova Geração, SNPs, Bioinformática.

Index

List of Figures

List of Tables

List of Abbreviations, symbols or acronyms

ABA	Abscisic Acid
ABI	Applied Biosystems
ACCase	Acetyl-CoA Carboxylase
ATP	Adenosine Triphosphate
BAM	Binary Alignment/Map
bp	Base pairs
BWA	Burrows Wheeler Alignment
BWA-MEM	Burrows-Wheeler Alignment of Maximal Exact Matches
BWA-SW	Burrows-Wheeler Aligner's Smith-Waterman Alignment
CEBAL	Centro de Biotecnologia Agrícola e Agroalimentar do Alentejo
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
cpDNA	Chloroplast DNA
ddNTPs	dideoxynucleotides
DNA	Deoxyribonucleic acid
dNTPs	deoxynucleotides
GA	Gibberellic Acid
GFF3	Genetic Feature Format version 3
Indels	Insertions/Deletions
lncRNA	Long Noncoding RNA
MAF	Minor Allele Frequency
Mbp	Mega base pairs
MEM	Maximal exact matches
MNPs	Multi-Nucleotide Polymorphisms
mtDNA	Mitochondrial DNA
ncRNA	Noncoding RNA
NEP	Nucleous Encoded RNA Polymerase
NGS	Next Generation sequencing
PCR	Polymerase Chain Reaction
PE	Paired-end
PEP	Plastid Encoded RNA Polymerase

PSI	Photosystem I
PSII	Photosystem II
RG	Read Group
RGID	Read Group Identification
RGLB	Read Group Library
RGPL	Read Group Platform
RGPU	Read Group Platform Unit
RGSM	Read Group Sample Name
RNA	Ribonucleic acid
rRNA	Ribossomal RNA
SAM	Sequence Alignment/Map
SBL	Sequencing By Ligation
SBS	Sequencing By Synthesis
SE	Single-end
SMRT	Single Molecule Real Time sequencing
SNPs	Single Nucleotide Polymorphisms
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
TGS	Third Generation Sequencing
tRNA	Transfer RNA
VCF	Variant Call Format
WGRS	Whole Genome Re-Sequencing

Chapter 1: Introduction

1.1. Next Generation Sequencing

The appearance of sequencing technologies has come to make easier, faster and more effective the study of genomic features of organisms. Sanger and his colleagues (1977) and Maxam and Gilbert (1977) were pioneers in the development of methods to sequence DNA molecules, the first by chain termination and the second by fragmentation techniques. Since the emergence of these techniques, for over 30 years the Sanger approach, named "Sanger Sequencing Technology", was the prevailing method for DNA sequencing since it was less dangerous than Maxam and Gilbert's method by having less contact with toxic chemicals and radioisotopes (van Dijk *et al.*, 2014; Kchouk *et al.*, 2017).

The Sanger method involves the synthesis of complementary DNA using dNTPs (2'-deoxynucleotides) and termination of synthesis using ddNTPs (2',3'-dideoxynucleotides). The work between these two processes generates a set of fragments of which the only difference is in the nucleoside monophosphate units. Then, based on their size, the fragments are separated using electrophoresis gel and the DNA sequence is discovered (Sanger *et al.*, 1977; Metzker, 2005). In case of automated Sanger sequencing, there is an identification with a fluorescent dye (typically four different dyes for each nucleotide), allowing the assignment of a base call and revealing the DNA sequence based on the fluorescent order. The last step before obtaining readable DNA sequence information is to eliminate noise, correct dye alterations and emission intensities normalization (Metzker, 2005; Kulski, 2016). Sanger was the most used sequencing method until the appearance of next generation methods.

With the increasing demand of faster, more economic and more effective and throughput technologies for sequencing larger and more complex genomes, a new era of sequencing technologies arose by the name of "Next Generation Sequencing" (NGS) or "High Throughput Sequencing Technologies". These were firstly introduced in 2005 by Roche's 454 technology (Kchouk *et al.*, 2017) being able to produce less expensive high throughput sequences.

The great prominence of NGS is the capacity to produce high quality parallel analysis from multiple samples at the same time, sequencing millions of reads in a single run in only a few hours or days. For example, using a first generation sequencing method,

Sanger sequencing, the entire Human genome took about 15 years to be completely sequenced with the cost of about 100 million US dollars and it demanded the cooperation of many laboratories around the world. On the other hand, the same goal was accomplished using NGS sequencers, 454 Genome Sequencer FLX, in only two months costing only, approximately 10 000 US dollars (Wheeler *et al.*, 2008; Kchouk *et al.*, 2017). Another big advantage of NGS technology is the production of paired-end (PE) sequencing reads. This type of sequencing involves sequencing both ends of the DNA fragment in the library and align the forward and reverse reads as read pairs. The distance between each paired read is known and so, the alignment algorithm uses this information to map the reads over repetitive regions in a more precise way, facilitating the ability to detect variations, such as indels, and also the ability to remove possible PCR duplicates from the library preparation step. This type of run also makes it easier to find information of position in the genome, ideal for *de novo* genome assembly studies, resolve structural re-arrangements, epigenetic modifications like methylation, the study of splicing variants and SNP identification – the reason why it is used in this study (Trapnell *et al.*, 2010; Illumina, 2015).

In contrast, there is also single-end (SE) sequencing, where reads are analyzed from one end to the other, providing faster, cheaper and are enough for profiling or counting studies like RNA-Seq or ChIP-Seq (Trapnell *et al.*, 2010).

Currently, NGS has been divided in second generation sequencing and third generation sequencing. The basic features of the second-generation sequencing are the generation of millions of short reads in parallel, the highest speed, low cost and the fact that electrophoresis is unnecessary once the output is directly detected. The short read sequencing approaches are divided into sequencing by ligation (SBL) and sequencing by synthesis (SBS) and there are three major platforms: Roche/454 (since 2005), Illumina/Solexa (since 2006) and ABI/SOLiD (since 2007). The third-generation sequencing (TGS) have the ability to give an even lower cost than the previous technology and eases the sample preparation by removing the necessity of PCR amplification. At the same time, it can produce longer reads which is better for assembly (Kchouk *et al.*, 2017). There are two main approaches characterizing TGS which are SMRT, Single Molecule Real Time sequencing approach, and the synthetic approach which depends on pre-existing short reads to construct longer reads. Nowadays, the most commonly used TGS

is the SMRT approach used by Pacific Biosciences and Oxford Nanopore sequencing technologies (Kchouk *et al.*, 2017).

The sequencing technology used in this study was Illumina, the most widely used sequencing platform. Illumina sequencing uses SBS method and, in particular, Illumina's HiSeq X Ten technology, is currently the most used sequencing technology offering the highest throughput and the lowest cost in the NGS market (Kchouk *et al.*, 2017).

The process of sequencing done by Illumina consists of 4 main steps (Figure 1):

- First, DNA suffers a random fragmentation and adapters are connected to each end of every sequence. These adapters are fixed in a complementary way to other adapters placed in a solid plate;
- Next, each sequence attached to the solid plate is amplified with "PCR bridge amplification" originating clusters;
- Afterwards, using SBS method the four altered nucleotides (which are labeled with a fluorescent specific), sequencing primers and DNA polymerases are added to the mix allowing the proper hybridization to the sequence. Clusters are then excited by laser to emit a light signal and will be detected by a device camera and computer programs, translating these signals to a nucleotide sequence (Heo, 2015; Reuter *et al.*, 2015).

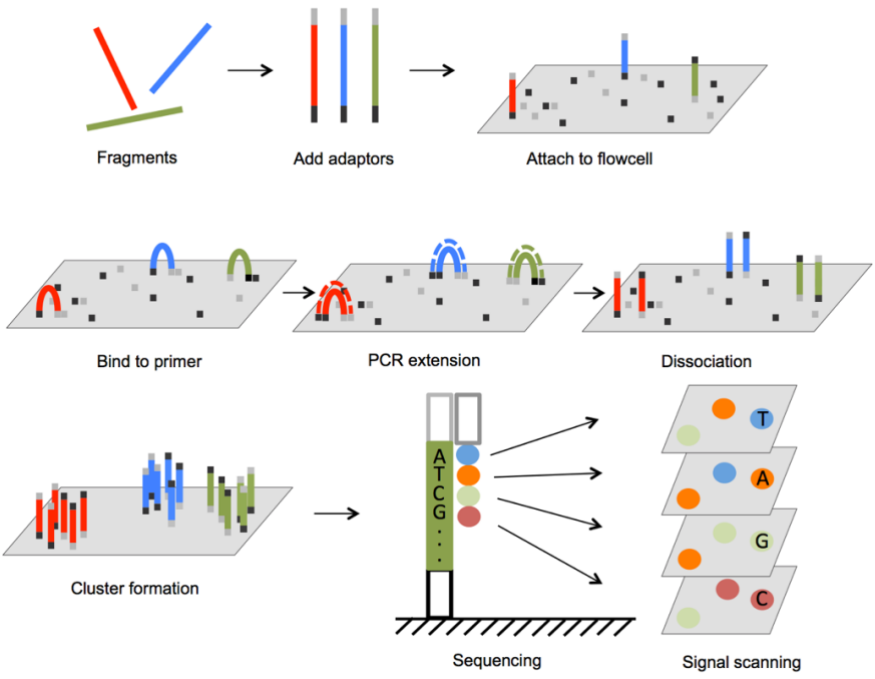

Figure 1 – Illumina sequencing process (Lu *et al.*, 2016).

Although Illumina sequencing helps the scientific community with the understanding of the relationships between genetic variation and phenotype, there is a downside associated with this technology. Illumina technology only sequences small DNA fragments and originates millions of small reads, turning assembly harder and demanding the use of high computing resources. Whole Genome Re-Sequencing (WGRS) at a population level has become attainable due to advances in throughput and cost reduction of sequencing technologies, as well as the progress in data management and development of bioinformatics tools for NGS data analysis due to the requirement of new algorithms for handling short reads, along with new programs for assembly, single nucleotide polymorphism (SNP) detection and other applications (Zhang *et al.*, 2011; Hatem *et al.*, 2013; Li *et al.*, 2013; van Dijk *et al.*, 2014).

1.2 *Quercus* genus

Comprising about 500 different species widespread across the globe, the genus *Quercus*, from *Fagaceae* family, is one of the most important woody plants existing (Sork *et al.*, 2016). In the Northern Hemisphere, these species are the most dominant plants based on features like species diversity, ecological dominance and economical value. Oaks can be found in a vast variety of habitats like temperate deciduous forest, temperate and subtropical evergreen forests, subtropical and tropical savannah to a variety of Mediterranean climate vegetations like chaparral, oak woodland and evergreen oak forests (Kappelle, 2006). In North and South America, there are 236 species (Chassé, 2016), in Europe 38 species and in Asia 156 (Gil-Pelegrín *et al.*, 2017). The economic importance of this genus is widely known. Various species are a source of different high-quality resources like hardwood used in house equipment, firewood and cork.

The cork oak species (*Quercus suber*) that can be found in the European southwest coastal region, Mediterranean Basin, has the unique feature of producing a natural, renewable and sustainable material which is its cork layer that comprehends the adequate properties for various industrial uses (Ramos *et al.*, 2018), like for example, wine closure. For this particular use, cork has the perfect features to protect wine qualities and to allow the development and improvement of its finest characteristics over time, especially for wines

that need to age in the bottle. These remarkable characteristics come from the impermeability of cork to liquids and gases (Silva *et al.*, 2005). The process of cork extraction happens for the first time when the tree is about 30 years old and harvest is practiced every 9-12 years until a fresh layer of bark has 30 mm of thickness. With that being said, through its life expectancy, these trees can be harvested over 16 times, maintaining the ability to regenerate its cork external layer as long as the vascular cambium remains stable (Oliveira and Costa, 2012; Kim *et al.*, 2017).

Alongside cork oak, the holm oak (*Quercus ilex*) is one of the four evergreen oak species growing in the Mediterranean area, where it is considered a fruit tree and has been selected for sweet acorn production for pig supplement food. These species comprise two main morphological types: *ilex* and *ilex rotundifolia.* The first happens to be distributed in Greece and France, being restricted to humid or sub-humid sites, and it's characterized by elongated and large leaf morph. On the other hand, the second type is only located in heartlands of Mediterranean areas as Portugal, Spain and North Africa, preferring semiarid to per-humid climates, having small and smooth-edged dense leaves (Lumaret *et al.*, 2002).

These two species of *Quercus* are the most abundant of evergreen broad-leaved trees characterizing the Mediterranean areas, in particular, areas called *dehesas* in Spain and *montados* in Portugal. Some studies report that these areas are threatened by the lack of regeneration due to excess of grazing and insolation (Soto *et al.*, 2007). The *montado* in the Alentejo region in Portugal is distinguished by its savanna-like appearance with changing densities of cork and holm oaks (Pinto-Correia *et al.*, 2011), or a mixture of both species coexisting with pasture and crops, constituting a seminatural ecosystem, a cultural landscape and a multifunctional system (Sá-Sousa, 2014). Resuming, this landscape plays an important role as a panacea system with ecosystem services like soil conservation, groundwater quality protection, carbon sequestration and with appropriate land-management practices it can provide food, fibers, cork, fuel, construction, livestock food, aromatic and medicinal plants, edible mushrooms among others (Sá-Sousa, 2014).

Despite their great importance, recent studies suggest that these ecosystems are facing different threads including the advanced age of individuals, overexploitation and deprived regeneration, incorrect livestock management, fungal attacks, extreme temperatures and climate changes and other factors (Surová *et al.*, 2018; Rey *et al.*, 2019). Science fields like biotechnology, bioinformatics and molecular genetics can play an important role

when trying to find solutions to solve some of these problems. However, the lack of knowledge on the biology of these trees at the molecular level, especially *Q. ilex*, may hamper this task.

1.3 Chloroplast

Chloroplast, among other plastids, are the most characteristic organelle within plant cells and living other eukaryotes algae (Wicke *et al.*, 2011). They have the unique characteristic of converting sunlight into chemical energy – photosynthesis and oxygen release – and carbon fixation, being an active metabolic center that sustain life on earth as we know it (Daniell *et al.*, 2016; Kersten *et al.*, 2016). It is conceptual nowadays that these organelles were incorporated into eukaryotic cells, during millions of years of species evolution, through primitive endosymbiosis of an autotrophic procaryotic cell, enabling the transition from heterotrophy to autotrophy, conceiving to these cells the capacity of utilizing photoenergy (Wicke *et al.*, 2011) and justifying the presence of its own genome (Greiner and Bock, 2013; Kersten *et al.*, 2016).

Although the key function of chloroplasts is the photosynthesis, they also are extremely important on the development and physiology of the plant once they can synthetize important metabolites. Those metabolites help the plant on the adaptation and interaction with the environment and defense against pathogens (Daniell *et al.*, 2016), being crucial the study of its genome (cpDNA). The plastome carries genes encoding proteins for genetic apparatus, like structural and transfer RNAs, and proteins for non-photosynthesis pathways, light-dependent (primary) reactions and light-independent (secondary) photosynthesis pathways (Wicke *et al.*, 2011).

Chloroplast genome is known to be highly conserved in gene order and gene content, being in the majority of plants, a circular genome whose size can range between 100 and 170 kb. Also, cpDNA exhibits a much lower substitution rate than nuclear DNA (Carbonell-Caballero *et al.*, 2015). For these reasons, is greatly suitable for phylogenetics studies of different plant families allowing researchers to isolate homologous loci for comparative studies over different evolution times (Atherton *et al.*, 2010; Carbonell-Caballero *et al.*, 2015; Daniell *et al.*, 2016; Sun *et al.*, 2019).

Additionally, some researchers point that some chloroplast genome sequences may have suffered some variation within and between plant species either in sequence or structural variation (Wambugu *et al.*, 2015; Brozynska *et al.*, 2016). For this reason, cpDNA can give relevant information on the adaptation of species to climate changes and help the scientific community to understand and select the traits conferring climatic adaptation in highly economical valuable plants and trees (Daniell *et al.*, 2016).

1.4 Mitochondria

Similarly, to chloroplasts, mitochondria is an organelle that has its own genome (mtDNA) and it is also derived from existing bacteria and through the process of endosymbiosis became part of the eukaryotic cell, but not only in plants. This organelle is maternally inherited, and it is responsible for the ATP production process through oxidative phosphorylation. Some plants (seed plants and land plants) have mitogenomes that are notably different from other eukaryotic species, having a mixture of both fast and slow rates of evolution (Chen *et al.*, 2017).

In plants, the mitogenome is distinguished by their size variation, structure and organization, in contrast with animal mitogenomes which are very conserved. This variation in its size is even observed at the level of the same genus, serving as an example *Silene latifolia* with 253kb genome size and *Silene conica* with 11Mb (Sloan *et al.*, 2010; Chen *et al.*, 2017). In seed plants, this genome comprises some very interesting traits like slow mutation rate, high RNA editing and *trans*-splicing of coding sequences, frequent uptake of foreign DNA both by intracellular and horizontal transfer, dynamic structure and large and variable sizes (Alverson *et al.*, 2010; Chen *et al.*, 2017).

When compared with other eukaryotes, plants have a larger and more complex mtDNA and all the features assigned to it include RNA editing, recombination, trans-splicing and external DNA insertion. Although its extraordinary variation in size, structure and sequence content, some essential genes are well preserved in these genomes such as NADH dehydrogenase, succinate dehydrogenase, ubichinol cytochrome c reductase, cytochrome c oxidase and ATP synthase (Ogihara *et al.*, 2005; Zhang *et al.*, 2011). As mentioned, although there's a high mtDNA variation size of between species, the gene

content is kept around the same quantity and one major characteristic of plant mitogenomes is the abundance of repeated sequences, either derived from parental transfer or horizontal transfer of exogenous sequences from the chloroplast, nuclear or viral genomes (Gualberto and Newton, 2017).

The mitochondrial DNA in plants vary from its animal equivalent mainly by the considerable amount of non-coding regions and dynamic genome structure induced by the recombination of repeated-mediated homologous. As a prime component of non-coding regions of angiosperm mitogenomes, repetitive sequences play an essential role in maintaining and shaping structure as they participate in rearrangements, recombination, duplications, insertions and deletions (Shi *et al.*, 2018).

The present study aims to investigate the chloroplast and mitochondria genomes of both species, in order to analyze and compare their genetic variation, in terms of single nucleotide polymorphisms (SNPs), among individuals within the same species and, also, between the two different species. The investigation is carried by the bioinformatics analysis of WGRS data derived from 47 samples collected in Portugal, the majority in Alentejo region.

Chapter 2: Objectives

In the present study, the goal is to identify and annotate single nucleotide polymorphisms (SNP) in the cork oak and holm oak organelle genomes. In more detail, the objectives include:

1. To review the available literature about the species, organellar genomes and genetics variations, as well as the analysis of suited bioinformatics tools for this pipeline;
2. To detect the extent of variation in *Quercus* chloroplast genomes;
3. To determine the amount of variation in *Quercus* mitochondrial genomes;
4. To proceed to the annotation and characterization of the variation found.

Chapter 3: Materials and Methods

This section will describe the pipeline applied in this study. The methods for SNP calling have a standard base guideline and the following Figure 2 summarizes the procedures.

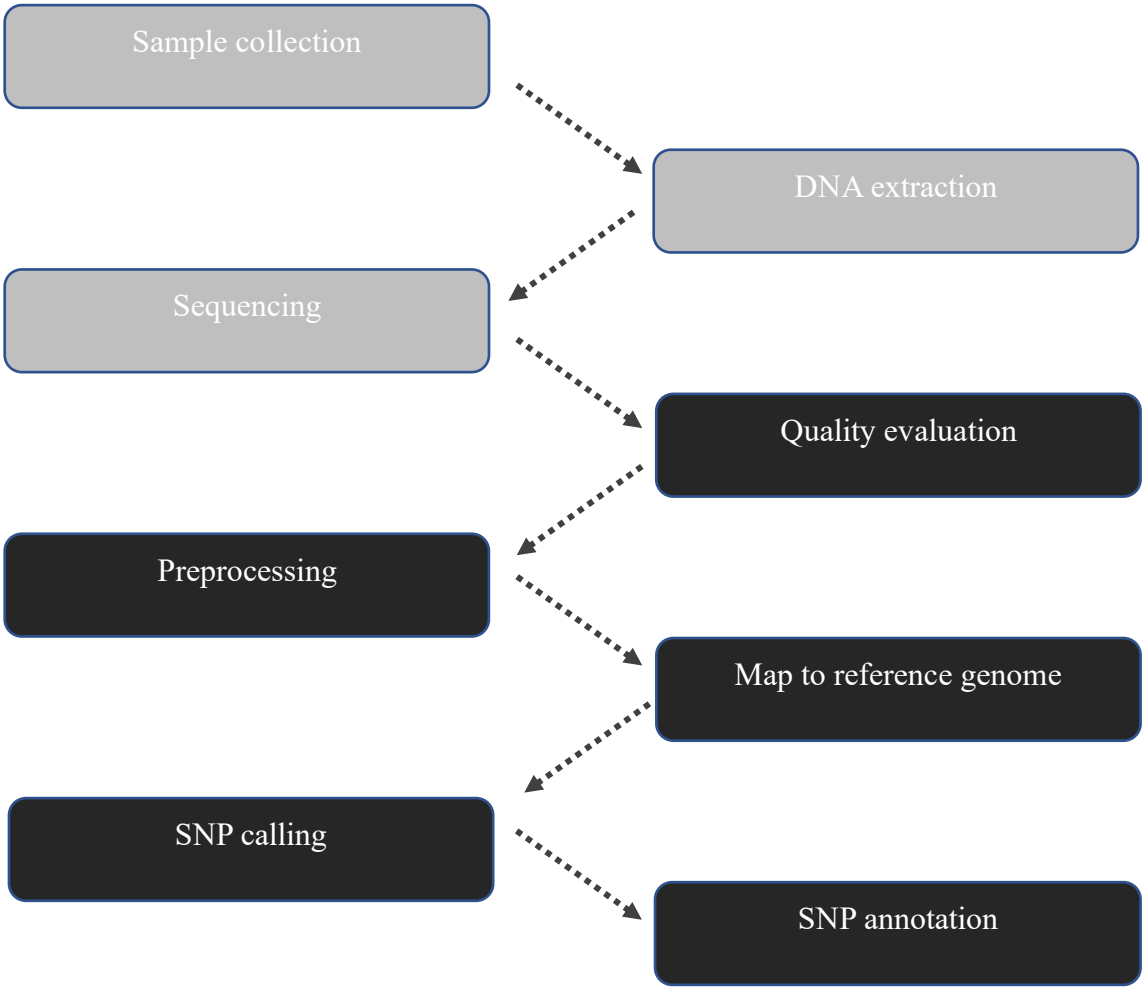

Figure 2 - Thesis pipeline with ordered steps performed. Grey background steps were performed by others while black background steps were performed by the author of this thesis.

3.1 Sample collection and high-throughput sequencing

In this thesis a total of 47 trees from two *Quercus* species were analyzed: 39 *Quercus suber* trees and 8 *Quercus ilex rotundifolia* trees. The information of where these trees are located is summarize in Table 1. The total DNA was extracted from leaf samples of each tree using a standard DNA extraction protocol. The 47 samples were divided in two datasets. The first dataset is composed by 17 samples (9 cork oaks, 8 holm oaks) while

the second dataset contained the remaining 30 samples, all cork oak trees. The samples from the second dataset were used in a previous study on identification and characterization of structural variation in *Quercus suber* species associated with cork quality (Magalhães, 2017).

Table 1 - Datasets information: Dataset identification, Sample name, species and localization of the sample.

Dataset	Sample Name	Species	Localization
1st	AB04	*Quercus ilex rotundifolia*	Abóbada, Abóbada
	AZ01	*Quercus ilex rotundifolia*	Azinhal
	CL1	*Quercus suber*	Companhia das Lezírias
	CL3	*Quercus suber*	Companhia das Lezírias
	HL11	*Quercus suber*	Herdade dos Leitões
	HL12	*Quercus suber*	Herdade dos Leitões
	HL14	*Quercus suber*	Herdade dos Leitões
	ISA1	*Quercus suber*	Instituto Superior de Agronomia, Lisboa
	L2	*Quercus suber*	Loulé
	L3	*Quercus suber*	Loulé
	MN03	*Quercus ilex rotundifolia*	Abóbada, Monte Novo
	MN04R	*Quercus ilex rotundifolia*	Abóbada, Monte Novo
	Q32	*Quercus ilex rotundifolia*	Quinta do Marquês
	QS32	*Quercus suber*	Quinta da Serra
	SN7	*Quercus ilex rotundifolia*	Parque Florestal de Monsanto
	VF03	*Quercus ilex rotundifolia*	Abóbada, Vale Formoso
	VF05	*Quercus ilex rotundifolia*	Abóbada, Vale Formoso
2nd	Ind113	*Quercus suber*	Herdade dos Leitões
	Ind115	*Quercus suber*	Herdade dos Leitões
	Ind118	*Quercus suber*	Herdade dos Leitões
	Ind120	*Quercus suber*	Herdade dos Leitões
	Ind19	*Quercus suber*	Herdade dos Leitões
	Ind19A	*Quercus suber*	Herdade dos Leitões
	Ind24	*Quercus suber*	Herdade dos Leitões
	Ind29	*Quercus suber*	Herdade dos Leitões
	Ind48	*Quercus suber*	Herdade dos Leitões
	Ind6	*Quercus suber*	Herdade dos Leitões
	Ind66	*Quercus suber*	Herdade dos Leitões
	Ind7	*Quercus suber*	Herdade dos Leitões
	Ind74	*Quercus suber*	Herdade dos Leitões
	Ind75	*Quercus suber*	Herdade dos Leitões
	Ind76	*Quercus suber*	Herdade dos Leitões
	Ind9	*Quercus suber*	Herdade dos Leitões

	Ind98	*Quercus suber*	Herdade dos Leitões
	IndHL12	*Quercus suber*	Herdade dos Leitões
	IndHL15	*Quercus suber*	Herdade dos Leitões
	IndHL16	*Quercus suber*	Herdade dos Leitões
	IndHL17	*Quercus suber*	Herdade dos Leitões
	IndHL18	*Quercus suber*	Herdade dos Leitões
	IndHL19	*Quercus suber*	Herdade dos Leitões
	IndHL20	*Quercus suber*	Herdade dos Leitões
	IndHL21	*Quercus suber*	Herdade dos Leitões
	IndHL3	*Quercus suber*	Herdade dos Leitões
	IndHL4	*Quercus suber*	Herdade dos Leitões
	IndHL5	*Quercus suber*	Herdade dos Leitões
	IndHL7	*Quercus suber*	Herdade dos Leitões
	IndHL9	*Quercus suber*	Herdade dos Leitões

In the first dataset, high-throughput sequencing was performed in the sequencing platform of Beijing Genomics Institute (BGI-SEQ500) using paired-end protocol, read length of 100bp and insert size of 300bp. The remaining 30 samples comprehending the second dataset, were sequenced using high-throughput sequencing in the Illumina HiSeq X Ten platform, using paired-end protocol, read length of 150bp and insert size of 300bp.

3.2 High-throughput sequence data quality evaluation

After verifying the integrity of the data received, all the WGRS reads were subjected to a quality control procedure using FastQC software, version 0.11.5 (Andrews, 2010). This software was conceived in 2010 by Simon Andrews at Babraham Institute with purpose to evaluate and control raw next generation sequencing data quality. The operating mode consists of a simple command line code and the output gives the user a report for each sample containing valuable information like basic statistics, per base sequence quality, per sequence quality score, per base sequence content, per sequence GC content, per base N content, sequence duplication levels, overrepresented sequences and adapter content, among others. Figure 3 shows the distribution of the quality scores along the read base-pairs (per base sequence quality) of one of the analyzed samples.

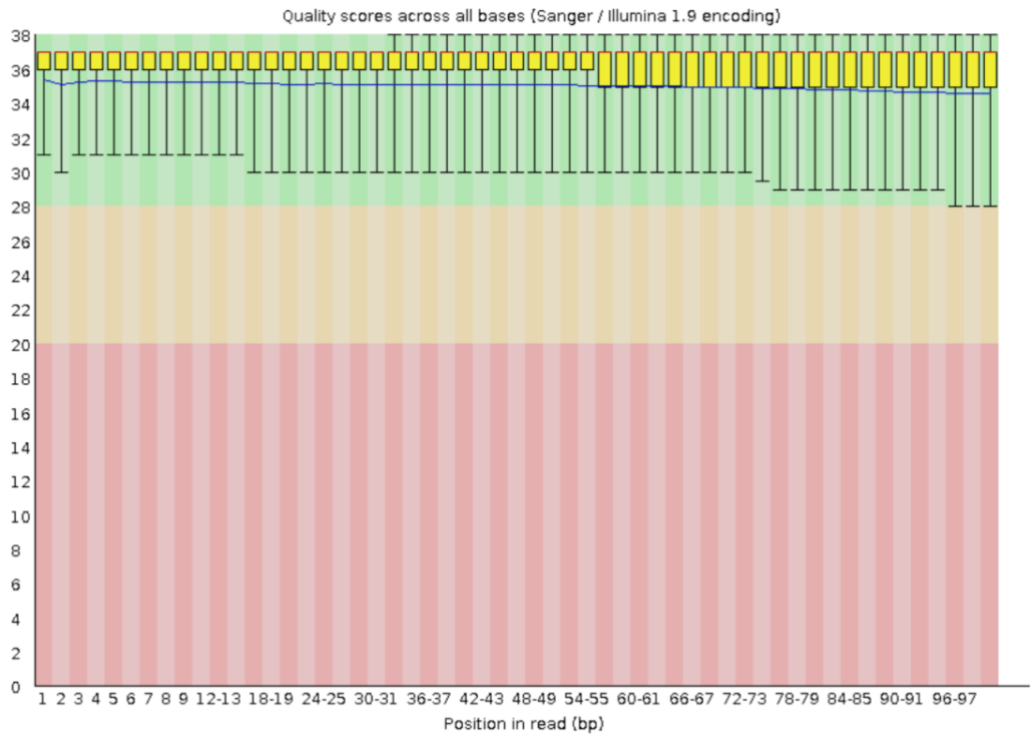

Figure 3 – Read quality evaluation using FastQC "Per base sequencing quality" graph.

When examining this graph, it is possible to identify at which position of the read sequence the quality score is below a certain threshold (defined by the user). These thresholds (read length and quality score) need to be defined in order to perform the preprocessing of the raw data.

Taking that into account, several combinations of length and quality thresholds were tested in order to identify which combination was more adequate to reject bad quality reads but also maintain a high number of reads. The values used for the quality threshold were 15, 20, 30 and 35, while the values for the read length were 80 % and 90 % of total read length.

3.3 Preprocessing of high-throughput sequence data

After quality checking and determination of the length and quality thresholds to be tested in the previous steps, it was used Trimmomatic software version 0.38 (Bolger *et al.*, 2014) for trimming the reads.

With most Illumina sequencing platforms, the quality of the reads will drop at the end of the fragment due to signal decay or phasing during the sequencing run. For that reason, Trimmomatic uses a sliding window approach, that starts scanning at the 5' end and clips the read once the average quality within the window falls below the threshold, trimming the end of the read. This decay of quality in reads can also be observed in Figure 3.

Trimmomatic was run defining three parameters: slidingwindow; minlen and quality. One of the operating modes of Trimmomatic is through the setting of a sliding window which acts through a certain percentage defined by the user of the read length window size for trimming low quality reads once the average quality value within that window falls below the threshold and that is where the algorithm will recognize the quality decline and the read is cut (3'-end). The remaining parameter (minlen) defines the minimum length that the read must have to be kept, that is, if the remaining sequence after trimming for quality is smaller than the minimum length defined (minlen), the read is rejected completely.

For further analysis, the preprocessing step has a major importance since it guarantees higher quality and length dataset of reads. In order to know which set of parameters values were the best for this study, a set of testes were experienced over a smaller set of data, equally representative. These tests are described in detail at appendixes and based on the results of those tests the chosen parameter values where: quality (Q) 20, minlen (L) 80 % and slidingwindow of 10 %.

3.4 Read mapping to the chloroplast and mitochondrion genomes

After the preceding phase has been completed, the data is then ready for alignment to the respective reference genomes. To do so, the reads were mapped using BWA-MEM (Li,

2013). BWA is a software for mapping sequences against a reference genome and consists of tree algorithms: BWA-backtrack, BWA-MEM, and BWA-SW. Based on literature, the BWA-backtrack is suitable for shorter reads (reads up to 100 bp), whereas for longer Illumina reads (from 70 bp to 1 Mbp), the BWA-MEM and BWA-SW are preferred. However, BWA-MEM is usually the ideal algorithm to use when high-quality queries are available as it is faster and more accurate, and for those reasons, it is the chosen algorithm.

BWA-MEM (Burrows-Wheeler Alignment of maximal exact matches) (Li, 2013) is an alignment algorithm whose purpose is to align sequence reads or assembly contigs against a large reference genome, such as the cork oak genome. In general, it is designed for Illumina sequence reads up to 100 bp and it aligns the sequences reads or long query sequences against the reference genomes automatically choosing between local and end-to-end alignment. BWA-MEM performs local alignment and may produce multiple primary alignments for different part of a query sequence. This algorithm implements two distinct read mapping methods, one for single ended reads (SE) and another for paired end reads (PE).

The chloroplast and mitochondrial genomes of *Quercus suber* were used as genome reference.

The cork oak chloroplast genome is represented in single circular molecule of 161 179 bp with a GC content of 36.8 % and contains a total of 137 genes annotated: 86 protein coding genes, 40 tRNA and 8 rRNA genes. The mitochondrial genome used, this is comprised by 3 average size contigs, one large contig (442 094 bp) and two smaller ones (52 064 bp and 37 700 bp) with a total genome size of 531 858 bp. A total of 69 genes were annotated, being 41 protein coding genes, 23 tRNA and 5 rRNA.

Initially, it was necessary to construct an index for the reference genome before mapping the reads. This index is usually constructed as a hash table for effective querying and is given as input for BWA-MEM along with the forty-seven samples. Once the reads were mapped against the reference genomes, the software outputs a Sequence Alignment/Map (SAM) file, which were then converted into BAM format (Binary Alignment/Map) the binary version of SAM files, since it is more efficient and less computational demanding to work with, saving disk space as well. Right after, the files were sorted by genomic coordinate where the alignments occur in "genome order", that is, ordered positionally based upon their alignment coordinates on each chromosome. This is done in order to facilitate the access of the data in a more efficiently way by other tools. Both file format

conversion and sorting were done using SAMtools v1.4.1 (Li *et al.*, 2009), as well as all the manipulation, analysis and retrieving of mapping statistics, using SAM bitwise flags. These flags allowed to count specific types of mapping results such as mapped/unmapped reads, mapped and paired reads, proper pair alignments and unique mapped reads.

Moreover, for variant calling the data need to contain specific information about read groups (RG), which are identified in the header of SAM/BAM files by different tags described in the official SAM specifications (Li *et al.*, 2009). The appropriate assignation of these tags can allow the differentiation of not only the samples, but also technical features associated with the technology. The most relevant tags for variant calling are: read group library (RGLB), read group platform (RGPL), read group platform unit (RGPU), read group sample name (RGSM) and read group ID (RGID). Reads with the same read group are a set of reads generated from a single run of sequencing. In a single library preparation from one biological sample which is run on a single lane of a flow cell, all these reads (from the same lane) belong to the same read group. In order to add this information, the command used is "AddOrReplaceReadGroups" of Piccard software ("Picard Toolkit." 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute).

3.5 Single nucleotide polymorphisms (SNPs) calling

In order to choose the best software for our data to perform the variant calling, two different software were tested: Freebayes (Garrison and Marth, 2012) and SAMtools Mpileup (Li *et al.*, 2009) using a subset of the data. See appendixes for details about the testing. At the end of the testing procedure, the Freebayes software was selected to perform the variant calling. The parameters used were those defined by default by the software, only taking into account the genome ploidy being this the only parameter to be defined.

SAMtool Mpileup works in two steps: it first collects information in the input BAM files, computes the likelihood of data given each possible genotype and stores the likelihoods in the BCF format - but it does not call variants. Instead, in a second step it requires

BCFtools commands, which will make the actual calling and converting the BCF file to VCF output.

On the other hand, Freebayes is able to perform variant calling alone. It is a Bayesian and haplotype-based genetic variant detector designed to find short polymorphisms like SNPs, indels, MNPs and other events. This software is haplotype-based because it reads short haplotypes from sequencing traces, i.e., it calls variants based on the literal sequences of reads aligned to a particular target, not a precise alignment. In this way, Freebayes offers benefits over other methods that operate on a single position at a time (Garrison and Marth, 2012). This software uses short-read alignments (BAM file) for any number of individuals from a population and a reference genome (in FASTA format) to determine a most-likely combination of genotypes for the population at each position in the reference. Finally, Freebayes reports polymorphic positions in variant call format file (VCF file).

After variant calling, a set of raw variants is obtained. Thus, in order to obtain the significant SNPs those variants were filtered using VCFtools v0.1.17 (Danecek *et al.*, 2011), indicating to remove indels, keep only bi-allelic sites (sites containing only two observed alleles, counting the reference allele as one allowing for one variant allele only), minimum depth coverage per sample of 15 and a SNP quality equal or above 30.

Once significant SNPs were obtained, they were grouped by species (*Q. suber* and *Q. ilex rotundifolia*) using a custom python script. The script output provides general information about position in the genome for each significant SNP, number of samples with valid genotype information, and number of samples per each represented genotype (0: representing the reference allele (REF field in the VCF file) or 1: representing the alternative allele (allele listed in ALT field in the VCF file)). Additionally, the same information is provided for each defined group.

3.6 SNP annotation

The SNP annotation was performed using the software ANNOVAR (version 2018Apr16) (Wang *et al.*, 2010). This program is a well stablished annotation software developed by

Kai Wang and colleagues (2010) and, among several features, it annotates functional effects of variants regarding genes, performs genomic region-based annotations and can compare variants to existing variation databases as long as they follow the standards for sequence-level feature annotation of Genetic Feature Format version 3 (GFF3). This format became a well-known standard format between several databases of model organisms and offers a suitable mode to exchange features of sequence annotation. By using this format, ANNOVAR is able to question any annotation database, making it an unquestionable software to use (Wang *et al.*, 2010).

ANNOVAR is a command-line tool which takes a text-based input file, as the commonly used VCF file, and outputs an annotated variant file in containing annotations for each variant in the input file. Each line of the input file represents a single SNP and must have at least 8 tab-delimited columns representing: chromosome name, start position, end position, reference nucleotide, observed nucleotide, zygosity status/allele frequency, genotype quality and read depth. For this reason, firstly it is necessary to convert the VCF file output from the previous variant calling software and convert it to "avinput format" – ANNOVAR input text file. The tab-delimited output file often contains several lines (each one for each variant) combining the information from the input file with additional annotation information like: genomic function, the affected gene and transcript, functional role of the coding variant, the transcript nucleotide change and, lastly, the protein amino-acid change (Yang and Wang, 2015).

A FASTA transcript sequence file from the original genome sequences and a proper genome database for the species are required to use ANNOVAR. Then it is possible to proceed to the variant's annotation using the program's perl script "annotate_variation.pl". Lastly, a python script was used to analyze the annotation output which is given in three possible ways referring to "genic", "intergenic" or "no annotation" types of SNPs. The difference between the first two types is that 'intergenic' variants are placed around genes and 'genic' variants are within the gene region. The first type can be placed in front of the gene (upstream intergenic) or after the gene (downstream intergenic), whereupon both upstream and downstream intergenic regions are related to regulatory functions of the gene, holding the core promoter and others elements upstream and less known regulatory elements downstream (Colinas *et al.*, 2008), as it is schematized in Figure 4.

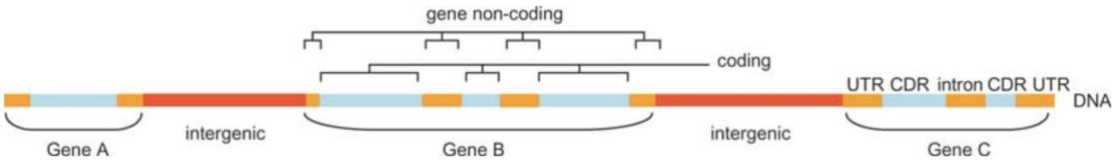

Figure 4 - Illustrative scheme of a protein coding gene in a DNA sequence (Colinas *et al.*, 2008).

Chapter 4: Results and Discussion

Chloroplast and mitochondrial genomes are an important source of information for studies and applications on genomics and biotechnology (Pinard *et al.*, 2019). It is unquestionable that these genomes play an important role in adaptation to environments and consequently evolution of plants, although it is not clear how much the organelle variants will affect the plants phenotype (Budar and Roux, 2011). *Quercus suber* and *Quercus ilex* are important tree species in Portugal for their socio-economic impact, and so, they are important organisms to study.

The following section describes and discusses the obtained results in the data preprocessing, mapping against the reference genomes, variant calling and SNP annotation including SNP characterization and selection.

4.1 Data preprocessing

When working with datasets of high-throughput sequences, to guarantee that we work with high quality data, the first and essential step is to properly filter the datasets removing low quality sequences. Therefore, preprocessing aims to filter and remove read sequences with low quality and small length, based on defined thresholds.

The results regarding the reads preprocessing were obtained after testing of different parameters values whose selection followed the extensive analysis of the per base sequence quality section of the FastQC reports for each sample (See appendixes for more details). The software used for this task was Trimmomatic (Bolger *et al.*, 2014), a well-known software for trimming bad quality reads. The values tested were 80 % and 90 % for the length parameter, combined with 15, 20, 30 and 35 for quality values. The full testing phase is well described in appendix section, and, at the end, the best set of parameters was: 80 for minlen (80 % of the total read length) and 20 for quality, with a sliding window of 10 % of the read length.

Given the high data quality and the combination of these parameters, which are customized to achieve high quality datasets, an average of 96 % of the reads in each sample of the first dataset and 88 % of the second dataset outlasted trimming. Due to the high quality that Illumina Platform technology offers, the percentage of reads saved is very high. Table 2 and appendixes table 10 show the preprocessing statistics.

Table 2 - Summary of the preprocessing statistics per dataset.

Dataset	Samples number	Average number of raw reads per sample	Average number of reads per sample after trimming
1st	17	287 525 393	275 091 686 (96 %)
2nd	30	170 884 695	151 196 809 (88 %)

4.2 Mapping against the reference genomes

After preprocessing, the reads from both species and datasets were mapped against the *Quercus suber* organelle reference genomes, once no organelle genome for *Quercus ilex* is available. The used software for this task was BWA-MEM (Burrows-Wheeler Alignment of maximal exact matches) (Li, 2013), which was recently proven by Yao and his colleagues (2020) to be more efficient when compared to other mapping software like Bowtie2 among others. In his study, the author stated that BWA-MEM mapped a higher number of reads and it also had a higher mapping rate of properly mapped PE read, i.e., both pairs of the same read mapped correctly in the correct insert distance and correct directions (Yao *et al.*, 2020). In our case study, no comparison was made at this stage of the protocol because BWA-MEM is a well-known standard software and the default parameters are well suited for the objective of this study, the identification of SNPs. Then, SAMtools (Sequence Alignment/Map tools) (Li *et al.*, 2009) was used to include the header to the alignment file, sort it by genomic position and make the binary conversion for the final output.

Table 3 contains the mapping statistics which for each dataset provides information about reads mapped, reads unmapped, reads mapped and paired (both forward and reverse reads are mapped), reads properly paired (both forward and reverse reads are mapped with the proper insert size distance between them) and, lastly, reads properly paired and unique (reads that mapped properly in only one location of the genome, i.e., in just only one loci).

Table 3 - Mapping statistics of chroloplast and mitochondrial genomes.

	Chloroplast				
	Mapped	**Unmapped**	**Mapped and paired**	**Properly paired**	**Properly paired and unique**
1st dataset	22 849 670 (8.29 %)	252 242 016 (91.71 %)	22 601 851 (8.20 %)	22 542 985 (8.18 %)	15 413 868 (5.59 %)
2nd dataset	7 219 073 (4.76 %)	143 977 737 (95.24 %)	7 149 919 (4.71 %)	7 126 401 (4.70 %)	4 798 836 (3.16 %)
Total	**30 068 743 (7.05 %)**	**396 219 753 (92.95 %)**	**29 751 770 (6.98 %)**	**29 669 386 (6.96 %)**	**20 212 704 (4.74 %)**
	Mitochondria				
	Mapped	**Unmapped**	**Mapped and paired**	**Properly paired**	**Properly paired and unique**
1st dataset	12 302 432 (4.49 %)	262 789 254 (95.51 %)	11 785 506 (4.30 %)	11 726 356 (4.28 %)	9 608 271 (3.50 %)
2nd dataset	2 728 235 (1.81 %)	148 468 575 (98.19 %)	2 557 937 (1.69 %)	2 545 241 (1.69 %)	2 081 003 (1.38 %)
Total	**15 030 667 (3.53 %)**	**411 257 828 (96.47 %)**	**14 343 442 (3.36 %)**	**14 271 597 (3.35 %)**	**11 689 275 (2.74 %)**

The results showed a low percentage of mapped reads in both organelles (7.05 % and 3.53 %, chloroplast and mitochondrion genomes, respectively) which was expected due to the DNA extraction protocol used. The total DNA obtained per each sample contained nuclear DNA, being the most abundant in the sample, and chloroplast and mitochondrial DNA. Additionally, it is also expectable to obtain a higher percentage of reads mapped against the chloroplast than against the mitochondrion genome, because the chloroplast is an organelle found in greater number in the cell compared to mitochondria (Scarcelli, 2020).

At the end, only the reads mapped properly paired and uniquely are kept for variant calling. The reason to discard multiple alignment reads for downstream analysis like SNP

calling, is that using them makes it harder to know which location carries the polymorphism, inducing bias and ignoring real genomic regions that may be biologically important.

4.3 Variant calling

In order to perform the variant calling in both organelles, the reads kept in the previous step (reads mapped properly paired and uniquely) were used as input for the variant calling tool. Two software were tested under the same parameters and data. Freebayes (Hwang *et al.*, 2015) was compared to SAMtools Mpileup, as it was described in appendixes, and in our case, it came up with better results. This software identified more SNPs under the same parameters and the SNP genotypes were represented for a larger number of samples. This fact is in accordance with other studies comparing different variant calling software, for instance, Hwang and colleagues (2015) work. In their study were compared different pipelines for variant calling combining different sequencing data types (from Illumina and Ion Proton platforms) and variant calling software, and in general they found that the pipelines with Freebayes show a higher performance for any type of data. By filtering the variants by their quality scores, i.e., rejecting false positives with low scores, the authors also state that Freebayes can output more true positive variants, suggesting this software for studies where only high-quality variants are considered. In this study's case, the performed filtering applied in the analysis of both organelles, removed variants that did not meet the requirements of minimum deep coverage per sample of 15 and minimum SNP quality of 30.

4.3.1 Chloroplast genome

The total number of raw variants identified in the chloroplast genomes of the 47 *Quercus* trees included in the WGRS dataset 1 and 2 was 1 070: 804 were SNPs, 69 MNPs (multi-nucleotide polymorphisms), 198 indels, 108 multiallelic sites and 15 multiallelic SNP

sites (Table 4). Note that the reason why the sum of the several variants types doesn't match the total number of raw variants, is because multi-allelic sites may contain both SNPs and indels, for instance, contributing a count for both cases.

After filtering for a minimum deep coverage of 15 per sample and a minimum SNP quality of 30 and removing non-biallelic SNPs and indels a total of 607 SNPs remained (Table 4).

When performing a more detailed analysis on the two species under study, the exclusive SNPs associated to each species were determined. A SNP is considered exclusive when it is represented by at least 80 % of the individuals of the species of interest and at most the 20 % of the individuals of the other species. Following this criterion, a total of 504 SNPs were found to be exclusive for the holm oak species while no exclusive SNPs were found for cork oak. Additionally, only one SNP was represented for both species simultaneously. These results clearly indicate that there is an extremely low level of variation in the cork oak chloroplast genome sequence. However, there is a clear identification of a high chloroplast variation between the cork oak and holm oak species.

Table 4 - Summary numbers of variants found in chloroplast genomes.

Chloroplast		
Variant Types	**Raw Variants**	**Filtered Variants**
number of samples:	47	47
number of records:	1070	607
number of SNPs	**804**	**607**
number of MNPs:	69	0
number of indels:	198	0
number of multiallelic sites:	108	0
number of multiallelic SNP sites:	15	0
number of others:	56	0

4.3.2 Mitochondrial genome

The number of raw variants found in mitochondria was 19 870, which was much higher than in the chloroplast. This difference is expected due to the difference in genome sizes, being the mitochondrion larger and less conserved among species than the chloroplast. These variants were composed by 17 596 SNPs, 3 248 MNPs, 785 indels, 2 455 multi-allelic sites, 331 multi-allelic SNP sites and 299 others (Table 5). After applying the same filtering criteria defined for the variants identified in the chloroplast genome, a total of 935 SNPs remained.

The number of SNPs found represented in both species (more than 80 % of individuals within the species present the variation) is 60, while 48 of those were represented by all the individuals within both species. Regarding the exclusive SNPs, 6 were found in cork oak and 471 in holm oak.

These results highlight the existence of variation between cork oak and holm oak mitochondrial genomes.

Table 5 - Summary numbers of variants found in mitochondrial genomes

Mitochondria		
Variant Types	**Raw Variants**	**Filtered Variants**
number of samples:	47	47
number of records:	19870	935
number of SNPs	**17596**	**935**
number of MNPs:	3248	0
number of indels:	785	0
number of multiallelic sites:	2455	0
number of multiallelic SNP sites:	331	0
number of others:	299	0

4.4 SNP annotation and characterization

The identified SNPs in both organelles were then annotated with ANNOVAR and divided by type of annotation (annotated or non-annotated). Then, the annotated SNPs were divided by loci: genic and intergenic.

4.4.1 Chloroplast genome SNPs

All the 607 valid SNPs identified in the chloroplast genome were annotated as occurring in genic regions (exonic annotation location).

The identified SNPs across all the individuals of both species occurred in an exonic region of the gene *rpl2* (encoding ribosomal protein L2) and the annotation of the implications of this variation is reported as unknown, being unclear if this alteration leads to changes in the amino acid. This gene is highly conserved, and it has a specific ribosomal promoter in angiosperms. Its functions may vary between, per example, the increasing of beta-galactosidase expression and the regulation of transcription in the chloroplast. The nonappearance of the protein encoded by this gene is a clear sign of ribosomal malfunction, once it is involved in the ribosomal enzymatic function (Njuguna *et al.*, 2019). Also, it is believed that *rpl2* encodes 50S ribosomal subunit components, containing only a group II intron. In a 2020 study, Zhang and colleagues (Zhang *et al.*, 2020) compared wild type rice and white stripe leaf (*wsl*) mutants, they report that the mutants reveal a reduced germination rate and lower shoot and root growth when treated with abscisic acid (ABA) but not with other components like gibberellic acid (GA), cytokinins and auxins. This discover indicate that ABA signaling process is specifically affected in *wsl* mutants suggesting that *rpl2* splicing will affect plants response to ABA (Zhang *et al.*, 2020), a hormone highly associated with plants' stress response (Chen *et al.*, 2020).

As mentioned before, no exclusive SNPs were found for cork oak while 502 were found for holm oak. Within this set of exclusive SNPs, 71 (about 14 %) have synonymous exonic alteration, 51 (about 10 %) are non-synonymous, i.e., alterations that can lead to

possible amino-acid changes, and the remaining 381 (about 76 %) have unknown exonic alterations. Due to its higher relevance, the focus of this analysis is on non-synonymous variants found in potential protein coding regions, although synonymous SNPs are also important for background mutation rate estimation in genomes.

The most annotated non-synonymous genes found were ribosomal protein genes, namely, *rpl2,* described above, and *rpl12.* This last gene encodes L12 proteins, which belongs to the ribosomal machinery components and are the only multicopy ribosomal protein involved in protein synthesis regulation (Nagaraj *et al.*, 2016). Besides this function, these proteins are also reported to be implied in stress signaling in a Nagaraj study in 2016. To validate that, the authors used *rpl12* gene silenced *Nicotiana benthamiana* plants (wild tobacco) and observed a late initiation of hypersensitive response to nonhost pathogens and it was established the hypothesis that the role of *rpl12* in plants defense could be their activation and involvement in particular protein synthesis for plant defence. Furthermore, they also affirm that *rpl12* plays a role in nonhost resistance besides the basal defense response in Arabidopsis (Nagaraj *et al.*, 2016).

In addition to the basic annotated genes report, it was estimated an additional variants information, the minor allele frequency (MAF). This parameter allows the restriction from all the variants set, the ones with low population frequency, presuming that common SNPs have fewer probabilities to cause severe alterations on individuals (Bao *et al.*, 2014). Thereby, when restricting the results to 25 % or less MAF value, the number of annotated variants falls to 75 exonic SNPs of which 20 are synonymous, 8 are nonsynonymous and 47 are unknown.

When focusing on the nonsynonymous single nucleotide variations filtered by the previous parameter, they all belong to holm oak affecting the following genes: *rps12, rps2, rpl2, rpl33, rpoB, psbZ, psbH, accD, ndhH* and *ndhE*.

The majority is attributed to ribosomal protein coding genes (*rps and rpl* genes). Also, the *rpoB* gene is known to play a role in chloroplast gene transcription. That is, the process of chloroplast gene transcription is naturally conducted by two types of RNA polymerases, nucleus encoded RNA polymerase (NEP) and plastid encoded RNA polymerase (PEP); this last type is the principal transcriptional machinery, composed by four core subunits and a promoter-recognizing subunit, and *rpoB* is one of the four genes (along with *rpoA, rpoC1* and *rpoC2*) encoding the core subunits of PEP (Zimmermann *et al.*, 2019; Zhang *et al.*, 2020). Chloroplast transcriptional regulation is crucial for the

overall well function of the chloroplast and the entire plant health under either normal or adverse conditions, and so, a possible malign variation in this gene could be equivalent to a bad function of chloroplast genes transcription.

Following this analysis, there is also two genes encoding photosystem II reaction center proteins Z and H, respectively: *psbZ*, which regulates the interaction of photosystem II cores (PSII) of the chloroplast with the light-harvesting antenna, and *psbH*, known to be o crucial for the stability and assembly of photosystem II complex (Consortium, 2018). PSII consists of a very conserved protein-pigment complex surrounded by distinct light harvesting centers whose activity is crucial for oxygenic photosynthesis. This complex is comprised by about fifteen proteins and its main functions are light-absorption, charge separation and electron transport from H_2O, subsequently generating O_2 and the proton gradient used for ATP formation. Its biogenesis is intricate due to PSII subunits encoding genes are dispersed between the chloroplast genome and the nuclear genome (Shen, 2015; Chotewutmontri *et al.*, 2020) and its deductible that a variation in two protein coding genes might compromise the well function of photosynthesis.

In its turn, there is also *accD* gene which is reported to be essential for leaf development and to maintain plastid compartment in tobacco, where its raised expression results in higher amount of ACCase (Acetyl-CoA carboxylase) in plastids and fatty acids (Madoka *et al.*, 2002; Kode *et al.*, 2005; Li *et al.*, 2018). On the other hand, it is also important for embryo development stage in Arabidopsis (Morinaka *et al.*, 2006). This gene is greatly distributed in plants and its loss from plastid genome in some plant families like Campanulaceae and Fabaceae was coherent with additional an ACCase equivalent in the nucleus (Li *et al.*, 2018).

Lastly, there's also genes encoding NDH complex subunits: *ndhH* gene encoding NAD(P)H-quinone oxidoreductase subunit 4L protein and *ndhE* gene encoding NAD(P)H-quinone oxidoreductase subunit H protein (Consortium, 2018). The NDH complex in chloroplast mediates the PSI cyclic electron flow (Kato *et al.*, 2018) and, once again, a nonsynonymous SNP in one of this genes might represent a problem for photosynthesis.

4.4.2 Mitochondrial genome SNPs

In the case of the mitochondrial genome, the annotated genes are divided into genic and intergenic variants.

The genic annotation file reports 86 occurrences, of which 3 variants are exclusive to *Q. suber*. These SNPs are annotated in a non-coding RNA exonic region (reported as "nc-RNA exonic") meaning they're placed in non-coding regions resulting in the absence of any phenotypic alteration. When analyzing exclusive variants to *Q. ilex*, 39 SNPs are annotated of which 30 are exonic, 3 are intronic and 6 are non-coding. From the 30 exonic variants, 20 are nonsynonymous, 8 are synonymous and 2 are unknown in terms of exonic alteration types.

In common, the species under study share 20 SNPs annotated with a MAF value minor to 25 %. These SNPs are located at non-coding exonic regions of 4 different transfer RNA (tRNA) genes - *trnP-TGG*, *trnW-CCA*, *trnN-GTT*, *trnD-GTC* - and 1 ribossomal RNA gene (rRNA) - *rrnS*. Given the fact that these variations are located at non-coding regions, no exonic alteration type is attributed by the software, i.e., it is not specified whereas the alteration is synonymous, nonsynonymous or unknown. In the plant mitogenome, tRNA genes may have many origins including native mitochondrial or chloroplast genome derived. Most of these genes transferred from plastid sequences are non-functional but there are several reported to be putatively functional tRNAs in angiosperms mitochondrion genomes, and those presented above are reported to be functional (Richardson *et al.*, 2013). Considering the importance of this organelle and its role in cellular imbalance and damage, in a review made by Cavalcante *et al* (2020), they examine the mitochondrial genetics and epigenetics complexity highlighting the role of ncRNAs inside mtDNA. For instance, the role of ncRNAs in regulatory processes like the lncRNA (long ncRNA) works in the modulation of mitochondrial metabolism and structure (Zhao *et al.*, 2018; Cavalcante *et al.*, 2020). Overall, noncoding variants are reported to be able to effect gene expression or gene function (French and Edwards, 2020).

Continuing this analysis with a value for minimum allele frequency of 25 % and following the criteria that an exclusive SNP to a specie needs to be represented by at least 80 % of

the individuals of one specie, there is no exclusive SNPs for neither cork oak nor holm oak individuals.

On the other hand, in the intergenic annotation file reports in total 849 SNPs from which only 6 are annotated as *Q. suber* exclusive (5 intergenic and 1 downstream intergenic). For holm oak exclusive SNPs annotation, the number is 526 (67 are downstream, 363 are intergenic, 74 are upstream and 22 are upstream of one gene and downstream of the following gene) where the minor allele frequency is high for all cases (over 80 %).

When restricting the value for MAF to 25 % or less, the number of overall SNPs is reduced to 217 (21 are downstream, 181 are intergenic, 10 are upstream and 5 are upstream/downstream).

Lei *et al.* (2013) reported that more than 50 % of the mitogenome consists of intergenic regions, with that being a reason for expansion and accumulation of repeated sequences in intergenic regions of the mitogenome (Lei *et al.*, 2013), and this can be the reason why there are so many more SNPs found in this regions when comparing to genic regions.

Commonly to both species there is 40 SNPs annotations from which in terms of annotation location 30 are intergenic and 10 are downstream located. Every variation in this group have very small MAF percentages.

4.5 Closing remarks

Our results suggested that cpDNA is more conserved than mtDNA, since there is a higher number of SNPs identified among mitochondrial genome, being in accordance with the literature. On the other hand, they also suggest that there is a high genetic variability between the cork oak and holm oak chloroplast, affecting important genes associated to the ABA and defense signaling, the photosynthesis machinery as well as chloroplast transcriptome proteins.

The mitogenome variation is reported to involve different gene contents, as genes and introns, intergenic regions and repeated sequences (Lei *et al.*, 2013). The principal genes of mtDNA are responsible for cell respiration (NADH dehydrogenase, cytochrome and ATP synthase) and also genetic machinery. Also, a very common feature in plants is that

mtDNA genes can be transferred to nuclear genome, representing up to 0,25 % of its genome (Scarcelli, 2020) .

Under some specific conditions, the holm oak shows higher levels of toughness and resistance when compared to cork oak. In previous works, researchers mention its capacity to resist to cold, drought and soil alterations (Soto *et al.*, 2007) and this can be justified with its higher genetic variability at the organelle level. Maybe the higher number of SNPs identified when compare against cork oak could be a consequence of its adaptation to the continuous environmental and climatic changes. This can also be a good sign for this species considering the climate changes that we are witnessing and the ones we will certainly feel in the future. On the other hand, this adaptation may lead to the loss of important characteristics. Regarding, *Q. suber*, the performed analysis shows and validates its conserved profile which is important to maintain its raw materials best quality while confirming as well as its weak capacity for adaptation.

Although, in order to confirm and find more accurate results, especially on holm oak samples, the organelle genomes of this specie should be assembled. That would not only give us the ability to identify variability between holm oak individuals but also to identify new SNPs, especially in mitochondrial genome which is less conserved than chloroplast genome. On the other hand, that would also enable the direct comparison between organelle genomes of these two species.

Chapter 5: Conclusion

Over the years, the progresses in sequencing technologies and bioinformatics tools are providing better quality data and enabling larger and deeper genomic studies. The constant development of new tools and workflows is improving the knowledge on the biological processes of several species, from microorganisms to big animals, their relationship with others and their habitats changes. This enlightenment is very important in areas like agriculture, forest recovery and, even, medicine.

The study of genetic variations gives insight about a species population evolution and adaptation to the environment. More specifically, SNPs are one of the most important genetic variations categories and can be influencers of diseases and predict ancestry history.

The pipeline described over this dissertation integrates many open-source software combined to obtain accurate results on each step, from preprocessing raw read data to the complete analysis of single nucleotide polymorphisms. Sequences of the organelle genomes of two well-known species of *Quercus* were studied.

The fact that cpDNA is more conserved may help in phylogenetic studies, whereas the higher variation on mitogenome may answer to evolutionary alterations. However, given the fact that very little is known and described in the literature about organelle genomes of these two species, this thesis serves as a starting point for a deeper analysis. The next step would be to make the assembly of organelle genomes of the holm oak and, for instance, do the selection of genetic markers or the selection of the individuals with the best characteristics to resist to environment changes and stresses given by the climatic changes.

Chapter 6: References

Alverson, A. J., Wei, X., Rice, D. W., Stern, D. B., *et al.* (2010). "Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae)." Molecular Biology and Evolution **27**(6): 1436-1448. DOI: 10.1093/molbev/msq029.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: . http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Atherton, R. A., McComish, B. J., Shepherd, L. D., Berry, L. A., *et al.* (2010). "Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform." Plant methods **6**(1): 22. DOI: 10.1186/1746-4811-6-22.

Bao, R., Huang, L., Andrade, J., Tan, W., *et al.* (2014). "Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing." Cancer Inform **13**(Suppl 2): 67-82. DOI: 10.4137/cin.S13779.

Bolger, A. M., Lohse, M. and Usadel, B. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120. DOI: 10.1093/bioinformatics/btu170.

Brozynska, M., Furtado, A. and Henry, R. J. (2016). "Genomics of crop wild relatives: expanding the gene pool for crop improvement." Plant Biotechnol J **14**(4): 1070-1085. DOI: 10.1111/pbi.12454.

Budar, F. and Roux, F. (2011). "The role of organelle genomes in plant adaptation: Time to get to work!" Plant signaling & behavior **6**: 635-639. DOI: 10.4161/psb.6.5.14524.

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., *et al.* (2015). "A Phylogenetic Analysis of 34 Chloroplast Genomes Elucidates the Relationships between Wild and Domestic Species within the Genus Citrus." Molecular Biology and Evolution **32**(8): 2015-2035. DOI: 10.1093/molbev/msv082.

Cavalcante, G. C., Magalhães, L., Ribeiro-dos-Santos, Â. and Vidal, A. F. (2020). "Mitochondrial Epigenetics: Non-Coding RNAs as a Novel Layer of Complexity." International journal of molecular sciences **21**(5): 1838. DOI: 10.3390/ijms21051838.

Chassé, B. (2016). "Eating acorns: what story do the distant, far and near past tell us, and why." International Oaks **27**: 107-135.

Chen, K., Li, G. J., Bressan, R. A., Song, C. P., *et al.* (2020). "Abscisic acid dynamics, signaling, and functions in plants." Journal of Integrative Plant Biology **62**(1): 25-54. DOI: 10.1111/jipb.12899.

Chen, Z., Zhao, N., Li, S., Grover, C. E., *et al.* (2017). "Plant mitochondrial genome evolution and cytoplasmic male sterility." Critical reviews in plant sciences **36**(1): 55-69. DOI: 10.1080/07352689.2017.1327762

Chotewutmontri, P., Williams-Carrier, R. and Barkan, A. (2020). "Exploring the link between photosystem II assembly and translation of the chloroplast psbA mRNA." Plants **9**(2): 152. DOI: 10.3390/plants9020152.

Colinas, J., Schmidler, S. C., Bohrer, G., Iordanov, B., *et al.* (2008). "Intergenic and Genic Sequence Lengths Have Opposite Relationships with Respect to Gene Expression." PloS one **3**(11): e3670. DOI: 10.1371/journal.pone.0003670.

Consortium, T. U. (2018). "UniProt: a worldwide hub of protein knowledge." Nucleic Acids Research **47**(D1): D506-D515. DOI: 10.1093/nar/gky1049.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., *et al.* (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158. DOI: 10.1093/bioinformatics/btr330.

Daniell, H., Lin, C.-S., Yu, M. and Chang, W.-J. (2016). "Chloroplast genomes: diversity, evolution, and applications in genetic engineering." Genome biology **17**(1): 134. DOI: 10.1186/s13059-016-1004-2.

French, J. and Edwards, S. (2020). "The role of noncoding variants in heritable disease." Trends in Genetics. DOI: 10.1016/j.tig.2020.07.004.

Garrison, E. and Marth, G. (2012). "Haplotype-based variant detection from short-read sequencing." arXiv: Genomics. https://ui.adsabs.harvard.edu/abs/2012arXiv1207.3907G

Gil-Pelegrín, E., Peguero-Pina, J. J. and Sancho-Knapik, D. (2017). Oaks Physiological Ecology: Exploring the Functional Diversity of Genus Quercus L. Springer.

Greiner, S. and Bock, R. (2013). "Tuning a ménage à trois: Co-evolution and co-adaptation of nuclear and organellar genomes in plants." BioEssays **35**(4): 354-365. DOI: 10.1002/bies.201200137.

Gualberto, J. M. and Newton, K. J. (2017). "Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation." Annual Review of Plant Biology **68**(1): 225-252. DOI: 10.1146/annurev-arplant-043015-112232.

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

Heo, Y. (2015). Improving quality of high-throughput sequencing reads. Electrical & Computer Eng. USA, University of Illinois at Urbana-Champaign. **PhD**.

Hwang, S., Kim, E., Lee, I. and Marcotte, E. M. (2015). "Systematic comparison of variant calling pipelines using gold standard personal exome variants." Scientific reports **5**(1): 17875. DOI: 10.1038/srep17875.

Illumina, I. (2015). "An introduction to next-generation sequencing technology."

Kappelle, M. (2006). Ecology and conservation of neotropical montane oak forests. Springer-Verlag Berlin Heidelberg. DOI: 10.1007/3-540-28909-7.

Kato, Y., Sugimoto, K. and Shikanai, T. (2018). "NDH-PSI Supercomplex Assembly Precedes Full Assembly of the NDH Complex in Chloroplast." Plant Physiology **176**(2): 1728-1738. DOI: 10.1104/pp.17.01120.

Kchouk, M., Gibrat, J.-F. and Elloumi, M. (2017). "Generations of sequencing technologies: from first to next generation." Biology and Medicine **9**(3). DOI: 10.4172/0974-8369.1000395.

Kersten, B., Rampant, P. F., Mader, M., Le Paslier, M.-C., *et al.* (2016). "Genome sequences of Populus tremula chloroplast and mitochondrion: implications for holistic poplar breeding." PloS one **11**(1). DOI: 10.1590/1678-4685-gmb-2019-0161

Kim, H. N., Jin, H. Y., Kwak, M. J., Khaine, I., *et al.* (2017). "Why does Quercus suber species decline in Mediterranean areas?" Journal of Asia-Pacific Biodiversity **10**(3): 337-341. DOI: 10.1016/j.japb.2017.05.004.

Kode, V., Mudd, E. A., Iamtham, S. and Day, A. (2005). "The tobacco plastid accD gene is essential and is required for leaf development." The Plant Journal **44**(2): 237-244. DOI: 10.1111/j.1365-313X.2005.02533.x.

Kulski, J. K. (2016). "Next-generation sequencing—an overview of the history, tools, and "Omic" applications." Next Generation Sequencing–Advances, Applications and Challenges: 3-60. DOI: 10.5772/61964.

Lei, B., Li, S., Liu, G., Chen, Z., *et al.* (2013). "Evolution of mitochondrial gene content: loss of genes, tRNAs and introns between Gossypium harknessii and other plants." Plant Systematics and Evolution **299**: 1889-1897. DOI: 10.1007/s00606-013-0845-3.

Li, H. (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." arXiv: Genomics. arXiv:1303.3997v2

Li, H., Handsaker, B., Wysoker, A., Fennell, T., *et al.* (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079. DOI: 10.1093/bioinformatics/btp352.

Li, J., Su, Y. and Wang, T. (2018). "The Repeat Sequences and Elevated Substitution Rates of the Chloroplast accD Gene in Cupressophytes." Frontiers in Plant Science **9**(533). DOI: 10.3389/fpls.2018.00533.

Li, Y., Chen, W., Liu, E. Y. and Zhou, Y.-H. (2013). "Single Nucleotide Polymorphism (SNP) Detection and Genotype Calling from Massively Parallel Sequencing (MPS) Data." Statistics in Biosciences **5**(1): 3-25. DOI: 10.1007/s12561-012-9067-4.

Lu, Y., Shen, Y., Warren, W. and Walter, R. (2016). "Next Generation Sequencing in Aquatic Models." Next Generation Sequencing-Advances, Applications and Challenges: 61-79. DOI: 10.5772/61657.

Lumaret, R., Mir, C., Michaud, H. and Raynal, V. (2002). "Phylogeographical variation of chloroplast DNA in holm oak (*Quercus ilex L.*)." Molecular ecology **11**(11): 2327-2336. DOI: 10.1046/j.1365-294X.2002.01611.x.

Madoka, Y., Tomizawa, K.-I., Mizoi, J., Nishida, I., *et al.* (2002). "Chloroplast transformation with modified accD operon increases acetyl-CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco." Plant and Cell Physiology **43**(12): 1518-1525. DOI: 10.1093/pcp/pcf172.

Magalhães, H. C. (2017). Identification and characterization of structural variation in the cork oak genome. Engineering School, Informatics Department. Braga, University of Minho. **Master Degree in Bioinformatics**.

Maxam, A. M. and Gilbert, W. (1977). "A new method for sequencing DNA." Proceedings of the national academy of sciences **74**(2): 560-564. DOI: 10.1073/pnas.74.2.560.

Metzker, M. L. (2005). "Emerging technologies in DNA sequencing." Genome Res **15**(12): 1767-1776. DOI: 10.1101/gr.3770505.

Morinaka, Y., Sakamoto, T., Inukai, Y., Agetsuma, M., *et al.* (2006). "Morphological alteration caused by brassinosteroid insensitivity increases the biomass and grain production of rice." Plant Physiology **141**(3): 924-931. DOI: 10.1104/pp.106.077081.

Nagaraj, S., Senthil-Kumar, M., Ramu, V. S., Wang, K., *et al.* (2016). "Plant Ribosomal Proteins, RPL12 and RPL19, Play a Role in Nonhost Disease Resistance against Bacterial Pathogens." Frontiers in Plant Science **6**(1192). DOI: 10.3389/fpls.2015.01192.

Njuguna, A. W., Li, Z.-Z., Saina, J. K., Munywoki, J. M., *et al.* (2019). "Comparative analyses of the complete chloroplast genomes of nymphoides and menyanthes species (menyanthaceae)." Aquatic Botany **156**: 73-81. DOI: 10.1016/j.aquabot.2019.05.001.

Ogihara, Y., Yamazaki, Y., Murai, K., Kanno, A., *et al.* (2005). "Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome." Nucleic Acids Research **33**(19): 6235-6250. DOI: 10.1093/nar/gki925.

Oliveira, G. and 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. DOI: 10.1016/j.foreco.2012.01.025.

Pinard, D., Myburg, A. A. and Mizrachi, E. (2019). "The plastid and mitochondrial genomes of Eucalyptus grandis." BMC Genomics **20**(1): 132. DOI: 10.1186/s12864-019-5444-4.

Pinto-Correia, T., Ribeiro, N. and Sá-Sousa, P. (2011). "Introducing the montado, the cork and holm oak agroforestry system of Southern Portugal." Agroforestry Systems **82**(2): 99. DOI: 10.1007/s10457-011-9388-1.

Ramos, A. M., Usié, A., Barbosa, P., Barros, P. M., *et al.* (2018). "The draft genome sequence of cork oak." Scientific data **5**: 180069. DOI: 10.1038/sdata.2018.69.

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015). "High-throughput sequencing technologies." Molecular cell **58**(4): 586-597. DOI: 10.1016/j.molcel.2015.05.004.

Rey, M.-D., Castillejo, M. Á., Sánchez-Lucas, R., Guerrero-Sanchez, V. M., *et al.* (2019). "Proteomics, Holm Oak (*Quercus ilex L.*) and Other Recalcitrant and Orphan Forest Tree Species: How do They See Each Other?" International journal of molecular sciences **20**(3): 692. DOI: 10.3390/ijms20030692.

Richardson, A. O., Rice, D. W., Young, G. J., Alverson, A. J., *et al.* (2013). "The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate." BMC Biology **11**(1): 29. DOI: 10.1186/1741-7007-11-29.

Sá-Sousa, P. (2014). "The Portuguese montado: conciliating ecological values with human demands within a dynamic agroforestry system." Annals of forest science **71**(1): 1-3. DOI: 10.1007/s13595-013-0338-0.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the national academy of sciences **74**(12): 5463-5467. DOI: 10.1073/pnas.74.12.5463

Scarcelli, N. (2020). Population Genomics of Organelle Genomes in Crop Plants. Population Genomics, Springer, Cham**:** 1-28. DOI: 10.1007/13836_2020_82.

Shen, J. R. (2015). "The Structure of Photosystem II and the Mechanism of Water Oxidation in Photosynthesis." Annu Rev Plant Biol **66**: 23-48. DOI: 10.1146/annurev-arplant-050312-120129.

Shi, Y., Liu, Y., Zhang, S., Zou, R., *et al.* (2018). "Assembly and comparative analysis of the complete mitochondrial genome sequence of Sophora japonica 'JinhuaiJ2'." PloS one **13**(8). DOI: 10.1371/journal.pone.0202485.

Silva, S. P., Sabino, M. A., Fernandes, E. M., Correlo, V. M., *et al.* (2005). "Cork: properties, capabilities and applications." International Materials Reviews **50**(6): 345-365. DOI: 10.1179/174328005X41168.

Sloan, D. B., Alverson, A. J., Štorchová, H., Palmer, J. D., *et al.* (2010). "Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm Silene latifolia." BMC Evolutionary Biology **10**(1): 274. DOI: 10.1186/1471-2148-10-274.

Sork, V. L., Fitz-Gibbon, S. T., Puiu, D., Crepeau, M., *et al.* (2016). "First Draft Assembly and Annotation of the Genome of a California Endemic Oak Quercus lobata Née (Fagaceae)." G3: Genes|Genomes|Genetics **6**(11): 3485-3495. DOI: 10.1534/g3.116.030411.

Soto, A., Lorenzo, Z. and Gil, L. (2007). "Differences in fine-scale genetic structure and dispersal in *Quercus ilex L.* and *Q. suber L.*: consequences for regeneration of mediterranean open woods." Heredity **99**(6): 601-607. DOI: 10.1038/sj.hdy.6801007.

Sun, J., Dong, X., Cao, Q., Xu, T., *et al.* (2019). "A systematic comparison of eight new plastome sequences from Ipomoea L." PeerJ **7**: e6563. DOI: 10.7717/peerj.6563.

Surová, D., Ravera, F., Guiomar, N., Martínez Sastre, R., *et al.* (2018). "Contributions of Iberian Silvo-Pastoral Landscapes to the Well-Being of Contemporary Society." Rangeland Ecology & Management **71**(5): 560-570. DOI: 10.1016/j.rama.2017.12.005.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., *et al.* (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature Biotechnology **28**(5): 511-515. DOI: 10.1038/nbt.1621.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). "Ten years of next-generation sequencing technology." Trends in Genetics **30**(9): 418-426. DOI: 10.1016/j.tig.2014.07.001.

Wambugu, P. W., Brozynska, M., Furtado, A., Waters, D. L., *et al.* (2015). "Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences." Scientific reports **5**(1): 1-9. DOI: 10.1038/srep13957.

Wang, K., Li, M. and Hakonarson, H. (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Research **38**(16): e164-e164. DOI: 10.1093/nar/gkq603.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., *et al.* (2008). "The complete genome of an individual by massively parallel DNA sequencing." Nature **452**(7189): 872-876. DOI: 10.1038/nature06884.

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F., *et al.* (2011). "The evolution of the plastid chromosome in land plants: gene content, gene order, gene function." Plant Molecular Biology **76**(3): 273-297. DOI: 10.1007/s11103-011-9762-4.

Yang, H. and Wang, K. (2015). "Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR." Nature Protocols **10**(10): 1556-1566. DOI: 10.1038/nprot.2015.105.

Yao, Z., You, F. M., N'Diaye, A., Knox, R. E., *et al.* (2020). "Evaluation of variant calling tools for large plant genome re-sequencing." BMC Bioinformatics **21**(1): 1-16. DOI: 10.1186/s12859-020-03704-1.

Zhang, W., Chen, J., Yang, Y., Tang, Y., *et al.* (2011). "A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies." PloS one **6**(3): e17915. DOI: 10.1371/journal.pone.0017915.

Zhang, Y., Zhang, A., Li, X. and Lu, C. (2020). "The Role of Chloroplast Gene Expression in Plant Responses to Environmental Stress." International journal of molecular sciences **21**(17): 6082. DOI: 10.3390/ijms21176082.

Zhao, Y., Sun, L., Wang, R. R., Hu, J. F., *et al.* (2018). "The effects of mitochondria-associated long noncoding RNAs in cancer mitochondria: New players in an old arena." Crit Rev Oncol Hematol **131**: 76-82. DOI: 10.1016/j.critrevonc.2018.08.005.

Zimmermann, H. H., Harms, L., Epp, L. S., Mewes, N., *et al.* (2019). "Chloroplast and mitochondrial genetic variation of larches at the Siberian tundra-taiga ecotone revealed by de novo assembly." PloS one **14**(7): e0216966. DOI: 10.1371/journal.pone.0216966.

Appendixes

A. Testing Trimmomatic parameters

In order to perform the selection of the best set of parameters for data preprocessing, two samples for each data set were randomly selected. From those samples a subset of the data was obtained, equally representative, with 1,000,000 of reads pairs each. This was performed randomly with seqtk tool (https://github.com/lh3/seqtk).

In order to run Trimmomatic (Bolger *et al.*, 2014), the sliding window size was set to 10 % of the read while the following parameters were tested for selection:

- Required minimum quality (Q) – 20, 30, 35 and 15.
- Reads minimal length (L or MINEN) – 80 % and 90 % (% of the total read length to keep).

The obtained results for each parameter combination tested are summarized in table 6. As the results were analyzed we observed that the number of reads kept is very similar between Q15 and Q20, either with L80 or L90. Based on the read's good quality, we are able to ignore Q15 since none of the read's quality value drop below 20. On the other hand, for quality values of 30 and 35 the number of reads kept is very low when compared to the previous values. This is due to the excessive restriction applied with those quality values which can lead to the loss of important reads. Regarding the minimum read length, the choice stands between values 80 % and 90 %. In this case, it needs to be considered that neither it can be too short, to avoid losing valuable read information, nor it can be too high as not to lose specificity. Therefore, it is possible to set that 20 for minimum average quality and 80 % for minimum read length as the best parameters to preprocess our data.

Table 6 – Complete set of tests statistics performed on Trimmomatic software. The table holds the number of reads kept after preprocessing with the various sets of parameters.

Samples	Raw reads	Q15L80	Q20L80	Q30L80	Q35L80	Q15L90	Q20L90	Q30L90	Q35L90
AB04	1 000 000	990 290 (99.03 %)	954 837 (95.48 %)	668 397 (66.84 %)	218 764 (21.88 %)	988 840 (98.88 %)	950 148 (95.01 %)	651 639 (65.16 %)	191 888 (19.19 %)
AZ01	1 000 000	991 822 (99.18 %)	966 403 (96.64 %)	759 042 (75.90 %)	254 397 (25.44 %)	990 835 (99.08 %)	963 630 (96.36 %)	745 991 (74.60 %)	222 428 (22.24 %)
Ind113	1 000 000	955 823 (95.58 %)	885 963 (88.60 %)	661 207 (66.12 %)	427 329 (42.73 %)	929 609 (92.96 %)	831 882 (83.19 %)	522 687 (52.27 %)	263 126 (26.31 %)
Ind115	1 000 000	953 693 (95.37 %)	882 711 (88.27 %)	681 682 (68.17 %)	459 884 (45.99 %)	928 510 (92.85 %)	833 141 (83.31 %)	543 594 (54.36 %)	278 014 (27.80 %)
Total	**4 000 000**	**3 891 628 (97.29 %)**	**3 689 914 (92.25 %)**	**2 770 328 (69.26 %)**	**1 360 374 (34.01 %)**	**3 837 794 (95.94 %)**	**3 578 801 (89.47 %)**	**2 463 911 (61.60 %)**	**955 456 (23.89 %)**

B. Testing Variant calling softwares: Freebayes or SAMtools Mpileup?

Both programs require a FASTA reference sequence file and a BAM-format alignment file sorted by reference position containing all samples for variant calling merged together.

For the testing purpose, the same subset of samples used for the preprocessing test were used. In order to choose the best tool, the testes were performed applying the variant calling on the subset of samples mapped against the chloroplast genome.

In the table below (Table 7) are represented the complete group of tests done with SAMtools Mpileup. To simplify the information, the tests are identified as follows:

- M_Set1 – Mpileup with its default parameters (max-depth 8000) and multiallelic call.
- M_Set2 – Mpileup with its default parameters (max-depth 8000) and consensus call.
- M_Set3 – Mpileup with Freebayes default parameters and multiallelic call.
- M_Set4 – Mpileup with Freebayes default parameters and consensus call.

The raw variants identified in the previous sets were then filtered to obtain the significant SNPs identified in each set. The filtrations made were to keep only significant SNPs, with minimum depth per sample of 15 and minimum SNP quality of 30, and also to keep only bi-allelic sites (sites with only two observed alleles, the reference allele and one variant allele only) and remove indels.

Table 7- Number of variants called among several tests using SAMtools Mpileup software.

	Raw				Filtered			
	M_Set1	**M_Set2**	**M_Set3**	**M_Set4**	**M_Set1**	**M_Set2**	**M_Set3**	**M_Set4**
number of samples:	4	4	4	4	4	4	4	4
number of records:	464	465	466	467	187	187	193	193
number of no-ALTs:	0	0	0	0	0	0	0	0
number of SNPs	**462**	**463**	**464**	**465**	**187**	**187**	**193**	**193**
number of MNPs:	0	0	0	0	0	0	0	0
number of indels:	2	2	2	2	0	0	0	0
number of others:	0	0	0	0	0	0	0	0
number of multiallelic sites:	0	0	0	0	0	0	0	0
number of multiallelic SNP sites:	0	0	0	0	0	0	0	0

As it was said before, the actual calling process for Mpileup is done with BCFtools, and the last version of it (version 1.11-24-g9718479+) allows the user to choose between the old SAMtools calling model, designed as consensus-caller and the new multiallelic calling model. The multiallelic calling model was designed to overcome the existing known limitations in the consensus calling which allows a better identification of multiallelic and rare variants. Besides, the later calling model is the most recommended by the developers for most tasks. For that reason, in this study were tested if the model used had any effect on the output results. M_Set3 and M_Set4 were the sets that identified a larger number of significant SNPs, and as there was no difference between multiallelic and consensus calling models in our tests, the selected model was the set applying the multiallelic call since it is the most recommended (M_Set3).

Next, in Table 8 are represented the group of tests performed with Freebayes software to call variants and VCFtools for filtering. The sets are identified as:

- F_Set1 – Freebayes with its default parameters.
- F_Set2 – Freebayes filtering out multi-nucleotide polymorphisms (MNPs).
- F_Set3 – Freebayes with Mpileup by default parameters (max depth 8000), ploidy 1 and filtering out MNPs.

Once again, the raw variants identified with these tests were filtered: keeping only significant SNPs and bi-allelic sites, removing indels, applying minimum depth per sample of 15 and minimum SNP quality of 30.

Table 8 - Number of variants called among several tests using Freebayes software.

	Raw			**Filtered**		
	F_Set1	**F_Set2**	**F_Set3**	**F_Set1**	**F_Set2**	**F_Set3**
number of samples:	4	4	4	4	4	4
number of records:	934	934	910	568	568	568
number of no-ALTs:	0	0	0	0	0	0
number of SNPs	**693**	**693**	**670**	**568**	**568**	**568**
number of MNPs:	46	44	44	0	0	0
number of indels:	181	181	180	0	0	0
number of others:	47	48	48	0	0	0
number of multiallelic sites:	77	76	71	0	0	0
number of multiallelic SNP sites:	12	12	7	0	0	0

Table 9 – Table showing the number of individuals presenting the variants outputted by the best tests from each software, either with the variants (N_IND) or number of individuals having the alternate allele for the variants (N_IND-Alt).

		Mpileup		**Freebayes**		
		M_Set3	**M_Set4**	**F_Set1**	**F_Set2**	**F_Set3**
N_IND	1	--	--	--	--	--
	2	--	--	--	--	--
	3	165	165	--	--	--
	4	28	28	568	568	568
N_IND-Alt	1	168	168	62	62	62
	2	25	25	505	505	505
	3	--	--	--	--	--
	4	--	--	1	1	1

Based on these testes presented in Tables 7, 8 and also Table 9, the choice for the variant call software fell on Freebayes based on its better results, faster running time and simpler commands usage. The only used parameter was to choose the ploidy of the samples (-p 1). Once the variant calling was finished, the filtering done to keep only SNPs was carried out by VCFtools removing indels, setting the minimum and maximum alleles (2), minimum depth per sample 15 and minimum SNP quality 30. VCFtools v0.1.13 (Danecek *et al.*, 2011) is a command-line tool for parsing, analyzing and manipulating VCF files (variant call format files) and among several operations it allows the filtering of variants, in this case, allowing to keep only SNPs in the file.

The main differences between both software were: 1) the mode of use, being simpler when using Freebayes, 2) the running time, being Freebayes significantly faster than Mpileup and 3) the types of variants they detected under the same parameters, identifying Freebayes more significant SNPs than Mpileup. Considering these different aspects, Freebayes was the chosen software for SNP calling,

C. Supplementary tables

The table below (Table 10) contains additional information on the preprocessing statistics presented in chapter 4, section 4.1.

Table 10 –Preprocessing summary information for each sample: Dataset, sample name, raw reads length and read percentage kept after preprocessing.

Dataset	Sample_name	Raw reads length	After preprocessing (%)
1st	AB04_1	308 647 388	95.47 %
	AZ01_1	313 635 776	96.67 %
	CL1	313 056 206	96.35 %
	CL3	292 032 544	96.44 %
	HL11	190 398 188	92.79 %
	HL12	294 849 380	95.73 %
	HL14	292 262 732	95.79 %
	ISA	296 532 648	95.50 %
	L2	286 157 126	96.01 %
	L3	300 257 290	97.15 %
	MN03	272 510 676	95.26 %
	MN04R	297 719 470	97.10 %
	Q32	245 363 110	93.27 %
	QS32	296 759 890	96.70 %
	SN7	309 101 530	95.69 %
	VF03	283 709 780	94.15 %
	VF05	294 937 946	94.80 %
2nd	Ind113	153 465 712	88.58 %
	Ind115	165 020 686	88.28 %
	Ind118	160 378 064	89.24 %
	Ind120	178 573 978	87.02 %
	Ind19	171 937 208	89.36 %
	Ind19A	160 332 834	87.79 %
	Ind24	133 833 250	88.95 %
	Ind29	152 675 486	88.07 %
	Ind48	184 982 870	88.79 %
	Ind6	179 854 826	89.36 %
	Ind66	227 017 034	88.41 %
	Ind7	169 636 824	88.01 %
	Ind74	156 632 266	88.55 %
	Ind75	149 125 312	90.23 %
	Ind76	174 692 310	89.82 %
	Ind9	173 614 318	88.31 %
	Ind98	170 151 536	87.60 %

	IndHL12	154 261 284	88.69 %
	IndHL15	183 424 876	89.72 %
	IndHL16	151 050 836	88.25 %
	IndHL17	175 397 652	89.03 %
	IndHL18	168 082 250	87.98 %
	IndHL19	189 927 134	87.57 %
	IndHL20	162 761 224	88.64 %
	IndHL21	150 698 750	87.47 %
	IndHL3	177 419 584	87.47 %
	IndHL4	193 294 276	88.24 %
	IndHL5	214 564 816	89.24 %
	IndHL7	184 700 700	87.73 %
	IndHL9	159 032 942	88.04 %