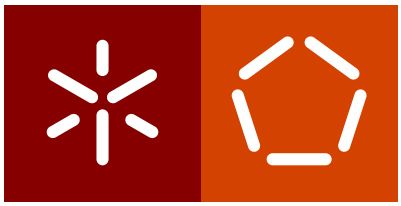
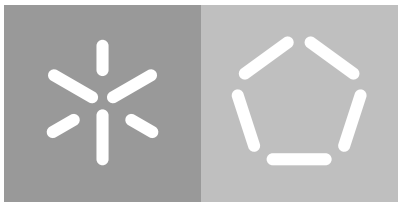


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Hugo Carvalho Magalhães

Identification and characterization of structural variation in the cork oak genome

October 2017



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Hugo Carvalho Magalhães

Identification and characterization of structural variation in the cork oak genome

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Miguel Francisco Almeida Pereira Rocha
António Marcos Costa do Amaral Ramos

October 2017

ACKNOWLEDGEMENTS / AGRADECIMENTOS

Gostaria de iniciar este documento a prestar o meu agradecimento a um conjunto de pessoas que contribuíram para a realização desta tese de mestrado.

Ao Doutor António Marcos Ramos pela orientação desta dissertação, pela oportunidade concedida, pelo acompanhamento diário e pelos ensinamentos científicos e não só.

Ao Professor Miguel Rocha, pela confiança demonstrada, pelo acompanhamento do decorrer do trabalho, ainda que à distância e por representar o corpo docente do Mestrado em Bioinformática.

Às Doutoras Ana Chimenos e Célia Leão e aos Mestres Brígida Meireles, Daniel Gaspar e Pedro Barbosa pela disponibilidade demonstrada para ajudar, pela amizade e pela integração e convivência no CEBAL.

Aos restantes membros do CEBAL pela convivência diária e ajuda na adaptação a uma nova cidade.

Aos meus pais por tornarem financeiramente possível a minha mudança para Beja, pelo apoio constante e por apostarem na minha formação, confiando sempre em mim.

Por último, à Catarina, a minha namorada, pelo apoio constante apesar da distância e pela paciência que tem comigo.

A todos, o meu sincero obrigado!

ABSTRACT

The appearance of high-throughput sequencing technologies revolutionized the study of genomics. The substantial volumes of data generated by these technologies allow a very comprehensive characterization of the species genomes and the genomic variation within individual genomes. This work focused on the identification and characterization of structural variants (SVs) in the cork oak genome, a class of variants described as genomic rearrangements that may be involved in several biological processes. There are many types of SVs, including insertions, deletions, inversions, translocations and duplications.

Cork oak trees are the only commercial source of cork, which is a renewable natural resource that has many applications, particularly in the production of stoppers and sound and thermal insulators, among others, due to its unique features. Cork is one of the most valuable non-wood forest products, putting this species among the most important trees with commercial relevance in the countries where it is naturally distributed.

The work pipeline followed the common steps used in this type of study, namely the sample collection, sequencing, data quality evaluation, preprocessing, read mapping to the reference genome, structural variation calling and, lastly, the identification and characterization of the SVs. To achieve this final goal of identifying and characterizing SV in the cork oak genome, several tests were performed, comparing the quality and length thresholds, for three software applications for SV calling and three sets of mapping parameters.

This present work was the first study performed in cork oak where whole genome resequencing was used, by the analysis of the whole genome of 30 individuals, which included 14 trees producers of good quality cork, along with 16 trees that produce cork with bad quality. This magnitude of genomes is the first step to construct the species pangenome, which will then be crucial to understand how SV determines the differences in cork quality, since this is the most important economic trait of these trees. The 93,980 SVs identified clearly indicated that SV is present in the cork oak genome.

RESUMO

O aparecimento de tecnologias de sequenciação de alto rendimento veio revolucionar o estudo da genómica. O volume substancial de dados gerado por estas tecnologias permite uma caracterização muito abrangente dos genomas das espécies e da variação genómica entre genomas individuais. Este trabalho focou-se na identificação e caracterização de variantes estruturais (VEs) no genoma do sobreiro, uma classe de variantes descrita como rearranjos genómicos que podem estar envolvidos em vários processos biológicos. Existem muitos tipos de VEs, incluindo inserções, deleções, inversões, translocações e duplicações.

Os sobreiros são a única fonte comercial de cortiça, que é um recurso natural renovável com muitas aplicações, particularmente na produção de rolhas, e isoladores sonoros e térmicos, entre outros, devido às suas características únicas. A cortiça é um dos mais valiosos produtos florestais que não a madeira, colocando esta espécie entre as árvores mais importantes com relevância comercial nos países onde esta ocorre naturalmente.

A estrutura deste trabalho seguiu os passos habituais usados neste tipo de estudo, nomeadamente a recolha de amostras, sequenciação, avaliação da qualidade dos dados, pré-processamento, mapeamento das *reads* contra o genoma de referência, a descoberta de variação estrutural e, por último, a identificação e caracterização das VEs. Para alcançar este objetivo final de identificar e caracterizar variação estrutural no genoma do sobreiro, vários testes foram efetuados, comparando valores limiares de qualidade e comprimento, usando três aplicações bioinformáticas de descoberta de VE e três conjuntos de parâmetros de mapeamento.

O presente trabalho foi o primeiro realizado em sobreiro onde foi usada a resequenciação completa do genoma, com a análise do genoma de 30 indivíduos, os quais incluíam 14 árvores produtoras de cortiça de boa qualidade, junto com 16 outras que produzem cortiça com má qualidade. Esta magnitude de genomas usados é o primeiro passo para construir o pangenoma da espécie, que será depois crucial para perceber como é que a VE determina as diferenças na qualidade da cortiça, visto que este é o traço económico mais importante destas árvores. As 93,980 VEs identificadas indicam claramente que a VE está presente no genoma do sobreiro.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BAM	Binary Alignment/Map.
bp	base pairs.
BWA	Burrows-Wheeler Alignment.
CA	contig assembly.
CEBAL	Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo.
CL	clustering.
CNVs	copy number variations.
CQ	cork quality.
CTX	intra-chromosomal translocations.
DEL	deletions.
DNA	deoxyribonucleic acid.
DP	discordantly-aligned read pairs.
DUP	duplications.
GRIDSS	Genomic Rearrangement IDentification Software Suite.
InDels	insertions/deletions.
INS	insertions.
INV	inversions.
ITX	inter-chromosomal translocations.
MAPQ	mapping-quality score.
MEM	maximal exact matches.
MP	mate-pair.
NGS	Next Generation Sequencing.

OEA one-end anchored.

PAVs presence/absence variations.
PCR Polymerase Chain Reaction.
PE paired-end.

RD read depth.
RNA ribonucleic acid.

SA split-reads alignment.
SAM Sequence Alignment/Map.
SC soft-clipped.
SE single-end.
SMEM supermaximal exact match.
SNPs single nucleotide polymorphisms.
ST statistical testing.
SV structural variation.
SVs structural variants.
SW Smith-Waterman's.

VCF Variant Call Format.

WGRS whole genome resequencing.
Wham Whole-genome Alignment Metrics.

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

The substantial volumes of data generated with high-throughput sequencing allow a very comprehensive characterization of the genomic variation that can be identified in individual genomes (Baker, 2012). Among the different types of variants that can be found, structural variants (SVs) represent a class of variants whose characterization in plant genomes is only just beginning (Saxena et al., 2014).

SVs are genomic rearrangements that may be involved in processes related to evolutionary mechanisms, adaptation to specific environmental conditions or associated with genetic disorders. There are several types of SVs, which include insertions/deletions (InDels), inversions (INV), duplications (DUP) and translocations, which can be inter-chromosomal translocations (ITX) or intra-chromosomal translocations (CTX), among others (Saxena et al., 2014; Liu et al., 2015). These variants can be identified and characterized through bioinformatics analyses of high-throughput sequence data, in particular sequence data obtained with paired-end (PE) sequencing protocols.

The pangenome represents the sum of the genes for a given species, including all types of variation identified, and to date only a few plant species had their pangenome characterized (Golicz et al., 2016). In cork oak, no information is available regarding the type, distribution and frequency of SVs in the species genome. Additionally, besides the identification of SVs, their effect on phenotypic traits under complex genetic regulation has never been studied. Lastly, even though there are already several bioinformatics software tools to execute this type of analysis, their performance in plant genomes is still poorly characterized. Eventually, the development of new software tools and/or analyses pipelines may be needed to efficiently extract all the information and potential contained in the large volumes of sequence data produced.

Given the exponential improvement of the sequencing technology, where a new technique becomes obsolete just a few years after its release, there is a need to keep the analysis tools up to date. That is a tough challenge due to the tools development time, which is normally long, given the complexity they require. Although, there are usually some tools

relatively adequate, whether it is previous software that can be adapted to the new data, or software developed by the sequencing company, finding the appropriate software for the study goals can nevertheless be hard, depending on the specificity of the data and objectives. Hereupon, the best solution in many cases is to try different tools and evaluate their performance concerning the study goals.

The Animal Genomics and Bioinformatics group of *Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo (CEBAL)* started its activities in January 2015, and has been heavily involved in the cork oak genome sequencing project, as well as in other studies whose objective is to develop genomic tools for cork oak, including the characterization of cork oak transcriptomic response to several biotic and abiotic stress factors, identification of genetic markers for economically important traits in cork production, and the development of molecular traceability systems for different cork oak origins.

Cork oak trees are the only commercial source of cork, a widely used product, with multiple functions. It consists on a thick periderm made of dead cells with empty lumens, which has an insulating and protective role (Pereira, 2011). As it is a product to be sold, the cork quality (CQ) is a very important factor. There are some criteria to determine if cork is good, most importantly the cork tissue homogeneity and higher thickness, the low porosity level (Pereira et al., 1996) and whether it was picked from a later harvest (Soler et al., 2007; Teixeira et al., 2014).

In order to identify genetic markers for CQ, the genomes of 30 cork oak individuals were sequenced, using a WGRS approach, with high-throughput sequence data generated using the Illumina HiSeq X10 platform (Illumina, 2015). These individuals were selected according to their phenotype for CQ, and two groups were formed, including one group formed by 14 trees with good CQ and a comparison group, with 16 individuals, whose CQ was characterized as bad.

The bioinformatics analyses of these data were performed using the draft sequence and annotation of the cork oak genome, whose publication is presently under peer-review. In this work, we initially tested and evaluated the performance of different bioinformatics approaches to identify and characterize the SVs present in the genomes of the 30 cork oak individuals, an effort that was then followed by an association study where we investigated the effect on CQ of the SVs identified.

1.2 OBJECTIVES

In the present study we aim to identify and characterize SV in the cork oak genome. The detailed objectives for the work include:

1. Review and survey the literature available regarding the identification of SV at the genome level, including bioinformatics analysis tools and their application.

2. Identification of all types of SVs present in the cork oak genome, to generate a comprehensive characterization of the catalogue of SV for the species.

3. Evaluate and determine the performance of different software tools and analysis strategies for the identification of SV in plant genomes, using cork oak as model species.

4. Determine the effect of SVs on CQ, by directly comparing the genomes of two groups of 14 and 16 individuals displaying, respectively, good and bad CQ.

1.3 STRUCTURE OF THE THESIS

Chapter 2: State of the art

The current state of the DNA sequencing techniques is evaluated, particularly Illumina sequencing, as well as some concepts concerning SV. Discussion about previous work done about SV and a description of the software used.

Chapter 3: Methods

All the steps of this work are discriminated, since the data quality evaluation until the SVs identification and characterization.

Chapter 4: Results

Comparisons of preprocessing thresholds, mapping software parameters and SV calling software are performed. Detailed results of the SVs identification and characterization are provided.

Chapter 5: Discussion

Discussion of the results, presenting possible reasons to explain them and comparing them with others from previous studies.

Chapter 6: Conclusions and future work

An overall analysis of this work, its contribution for cork oak genomic research and its limitations. Discussion on further work that can be made from this study.

2

STATE OF THE ART

2.1 RELEVANT CONCEPTS

2.1.1 *DNA Sequencing*

The concept of DNA sequencing was first introduced by Frederick Sanger in 1975 (Sanger and Coulson, 1975). The method was based on the incorporation of chemically altered nucleotides, called dideoxynucleotides, which tricked the DNA polymerase, yielding a collection of nucleotide-specific terminated fragments for each of the four bases. The resultant fragments were then separated by size by electrophoresis, making possible to identify the positions of each nucleotide.

Sequencing by radiolabelled methods underwent numerous improvements following its invention until the mid 1980s, when a company called Applied Biosystems, Inc. commercialized a fluorescent DNA sequencing instrument invented in Leroy Hoods laboratory at the California Institute of Technology (Smith et al., 1986b). The laborious processes of gel drying, X-ray film exposure and developing, reading autoradiographs, and performing hand entry of the resulting sequences were replaced by an instrument that used reactions primed by fluorescently labeled primers. In this instrument, a raster scanning laser beam crossed the surface of the gel plates to provide an excitation wavelength for the differentially labelled fluorescent primers to be detected during the electrophoretic separation of the fragments. These high-throughput slab gel fluorescence instruments largely contributed to the sequencing of several model organism genomes, like *Caenorhabditis elegans* (Sulston et al., 1992) and *Arabidopsis thaliana* (Xu et al., 1995). However, and although they were impressive in their data producing capacity, there were still several manual and, hence, labor-intensive and error-prone steps.

With the introduction of capillary sequencing instruments, like MegaBACE (Marsh et al., 1997) and ABI PRISM 3700, the limitations originated by manual steps in slab gels were eliminated. These instruments provided single-nucleotide resolution by directly injecting a polymeric separation matrix into capillaries. The ABI PRISM 3700 instruments were the main data-generating instruments for the human genome project, among others. Even

so, this Sanger-based approach still had many limitations, empowering the appearance of new sequencing approaches, called Next Generation Sequencing (NGS).

In NGS instruments, the DNA to be sequenced is used to construct a library of fragments that have synthetic DNA (adapters) added covalently to each fragment end using DNA ligase, by contrast with the cloning step performed by Sanger-based methods. Also, the library fragments are amplified *in situ* on a solid surface containing adapter sequences that are complementary to those on the library fragment, instead of using microtiter plate wells.

The first NGS platform was launched in 2005 by the company 454 Life Sciences. This sequencer relies on pyrosequencing technology, in which, instead of using dideoxynucleotides to terminate the chain amplification, as the Sanger-based methods do, it depends on the detection of pyrophosphate released during nucleotide incorporation. Its main applications are the sequencing of bacterial and viral genomes, multiplex-Polymerase Chain Reaction (PCR) products, validation of point mutations and targeted somatic-mutation detection, producing single-end (SE) reads of 400 base pairs (bp) (Mardis, 2017).

Over time, new platforms have emerged. ABI SOLiD, which produces 75 bp reads for SE data and 50 bp for PE data, is applied in complex genomes and genome-wide NGS applications, ribonucleic acid (RNA)-seq, hybrid capture or multiplex-PCR products and somatic-mutation detection. Pacific Biosciences produces reads either SE or circular consensus up to 40,000 bp in length, being used in complex genomes, microbiology and infectious-disease genomes, transcript-fusion detection and methylation detection. Ion Torrent is used in multiplex-PCR products, microbiology and infectious diseases, somatic-mutation detection and validation of point mutations, producing SE reads between 200-400 bp. Oxford Nanopore is used in pathogen surveillance, targeted mutation detection, metagenomics and bacterial and viral genomes, with reads length produced reaching the tens of thousands of bp (Goodwin et al., 2015), depending on library preparation (Mardis, 2017).

The most widely used platform is Illumina, which was the platform used in this study, reason why it is described in detail below. Illumina sequencers generate 150-300 bp PE reads, and its main applications are complex genomes and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection, forensics and noninvasive prenatal testing.

Illumina Sequencing

Illumina was the first platform to achieve sequencing of the first human genome at 30x coverage, the first cancer genome and the first genome in a single day (Ley et al., 2008; Saunders et al., 2012; Bentley et al., 2008). The Illumina HiSeq X10 platform (Illumina, 2015) was the first one to break the $1000 barrier for human whole-genome sequencing. Since it can sequence more than 18,000 genomes per year at a low cost (for example, the first human

genome sequence consumed 12 years and cost nearly $3 billion (S. Collins et al., 2004)), this platform is ideal for population-scale genome sequencing, resorting to ten instruments that together form the HiSeq X10 system. The performance parameters include 1.6-1.8Tb of output per run, 5.3-6 billion of single reads passing filter, 2x150 bp supported read length, run time less than 3 days and over 75% of the bases with quality above 30. It supports two library preparation kits, the TruSeq DNA PCR-Free Library Prep Kit, which provides a fast, gel-free protocol with superior coverage of areas difficult to sequence, like GC-rich regions, promoters and repetitive regions; and the TruSeq Nano DNA Library Prep Kit, which, with just 100 nanograms of DNA, allows for efficient sequencing (Illumina, 2015). This all combines to provide a faster, more accurate and cheaper way to perform population-scale whole genome sequencing.

2.1.2 *Structural Variation*

The development of high-throughput sequencing techniques revolutionized the genomic studies aiming to better understand how genomic variations influence individuals. These studies have mostly focused on evaluating the effects of single nucleotide polymorphisms (SNPs) (Varshney et al., 2009; Lai et al., 2012). However, it is known that SNPs do not capture all the meaningful genomic variations, with SV also playing an important role concerning to phenotypic differences (Vacic et al., 2011; Nishida et al., 2013).

Types of structural variants

SV is defined as a group of genomic variations, SVs, with many types, including InDels, INV, ITX, CTX, tandem DUP and dispersed DUP (Figure 1).

InDels consist on the addition of a sequence comparing with the reference genome, insertions (INS), or an exclusion of a sequence present in the reference genome, deletions (DEL). An INV is a change in the sequences order, inverting it. On the other hand, a translocation, whether it is ITX or CTX, is also a change in the sequences order, but by moving a sequence from one position to another in the genome. There are two types of duplications, the tandem duplications, consisting on the adjacent sequence repetition, and dispersed duplications, in which the sequence is repeated apart.

Deletions, insertions and duplications alter the copy number of the genome and are thus called unbalanced SVs. Inversions and translocations dont change the copy number and are called balanced SVs.

A large group within the SVs are the copy number variations (CNVs), which contemplates sequences that demonstrate a variable copy number between individuals, including more than one type of SVs, such as InDels and duplications (Scherer et al., 2007). A significant amount of phenotypic effects have been correlated with CNVs, either in human

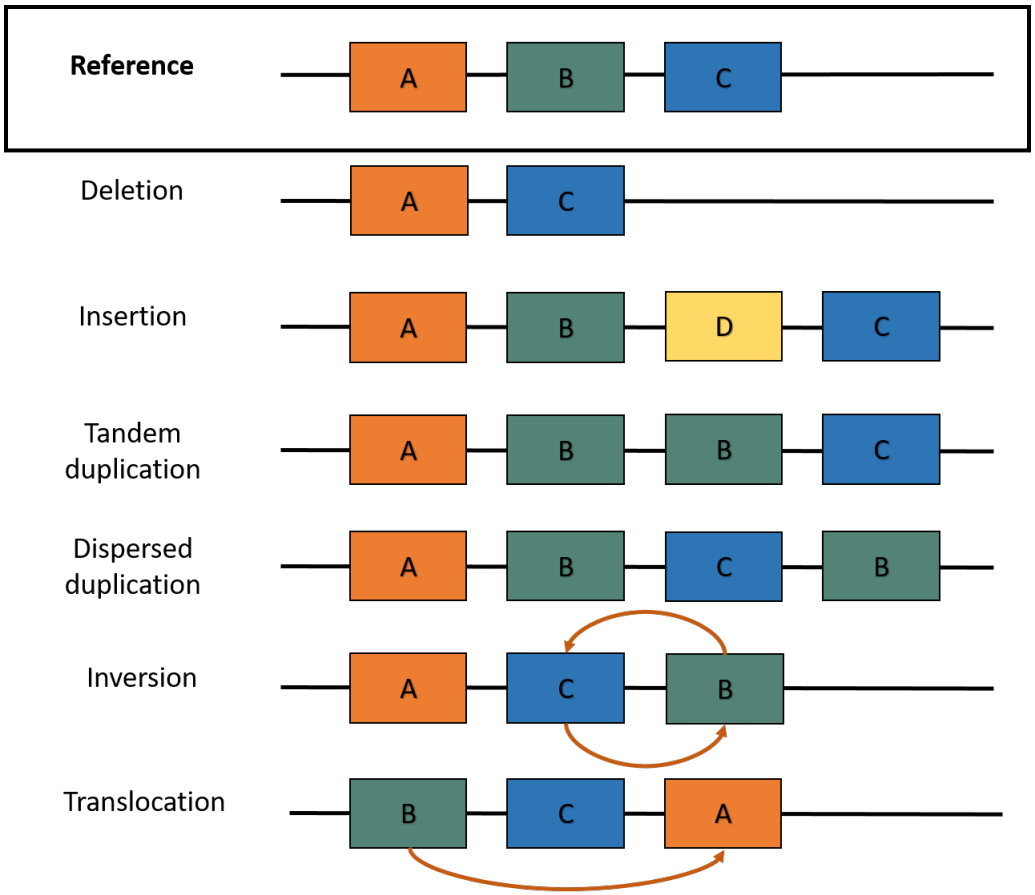


Figure 1: Types of structural variations. Each block (A, B and C) represents a DNA segment.

diseases (Miller et al., 2010a; Stobbe et al., 2013) or in plants' regulation mechanisms (Beló et al., 2010; McHale et al., 2012).

Another group includes the presence/absence variations (PAVs), where there are some sequences that are either present or absent, depending on the individuals analyzed. PAVs can be considered to be extreme CNVs, where some individuals have multiple copy numbers and the others have null copy numbers. PAVs have also been associated with phenotypic effects in plants, like contributing to heterosis (Springer et al., 2009) and stress response (Haun et al., 2011).

Considering all this information, it is important to identify the SVs that can be present in the individuals studied in this work, and characterize them in order to investigate the presence of SVs on the cork oak genome and their possible effect on CQ.

Pangenome

The concept of pangenome was first introduced by Tettelin et al. (2005). The pangenome is the complete set of genes that represent a species. Pangenome genes can be divided into

two groups, including the core genes, which are present in all individuals, and the variable or dispensable genes, which are present only in some individuals or even just in one individual, being called the unique genes.

The core genes can be considered the species essence, whereas the variable genes represent the species diversity (Medini et al., 2005). The core genes include the fundamental genes for the species survival, being thus necessary their presence in all individuals. On the other hand, variable genes are distinctive genes that may encode functions that are not essential for survival, but which may confer selective advantages.

The pangenome can also be classified as open or closed. In the open ones, the number of genes appears to be infinite, because the pangenome keeps increasing with the addition of new individuals, revealing high diversity. In contrast, the closed pangenome reveals low diversity. After a certain number of individuals, new additions do not result in an expansion of the pangenome, showing a limited gene pool.

So, in an open pangenome, it is expected that the core genome decreases with the addition of new individuals. On the other hand, more individuals results in a bigger pangenome. The relations between the number of individuals and the number of clusters are shown in Figure 2.

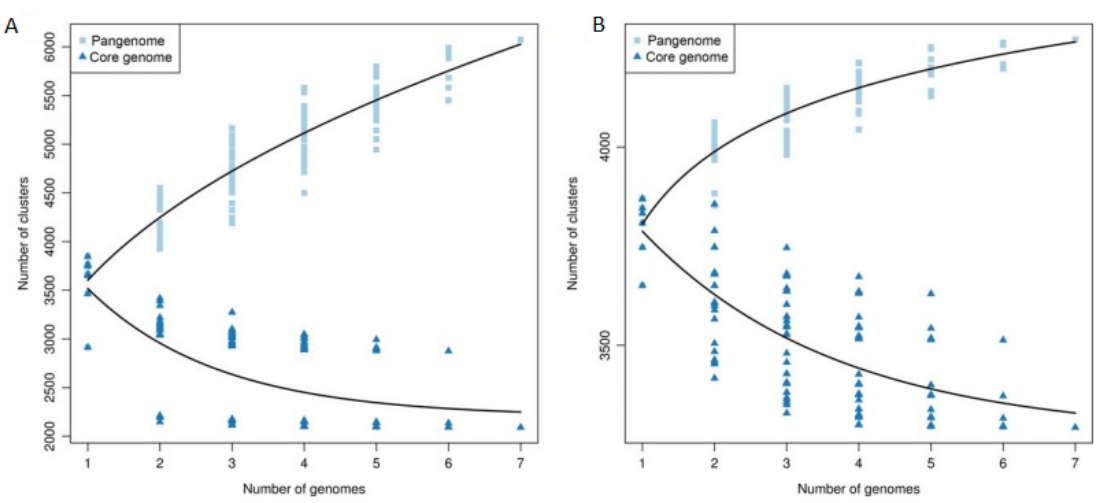


Figure 2: Comparing two pangenomes with seven individuals each. Seven diverse individuals (A) and seven related individuals (B). The pangenome is much bigger in A. On the other hand, the core genome is bigger in B. This suggests an open pangenome in A and a closed one in B. Adapted from Golicz et al. (2016)

Therefore, to identify and characterize the SVs present within a species is a crucial step to build its pangenome, in order to determine whether they occur in the core or the variable genome, and how they might have an impact on the individual. This is still very unexplored, due to the low amount of complete, well-annotated genome sequences. How-

ever, new technologies, such as long-read single molecule sequencing, have the potential to overcome these flaws, which will boost pangenomic studies and, consequently, the characterization of SV.

2.2 PREVIOUS WORK

2.2.1 *Previous studies on structural variation*

The associations between SVs and their phenotypic effects have been comprehensively characterized in humans. For instance, Vacic and colleagues (2011) showed the association between duplications of the *VIPR2* gene and a significant risk for schizophrenia. Some autism spectrum disorders can also be caused by SV (Marshall et al., 2008; Stobbe et al., 2013). Other diseases have been related with SV presence, like Smith-Magenis syndrome (Smith et al., 1986a), Williams-Beuren syndrome (Pober, 2010) and Angelman syndrome (Buiting et al., 1995).

As demonstrated, association between SV and human diseases has been widely studied, since it can lead to great improvements in human population health. However, SV has been associated with phenotypic features other than disease, like metabolic processes and physical features. For example, and regarding metabolic processes, a duplication in the *AMY1* gene increases the salivary amylase production (Perry et al., 2007), while a deletion in the *APC* gene increases the bone mineral density (Chew et al., 2012). Concerning physical features, the short stature of an individual is determined by deletions in multiple *loci* (Dauber et al., 2011), and the olive color skin is associated with a partial duplication on the promoter region of *MATP* (Graf et al., 2007).

On the other hand, studies on other animals have been performed, associating SV also with disease, metabolic processes and physical features. Studies in cow showed an association between a duplication in *CATHL4* and *ULBP17* genes increase the resistance to pathogens and parasites (Bickhart et al., 2012), preventing diseases. Also in cow, bull fertility increasing was associated with a duplication in the *TSPY* gene (Hamilton et al., 2012), and different cow colors were associated with different SVs, with a translocation in the *KIT* locus within different chromosomes leading to color sidedness (Durkin et al., 2012), a deletion in the *MSHR* gene originating a red coat color (Joerg et al., 1996) and a dilute-colored coat color being associated with a deletion in the *PMEL* gene. Besides these, studies in other animals, like dogs (Axelsson et al., 2013), Rhesus macaques (Degenhardt et al., 2009) or goats (Fontanesi et al., 2009), associated SV with both disease, metabolic processes and physical features.

Structural variation in plants

In plants, even though several examples are available, much less is known about the association between SV and phenotypic traits.

In maize, there have been some studies that proved the impact of SV on phenotype. Beló and colleagues (2010) suggest that CNVs might have a considerable impact in disease response and heterosis, while Springer and colleagues (2009) also suggest that SVs, in particular PAVs, may contribute to heterosis. Additionally, 570 genes were found to be absent from the B73 maize line, when comparing with another line, Mo17 (Lai et al., 2010).

SVs found in soybean were specifically localized to gene-rich regions that harbour clustered multigene families, being the most abundant classes of gene families associated with these regions important for plant biotic defense (McHale et al., 2012). Additionally, comparison between the assembled sequence of a wild soybean genome and the reference genome revealed a total of 4,444 and 1,148 large PAVs (over 500 bp) that were absent in the reference and the wild soybean genomes, respectively (Lam et al., 2010).

The impact of SVs has also been demonstrated in wheat, a species where SVs detected in the photoperiod-insensitive alleles Ppd-A1a and PpdB1a were shown to affect heading time (Nishida et al., 2013), while SVs were also involved in flowering regulation (Díaz et al., 2012).

The sequencing of eighteen *Arabidopsis thaliana* genomes revealed that between 2.1 and 3.7 Mb of sequence present in the reference genome were missing in these accessions. In each accession, there were on average 319 novel genes or gene fragments (Gan et al., 2011). Another survey of 80 accessions showed that 10% of the genes from the reference sequence were absent in one or more accessions, with an average of 444 genes per accession (Tan et al., 2012).

Overall, there are many SV studies in humans, for obvious reasons, since they usually aim to find solutions to improve diseases treatment and diagnosis. The number of studies in plants is smaller, as plant genomes sequencing came long after the human genome sequencing, and, in many species, such as cork oak tree, it was not even possible to perform this type of analysis, since there was no genome available. However, with sequencing progress (NGS) and the consequent increase on the number of species with their genome sequenced, it is anticipated that characterization of SV and its phenotypic effects will increase considerably.

2.2.2 *Bioinformatics structural variation pipeline*

Due to the potential biological impacts SVs can have, we need technologies to call them. Currently, one of the most used technologies to study SV is the NGS, which allow us to perform whole genome resequencing (WGRS), providing means to identify SVs by

mapping the NGS reads to the reference genome or building a pangenome with these reads.

The NGS protocol is formed by four simple steps. The first one is to break the genome into short fragments by sonication or enzymatic cutting. After that, the next step consists in selecting fragments of a certain size, the insert size. Then, the two ends of each fragment are sequenced and then analyzed, in order to identify SVs.

Different strategies exist to compare sample DNA sequences with the other ones from the pangenome, and include mapping the NGS reads to the reference genome or aligning the reads from different individuals. There are then two categories where SV calling methods can be classified, including mapping-free and mapping-based approaches. In this work, we will focus on the mapping-based approaches, since these are expected to account for the vast majority of future SV studies (Guan and Sung, 2016).

Mapping-based SV calling pipelines include four ordered basic steps: read mapping and classification, SVs discovery, SVs verification and SVs annotation and visualization. The given input consists of a set of PE reads from the NGS process applied to several individuals. The data is then organized and filtered to posteriorly find the SVs candidates. After that, the candidates are examined and false discoveries are removed. Finally, the SVs are annotated and visualized, providing additional information for biological analysis or validation.

The read mapping and classification step, is divided in three parts. The first one consists of mapping the reads, where, if there is no SV breaking point in the DNA fragment and its quality is high, the two reads are mapped in the correct orientation and strand, and the distance among them is consistent with the insert size distribution. On the other hand, if these requirements are not met, the paired read is anomalous. The more frequent read mappers used are *BWA-MEM* (Li, 2013), *Bowtie2* (Langmead and Salzberg, 2012) and *Mosaik* (Lee et al., 2014). It should be emphasized that certain SV callers require specific read mappers.

The second stage of the read mapping and classification step is the filtering of the read mapping results, because reads may be incorrectly aligned by the mappers. Most of the reads mappers have a mapping-quality score (MAPQ), which indicates the probability of a wrong mapping. Several of them have a threshold, keeping only the mappings above that threshold. Although this filtering process removes mapping errors, it can also potentially remove SV signals.

The last stage includes the reads classification, when the anomalously mapped reads are identified, showing evidence of the existence of SV. The three types of anomalously mapped reads include the discordant ones, which were mapped either on a different chromosome, incorrect strands, incorrect orientation or incorrect insert size, the soft-clipped ones, which contains the breakpoint signal and were mapped partially, and the one-end

anchored (OEA) reads, also containing the signal for breakpoint and were not mapped although their mates were.

Upon completion of the read mapping and classification steps, the SVs discovery comes into play, aiming to identify regions of possible SVs candidates, using the information from the anomalously mapped reads discovered in the previous step. For that purpose, there are four commonly used techniques, namely, clustering (CL), split-reads alignment (SA), contig assembly (CA) and statistical testing (ST) (Medvedev et al., 2009).

The goal of the CL technique is to group reads supporting the same SVs together (Chen et al., 2009). This method is less sensitive to small InDels. The SA approach tries to align two types of the anomalously mapped reads that result from the read mapping and classification stage, the soft-clipped reads and the OEA reads, to find the matching breakpoints, or refine the breakpoints identified by the other type of anomalously mapped reads, the discordant ones (Zhang et al., 2011). In turn, CA proposes the assembling of paired reads into contigs (Hajirasouliha et al., 2010) which, when long enough, facilitate locating the SV breakpoints. Some SV callers are designed for whole genome assembly, producing long contigs, building them by existing *de novo* assemblers, such as *Velvet* (Miller et al., 2010b) or *Cortex* (Iqbal et al., 2012). Others, like *TIGRA* (Chen et al., 2014), use an assembler specifically designed for SV breakpoint assembly. Finally, the ST technique, commonly used to find CNVs, uses statistical models to verify the read-depth against null distributions.

There are also some SV callers that use an hybrid-approach for SVs discovery (Rausch et al., 2012; Bartenhagen and Dugas, 2015). They consist of the integration of different types of anomalously mapped reads, to improve SV calling sensitivity.

The third step of the SV calling pipelines is the SVs verification. The predictions may need to be supported with more evidence, because the mappers can produce many wrong mappings leading to false predictions. Some methods used include the use of split-reads to filter out noisy candidates (Jiang et al., 2012), the use of read depth (RD) for validation of the SV breakpoints resorting statistical methods (Sindi et al., 2012), the use of the known features of genomic regions (Sboner et al., 2010) and the use of *ad hoc* quantitative filters based on the number of discordant reads (Chen et al., 2009) or the number of split-reads (Schröder et al., 2014).

Lastly, the pipelines end with the SVs annotation and visualization. Regarding annotation, SV callers generally provide information on the SVs location, their nearest genes, their formation mechanisms and sequence features. This is very helpful for understanding the formation mechanisms and downstream implications.

The visualization step is very important, allowing the visual inspection of the SVs, and assisting in data interpretation, showing useful information regarding the mechanisms of SV formation.

Some factors may have an impact on SV calling, such as SVs properties (type, size and frequency), library properties (coverage, insert size and read length of the paired reads), sequencing errors (impacting both sensitivity and specificity of SV calling), quality of the reference genome (needs to be complete and accurately assembled) and sequence context (certain genome regions are difficult to sequence).

In order to achieve better sensitivity, it is possible to combine prediction results from different SV callers to get the SVs candidates. Another combination potentially useful includes SV detection combining multiple samples to achieve population-scale studies (Guan and Sung, 2016).

Currently, there are several bioinformatics tools aiming to identify and characterize SVs, designated SVs callers, which differ on the approach they implement and their function. This information is summarized in Table 1.

Table 1: Some of the existent SV callers and their main functions, which distinguish them from each other

SV caller	Function
BreakDancer (Chen et al., 2009)	Predicts a wide variety of structural variants including indels, inversions and translocations
TIGRA (Chen et al., 2014)	Implements a set of novel data analysis routines to achieve effective breakpoint assembly from NGS data
Socrates (Schroder et al., 2014)	Highly efficient and effective method for detecting genomic rearrangements in tumours that uses only split-read data
PRISM (Jiang et al., 2012)	Identifies SVs and their precise breakpoints from whole-genome resequencing data
GASVPro (Sindi et al., 2012)	An algorithm combining both paired read and read depth signals into a probabilistic model that can analyze multiple alignment of reads
GRIDSS (Cameron et al., 2017)	A high-speed SV caller that performs genome-wide break-end assembly prior to variant calling using a positional de Bruijn graph assembler
Wham (Kronenberg et al., 2015)	Provide a single integrated framework for both structural variant calling and association testing

Although tools to identify and characterize SVs are already available, there are still some improvements to achieve. For example, the efficiency of the SV calling algorithms is one of the aspects that could be improved. New technologies, including long-read single molecule sequencing, promise reads hundreds of times longer, bringing up new challenges and opportunities with the development of new algorithms enabling assembly of long reads promising to deliver high-quality genomes.

Yet, lowering sequence costs will lead to larger volumes of data, creating the need for better databases, especially for plants genomes and pangenomes, species for which these resources are less developed, when compared with others, humans in particular, for which there are already some databases of international collaborations for human NGS data, such as The Cancer Genome Atlas (Weinstein et al., 2013) and the International Cancer Genome Consortium (Joly et al., 2012).

2.2.3 *Software tools*

In this section, the software used in this work will be described in detail. The first software, *BWA-MEM* (Li, 2013), is a mapping tool widely used in NGS studies, while the other three software described are SV callers, i.e., tools used to call potential SVs using their own methods and algorithms. The three tested SV callers where *BreakDancer* (Chen et al., 2009), *Wham* (Kronenberg et al., 2015) and *GRIDSS* (Cameron et al., 2017).

BWA-MEM

BWA-MEM (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. Its name stands for Burrows-Wheeler Alignment (BWA) of maximal exact matches (MEM) and it was developed due to some flaws and difficulties that some of the existent tools at the time were facing.

With the appearance of NGS techniques, tools were developed to align the sequences produced against reference genomes. However, by then the sequence reads were about 36 bp in length, and for such small read length it makes sense to require that all bases from the sequence reads align to the subject (end-to-end alignment).

However, with the constant increase on sequence reads length, brought by the evolution of the NGS techniques, a new challenge was presented to alignment tools, consisting on the need of a more comprehensive algorithm. In order to respond to this, some new tools were developed, with the goal of better adjusting the aligning algorithm to the new reads requirements.

Some of the aligning algorithms developed included the BWA using Smith-Waterman's (SW) alignment, *BWA-SW* (Li and Durbin, 2010), *Bowtie2* (Langmead and Salzberg, 2012) and *Cushaw2* (Liu and Schmidt, 2012). The three were developed as improved versions of previous tools, having the ability to map longer reads as their main innovative feature.

BWA-MEM implements two distinct read mapping methods, one suited to SE sequencing reads and the other to perform the alignment on PE reads. The PE method can be considered as a continuation of the SE one, since the SE is always made and, if the query consists on PE sequences, the second method is performed (Li, 2013).

The SE method is composed by three steps, namely seeding and re-seeding, chaining and chain filtering, and the seed extension, whereas the PE method consists of only two steps, the one where missing hits are rescued and the pairing one.

The seeding and re-seeding step consists on seeding an alignment with supermaximal exact match (SMEM), where, for each query position, the algorithm tries to obtain the longest exact match covering the position. However, the existence of SMEMs is not guaranteed to happen in the true alignment. So, in order to reduce the number of mismappings

due to missing seeds, it is necessary to re-seed. This is accomplished by using the longest exact matches that cover the middle base of the SMEM and occur at least 1 more time in the genome than the original SMEM.

The next step is chaining and chain filtering. A chain is a group of seeds that are collinear and close to each other. The seeds are chained during the seeding process and then the short chains that are largely contained in a long chain and are simultaneously 50% and 38 bp shorter than the longest ones are filtered out. This chain filtering has the goal of reducing unsuccessful seed extension at the next step.

Finally, the last step of the first method is seed extension. In this step, seeds are ranked by the length of the chain they belong to and by seed length. Then, from best to worst, seeds are dropped if they are already contained in an alignment found before, otherwise they are extended as a potential new alignment. Here, *BWA-MEM* differs from the standard seed extension, avoiding extension through poorly aligned regions with good flanking alignment and trying to keep track of the best extension score reaching the end of the query sequence. If the difference between this score and the best local alignment score is below a threshold, the local alignment will be discarded. This aims to automatically choose between local and end-to-end alignments (Li, 2013).

To perform the PE method requires the execution of the previous steps and then proceed with the rescue of the missing hits. In *BWA-MEM*, batches of reads are processed at a time. For each one, the mean μ and the variance v^2 of the insert size distribution from trusty SE hits are estimated. For the top 100 hits of either end, if the mate is unmapped in a window $[\mu - 4v, \mu + 4v]$ from each hit, the software performs SSE2-based SW alignment (Farrar, 2006) for the mate within the window. In order to detect potential mismapping in a long tandem repeat, the second best SW alignment score is recorded. The SW alignment rescuing, along with hits found in SE method, will be used for pairing.

The last step is pairing, when the software attributes scores to a given pair. The scoring process takes into account the SW alignment score of the two hits from the pair, their matching score and the probability of observing an insert size longer than the distance between the hits assuming normal distribution. In the end, *BWA-MEM* chooses the pair that maximizes this score as the final alignment for both ends.

BreakDancer

BreakDancer (Chen et al., 2009) is a software package for SV detection that was developed to address the need for an efficient way to call SV, using CL as the calling technique. This need emerged due to the appearance of NGS technologies, which have made PE WGRS much cheaper, providing larger data volumes to analyze.

BreakDancer is divided into two algorithms, *BreakDancerMax* and *BreakDancerMini*. The former can detect INS, DEL, INV, and both ITX and CTX, while the latter focuses on

detecting small InDels (between 10-100 bp) that are normally not detected by *BreakDancerMax*.

BreakDancerMax takes mapping results as input and the read pairs mapped to the reference genome with a minimum MAPQ are used. Based on the separation distance and alignment orientation between the reads within a pair, the given threshold (specified by the user) and the empirical insert size distribution estimated from the alignment of each library contributing genome coverage, the read pairs are then classified into six types, defining whether they are normal, deletion, insertion, inversion, inter-chromosomal translocation or intra-chromosomal translocation (Chen et al., 2009).

Then, the algorithm tries to find genomic regions with more anomalous read pairs than expected on average and, if one or more regions are interconnected by at least two anomalous read pairs, a putative SV is derived. The algorithm output includes the SVs type, size and their start and end coordinates.

The SVs are determined by the dominant type of associated anomalous read pairs in a particular region, supported by a confidence score for each variant. This score is calculated taking into account the number of supporting anomalous read pairs, the size of the anchoring regions and the coverage of the genome.

On the other hand, the size is estimated by subtracting the mean insert size from the average spanning distance in each library and then averaging across libraries. Finally, the start and end coordinates are defined as the inner boundaries of the SVs constituent regions that are closest to the suspected breakpoints.

With regard to the second algorithm, *BreakDancerMini*, it takes the normally mapped read pairs ignored by *BreakDancerMax* and analyzes them. Genomic regions of size equivalent to the mean insert size are then classified as either normal or anomalous based on a sliding window test. The window slides along the read pairs and compare the separation distances between the read pairs within the window against those in the entire genome. As in *BreakDancerMax*, putative SVs are derived from anomalous genomic regions that are interconnected by at least two common read pairs, and the output is similar, since it includes the SVs type, size and their start and end coordinates. The SVs type is determined by a confidence score based on the significance value of the sliding window test. The size is also estimated by subtracting the mean insert size from the average spanning distance in each library, and then averaging across libraries, as in *BreakDancerMax*. On the other hand, start and end coordinates are defined as the outer boundaries of the constituent regions, contrary to what happens in *BreakDancerMax* (Chen et al., 2009).

Wham

Whole-genome Alignment Metrics (Wham) (Kronenberg et al., 2015) was created to address some of the problems related with the identification of SV, such as the high false

positive and false negative rates of several SV callers, the high variability of SVs breakpoints, which make difficult to detect an association between a phenotype and a complex ensemble of overlapping SVs, and the absence of a SV caller capable of identifying SV enrichment in cases vs. controls studies.

The *Wham* suite consists of two programs, *wham* and *whamg*. The former is the original tool, a very sensitive method with a high false discovery rate. The latter is an improved version of *Wham*, being more accurate and better suited for general SV calling. The basic implementation behind both is very similar, therefore no distinction will be made between them.

Wham is designed for PE Illumina libraries with standard insert sizes (∼300-500 bp). The prediction of SV breakpoints integrates both mate-pair (MP) mapping, SA, soft-clipped (SC) reads re-alignment, alternative alignment and consensus sequence based evidence as calling techniques, producing results with single-nucleotide accuracy (Kronenberg et al., 2015).

To identify the SV breakpoints, a combined pileup for all Binary Alignment/Map (BAM) files provided is generated. Reads are then hashed by position to identify shared breakpoints, and putative SVs are inferred from positions in the pileup where three or more reads share the same breakpoint. The reads for each breakpoint are aligned in order to form a consensus sequence. If this sequence is shorter than 10 bp or contain more than 50% mismatches in the alignments, the breakpoint is not reported, as it is more likely to be a consequence of mapping errors than a result of SV. In addition, overlapping alleles that do not share the same breakpoints are reported as independent records, and different alleles with the same breakpoints that fail the mismatch consensus filter are discarded.

For the breakpoints not present in the initial pileup position, *Wham* uses SA, MP positional information and alternative alignment to find them. It also processes the "SA" and "XA" tags from the BAM files to identify shared positions as candidate endpoints of the reported SV. Then, all the candidate breakpoints are clustered and their positions rounded to the nearest tenth bp, reporting the position with the highest read support. If possible, the consensus sequence will be aligned to the putative breakpoint region using the information from the SW alignment. This will result in a breakpoint improvement to the location of the consensus sequence alignment (Kronenberg et al., 2015).

SVs bigger than 1Mb can be highly deleterious genomic aberrations, hence additional filtering is required. Therefore, in these cases, the other breakpoint, not present in the initial pileup position, must have at least two reads supporting the exact breakpoint. Furthermore, in the case of translocations, the SV is discarded if the split reads in the pileup map to more than three different chromosomes, in order to remove many false positive calls. These false positive calls result from inter-chromosomal mapping errors introduced by repetitive sequences.

The classification of the SV type is made with resource to a random forest of decision trees. The resulting output is a Variant Call Format (VCF) file, which includes the relevant information for each call, which includes SV type, SV length, end position, among others, for the INFO field, and the tags ”Genotype”, ”Read depth” and ”Per sample SV support”, for the FORMAT field.

GRIDSS

The *Genomic Rearrangement IDentification Software Suite (GRIDSS)* (Cameron et al., 2017) is a novel approach to predict genomic rearrangements from DNA sequencing data, providing high-speed SV calling using a combination of assembly, SA and read pair support, achieving high sensitive and specific results.

Split-reads alignment based methods, where *GRIDSS* is included, have the ability of identifying of the exact breakpoint location, having single nucleotide resolution. This is achieved resorting to SA identification, either through direct mapping by the aligner, re-alignment of soft-clipped bases or SA of the unmapped read in one-end anchored read pairs.

GRIDSS takes a three-step approach to maximize sensitivity and prioritize calls into high or low confidence, therefore maintaining specificity in the high confidence call set. First, properly aligned reads are filtered out, keeping only the reads that might provide any evidence for underlying genomic rearrangements. Then, these remaining reads are assembled using information from the alignment to constrain this assembly, a process called genome-wide break-end assembly, since each contig corresponds to a break-end and the underlying breakpoint and partner break-end are identified only after the assembly. Lastly, a probabilistic model combining break-end contigs with split reads and discordantly-aligned read pairs (DP) evidence is applied, with the objective of scoring and calling SVs (Cameron et al., 2017).

Genome-wide break-end assembly is performed using a new approach developed by extending a positional de Bruijn graph data structure. Unlike a traditional de Bruijn graph, originally developed for small indel and base calling error correction of *de novo* assembly contigs (Ronen et al., 2012), this new approach is a directed acyclic graph due to the addition of positional information to each node. It also reduces depth of coverage needed and memory requirements for accurate assembly, making it computationally efficient at the genome scale.

Regarding the probabilistic model, the quality of predicted SVs is scored according to the Phred-scale probability of originating from the mapped locations without underlying any SVs. The Phred score Q of a probability P is given by $Q = -10\log_{10}(P)$ (Cameron et al., 2017).

So, SVs are scored according to the level of support given by both SA, DP and assembly evidence combined. With each piece of evidence being considered independent, and considering that evidence scores are expressed as Phred scores, the sum of the scores of evidence supporting the SVs breakpoints represents the final score of the SVs called. This approach can result in reads providing support for more than one breakpoint. To fix this problem, each piece of evidence is assigned to only the highest scoring SV it supports.

2.3 CORK PRODUCTION AND ITS BIOLOGICAL FUNCTION

The reason why cork oak trees play an important role in the economy of several Mediterranean countries economy, with a vast amount of industries using them, is their phellem, or cork tissue, as it is better known. It is a renewable natural resource that has many applications, particularly in the production of stoppers, and sound and thermal insulators, among others. This is due to the unique features of cork, as it is light to the point of floating, waterproof, isolator, flexible, compressible, hypoallergenic and resistant to friction, temperature and time wear. It is extracted by peeling off the cork layer from adult trees at regular intervals of at least 9 years. Cork is one of the most valuable non-wood forest products, putting this species among the most important trees with commercial relevance in the countries where it is naturally distributed.

The thick cork layer works as a protective multi-layered dead tissue, composed of suberized cells located within the periderm of the bark system's outer face and produced towards the outside (Pereira, 2011; Teixeira et al., 2014; Verdaguer et al., 2016). It is formed as the end result of meristematic activity of the cork cambium, a specialized phellogen tissue, followed by both cell expansion and cell wall deposition of suberin and waxes, and, lastly, with an irreversible program of senescence resulting in cell death (Soler et al., 2007).

During the first year, the phellogen originated in the subepidermis differentiates, leading to the formation of phelloderm cells towards inside and of phellem cells to the outside. Subsequently, this meristematic layer sustainably produces cork (phellem) cells which form a thick continuous layer of cells covering the tree trunk, roots and branches. A new thick layer of suberized cells is formed every year, accumulating with those of previous years in the form of annual rings (Caritat et al., 2000). The phellogen is located between the inner bark and cork, with its activity period ranging from April to October (Graça and Pereira, 2004). In the beginning of the spring, it is highly active, starting the production of new cells and, by the time winter starts, this production ceases (Silva et al., 2005).

The cork tissue's peculiar properties, mentioned above, are due to the chemical composition of its cells walls, with suberin as the main component, representing about 50% of the total material present in the cork oak's periderm, and lignin as the second most important one (Pereira, 2011; Marques and Pereira, 2013; Lourenço et al., 2016). Differ-

ential expression studies revealed that four main secondary metabolic pathways, namely the synthesis of acyl-lipids, phenylpropanoids, isoprenoids, and flavonoids, are related to cork formation and suberin biosynthesis (Soler et al., 2007). Also, the biosynthesis of both suberin and lignin derives from the general phenylpropanoid pathway and shares the same basic reactions (Rahantamalala et al., 2010).

Cork's biological function is to protect the internal tissues and prevent water loss. As it is a thermal isolator, the plant's inner tissues are protected against environmental conditions, whether it is heat or cold. Its waterproof property confers the ability to maintain a regular amount of water in the plant, resisting to possible extreme weather conditions. On the other hand, lenticels, small regions in the periderm of most plants, provide aeration to the inner tissues. Lenticels are made up of relatively loosely arranged cells and formed by the lenticular phellogen, a specific zone of the phellogen. These areas contain many intercellular open spaces which, along the years, by radial extension from the phellogen to the external surface of the periderm, form the lenticular channels, responsible for the mentioned aeration (Pereira, 2011).

Several features which determine the CQ can be observed and measured to classify individuals according the quality of the cork they produce. Experienced operators, well acquainted with cork, are able to determine the CQ just by visually inspecting the samples (Figure 3). However, a quantitative analysis of these features can be made, to give scientific support to this selection. The features usually evaluated to determine whether a sample of cork has good or bad quality include the cork tissue homogeneity, thickness and porosity level (Pereira et al., 1996), as well as the number of times the crop was harvested (Soler et al., 2007; Teixeira et al., 2014).

Good quality cork is defined as being an homogeneous tissue, with higher thickness and a low porosity level (Figure 3A), with the latter being determined using a porosity coefficient, defined as the area of pores (or lenticular channels) in percentage of total area (Pereira et al., 1996), the number of pores and the mean pore area, features observed in transverse cuts of the cork tissue (Teixeira et al., 2014). The "virgin cork" corresponds to the first cork layers produced by the original phellogen of the cork oak tree, which are removed from the tree at an age between 20-30 years, being a bad CQ harvest (Pereira et al., 1987). After that, new cork layers are removed every 9 years, with the second harvest still presenting bad CQ (Figure 3B) and the "reproduction cork", which corresponds to cork layers removed from the tree from the third harvest onwards, being the one able to present good CQ (Pereira et al., 1987). So, the combination of all these quantitative features determines the qualitative cork quality classification.

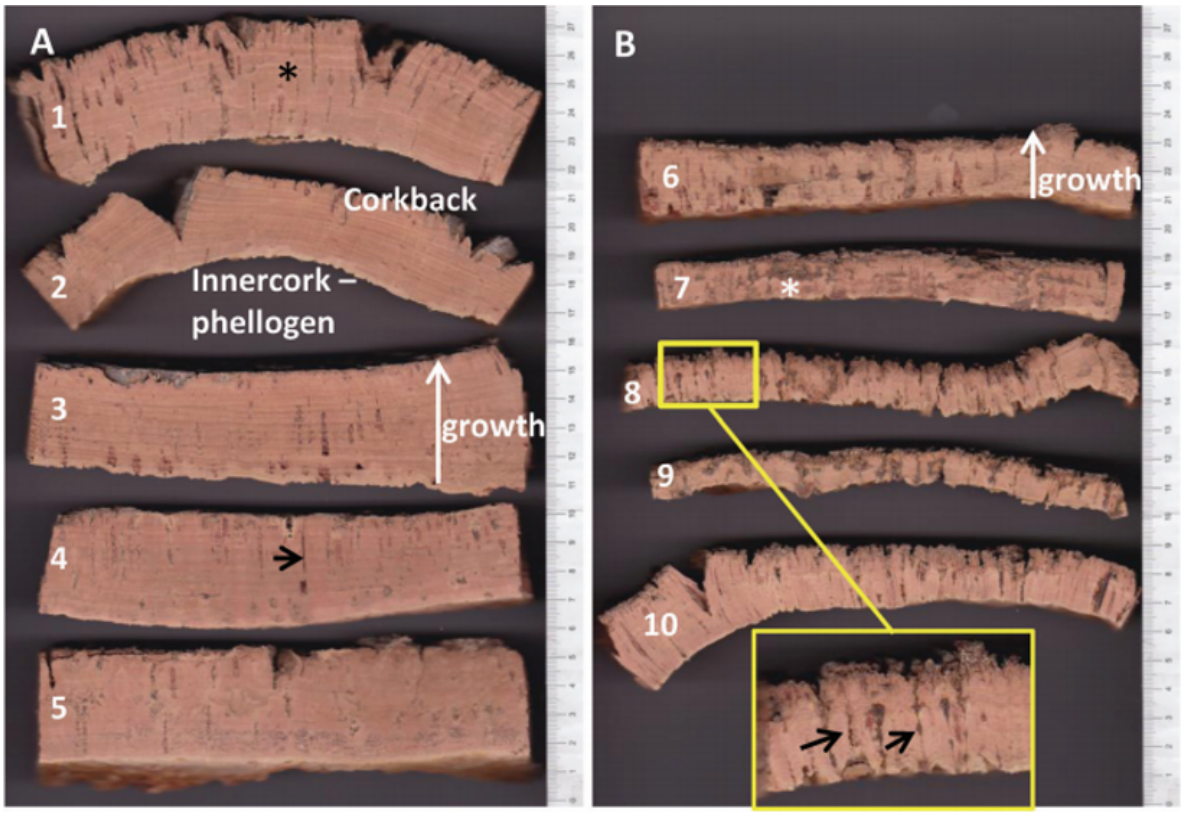


Figure 3: Transverse cuts of cork planks with (A) good CQ and (B) bad CQ where the radial growth of the suber layer (white arrow) can be observed. Planks in (A) show a a minimum thickness of 28mm with a reduced number of lenticels (arrow in plank 4). Cork rings of growth are clearly visible (* in plank 1). Planks in (B) show a maximum thickness of 17mm with a high number of lenticels (arrows in plank 8), as well as other cork defects (* in plank 7). Adapted from Teixeira et al. (2014).

2.4 GROUP'S PREVIOUS WORK

At CEBAL, the Animal Genomics and Bioinformatics group has been the leader of the bioinformatics tasks related to the cork oak genome sequencing project (Genosuber), which incorporates a variety of sequence data generated using different high-throughput sequencing platforms, as well as many different types of bioinformatics analyses. The draft sequence and annotation of the cork oak genome was finalized in August 2017, and the generation of the final version of the genome is ongoing. In addition to the work being developed in cork oak, projects in pigs, stone pine and wheat are also presently underway.

In cork oak, many bioinformatics analyses pipelines have been developed, which integrate all the steps related with data pre-processing, genome and transcriptome de novo assembly, genome and transcriptome annotation, read mapping, SNP identification, detection of differential expression and discovery of alternative splicing events. The identification of structural variation in cork oak is a logical step on the groups strategy to develop and

implement a wide variety of bioinformatics tools for analysing high-throughput sequence data developed in this species.

This effort is carried within the frame of research projects where the group collaborates with other research teams, as well as producers and the cork industry, in order to identify genetic markers associated with the more relevant economic traits in cork oak.

3

METHODS

Since the underlying methods for SV calling have a standard base guideline, the pipeline created for this work (Figure 4) is based on the standard procedures described in the previous chapter. The WGRS (whole genome resequencing) data is submitted to a quality evaluation step, followed by a preprocessing to remove the bad quality reads. Then, the reads are mapped to the reference genome, with SVs being identified by SV callers. Finally, the SVs are annotated in order to evaluate possible associations between the SVs and the CQ phenotypes.

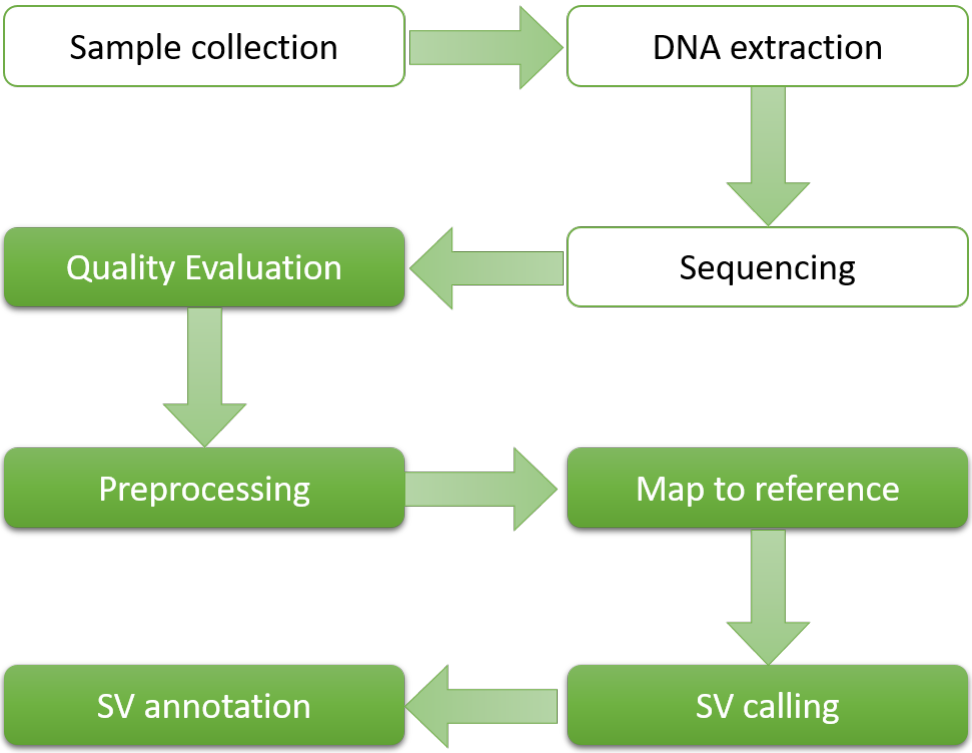


Figure 4: Thesis pipeline with the sequence of steps performed. White background steps were performed by others, while the green background ones were performed by the author of the thesis.

3.1 SAMPLE COLLECTION AND HIGH-THROUGHPUT SEQUENCING

The cork oak trees analysed in this study are located in the Herdade dos Leitões, Montargil, Ponte de Sor, Portugal. A total of 30 trees were sampled, which included 14 trees producers of good quality cork, along with 16 trees that produce cork with bad quality. The selection and assignment of the trees to each of the cork quality groups was based on the historical records available for each tree, routinely collected by Herdade dos Leitões. Leaf samples were collected from each tree and DNA was extracted using a standard DNA extraction protocol. High-throughput sequencing was performed in the Illumina HiSeq X10 platform, at the Beijing Genomics Institute, using a paired-end protocol and a read length of 150 bp.

3.2 HIGH-THROUGHPUT SEQUENCE DATA QUALITY EVALUATION

The quality of the WGRS reads from the 30 sequenced libraries was evaluated using FastQC software, version 0.11.5 (Andrews, 2010). FastQC is a quality evaluation tool for high throughput sequence data, performing a wide variety of analysis on the quality of the sequenced libraries, which include basic statistics, per base sequence quality and overrepresented sequences, the more relevant in SV studies.

The FastQC report for all libraries was carefully analyzed, including read length, encoding of quality scores and per base sequence quality. The distribution of the quality scores along the read length is shown in Figure 5. The interpretation of this graphic consists of identifying at which position in the read the quality score gets below a chosen threshold. With the quality and length thresholds defined, the data were then ready to enter the preprocessing step.

In order to decide which length and quality thresholds to use, we tested four combinations of thresholds, which included 110 bp length threshold and 20 quality threshold, 120 bp and 20, 130 bp and 20, and finally 110 bp and 15. The goal was to identify which combination was sufficiently restrictive to discard the bad quality reads but, at the same time, was capable of maintaining a high number of reads.

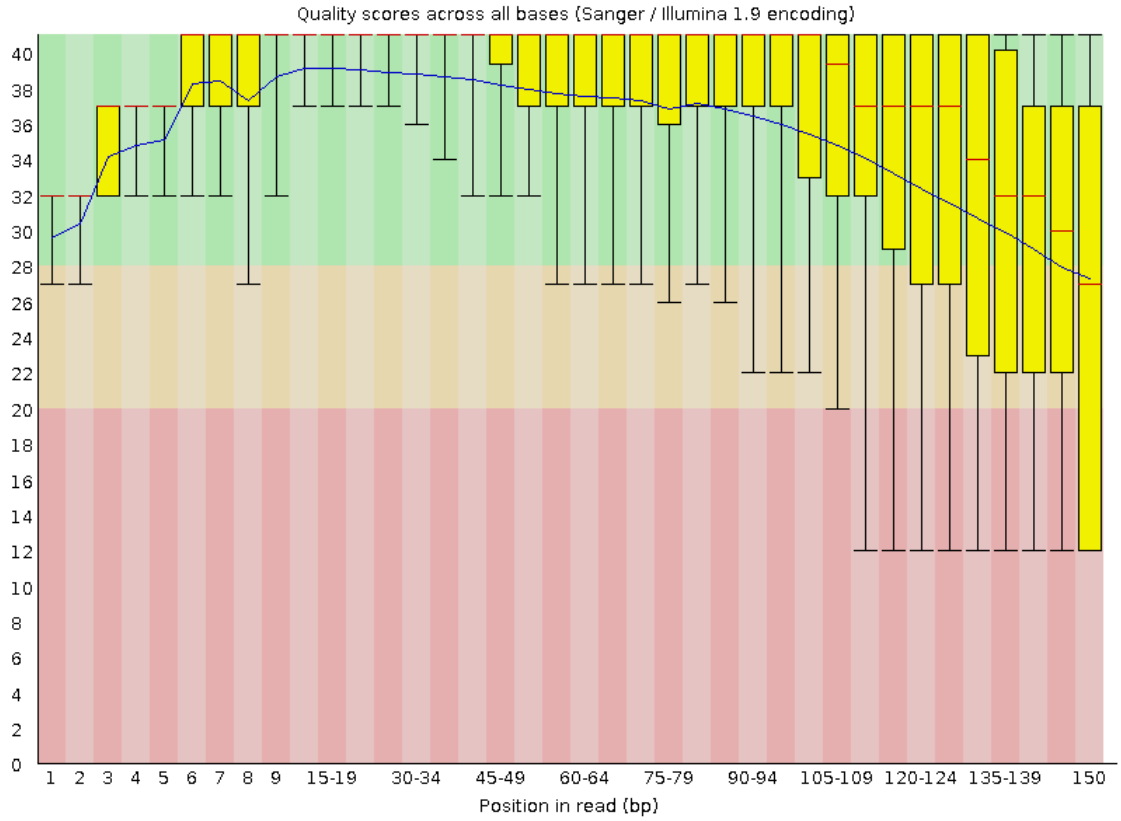


Figure 5: WGRS files quality evaluation using FastQC

3.3 Preprocessing of high-throughput sequence data

The quality and length thresholds determined in the quality evaluation step were used as trimming input parameters. Trimming was accomplished using Sickle version 1.33 (Joshi and Fass, 2011), a tool that resorts to sliding windows which, along with the given thresholds, will trim the reads when quality does not fit its threshold and discard the reads smaller than the length threshold.

Each window is 10% of the read length, and it starts sliding until the average quality within it rises above the threshold. Then, the algorithm determines where the rise occurred and cuts the read there for the 5'-end cut. The window keeps sliding until the average quality in the window drops below the threshold. When this happens, the algorithm determines where the drop occurred and the read is cut there for the 3'-end cut. If the remaining sequence is smaller than the length threshold, the read is discarded entirely.

The preprocessing step guaranteed that the remaining reads had high quality and length, generating a higher quality dataset for downstream analysis.

3.4 READ MAPPING TO THE REFERENCE GENOME

Upon completion of the preprocessing stage, the reads were then ready to be aligned to the reference genome. The reads were mapped to the reference genome using BWA-MEM (Li, 2013), a software extensively described in the previous chapter. The sequences used as reference genome the draft version of the cork oak genome, generated by the Animal Genomics and Bioinformatics group at CEBAL, as part of the cork oak genome sequencing project (Genosuber). The cork oak draft genome contained 23,344 scaffolds greater than 1,000 bp, which represented an assembly length of 953.3 Mb. The N50 observed was 465.2 Kb, while the longest scaffold was 2,284,287 bp in length. There were 79,752 genes annotated for the cork oak draft genome, for which 83,814 transcripts were identified. Functional annotation was determined for 69,218 transcripts, which represented 82.6% of the total number of transcripts.

The first step consisted in building the index over the reference genome, which was given as input to the software, along with the thirty PE libraries. Initially, three different sets of parameters were tested,in order to produce higher quality mapping results. The sets were: (1) all the parameters with their default values; (2) a minimum seed length of 40, instead of the default value of 19, and all the other parameters with their default values; and (3) a mismatch penalty of 8, in opposition to the default value of 4, and again all the other parameters with their default values. Since no significant differences between the read datasets derived from each individual were expected, and considering that it would be very time consuming and computationally more demanding to test the three sets of parameters in all libraries, the comparison was not performed using all the 30 libraries. Therefore, a subset of four libraries, corresponding to two individuals with bad CQ and two others with good CQ, was used, so that the comparison was made in a reasonable time frame.

The software outputted a Sequence Alignment/Map (SAM) file, which was then converted into the BAM format, since it is more efficient to work with binary files and it also saves disk space. After that, the file was sorted by genomic coordinate to facilitate other tools to access the data. Both file format conversion and sorting were performed using SAMtools (Li et al., 2009), a set of tools for manipulating alignments in SAM/BAM/CRAM format. Mapping statistics were then collected also using SAMtools, in particular the SAM bitwise flags, a data structure used in computer programming, which facilitates counting some specific types of results, such as the number of mapped/unmapped reads, chimeric alignments or proper pairs aligned.

3.5 STRUCTURAL VARIATION CALLING

Before mapping the whole WGRS dataset to the reference genome, an evaluation and subsequent selection of the SV calling software was carried out, since mapping parameters and SV software depended on each other. For that purpose, three software tools were tested together with the subsets of mapping parameters, making it a three software by three parameter sets comparison, where just one of the nine sets was chosen. For the same reasons mentioned before in the read mapping section, the same subset of four libraries was used in this comparison.

BreakDancer (Chen et al., 2009), *GRIDSS* (Cameron et al., 2017) and *Wham* (Kronenberg et al., 2015) were used for SV calling for the four libraries. The three software were chosen according to some criteria that allowed the comparison of tools with different approaches to SV calling and integrated in different chronological and technological contexts.

The differences between the three software begin in the types of SV each one can detect. All of them identify INS, DEL, INV and ITX, but while *BreakDancer* calls CTX, *GRIDSS* and *Wham* do not and, on the other hand, the latter two software call for DUP and *BreakDancer* does not. Furthermore, the reads used in the SV detection differed between them. Whereas *BreakDancer* only uses DP (discordantly-aligned read pairs) to both discover and validate SVs, *Wham* uses RD (read-depth) evidence and SC (soft-clipped) reads, and *GRIDSS* uses SC reads, OEA (one-end anchored) reads and DP support.

Concerning the SV calling techniques used by each one tool, *BreakDancer* uses CL and a statistical test, *Wham* uses MP mapping, SA, SC reads re-alignment, alternative alignment and consensus sequence based evidence, whilst *GRIDSS* makes a genome wide break-end assembly, SA and a probabilistic model to score the calls.

As noted, the three software have many differences between their methods, mainly when comparing *BreakDancer* with the other two. This can be explained by the chronological and technological context that existed at the time they were developed. *BreakDancer* is much older than the others, when the sequencing technologies were less advanced. Hence, with the improvement of these technologies, software had to be improved as well, containing new upgraded approaches to better fit the new requirements. This emphasizes the importance of comparing different software with different approaches.

Considering the number of SVs identified for the subset of four libraries by each software, along with the mapping statistics and resulting number of SVs identified using each of the mapping parameters subsets, the combination of *GRIDSS* with a mismatch penalty value of 8 was chosen. Detailed information on the reason why this combination was selected among the nine tested will be provided in the next chapters. Next, SV calling was performed for all the 30 libraries using *GRIDSS* and parameters selected, which generated an output consisting of a VCF file of break-end points.

3.6 IDENTIFICATION AND CHARACTERIZATION OF THE STRUCTURAL VARIANTS

To identify SVs in the *GRIDSS* output file, which is a VCF file of break-end points, the information needed to be translated into simple SV type. This was accomplished using an R (R Development Core Team, 2008) script available in the *GRIDSS* repository. The results were then filtered by quality, and calls with score less than 500 were discarded, following the recommendations of the *GRIDSS* (Cameron et al., 2017) developers, which mention that such calls are unreliable. In addition, each call is a breakpoint consisting of two break-ends, with one being the reciprocal record of the other, but, although VCF format required both break-end to be written as separate records, each record fully defines the call. Therefore, in order to have a single record for each call, only the first record for each call was kept, and subsequently counting of the remaining SVs was performed.

After completion of the identification and counting of the SVs, their positions were compared with the ones of the transcripts from the reference genome annotation, to check if some of the SVs were located in cork oak transcripts. The SVs located in transcripts were compared with the functional annotation of the reference genome, to determine whether there were some SVs affecting genes believed to be involved in cork production and quality. Both comparisons were made in two separate sets, with good CQ individuals on one side and bad CQ individuals in the other one, making it easier to distinguish SVs exclusive for each set.

4

RESULTS

In this section, we present the results for the comparisons of thresholds for read quality and length, read mapping parameters, and software for SV calling, together with the results of the identification and characterization of SVs using the whole WGRS dataset.

4.1 COMPARISON OF THRESHOLDS FOR READ QUALITY AND LENGTH

The results regarding the testing of different read quality and length thresholds are indicated in Table 2, which includes the comparison between the number of reads originally present in the WGRS dataset and the number of reads that met the requirements for each combination of quality and length thresholds, as set in section 3.2.

These combinations were chosen as a result of the individual analysis of each FastQC report, namely the analysis of the per base sequence quality section, where it was verified that the quality of the data were assured with a length between 110 bp and 130 bp and a quality score about 15 or 20.

A quality threshold of 20 resulted in the loss of a significant amount of reads. The percentage of reads removed from the dataset exceeded 10% for all the length thresholds, namely 12.3% for a 110 bp length, 15.9% for 120 bp and 20.8% for 130 bp. As expected, an increase in the length threshold was leading to a higher percentage of reads removed.

The combination of 110 bp length and 15 quality thresholds generated an average of 94.9% of reads passing the filters, i.e., with both length and quality above the thresholds. Given that 15 is still a very reasonable quality threshold that does not compromise the quality of the downstream analysis, and considering the higher percentage of reads kept, this was the combination of parameters selected for the preprocessing of the WGRS dataset.

Table 2: Comparison of thresholds for read quality and length. The initial number of reads is presented, as well as the number of reads obtained after filtering with each set of thresholds and the corresponding percentage of reads kept relative to the initial number of reads, for all 30 individuals, with average values in the bottom of each column.

Individual	Number of reads	Q>=15, L>=110		Q>=20, L>=110		Q>=20, L>=120		Q>=20, L>=130	
		Number	%	Number	%	Number	%	Number	%
1	76,732,856	72,798,182	94.9	67,325,333	87.7	64,737,271	84.4	61,249,717	79.8
2	82,510,343	78,322,432	94.9	72,466,056	87.8	69,893,376	84.7	66,329,884	80.4
3	80,189,032	76,358,396	95.2	70,954,860	88.5	68,202,387	85.1	64,510,690	80.4
4	89,286,989	84,286,926	94.4	76,932,447	86.2	73,266,309	82.1	68,392,116	76.6
5	80,166,417	75,763,277	94.5	69,630,397	86.9	66,325,239	82.7	61,732,728	77.0
6	85,968,604	81,873,521	95.2	76,130,525	88.6	73,416,840	85.4	69,745,306	81.1
7	66,916,625	63,682,077	95.2	58,953,755	88.1	56,548,333	84.5	53,303,645	79.7
8	76,337,743	72,334,694	94.8	66,638,419	87.3	63,864,933	83.7	60,197,881	78.9
9	92,491,435	88,065,302	95.2	81,507,030	88.1	78,269,669	84.6	73,842,487	79.8
10	113,508,517	107,652,036	94.8	99,486,157	87.6	95,448,809	84.1	90,090,353	79.4
11	89,927,413	85,614,390	95.2	79,525,754	88.4	76,302,142	84.8	71,769,354	79.8
12	78,316,133	74,465,630	95.1	68,837,752	87.9	66,139,219	84.5	62,448,401	79.7
13	74,562,656	71,371,400	95.7	66,679,900	89.4	64,169,344	86.1	60,653,165	81.3
14	87,346,155	83,457,422	95.5	77,732,770	89.0	74,734,052	85.6	70,551,779	80.8
15	84,818,412	80,244,694	94.6	74,110,950	87.4	71,204,146	83.9	67,159,492	79.2
16	85,075,768	80,592,525	94.7	73,784,587	86.7	70,352,266	82.7	65,760,801	77.3
17	86,807,159	82,323,017	94.8	75,984,205	87.5	73,020,447	84.1	69,015,619	79.5
18	77,130,642	73,200,883	94.9	67,691,134	87.8	64,682,497	83.9	60,485,687	78.4
19	91,712,438	87,519,132	95.4	81,549,044	88.9	78,375,614	85.5	73,936,569	80.6
20	75,525,418	71,563,110	94.8	65,927,249	87.3	62,904,087	83.3	58,697,841	77.7
21	87,698,826	83,599,475	95.3	77,510,836	88.4	74,546,159	85.0	70,476,925	80.4
22	84,041,125	79,523,113	94.6	73,309,634	87.2	70,338,143	83.7	66,400,166	79.0
23	94,963,567	89,742,662	94.5	82,457,301	86.8	78,998,908	83.2	74,335,512	78.3
24	81,380,612	77,317,990	95.0	71,653,797	88.0	69,015,774	84.8	65,404,806	80.4
25	75,349,375	71,212,069	94.5	65,271,480	86.6	62,519,706	83.0	58,851,258	78.1
26	88,709,792	83,925,794	94.6	76,845,200	86.6	73,294,308	82.6	68,555,058	77.3
27	96,647,138	91,567,195	94.7	84,500,979	87.4	80,820,597	83.6	75,738,673	78.4
28	107,282,408	102,053,794	95.1	94,754,363	88.3	90,763,554	84.6	85,168,873	79.4
29	92,350,350	87,331,167	94.6	80,266,904	86.9	76,967,322	83.3	72,547,230	78.6
30	79,516,471	75,301,096	94.7	69,400,027	87.3	66,549,914	83.7	62,784,816	79.0
Average	85,442,347	81,102,113	94.9	74,927,295	87.7	71,855,712	84.1	67,671,228	79.2

4.2 Comparison of read mapping parameters

As mentioned in the previous chapter, a subset of four libraries was used to compare the three subsets of mapping parameters. The subset with all the values set to their default is simply referred to as *default*, whereas the one with a minimum seed length of 40 and the other with a mismatch penalty of 8 are named based on their differing parameters,

respectively *k40* and *B8*. Table 3 contains the results of the comparison for all the mapping parameters tested.

The comparison between these three parameter subsets revealed a decrease in the number of mapped reads when comparing both *k40* and *B8* subsets with the *default* subset, particularly in the *k40* subset, where the percentage of mapped reads decreased 5%. On the other hand, the *B8* subset not only presented a 1.4% decrease in the number of mapped reads, but also an increase in the number of chimeric alignments (0.2% more when compared with the *default* subset), which are considered to be a sign of SV presence.

Table 3: Comparison of read mapping parameters. Number of alignments, mapped reads, unmapped reads and chimeric alignments for each subset of mapping parameters are presented for each one of the 4 individuals tested. In addition, percentage of mapped reads and chimeric alignments relative to the number of alignments are presented for each subset of mapping parameters.

Individuals	Stats								
	Number of alignments			Number of mapped reads					
	Default	k 40	B 8	Default	%	k 40	%	B 8	%
1	149,024,116	146,335,603	149,586,926	140,398,597	94.2	131,126,611	89.6	139,041,644	93.0
2	160,925,612	157,571,418	161,198,023	152,912,121	95.0	141,643,832	89.9	150,903,268	93.6
3	156,939,014	153,418,120	157,049,259	148,928,304	94.9	136,396,774	88.9	146,633,597	93.4
4	172,505,053	169,386,043	172,998,539	163,396,414	94.7	153,129,414	90.4	161,347,374	93.3
Average	159,848,449	156,677,796	160,208,187	151,408,859	94.7	140,574,158	89.7	149,481,471	93.3
Individuals	Unmapped reads			Chimeric alignments					
	Default	k 40	B 8	Default	%	k 40	%	B 8	%
1	5,197,767	14,469,753	6,554,720	3,427,752	2.3	739,239	0.5	3,990,562	2.7
2	3,732,743	15,001,032	5,741,596	4,280,748	2.7	926,554	0.6	4,553,159	2.8
3	3,788,488	16,320,018	6,083,195	4,222,222	2.7	701,328	0.5	4,332,467	2.8
4	5,177,438	15,444,438	7,226,478	3,931,201	2.3	812,191	0.5	4,424,687	2.6
Average	4,474,109	15,308,810	6,401,497	3,965,481	2.5	794,828	0.5	4,325,219	2.7

Although these results provided some insight on the relevance each subset could have for SV detection, no subset was selected in this step, since that selection only took place when the results from both the mapping parameters and SV calling software were compared together. Nevertheless, this comparison showed good signs from both *default* and *B8* subsets, which showed an average high percentage of mapped reads (94.7% and 93.3%, respectively) and chimeric alignments (2.5% and 2.7%, respectively), whereas *k40* did not seem to be a solution, since it had lower values (89.7% average percentage of mapped reads and 0.5% average percentage of chimeric alignments) when compared with the latter ones.

4.3 Comparison of software for structural variation calling

In order to find the combination of read mapping parameters and SV calling software producing the best results, each of the three SV calling software was used with each of the

three read mapping parameter sets. For this purpose, the same subset of four libraries used in the previous section was used.

Initially, the total number of SVs identified in all individuals was determined, in order to have an overview of the frequency of this type of variation in the cork oak genome (Figure 6). This analysis was performed without including the phenotypic information for CQ that was available for all individuals. However, considering the relevance of possible associations between SVs and CQ, only the variants exclusively present in one of the phenotypic groups were taken into account in the selection of the software for calling SV to be used with the data derived from the 30 individuals. The information regarding the performance of each software for each read mapping parameter set is indicated in Figure 7, on which only the SVs exclusive to one cork quality group are indicated.

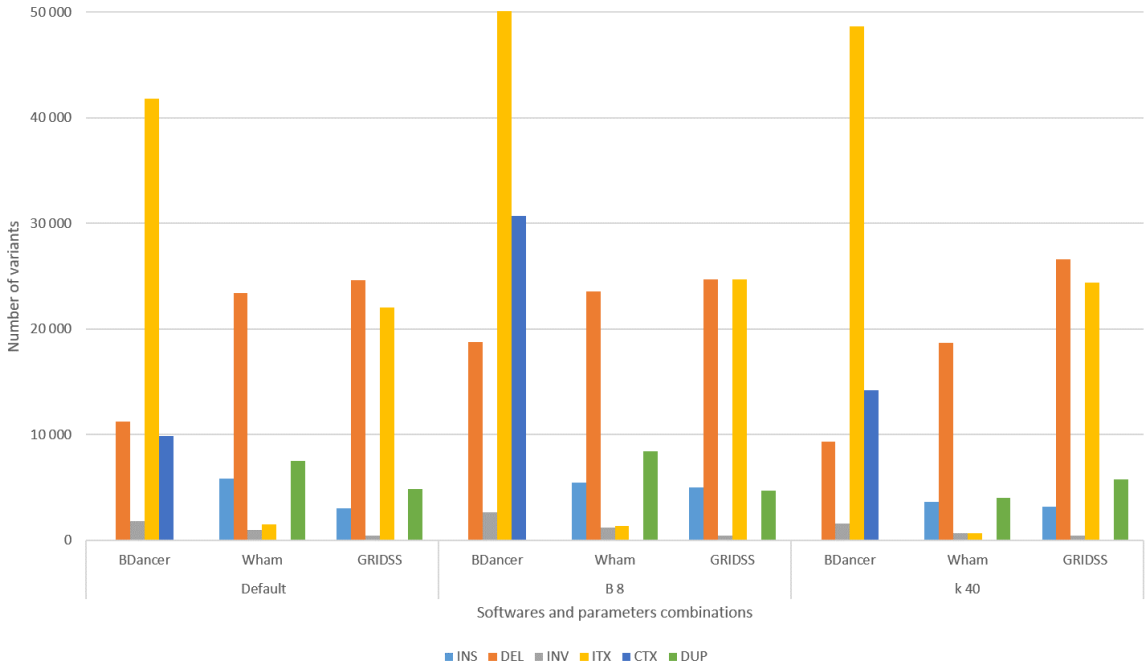


Figure 6: Software and mapping parameters comparison for SV calling using the whole WGRS dataset for the 30 individuals. Each bar represents a type of SV, and the graph is divided in three parts corresponding to the three sets of mapping parameters, which are also divided in three parts, one for each software used.

The results show that *BreakDancer* displayed the largest number of translocations, both ITX and CTX, in each set of parameters, including 9,071 ITX and 2,237 CTX in the *default* set, 7,151 ITX and 3,913 CTX in the *B8* set, and 8,384 ITX and 3,585 CTX in the *k40* set, even though *BreakDancer* is the only software that calls CTX. Moreover, a common pattern was identified for the three software, with few differences between their performances for the different sets of mapping parameters. Also, the *k40* subset provided similar results as

the other subsets of parameters, somewhat contradicting the signs given in the previous section.

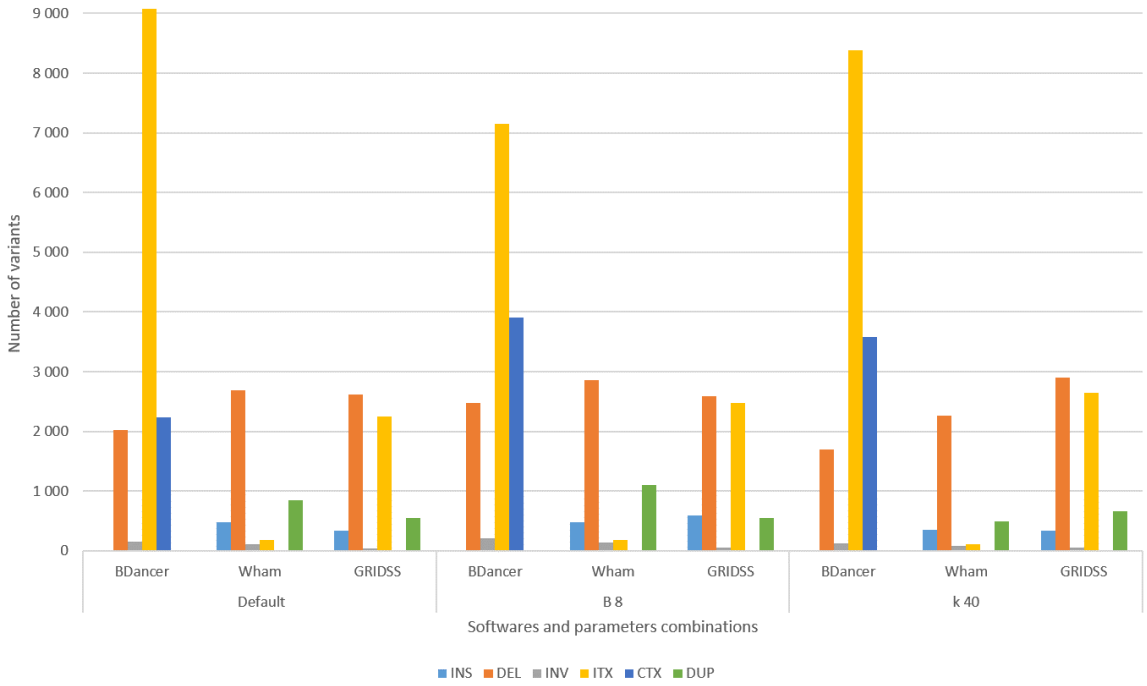


Figure 7: Software and mapping parameters comparison. Only exclusive variants for either good or bad CQ individuals were accounted for this comparison. Each bar represents a type of SV, and the graph is divided in three parts corresponding to the three sets of mapping parameters, which are also divided in three parts, one for each software used.

The counts shown in Figures 6 and 7 included all the predictions (Figure 6) for SVs, as well as the ones exclusive for each phenotype (Figure 7), even if there was only one individual supporting that call. Therefore, to make a more precise selection on the combination of software and mapping parameters to be used, the counts were made using as threshold a minimum coverage of 10, meaning that only SVs with a minimum of 10 reads supporting the call (still with phenotypic exclusivity) were taken into account. These results are presented in Figure 9.

When the minimum number of reads supporting the call threshold was applied, *BreakDancer* had a lower number of SVs being called, with a total of 34 SVs when using both the *default* and the *B8* set, and 15 SVs when using the *k40* set of mapping parameters, whereas the other two software had a better performance, especially *GRIDSS*, which accounted for a total of 516, 565 and 345 SVs called when using, respectively, the *default* set, the *B8* set and the *k40* set. Also, the combinations using the *k40* set of mapping parameters had a visibly decrease in the number of SVs called when compared with the other two,

accounting for a total of 523 SVs called with all software, against the 959 SVs called using the *default* set and the 874 SVs using the *B8* set of mapping parameters.

Hence, due to the higher number of SVs called with more stringent conditions, *GRIDSS* was selected as the software to be used for the analysis of the whole dataset. Moreover, since the *k40* subset has been discarded, the combinations of *GRIDSS* with *default* mapping parameters and *B8* mapping parameters were further analyzed. Since the *B8* set of mapping parameters showed a lower percentage of mapped reads and yet a higher number of SVs called (565 using the *B8* set against 516 when using the *default* set), the combination of *GRIDSS* with the *B8* set of mapping parameters was selected.

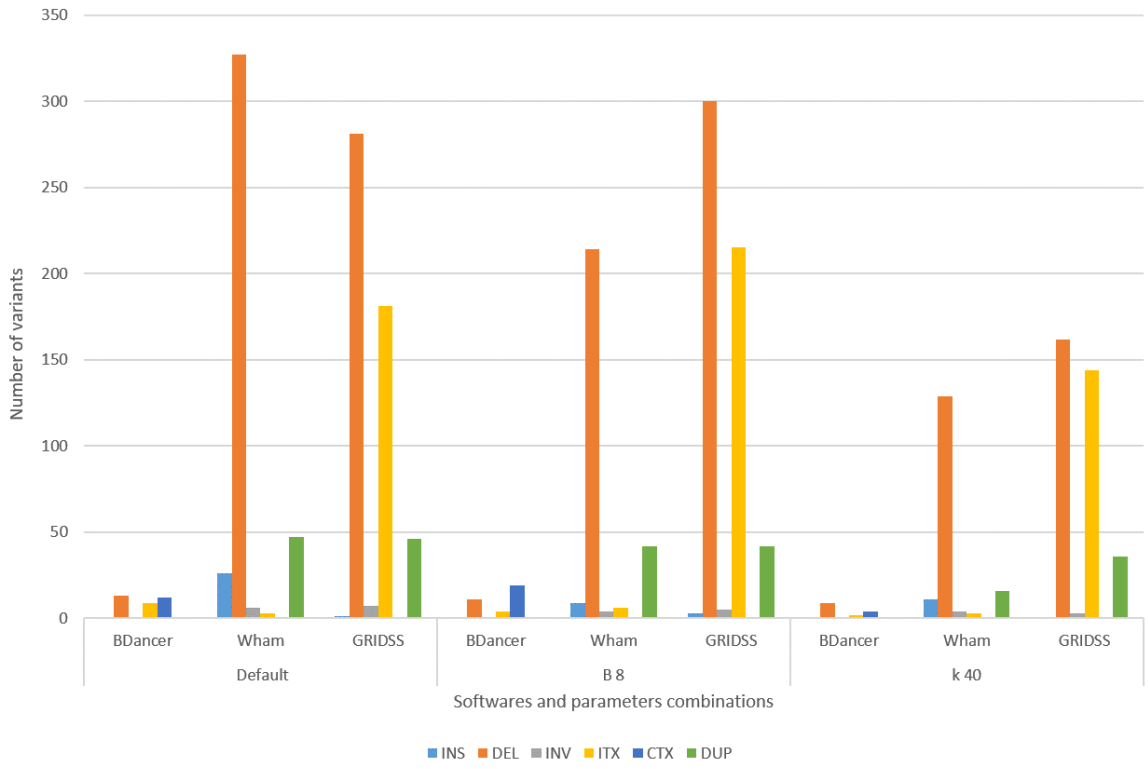


Figure 8: Software and mapping parameters comparison where all the SVs have a minimum coverage of 10 reads. Only exclusive variants for either good or bad CQ individuals were accounted for this comparison. Each bar represents a type of SV, and the graph is divided in three parts corresponding to the three sets of mapping parameters, which are also divided in three parts, one for each software used.

4.4 IDENTIFICATION AND CHARACTERIZATION OF THE STRUCTURAL VARIANTS USING THE WHOLE WGRS DATASET

Using the combination of mapping parameters and software to use, all the libraries were mapped to the reference with the selected parameters and, after that, SV calling was performed using *GRIDSS*.

The total number of SVs identified in the genomes of the 30 cork oak trees included in the WGRS dataset was 93,980. The more frequent type of SVs were deletions (DEL), which accounted for 38,865, followed by translocations (ITX), with a total number of 34,745 variants. On the other hand, the least common SVs were inversions (INV), for which 654 variants were detected. This information is summarized in Table 4.

Table 4: Number of SVs identified for each SV type in the genomes of the 30 cork oak trees included in the WGRS dataset.

Structural variation types	Number of SVs identified
Insertion	10,990
Deletion	38,865
Inversion	654
Duplication	8,726
Translocation	34,745
Total calls	93,980

The SVs exclusive for each phenotype (bad and good CQ) were determined, considering the biological and economical importance of identifying SVs associated with CQ. For this purpose, two approaches were followed, having in consideration the number of reads per individual supporting the SVs. The first approach considered all the individuals with at least 1 read supporting the call, whereas the second one considered only the individuals with at least 10 reads supporting the SV (Table 5). In each approach, two sets of SVs were produced, taking the number of individuals supporting the SVs into account.

In the first approach, a minimum coverage of a single read per individual supporting the variant was considered, and all the SVs exclusive for a phenotype were counted, regardless of the number of individuals supporting it. This meant that at least one individual supported the call and no individual with the other phenotype supported it. A total of 7,481 SVs were called for individuals with bad CQ and 6,697 for individuals with good CQ (Table 5). On the other hand, when counting the SVs in the same conditions, but requiring a minimum coverage of 10 reads per individual supporting the variant, there were a total of 1,480 SVs exclusive of bad CQ individuals and 1,435 of good CQ trees (Table 5).

Table 5: SV calling counts for each type of SV for two distinct coverage thresholds, namely 0 and 10. All calls are for variants exclusive for each phenotype (no individuals from the other phenotype report that SV).

SV types	Supporting individuals > 0			
	Coverage > 0		Coverage > 10	
	Bad CQ	Good CQ	Bad CQ	Good CQ
INS	1,630	1,479	12	15
DEL	2,375	2,019	672	646
INV	64	41	28	21
DUP	630	579	163	171
ITX	2,782	2,579	605	582
Total calls	7,481	6,697	1,480	1,435

Additionally, the number of individuals supporting the same variant was also taken into account, and a new set of results was produced, using as thresholds a minimum of six individuals for bad CQ and five individuals for good CQ. This meant that only SVs with at least seven or six individuals, respectively, for bad and good CQ individuals, were accounted for. The different thresholds for each phenotype were selected considering that there were more libraries from bad CQ individuals (16) than from good CQ ones (14).

For the first approach with a minimum coverage of one read per individual supporting the variant, a total of 49 SVs were called for bad CQ individuals with at least seven individuals supporting the call, and 37 SVs for good CQ individuals with a minimum of six individuals supporting the call (Table 4). Regarding the second approach, the number of SVs detected for each phenotype was smaller, and only 1 SV was found for each phenotype, with both of them being deletions (Table 6).

After defining the list of the SVs called, these data were checked against the reference functional annotation available for the cork oak genome draft, comparing the reference genome coordinates corresponding to the genes with the genomic coordinates of the SVs, in order to evaluate the presence of SVs in genes associated with cork production and quality. Tables 7 and 8 present the most relevant results for the bad and good CQ individuals, respectively. The proteins shown were previously associated with cork production (Ricardo et al., 2011; Soler et al., 2007; Baxter and Stewart Jr, 2013), therefore making relevant their presence in the results. Each record consists in the SV call ID given by *GRIDSS*, the type of SV, the Swissprot (Apweiler et al., 2004) accession number for the protein, the number of individuals supporting the call and number of reads per each individual, and, lastly, the Swissprot description of the protein.

Table 6: SV calling counts for each type of SVs with at least six and five individuals, with respectively bad and good CQ supporting the call, for two distinct coverage thresholds, namely 0 and 10. All calls are for variants exclusive for each phenotype (no individuals from the other phenotype report that SV).

SV types	Bad Supporting indiv. > 6 & Good Supporting indiv. > 5			
	Coverage > 0		Coverage > 10	
	Bad CQ	Good CQ	Bad CQ	Good CQ
INS	2	2	0	0
DEL	24	16	1	1
INV	0	0	0	0
DUP	3	4	0	0
ITX	20	15	0	0
Total calls	49	37	1	1

Table 7: SVs called from bad CQ individuals present in genes associated with cork production.

SV ID	Type of SV	Swissprot Accession	N° SupInd (Reads/each)	Protein (Swissprot)
gridss75_42210o	INS	Q6YW62	2 (1 & 1)	ABC transporter G family member 44
gridss641_19540o	INS	Q84M24	2 (4 & 2)	ABC transporter A family member 1
gridss594_112406o	INS	Q42093	3 (1, 1 & 1)	ABC transporter C family member 2
gridss319_3513o	INS	Q9ZVN2	1 (1)	Acyltransferase-like protein At1g54570
gridss172_84271o	INS	Q8VZ42	1 (8)	Zinc finger AN1/C2H2 stress-associated
gridss101_25681o	INS	Q9LK64	2 (1 & 7)	ABC transporter C family member 3
gridss214_94364o	DEL	Q9T048	5 (4, 16, 5, 8 & 10)	Disease resistance protein At4g27190
gridss150_36369o	DEL	Q6Z4U4	7 (1, 1, 1, 1, 1, 1 & 1)	LRR receptor kinase BAK1
gridss326_96648o	DEL	Q9S9P3	2 (2 & 1)	Factor of DNA methylation 1

A total of six INS and three DEL located in intragenic regions (introns) were identified (Table 7), with the DEL gridss214_943640 being the SV with most reads supporting it (43), and the DEL gridss150_363690 being the one with more individuals supporting the call (7). On the other hand, the INS gridss319_35130 was the one with fewer individuals (1, together with the INS gridss172_842710) and reads (1) supporting the call.

Regarding the SVs identified in the individuals with good CQ (Table 8), six DEL, three INS and one DUP were detected. The DEL gridss39_711520 was the one with more reads supporting the call (33), whereas the DEL gridss358_180040 was the one with more individuals supporting it (5). In turn, five SVs had only 1 individual supporting the call, namely DEL gridss594_1118900, DEL gridss601_126150, DEL gridss14_400990, INS gridss621_159770 and INS gridss284_1237010. Between them, DEL gridss594_1118900 and INS gridss621_159770 were the ones with less reads supporting the call (1). It is noteworthy to point that there were 2 calls for the same protein, ABC transporter C family member

Table 8: SVs called from good CQ individuals present in genes associated with cork production.

SV ID	Type of SV	Swissprot Accession	N° SupInd (Reads/each)	Protein (Swissprot)
gridss39_71152o	DEL	Q9FWX7	2 (15 & 18)	ABC transporter B family member 11
gridss358_18004o	DEL	O81970	5 (1, 4, 3, 1 & 3)	Cytochrome P450 71A9
gridss594_111890o	DEL	Q42093	1 (1)	ABC transporter C family member 2
gridss601_12615o	DEL	Q91VC0	1 (5)	dCTP pyrophosphatase 1
gridss171_38777o	INS	Q7Z5U6	2 (1 & 2)	WD repeat-containing protein 53
gridss14_40099o	DEL	Q9LX30	1 (22)	eIF-2-alpha kinase GCN2
gridss621_15977o	INS	F4HQD4	1 (1)	Heat shock 70 kDa protein 15
gridss135_68313o	DEL	Q9FNV9	2 (3 & 1)	Transcription factor MYB113
gridss384_187813o	DUP	Q9C8L2	2 (1 & 9)	Fatty-acid-binding protein 3
gridss284_123701o	INS	Q9LXV2	1 (6)	Transcription factor MYB46

2, an INS for bad CQ individuals and a DEL for good CQ individuals. Concerning the location of these variants, they were all identified in introns, like the ones identified in the individuals with bad CQ. However, gridss135_683130, a 602 bp DEL was also located in a coding region of the gene coding for the transcription factor MYB113.

Finally, after this comparison between SV calls and functional annotation of the reference, some SVs were visualized using Integrative Genomics Viewer (Robinson et al., 2011). Figure 9 shows an example of one of those SVs visualized, a DEL present in some reads, whereas the other ones mapped concordantly.

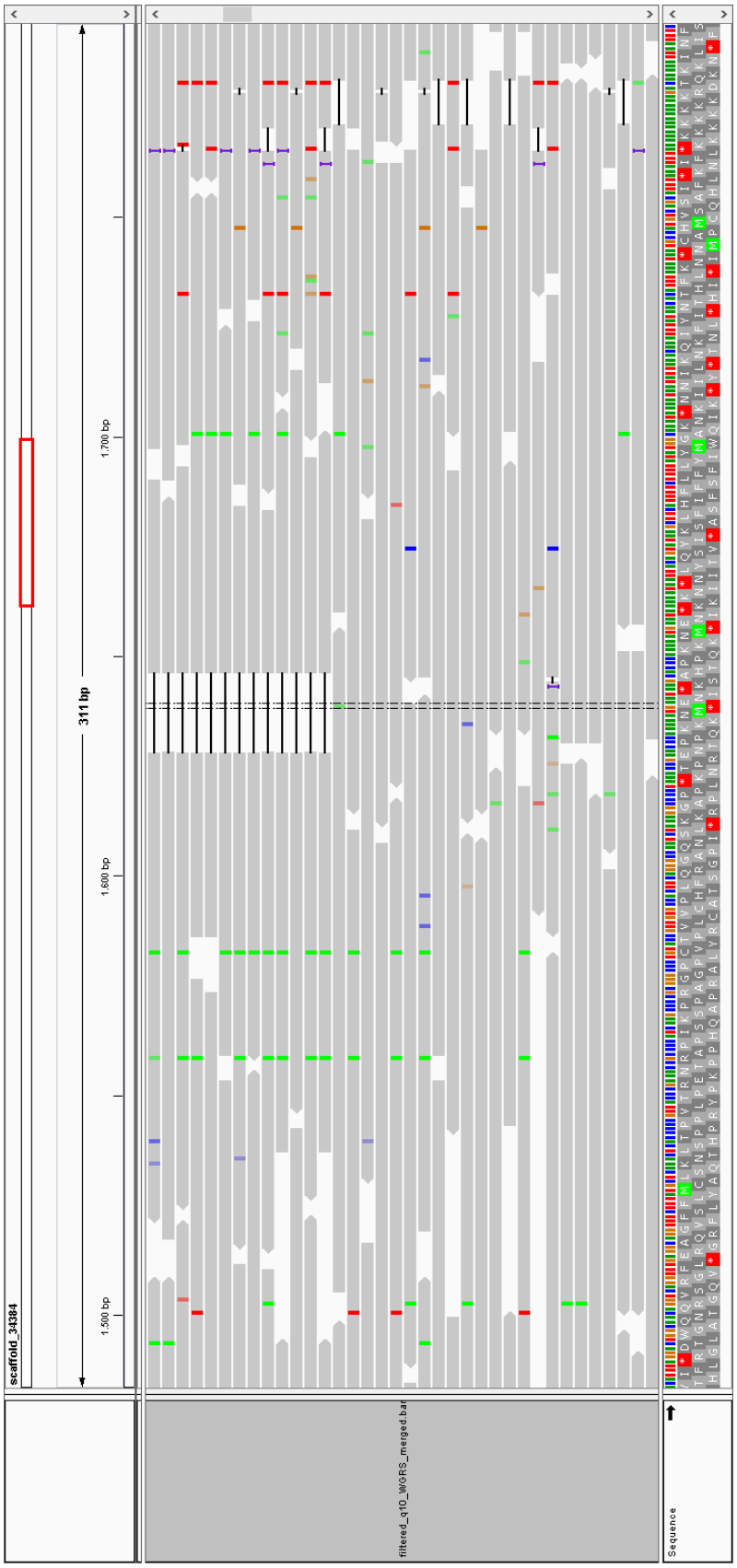


Figure 9: Deletion visualization using Integrative Genomics Viewer (Robinson et al., 2011)

5

DISCUSSION

The quality control of any high-throughput sequence dataset is essential to guarantee that results are not compromised by the inclusion of low quality data. In particular, the quality and length thresholds were selected with the goal of filtering out the bad quality reads but, at the same time, keeping a high percentage of reads, as this was a pipeline with several filtering steps, which progressively discarded the low quality reads. Therefore, an effort was made to balance these two concerns, to avoid discarding a high percentage of reads.

Three initial sets of thresholds were used at the beginning, all with a quality threshold value of 20, and a variable length of 110, 120 and 130 bp, respectively. As the percentage of discarded reads observed was high, averaging between 12.3% for the 110 bp length threshold and 21.8% for the 130 bp one, the quality threshold was placed at a lower level of 15, which produced a dataset with a lower percentage of discarded reads.

The selected quality threshold is in agreement with other studies (Kircher, 2012) and can assure good quality results. Regarding the length threshold, there is a general consensus in the bioinformatics community that a minimum of approximately 70-80% of the length of the original reads size should be maintained, which was the approach followed in this study.

Another filtering process was performed with the alteration of the mapping parameters. In order to increase the stringency, two BWA-MEM default parameters were altered to more stringent values and their performance was evaluated.

The minimum seed length, which is 19 bp by default, was set to 40 bp. Reads are partitioned into several short, non-overlapping segments called seeds. Every seed is a subsequence of the read that contains it, hence every correct mapping for a read will also be mapped by the seed and its location generates a set of potential mappings of the read (Xin et al., 2015). Therefore, if the length of the seed is increased, there will be less potential mappings for the read, however, the confidence of those mappings will be higher.

The other altered parameter was the mismatch penalty, set to 8 instead of the default value, 4. With this alteration, the mapping of a read containing mismatches became more difficult, as it was more penalized. Consequently, the number of mapped reads decreased

when comparing to the default value, although this parameter showed less influence in the decrease of the number of mapped reads than the minimum seed length. However, the big advantage of this parameter was that it increased the number of SA, or chimeric alignments. These alignments are associated with SV, as indicated in several studies (Mills et al., 2006, 2011; Zhang et al., 2011), since they consist of a single read aligning to two distinct portions of the genome with little or no overlap, which indicates the possible existence of a SV in that specific location.

Even though the set of mapping parameters with a mismatch penalty of 8 had less mapped reads, more SA reads, typically indicative of SV presence, were detected. The final decision on the set of parameters to be used was only made after comparing the sets of mapping parameters and the three SV calling software tested, since the main goal was to get the higher number of SVs possible.

The SV calling software chosen included different algorithmic approaches, to verify the influence of these different approaches on the results obtained. Furthermore, the chronological context where they were developed also played a role in the differences observed.

BreakDancer (Chen et al., 2009) was one of the first SV calling software tools to be developed, reason why it has been widely used (Ellis et al., 2012; Govindan et al., 2012; Ding et al., 2010). However, it was developed in a context where sequencing technologies were less advanced than what they currently are. Since sequencing technologies are being improved at a high speed, the need for software to keep up with this development arises.

Algorithms like *BreakDancer*, predicting SVs based only on paired-reads information, namely the insert size and the start and end positions, may not be accurate enough, since they do not use all the information provided by the NGS data. This calling method can explain the high number of ITX predicted, since the reference genome is a draft and, therefore, organized in thousands of scaffolds (23,347) instead of chromosomes. *BreakDancer* assumes each scaffold is a chromosome, and so a pair of reads with one read mapping to a scaffold and the other one to a different scaffold is called an ITX. However, these putative ITX may represent situations where a pair of scaffolds will be arranged in a contiguous order, but at the moment are separated because of the fragmentation observed in the draft genome. This residual number of calls was not exclusive for ITX but for every SV type, showing that *BreakDancer* was not a reliable option for SV detection, due to the lack of support for many of its predictions.

Wham (Kronenberg et al., 2015) was developed more recently, therefore being more suited for recent sequencing technologies than *BreakDancer*. Since it is a recent software, it was used in only a few studies until now (Huddleston et al., 2017; Cone et al., 2016). The method proposed by *Wham* was considered a good fit to this study goals, since it is designed to be performed in a comparison between two groups.

Unlike *BreakDancer*, *Wham* uses additional information than the paired-reads one from the NGS data. It also integrates MP mapping, SA and SC reads re-alignment, together with the use of an alternative alignment and a consensus sequence based evidence to perform the calls with single-nucleotide accuracy. This increases the number of reads being used in the SVs prediction, making use of SA, which are, as pointed before, indicative of SV presence. Also, SC reads re-alignment makes the call more precise, as the alignment of the consensus sequence of these reads to the putative breakpoint makes the breakpoint to be refined to the location of that alignment.

Moreover, SVs bigger than 1,000 bp undergo additional filtering, in order to prevent false positives, since these variants can represent highly deleterious genomic aberrations. This filtering can explain the reduced number of ITX called by *Wham*, as they were possibly resulting from inter-chromosomal (or, in this particular case, inter-scaffold) mapping errors introduced by repetitive sequences. This hypothesis is supported by the *BreakDancer* results, where the majority of ITX called were bigger than 1,000 bp and with low support, as mentioned before with the reduction of predictions when only SVs with a minimum coverage of 10 reads were called.

Concerning the mapping parameters, when there was no coverage filter applied, the *B8* set of parameters was the one with most SVs called by *Wham*, as expected, since this set increased the number of SA, which are then used by *Wham* to call SVs. However, when filtering out predictions with less than 10 reads supporting them, there were more calls with the *default* set of parameters. So, although *B8* favours *Wham* prediction, the raise in the coverage threshold led to a decrease on the number of calls, which could be a result of a higher number of reads discarded by BWA-MEM when mapping with *B8*, altering their "SA" and "XA" tags in the BAM file, information that is used in the prediction by *Wham*.

In turn, *GRIDSS* (Cameron et al., 2017) is the most recent among the three software compared, reason why it has not been used much so far (Ryland et al., 2017). Just like *Wham*, it takes SC reads and SA in consideration, but *GRIDSS* also uses OEA and DP to predict SVs. The use of additional types of non-properly aligned reads to call the variants may be regarded as a potential advantage. These reads were extracted from the BAM file and then assembled by the novel approach presented by *GRIDSS*, which consists on an extension of a positional de Bruijn graph data structure. Finally, SVs were scored and called by applying a probabilistic model.

The fact that *GRIDSS* used all the abnormal reads can explain the *k40* set results, which presented the highest number of calls from this particular software when no coverage filter was applied. Since the *k40* set of parameters increased the number of unmapped reads, and most of them were used by *GRIDSS*, this result was in agreement with the expectations. However, when applying the coverage filter, the *k40* set had the lowest volume of results, which implied that the previous results were due to a higher stringency when mapping.

When the minimum coverage threshold of 10 reads was applied, *GRIDSS* with the *B8* set combination was the one with the best results, not only for this software, but among all software and mapping parameters combinations. Therefore, this was the software chosen to continue with all downstream analyses. Similar to what had been observed with *Wham*, the higher number of SA caused by the *B8* set of parameters led to an increase in the number of SVs called, with *GRIDSS* having an advantage over *Wham*, because it used more types of non-properly aligned reads.

As demonstrated before, it was highly relevant to perform the software and mapping parameters comparison with a minimum coverage of reads, to eliminate some false positives from the SVs counts, since without this minimum threshold, all the predictions are being considered, even the ones with few support. Since the goal was to choose the most reliable combination, it was important to verify which one produced more results with a higher level of confidence.

Given the fact that the VCF file outputted by *GRIDSS* was a file of break-end points, instead of SVs, this VCF file underwent an additional processing step, which consisted in using an R script available in the *GRIDSS* repository. This was made under advice from the software developer himself who, when confronted with the details of this study and its aims, pointed this translation from break-end points to SV types as the best solution.

With both the set of mapping parameters and the software selected, SV calling was performed for all the 30 libraries. Subsequently, the total number of SVs, as well as the SVs exclusive for each phenotype of CQ, were counted.

The total number of SVs found in the cork oak genome accounted for a total of 93,980 variants identified, of which 78% were deletions and translocations, with deletion being the most representative type of SV with 38,865 occurrences identified. On the other hand, inversion was the less representative type, with only 654 SVs called. It is difficult to compare these results with studies performed in other species, with this being one of the first studies for the identification of these types of SV in plants, since the published studies were mainly focused in the identification of CNVs and PAVs (Saxena et al., 2014; Zhang et al., 2017).

The results showed differences between the two CQ phenotypes, supported by the 7,481 SVs exclusive for bad CQ individuals and the 6,697 ones exclusive for good CQ individuals. Like in the previous comparison, the SVs were also filtered by coverage, but this time the filter applied was a minimum of 10 reads per individual supporting the call. Despite this increased stringency, a total of 1,480 SVs exclusive for bad CQ individuals and 1,435 for good CQ individuals were found, supporting the idea that the differences between the phenotypes may be related with SV present in the cork oak genome.

In addition, the number of individuals supporting the call, with and without the minimum coverage threshold, was used as an extra filter to further evaluate the results.

Without applying the threshold, there were 49 SVs called for a minimum of 7 out of 16 bad CQ individuals supporting them and with none of the 14 good CQ individuals supporting them. On the other hand, 37 SVs were called on a minimum of 6 out of 14 good CQ individuals, with none of the 16 bad CQ individuals supporting them. This emphasizes the idea that the differences observed between the phenotypes may be regulated by the presence of specific SVs in the cork oak genome. Regarding the comparison with the minimum coverage threshold, only a single DEL was found for each exclusive group with 7 and 6 supporting individuals, respectively for bad and good CQ individuals. This decrease was expected, because the additional stringency applied for the read coverage threshold would always reduce the number of SVs. In the future, WGRS datasets will be produced using higher genome coverage per individual, as the sequencing costs keep decreasing, which will alleviate this problem and allow the detection of relevant SVs with higher read coverage thresholds.

Finally, when crossing the information from SV calling with the transcripts from the functional annotation of the draft genome, it was possible to identify SVs occurring in genes somehow connected with cork production. There were four INS called for the bad CQ individuals and two DEL for the good CQ ones in transcripts posteriorly translated into ABC transporter proteins, which have already been associated with cork production (Soler et al., 2007). The fact that there were INS for bad CQ phenotype and DEL for good CQ one, can mean that a variation in this type of transporters may affect the CQ, although the individuals and reads support was somewhat insufficient to make this claim with more certainty.

Regarding the bad CQ individuals, the two transcripts with more support for the SV called were the ones coding for a disease resistance protein At4g27190 and for a LRR receptor kinase BAK1. This disease resistance protein belongs to the disease resistance NBS-LRR family (Apweiler et al., 2004), responsible for the host defense responses (Marone et al., 2013). The LRR receptor kinase BAK1, or SERK3 is involved in defense response (Park et al., 2011), which, together with SERK1, SERK2, SERK4 and SERK5, are co-receptors to regulate extracellular multiple signal transduction (Albrecht et al., 2008). The other three remaining proteins, acyltransferase-like protein At1g54570, zinc finger AN1/C2H2 stress-associated protein and factor of DNA methylation 1, were found to be related with cork production in an ongoing study of the group.

Concerning to good CQ individuals, the SV with most individuals support was a DEL in the gene coding for the cytochrome P450 71A9, already related with suberin biosynthesis and cork production (Soler et al., 2007). Also, INS were called in WD repeat-containing protein 53 and heat shock protein 15, two xylem proteins possibly related with cork production (Ricardo et al., 2011). Transcription factors MYB113 and MYB46, previously associated with cork production (Soler et al., 2007; Baxter and Stewart Jr, 2013), also had

SVs present in their genes, respectively a DEL and an INS, with special relevance for the DEL identified in the transcription factor MYB113, since it was located in a coding region, which consequently may be associated with CQ. Additionally, like in the bad CQ individuals, the remaining proteins, dCTP pyrophosphatase 1, eIF-2-alpha kinase GCN2 and the fatty-acid-binding protein 3 were associated with cork production in an ongoing study of the group.

Overall, some evidence was found regarding the role of SV on CQ, since SVs were found on cork production associated genes. These results are now available for validation studies using a larger number of cork oak individuals with phenotypic information regarding CQ, and sampled in different geographical locations.

6

CONCLUSIONS

The advances in sequencing technology are extending the reach of genomic studies, providing better data that, with the help of newly developed tools, will expand and improve the knowledge on all relevant species, and subsequent understanding of biological processes that might be important in many areas like medicine, agriculture and engineering. In particular, SV certainly plays a role in many biological processes that can be leading to phenotypic differences, and will greatly benefit from these technological advances. Bioinformatic analysis of NGS data will be the key element to find and characterize these effects.

This present work was the first study performed in cork oak where WGRS was used, by the analysis of the whole genome of 30 individuals. This magnitude of genomes is the first step to construct the species pangenome, which will then be crucial to understand how SV determines the differences in CQ, since this is the most important economic trait of these trees. The results obtained clearly indicated that SV is present in the cork oak genome. The future release of an improved version of the cork oak genome, with a higher degree of genome organization and lower number of scaffolds, will substantially improve the accuracy with which SV will be performed in cork oak.

This study will enhance our ability to determine whether the CQ associated genes belong to the set of core or variable genes, which will then enable the knowledge regarding the role played by CNVs, PAVs and SV in the cork oak genome and in the regulation of CQ. A number of SVs were specific to one of the CQ groups analysed in this study. With the final version of the genome available also including an improved annotation it will be possible to originate a more accurate prediction of the effect SVs may have in gene regulation, and its effects on relevant cork oak phenotypes, such as CQ and resistance to (a)biotic stresses. Future studies should also include cork oaks sampled from different geographical regions, to further understand the global distribution of SV in cork oak and the effect of SVs in phenotypes under a complex genetic regulation. Finally, the larger volumes of data that are anticipated for the coming years, including short and long-read sequencing technology, will require the update of the current bioinformatic tools currently available for SV detection, and eventually the development of new tools, better suited to deal with this new big-data scenario.

BIBLIOGRAPHY

Albrecht, C., Russinova, E., Kemmerling, B., Kwaaitaal, M., and de Vries, S. C. (2008). Arabidopsis somatic embryogenesis receptor kinase proteins serve brassinosteroid-dependent and-independent signaling pathways. *Plant physiology*, 148(1):611–619.

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

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl_1):D115–D119.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., Liberg, O., Arnemo, J. M., Hedhammar, Å., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, 495(7441):360–364.

Baker, M. (2012). Structural variation: the genome's hidden architecture. *Nature methods*, 9(2):133–7.

Bartenhagen, C. and Dugas, M. (2015). Robust and exact structural variation detection with paired-end and soft-clipped alignments: Softsv compared with eight algorithms. *Briefings in bioinformatics*, 17(1):51–62.

Baxter, H. L. and Stewart Jr, C. N. (2013). Effects of altered lignin biosynthesis on phenylpropanoid metabolism and plant stress. *Biofuels*, 4(6):635–650.

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., and Rafalski, A. (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics*, 120(2):355–367.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53.

Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., Matukumalli, L. K., Song, J., Schnabel, R. D., Ventura, M., Taylor, J. F., et al. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research*, 22(4):778–790.

Buiting, K., Saitoh, S., Gross, S., Dittrich, B., Schwartz, S., Nicholls, R. D., and Horsthemke, B. (1995). Inherited microdeletions in the angelman and prader–willi syndromes define an imprinting centre on human chromosome 15. *Nature genetics*, 9(4):395–400.

Cameron, D. L., Schroeder, J., Penington, J. S., Do, H., Molania, R., Dobrovic, A., Speed, T. P., and Papenfuss, A. T. (2017). Gridss: sensitive and specific genomic rearrangement detection using positional de bruijn graph assembly. *bioRxiv*, page 110387.

Caritat, A., Gutiérrez, E., and Molinas, M. (2000). Influence of weather on cork-ring width. *Tree physiology*, 20(13):893–900.

Chen, K., Chen, L., Fan, X., Wallis, J., Ding, L., and Weinstock, G. (2014). Tigra: a targeted iterative graph routing assembler for breakpoint assembly. *Genome research*, 24(2):310–317.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., et al. (2009). Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–681.

Chew, S., Dastani, Z., Brown, S. J., Lewis, J. R., Dudbridge, F., Soranzo, N., Surdulescu, G. L., Richards, J. B., Spector, T. D., and Wilson, S. G. (2012). Copy number variation of the apc gene is associated with regulation of bone mineral density. *Bone*, 51(5):939–943.

Cone, K. R., Kronenberg, Z. N., Yandell, M., and Elde, N. C. (2016). Accelerated selection of a viral rna polymerase variant during gene copy number amplification promotes rapid evolution of vaccinia virus. *bioRxiv*, page 066845.

Dauber, A., Yu, Y., Turchin, M. C., Chiang, C. W., Meng, Y. A., Demerath, E. W., Patel, S. R., Rich, S. S., Rotter, J. I., Schreiner, P. J., et al. (2011). Genome-wide association of copy-number variation reveals an association between short stature and the presence of low-frequency genomic deletions. *The American Journal of Human Genetics*, 89(6):751–759.

Degenhardt, J. D., De Candia, P., Chabot, A., Schwartz, S., Henderson, L., Ling, B., Hunter, M., Jiang, Z., Palermo, R. E., Katze, M., et al. (2009). Copy number variation of ccl3-like genes affects rate of progression to simian-aids in rhesus macaques (macaca mulatta). *PLoS genetics*, 5(1):e1000346.

Díaz, A., Zikhali, M., Turner, A. S., Isaac, P., and Laurie, D. A. (2012). Copy number variation affecting the photoperiod-b1 and vernalization-a1 genes is associated with altered flowering time in wheat (triticum aestivum). *PLoS One*, 7(3):e33234.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., et al. (2010). Genome remodeling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291):999.

Durkin, K., Coppieters, W., Drögemüller, C., Ahariz, N., Cambisano, N., Druet, T., Fasquelle, C., Haile, A., Horin, P., Huang, L., et al. (2012). Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature*, 482(7383):81.

Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., Van Tine, B. A., Hoog, J., Goiffon, R. J., Goldstein, T. C., et al. (2012). Whole genome analysis informs breast cancer response to aromatase inhibition. *Nature*, 486(7403):353.

Farrar, M. (2006). Striped smith–waterman speeds database searches six times over other simd implementations. *Bioinformatics*, 23(2):156–161.

Fontanesi, L., Beretti, F., Riggio, V., González, E. G., DallOlio, S., Davoli, R., Russo, V., and Portolano, B. (2009). Copy number variation and missense mutations of the agouti signaling protein (asip) gene in goat breeds with different coat colors. *Cytogenetic and genome research*, 126(4):333–347.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., et al. (2011). Multiple reference genomes and transcriptomes for arabidopsis thaliana. *Nature*, 477(7365):419–423.

Golicz, A. A., Batley, J., and Edwards, D. (2016). Towards plant pangenomics. *Plant Biotechnology Journal*, 14(4):1099–1105.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., and McCombie, W. R. (2015). Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11):1750–1756.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N. D., Kanchi, K. L., Maher, C. A., Fulton, R., Fulton, L., Wallis, J., et al. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6):1121–1134.

Graça, J. and Pereira, H. (2004). The periderm development in quercus suber. *Iawa Journal*, 25(3):325–335.

Graf, J., Voisey, J., Hughes, I., and Van Daal, A. (2007). Promoter polymorphisms in the matp (slc45a2) gene are associated with normal human skin color variation. *Human mutation*, 28(7):710–717.

Guan, P. and Sung, W.-K. (2016). Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods*, 102:36–49.

Hajirasouliha, I., Hormozdiari, F., Alkan, C., Kidd, J. M., Birol, I., Eichler, E. E., and Sahinalp, S. C. (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, 26(10):1277–1283.

Hamilton, C., Verduzco-Gómez, A., Favetta, L., Blondin, P., and King, W. (2012). Testis-specific protein y-encoded copy number is correlated to its expression and the field fertility of canadian holstein bulls. *Sexual Development*, 6(5):231–239.

Haun, W. J., Hyten, D. L., Xu, W. W., Gerhardt, D. J., Albert, T. J., Richmond, T., Jeddeloh, J. A., Jia, G., Springer, N. M., Vance, C. P., et al. (2011). The composition and origins of genomic variation among individuals of the soybean reference cultivar williams 82. *Plant physiology*, 155(2):645–655.

Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5):677–685.

Illumina (2015). HiSeq X Series of Sequencing Systems. pages 1–4.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature genetics*, 44(2):226–232.

Jiang, Y., Wang, Y., and Brudno, M. (2012). Prism: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, 28(20):2576–2583.

Joerg, H., Fries, H., Meijerink, E., and Stranzinger, G. (1996). Red coat color in holstein cattle is associated with a deletion in the mshr gene. *Mammalian genome*, 7(4):317–318.

Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M., and Chalmers, D. (2012). Data sharing in the post-genomic world: the experience of the international cancer genome consortium (icgc) data access compliance office (daco). *PLoS Comput Biol*, 8(7):e1002549.

Joshi, N. and Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for fastq files (version 1.33) [software]. *Available online at: http://github.com/najoshi/sickle*.

Kircher, M. (2012). Analysis of high-throughput ancient dna sequencing data. *Ancient DNA: methods and protocols*, pages 197–228.

Kronenberg, Z. N., Osborne, E. J., Cone, K. R., Kennedy, B. J., Domyan, E. T., Shapiro, M. D., Elde, N. C., and Yandell, M. (2015). Wham: identifying structural variants of biological consequence. *PLoS computational biology*, 11(12):e1004572.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature genetics*, 42(11):1027–1030.

Lai, K., Duran, C., Berkman, P. J., Lorenc, M. T., Stiller, J., Manoli, S., Hayden, M. J., Forrest, K. L., Fleury, D., Baumann, U., et al. (2012). Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant biotechnology journal*, 10(6):743–749.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.-W., He, W., Qin, N., Wang, B., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature genetics*, 42(12):1053–1059.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., and Marth, G. T. (2014). Mosaik: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS one*, 9(3):e90581.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., et al. (2008). Dna sequencing of a cytogenetically normal acute myeloid leukemia genome. *Nature*, 456(7218):66.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Liu, S., Huang, S., Rao, J., Ye, W., Krogh, A., and Wang, J. (2015). Discovery, genotyping and characterization of structural variation and novel sequence at single nucleotide resolution from de novo genome assemblies on a population scale. *GigaScience*, 4(1):64.

Liu, Y. and Schmidt, B. (2012). Long read alignment based on maximal exact match seeds. *Bioinformatics*, 28(18):i318–i324.

Lourenço, A., Rencoret, J., Chemetova, C., Gominho, J., Gutiérrez, A., del Río, J. C., and Pereira, H. (2016). Lignin composition and structure differs between xylem, phloem and phellem in quercus suber l. *Frontiers in plant science*, 7.

Mardis, E. R. (2017). Dna sequencing technologies: 2006-2016. *Nature protocols*, 12(2):213–218.

Marone, D., Russo, M. A., Laidò, G., De Leonardis, A. M., and Mastrangelo, A. M. (2013). Plant nucleotide binding site–leucine-rich repeat (nbs-lrr) genes: active guardians in host defense responses. *International journal of molecular sciences*, 14(4):7302–7326.

Marques, A. V. and Pereira, H. (2013). Lignin monomeric composition of corks from the barks of betula pendula, quercus suber and quercus cerris determined by py–gc–ms/fid. *Journal of analytical and applied pyrolysis*, 100:88–94.

Marsh, M., Tu, O., Dolnik, V., Roach, D., Solomon, N., Bechtol, K., Smietana, P., Wang, L., Li, X., Cartwright, P., et al. (1997). High-throughput dna sequencing on a capillary array electrophoresis system. *Journal of capillary electrophoresis*, 4(2):83–89.

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., et al. (2008). Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, 82(2):477–488.

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., Gerhardt, D. J., Jeddeloh, J. A., and Stupar, R. M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant physiology*, 159(4):1295–1308.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Current opinion in genetics & development*, 15(6):589–594.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, 6:S13–S20.

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., et al. (2010a). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *The American Journal of Human Genetics*, 86(5):749–764.

Miller, J. R., Koren, S., and Sutton, G. (2010b). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (indel) variation in the human genome. *Genome research*, 16(9):1182–1190.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., et al. (2011). Mapping copy number variation by population scale genome sequencing. *Nature*, 470(7332):59.

Nishida, H., Yoshida, T., Kawakami, K., Fujita, M., Long, B., Akashi, Y., Laurie, D. A., and Kato, K. (2013). Structural variation in the 5??? upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (Triticum aestivum L.), and their effect on heading time. *Molecular Breeding*, 31(1):27–37.

Park, H. S., Ryu, H. Y., Kim, B. H., Kim, S. Y., Yoon, I. S., and Nam, K. H. (2011). A subset of osserk genes, including osbak1, affects normal growth and leaf development of rice. *Molecules and cells*, 32(6):561–569.

Pereira, H. (2011). *Cork: biology, production and uses*. Elsevier.

Pereira, H., Lopes, F., and Graga, J. (1996). The Evaluation of the Quality of Cork Planks by Image Analysis. *Holzforschung*, 50(2):111–115.

Pereira, H., Rosa, M. E., and Fortes, M. (1987). The cellular structure of cork from quercus suber l. *IAWA Journal*, 8(3):213–218.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10):1256.

Pober, B. R. (2010). williams–beuren syndrome. *New England Journal of Medicine*, 362(3):239–252.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rahantamalala, A., Rech, P., Martinez, Y., Chaubet-Gigot, N., Grima-Pettenati, J., and Pacquit, V. (2010). Coordinated transcriptional regulation of two key genes in the lignin branch pathway-cad and ccr-is mediated through myb-binding sites. *BMC plant biology*, 10(1):130.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339.

Ricardo, C. P., Martins, I., Francisco, R., Sergeant, K., Pinheiro, C., Campos, A., Renaut, J., and Fevereiro, P. (2011). Proteins associated with cork formation in quercus suber l. stem tissues. *Journal of proteomics*, 74(8):1266–1278.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26.

Ronen, R., Boucher, C., Chitsaz, H., and Pevzner, P. (2012). Sequel: improving the accuracy of genome assemblies. bioinformatics 28: i188–i196.

Ryland, G. L., Jones, K., McBean, M., Khot, A., Seymour, J. F., and Blombery, P. (2017). Comprehensive genomic characterization dissects the complex biology of a case of synchronous burkitt lymphoma and myeloid malignancy with shared hematopoietic ancestry. *Leukemia & Lymphoma*, pages 1–4.

S. Collins, F., S. Lander, E., Rogers, J., H. Waterston, R., and H. G. S. Conso, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441IN19447–446IN20448.

Saunders, C. J., Miller, N. A., Soden, S. E., Dinwiddie, D. L., Noll, A., Alnadi, N. A., Andraws, N., Patterson, M. L., Krivohlavek, L. A., Fellis, J., et al. (2012). Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Science translational medicine*, 4(154):154ra135–154ra135.

Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant genomes. *Briefings in functional genomics*, 13(4):296–307.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., Tewari, A. K., Kitabayashi, N., Moss, B. J., Chee, M. S., et al. (2010). Fusionseq: a modular framework for finding gene fusions by analyzing paired-end rna-sequencing data. *Genome biology*, 11(10):R104.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., and Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature genetics*, 39:S7–S15.

Schröder, J., Hsu, A., Boyle, S. E., Macintyre, G., Cmero, M., Tothill, R. W., Johnstone, R. W., Shackleton, M., and Papenfuss, A. T. (2014). Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, 30(8):1064–1072.

Silva, S. P., Sabino, M. a., Fernandes, E. M., Correlo, V. M., Boesel, L. F., and Reis, R. L. (2005). Cork: properties, capabilities and applications. *International Materials Reviews*, 50(4):256–256.

Sindi, S. S., Önal, S., Peng, L. C., Wu, H.-T., and Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome biology*, 13(3):R22.

Smith, A., McGavran, L., Robinson, J., Waldstein, G., Macfarlane, J., Zonona, J., Reiss, J., Lahr, M., Allen, L., Magenis, E., et al. (1986a). Interstitial deletion of (17)(p11. 2p11. 2) in nine patients. *American Journal of Medical Genetics Part A*, 24(3):393–414.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., and Hood, L. E. (1986b). Fluorescence detection in automated dna sequence analysis. *Nature*, 321(6071):674–679.

Soler, M., Serra, O., Molinas, M., Huguet, G., Fluch, S., and Figueras, M. (2007). A Genomic Approach to Suberin Biosynthesis and Cork Differentiation. *Plant Physiology*, 144(1):419–431.

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., et al. (2009). Maize inbreds exhibit high levels of copy number variation (cnv) and presence/absence variation (pav) in genome content. *PLoS Genet*, 5(11):e1000734.

Stobbe, G., Liu, Y., Wu, R., Hudgings, L. H., Thompson, O., and Hisama, F. M. (2013). Diagnostic yield of array comparative genomic hybridization in adults with autism spectrum disorders. *Genetics in Medicine*, 16(1):70–77.

Sulston, J., Du, Z., et al. (1992). The c. elegans genome sequencing project: a beginning. *Nature*, 356(6364):37.

Tan, S., Zhong, Y., Hou, H., Yang, S., and Tian, D. (2012). Variation of presence/absence genes among arabidopsis populations. *BMC evolutionary biology*, 12(1):86.

Teixeira, R. T., Fortes, A. M., Pinheiro, C., and Pereira, H. (2014). Comparison of good- and bad-quality cork: Application of high-throughput sequencing of phellogenic tissue. *Journal of Experimental Botany*, 65(17):4887–4905.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al. (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955.

Vacic, V., McCarthy, S., Malhotra, D., Murray, F., Chou, H. H., Peoples, A., Makarov, V., Yoon, S., Bhandari, A., Corominas, R., Iakoucheva, L. M., Krastoshevsky, O., Krause, V., Larach-Walters, V., Welsh, D. K., Craig, D., Kelsoe, J. R., Gershon, E. S., Leal, S. M., Dell Aquila, M., Morris, D. W., Gill, M., Corvin, A., Insel, P. A., McClellan, J., King, M. C., Karayiorgou, M., Levy, D. L., DeLisi, L. E., and Sebat, J. (2011). Duplications of

the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature*, 471(7339):499–503.

Varshney, R. K., Nayak, S. N., May, G. D., and Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in biotechnology*, 27(9):522–530.

Verdaguer, R., Soler, M., Serra, O., Garrote, A., Fernández, S., Company-Arumí, D., Anticó, E., Molinas, M., and Figueras, M. (2016). Silencing of the potato stnac103 gene enhances the accumulation of suberin polyester and associated wax in tuber skin. *Journal of experimental botany*, 67(18):5415–5427.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Xin, H., Nahar, S., Zhu, R., Emmons, J., Pekhimenko, G., Kingsford, C., Alkan, C., and Mutlu, O. (2015). Optimal seed solver: optimizing seed selection in read mapping. *Bioinformatics*, 32(11):1632–1642.

Xu, W., Purugganan, M. M., Polisensky, D. H., Antosiewicz, D. M., Fry, S. C., and Braam, J. (1995). Arabidopsis tch4, regulated by hormones and the environment, encodes a xyloglucan endotransglycosylase. *The Plant Cell*, 7(10):1555–1567.

Zhang, X., Chen, X., Liang, P., and Tang, H. (2017). Cataloguing plant genome structural variations. *Current issues in molecular biology*, 27:181.

Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). Identification of genomic indels and structural variations using split reads. *BMC genomics*, 12(1):375.