A case study for the *de novo* genome assembly of plant genomes from Illumina short reads

by

Godwin Mafireyi

Submitted in partial fulfilment of the requirements of the degree

Magister Scientiae in Bioinformatics

Department of Biochemistry, Microbiology and Genetics

Faculty of Natural and Agricultural Sciences

University of Pretoria

August 2018

Supervisor: Dr Charles A Hefer

Co-supervisor: Prof Fourie Joubert

Declaration

I, Godwin Mafireyi, declare that the dissertation, which I hereby submit for the degree *Magister Scientiae* at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date: 10 June 2018

Acknowledgements

I would like to thank the following people for their contribution and support during this project

- My supervisor, **Dr Charles Hefer.** For his support from the start to the end of this project. Thank you for your guidance. I learnt a lot from you.
- My co-supervisor, **Prof Fourie Joubert**. For believing in me and giving me a chance. I will forever be grateful for your support.
- The **National Research Foundation** – For funding me.
- The **Agricultural Research Council** – for housing me, funding and supporting my project.
- **Dr Jasper Rees** and **Mrs Anita Severn-Ellis** for graciously sharing the sequence data for the project with me.
- The **ARC-Biotechnology Platform**. All my colleagues from BTP have been family and helped be grow as a researcher.
- The **Centre for Bioinformatics and Computational Biology**, University of Pretoria. Colleagues from the Bioinformatics and Genetics were a great support during this project. I am thankful.
- My wife, **Kutlwano Constance Mafireyi**, for her love and support during the project. For always believing in me and being by my side when working on this project
- My son, **Anashe Oanthata Mafireyi**, for always being a source of inspiration.

Contents

Table of Figures

Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	messenger Ribonucleic acid
NGS	Next Generation Sequencing
CDS	Coding sequence
WGS	Whole Genome Sequencing
GWA	Guava Wilt Disease
ARC	Agricultural Research Council
ARC-ITSC	ARC's Institute for Tropical and Subtropical Crops
TPS	Terpene Synthase
ABySS	Assembly By Short Sequences
MaSurCA	Maryland Super-Read Celera Assembler
SGA	String Graph Assembler
BUSCO	Benchmarking Universal Single-Copy Orthologs
CEGMA	Core Eukaryotic Genes Mapping Approach
QUAST	Quality Assessment Tool

List of Tables

Abstract

The aim of this study was to create a genomic resource for a typical plant genome from Illumina short reads, using *Psidium guajava* as a case study. Here we present a bioinformatics approach to produce a *de novo* plant genome assembly, perform annotation, and compare the newly assembled and annotated genome to that of a reference genome, in this case *Eucalyptus grandis*. The assembly pipeline was constructed using a combination of the best results from four different assemblers namely (ABySS, Allpaths-LG, SGA and MaSurCA) with a combination of Illumina paired end and mate pair reads. Each assembler used a different graph-based approach in their assembly strategy, and the output from these assemblers were merged by Metassembler to produce a best assembly. We manage to create comprehensive genomic resource for the guava fruit tree from Illumina short reads. The annotated genome of *Psidium guajava* will serve a major genomic resource in the investigation of the interaction between the plant and pathogens such as *Nalanthamala psidii* (*N. psidii).* Also,our comparative genomics work is a starting point to learn more about the genetic diversity in the Myrtaceae family.

In Chapter 1 a comprehensive literature review of the current state of sequencing technologies, with a focus on third generation sequencing technologies is presented. This is followed by a discussion on different whole genome assembly approaches and techniques, with examples of each type of approach implemented as a software package. The relevance and importance of the non-model organism that we used as a case study, *Psidium guajava*, is also discussed in Chapter 1

In Chapter 2 the genome assembly and annotation pipelines and processes are discussed. Detailed materials and methods used in this study are provided.

The main findings and results of the study is discussed in Chapter 3, with a concluding remarks chapter presented as the last chapter of this dissertation.

The work presented here has been presented at the following conferences, and a manuscript on the genome resource is in preparation:

1. Poster Presentation – SAGS/SASBI conference 2016 (Durban)

1 Chapter 1: Literature Review

1.1 Introduction to whole genome sequencing

Whole genome sequencing (WGS) is a method of determining the complete sequence identity of nucleotides that compose the genome from a DNA molecule. Technological advancements to perform WGS has developed exponentially over the past 10 years, a field that is continuously improving and innovating. In earlier years, technology used for WGS required extensive resources, was expensive and time consuming. Now, sequencing an entire genome is much less expensive and requires less time with the price dropping to less than US$1000 per human genome at 30x coverage from about 3 billion US dollars for the first human genome (Muir *et al.*, 2016).

The first papers published on sequencing of entire genomes or genes were by Fred Sanger and Alan R. Coulson in 1977 (Schuster, 2007). DNA laboratories used Sanger sequencing for about 30 years. Demand for high DNA throughput led to development of technologies such as automated capillary electrophoresis. The first automatic sequencing machine (namely the ABI 370) which used capillary electrophoresis, was introduced by Applied Biosystems in 1987 and it made the sequencing more accurate and faster. The current model AB3730xl can output 2.88 M bases per day and read length could reach 900 bases since 1995 (Liu *et al.*, 2012).

The first genome to be sequenced was the viral genome Bacteriophage fX174 with a total of 5 368 base pairs (bp). The Human Genome Project (HGP) was introduced in 1990 (Tsui and Scherer W., 2001) and 11 years later in February 2001, the first draft of the human genome was published (Lander *et al.*, 2001). The cost of the project was about 3 billion US dollars. During this time, due to the high throughput requirements of the HGP, a new massive parallel technology referred to as Next Generation Sequencing (NGS) was introduced to meet some of these requirements.

1.1.1 Next generation sequencing

The NGS technologies differ from the Sanger method in the technology used (simultaneous fluorescence technology in NGS and single dye termination system in Sanger), cost (NGS is significantly cheaper) and that NGS uses massive parallel analysis with high throughput. Following the human genome project, three companies

454, Solexa and Agencourt,released three massive parallel sequencing systems with lower costs, higher accuracy and performance compared with Sanger sequencing. 454 relased the 454 sequencer, Solexa released the Genome Analyzer and Agencourt released the Sequencing by Oligo Ligation Detection (SOLID) instruments in 1995. Applied Biosystems acquired Agencourt in 2006, followed by 454, which was bought by Roche in 2007, and Solexa bought by Illumina. In 2010 Pacific Biosciences released yet another sequencer, the PacBio RS and a new version PacBio RS II in April 2013. This was the first long read sequencing technology that produced reads of potentially several kilobases long. More long read sequencing platforms are currently being introduced, an example being the MinION introduced by Oxford Nanopore technologies. Table 1.1 shows a comparison of the different sequencing technologies. Sequencing technologies have greatly improved since their inception, with each company modifying their technologies to produce longer and more accurate reads. Below are examples of different systems used in NGS technology. Figure 1.1 shows examples of different sequencing machines. Figure 1.1a shows the HiSeq 3000 machine from Illumina, Figure 1.1b shows the Minion sequencer from Oxford Nanopore, Figure 1.1c is the Roche 453 sequencer, Figure 1.1d shows the PacBio RSII sequencer from PacBio and finally the Solid sequencer (Figure 1.1e).

Figure 1.1: Examples of sequencing technologies used in genomic projects. The Illumina HiSeq 3000 (a), the pocket-sized MinIon from Oxford Nanopore Technologies (b), The Roche 454 seqeuncer (c), the PacBio RSII and the SOLiD sequencer (e).

Illumina systems

The Illumina systems account for the largest share of the market (Fedurco *et al.,* 2006). It has been the main sequencing platform in many genome projects including humans, plants and animals such as fish genomes (Narum, 2015). The Illumina sequencer adopted the technology of sequencing by synthesis (SBS) (Liu *et al*., 2012). The SBS technology takes advantage of the base-by-base sequencing where each dNTP is attached to a fluorescently labelled terminator. The reaction image from the terminator is easily taken before another base is added (Ju *et al.*, 2006).

Solexa introduced the Genome Analyzer (GA) in 2006 and it had an output of 1 million bases, or 1Gbp per run. Through improvements in software, flow cell, polymerase and buffer, the output increased to 50Gpb/run in December 2009 and the last interation in the GAIIx series eventually could attain 85Gbp/run and produce read lengths of up to 100 base pairs (Quail *et al.*, 2012). However, due to the high throughput requirements of most genomic projects, greater improvements needed to be made to the system.

In early 2010, (after acquiring Solexa in 2006) Illumina launched the HiSeq 2000, which could output 600Gbp per run in 8 days. With that high throughput, Illumina posed an advantage over Sanger and the GA technologies in meeting high throughput requirements of most genome projects. Also, compared to other sequencing techniques, Illumina was relatively cheap with a cost of about US$0.07/million bases (Illumina Inc., 2012). However, with a relatively short read length of up to 150bp (HiSeq Nextera kit), which is significantly shorter than that of Sanger (550bp – 900bp), it posed many bioinformatics related challenges in tasks such as genome assembly.

In 2011, Illumina launched the MiSeq (Quail *et al.,* 2012). It produced an output of 15G/run and read lengths of up to 2x300bp. It is therefore most applicable for bacterial sample and amplicon sequencing. The MiSeq sequence data is better for contig assembly than HiSeq mainly because of the improvement in read length even though the low output is still a shortcoming. Over the years Illumina has further improved its platforms and introduced the HiSeq 3000 and HiSeq 4000 which could produce 1500G/run (5 billion 2x150bp reads) in 1-3.5 days, the HiSeq X series that produces 1800G/run in less than 3 days, the NextSeq series that produces 120Gbp/run (400 million 2x150bp reads) in 12-30hours and most recently the NovaSeq series which produces 6000G/run (20 billion 2x150bp reads) in 19-40 hours) and the MiniSeq which in 4-20 hours can produce 4.5Gbp/run (25 million 2x150bp reads).

Roche 454 system

The Roche 454 system was first introduced in 2000 by CuraGen before being acquired by Roche in 2007 but was shut down in 2013 as the technology became non-competitive (http://www.bio-itworld.com/2013/10/16/six-years-after-acquisition-roche-quietly-shutters-454.html). The 454-sequencing system used pyrosequencing technology to perform the sequence analysis (Willer, GM and CJ, 2008). The 454 systems were used for many genome sequencing projects including the Neanderthal genome (Green *et al.*, 2010).

Pyrosequencing technology relied on the luminometric detection of pyrophosphate that is released during primer-directed DNA polymerase-catalyzed nucleotide incorporation (Chowdhury *et al.,* 2012). It used four enzymes to detect nucleic acid sequences during the synthesis. First a DNA segment is amplified, biotinylated and mixed with four enzymes; DNA polymerase, ATP sulfurylase, luciferase and apyrase, and the substrates adenosine 5' phosphosulfate (APS) and luciferin (Gharizadeh *et al.*, 2006). Then on a picotiter plate, one of the four dNTPs will complement to the bases of the template strand with the help of the four enzymes and release pyrophosphate (PPi) which equals the number of incorporated nucleotides. The ATP transformed from PPi drives the luciferin into oxyluciferin and generates visible light. The locations of these light signals are detected and used to determine which beads the nucleotides are added to. At the same time, the unmatched bases are degraded by apyrase. Then another dNTP is added into the reaction system and the pyrosequencing reaction is repeated (Liu, *et al.*, 2012).

Roche 454 sequencing could sequence much longer reads than Illumina, achieving over 700bp read lengths. Like Illumina, it does this by sequencing multiple reads at once by reading optical signals as bases are added. On top of the long-read length, another advantage of Roche 454 was that it can produce an output of 0.7Gbp in 24 hours which was improved to 14Gbp/run in 2009. However, one major disadvantage of the 454 system was that the output of 0.7Gbp and 14Gbp per run was low, and the cost a lot higher per base generated than Illumina technology, at an estimated US$10 per million bases (Liu *et al.*, 2012). The Roche 454 also has a relatively high error rate in terms of poly-bases longer than 6bp, which was a major shortcoming. Roche, however, shut down 454 sequencing in 2013 when its technology became noncompetitive (Holmer, 2013)

SOLiD system

The SOLiD sequencer adopts the di-base technology based on ligation sequencing (Huang *et al.*, 2012). The SOLiD flow cell consists of the ligation site, cleavage site and 4 different flourent dyes. The fluorescent signal will be recorded via the probes complementary to the template strand and vanished by the cleavage of probes' last 3 bases. The sequence of the fragment can be deduced after 5 rounds of sequencing using ladder primer sets (Liu *et al.*, 2012). The current version of SOLiD, the SOLiD 5500W has a read length, accuracy, and data output of 50bp, 99.99%, and 160Gbp per run respectively. A complete run could be finished within 6 days for single end reads and 12 days for paired end reads. The high accuracy of the AB SOLID system is a major advantage to genome sequencing. However, the short reads cause even more bioinformatics challenges than Illumina when applying it to processes like genome assembly, especially challenges in repeat resolution.

Ion Torrent: Proton / PGM / S5 sequencing

Ion Torrent launched the Personal Genome Machine (PGM) in the end of 2010 and uses semiconductor sequencing technology (Liu *et al.*, 2012). Instead of optical signals, Ion torrent and Ion proton use hydrogen the $(H)^+$ ion released when dNTP is added to a DNA polymer. This $(H)^+$ ion decreases pH and changes in pH allow sensors to determine if that base, and how many thereof, was added to the sequence read. Since the PGM does not require florescence and camera scanning, it produces reads of stable quality at a lower cost and less time. The PGM can produce reads of 200bp in 2 hours. The PGM has an output of 10Mbp-100Mbp per run, a throughput that is very low when considering the amount of data needed for a large-scale genome project.

Recently, in 2015, Ion Torrent launched and the Ion GeneStudio S5 series (Shin *et al.*, 2017) which produces from 2-3 million reads in approximately 3hours (Ion 510 Chip) to 100–130million reads in approximately 11hours (Ion 550chip). This is more data in less time compared to Proton and PGM.

1.1.2 Long read sequencers

Due to the shortcomings that come with short reads in genome assemblies, long fragment reads became necessary to address some of these challenges such as scaffolding and resolving repeat regions within the genome. Currently, sequencers have been introduced that produce long reads that are several kilobases long. There are two main types of long-read technologies: real-time sequencing approaches and synthetic approaches. Examples of real-time sequencing long read platforms are the PacBio sequencer from Pacific Biosciences (Quail *et al.*, 2012) and Oxford Nanopore Technologies sequencers such as the MinION (Jain *et al.*, 2015), and PromethION (Datema *et al.*, 2016). Examples of synthetic long read platforms are the Illumina synthetic long-read sequencing platform and the 10x Genomics emulsion-based system.

Real-time long-read sequencing

PacBio

The PacBio RS II uses Single Molecule Real Time sequencing (SMRT) (Quail *et al,* 2012). A single DNA polymerase enzyme is immobilized at the bottom of a reaction cell called a zero-mode waveguide (ZMW) cell, with a single molecule of DNA as a template (Eid *et al.*, 2009). Each sequencing plate contains ~3 000 individual cells with each holding only a single DNA molecule. During sequencing, a single labelled dNTP (each dNTP containing a different fluorophore) enters the polymerase and is held in position for a short period of time. A fluorescence signal is emitted in the ZMW during that time and signals are then collected from each ZMW. The dNTP leaves and then another molecule enters and the process continues. The DNA sequence of single molecule is determined by sequence of light pulses. The major advantage of PacBio technology is that it has a read length of about 10kb, which is significantly longer than all other short read sequencing technology available. However, the raw error rate is significantly higher than all the other sequencing technologies at 12.86% (Quail *et al,* 2013). Illumina HiSeq 2000 for instance has a raw error rate of 0.26% (Quail *et al,* 2013). Also, at US$2000 per Gb, the PacBio technology is more expensive than the other sequencing technologies.

Oxford nanopore technologies

Oxford Nanopore Technologies are currently developing a number of platforms that use nanopore technology to produce very long reads. Nanopore technology is often referred to as fourth generation sequencing. This technology utilizes nanometer sized pores that are either embedded in a biological membrane or formed in solid-state film which separates the reservoirs containing conductive electrolytes into cis and trans compartments (Feng *et al.*, 2015). In nanopore sequencing the DNA strand is analysed directly as the molecule is drawn through a tiny pore (nanopore) suspended in a membrane. Changes in electrical current, or tunnelling currents, are read off a chain of bases and interpreted as particular k-mers (Wilhelm, 2015). Oxford has introduced a number of platforms that use nanopore technology. These include the MinION (Jain *et al.,* 2016), PromethION (Datema *et al.*, 2016), GridION (Karow*,* 2017) and the SmidgION which is still under development. The advantage of nanopore sequencers is the direct detection the DNA composition of native ssDNA and not a secondary signal like colour, pH or light as in other platforms. Also, it does not need to monitor incorporations or hybridizations of nucleotides guided by a template DNA strand. Nanopore DNA sequencing do not require a great deal of sample preparation and complicated algorithms for data processing as is the case with most non-nanopore DNA sequencers (Feng *et al.,* 2015).

The MinION sequencer was introduced in 2014 and can produce data from 10G – 20Gbp with reads of up to 200kb in 48 hours. The PromethION uses the same technology as the MinION but produces larger amounts of data for users who require sequencing of many samples in parallel or the same sample in larger depth. The GridION X5, whose release was announced on the 14th of March 2017, is able to run up to five flow cells at a time, enabling it to generate up to 50Gbp of sequence data per 48-hour run with current chemistry and software (Karow, 2017). However, as with PacBio sequencer, the error rate for the MinION and PromethION was initially discouragingly high (~30%) (Feng *et al.,* 2015) but fortunately, recent developments in the chemistry and base calling algorithms are improving accuracy with a current error rate between 2-13%.

Synthetic long-reads sequencers

The synthetic long-read sequencing platforms rely on a system of barcoding to associate fragments that are sequenced on existing short-read sequencers (Voskoboynik *et al.*, 2013). Currently there are two systems available for generating long-reads; the Illumina synthetic long-read sequencing platform and the 10X Genomics system. The main difference between the Illumina system and the 10X Genomics system is that the Illumina system does not require special instrumentation to partition DNA into a microtiter plate while the 10X Genomics system uses emulsion to partition DNA and requires the use of a microfluidic instrument to perform pre-sequencing reactions. The error rate for Illumina long read sequencers as well as the 10X Genomics is similar to that of the existing short read sequencers but since the long-read sequencers require more coverage than the current short read systems, the cost of sequencing long reads is higher (Feng *et al.*, 2015). The cost of the 10X Genomics sequencer is slightly higher because of the additional cost of the microfluidic instrument.

Table 1.1: Comparison between different sequencing technologies. This table shows different sequencing mechanisms, read length and output per run for different sequencers.

Platform	Read length (bp)	Throughput	Reads	Runtime	Sequencing technology
SHORT READ SEQUENCING					
Illumina Systems					
Illumina MiniSeq Mid output	150 (SE)	2.1–2.4 Gb	14–16 M	17 h	Sequencing by synthesis
Illumina MiniSeq High output	75 (SE)	1.6–1.8 Gb	22–25 M (SE)	7 h	
	75 (PE)	3.3–3.7 Gb	44–50 M (PE)	13 h	
	150 (PE)	6.6–7.5 Gb		24 h	
Illumina MiSeq v2	36 (SE)	540–610 Mb	12–15 M (SE)	4 h	
	25 (PE)	750–850 Mb	24–30 M (PE)	5.5 h	
	150 (PE)	4.5–5.1 Gb		24 h	
	250 (PE)	7.5–8.5 Gb		39 h	
Illumina MiSeq v3	75 (PE)	3.3–3.8 Gb	44–50 M (PE)	21–56 h	
	300 (PE)	13.2–15 Gb			
Illumina NextSeq 500/550 Mid output	75 (PE)	16–20 Gb	Up to 260 M (PE)	15 h	
	150 (PE)	32–40 Gb		26 h	
Illumina NextSeq500/550 High output	75 (SE)	25–30 Gb	400 M (SE)	11 h	
	75 (PE)	50–60 Gb	800 M (PE)	18 h	
	150 (PE)	100–120 Gb		29 h	
Illumina HiSeq 2500 v2 Rapid run	36 (SE)	9–11Gb	300 M (SE)	7 h	

	50 (PE)	25–30 Gb	600 M (PE)	16 h	
	100 (PE)	50–60 Gb		27 h	
	150 (PE)	75–90 Gb		40 h	
	250 (PE)	125–150 Gb		60 h	
Illumina HiSeq 2500 v3	36 (SE)	47–52 Gb	1.5 B (SE)	2 d	
	50 (PE)	135–150 Gb	3 B (PE)	5.5 d	
	100 (PE)	270–300 Gb		11 d	
Illumina HiSeq 2500 v4	36 (SE)	64–72 Gb	2 B (SE)	29 h	
	50 (PE)	180–200 Gb	4 B (PE)	2.5 d	
	100 (PE)	360–400 Gb		5 d	
	125 (PE)	450–500 Gb		6 d	
Illumina HiSeq 3000/4000	50 (SE)	105–125 Gb	2.5 B (SE)	1–3.5 d	
	75 (PE)	325–375 Gb			
	150 (PE)	650–750 Gb			
Illumina HiSeq X	150 (PE)	800–900 Gb per flow cell	2.6–3 B (PE)	<3 d	
Solid Systems					
SOLiD 5500 Wildfire	50 (SE) and 75 (SE)	up to 160Gb	~700 M	6 d	Sequencing by ligation
SOLiD 5500 xl	50 (SE) and 75 (SE)	up to 320Gb	~1.4 B	10 d	
BGI Systems					
BGISEQ-500 FCS155	50–100 (SE/PE)	8–40 Gb	NA	24 h	Sequencing by Ligation
BGISEQ-500 FCL155	50–100 (SE/PE)	40–200 Gb	NA	24 h	
Roche system					
454 GS FLX Titanium XL+	Up to 1,000; 700 mode (SE, PE)	up to 700 Mb	~1 M	23 h	*Sequencing by synthesis: SNA*
Ion Torrent Sytems					
Ion PGM 314	200 (SE)	30–50	400 000–550 000	23 h	*Sequencing by synthesis: SNA*
	400 (SE)	60–100 Mb		3.7 h	
Ion PGM 316	200 (SE)	300–500 Mb	2–3 M	3 h	
	400 (SE)	600 Mb–1 Gb		4.9 h	
Ion PGM 318	200 (SE)	600 Mb–1 Gb	4–5.5 M	4 h	
	400 (SE)	1–2 Gb		7.3 h	
Ion Proton	Up to 200 (SE)	Up to 10 Gb	60–80 M	2–4 h	
Ion S5 520	200 (SE)	600 Mb–1 Gb	3–5 M	2.5 h	
	400 (SE)	1.2–2 Gb		4 h	
Ion S5 530	200 (SE)	3–4 Gb	15–20 M	2.5 h	

	400 (SE)	6–8 Gb		4 h	
Ion S5 540	200 (SE)	10–15 Gb	60–80 M	2.5 h	
LONG READ SEQUENCING					
Pacific BioSciences Systems					
Pacific BioSciences RS II	~20 Kb	500 Mb–1 Gb	~55,000	4 h	*Single-molecule real-time long reads*
Pacific Biosciences Sequel	8–12 Kb	3.5–7 Gb	~350,000	0.5–6 h	
Oxford Nanopore Technology					
Oxford Nanopore MK 1 MinION	Up to 200 Kb	Up to 1.5 Gb159	>100,000	Up to 48 h160	*Single-molecule real-time long reads*
Oxford Nanopore PromethION	up to 200 Kb159	Up to 4 Tb			
Synthetic long reads					
Illumina Synthetic Long-Read	~100 Kb synthetic length	Same as HiSeq 2500	same as HiSeq 2500	See HiSeq 2500	*Synthetic read technology*
10X Genomics	Up to 100 Kb synthetic length	Same as HiSeq 2500	Same as HiSeq 2500	See HiSeq 2500	

1.2 The relevance of guava

Guava (*Psidium guajava*), a member of the Myrtaceae family, is one of the most important fruit crops grown commercially across the tropics and sub-tropics (Hayes, 1966; Pathak and Ojha, 1993). High in vitamin A and B, the fruit is exceptionally rich in vitamin C when compared to other common winter fruits (Table 1.2). Guava fruit is estimated to be higher in vitamin C (184 mg/100g), calcium (20g/100g) niacin (1.2mg/100g) and fiber, while having a comparative higher level of iron (0.3mg/100g) and beta carotene than most other winter fruit. Colloquially, the guava fruit is generally known as the ‘Apple of Tropics and Sub-tropics’(Prakash, Narayanaswamy and Sondur, 2002; Rai *et al.*, 2010).

Table 1.2: The nutritional value of well-known winter fruits. Guava has the highest levels of Vitamin C, calcium and niacin when compared to more commercialized fruits (source http://www.guavaproducers.co.za/all-about-guavas_history.html).

Fruit	**Vitamin C (mg/100g)**	**Iron (mg/100g)**	**Beta Carotene. Re-retinol Equivs g/100g**	**Fibre g/100g**	**Calcium g/100g**	**Niacin mg/100g**
Guava	184	0.3	79	5.6	20	1.2
Paw paw	62	0.1	201	0.9	0.1	0.3
Orange	53	0.1	21	2.0	0.1	0.3
Grapefruit	34	0.1	12	0.6	12	0.3
Banana	9	0.3	8	3.0	6	0.5
Apple	6	0.3	5	3.1	7	0.1

In addition to the nutritional value of guava, it exhibits some pharmaceutical properties when used in traditional medicines. The fruit, bark and leaves have been used in folk medicine in treatment of ailments such as wounds, ulcers, bowls and cholera (Begum, Hassan and Siddiqui, 2002). The pharmacological properties of these plant parts have also been investigated for their antibacterial, hypoglycemic, anti-inflammatory, antipyretic, spasmolytic and central nervous system depressant activities (Begum, Hassan and Siddiqui, 2002).

Guava (*Psidium guajava*) is grown in Mpumalanga, Limpopo and the Western Cape provinces of South Africa (Schoeman, 2011). The origin of the cultivated guava, *Psidium guajava Linn,* can be traced to South and Central America and its sister species include the Brazilian guava (*Psidium guineense*), mountain guava (*Psidium montanum*), strawberry or cherry guava (*Psidium cattleianum*), Pineapple guava (*Acca sellowiana*) and Chilean guava (*Ugni myricoides*) (Mehmood *et al.*, 2013). In South Africa, approximately 1 200 ha are currently under guava production, mainly in the Western Cape, Limpopo and Mpumalanga. Total production is estimated at 27 000 tons, of which most being used in the food processing industry (Schoeman, 2011). In 2015, South Africa produced 33 574 tons of guava fruit with a gross value of R53 439 000 (Department of Agriculture, Forestry and Fisheries, 2015).

The guava fruit is botanically a berry and can vary in shape. It may be rounded, ovate, or pear shaped (Fig 1.2). The fruit may also vary in diameter and weight, from 25 to 102 mm and from 56 to about 450 respectively. The skin color of the ripe fruit is usually yellow, and the flesh color may be white, pink, yellow or cream. Guavas vary from thick-fleshed fruits with only a few seeds in a small central cavity, to thin fleshed fruits with numerous seeds imbedded in a large mass of pulp (Menzel, 1985).

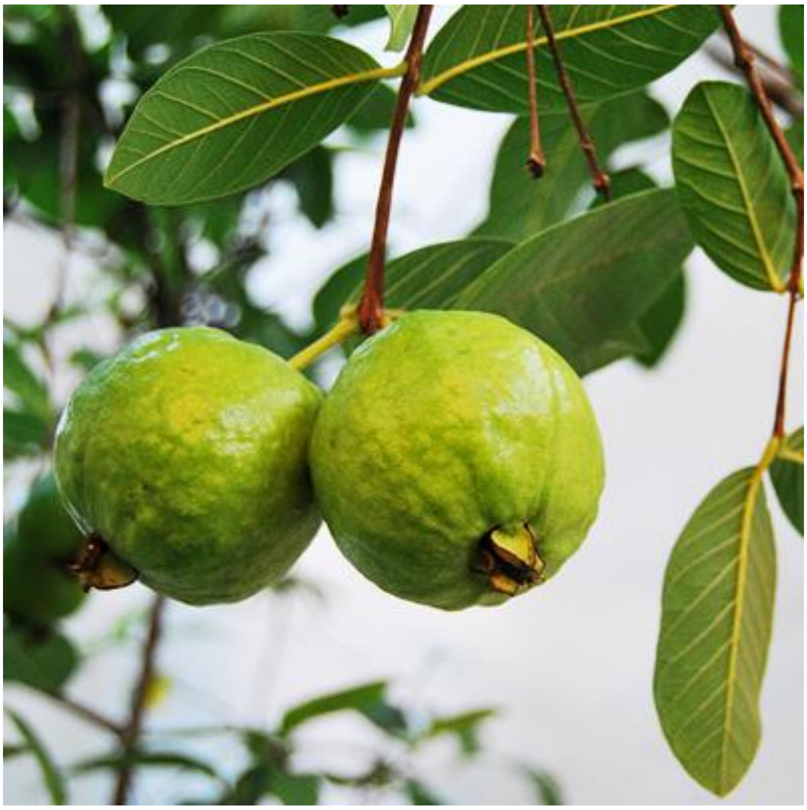


Figure 1.2: Photograph of a guava fruit (source: https://nurserylive.com).

1.3 The guava genome

The chromosome studies done on four species of *Psidium (P. acutangulum, P. catteyanum, P. cinereum and P. guajava)* from different populations showed that in the *Psidium* family only *P. guajava* is diploid (2n=22) while the other three family members are tetraploid (2n=4x=44) (da Costa Itayguara and Forni-Martins, 2006). Within *Psidium guajava,* studies have identified many diploid individuals or populations (2n=21, 22, 28, 20, 32 and 34) (D'Cruz and G.B, 1962; Majumder and Mukherjee, 1972), as well as polyploids, with 2n=33 (Kumar & Ranade, 1962) and 2n=44 (Srivastava, 1977b). The *Psidium* can therefore be described as having several cytotypes (chromosomic races), with the chromosome number varying in a diploid series (from 2n=21-34) or with different levels of ploidy (2n=22, 33 and 44) (da Costa Itayguara and Forni-Martins, 2006).

The C-value of an organism is the amount of nuclear DNA in its unreplicated gametic nucleus (Swift, 1950), irrespective of the ploidy level of the taxon (da Costa, Dornelas and Forni-Martins, 2008). The first species of fleshy-fruited Myrtaceae to be investigated was *Psidium guajava L.*, where the authors found different 2C-values, as estimated by Feulgen microdensitometry of 0.7pg (Bennett & Leitch, 2004) and 1.3pg (Ohri and Kumar, 1986) respectively, both samples having 2n = 22. Table 1.3 shows the genome size estimations in the *Psidium* genus. The values are given as the mean (at least 10.000 nuclei) and standard deviation of the mean of the haploploid nuclear DNA content (2C, pg DNA) of each species. The 2C range is presented by the minimum (Min.) and maximum (Max.) value obtained for each species. The monoploid nuclear DNA content (1Cx) in mass values (pg) and Mbp and the mean sample coefficient of variation of G0/G1 DNA peak (c.v., %) are also provided for each species. 1 pg DNA = 978 Mbp (Dolezel *et al.*, 2003).

Table 1.3: Chromosome numbers (2n) and genome size estimations for the fleshy-fruited *Psidium* genus. Two cultivars of P. guajava, the white cultivar and red cultivar genome sizes are 247.92Mbps and 269.44Mbps respectively (adapted from da Costa Itayguara & Forni- Martins, 2006).

Species	**2**	**Ploidy**	**Nuclear DNA content**					**c.v. (%)**
			2C(pg)	2C(range)		1Cx	1Cx	
				min	max	(pg)	(Mbps)	
P. pseudocariophyllus	22	2x	0.523±0.020	0.503	0.543	0.262	255.75	5.24
P. acutangulum	44	4x	1.167 ± 0.044	1.123	1.211	0.584	572.32	3.27
P. cattleianum	44	4x	1.053 ± 0.040	1.013	1.093	0.526	515.48	3.89
P. guajava (white)	22	2x	0.507 ± 0.019	0.488	0.526	0.254	247.92	4.32
P. guajava (red)	22	2x	0.551 ± 0.021	0.530	0.572	0.276	269.44	5.03

Genomics is an area of genetics that concerns the sequencing and analysis of an organism's genome (Miko and LeJeune, 2009). It applies DNA sequencing methods and bioinformatics algorithms to sequence, assemble and analyze the function and structure of genomes. We aim to explore bioinformatics methods used to create resources that will serve as a starting point to utilize genomics resources in guava breeding and cultivation.

The guava industry faces many challenges, one being the recent outbreak of Guava Wilt Disease (GWD), which is causing severe economic losses to the industry. Successful breeding techniques employed by the Agricultural Research Council's Institute for Tropical and Subtropical Crops (ARC-ITSC) in the 1990's (Schoeman, 2011) produced rootstocks resistant to the disease, but a resistant strain of the causal pathogen *Nalanthamala psidii* is currently causing severe losses in the industry. A genomic resource for guava will be a starting point in addressing this problem and others using genomics and bioinformatics.

1.4 Guava wilt disease

Guava wilt disease (GWD) was first reported in South Africa in the 1980s (Grech, 1985) in Malane, Mpumalanga. The outbreak of the disease caused severe losses in the guava industry, both to guava plantations and in monetary terms. It spread across the whole Mpumalanga and Limpopo as the guava industry in these areas relied solely on a single, highly susceptible cultivar, 'Fan Retief' (FR) (Grech, 1987). The fungus causing GWD is classified as *Nalanthamala psidii* (Schroers *et al.*, 2005). GWD causes the leaves of the plant to shrivel (Fig 1.4) and die resulting in complete defoliation (Schoeman *et al.*, 1997). In 1995 two resistant rootstocks and one tolerant guava rootstock were developed by the Agricultural Research Council's Institute for Tropical and Subtropical Crops (ARC-ITSC) (Schoeman, 2011). However, in 2009 there was a renewed outbreak caused by a resistant strain of *N. psidii,* placing the South African guava industry once again under threat.

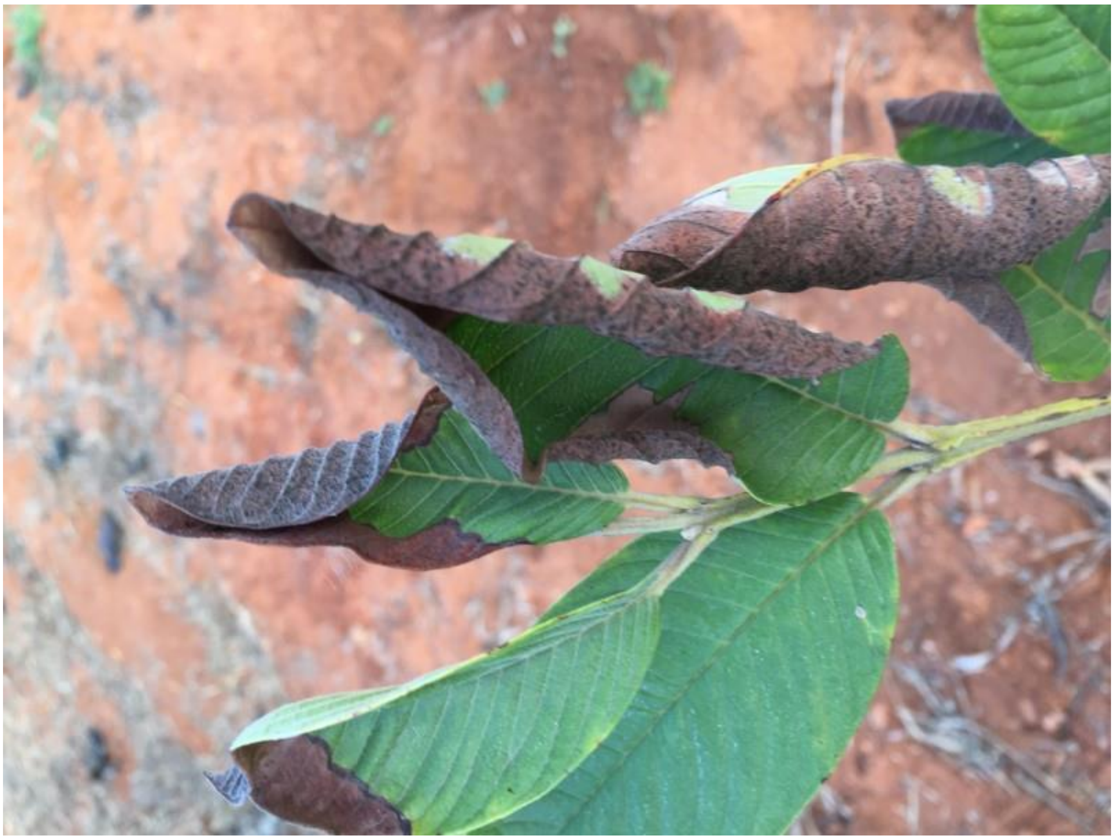


Figure 1.3: Picture of a guava fruit tree affected by GWD. (Source: https://discuss.farmnest.com/t/guava-wilt-any-organic-remedies/3699)

Genome and transcriptome work on the guava genome present an opportunity to address the GWD problem and other challenges facing the guava industry. Understanding the molecular nature of host-pathogen interactions assists breeders to identify genomic resources for breeding against diseases (Meyer *et al.*, 2005). Creating a genomic resource for *Psidium guajava* would start by determining the sequence of the host's (*P. guajava*) genome, as well as that of pathogens such as *N. psidii.*

1.5 Whole genome assembly

Whole Genome Shotgun Assembly (often called simply Whole Genome Assembly or WGA) is the process of sequencing the entire genome of an organism by ordering and orienting sequenced reads. Overlapping reads are aligned together to form a contiguous sequence of base pairs called contigs. These are then used to make super contigs/scaffolds, which consist of many contigs separated by gaps of known or estimated sizes. Figure 1.4 shows the general workflow for whole genome assembly.

This illustrates how a genome is first fragmented into fragments, sequenced by sequencers to form reads and then finally stitching the reads to form the original genome sequence. The stitching of the reads involves various bioinformatics algorithms.

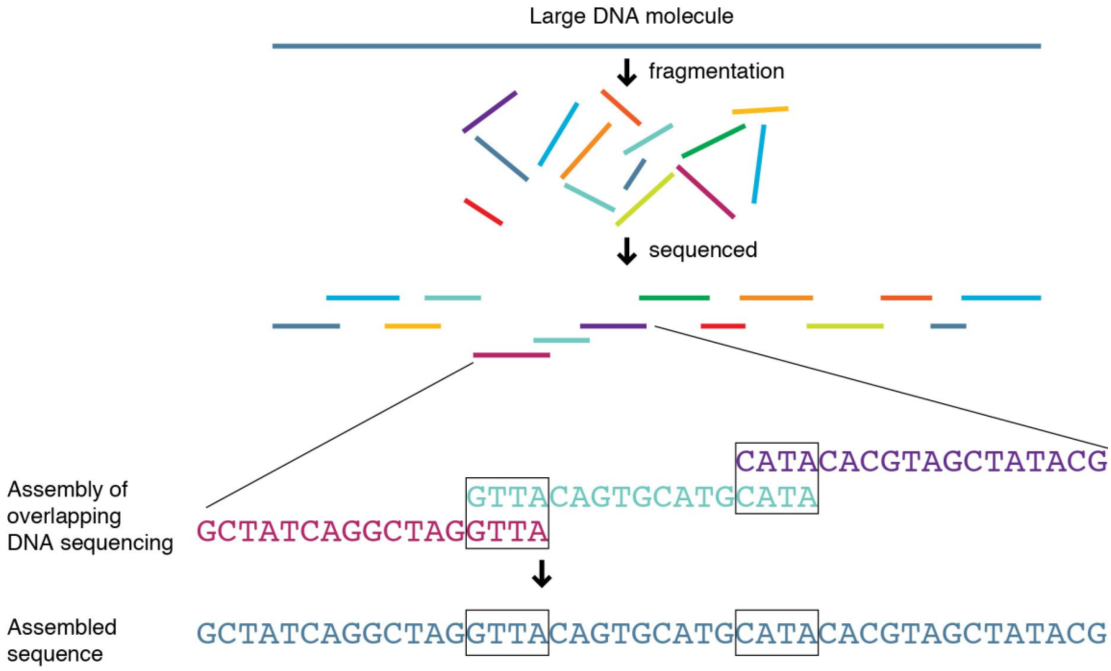


Figure 1.4: Diagram showing the whole genome assembly process (Source: http://www.genome.gov).

Studies done by Lander & Waterman (1988) explored conditions that are necessary for an assembly to be possible. They explored how read length, coverage of the genome and overlap between two reads can affect an assembly. The coverage of a genome is number of total bases in the set of reads divided by the length of the genome. Their work showed that genome sequences could be effectively reconstructed with as low as tenfold coverage of sequence reads (Simpson and Pop, 2015) when Sanger reads are used. Also, work by Ukkonen and others showed that finding the correct solution for a genome assembly may require exploring an exponential number of possible solutions. Research done to find the best solution led to the first practical genome sequence assemblers (Simpson & Pop, 2015).

1.5.1 Challenges with whole genome assembly

WGA is affected by many inherent challenges including sequencing errors, repeats and polymorphisms. Sequencing errors arise from miscalls from different sequencing methods. These sequencing errors result in assemblers constructing contigs that consisting of unambiguous, unbranching regions of the genome (Simpson & Pop, 2015). NGS reads come with quality scores associated with each nucleotide. These quality scores (PHRED scores) are values that are given to each base to show the likelihood of the base not being a miscall. Various packages like Trimmomatic (Bolger, Lohse and Usadel, 2014), Quake (Kelley, Schatz and Salzberg, 2010) and cut-adapt (Martin, 2011) are used to trim or filter these miscalls and improve the quality of reads.

Repeats are sequences that occur in more than one place in the genome. Repetitive elements pose a serious problem for WGS assembly since it is mathematically impossible to reads sequenced from different but identical-looking regions (Simpson & Pop, 2015). Different reads with identical looking reads may form false overlaps which ultimately form misassembles. Moreover, genomes with repeats longer than read length are particularly harder to assemble as it is impossible to bridge that gap. However, mate-pair information can be used to resolve these repeats. Since mate-pair information spans are made from large DNA fragments (between 3kb and 20kb in size), they offer valuable constraints on the relative placement of sequence reads in an assembly (Simpson & Pop, 2015).

Genetic polymorphism can be defined as the existence of multiple alleles or forms at a genomic locus for genomes in the same population. A high level of with-in population polymorphisms may significantly increase the complexity of genome assembly and most assemblers assume a polymorphic rate of <1%. Assemblers such as Hapsembler (Donmez and Brudno, 2011), have been developed specially for highly polymorphic genomes by utilising a mate-pair graph that is essentially an overlap graph built from read pairs instead of single reads and is very useful in resolving repeats in addition to the ambiguities caused by polymorphisms (Donmez and Brudno, 2011).

1.5.2 Types of whole genome assembly

There are three general approaches to whole genome assembly. These include *de novo* genome assembly, reference-based genome assembly and combined genome assembly. NGS technologies typically produces libraries with fastq files containing nucleic acid sequence reads together with quality scores that are used for performing *in silico* assemblies.

1.5.2.1 De novo genome assembly

De novo genome assembly is a form of assembly where no previously assembled genome (a reference genome) is present. There are various publicly available packages for *de novo* genome assembly. Examples of such packages (assemblers) are ALLPATHS-LG (Butler *et al.*, 2008), SOAP-Denovo (Luo *et al.*, 2012), AbySS (Simpson *et al.*, 2009) and SGA (Simpson and Durbin, 2012) among others. The two basic approaches that assemblers use are (a) greedy-based approaches and (b) graph-based approaches.

(a) Greedy-based approach

This strategy for assembling a genome sequence involves iteratively joining together the reads in decreasing order of the quality of their overlaps (Simpson & Pop, 2015) consequently growing contigs. The greedy approach has shown to be a good approximation for the optimal assembly and is the underlying approach to assemblers like phrap (Green, 1994) and TGIR Assembler (Sutton *et al.*, 1995), which were used in the Human Genome Project. The greedy approach, however, has its limitations. It fails to effectively handle repetitive regions and this is due to its local assembly nature. It also has high computational times. Because of this, the greedy approach has been replaced by more effective graph-based approaches.

(b) Graph-based approach

The graph-based approach was developed from theoretical studies of graphs and graph theory. This approach converts reads into nodes/vertices and edges and the genome to be a path of ordered nodes joined by edges. Three basic models that use the graph-based approach are (i) overlap graph construction, (ii) the *de Bruijn* method and (iii) string graphs.

(i) The overlap graph approach

The overlap graph construction method was developed for Sanger reads, although some assemblers for next-generation sequence data also use this approach (Illumina Inc, 2012). The overlap graph approach assemblers compute all pair-wise overlaps between the reads and capture this information in a graph. Each node of the graph corresponds to a read, and an edge denotes an overlap between two reads. The graph construction is in three stages (i) overlapping reads are detected (ii) the graph is constructed, and an appropriate ordering and orientation (layout) of the reads are found and (iii) a consensus sequence is computed from the ordered and oriented reads (Simpson & Pop, 2015). This method is however computationally intensive such that only a small percentage of the available assemblers use this method (Illumina Inc., 2012). The overlap graph has been successful and has been used by assemblers such as the Celera assembler (Myers *et al,* 2000) and Newbler (Wheeler *et al.*, 2008). Despite its success, however, the overlap graph strategy struggles when presented with the vast amounts of short-read data generated by the next generation of DNA sequencing instruments (Simpson and Pop, 2015).

(ii) The *de Bruijn* method

The *de Bruijn* method was first introduced by Pevzner (Pevzner, MYu and Mironov, 1989) in the 1980s. It reduces computational effort by breaking reads into smaller sequences of DNA called *k*-mers, where *k* is the length of the small sequences. The *de Bruijn* graph captures overlaps of length *k*-1 between these *k*-mers and not between actual reads (Illumina Inc., 2012). The assembly problem can then be formulated as finding a walk through the graph that visits each edge in the graph once—also known as an Eulerian path problem (Simpson and Pop, 2015). However, in a typical genome assembly problem, the Eulerian path will produce an exponential number of paths. Because of this, most assemblers that use the *de Bruijn* method do not seek the completion of the Eulerian path, rather the assembler attempts to construct contigs consisting of the unambiguous, unbranching regions of the graph. *De Bruijn* graphs also reduce the challenges that repeats in the genome pose to genome assembly, as it can collapse repeats in the genome in the graph and does not lead to many spurious overlaps. Given a *k*-mer substring of a genome, coverage is defined as the number of reads to which this *k*-mer belongs (Compeau and Pevzner, 2015). Breaking reads into short *k*-mers ensures a better coverage of the *k*-mers. However, the smaller *k* results in a more tangled *de Bruijn* graph, making it difficult to infer the genome from this graph (Compeau and Pevzner, 2015). A bigger *k* on the other hand, is helpful in repeat resolution. Again, having a *k* value that is too big will lead to imperfect coverage reducing the chances of obtaining a correct assembly size. Tools such as jellyfish (Marcais and Kingsford, 2012) and kmergenie (Chikhi and Medvedev, 2014) are used to estimate the best *k* value to use for a particular set of reads. Most recent assemblers use the *de Bruijn* construction. Examples are: The ABySS (Assembly by Short Sequences) assembler (Simpson *et al.*, 2009), SOAP-Denovo assembler (Luo *et al.*, 2012) and ALLPATHS-LG (Butler *et al.*, 2008).

(iii) String graphs

The approach is based on the overlap graph but performs two other transformations to collapse repeats just like the *de Bruijn* graph. The two transformations to the overlap graph are: firstly, contained reads that are substrings of some other read are removed. Secondly, transitive edges are removed from the graph (Simpson and Pop, 2015). This forms a graph (string graph), which has properties of *de Bruijn* without the need to break reads into *k*-mers. The Edena assembler (Hernandez *et al.*, 2008) is the first assembler to use the string graph approach but other string graph algorithms such as the String Graph Assembler (Simpson and Durbin, 2012) have been developed since.

1.5.2.2 Reference-based assembly

As the name implies, the reference-based approach entails the use of a genome as a blueprint for assembly. This strategy requires three steps: read alignment; overlap graph construction and isoform resolution (Martin and Wang, 2011; Florea and Salzberg, 2013).

A wide range of aligners exists for alignment of various types of nucleic acid and protein sequence information (Flicek and Birney, 2009; Martin and Wang, 2011). Aligners such as BWA (Li and Durbin, 2009) or Bowtie (Langmead *et al.*, 2009) can be subdivided into two classes based on their underlying algorithms (Flicek and Birney, 2009); seed-and-extend aligners use a hash table-based approach that relies on heuristic techniques to align 'seed' sequences to the genome followed by Smith-Waterman alignment algorithms to extend local alignments, and Burrows-Wheeler transform-based (Burrows and Wheeler, 1994) aligners that use a condensed suffix array-based approach (Flicek and Birney, 2009).

A typical BWA alignment will start by creating an index/database of the reference genome to speed up the mapping. Paired-end reads are aligned separately and then combined. The next steps involve creating a sequence alignment map file and finally an index for the alignment file. Any unmapped reads can then be identified and used for a *de novo* assembly if necessary, to improve assembly.

1.5.2.3 Combined assembly

Both *de novo* and reference-based strategies of genome assembly have distinctive advantages and disadvantages. Consequently, if a suitable genome sequence is available, these strategies can be combined to complement each other (Martin and Wang, 2011; Jain, Krishnan and Panda, 2013). Conversely, a combined approach runs the risk of losing sensitivity due to compounded errors in the reference genome (Jain, Krishnan and Panda, 2013). Two approaches to combined assembly are align-then-assemble or assemble-then-align. (Martin *et al.*, 2010).

The align-then-assemble approach entails performing reference-guided assembly and then using the unmapped reads as input for *de novo* assembly. Alternatively, the align-then-assemble approach can use both reference-guided assembled contigs and the failed reads in cases where the *de novo* assembler can use long reads (Martin and Wang, 2011). The assemble-then-align approach is the reverse, where data is first used for *de novo* assembly of a genome, followed by scaffolding and extension of assembled contigs by aligning them to the reference genome. This approach is mainly used if the quality of the genomic data is in question. Tools such as AlignGraph (Bao, Jiang and Girke, 2014) use an algorithm for secondary *de novo* genome assembly guided by closely related references. This algorithm uses contigs obtained from a *de novo* genome assembler such as ALLPATHS-LG and an assembled genome of a closely related species. First it aligns paired-end reads to both the pre-assembled contigs and the closely related reference genome. It also aligns contigs to the reference genome. Secondly, the alignment mapping results are used to construct a positional variant of the *de Bruijn* graph, called the paired-end multipositional *de Bruijn* graph (Bao, Jiang and Girke, 2014). Finally, the resulting graph is edited and traversed to obtain extended contigs.

1.5.3 Assembly pipeline

Assembling a genome requires prior planning. An assembly pipeline can be proposed, implemented and modified depending on assembly output quality and challenges. A typical pipeline is shown below in Figure 1.5. A major step in the assembly is the pre-assembly process that involves quality control of NGS reads and trimming. Also, post-processing like quality assessment is also essential.

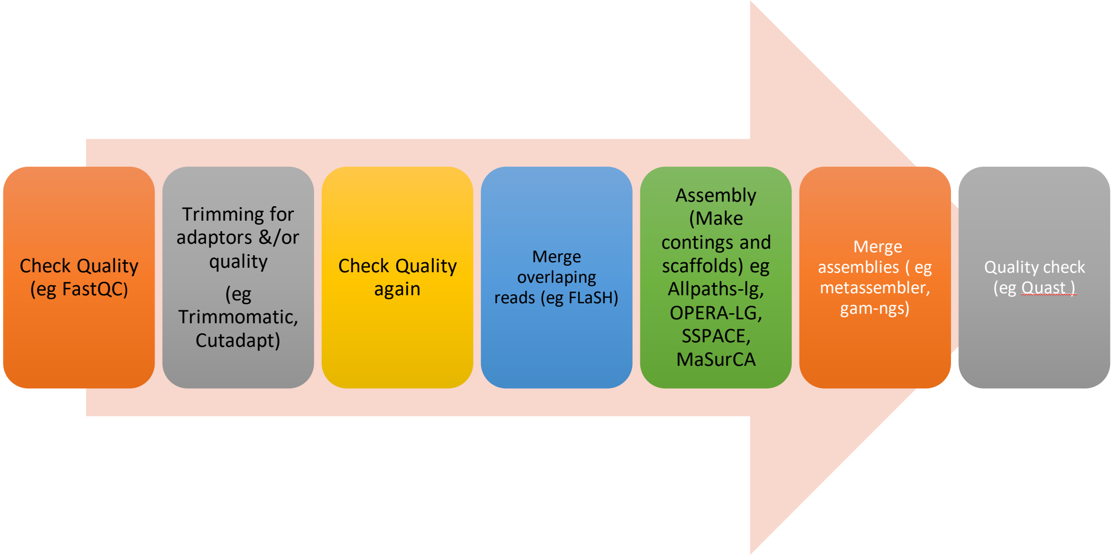

Check Quality (eg FastQC)
Trimming for adaptors &/or quality (eg Trimmomatic, Cutadapt)
Check Quality again
Merge overlaping reads (eg FLaSH)
Assembly (Make contings and scaffolds) eg Allpaths-lg, OPERA-LG, SSPACE, MaSurCA
Merge assemblies (eg metassembler, gam-ngs)
Quality check (eg Quast)

Figure 1.5: Typical assembly pipeline showing the pre-assembly steps (first 4 steps) the assembly and post assembly steps. Examples of tools used at each stage are given.

1.5.3.1 Pre-assembly processing

Before an assembler is used on NGS reads, the reads have to be assessed for quality using quality control tools such as FastQC (Andrew, 2016) and Prinseq (Schmieder and Edwards, 2011). These tools assess the quality of the raw reads by looking at aspects such as per base sequence quality, *k*-mer content and adapter content among others.

Based on the results of quality control tools, appropriate action will be taken on the NGS reads to correct/clean up the reads depending on the assembler to be used. Trimming is a common action taken to reads. Several tools are publicly available and are used for removing adaptors and for trimming on quality of the reads. Examples of these trimming tools are Trimmomatic (Bolger, Lohse and Usadel, 2014) and Cutadapt (Martin, 2011). Other tools like FlaSh (Magoc and Salzberg, 2011) which is used to merge overlapping paired end reads can also be used to improve the length of input reads, hence improving assembly. A recommended practice is to use quality control tools on the reads again after trimming and merging reads before one can use them for assembly. The processing done on the reads before assembly depends on the assembler to be used. For example, ALLPATHS-LG only requires reads to have its adapters removed since it does its quality trimming and merging on its own. SOAP-Denovo, however, requires reads to be trimmed for quality and merged if possible.

1.5.3.2 Assessment of assembly quality

Different assemblers use different methods to assemble the same genome. Because of this, the quality of assemblies tends to differ. Quality assessment of assemblies as well as comparing different assemblies is therefore essential. Recently, there has been a lot of work on developing comprehensive ways to compare different assemblers (Gurevich *et al.*, 2013). Packages like Plantagora (Barthelson *et al.*, 2011), GAGE (Salzberg *et al.*, 2012) and QUAST (Gurevich *et al.*, 2013) have been used to assess different assemblers. Plantagora and GAGE can only be used to evaluate assemblies of datasets with a known reference genome; thus, they are not suitable for evaluating assemblies of previously unsequenced genomes. QUAST however, is not restricted to assemblies with reference genomes. These use a number of metrics to assess the assembly quality and compare it with other assemblies. Comparing metrics that are given in each assembly report can also be done by hand. There are several metrics that can be used for assessment. These include: (i) contig sizes, (ii) genome misassemblies and structural variations, (iii) genome representation and its functional elements and (iv) variations of N50 based on aligned blocks among others. The contig N50 of an assembly is the length of the shortest contig in a list of contigs ordered in descending order of size where the cumulative length of the list is at least 50% of the total length.(Yandell and Ence, 2012). The longer the scaffold N50, the better the assembly, although erroneous N50 values can be calculated when unrelated reads are used. The percentage genome coverage and gene coverage can also determine the quality of an assembly. Genome coverage is the fraction of the genome represented in the assembly based on genome size estimates while gene coverage is the fraction of the genes in the genome that are contained in the assembly. A good genome coverage is between 90-95%. Based on these metrics a general comparison can be made between different assemblies. However, it should be noted that one assembly could be better when considering one metric and worse when considering another. This makes comparing assemblies a complicated task. Another way to determine the completeness of an assembly is using CEGMA (Parra, Bradnam and Korf, 2007), which checks the percentage of the universal eukaryotic single-copy genes found in the assembly and also determines the percentage of each gene lying on a single scaffold. Also, BUSCO, **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs (Simão *et al.*, 2015) can be used to assess the completeness of a genome by using expectations of gene content from near-universal single-copy orthologs which are evolutionarily-informed.

1.6 Genome annotation

Genome annotation can be defined as a subfield in the general field of genome analysis, which includes the downstream analysis performed with genome sequences by computational means (Koonin and Galperin, 2003). It can also be defined as the description of an individual gene and its protein (or RNA) product. The curation of the genome assemblies by humans can be inconsistent and error-prone, hence the incentive for automating as much of the annotation process as possible.

There are a lot of packages publicly available for genome annotation. These include MAKER (Campbell *et al.*, 2002), GeneQuiz (Scharf *et al.*, 1994), PEDANT (Walter *et al.*, 2009) and MAGPIE (Gaasterland and Sensen, 1996) among others. Genome annotations are referred to as pipelines since a lot of tools are involved in the process (Yandell and Ence, 2012). These genome annotation pipelines share a set of common features that can be divided into two phases: (i) computation phase and (ii) annotation phase.

1.6.1 The computational phase

In this phase, evidence-driven gene predictions and/or *ab initio* evidence is generated by first identifying repeats in the genome and then aligning proteins expresses sequence tags (ESTs), transcriptomic data and proteins to the genome. Repeat detection is done by tools such as RepeatMasker (Smit *et al.*, 2013) and alignment is done by various tools such as TBLASTX, BLAST (Korf, Yandell and Bedell, 2003), BLAT (Kent, 2002), TopHat (Trapnell, Pachter and Salzberg, 2009) and GMAP (Wu and Watanabe, 2005).

Ab initio gene predictors use mathematical and statistical models rather than external evidence to identify genes and to determine their intron–exon structures while evidence driven gene predictors use ESTs, for example, to identify exon boundaries unambiguously. (Yandell and Ence, 2012). *Ab initio* gene predictors need to be trained on the genome that is under study, as even closely related organisms can differ with respect to intron lengths, codon usage and GC content. The MAKER pipeline provides a simplified process for training the predictors Augustus (Stanke and Waack, 2003) and SNAP (Korf, 2004) using the EST, protein and mRNA-seq alignments that MAKER has produced (Yandell and Ence, 2012).Table 1.4 below shows four basic categories of gene identification programs.

Table 1.4: A table showing four basic categories of gene prediction programs (Source: Wei *et al.*, 2002).

Category	Algorithm	url
Based on direct evidence of transcription	EST_GENOME	http://www.hgmp.mrc.ac.uk/Registered/Option/est_genome.html
	Sim4	http://globin.cse.psu.edu/
Based on homology with known genes	PROCRUTES	http://igs-server.cnrs-mrs.fr/igs/banbury/Procrustes-about.html
Statistical/ *ab initio* approaches	GeneScan	http://genes.mit.edu/GENSCAN.html
	Genie	http://www.fruitfly.org/seq_tools/genie.html
	FGENESH	http://genomic.sanger.ac.uk/gf/Help/fgenes.html
	GeneMark_hm	http://opal.biology.gatech.edu/GeneMark/
	HMMGene	http://www.cbs.dtu.dk/services/HMMgene/
	Glimmer	http://www.tigr.org/software/glimmer/glimmer.html
Using genome comparison	Twin Scan	http://genes.cs.wustl.edu/
	Rossetta	http://crossspecies.lcs.mit.edu/
	SGP-1	http://soft.ice.mpg.de/sgp-1

1.6.2 The annotation phase.

In this phase, the evidence for each predicted gene is reviewed, either manually or by automated methods to decide on their intron-exon structures. In automated annotation, a 'chooser algorithm' is used to select the best prediction from different gene predictors (Yandell and Ence, 2012). This is the process used by JIGSAW, EvidenceModeler (EVM) (Haas *et al.*, 2008) and GLEAN (Elsik *et al.*, 2007). Alternatively, alignment evidence can be fed to the gene predictors at run time to improve the accuracy of the prediction process and a chooser can then be used to identify the most representative prediction. This is the process used by PASA, Gnomon and MAKER (Yandell and Ence, 2012).

1.7 Comparative genome analysis

Comparative genomics is a process of comparing two or more genomes to discover the similarities and differences between the genomes and to study the biology of the individual genomes (Wei *et al.*, 2002). It can be done on whole genomes or synthenic regions of different species/ subspecies/ strains of the same species. Comparisons can be done for: (i) genome structure (ii) coding regions and (iii) noncoding regions.

1.7.1 Comparing genome structure

Comparative genomics on genome structure can explore overall nucleotide statistics or genome structure at DNA level and gene level. Nucleotide statistics include genome size, overall GC content and regions of different GC content. Comparison of genome structure at DNA level can compare breakage and exchange of chromosomal fragments, which are common mode of genome evolution. This can also cause disruption of gene order. Comparisons of the conservation of these genomes can also be done and computational tools available, such as GRIMM (Tesler, 2002) can be used for this.

1.7.2 Comparing coding regions

Comparison of coding regions between two or more genomes can be used to discover similarities or differences between genomes by calculating the percentage of genes that are common among the genomes, unique genes for each genome compared to known sequences in all other species in databases such as Genbank and genes that are unique to each genome compared to the other genomes. Typically, performing such a comparison involves identifying of gene-coding regions, comparing gene and protein content (Wei *et al.*, 2002). This is often done using a pairwise sequence comparison tool such as BLASTN or TBLASTX (Wei *et al.*, 2002).

Comparative genomic approaches can be used to assist in the functional assignment of genes in a non-similarity-based manner. These approaches include co-conservation across genomes, conservation of gene clusters and genomic context across species, and physical fusion of functionally linked genes across species. Co-conservation across genome approach allows one to establish phylogenetic profiles for a gene by observing the presence or absence of a gene in a genome across many genomes. The conservation gene clusters approach uses the fact that functionally related genes tend to be located in close proximity to each other, and often in specific order.

The "Rosetta Stone" sequence or "composite" protein is a protein chain formed when certain pairs of interacting and functionally related proteins are fused into one protein chain (Wei *et al.*, 2002). If this composite protein is similar to two component proteins in another species, the two proteins are likely to be interacting and/or functionally related (Marcotte *et al.*, 1999).

1.7.3 Comparing noncoding regions

Non–coding regions can comprise up to 97% of some genomes. These have been identified to be mostly regulatory elements for processes like replication. The functional non-coding regions are conserved regions mainly due to selective pressure which causes regulatory elements to evolve at a slower rate than that of non-regulatory sequences in the noncoding region. The specificity for regulatory region detection increases significantly when more than two species are used in the comparative analysis (Wei *et al.*, 2002).

1.8 Conclusion

Creating a genomic resource for a tropical plant, *Psidium guajava*, is an extremely involved process, from sequencing of the DNA to annotation to comparative genomics analysis of the assembled genome. The choice of sequencing technology and library type is crucial for the other downstream processes as it can make an enormous difference in the budget and application of reads. Illumina technology has proved to be a cost efficient and effective way in sequencing a tropical plant like guava. However, strict quality control is required during library construction and pre-assembly to ensure that a good quality assembly is produced. A careful selection and careful use of bioinformatics packages is to be done for all the processes in the creation of this genomic resource. A lot of publicly available software packages and pipelines offers a wide range of choices in all the process involved. However, it should be noted that not one software package has all the best qualities; a software package can be good in one aspect but not so good in another. Therefore, experimenting with these different packages is essential to come up with an optimal set of processes and packages that will produce the best result.

The aim of this study is to create a genomic resource for *Psidium guajava* using NGS technologies and several bioinformatics packages. We will create the first annotated draft assembly of the guava fruit tree and perform some comparative genomics work between the guava genome and another member of the Myrtaceae family, *Eucalyptus grandis*. This work will serve as a starting point in addressing problems in the guava industry such as the Guava Wilt Disease. Having a reference genome of *P. guajava* can be used to design SNP markers that will be used in marker assisted selection in breeding projects.

1.9 Bibliography

Andrew, S. (2016) *FastQC A Quality Control tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bao, E., Jiang, T. and Girke, T. (2014) 'AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references', *Bioinformatics*. Oxford University Press, 30(12), pp. i319–i328. doi: 10.1093/bioinformatics/btu291.

Barthelson, R. *et al.* (2011) 'Plantagora: Modeling Whole Genome Sequencing and Assembly of Plant Genomes', *PLOS ONE*. Public Library of Science, 6(12), p. e28436. doi: 10.1371/journal.pone.0028436.

Begum, S., Hassan, S. I. and Siddiqui, B. S. (2002) 'Two new triterpenoids from the fresh leaves of Psidium guajava.', *Planta medica*. Germany, 68(12), pp. 1149–1152. doi: 10.1055/s-2002-36353.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Genome analysis Trimmomatic : a flexible trimmer for Illumina sequence data', pp. 1–7. doi:10.1093/bioinformatics/btu170.

Burrows, M. and Wheeler, D. J. (1994) 'A block-sorting lossless data compression algorithm', *Systems Research*, Research R(124), p. 24. doi: 10.1.1.37.6774.

Butler, J. *et al.* (2008) 'ALLPATHS: De novo assembly of whole-genome shotgun microreads', *Genome Research*, 18(5), pp. 810–820. doi: 10.1101/gr.7337908.

Campbell, M. S. *et al.* (2002) 'Genome Annotation and Curation Using MAKER and MAKER-P', in *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. doi: 10.1002/0471250953.bi0411s48.

Chikhi, R. and Medvedev, P. (2014) 'Informed and automated k-mer size selection for genome assembly', *Bioinformatics*, 30(1), pp. 31–37. doi: 10.1093/bioinformatics/btt310.

Compeau, P. and Pevzner, P. (2015) *Bioinformatics Algorithms: An Active Learning Approach*. Active Learning Publishers. Available at: https://books.google.co.za/books?id=ddHqjwEACAAJ.

da Costa, I. R., Dornelas, M. C. and Forni-Martins, E. R. (2008) 'Nuclear genome size variation in fleshy-fruited Neotropical Myrtaceae', *Plant Systematics and Evolution*, 276(3–4), pp. 209–217. doi: 10.1007/s00606-008-0088-x.

da Costa Itayguara, R. and Forni-Martins, E. R. (2006) 'Chromosome studies in Brazilian species of *Campomanesia* Ruiz &amp; Pávon and *Psidium* L. (Myrtaceae Juss.)', *Caryologia*, 59(1), pp. 7–13. doi: 10.1080/00087114.2006.10797891.

D'Cruz, R. and G.B, R. (1962) 'Cytogenetic studies in two guava aneuploids', *Journal of the Indian Bo- tanical Society*, 41(2), pp. 316–321.

Datema, E. *et al.* (2016) 'The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only.', *bioRxiv*. Available at: http://biorxiv.org/content/early/2016/11/01/084772.abstract.

Dolezel, J. *et al.* (2003) 'Nuclear DNA content and genome size of trout and human.', *Cytometry. Part A : the journal of the International Society for Analytical Cytology*. United States, p. 127–8; author reply 129. doi: 10.1002/cyto.a.10013.

Donmez, N. and Brudno, M. (2011) 'Hapsembler: An assembler for highly polymorphic genomes', *Lecture Notes in Computer Science (including subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6577 LNBI, pp. 38–52. doi: 10.1007/978-3-642-20036-6_5.

Eid, J. *et al.* (2009) 'Real-time DNA sequencing from single polymerase molecules', *Science*, 323. doi: 10.1126/science.1162986.

Elsik, C. G. *et al.* (2007) 'Creating a honey bee consensus gene set', *Genome Biology*, 8(1), p. R13. doi: 10.1186/gb-2007-8-1-r13.

Feng, Y. *et al.* (2015) 'Nanopore-based Fourth-generation DNA Sequencing Technology', *Genomics, Proteomics & Bioinformatics*. Elsevier, 13(1), pp. 4–16. doi: 10.1016/j.gpb.2015.01.009.

Flicek, P. and Birney, E. (2009) 'Sense from sequence reads: methods for alignment and assembly', *Nat Meth*. Nature Publishing Group, 6(11s), pp. S6–S12. Available at: http://dx.doi.org/10.1038/nmeth.1376.

Florea, L. D. and Salzberg, S. L. (2013) 'Genome-guided transcriptome assemble in the age of next-generation sequencing', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(5), pp. 1234–1240. doi: 10.1016/j.micinf.2011.07.011.Innate.

Gaasterland, T. and Sensen, C. W. (1996) 'MAGPIE: automated genome interpretation.', *Trends in genetics : TIG*. England, 12(2), pp. 76–78.

Gharizadeh, B. *et al.* (2006) 'Large-scale Pyrosequencing of synthetic DNA: A comparison with results from Sanger dideoxy sequencing', *ELECTROPHORESIS*. WILEY-VCH Verlag, 27(15), pp. 3042–3047. doi: 10.1002/elps.200500834.

Grech, N. M. (1985) 'First report of guava rapid death syn - drome caused by Septofusidium sp. In South Africa', *Plant Disease,*.

Grech, N. M. (1987) 'Guava Wilting Disease: The Cape Scenario', *Citrus and Subtropical FruitInformation Bulletin*, 179.

Green, R. E. *et al.* (2010) 'A Draft Sequence of the Neandertal Genome', *Science*, 328(5979), p. 710 LP-722. doi: 10.1126/science.1188021.

Gurevich, A. *et al.* (2013) 'QUAST: quality assessment tool for genome assemblies.', *Bioinformatics (Oxford, England)*, 29(8), pp. 1072–5. doi: 10.1093/bioinformatics/btt086.

Haas, B. J. *et al.* (2008) 'Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.', *Genome biology*. England, 9(1), p. R7. doi: 10.1186/gb-2008-9-1-r7.

Hayes, W. B. (1966) *Fruit Growing in India*. Kitabistan. Available at: https://books.google.co.za/books?id=ksvPAAAAMAAJ.

Hernandez, D. *et al.* (2008) 'De novo bacterial genome sequencing : Millions of very short reads assembled on a desktop computer De novo bacterial genome sequencing : Millions of very short reads assembled on a desktop computer', pp. 802–809. doi: 10.1101/gr.072033.107.

Holmer, M. (2013) 'Roche to close 454 Life Sciences as it reduces gene sequencing focus', *Fierce Biotech.*

Huang, Y.-F. *et al.* (2012) 'Palindromic sequence impedes sequencing-by-ligation mechanism', *BMC Systems Biology*. BioMed Central, 6(Suppl 2), pp. S10–S10. doi: 10.1186/1752-0509-6-S2-S10.

Illumina Inc. (2012) 'Data Processing of Nextera ® Mate Pair Reads on Illumina Sequencing Platforms'. Available at: http://res.illumina.com/documents/products/datasheets/datasheet_nextera_mate_pair.pdf.

Jain, M. *et al.* (2015) 'Improved data analysis for the MinION nanopore sequencer', *Nat Methods*, 12. doi: 10.1038/nmeth.3290.

Jain, P., Krishnan, N. M. and Panda, B. (2013) 'Augmenting transcriptome assembly by combining *de novo* and genome-guided tools', *PeerJ*, 1, p. e133. doi: 10.7717/peerj.133.

Ju, J. *et al.* (2006) 'Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators.', *Proceedings of the National Academy* *of Sciences of the United States of America*, 103(52), pp. 19635–19640. doi: 10.1073/pnas.0609513103.

Karow, J. (2017) ‘Oxford Nanopore Launches GridIon X5 Nanopore Sequencer, Details Product Improvements’, *GenomeWeb*.

Kelley, D. R., Schatz, M. C. and Salzberg, S. L. (2010) ‘Quake: quality-aware detection and correction of sequencing errors’, *Genome Biology*, 11(11), p. R116. doi: 10.1186/gb-2010-11-11-r116.

Kent, W. J. (2002) ‘BLAT--the BLAST-like alignment tool.’, *Genome research*. United States, 12(4), pp. 656–664.

Koonin, E. V and Galperin, M. Y. (2003) *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics.* Boston.

Korf, I. (2004) ‘Gene finding in novel genomes’, *BMC Bioinformatics*, 5(1), p. 59. doi: 10.1186/1471-2105-5-59.

Korf, I., Yandell, M. and Bedell, J. (2003) ‘Sequence Similarity’, *BLAST: An Essential Guide to the Basic Local Alignment Search Tool*, pp. 55–71. doi: 10.1177/0049124103253373.

Lander, E. S. *et al.* (2001) ‘Initial sequencing and analysis of the human genome.’, *Nature*. England, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Langmead, B. *et al.* (2009) ‘Ultrafast and memory-efficient alignment of short DNA sequences to the human genome’, *Genome Biology*. BioMed Central Ltd, 10(3), p. R25. doi: 10.1186/gb-2009-10-3-r25.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform.’, *Bioinformatics (Oxford, England)*. England, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Liu, L. *et al.* (2012) ‘Comparison of Next-Generation Sequencing Systems’, *Journal of Biomedicine and Biotechnology*. Hindawi Publishing Corporation, 2012, p. 251364. doi: 10.1155/2012/251364.

Luo, R. *et al.* (2012) ‘SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler’, *GigaScience*. BioMed Central, 1, p. 18. doi: 10.1186/2047-217X-1-18.

Magoc, T. and Salzberg, S. L. (2011) ‘FLASH: fast length adjustment of short reads to improve genome assemblies.’, *Bioinformatics (Oxford, England)*. England, 27(21), pp. 2957–2963. doi: 10.1093/bioinformatics/btr507.

Majumder, P. . and Mukherjee, S. . (1972) ‘Aneuploidy in Guava (Psidium guajava L .) I. Mechanism of variation in chromosome number’, *Division of Horticulture, Indian Agricultural Research Institute, New Delhi, India*, pp. 541–548.

Marcais, G. and Kingsford, C. (2012) ‘Jellyfish : A fast k-mer counter’, *Tutorialis e Manuais*, (1), pp. 1–8.

Marcotte, E. M. *et al.* (1999) ‘Detecting protein function and protein-protein interactions from genome sequences.’, *Science (New York, N.Y.)*. United States, 285(5428), pp. 751–753.

Martin, J. *et al.* (2010) ‘Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads.’, *BMC genomics*. England, 11, p. 663. doi: 10.1186/1471-2164-11-663.

Martin, J. a and Wang, Z. (2011) ‘Next-generation transcriptome assembly.’, *Nature reviews. Genetics*. Nature Publishing Group, 12(10), pp. 671–682. doi: 10.1038/nrg3068.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Mehmood, A. *et al.* (2013) ‘iPBS PRIMERS’, 50(4), pp. 591–597.

Meyer, P. *et al.* (2005) ‘Crystal structure and confirmation of the alanine:glyoxylate aminotransferase activity of the YFL030w yeast protein.’, *Biochimie*, 87(12), pp. 1041–1047.

Miko, I. and LeJeune, L. (2009) ‘eds. Essentials of Genetics. Cambridge’, *MA: NPG Education*.

Muir, P. *et al.* (2016) ‘The real cost of sequencing: scaling computation to keep pace with data generation’, *Genome Biology*, 17(1), p. 53. doi: 10.1186/s13059-016-0917-0.

Narum, S. (2015) ‘Restoring and Managing Historical Columbia River Basin Fish Populations with Genomics’, *iCommunity Newsletter*, (April).

Ohri, D. and Kumar, A. (1986) ‘Nuclear DNA amounts in some tropical hardwoods’, *Caryologia*, 39(3–4), pp. 303–307. doi: 10.1080/00087114.1986.10797792.

Parra, G., Bradnam, K. and Korf, I. (2007) ‘CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.’, *Bioinformatics (Oxford, England)*. England, 23(9), pp. 1061–1067. doi: 10.1093/bioinformatics/btm071.

Pathak, R. K. and Ojha, C. M. (1993) ‘Genetic resources of guava.’, in *Advances in horticulture: fruit crops - Volume 1.* New Delhi: Malhotra Publishing House, pp. 143–147. Available at: https://www.cabdirect.org/cabdirect/abstract/19941601691.

Pevzner, P. A., MYu, B. and Mironov, A. A. (1989) ‘Linguistics of nucleotide sequences. II: Stationary words in genetic texts and the zonal structure of DNA.’, *Journal of biomolecular structure & dynamics*. England, 6(5), pp. 1027–1038. doi: 10.1080/07391102.1989.10506529.

Prakash, D. P., Narayanaswamy, P. and Sondur, S. N. (2002) ‘Analysis of molecular diversity in guava using RAPD markers’, *The Journal of Horticultural Science and Biotechnology*. Taylor & Francis, 77(3), pp. 287–293. doi: 10.1080/14620316.2002.11511494.

Quail, M. A. *et al.* (2012) ‘A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers’, *BMC Genomics*. BioMed Central, 13(1), p. 341. doi: 10.1186/1471-2164-13-341.

Rai, M. K. *et al.* (2010) ‘Biotechnological advances in guava (Psidium guajava L.): recent developments and prospects for further research’, *Trees*, 24(1), pp. 1–12. doi: 10.1007/s00468-009-0384-2.

Salzberg, S. L. *et al.* (2012) ‘GAGE: A critical evaluation of genome assemblies and assembly algorithms', *Genome Research*, 22(3), pp. 557–567. doi: 10.1101/gr.131383.111.

Scharf, M. *et al.* (1994) 'GeneQuiz: a workbench for sequence analysis', *Proceedings of the International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2, pp. 348–53. Available at: http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=7584411.

Schmieder, R. and Edwards, R. (2011) 'Quality control and preprocessing of metagenomic datasets.', *Bioinformatics (Oxford, England)*. England, 27(6), pp. 863–864. doi: 10.1093/bioinformatics/btr026.

Schoeman, M. H. (2011) 'The current status of guava wilt disease in South Africa', *SA Vrugte Joernal*, pp. 46–49.

Schroers, H. J. *et al.* (2005) 'Classification of the guava wilt fungus Myxosporium psidii, the palm pathogen Gliocladium vermoesenii and the persimmon wilt fungus Acremonium diospyri in Nalanthamala.', *Mycologia*. England, 97(2), pp. 375–395.

Schuster, S. C. (2007) 'Next-generation sequencing transforms today's biology', *Nature Methods*. United States, 5(1), pp. 16–18. doi: 10.1038/nmeth1156.

Shin, S. *et al.* (2017) 'Validation and optimization of the Ion Torrent S5 XL sequencer and Oncomine workflow for BRCA1 and BRCA2 genetic testing', *Oncotarget*. Impact Journals LLC, 8(21), pp. 34858–34866. doi: 10.18632/oncotarget.16799.

Simão, F. A. *et al.* (2015) 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Simpson, J. T. *et al.* (2009) 'ABySS: a parallel assembler for short read sequence data.', *Genome research*. United States, 19(6), pp. 1117–1123. doi: 10.1101/gr.089532.108.

Simpson, J. T. and Durbin, R. (2012) ‘Efficient de novo assembly of large genomes using compressed data structures.’, *Genome research*. United States, 22(3), pp. 549–556. doi: 10.1101/gr.126953.111.

Simpson, J. T. and Pop, M. (2015) ‘The Theory and Practice of Genome Sequence Assembly’, *Annual Review of Genomics and Human Genetics*, 16(1), pp. 153–172. doi: 10.1146/annurev-genom-090314-050032.

Stanke, M. and Waack, S. (2003) ‘Gene prediction with a hidden Markov model and a new intron submodel.’, *Bioinformatics (Oxford, England)*. England, 19 Suppl 2, pp. ii215-25.

Sutton, G. G. *et al.* (1995) ‘TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects’, *Genome Science and Technology*, 1(1), pp. 9–19. doi: 10.1089/gst.1995.1.9.

Swift, H. (1950) ‘The Constancy of Desoxyribose Nucleic Acid in Plant Nuclei’, *Proceedings of the National Academy of Sciences of the United States of America*, 36(11), pp. 643–654. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1063260/.

Tesler, G. (2002) ‘GRIMM : Genome Rearrangements Web Server’, 18(March), *Bioinformatics(8)*, pp. 492–493.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) ‘TopHat: discovering splice junctions with RNA-Seq’, *Bioinformatics*, 25. doi: 10.1093/bioinformatics/btp120.

Tsui, L. C. and Scherer W., S. (2001) ‘The Human Genome Project’, *Biotechnology Set*, pp. 41–60. doi: 10.1002/9783527620999.ch2e.

Voskoboynik, A. *et al.* (2013) ‘The genome sequence of the colonial chordate, *Botryllus schlosseri*’, *eLife*. Edited by M. E. Bronner. eLife Sciences Publications, Ltd, 2, p. e00569. doi: 10.7554/eLife.00569.

Walter, M. C. *et al.* (2009) ‘PEDANT covers all complete RefSeq genomes’, 37(October 2008), *Nucleic Acids Research (37),* pp. 408–411. doi: 10.1093/nar/gkn749.

Wei, L. *et al.* (2002) ‘Comparative genomics approaches to study organism similarities and differences’, *Journal of Biomedical Informatics* (35), pp. 142–150.

Wheeler, D. A. *et al.* (2008) ‘The complete genome of an individual by massively parallel DNA sequencing’, *Nature*. Nature Publishing Group, 452(7189), pp. 872–876. doi: 10.1038/nature06884.

Wilhelm, A. j (2015) ‘Next Generation DNA Sequencing (II): Techniques, Applications’, *Journal of Next Generation Sequencing & Applications*, 01(S1), pp. 1–10. doi: 10.4172/2469-9853.S1-005.

Willer, M., GM, F. and CJ, S. (2008) ‘Sec61p is required for ERAD-L: genetic dissection of the translocation and ERAD-L functions of Sec61P using novel derivatives of CPY.’, *The Journal of biological chemistry*, 283(49), pp. 33883–33888.

Wu, T. D. and Watanabe, C. K. (2005) ‘GMAP: a genomic mapping and alignment program for mRNA and EST sequences’, *Bioinformatics*, 21(9), pp. 1859–1875. doi: 10.1093/bioinformatics/bti310.

Yandell, M. and Ence, D. (2012) ‘A beginner’s guide to eukaryotic genome annotation’, *Nature Reviews Genetics*. Nature Publishing Group, 13(5), pp. 329–342. doi: 10.1038/nrg3174.

2 Chapter 2: *De novo* genome assembly and annotation of *Psidium guajava*

2.1 Introduction

De novo genome assembly involves joining together the contiguous regions contributing to an organism's chromosomes from fragmented reads of DNA and performs some of the most complex computations in all of biology (Baker, 2012). Next generation sequencing technology has made the cost of generating massive amounts of DNA sequence data significantly cheaper compared to Sanger sequencing. However, despite these advances in technology, modern instruments can read only relatively small segments of the genomes of most organisms, ranging from approximately 100 base pairs (bp) (e.g., Illumina technology) to approximately 300bp and now recently longer reads of approximately 10–20 kb (e.g., Pacific Biosciences technology) (Simpson and Pop, 2015) making it computationally difficult to join it back together during *de novo* genome assembly. Approaches to address the assembly problem are grouped under two main headings, these are greedy approaches (Jeck *et al.*, 2007) and graph-based approaches (Kececioglu and Myers, 1995).

Greedy approaches involve iteratively joining together the reads in decreasing order of the quality of their overlaps, while graph-based approaches represent sequence reads and their inferred relationships to one another as vertices and edges in a graph and attempts to find a walk that best reconstructs the underlying genome while avoiding generating misassemblies (Simpson & Pop, 2015).

In this chapter, we aim to generate an assembly of the guava fruit tree, *Psidium guajava* from Illumina reads as part of creating a genomic resource for the species. This will be done to provide a starting point in genomic studies on guava and its interaction with pathogens at genomic level. The Illumina sequence data was already available at the Agricultural Research Council, Biotechnology Platform (ARC-BTP). Since there is no publicly available genome for guava at the moment, the assembly approach used was *de novo* genome assembly, although a closely related genome of *Eucalyptus* was used to improve on the assembly. The divergence time between eucalyptus and *P. guajava* is ~67mya (Biffin *et al.*, 2010; Thornhill *et al.*, 2015) The genome size of guava is estimated to be ~269.44Mbps (da Costa, Dornelas and Forni-Martins, 2008), with a chromosome number 2n = 22 (diploid state) (da Costa, Dornelas and Forni-Martins, 2008).

One of the main challenges that the guava industry faces is Guava Wilt Disease (GWD), which has caused major losses in the guava industry. Genomics work on guava can apply bioinformatics methods to sequence, assemble and analyze the function and structure of genomes. This can serve as a starting point to utilize genomics resources in guava breeding and cultivation.

Our approach to perform the *de novo* genome assembly was to utilize the three graph-based approaches namely (i) OLC (overlap, layout, consensus) graphs (Kececioglu and Myers, 1995), (ii) *De Bruijn* graphs (Pevzner, MYu and Mironov, 1989) and (iii) String graphs (Myers, 2002) to generate contigs (contiguous sequences) from the Illumina reads. Various open source software packages are available to perform these assemblies. Examples of these packages are: Allpaths-LG (Butler *et al.*, 2008) (de *Bruijn* graph), AbySS (Simpson *et al.*, 2009) (de *Bruijn* graphs), String Graph Assembler (SGA) (Simpson and Durbin, 2012), MaSuRCA (Yorke, J.A *et al,* 2013) (Overlap graph and de *Bruijn* graph) and StriDe (Huang and Liao, 2016). Pre-assembly processes will include quality checking of reads, trimming, error correction and merging of overlapping reads using open source software like FastQC (Andrew, 2016), Trimmomatic (Bolger, Lohse and Usadel, 2014) and FlaSh (Magoc and Salzberg, 2011). Post assembly processes will include ordering of contigs to form scaffolds, gap filling, merging assemblies, and evaluation of assemblies. This is done using various open source software including OPERA-LG (Gao *et al., 2015),* GapFiller (Nadalin *et al.,* 2011), Metassember (Wences *et al.*, 2015), and AlignGraph (Bao, Jiang and Girke, 2014).

The quality of assembly is difficult to assess since the available assembly methods are untraceable. However, certain metrics of an assembly can give an indication of the quality of an assembly. These metrics include the N50 (defined as the shortest sequence length at 50% of the genome), number of contigs, largest contig and the presence of core genes in the assembly. The presence of core genes can be assessed by CEGMA (Parra, Bradnam and Korf, 2007) or BUSCO (Simão *et al.*, 2015), which are both open source software packages. Reapr (Recognising Errors in Assemblies using Paired Reads) (Hunt *et al.,* 2013) can also be used to find errors by re-mapping paired reads to the assembly. A representation of the bioinformatics pipeline used in the genome assembly process is shown in Figure 2.1.

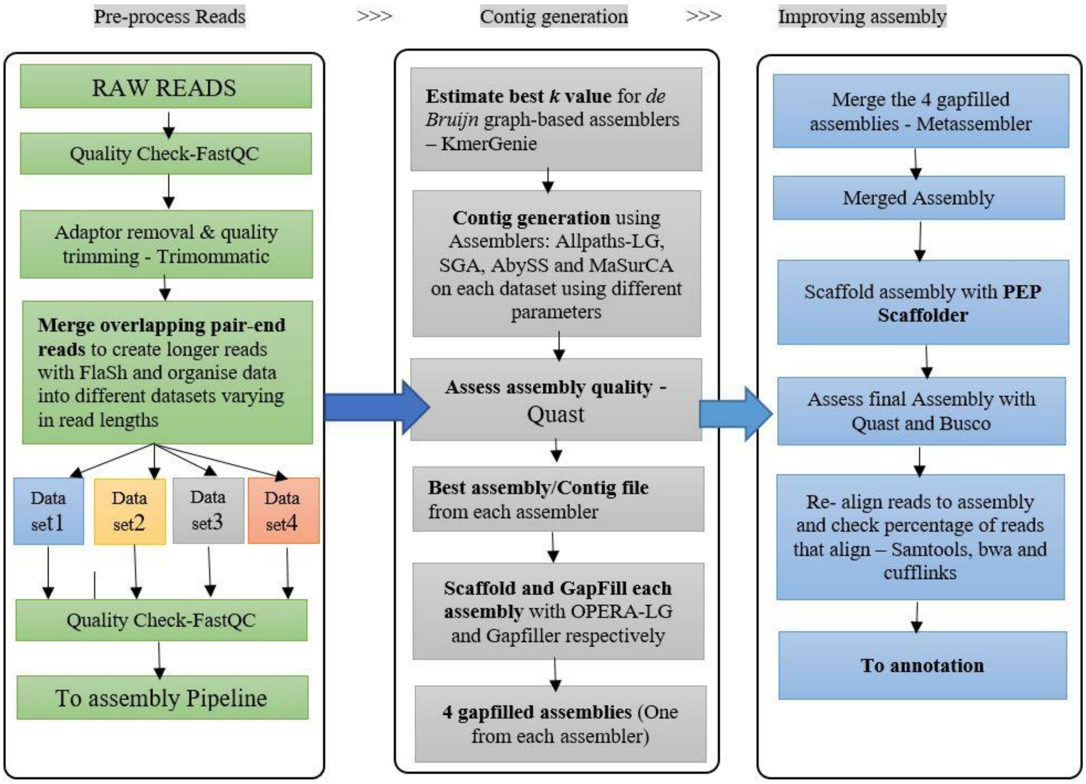

Pre-process Reads
>>>
Contig generation
>>>
Improving assembly
RAW READS
Quality Check-FastQC
Adaptor removal & quality trimming - Trimommatic
Merge overlapping pair-end reads to create longer reads with FlaSh and organise data into different datasets varying in read lengths
Data set1
Data set2
Data set3
Data set4
Quality Check-FastQC
To assembly Pipeline
Estimate best k value for de Bruijn graph-based assemblers – KmerGenie
Contig generation using Assemblers: Allpaths-LG, SGA, AbySS and MaSurCA on each dataset using different parameters
Assess assembly quality - Quast
Best assembly/Contig file from each assembler
Scaffold and GapFill each assembly with OPERA-LG and Gapfiller respectively
4 gapfilled assemblies (One from each assembler)
Merge the 4 gapfilled assemblies - Metassembler
Merged Assembly
Scaffold assembly with PEP Scaffolder
Assess final Assembly with Quast and Busco
Re- align reads to assembly and check percentage of reads that align – Samtools, bwa and cufflinks
To annotation

Figure 2.1: Flow diagram showing the assembly process for the guava genome. The first column shows how reads are preprocessed and organised into four datasets according read length (merged reads and un-merged reads) and number of reads. The second column shows how each dataset is assembled using four assemblers and how the config files are scaffolded and gap-filled. The third column describes how the gap=filled assemblies are then merged with Metassembler and scaffolded using PEP scaffolder. The assembly is then assessed for quality using Quast, BUSCO and samtools.

Genome annotation are descriptions of different features of the genome, and they can be structural or functional in nature (Yandell and Ence, 2012). These features include positions of introns, exons and genes among other features of the genome. To annotate a new genome, various stages are involved including repeat masking, aligning ESTs and proteins to a genome and to perform *ab-initio* gene predictions. To annotate *P. guajava* the Maker-P pipeline which uses tools like RepeatMasker for masking repeat regions, ncbi-blast and rmblast for aligning nucleotides and proteins to the genome, exonerate to polish blast hits and gene prediction softwares such as snap, augustus, GeneMark and FGENESH was used. The Maker-P algorithm finally curates all gene predictions and finds the best match evidence for every gene model.

An annotated genome of *P. guajava* will serve as a good genomic resource for *P. guajava* and this is a good starting point in solving the problems facing the guava industry. This resource can be used to investigate the host-pathogen interaction between *P. guajava* and pathogens such as *N. psidii.* Also, this genomic resource can be used for comparative genomics between *P. guajava* with other members of the Myrtaceae family such as *Eucalyptus and Metrosideros polymorpha.*

The term "Comparative genomics" describes the analysis of the similarities and differences between the genome sequences and resultant features of related biological strains or species (Bachhawat, 2006). A comparison between members of the Myrtaceae family can give us an indication of how closely related the different members of the family are. Besides *Psidium guajava,* two other genomes in the Mytaceae family have been sequenced assembled and annotated: *Metrosideros polymorpha* (Izuno *et al.*, 2016) and *Eucalyptus grandis* (Myburg *et al.*, 2014). In this chapter, we will also perform comparative genomics between *Eucalyptus grandis* and *Psidium guajava* to explore synteny and collinearity of homologous genes between the two genomes.

Synteny can be referred to as the conservation of blocks of order within two sets of chromosomes that are being compared with each other (Myers, 2008). Various open source packages can be used to investigate synteny between genomes in comparative genomics. These include ADHoRe (Proost *et al.*, 2012), Mauve (Darling *et al.*, 2004), Cyntenator (Rödelsperger and Dieterich, 2010), GRIMM-Synteny (Pevzner and Tesler, 2003) DiagHunter (Cannon *et al.*, 2003), MCScanX (Wang *et al.*, 2012) and SyMAP (Soderlund *et al.*, 2006) among others. These packages are used to compare genomes by detecting homologous genes in conserved order or with micro rearrangements allowed.

Comparisons of eukaryotic genomes has revealed various degrees to which homologous genes remain on corresponding chromosomes (synteny) and in conserved orders (collinearity) during evolution (Wang *et al.*, 2012). To compare the guava and eucalyptus genomes, SyMAP was used. SyMAP computes the syntenic blocks between the genomes and has the advantage of interactive java displays that allow graphical displays of alignments and syntenic blocks. Secondly, the distribution of terpene synthase gene family (TPS) in the guava genome was explored and compared to the *Eucalyptus grandis* assembly (v2.0 www.phytozome.net).

Compounds associated with TPS gene family play a major role in many plant process such as attractants (Hume and Esson, 1993), mitigators to heat stress (Sharkey and Yeh, 2001), determinants of leaf litter decomposition rates (Molina *et al.,* 1991) and cues to other toxic constituents (Lawler *et al.*, 1999) among others. The TPS gene family is divided into three classes: Class I consists of TPS-c (copalyl diphospate and ent-kaurene), TPS-e/f (ent-kaurene and other diterpenes as well as some mono- and sesquiterpenes) and TPS-h (*Selaginella* specific); class II consists of TPS-d (gymnosperm specific) and class III of TPS-a (sesquiterpenes), TPS-b (cyclic monoterpenes and hemiterpenes) and TPS-g (acyclic monoterpenes) (Külheim *et al.*, 2015). A comparison of the TPS gene family between *Eucalyptus grandis* and *Psidium guajava* can further give us an indication on how closely related the two plants are.

2.2 Materials and methods

2.2.1 Library construction and pre-processing of data

2.2.1.1 NGS Data available

At the start of the project, two paired end libraries with varying insert sizes were available together with one mate pair library. The paired-end (PE) libraries were prepared using the HiSeq 2000 and HiScan Illumina machines. From the HiSeq 2000, a DNA library (HiSeq_Run14) was constructed using a size selection of approximately 250bp and consisted of 550 292 550 (125bp x 125bp) PE reads. A mate pair library (HiSeq_Run12) was also constructed using a size selection of 5kbp on the HiSeq 2000, consisting of 186 642 714, 125bp x 125bp reads. From the HiScan machine, a paired-end library was created from a size selection of 500bp and consisted of 180 995 474, 100bp x 100bp reads. Table 2.1 below shows a summary of these DNA libraries used during the assembly.

Two more mate pair libraries were sequenced from the HiSeq 2000. The first library was constructed using a size selection of approximately 6 kb and consisted of 43 418 988, 125bp x 125bp paired end reads. The second library was constructed using a size selection of approximately 9 kb and consisted of 93 306 360, 125bp x 125bp reads.

Table 2.2.1: Summary of initial libraries used in the *Psidium guajava* genomics project.

Library name	Number of reads	Type of reads	Read length	Size selected
HiSeq_Run14	550 292 550	Paired-end	125bp	250bp
HiSeq_Run12	186 642 714	Mate-pair	125bp	5kbp
HiScan_Run19	180 995 474	Paired-end	100bp	500bp
6kb_Mate_Pair	43 418 988	Mate Pair	125bp	6000bp
9kb_Mate_Pair	93 306 360	Mate Pair	125bp	9000bp

2.2.1.2 Quality Control of NGS data

Before reads could be used for any downstream processes, the quality of the reads was assessed using FastQC (v0.11.3, Andrew, 2015). Adapter removal and quality trimming was done using Trimmomatic v0.35 (Bolger *et al.,* 2014). Trimmomatic make use of a fasta file containing Nextera adapter sequences to remove Illumina adapters from the reads by identifying seed matches (16 bases) allowing maximally 2 mismatches between the adapters and the reads. These seeds were extended and clipped if a score of 30 was reached. Trimmomatic would also remove leading and trailing low quality or N bases (below quality 3). Also, Trimmomatic scanned the reads with a 4-base wide sliding window, cutting when the average quality per base dropped below 15 and then finally dropping reads shorter than 36bps. Trimmomatic outputs four files, 2 for the 'paired' output where both reads survived the processing, and 2 for corresponding 'unpaired' output where a read survived, but the partner read did not. After trimming, quality of reads was again assessed using FastQC.

Actual insert sizes of all the libraries were estimated by aligning the reads to the results from the first assembly and using Picard tools (http://broadinstitute.github.io/picard/) to estimate the insert sizes. Overlapping paired-end reads (trimmed) were then merged using FLASH (Fast Length Adjustment of SHort reads), (Magoc and Salzberg, 2011) before using these reads in the assembly. FLASH processes each read pair separately and searches for the correct overlap between the paired-end reads. When the correct overlap is found, the two reads are merged, producing an elongated read that matches the length of the original DNA fragment from which the paired-end reads were generated (Magoc & Salzberg, 2011). The extending of reads was meant to improve assembly by providing the assembler with longer reads.

2.2.2 *De novo* genome assembly

De novo genome assembly was done using four different assemblers namely ALLPATHS-LG (Butler *et al.*, 2008), ABySS (Simpson *et al.*, 2009), SGA (Simpson and Durbin, 2012) and MaSuRCA (Yorke, J.A *et al.,* 2013)). To investigate the effects of types of dataset (read length, type of reads and coverage) on assembly quality, different datasets and parameters were used for each of the assemblers and basic metrics of the assemblies were obtained using QUAST (Gurevich *et al.*, 2013).

Four different datasets were prepared from the available Illumina reads. Datasets differed in number of libraries used and whether they were merged with FlaSh or not. This was done to vary number of reads and the length of reads. However, for MaSuRCA and Allpaths-LG, FlaSh was not used on any of the datasets since the assembler requirements suggest that no trimming or merging of reads should be done when using these assemblers. The following datasets were prepared for each assembler.

Table 2.2.2: Datasets used for each of the four assemblers.

Dataset	Type of reads
Dataset 1	HiSeq reads (NOT merged with FlaSh) + Mate pair reads
Dataset 2	HiSeq reads (Merged with FlaSh) + Mate pair reads
Dataset 3	HiSeq reads, HiScan reads (Not merged with FlaSh) + Mate pair reads
Dataset 4	HiSeq reads, HiScan reads (Merged with FlaSh) + Mate pair reads

Before any of the assemblers could be used, KmerGenie (Chikhi and Medvedev, 2014) was used to predict the *k* value that maximizes the assembly size. KmerGenie first computes the *k*-mer abundance histogram for many values of k on a set of reads. It then computes the number of distinct *k*-mers for each value of *k* and returns the *k*-mer value that maximizes this number. For each de *Bruijn* graph-based graph assembler, two *k* values were used, (i) the assembler's default *k*-value and (ii) the *k* value suggested by KmerGenie.

Scaffolds were formed from the contigs formed by each assembler using the assemblers' own scaffolder and OPERA-LG (Gao *et al.*, 2016). In addition to scaffolding the assemblies, a tool called GapFiller (Nadalin, Vezzi and Policriti, 2012) was used to improve each assembly by closing the gap within paired reads. GapFiller differs from FlaSh in that FlaSh closes gaps between overlapping reads while GapFiller fills gaps between reads that do not overlap. Both contig assembly, scaffolding and gap filling was done on the High-Performance Cluster (HPC) at the Agricultural Research Council, Biotechnology Platform (ARC-BTP), University of Pretoria (UP) and the Centre of High-Performance Computing (CHPC) servers in Cape Town. To assess and compare the quality of the assembly QUAST was used. QUAST displays essential metrics used to assess assembly quality in a way that make it easy to compare the metrics.

2.2.2.1 Improving assembly: Merging Assemblies, Gap filling and PEP Scaffolder

The available assembly algorithms vary most significantly in the techniques and heuristics applied to assemble repetitive sequences and resolve errors present, especially in response to the ever-changing landscape of available biotechnologies (Schatz, Delcher and Salzberg, 2010). As a result, the performance of different assemblers varies greatly even with the same dataset. Therefore, to get a consensus of all these different assemblies, contigs and scaffolds produced by different assemblers from *de novo* assembly step were merged using Metassembler (Wences *et al.*, 2015). The best gap filled assemblies from each assembler in the assembly step were merged in order to enhance contiguity and correctness between them. Also, PEP Scaffolder (Zhu *et al.*, 2016) was used to scaffold the contigs produced. PEP scaffolder uses proteins from *E. grandis* and BLAT to scaffold contigs and ensuring completeness of gene regions.

2.2.2.2 Quality assessment of assembly

To assess the quality of assembly, different metrics about the assembly including the N50, the number of contigs and the size of the largest contigs are generally assessed. QUAST was used to calculate and display these metrics in a way that a comparison could be made between any two different assemblies. Also, as a way of assessing the completeness of assembly, BUSCO (Simão *et al.*, 2015) was used to check the presence of core eukaryotic genes in the assembly. This was done using the plant dataset. The transcriptome data was also mapped on to the assembly using BWA (Li and Durbin,

2009) and the SAMtools package (flagstat) (Li *et al.*, 2009) was used to calculate mapping rates.

2.2.3 Annotation

Annotation was performed using Maker-P (Campbell *et al.*, 2014). Maker is a pipeline that uses a variety of other tools to predict genes and curate all annotations. The assembled genome of *P. guajava*, RNAseq data extracted from different parts of the guava plant at different stages of growth, and a protein database from Uniprot were used as input to annotate the guava genome. Tools used in the Maker Pipeline included (i) Perl and many perl modules, (ii) RepeatMasker, (iii) NCBI–BLAST, (iv) SNAP (Korf, 2004), (v) Augustus (Stanke and Waack, 2003) (vi) GeneMark-ES and GeneMark-ET (Lomsadze, Burns and Borodovsky, 2014) (vii) Exonerate (Slater and Birney, 2005) and (viii) BRAKER (Hoff *et al.*, 2016). The steps followed to annotate the guava genome were as follows (i) assemble RNAseq data (ii) train Augustus using BRAKER (ii) run the Maker Pipeline with GeneMark, SNAP and Augustus as gene predictors.

2.2.3.1 Transcriptome assembly of RNAseq data

RNAseq data was extracted from different parts of the plant namely leaves, roots, fruit, flowers and stem at different stages of growth. A total of 336 472 772 RNA-seq reads, trimmed on quality using trimmomatic, were used for transcriptome assembly. The Trinity transcriptome assembler (Grabherr *et al.*, 2011) was used to assemble the transcriptome of RNA-seq reads from each plant part used. These assembled mRNA were then concatenated into one fasta file. The resulting fasta file was deduplicated using CD-HIT (Li and Godzik, 2006) to produce a fasta file with unique transcripts. This assembled fasta file was then used as one of the inputs in the Maker-P pipeline.

2.2.3.2 Training Augustus

To train Augustus, the tool Braker was used. Braker uses Genemark-ET to train and evaluate Augustus and Augustus makes the final gene predictions. Braker was given the final assembly from Metassembler as input together with a bam file of RNA-seq data to the assembled genome. RNA-seq data is aligned to the assembly using Tophat

(Trapnell, Pachter and Salzberg, 2009) to create the bam file. First, GeneMark-ET performs iterative training and generates initial gene structures. Second, AUGUSTUS uses predicted genes for training and then integrates RNA-Seq read information into final gene prediction (Hoff *et al,* 2011). Braker creates a new species gene model in the config directory of Augustus which will be used for gene prediction in Maker.

2.2.3.3 Running Maker-P

Maker v2.32 was used to annotate the assembled genome of the guava fruit tree. To run maker, a few pre-requisite tools are installed. These include: (i) RepeatMasker – To identify and mask repeats in the genome (ii) Exonerate – polished protein and EST alignment to genome using protein2genome and est2genome respectively (iii) NCBI BLAST - to align proteins and ESTs to genome using TBLASTX and BLASTN respectively and (iv) Gene prediction software namely (a) SNAP, (b) GeneMark and (c) Augustus. As input, Maker takes (i) the assembled genome of guava, (ii) an assembled mRNA-seq fasta file, (iii) a protein sequence fasta file from Uniprot and (iv) an Augustus species created from the Braker run, for gene prediction.

A bootstrap method was used to train SNAP. First, the initial Maker run was run without a snap model. The GFF3 models produced are then converted to ZFF format using a maker script make2zff. The ZFF is then converted to a hmm model using snap scripts fathom, forge and hmm-assembler.pl. Maker is then run with snap supplied with the hmm model produced. The same process is done again after the second maker run and the new hmm file produced is fed into snap and maker is run again. The output of Maker, gff files and fasta files, are then analyzed by SOBA (Moore, Fan and Eilbeck, 2010). SOBA (sequence ontology bioinformatics analysis) generates general metrics of the gff file produced by Maker.

2.2.4 Comparative genomics

2.2.4.1 Synteny

Synteny between *Psidium guajava* and *Eucalyptus grandis* was explored with the aid of an open source package called SyMAP. SyMAP uses the assembly fasta files and annotation gff files for both species as input. It uses MUMmer for whole genome alignment and finding genome or amino acid matches. MUMmer uses NUCmer for finding nucleotide matches and PROmer for finding amino acid matches when genomes are different. Interactive graphical output of alignments was produced from SyMAP. The java views produced by symap are; 1) a circle two-genome display, 2) a block-to-chromosome display, 3) an annotation and location search page and 4) a summary page of statistics with a table of blocks. These views can be used to explore synteny blocks from multiple chromosomes that may be displayed in a high-level dot plot or three-dimensional view where details of interesting regions. SyMAP was run using default settings except for the minimum contig size which was set to 500 000. It used 8 threads on a 64bit intel Core i7 machine with 32G ram and a 3.40GHz processor and took under two hours to complete. Assembly and annotation files used for *E. grandis* were obtained from www.phytozome.com and version 2.1 was used. Sixty-three scaffolds of *P. guajava* and nineteen scaffolds of *E. grandis* with scaffold size of more than 500 000, with 83 161 and 309 554 annotations respectively, were loaded into SyMAP and run with default settings.

Terpene synthase gene family (TPS)

Comparative genomics was performed on the TPS gene family in both *E. grandis* and *P. guajava.* The 133 *E. grandis* TPS genes that were found in the paper by Külheim *et al.,* (2015) were used as a starting point in this comparison. The aim was to find the compare availability and distribution of TPS genes between *E. grandis* and *P.* guajava. Besides, the fact that TPS is one of the few well studied gene families in *E.grandis,* we also chose the TPS family to investigate the relationship the effects of the number of TPS genes in plants, such as branch size. A method similar to the method used to discover TPS genes in *Eucalyptus globus* described in the the paper by Külheim *et al.,* (2015) was adapted to find the TPS genes present in the guava genome. First, genomic sequences of the *E. grandis* TPS genes were blasted to the guava genome and all hits with e-values of less than 10^{-10} were considered. Unique genomic regions surrounding the blast hits were then extracted and reverse blasted against the non-redundant blast database. Only regions that returned blast hits to TPS genes with an e value of less than 10^{-10} were considered. Full and partial genes found the TPS gene containing genomic regions were extracted and blasted again to the non-redundant nucleotide blast database and only TPS genes were considered.

2.3 Bibliography

Bachhawat, A. K. (2006) 'Comparative genomics', *Resonance*, 11(8), pp. 22–40. doi: 10.1007/BF02855776.

Baker, M. (2012) 'De novo genome assembly: what every biologist should know', *Nature Methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 9(4), pp. 333–337. doi: 10.1038/nmeth.1935.

Bao, E., Jiang, T. and Girke, T. (2014) 'AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references', *Bioinformatics*. Oxford University Press, 30(12), pp. i319–i328. doi: 10.1093/bioinformatics/btu291.

Biffin, E. *et al.* (2010) 'Evolution of exceptional species richness among lineages of fleshy-fruited Myrtaceae.', *Annals of botany*. England, 106(1), pp. 79–93. doi: 10.1093/aob/mcq088.

Butler, J. *et al.* (2008) 'ALLPATHS: De novo assembly of whole-genome shotgun microreads', *Genome Research*, 18(5), pp. 810–820. doi: 10.1101/gr.7337908.

Campbell, M. S. *et al.* (2014) 'MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations.', *Plant physiology*. United States, 164(2), pp. 513–524. doi: 10.1104/pp.113.230144.

Cannon, S. B. *et al.* (2003) 'DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization', *Genome Biology*. London: BioMed Central, 4(10), pp. R68–R68. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC328457/.

Gao, S. *et al.* (2016) 'OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees', *Genome Biology*, 17(1), p. 102. doi: 10.1186/s13059-016-0951-y.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome.', *Nature biotechnology*. United States, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Hoff, K. J. *et al.* (2016) 'BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.', *Bioinformatics (Oxford, England)*. England, 32(5), pp. 767–769. doi: 10.1093/bioinformatics/btv661.

Huang, Y. and Liao, C. (2016) 'Integration of String and de Bruijn Graphs for Genome Assembly', *Bioinformatics 32(9)*, pp 1301-1307.

Hume, I. D. and Esson, C. (1993) 'Nutrients, antinutrients and leaf selection by captive koalas (Phascolarctos cinereus)', *Aust J Zool*, 41, pp 317-323 doi: 10.1071/ZO9930379.

Izuno, A. *et al.* (2016) 'Genome sequencing of Metrosideros polymorpha (Myrtaceae), a dominant species in various habitats in the Hawaiian Islands with remarkable phenotypic variations', *Journal of Plant Research*. Springer Japan. pp 727-736 doi: 10.1007/s10265-016-0822-3.

Jeck, W. R. *et al.* (2007) 'Extending assembly of short DNA sequences to handle error.', *Bioinformatics (Oxford, England)*. England, 23(21), pp. 2942–2944. doi: 10.1093/bioinformatics/btm451.

Kececioglu, J. D. and Myers, E. W. (1995) 'Combinatorial algorithms for DNA sequence assembly', *Algorithmica*, 13(1), p. 7. doi: 10.1007/BF01188580.

Külheim, C. *et al.* (2015) 'The Eucalyptus terpene synthase gene family', *BMC Genomics*, 16(1), p. 450. doi: 10.1186/s12864-015-1598-x.

Lawler, I. R. *et al.* (1999) 'Ecological example of conditioned flavor aversion in plant-herbivore interactions: effect of terpenes of Eucalyptus leaves on feeding by common ringtail and brushtail possums', *J Chem Ecol*, 25. doi: 10.1023/A:1020863216892.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. Available at: http://dx.doi.org/10.1093/bioinformatics/btp352.

Li, W. and Godzik, A. (2006) ‘Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences’, *Bioinformatics*, 22(13), pp. 1658–1659. Available at: http://dx.doi.org/10.1093/bioinformatics/btl158.

Lomsadze, A., Burns, P. D. and Borodovsky, M. (2014) ‘Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm’, *Nucleic Acids Research*, 42(15), pp. e119–e119. Available at: http://dx.doi.org/10.1093/nar/gku557.

Moore, B., Fan, G. and Eilbeck, K. (2010) ‘SOBA: sequence ontology bioinformatics analysis’, *Nucleic Acids Research*. Oxford University Press, 38(Web Server issue), pp. W161–W164. doi: 10.1093/nar/gkq426.

Myburg, A. *et al.* (2014) ‘The genome of *Eucalyptus grandis*’, *Nature*, 510. pp 356-362. doi: 10.1038/nature13308.

Myers, G. (2002) ‘Whole genome DNA sequencing’, *Computing in Science & Engineering*, 1(3), pp. 33–43. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=764214.

Myers, P. (2008) ‘Synteny: Inferring Ancestral Genomes’, *Nature Education* 1(1):47.

Nadalin, F., Vezzi, F. and Policriti, A. (2012) ‘GapFiller: a de novo assembly approach to fill the gap within paired reads.’, *BMC bioinformatics*. England, 13 Suppl 1, p. S8. doi: 10.1186/1471-2105-13-S14-S8.

Pevzner, P. and Tesler, G. (2003) ‘Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes’, *Genome Research*. Cold Spring Harbor Laboratory Press, 13(1), pp. 37–45. doi: 10.1101/gr.757503.

Proost, S. *et al.* (2012) ‘i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets’, *Nucleic Acids Research*, 40(2), pp. e11–e11. Available at: http://dx.doi.org/10.1093/nar/gkr955.

Rödelsperger, C. and Dieterich, C. (2010) ‘CYNTENATOR: Progressive Gene Order Alignment of 17 Vertebrate Genomes’, *PLOS ONE*. Public Library of Science, 5(1), p. e8861. Available at: https://doi.org/10.1371/journal.pone.0008861.

Schatz, M. C., Delcher, A. L. and Salzberg, S. L. (2010) ‘Assembly of large genomes using second-generation sequencing', *Genome Research*. Cold Spring Harbor Laboratory Press, 20(9), pp. 1165–1173. doi: 10.1101/gr.101360.109.

Sharkey, T. D. and Yeh, S. S. (2001) 'Isoprene emission from plants', *Annu Rev Plant Physiol Plant Mol Biol*, 52. doi: 10.1146/annurev.arplant.52.1.407.

Slater, G. S. C. and Birney, E. (2005) 'Automated generation of heuristics for biological sequence comparison', *BMC bioinformatics*, 6, p. 31. doi: 10.1186/1471-2105-6-31.

Soderlund, C. *et al.* (2006) 'SyMAP: A system for discovering and viewing syntenic regions of FPC maps', *Genome Research*. Cold Spring Harbor Laboratory Press, 16(9), pp. 1159–1168. doi: 10.1101/gr.5396706.

Stanke, M. and Waack, S. (2003) 'Gene prediction with a hidden Markov model and a new intron submodel.', *Bioinformatics (Oxford, England)*. England, 19 Suppl 2, pp. ii215-25.

Thornhill, A. H. *et al.* (2015) 'Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny.', *Molecular phylogenetics and evolution*. United States, 93, pp. 29–43. doi: 10.1016/j.ympev.2015.07.007.

Wang, Y. *et al.* (2012) 'MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity', *Nucleic Acids Research*. Oxford University Press, 40(7), pp. e49–e49. doi: 10.1093/nar/gkr1293.

Wences, A. H. *et al.* (2015) 'Metassembler: Merging and optimizing de novo genome assemblies', *Genome Biology*. Genome Biology, 16(1), p. 207. doi: 10.1186/s13059-015-0764-4.

Yorke, J. A. *et al.* (2013) 'The MaSuRCA genome assembler', *Bioinformatics,* 29(21), pp. 2669–2677. doi: 10.1093/bioinformatics/btt476.

Zhu, B.-H. *et al.* (2016) 'PEP_scaffolder: using (homologous) proteins to scaffold genomes', *Bioinformatics*. Oxford University Press, 32(20), pp. 3193–3195. doi: 10.1093/bioinformatics/btw378.

3 Chapter 3: Results

In this chapter we evaluate the results from our assembly, annotation and comparative genetics. First, we look the quality of the reads used in the assembly as assessed by FastQC. We then explore quality metrics of all assemblies assembled from all assemblers. Metrics such as N50, contig size and number of contigs, all calculated using the tool QUAST, gave an indication of the quality of the assembly. Assemblies with a larger N50 metric, bigger contig sizes and smaller number of contigs are considered better assemblies as most reads will have been stitched together to make contigs. We also proceed check completeness of the assembly and annotation. We analyse the results from BUSCO and Maker to assess the number of genes found in our best assembly. Finally, we look at the results from our comparative genomics studies done between the assembled and annotated genomes of *P. guajava* and *E. grandis.* We explore the results from synteny calculations between the two genomes and also terpene synthase gene distribution in *P. guajava*.

3.1.1 Quality control, trimming and merging

The FastQC reports for raw reads, trimmed reads and merged reads showed that Nextera adapters were present in the raw reads and removed in the trimmed and merged reads. After adapter removal of HiSeq 2000 PE reads, 493 242 216 (89.6%) reads were properly paired and 473 941 948 (86.1%) PE reads after quality trimming. Only 178 548 075 (65%) of the properly paired reads overlapped and were merged with FlaSh. Fig 5.1 – Fig 5.7 (Supplementary files) show basic statistics and per base sequence quality of raw reads, trimmed reads and merged reads extracted from the FastQC report. After adapter removal of HiSeq 2000 mate-pair reads, 38 355 096 (80%) were properly paired and 34 548 616 (72%) reads after quality trimming. The HiScan library was also trimmed and yielded 163 406 356 (90.3%) properly paired reads after adapter removal and 152 515 260(84%) properly paired reads after quality trimming. When FlaSh was used to merge the trimmed reads, 17 466 027 (19.2%) between 36 and 190bp were merged. Table 3.1 below show the basic statistics regarding the number of reads before and after processing.

Table 3.1: Table showing the number of PE reads before and after adapter removal and quality trimming.

Library name	Number of raw reads	Number of properly paired reads with only adapters removed	Number of properly paired reads after quality trimming	Number of reads merged
HiSeq_Run14	550 292 550	493 242 216	473 941 948	178 548 075(37-
HiScanRun12	47 966 562	38 355 096	34 548 616	3 749 681 (36-
HiScan_Run19	180 995 474	163 406 356	152 515 260	17 466 027 (36-

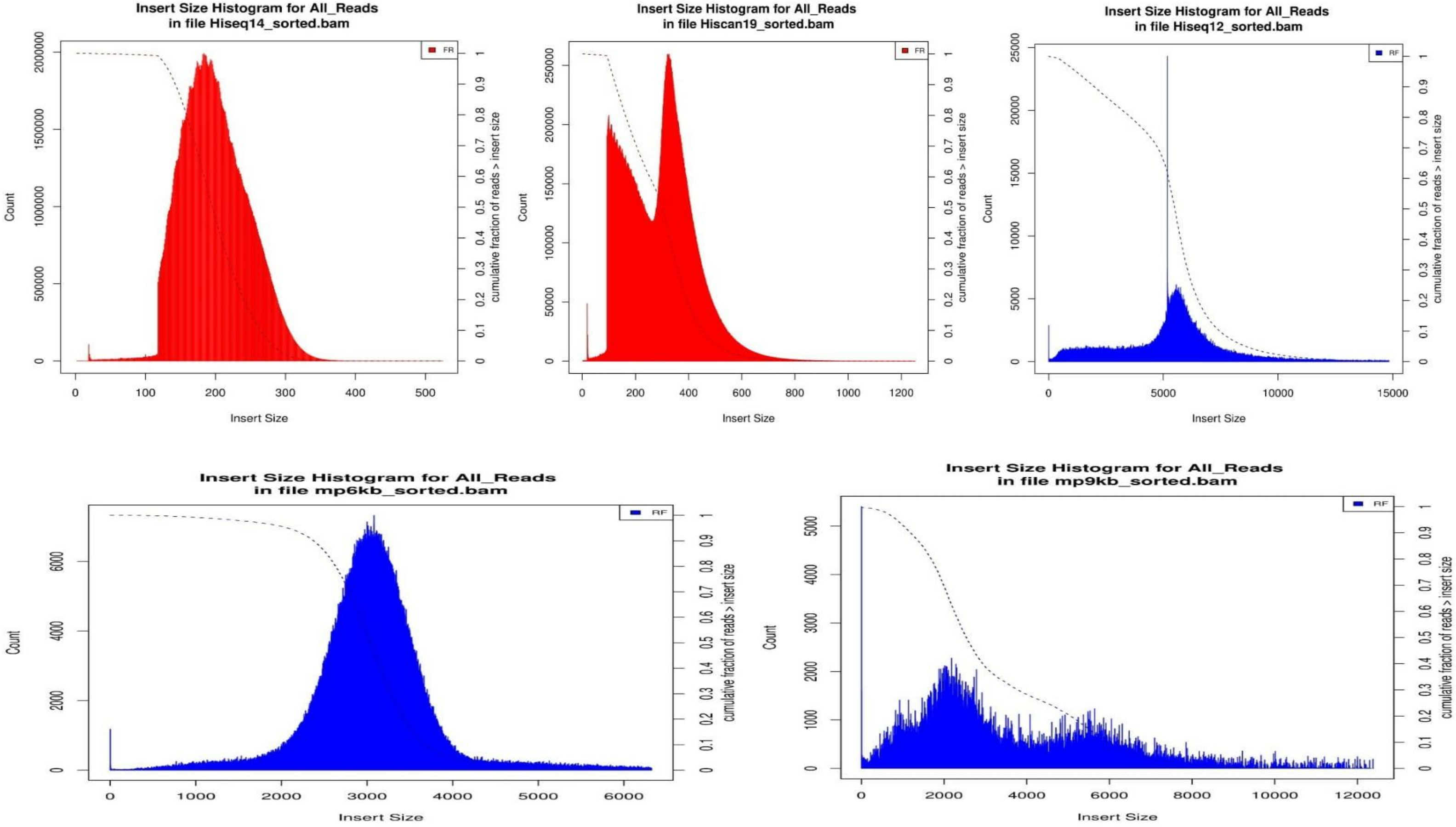
The insert sizes of each library were estimated using Picard tools and insert size distribution were plotted and insert size metrics calculated for each library. Picard tools produces an insert size metric file with metrics such as mean insert size, standard deviation and median from bam files and a histogram showing insert size distribution. Fig 3.1 shows the insert size histograms of four of the main libraries used in this project. As seen in the figure, the insert sizes of some of the mate-pair libraries were smaller than the expected. For instance, the 6kbp library was actually thought to have an insert size of 6kbp but turned out to have an insert size of ~3kbp.

KmerGenie predicted a *k* value of 99 as the best *k* value to maximize genome size. It also predicted a genome size of 334 274 499bp. For each *de Bruijn* graph-based assembler, the predicted *k* value of 99 was used together with the assembler's default *k* value, to test the effects of *k* on assembly quality.

3.1.2 Assembly quality metrics

Each assembler produced a contig file and a scaffold file for each assembly. To ensure a fair comparison of the assembly quality, QUAST was used to calculate the metrics used evaluate assembly quality. Each assembler was run on each of the four datasets at different parameters. Each of the contig files produced was scaffolded with OPERA-LG. A summary of the quality metrics (N50, number of contigs, largest contig/scaffold and others) are shown in Table 3.2. The table shows the statistics of each assembly produced by each assembler for each dataset and using different k-mer sizes and other parameters. Metrics such as number of contigs, number of scaffolds, number of Ns and N50 were used to compare these assemblies. For instance, ABySS was run on each of the four datasets and for each dataset 2 *k*-mer sizes were used, k = 64 and k = 99. The quality metrics from the contigs fasta files produced in each run were evaluated using Quast to calculate the number of contigs in each fasta file (> 500bp), the length of the largest contig and the N50 statistic of each fasta file. These contigs are then scaffolded using OPERA-LG and the quality metrics of each of the fasta files produced was again evaluated using Quast to calculate the number of scaffolds (>500bp), the N50 statistic, the largest scaffold and the number of Ns in each fasta file.

The four best assemblies, shown in Table 3.3, were merged with Metassembler and produced 8 480 scaffolds with an N50 of 106 594bp and a total length of 385 786 388bp with the largest scaffold being 5 288 581bp (Table 3.4). The criteria used to select the best assemblies was as follows: a good assembly should have a lower number of contigs/scaffolds, a bigger N50 and fewer Ns. When the PEP scaffolder was used to scaffold contigs produced by Metassembler, the N50 improved to 111 511bp. These results are shown in Table 3.5. BUSCO was used to assess the Metassembler assembly completeness and showed that 1349 out of the 1440 BUSCO plant genes (93,7%) were found within the genome, an indication that 93.7% of gene regions were covered by the assembly. BUSCO was also run on the Metassembler assembly which was scaffolded with PEP scaffolder and showed that 1351 out of the 1440 BUSCO plants genes (93,9%) were found in that assembly, a slight improvement from the 1349 found in the assembly that was not scaffolded with PEP scaffolder.



Insert Size Histogram for All_Reads in file Hiseq14_sorted.bam

Insert Size Histogram for All_Reads in file Hiscan19_sorted.bam

Insert Size Histogram for All_Reads in file Hiseq12_sorted.bam

Insert Size Histogram for All_Reads in file mp6kb_sorted.bam

Insert Size Histogram for All_Reads in file mp9kb_sorted.bam

Figure 3.1: Histograms showing insert size distributions of reads in four datasets. These show the count of each insert size found in each data set.

Table 3.2: Summary of assembly quality metrics for each assembler used of the different datasets.

		Contig statistics				**Scaffold Statistics (using Opera-LG except for MaSurca**			
	Assembler	**Parameters**	**#Contigs**	**Largest Contig**	**N50**	**#Scaffolds**	**Largest Scaff**	**N50**	**#Ns**
Dataset 1	ABySS	k = 64	130 126	56 368	4 214	66 158	167 303	24 087	72 481 244
		k = 96	194 983	33 593	2 942	78 240	130 278	24 200	81 520 311
	Allpaths-LG	k = 96	49 858	88 728	5 509	36 245	445 331	23 395	30 568 283
		k= 96 (Haplodify)	61 043	70 160	6 098	27 348	234 320	31 777	50 228 535
	MaSurCA	k = auto	143 701	136 607	7 862	87 521	12 552 775	31 905	71 110 658
		k = 99	147 326	98 801	7 350	89 642	6 906 985	29 492	74 879 807
	SGA	m =75 & k = 41	199 424	39 661	2 148	110 826	135 952	19 064	127 323 142
Dataset 2	ABySS	k = 64	132 424	41 763	3 637	55 010	153 761	30 116	82 257 793
		k = 96	154828	45 100	3 632	72 214	140 468	27 636	88 821 034
	SGA	m= 75 & k = 41	298 966	39 709	1 704	49 300	133 500	26 712	81 715 003
Dataset 3	ABySS	k = 64	117 953	66 797	5 130	70 475	193 515	21 546	62 712 790
		k = 96	184 471	33 734	3 343	135 980	104 562	8 540	87 997 404
	Allpaths-LG	k = 96	47 560	102 204	5 933	26 861	323 901	26 743	33 876 984
		k = 96 (Haplodify)	58 650	135 442	7 022	26 661	256 825	33 154	47 742 440
	MaSurCA	k = auto	107 877	148 599	11 549	61 801	3 862 027	53 841	40 946 738
		k = 99	107 659	191 598	11 697	61 182	5 312 630	56 932	41 497 850
	SGA	m = 75 & k = 41	193 197	41 074	2 205	102 545	162 164	20 266	122 239 837
Dataset 4	ABySS	k = 64	116 010	66 909	4 888	55 217	250 301	28 887	67 602 762
		k = 96	143 129	59 711	4 506	74 916	145 459	23 742	75 257 971
	SGA	m= 75 & k = 41	161 609	24 376	1 870	80 013	159 542	18 099	102 055 802

Table 3.3: The best assemblies from each assembler following gap filling.

		Scaffold Statistics before new mate pair data					**Scaffold statistics with Mate Pair data (Gap filled)**				
Datasets	Assembler	Parameters	#Scaffolds	Largest Scaffold	N50	#Ns	#Ns	#Scaffolds	Largest Contig	N50	Total Length
Dataset 4	ABySS	k = 64	55 217	250 301	28 887	67 602 762	26 816 773	40 812	423 026	43 787	371 049 004
Dataset 3	Allpaths - LG	k=96(Haplodify)	26 661	256 825	33 154	47 742 440	34 667 727	18 006	505 250	51827	281 486 143
Dataset 3	MaSurCA	k = 99	61 182	5 312 630	56 932	41 497 850	36 064 996	61 181	5 382 318	58 058	463 304 571
Dataset 3	SGA	m = 75 & k = 41	102 545	162 164	20 266	122 239 837	75 019 057	102 545	162 729	20 839	424 335 802

Table 3.4: Statistics of the Metassembler assembly.

Assemblers Used	**Contigs Statistics**			**Scaffolds statistics**				
	#Contigs	**Largest contig**	**N50**	**#Scaffolds**	**Largest Scaffold**	**N50**	**Total Length**	**#Ns**
Best assembly from each assembler	20 954	447 404	43 276	8 480	5 288 581	106 594	385 786 388	19 984 729

Table 3.5: Statistics of assembly scaffolded by PEP Scaffolder**.**

Scaffold Statistics				
#Scaffolds	**Largest Scaffold**	**N50**	**Total Length**	**#Ns**
8357	5 288 581	111 511	385 798 688	19 996 850

3.1.3 Annotation

A total of 336 472 772 RNA-seq reads from different parts of the guava fruit tree, collected at different times were assembled with Trinity and the resulting fasta file de-duplicated by CD-HIT. In total, 369 742 contigs were produced, with the smallest being 201bp and the largest being 16 599bp, with an N50 for the transcript of 1 959bp. This assembled mRNA set was used as input in the Maker annotation pipeline. Maker was run using Augustus version 3.2.2 and SNAP as *ab-initio* gene predictors and produced a gff file with gene positional information and a fasta files containing transcripts and proteins as output. SOBA was used to count the number of genes found in the assembly annotation together with other metrics. There were 24 134 genes from the 8 435 scaffolds found in the assembly. Table 3.6 below shows basic statistics calculated by SOBA from the Maker gff output. Maker also generates an Annotation Edit Distance (AED) score associated with every gene to measure the gene quality. The AED score ranges from zero to one, with zero being the best quality score and one being the worst. AED provides a means to distinguish between a new gene release with no changes, and one wherein the intron-exon coordinates alone have been altered and also provides a means to quantify the extent of these changes (Eilbeck *et al.*, 2009). A cumulative frequency plot of the AED scores (Figure 3.2) shows that more than 80% of the genes are of high quality with AED scores < 0.4. GenomeTools (Gremme *et al*., 2013) was used to calculate the lengths distribution of genes and exons in the genome annotations and summary statistics calculated using R (Table 3.7).

Table 3.6: SOBA output showing counts of different genome features

Feature	Count
CDS	143 629
Contig	8 435
Exon	152 657
Expressed sequence match	2 663 140
Five prime UTR	18 083
Gene	24 134
mRNA	24 134
Match	309 185
Match part	7 278 678
Protein match	662 941
Three prime match	19 183

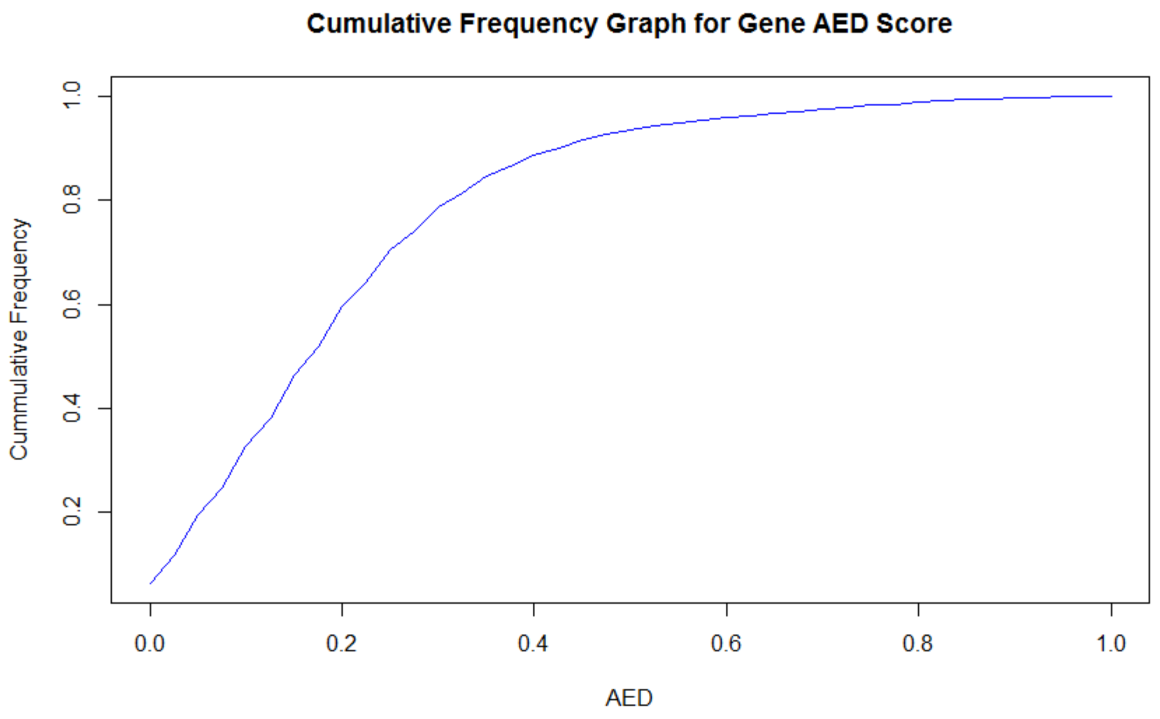


Figure 3.2: Cumulative frequency graph of AED scores of genes in the newly annotated genome**.**

Table 3.7: Summary statistics of gene and exon lengths distribution in *P. guajava* in bp.

Feature	Min length (bp)	Mean length (bp)	Median length (bp)	Max length (bp)
Gene	180	6726	5428	56500
Exon	2	293.9	158.0	7905

3.1.4 Comparative genomics

3.1.4.1.1 Synteny

SyMAP ran for under two hours and produced plots that showed syntenic blocks between *E. grandis* and *P. guajava*. Figure 3.3 below show a dot plot produced by SyMAP showing syntenic blocks scaffold by scaffold. The numbers in each of the axes represent the scaffold/chromosome number for each of the genomes, *E. grandis* scaffolds/chromosomes in the x-axis and *P. guava's* in the y-axis. The synteny blocks between a pair of scaffolds is represented by blue boxes and the anchors by dots. A synteny block is a collection of contiguous genes located on the same chromosome (Sinha & Meller, 2007). The number of synteny blocks in any scaffold intersection show strongly syntenic the two scaffolds are. The syntenic blocks between the two genomes give us an indication of how closely related the two plants are, at the same time exploring which contigs/scaffolds from guava are found on the same chromosome. For instance, *E. grandis* chromosome 1 has relatively stronger synteny blocks with *P. guajava* scaffold 3, 4, 12, 15 and 16 than any other scaffold as shown by the number and size of the syntenic blocks in Figure 3.3. This suggests that the scaffolds 3, 4, 12, 15 and 16 are most likely to be from the same chromosome. Figure 3.3 – 3.5 give pictorial views of how best each of the contigs from guava match the eucalyptus chromosomes. Figure 3.4 shows a pictorial block view of how the *P. guajava* blocks anchor on to the 11 *E. grandis* chromosomes. Figure 3.5 is a circular view of syntenic blocks between the two genomes. Figure 3.6 is a 3D representation of guava contigs anchoring around chromosome 1 of eucalyptus. Here we zoomed into only the synteny on chromosome 1 of *E. grandis* to look in more detail on the synteny with *P. guajava.* The same can be done to other chromosomes to explore them in more detail. Table 3.8 - 3.10 show summary statistics for synteny between the two genomes. Table 3.11 contains descriptions and explanations for each of the terms / metrics used to describe the synteny in Table 3.8 – 3.10. Table 3.8 are basic statistics of the scaffolds used for synteny studies. Only 16 scaffolds and 63 scaffolds from eucalyptus and guava, respectively, were used for synteny. Table 3.9 shows that only 5% of the eucalyptus genome was covered by alignment regions while only 28% of the guava genome was covered. Table 3.10 shows that 64% of the *P. guajava* genome and 39% of the *E. grandis* genome are covered by synteny blocks.

3.1.4.1.2 Terpene synthase gene family (TPS)

1 543 unique hits where found when 113 *E. grandis* TPS genes were blasted on to the guava genome with e-value < 10^{-10}. All the annotated genes in genomic regions +/- 2 500bp around 1 543 hits on guava, were collected and blasted to the non-redundant nucleotide blast database, 36 of the genes found in those regions were TPS genes. This implies that, of the 113 TPS genes found in *E.grandis,* only 36 were found in guava.

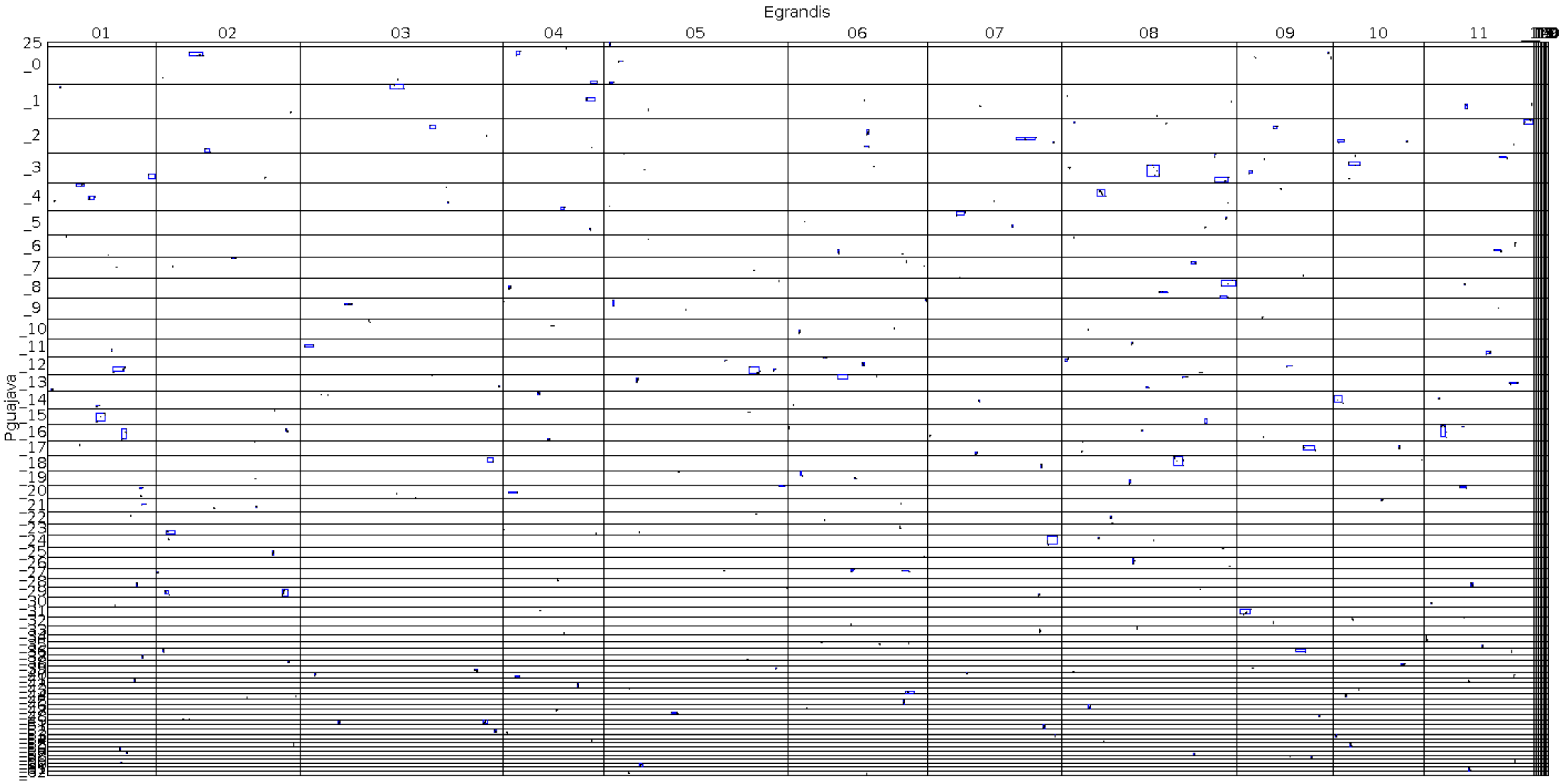


Figure 3.3: Dot plot showing synteny between *E. grandis* and *P. guajava*. The blue boxes show synteny between the 2 chromosomes. More or bigger the blocks indicate stronger synteny between the 2 chromosomes.

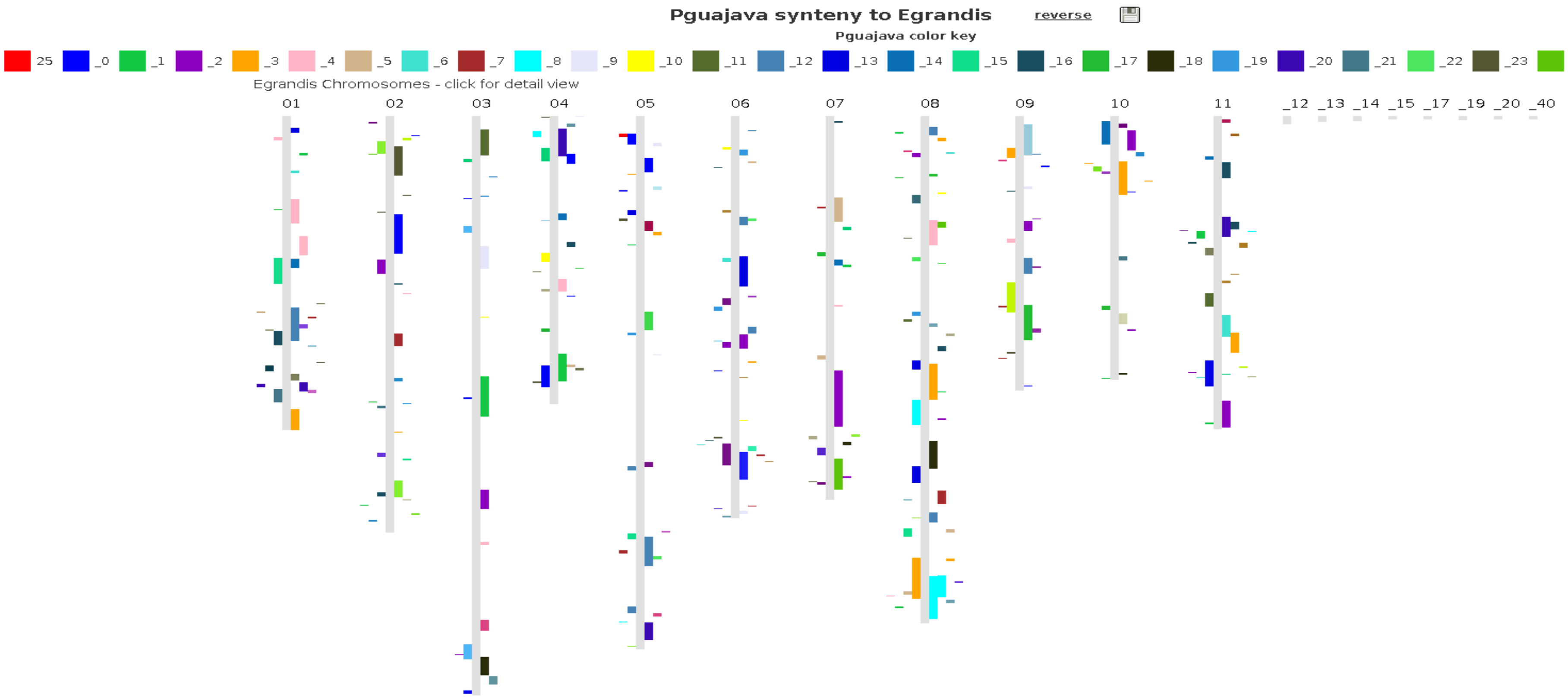


Figure 3.4: Symap Block view. This shows how the different blocks of *P. guajava* genome anchor to *E. grandis* chromosomes.

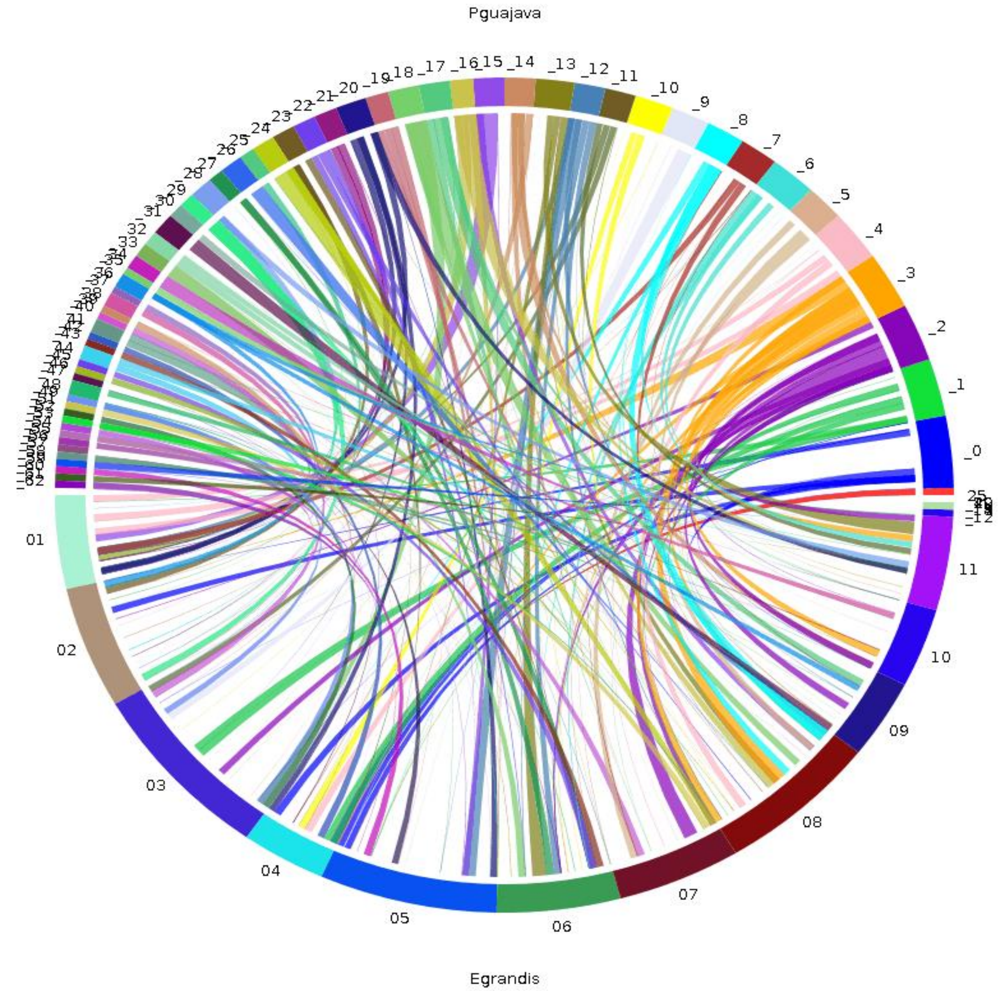


Figure 3.5: The Symap circle view, showing every syntenic block between *E. grandis* and *P. guajava.*

Summary statistics for synteny between P. guajava and E. grandis. (Symap)

Table 3.8: Genome and Annotation Statistics.

Species	#Seqs	Total Kb	#genes	Max Kb	Min Kb	<100kb	100kb-1MB	1Mb-10Mb	> 10Mb
E. grandis	19	618334	36376	83952	503	0	7	34	11
P. guaiava	63	101116	6328	5288	511	0	29	0	0

Table 3.9: Anchor Statistics.

Species	#Anchors	%InBlocks	%Annotated	%Coverage	<100bp	100bp - 1kb	1kb-1Okb	>10kb
E. grandis	15725	42%	64%	5%	188	6526	8761	250
P. guajava	15725	42%	34%	23%	204	6714	8492	315

Table 3.10: Block Statistics.

Species	#Blocks	%Coverage	%DoubleCov	Inverted	%GenesHit	<100kb	100kb-1 Mb	1Mb-10Mb	>10Mb
E. grandis	290	39%	6%	152	25,0	40	171	79	0
P. guajava	290	64%	15%	152	60,0	75	209	6	0

Table 3.11: Key: Descriptions of key terms used in Table 3.8 – 3.10.

Genome and Annotation Statistics

# Seqs	Total number of sequences loaded for the project
Total Kb	Total kilobases of the loaded sequences
#genes	Number of annotated genes
Max Kb, Min Kb	Size of largest and smallest sequences
Size range columns	Number of sequences in these size ranges

Anchor Statistics

#Anchors	Total number of anchors loaded between the two projects (same for each project)
%1nBlocks	Percentage of anchors in synteny blocks (same for each project)
%Annotated	Percentage of anchors intersecting a gene annotation on the project
%Coverage	Percent of total project sequence length covered by anchor alignment regions
Size range columns	Number of anchors having alignment lengths in the given ranges (can be slightly different between projects)

Block Statistics

#Blocks	Total number of synteny blocks found between the two projects (same for each project)
%Coverage	Percent of total project sequence length covered by synteny blocks
%DoubleCov	Percent of total project sequence length covered by two or more synteny blocks (i.e.,mapping to a putative duplication in the other genome)
Inverted	Number of blocks which are inverted (note, blocks may contain local regions of opposite orientation)
%GenesHit	Percentage of genes located in synteny block regions which have syntenic hits ("gene retention")
Size range columns	Number of blocks whose total coverage region on the project is in the given range

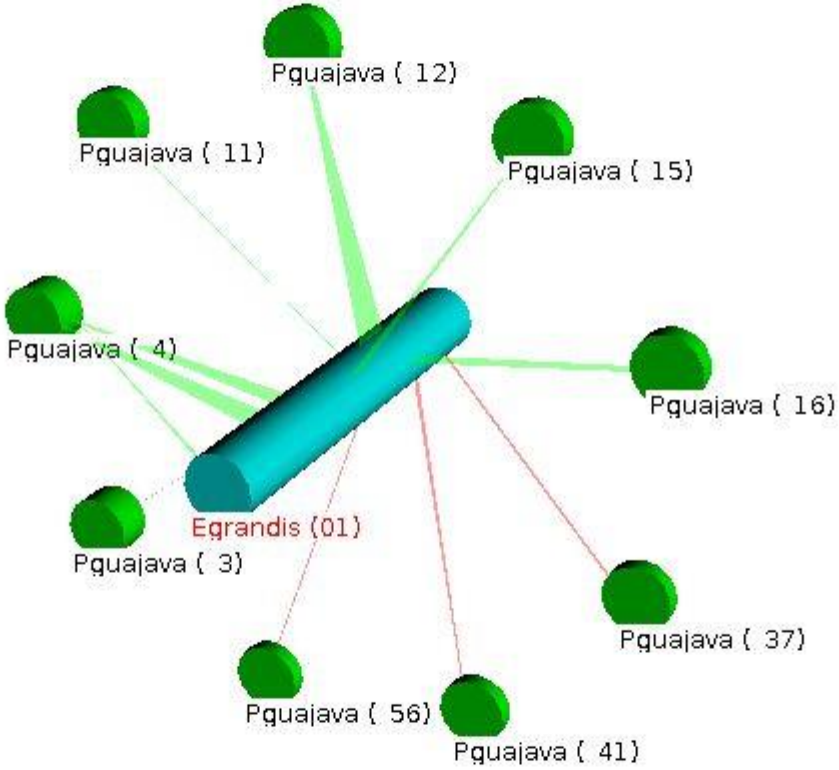


Figure 3.6: A 3D view from SyMAP showing synteny between *E. grandis* chromosome 1 and *P. guajava* scaffolds. The central bar is *E. grandis* chromosome 1 surrounded by selected *P. guajava* scaffolds with strong synteny.

3.2 Discussion

Psidium guajava (Guava) is one of the most important crops grown commercially in South Africa and Southern Africa. Production of the fruit amounted to about 33 574 tons of guava fruit in South Africa in 2013 with an estimated gross value of R53 439 000. Guava has been shown to have very high nutritional value with exceptionally high vitamin C content (Prakash *et al.,* 2002; Rai *et al.,* 2010) and possess various pharmaceutical properties such that it has been used for treatment of ailments such as wounds ulcers, bowls and cholera (Begum *et al.*, 2002). The outbreak of Guava Wilt Disease caused by a resistant strain of the fungus *Nalanthamala psidii,* has caused major loses in the guava industry. A genomic resource for guava will be a starting point in addressing this problem and others using genomics and bioinformatics. Having an annotated genome of the guava tree can be very useful resource in development of markers for Marker Assisted Selection (MAS) that could lead to breeding of a Guava Wilt Disease resistant plant.

Creating such a genomic resource would involve assembling and annotating the genome of *Psidium guajava.* This study has produced the first annotated genome of *Psidium guajava.*

3.2.1 *De-novo* genome assembly

Many factors affect the quality of assembly. These include (i) the dataset used e.g. size of reads or type of data (paired end (PE), single end (SE) or mate pair reads (MP)), (ii) assembly method used by assembler and (iii) parameters used by the assemblers e.g. *k*-mer size. Our *de-novo* genome assembly approach was to use multiple assemblers on different datasets using various assembly parameters and then merging the best assemblies with Metassembler. The length, amount and quality of reads can greatly affect the quality of assembly. With this in mind, four datasets were created from the two pair-end libraries available. These datasets differed in read length and number of reads. To vary the read lengths between datasets, the tool FlaSh was used to merge overlapping reads in some of the datasets to increase the read length. To vary number of reads in datasets, some datasets had reads from the Illumina HiSeq 2500 machine only and other datasets included the ones from the Illumina HiScan machine as well as the Illumina HiSeq 2500 machine. The assemblers were selected based on the graph-based approach used. All assemblers used some variation of graph-based approaches to generate contigs. The four assemblers used included Allpaths-LG, ABySS, SGA and MaSurCA. Allpaths-LG and ABySS use the *de-bruijn* graph method for contig generation, SGA uses the string graph method and MaSurCA used both *De Bruijn* graph method and Overlap Layout Consensus method. Since these assemblers use different contig generation and repeat resolving methods, the contigs and scaffolds produced differed from assembler to assembler. Metassember was therefore used to construct a consensus assembly by merging the best assembly from each of the four assemblers. As expected, the merged assembly was significantly better that all the individual assemblies when considering quality metrics such as N50, largest conting and number of contigs. The best assembly among the individual assemblies was the MaSuCA assembly with 61 181 scaffolds with a N50 of 58 058bp and a largest scaffold of 5 312 630bp while the merged assembly had 8 480 scaffolds with a N50 of 106 594bp and a largest contig of 5 288 581bp (Table 3.2). PEP Scaffolder then also improved the Metassembler assembly as the N50 statistic improved to 111 511bp and the number of scaffolds dropped to 8 357. A smaller number of scaffolds show that contig generation was more effective in stitching together the reads and the scaffolding more effective in resolving repeats and laying out the contigs into scaffolds. A larger N50 reveals that there were generally longer scaffolds in the merged assembly and the PEP scaffolder assembly. PEP scaffolder greatly improved the completeness of gene regions (Zhu *et al.,* 2016) by using homologous proteins from a closely related species. In this case, well curated proteins from *E. grandis* were used and this greatly improved the assembly.

3.2.2 Annotation

The final assembly was annotated using the MAKER pipeline. The MAKER pipeline uses both *ab- initio* and EST and protein-based evidence. *Ab- initio* gene predictors utilizes genomic data only and MAKER combines the genes/exons predicted from these predictors with EST evidence to produce final gene models. The *ab-initio* gene predictors were SNAP and Augustus. Both predictors were used to make gene models prediction more reliable since gene models with support from different sources are more reliable. 24 134 genes were predicted with a mean length of 6 726bp. This is slightly less that the 36 376 genes found in *Eucalyptus grandis.* The MAKER gff output contains AED scores to show the quality of each gene. A plot of the AED scores shows a significantly high number of genes with high quality genes (> 0.4). This indicates that over 80% of the annotated genes were of high quality. PEP scaffolder played an important role in resolving gene regions in the Metassembler assembly as it uses a protein database to resolve gene regions. This is shown by the number of BUSCO plant genes found in the scaffolded genome which had 1351 genes out of the 1440 BUSCO plant genes which is

3 genes more than the original metassembler assembly. Since 1351 out of the 1440 BUSCO plants genes (93,9%) were found in that assembly, this gives an indication that our assembly was about 93.9% complete.

3.2.3 Comparative Genomics

Comparative genomics was performed to give us an indication of the genetic difference between *E. grandis* and *P. guajava.* According to Biffin *et al.*, 2010 and Thornhill *et al.*, 2015 the divergence time between *E. grandis* and *P. guajava* is ~67mya. We first explored synteny between the two genomes to see which blocks are conserved between the two genomes. Syntenic blocks were found between each of *E. grandis* chromosomes and at least one of the *P. guajava* scaffolds. We used 16 *E.grandis* scaffolds with 42% of its sequence lengths covered with genes, and 64 *P.guajava* scaffolds with only 28% covered with genes. Symap searched for syntenic blocks between *E. grandis* and *P. guajava* and only 39% of *E. grandis* sequence length was covered in syntenic blocks while 64% of *P.grandis* sequence length was covered in syntenic blocks. 42% of these blocks were covered in anchors, meaning 23% of total *P. guajava* sequence length and 5% of *E. grandis* sequence length were covered by anchor alignment regions. This gives an indication of how genetically similar *P. guajava* is to *E. grandis,* at least 23% similar.

To further explore comparative genomics between *P. guajava* and *E. grandis* the terpene synthase (TPS) gene family was chosen to compare the presence and distribution of TPS genes in these two genomes. The 113 Eucalypus TPS genes that were identified in Külheim *et al.,* 2015 were used. We found 36 out of the 113 *E. grandis* TPS genes in *P. guajava*. The lower number of TPS genes was expected since the number of TPS genes in a plant has been shown to be inversely proportional to the branch length (Külheim *et al.,* 2015).

The comparative genomics work done between *P. guajava* and *E. grandis* is only a starting point in investigating the genetic diversity between members of Mytaceae family. However, this work is a good indication of the genetic diversity between the two genomes.

4 Chapter 4: Concluding Remarks

The guava fruit tree is one of the most important commercial fruit trees in South Africa and Sub Saharan Africa (Hayes, 1966; Pathak & Ojha, 1993). It has the absence of a genetic resource for guava had made it difficult to tackle problems the guava industry faces using genomics. This project was aimed at creating a genomic resource that will be used as a starting point in tackling problems such as the guava wilt disease (Grech, 1985).

The genomic resource was made by first assembling the genome of the guava fruit tree using next generation sequence data and bioinformatic tools to create the first draft genome assembly of guava. Next, the assembly was annotated with the aid of RNA-seq data and bioinformatic tools. Finally, comparative genomics work was done between *P. guajava* and *E. grandis* which another member of the Myrtaceae family which guava belongs to.

The pipeline used for de *novo* genome assembly harnessed the power of all three graph-based genome assembly techniques, namely *De-bruijn* graph method, string graph method and Overlap graph assembly (Pevzner *et al.*, 1989; Myers, 2002; Hernandez *et al.*, 2008; Simpson & Durbin, 2012). The assembler Metassembler (Wences *et al.*, 2015) was used to create a consensus genome assembly of different assemblies made from different datasets and different assembly methods. PEP scaffolder (Zhu *et al.*, 2016) was used to improve the assembly by making scaffolds using BLAT (Kent, 2002) and protein sequences, ensuring that gene regions were properly assembled. Assembly completion was assessed by using BUSCO (Simão *et al.*, 2015) which checks the percentage of BUSCO genes present in the assembled genome. Also, metrics such as N50 and length of longest scaffolds were used to check the quality of assembly. The final assembly had 8 357 scaffolds, an N50 statistic of 111 511bp and the longest scaffold was 5 288 581bp. BUSCO predicted that 93,9% of core plant genes were found in the final assembly. This implies that the assembly is at least 93.9% complete.

Maker-P (Campbell *et al.*, 2002) was used to annotate the genome assembly with the assistance of RNA-Seq data extracted from plant tissue obtained on different parts of the guava tree and at different stages of development. Obtaining genetic material for RNA-Seq this way ensured that we maximize RNA extraction since different genes are expressed from different parts of the plant at different stages of development. Different gene predicters were used such as SNAP (Korf, 2004), Augustus (Stanke & Waack, 2003) and GeneMark-ET (Lomsadze *et al.*, 2014).

These were all trained in a bootstrap fashion till a final gene model for guava was created. A total of 24 134 genes were found in the guava assembly. Over 80% of these had an AED score of less than 0.4. These annotation quality metrics show our pipeline captured high quality full length genes.

The comparison genomics work done on guava and eucalyptus was aimed to explore the genetic diversity of members of the Mytaceae family. Using SyMap (Soderlund *et al.*, 2006) to calculate synteny between the genome of *Psidium guajava* and *Eucalyptus grandis* (Myburg *et al.*, 2014a), we observed that guava is at least 23% similar to eucalyptus since SyMap showed that only 23% of the *P. guajava* genome was covered by anchor alignment regions. We discovered that only 36 of the 113 Eucalyptus terpene synthase genes (TPS) (Külheim *et al.*, 2015) were found in the new guava assembly. This small number may be attributed to many factors including incomplete assembly and annotation.

The pipeline used in this project can be used to further create more genomic resources for other plants in the Mytaceae family such as *Metrosideros* (Izuno *et al.*, 2016). Having these genomic resources will not only work as a starting point in solving problems faced by plants in the Mytaceae family but will also help us to explore genetic diversity between members of this family and others. One such use of such a resource is for marker assisted selection (MAS) during breeding. An annotated genome will be pivotal in marker development. The first annotated genome of guava will act as a reference genome for guava where other sequence data from different guava lines can be aligned to and this will form the basis of polymorphism detection. Detecting polymorphism such as SNPs and INDELs is pivotal in marker development and creating trait associated markers is the essence of MAS. The assembly of *P. guajava* is therefore a starting point to solving problems in the guava industry such as breeding for the guava wild disease using MAS and also understanding how closely related guava is to other plants in the Myrtaceae family.

4.1 Bibliography

Eilbeck, K. *et al.* (2009) ‘Quantitative measures for the management and comparison of annotated genomes’, *BMC Bioinformatics*. BioMed Central, 10, p. 67. doi: 10.1186/1471-2105-10-67.

Sinha, A. U. and Meller, J. (2007) ‘Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms’, *BMC Bioinformatics*. London: BioMed Central, 8, p. 82. doi: 10.1186/1471-2105-8-82.

5 Supplementary files

5.1 FastQC files

HiSeq2000 raw reads (with adaptors)

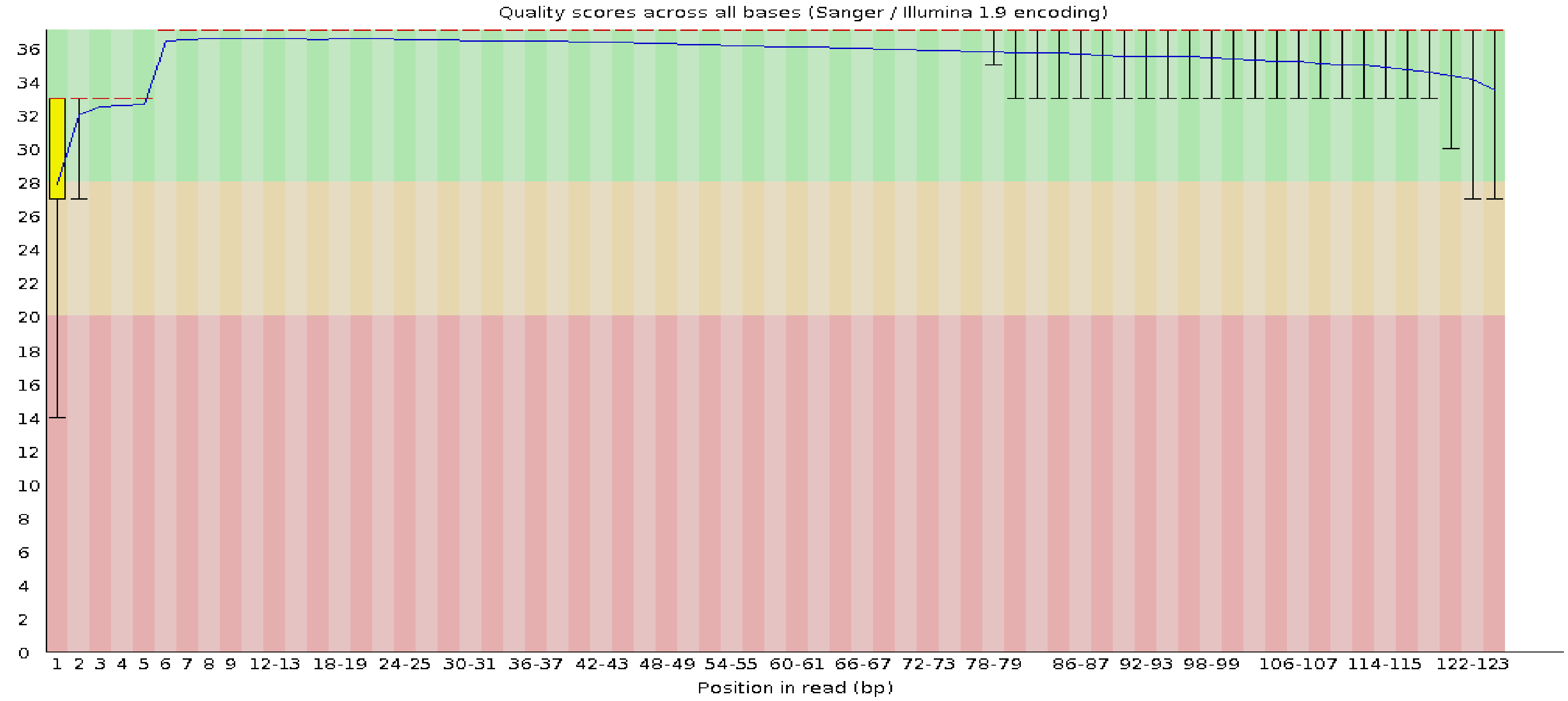


Figure 5.1: Summary and basic statistics of the HiSeq 2000 forward reads (raw reads) quality statistics extracted from the FastQC report.

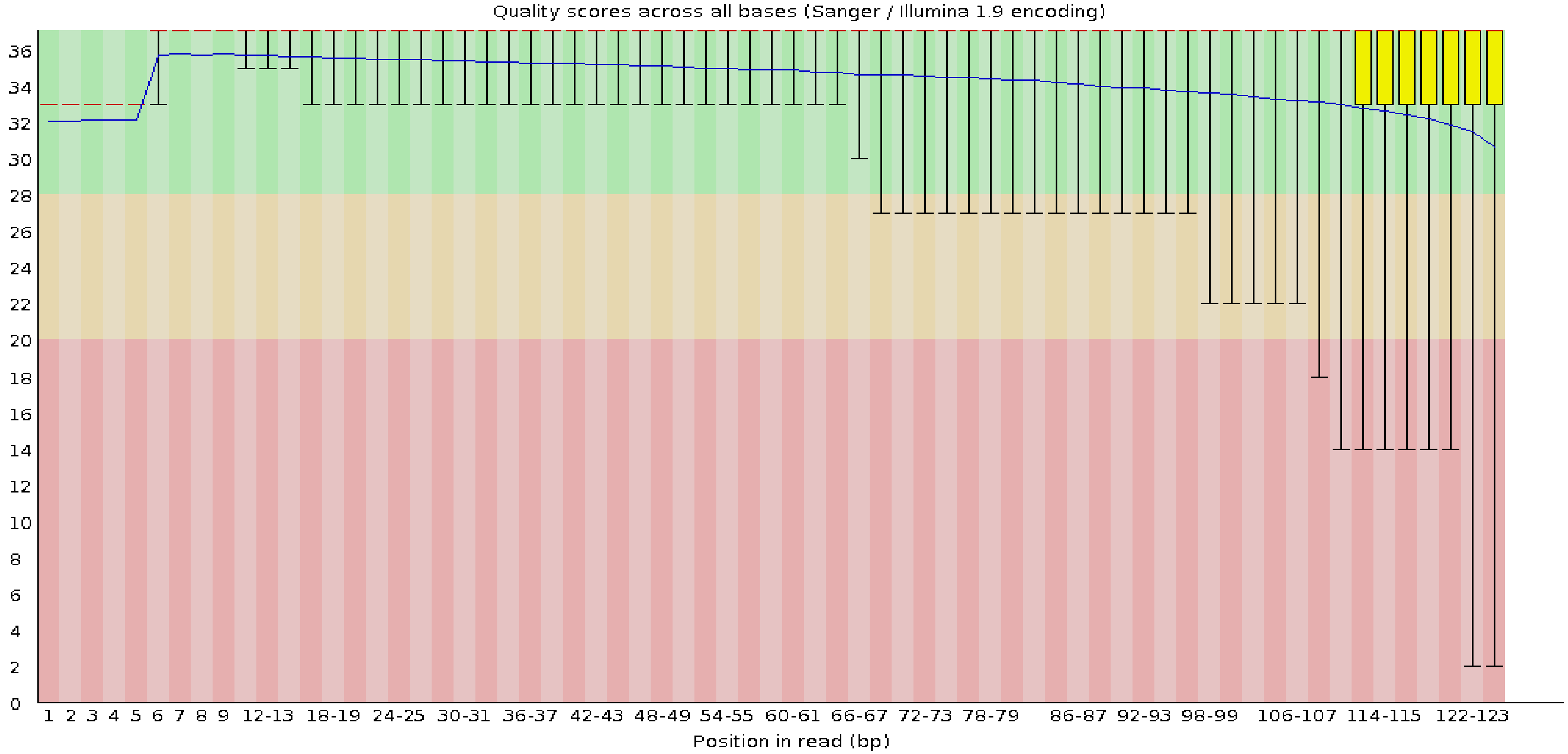


Figure 5.2: Summary and basic statistics of the HiSeq 2000 reverse reads (raw reads) quality statistics extracted from the FastQC report.

HiSeq 2000 reads trimmed

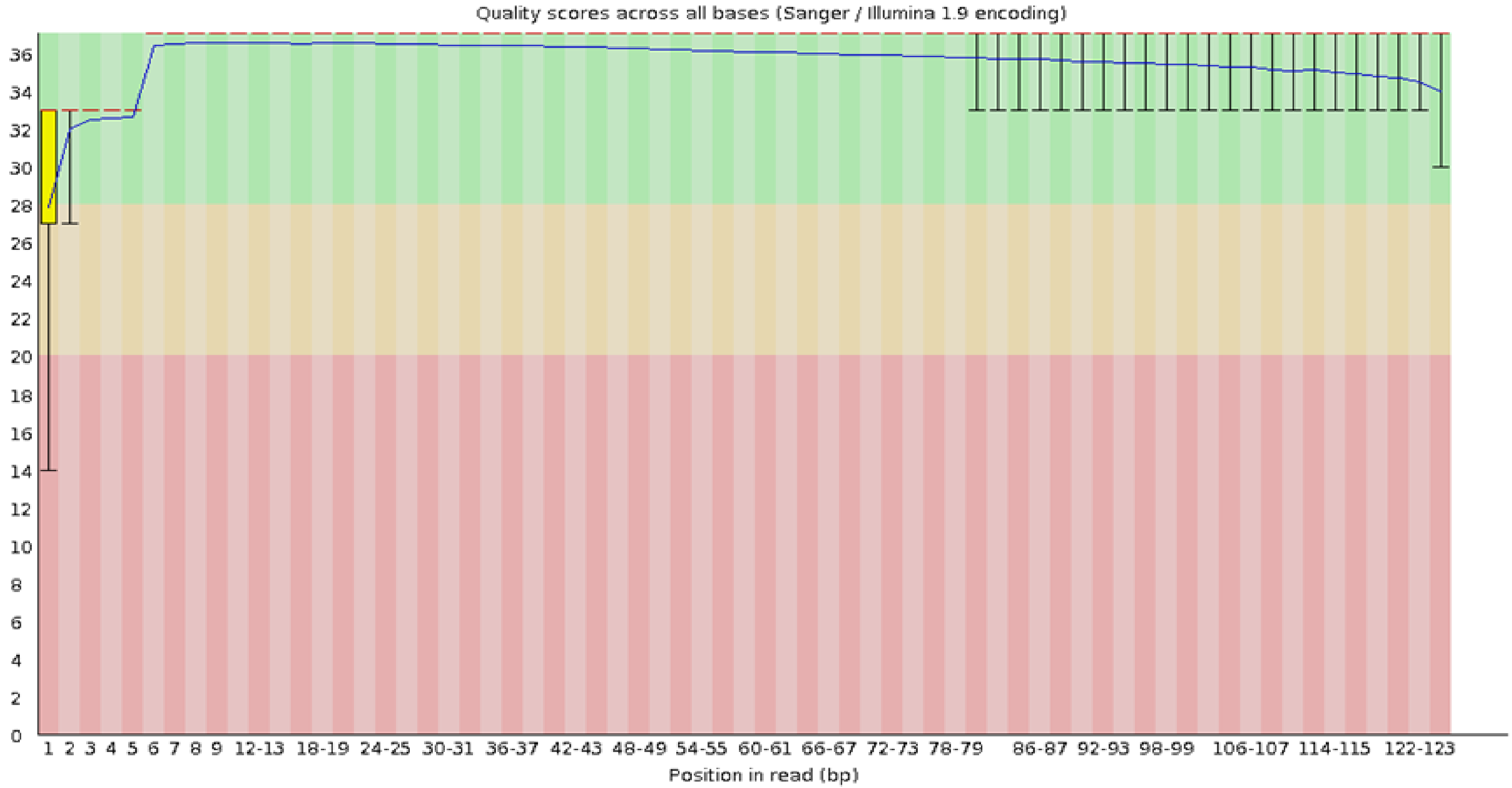


Figure 5.3: Per base sequence quality statistics for HiSeq 2000 forward, properly paired reads with adapters removed. The PHRED scores (y axis) represent quality of base calls. The bigger the PHRED score, the better the base quality of bases at that position.

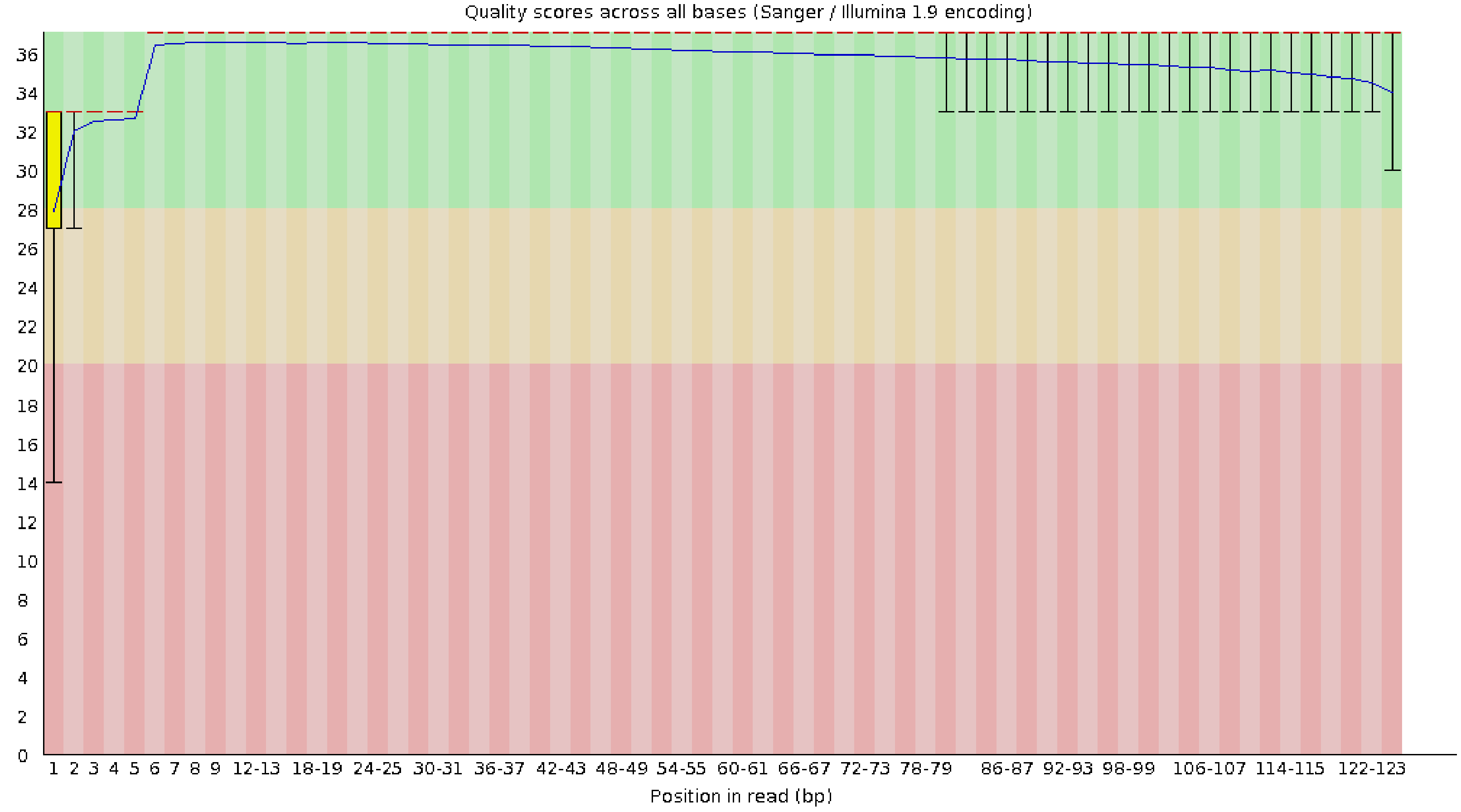


Figure 5.4: Per base sequence quality statistics for HiSeq 2000 forward, properly paired reads with adapters removed. The PHRED scores (y axis) represent quality of base calls. The bigger the PHRED score, the better the base quality of bases at that position.

FastQC report for HiSeq2000 PE overlapping reads, merged with FlaSh

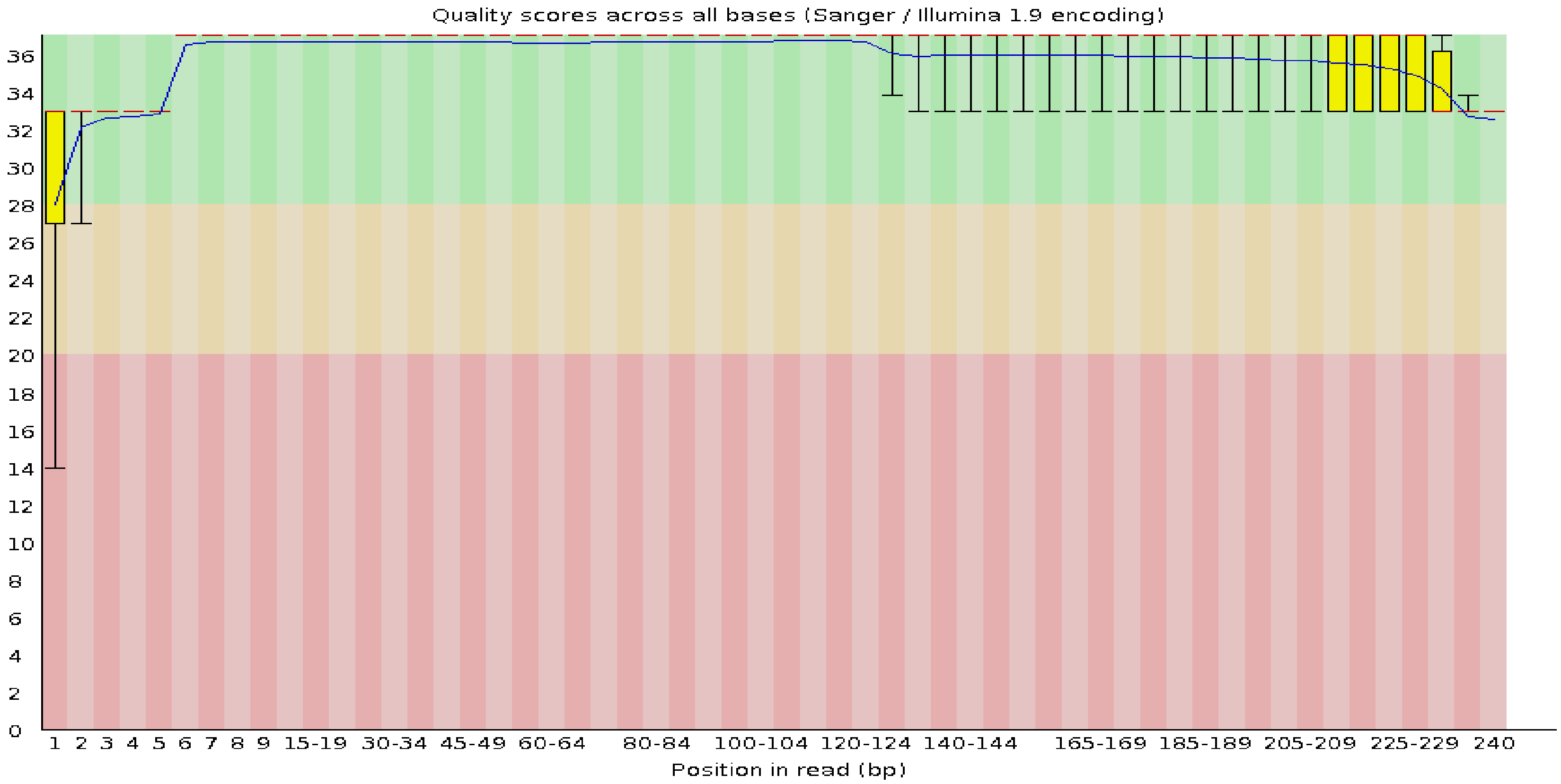


Figure 5.5: Per base sequence quality statistics for HiSeq2000 overlapping pair-end reads merged with FlaSh. The PHRED scores (y axis) represent quality of base calls. The bigger the PHRED score, the better the base quality of bases at that position.

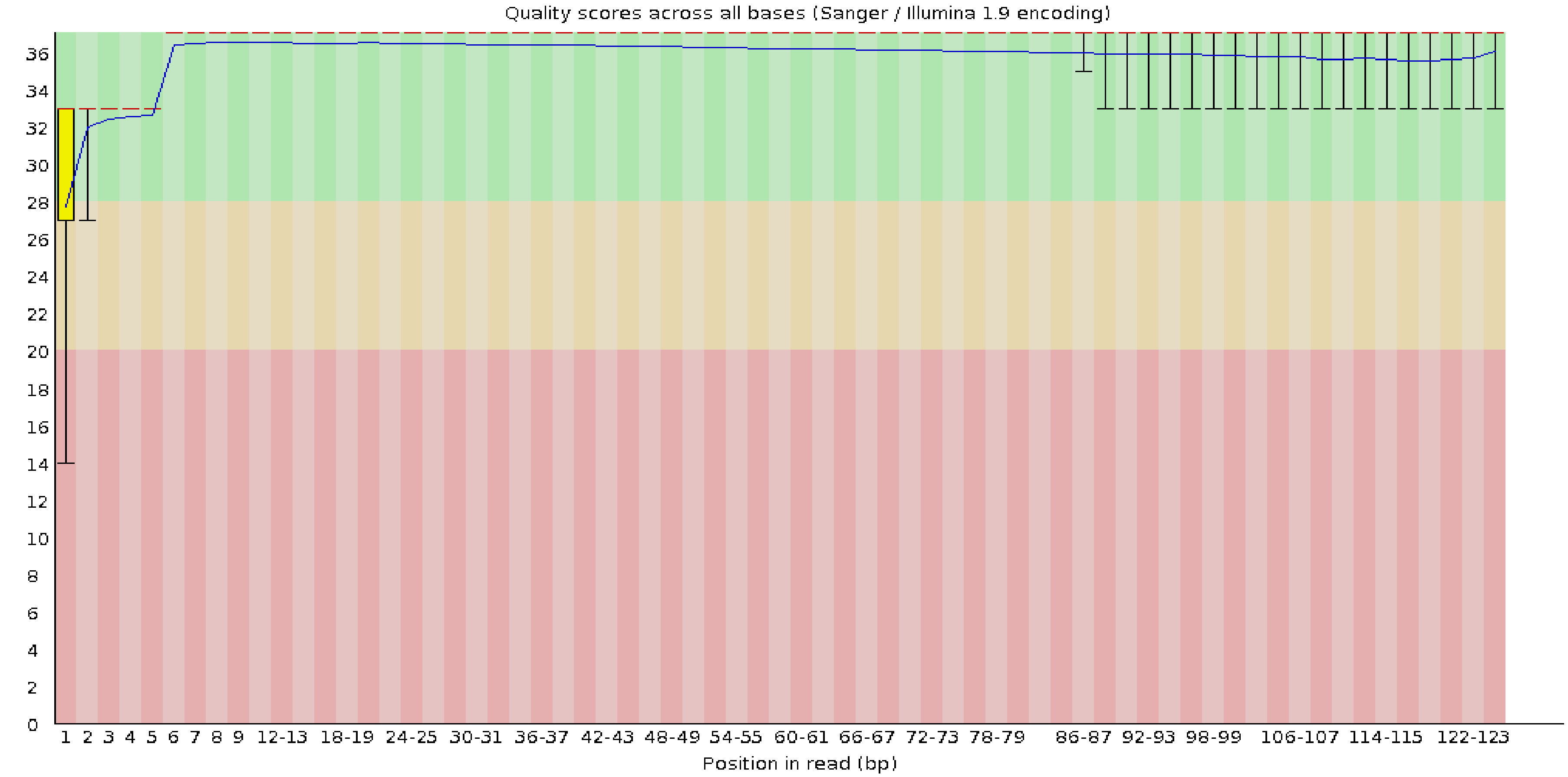


Figure 5.6: Per base sequence quality statistics for HiSeq 2000 unmerged forward reads. The PHRED scores (y axis) represent quality of base calls. The bigger the PHRED score. the better the base quality of bases at that position.

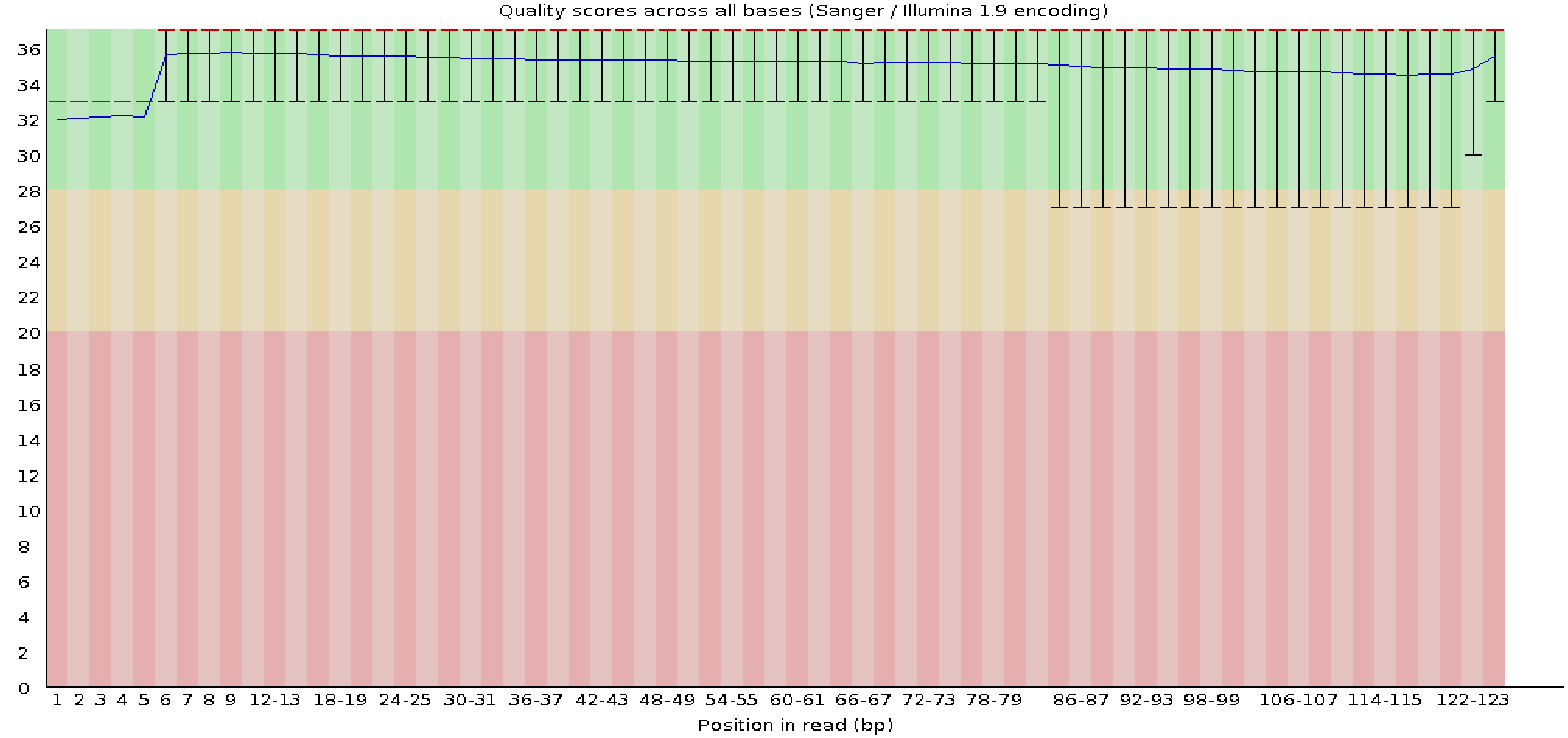


Figure 5.7: Per base sequence quality statistics for HiSeq 2000 unmerged reverse reads. The PHRED scores (y axis) represent quality of base calls. The bigger the PHRED score, the better the base quality of bases at that position.

Project data (assemblies and annotations) found at: https://www.dropbox.com/sh/v6vb8e7nocw5gkg/AADloBXU0Gc_3XIaAhPpz2nda?dl=0