IMPROVING EUKARYOTIC GENOME ASSEMBLY THROUGH APPLICATION OF SINGLE MOLECULE REAL-TIME SEQUENCING DATA GENOME: COFFEE LEAF RUST FUNGUS, *H. vastatrix*

by

Modupeore O. Adetunji

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics and Computational Biology

Fall 2014

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

IMPROVING EUKARYOTIC GENOME ASSEMBLY

THROUGH APPLICATION OF

SINGLE MOLECULE REAL-TIME SEQUENCING DATA

GENOME: COFFEE LEAF RUST FUNGUS, *H. vastatrix*

by

Modupeore O. Adetunji

Approved: _____
Shawn Polson, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Errol Lloyd, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved: _____
Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____
James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Shawn Polson, for all his support, guidance and encouragement throughout my Master' Degree. My deepest thanks to Prof. Cathy Wu, Dr. Chuming Chen, and my PhD. advisor, Dr. Carl Schmidt, for serving on my thesis committee and their encouragement and valuable suggestions throughout this thesis process.

I would like to give special thanks to Dr. Karol Miaskiewicz for making sure all the tools I needed were properly installed and my numerous jobs on the Biohen-Cluster executed smoothly, and to my colleague, Dan Nasko, for his assistance in understanding and implementing most of the tools used. I would also like to thank my friend, David Okoh who has supported me and enriched my life.

My sincerest thanks to Dr. Nicole Dinofrio and Dr. Mario Lucio Vilela Resende, who made this thesis possible and for their interest in sequencing the genome and studying the pathogenicity of the Coffee leaf rust fungus, *H. vastatrix*.

This thesis is dedicated to my parents, Arch. & Mrs Adetunji for their love, support, encouragement and guidance in all things me, and to my vivacious sisters, Damilola and Doyin Adetunji for being fun 24/7.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Coffee production is globally threatened by Coffee Leaf Rust disease. The fungal pathogen, *Hemileia vastatrix*, has been estimated to have the largest fungal genome known. With the absence of an available draft genome, genome sequencing and assembly is a fundamental step in understanding the infectious mechanism of the disease.

Next Generation Sequencing technologies (NGS) have been successfully applied for the whole genome sequencing and assembly of many genomes. Second-generation sequencing technologies, such as Illumina, are known for their high throughput but limited by short read lengths and systematic biases. The application of such technologies on large and more complex genomes result in numerous inaccuracies due to the inability to handle repeat regions and sequencing errors. Longer sequence data produced by third generation sequencing technologies, notably PacBio RS-II (Pacific Biosciences Inc.), show promise for overcoming such issues, demonstrated through accurate bacterial-scale genome assemblies and improvements to existing eukaryotic genomes by filling gaps and sequencing through repetitive sequence regions, but are limited by a high error rate and lower throughput.

In this study, we developed a three-stage pipeline to assess the performance of various *de novo* assembly algorithms, SOAPdenovo2, CLC Genomics Workbench (CLC), and Velvet; error correction tools, LSC and PacBioToCA; and the whole-genome shotgun assembler, Celera, for the whole genome assembly of large eukaryotic genomes using synthetic PacBio RS II CLR (Continuous Long Reads) and

Illumina paired-end reads created from the *Arabidopsis thaliana* genome as a proxy for *H. vastatrix*. At each stage, performance was assessed by reference genome mapping using BLASR and BWA-MEM, and was visualized using SeqMonk and CLC. The results showed the ability of the pipeline to produce long scaffolds with low nucleotide mapping error; the best performance overall was seen with the whole-genome shotgun assembly of SOAPdenovo2 scaffolds and PacBioToCA contigs, producing long genome scaffolds (>1.8Mb) with high N50, no captured gaps and spanning 93% of the reference genome with 1% nucleotide mapping error. These findings demonstrate that creating long genomic scaffolds for complex eukaryotic genomes such as *H. vastatrix* by NGS can be achieved with implementation of appropriate *de novo* assembly algorithms.

Chapter 1

INTRODUCTION

Coffea arabica is the one of the most important agricultural products in international trade, and a major source of income to many countries. Coffee production is threatened by the outbreak of the coffee leaf rust fungus, *Hemileia vastatrix*. *H. vastatrix* is the most destructive pathogen of coffee (15), causing a significant yield reduction in coffee production in major coffee producing regions, including South and Central America countries, thus having a huge effect in agricultural international trade (3) and is major contributing factor to the price increase in coffee futures according to the International Coffee Organization (ICO) March 2013 and May 2013 market report (31, 32).

Various genomic and transcriptomic analyses have been applied to understand this organism and it's interaction with its host, coffee (7). *Hemileia vastatrix* exists primarily as dikaryotic and an obligate parasite to coffee. Its life cycle is unique and complex due to not only its characteristic asexual production of urediniospores, but its ability to undergo "a hidden" sexual reproduction within the asexual spore, known as crytosexuality (11, 39). The urediniospore is the most common spore form of fungus for growth and reproduction, and this spore form is the material used for genomic sequencing.

H. vastatrix has been shown to have several strains/races; about eighteen strains have been identified for the fungus (36). The fungus has been estimated to have a genome size of 733.5Mb by flow cytometry (12), making it the largest fungal genomes

known to date. It is hypothesized that the large size of the genome is as a result of one or more whole-genome duplication events, with its first genome being least repetitive, thus postulated to contain its original ancestral genome of the species, while its' duplicate(s) is highly repetitive and highly mutative, thus attributing to defense and pathogenicity of the fungus (24). One of the major barriers to understanding *H. vastatrix* is the lack of an existing draft genome for this fungus.

A study for the whole genome sequencing and assembly was previously applied on nine isolates of *Hemileia* vastatrix. These genomes were sequenced using Illumina and 454 Roche next generation sequencing technologies and assembled using CLC assembler. Though the data is not available, Cristancho M. et al reported the *H. vastatrix* race II genome resulting assembly had an approximate genome size of 250Mb using CEGMA; genome coverage of 92%, 32% GC content with a huge number of repeated sequences, over 74%, but the average contig size below 900bp, N50 at 1,590bp and maximum contig size below 90KB (16). This assembled genome size and contig size distribution information shows better strategies for assembling this complex genome are required.

Whole genome sequencing and assembly is a computational problem for genome construction, different sequencing technologies have been developed to handle this problem. Using next generation sequencing reads, especially for large genomes, and given the limitation on their short read length, of less than 700bp, and systematic error bias from second generation sequencing technologies such as Illumina (HiSeq), Roche (GS FLX) or Life Technologies (SOLiD, Ion Torrent). A new sequencing technology platform, called PacBio RS (Pacific Biosciences Inc.), performs single molecule real-time (SMRT) sequencing to produce long read length libraries. The

technology has two general modes: the multiple pass Circular Consensus Sequencing (CCS) mode, this is characterized by small insert size (<3000bp) and low error rate; and the single pass Continuous Long Read (CLR) mode, which utilizes long insert lengths (up to 20kb) to produce very long read length libraries with an average of >6000bp, but with a higher random error rate (5, 20).

PacBio Biosciences RS is the first sequencing technology to produce very long reads (>1000bp), though it has its' shortcoming of low throughput and reduced accuracy, PacBio offers the opportunity to create accurate genome assemblies and improve draft genomes efficiently and at low cost. The successes of PacBio in whole genome assembly have been shown clearly with various prokaryotes, such as *Escherichia coli* (63), *Meiothermus ruber* (14), *Pedobacter heparinus* (14), *Salmonella enterica* subsp. *enterica serovar* Typhimurium (28), and some slightly larger eukaryotic genomes, Atlantic Cod fish (57).

With the advent of third-generation sequencing technologies, the initial bottleneck of incomplete and inaccurate genomes based on second-generation, short-read data can be addressed, because PacBio RS sequencing method generates very long reads which facilitate complete genome assembly by filling gaps and joining existing contigs (5). However, third-generation sequencing accuracy and coverage is much lower than that of second-generation methods, thus hybrid genome assembly protocols using both second- and third-generation reads are being applied for creating near-complete and accurate genomes (37).

Novel bioinformatic algorithms, such as Celera Assembler PBcR pipeline: PacBioToCA (38), LSC (73), Cerulean (17) and ALLPATHS-LG (26), have been developed and implemented to handle either or both second- and third- generation

sequencing technologies simultaneously. These algorithms primarily function in combining the high quality second-generation short reads with the third-generation longer reads to produce very long genomic contigs, and have been shown to be largely successful with prokaryotes.

Due to the coffee rust fungus genome's complexity and large size; PacBio RS SMRT DNA sequencing provides an opportunity to ensure better and longer assemblies of this difficult genome using a hybrid *de novo* assembly approach. The coffee leaf rust fungus is currently being sequenced using a second-generation sequencing technology, Illumina HiSeq, at 100X coverage to produce high quality paired-end short read datasets, and third-generation SMRT DNA sequencing technology, PacBio RS II, at 20X coverage to produce long-read datasets.

Thus, our overall objective is to create an extensive whole genome assembly pipeline that utilizes various assembly tools that can efficiently handle both second- and third- generation sequencing technology datasets to produce long, accurate and high-quality assemblies of this large fungus genome.

Three strategies were developed for this objective; The first strategy adopts the de Bruijn graph mapping algorithm, which involves whole genome *de novo* assembly of the short high quality reads using three different *de novo* assemblers; CLC genomics workbench (http://www.clcbio.com/), SOAPdenovo (46) and Velvet (74) to generate contigs/scaffolds. After which these contigs/scaffolds were improved using the PacBio CLR reads to fill or reduce as many captured gaps as possible and scaffolding (i.e. joining contigs) using PBJelly (22) . The second strategy creates high quality long reads and utilizes mapping algorithms to error-correct the long PacBio reads with the higher quality Illumina short-reads using the Celera Assembler PBcR pipeline: PacBioToCA

utility (38) and LSC (73). The third and final strategy involves the final scaffolding stage; which evaluates and merges the results generated from the first strategy and the second strategy to create genomic scaffolds. This will be done using the OLC (Overlap-Layout-Consensus) assembly algorithm in Celera Assembler (54).

The specific aims for this project are to; generate *de novo* contigs/scaffolds with high quality Illumina paired-end reads and improve them with SMRT long reads; error-correction of SMRT long reads with Illumina paired-end reads; and the final aim is to develop an overlap assembly of the improved-*de novo* contigs and error-corrected SMRT reads to create long DNA fragments.

Since DNA sequencing of the Coffee Leaf Rust fungus was not yet available; a fully annotated and reasonably complete eukaryotic genome was used as a proxy to ascertain the efficiency of our pipeline. The genome chosen was the *Arabidopsis thaliana*, as it is fairly complete and well annotated, and comparably sized to the initial expected genome size of over 100-200Mb. Subsequent data has since indicated that the genome may be much larger, ca. 733.5Mb (12), but this should not reduce the usefulness of using *A. thaliana* as a proxy. Pacbio RS II continuous long reads and Illumina paired-end reads datasets were generated using the Profile-based Illumina paired-end reads simulator (pIRS) (29) and the PacBio reads simulator (PBSIM) (59) tools respectively. These reads are simulated with similar simulation coverage as the estimated sequencing coverage for *H. vastatrix* in relation to genome size.

This thesis is structured as follows: First, we provide a detailed background and literature review on the coffee rust disease and whole genome sequencing and assembly. Second, we explain the pipeline and synthetic sequencing data used for our

genome assembly pipeline. Finally, we discuss extensively the different strategies applied for the workflow and conclude the thesis, along with the future work.

Chapter 2

BACKGROUND AND LITERATURE REVIEW

Genome sequencing and assembly is an imperative first step in whole genome analysis. Not only is all the genetic information (in the form of DNA) of an organism determined, it provides the benchmark for important approaches such as genomic, variant, transcriptome analysis and other tools that address issues related to DNA genome sequences, such as sequence alignment, gene prediction, protein structure prediction and function prediction.

After the conclusion of the Human Genome Project, the approach for whole-shotgun genome sequencing and assembly encountered major changes with the development of high-throughput sequencing technologies, such as Roche 454 Pyrosequencing, SOLiD, IonTorrent and Illumina. These second generation sequencing technologies provided fast, relatively inexpensive, and high throughput of data through amplification, thus enabling the commencement of numerous genome sequencing projects and a multitude of related projects for the study of genomes. The limitations of these high throughput sequencing platforms; such as very short-read length, and amplification bias, brought about the introduction of single-molecule sequencing technologies, which are capable of sequencing very long sequence read-lengths, examples of such are PacBio RS Biosciences, Oxford nanopore, and GnuBIO; with PacBio RS currently being the only commercially available sequencer. In this thesis, these sequencing platforms are applied through an array of whole genome *de novo*

assemblers that can overcome the bias of each individual platform and implemented together can improve assemblies of large eukaryotic genomes.

This chapter covers the background of the coffee rust disease, and a literature review for genome sequence assembly, which is necessary for understanding the computational problems in genome sequencing and solutions made for these problems. An overview of next generation sequencing technologies and its' application in whole-genome shotgun sequence assembly are also presented.

2.1 Coffee Rust Disease

Coffee leaf rust was first reported in Ceylon (Sri Lanka) in 1869 and was known for its devastating effects on *Coffea arabica* crop production in that country (58). In the 1950s, the disease was widespread across Africa and many Asian countries where coffee was grown commercially, and in 1970, it was found in Bahia, Brazil and has been discovered in most of the South and Central American countries, making the coffee leaf rust the most widely spread tropical plant disease in all coffee producing regions (Figure **2.1**) (8). However, the prevalence of the disease is dependent on weather and farming efforts. In recent years, coffee rust has been reported to occur at unusually high and severe levels in all major coffee growing regions, thereby causing an incremental increase in coffee price futures and yield reduction in coffee production (4).

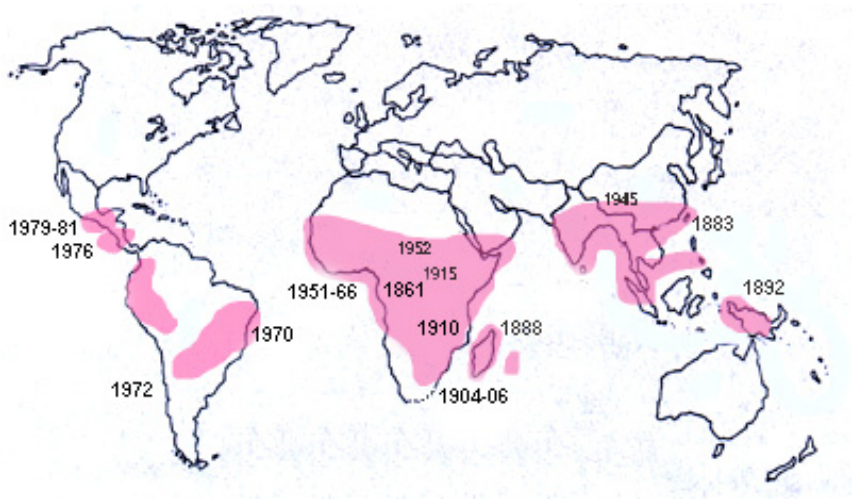


Figure 2.1 World distribution of the coffee rust fungus, with the dates it was first discovered, from ref. (3).

The coffee leaf rust disease, also known as "roya", is caused by the fungus *Hemileia vastatrix*. It is a dikaryotic obligate parasite known to thrive on coffee trees, free water and in tropical climates. The fungus can only complete its life cycle on the leaves by which the urediniospores (asexual spores) attach to the underside of the leaves and in the presence of appropriate conditions, i.e. free water and high humidity, germinate through the stomata, taking over the leaf's nutrition (Figure **2.2**) (64). The fungus can survive as mycelium in the living tissues of the host, as dry urediniospores for about 6 weeks while the basidiospores (sexual spores) do not infect the plant. It is typically recognized by the yellow-orange powdery lesions or spots on the underside of coffee leaves and chlorosis on the upper side causing impaired photosynthetic capacity of infected leaves, premature defoliation or leaf drop, and reduction in vegetative and berry growth (3, 58).

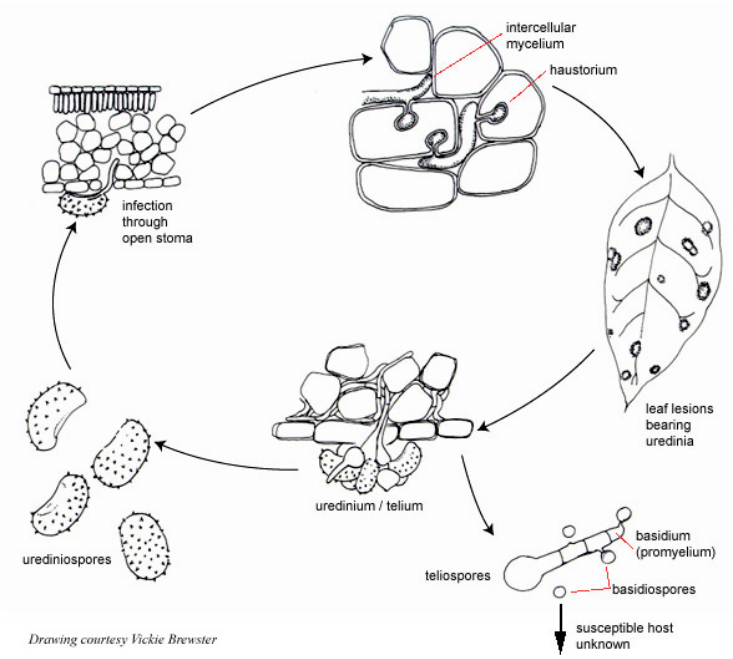


Figure 2.2 Life cycle of Coffee Leaf Rust Fungus, from ref. (3).

Hemileia vastatrix is morphologically distinct from other rust fungi, which are kidney-shaped with half-smooth (and half-rough) spores not the typically round to oval shape with fine spines over their entire surface, and they exists primarily as a dikaryotic (two nuclei in each compartment of a hypha). Its life cycle is unique and complex due to not only its characteristic asexual production of urediniospores, but its ability to undergo "a hidden" sexual reproduction disguised within the asexual stage, or cryptosexuality (11). This explains why multiple physiological races of the rust exist in all coffee-producing countries (genetic diversity); with at least 45 races identified in Portugal, 6 races in Colombia, 18 races in Tanzania, and over 20 races in Kenya, thus the rapid genetic change and emergence of new races make it possible to infect initially-resistant plants (25, 36, 65).

Advancements have been made in understanding this organism and it's interaction with its host, coffee. An important advancement was its genome size estimation using flow cytometry (19) to be about 733.5Mb (C-value of 0.75pg); making it the largest fungal genome known to date, compared to the average fungal genome (*Ascomycetes sp.*) of *ca.* 37Mb (40). Furthermore, with a base composition of AT = 65.4% and GC = 34.6%, indicates a significantly lower GC content than for other rusts, and higher AT content, which suggests a high ratio of repetitive regions, and it's dependence on a living host to complete it's life cycle (biotrophy) (12). Also, the large C-value could be as a result of whole-genome duplication event(s); with the ancestral genome being least repetitive and the duplicates being highly repetitive and highly mutative, thus attributing to its genetic diversity and virulence of the fungus (24).

Another important advancement was the application of next generation sequencing technologies, such as 454 Roche and Illumina sequencing platforms, for genome studies on nine races of the coffee rust fungus, and protein prediction in comparison with different similar fungal genomes by Cristancho et al (16). Though the data derived are not publicly available, the paper showed agreement in the mean GC content with the flow-cytometry approach of 33%, and read duplications were observed in about 20% of the Illumina dataset. Furthermore, a draft assembly for the fungus was created using the CLC assembler, resulting in 396,264 contigs; having total length = 333,481,311bp, maximum length = 85,126bp, average length = 841.56bp and N50 contig size = 1590bp. Some of the contigs generated were homologous to other fungi genomes; especially with *Puccinia graminis* genomes, showing several blocks of genome conversation, however most of the contigs generated were not similar to any organism (16).

2.2 Genome Sequencing Technologies

The relevance of genome sequencing has been ever increasing over the past decades, from the arduous and time-consuming Sanger biochemistry to the cost effective, high throughput Next Generation Sequencing technologies. Genome sequencing has transformed every area of biological research (45). This section discusses the genome sequencing platforms commercially available and the ones applied in this study.

DNA sequencing was first described by Maxam and Gilbert in 1977, known as the Maxam-Gilbert or "chemical" sequencing method, the principle involves the use of different chemicals to cleave radiolabeled DNA between specific bases, and the resulting fragmented DNA are run on a polyacrylamide gel and analyzed to determine the DNA sequences (49). In the same year, the Sanger or chain-termination sequencing method was also developed by Frederick Sanger, the principle employs the use of radiolabeled dideoxynucleotides-triphosphate (ddNTPs), also known as terminating triphosphates, that prevent the formation of phosphodiester bonds during DNA elongation based on the respective terminating nucleotide and by polyacrylamide gel electrophoresis, the DNA sequencing is then read along the gel (66). The classical Sanger biochemistry is able to sequence long reads (650 - 800bp), with high accuracy, but at very low throughput and a very expensive and time-consuming process (48).

The need for a faster and cost effective method of DNA sequencing brought about the development of "Shotgun Sequencing" approach to sequence longer sections of genomic DNA, this approach involves the use of enzymes or physical sheering to break down long DNA fragments, sequencing each smaller fragment and aligning these

sequenced fragments based on overlap, this technique coupled with parallel sequencing characterized the next generation of high throughput sequencing platforms (75). Several strategies for low-cost sequencing have then been developed and can be grouped into five categories, which includes microelectrophoretic methods, 'sequencing by hybridization', cyclic-array sequencing on amplified molecules, cyclic-array sequencing on single molecules, and non-cyclical, single-molecule, real-time methods (68).

After the completion of the human genome project (HGP), various implementations of the cyclic-array sequencing strategy were developed to allow larger-scale DNA sequencing; the principle of cyclic-array sequencing involves a repetitive process of enzymatic manipulation and imaging-based data collection (69). These sequencing platforms were tagged "second-generation" or, the so-called, "High Throughput" or "Next Generation" Sequencing (HTS/NGS) technology (69), and applications of these platforms increased speed, throughput capacities and reduced overall sequencing costs. Though these NGS instruments apply similar principle (cyclic-array sequencing), their sequencing biochemistry are diverse, as a result generate differences in sequence read lengths, error rates and error profiles between each other (48), also relative to Sanger sequencing data, the major drawback of NGS is their shorter read length (35-250bp), higher base-call error rates and novel platform-specific artifacts (45). Commercially available NGS platforms include Roche - 454 GS FLX Pyrosequencer, ABI - SOLiD, Illumina/Solexa - Genome Analyzer (I, II) and HiSeq, Polonator and Helicos (HeliScope Single Molecule Sequencer technology) platforms; among which, only three of these NGS platforms – Roche/454 FLX, Illumina/Solexa and SOLiD Analyzer – were widely applied (Figure **2.3**).

NGS TECHNOLOGIES	AMPLIFICATION APPROACH	SEQUENCING CHEMISTRY	COST PER MEGABASE	COST PER INSTRUMENT	1ST ERROR MODALITY	AVERAGE READ LENGTH	READS/RUN	TIME/RUN
454 - GS FLX	Emulsion PCR	Polymerase (pyrosequencing)	~$10	$500,000	Indel	700bp	450Mb	7 hours
ILLUMINA/SOLEXA - HiSeq	Bridge PCR	Polymerase (sequencing-by-synthesis using reversible terminators)	~$0.07	$540,000	Substitution	50 - 100bp	600Gb	4 - 7 days
SOLID	Emulsion PCR	Ligase (octamers with two-base encoding)	~$0.13	$595,000	Substitution	50 - 100bp	120Gb	5 - 14 days
POLONATOR – G.007	Emulsion PCR	Ligase (nanomers)	~$1	$170,000	Substitution	13bp	12Gb	~5 days
HELISCOPE	Single molecule	Polymercase (asynchromous extensions)	~$1	$990,000	Deletion	30bp	37Gb	~8 days

Figure 2.3 Next generation Sequencing Technologies. Adapted from ref. (48, 51, 69).

The Roche (454) GS-FLX Genome Sequencer was the first commercial NGS platform introduced in 2004 by 454 Life Sciences. The sequencer works on a principle of sequencing-by-synthesis known as “pyrosequencing” technology; using emulsion PCR, the DNA is amplified on the surfaces of numerous agarose beads with complimentary ligated oligomer adaptor sequences, then by PCR amplification each bead will contain over a million copies of the original annealed DNA fragment in order to produce detectable signal for the sequencing reaction (48), and during pyrosequencing via ATP sulfurylase and luciferase, imaging of the light emission from pyrophosphate molecules released during polymerase addition of nucleotides are recorded (60). The GS-FLX can generate ~100Mb reads per 7 hour-run, at average read length of ~250bp, with a maximum capacity of ~600bp, also with a very high base accuracy (over 99%), it is however prone to higher error rate due to homopolymers,

resulting in insertion-deletions. Thus the 454 sequencer has the longest short reads among all NGS platforms, but the lowest yield per run making the per-base cost of sequencing much greater compared to other NGS platforms (69, 75).

The Illumina Genome Analyzer (GA) was the second commercial platform introduced in 2006, and is currently the most widely used platform. The sequencer also works on the principle of sequencing-by-synthesis approach by which the library consist of mixture of adaptor-flanked fragments; using bridge PCR, the four fluorescence-tagged nucleotides, or reversible terminators and the forward and reverse PCR primers, the amplicons derived from the single template molecules are clustered and immobilized to a single location, with each cluster containing over a thousand clonal amplicons, the resulting 'clusters' are further amplified with which the fluorescently labeled reversible terminators is imaged. The Illumina platform could generate 36bp average read length, but over the years, the Illumina GA platform underwent a lot of improvement in read throughput, sequence length and time of sequencing. In 2010, the Illumina HiSeq platform was introduced generating over 200Gb per about a week run, at read length of 150bp and raw base accuracy greater than 99.5% (21). The major drawback of this platform is short read length, but it is the most adaptable and often considered easiest to use sequencing platform with high throughput and inexpensive costs (75).

The SOLiD (Sequencing by Oligo Ligation and Detection) Analyzer, which was commercially released in 2007 by Applied Biosystems Inc. (ABI), uses a unique sequencing-by-ligation and detection technology; using oligo adapter-linked DNA fragments and magnetic beads with complimentary oligoes are amplified by emulsion PCR, followed by ligation-based sequencing. ABI SOLiD platform can generate ~3Gb

reads per run, at average read length of ~35bp, with 99.85% accuracy (48), in following years, the SOLiD 5500 series was introduced having immense upgrades to the sequencing systems in not only improving the read length to 85bp and output of 30Gb per a week run with 99.99% accuracy, also in reducing sequencing costs, but similar limitations as with other platforms of short read length (21).

Similar to conventional sequencing (i.e. Sanger sequencing), the second generation sequencing platforms require PCR amplification to generate high-throughput DNA sequencing, however the amplification process introduces sequencing errors as well as amplification bias (67). In addition, the massive amounts of short read data generated by NGS platforms become a major challenge for data storage and informatics operations, such as sequence alignment and assembly across repetitive regions (44). To address these issues, a new strategy of DNA sequencing called third generation sequencing technologies was developed; it incorporates a non-cyclical, single-molecule, real-time sequencing strategy in which PCR amplification is not required, the DNA molecule is captured in real-time by fluorescent or electric current signal (44).

The Single-molecule real-time (SMRT) sequencing approach, developed by Pacific Biosciences in 2011, is the first commercially available third generation sequencing platform that made use of modified enzyme and direct observation of a single molecule to sequence a strand of DNA. It performs single-molecule sequencing by identifying nucleotides, which are phospholinked with distinctive flours, emitting fluorescence as each nucleotide incorporates into a single DNA strand (75). SMRT cells consist of thousands of zero-mode waveguides (ZMW), which comprises of over 75,000 ZMWs on a 100nm metal file deposited on a glass substrate. A ZMW is a tiny

hole serving as a protective screen that premits passage of visible laser light to the DNA molecule. Within each ZMW, a DNA polymerase molecule is anchored to the bottom glass surface, then the holes are flooded with fluorescence-labeled nucleotides, which diffuse through the bottom of the ZMW, and by laser excitation, when the correct nucleotide is detected, it fluorescence (67). This platform requires minimal amounts of reagents and sample preparation, with sequencing run taking only minutes instead of days as with prior sequencing technologies. Besides the several advantages over previous platforms, such as no PCR-amplification, single molecule sequencing, and shorter sequencing time (47), PacBio RS II SMRT sequencer can generate very long read lengths ranging to over 20kb with average read length > 5kb for raw reads (CLR - Continuous Long Reads) and of about 2.5kb for error-corrected reads (CCS – Circular Consensus Sequence), thus resolving initial bottlenecks of using shorter reads, enabling more accurate *de novo* assembly and improvement of pre-assembled genomes (14, 35). However, some challenges compared to prior sequencing platforms are evident such as low read accuracy of 83-87% (14, 51), high random sequencing errors; majority of which are contributed by insertions and deletions, and low throughput (10).

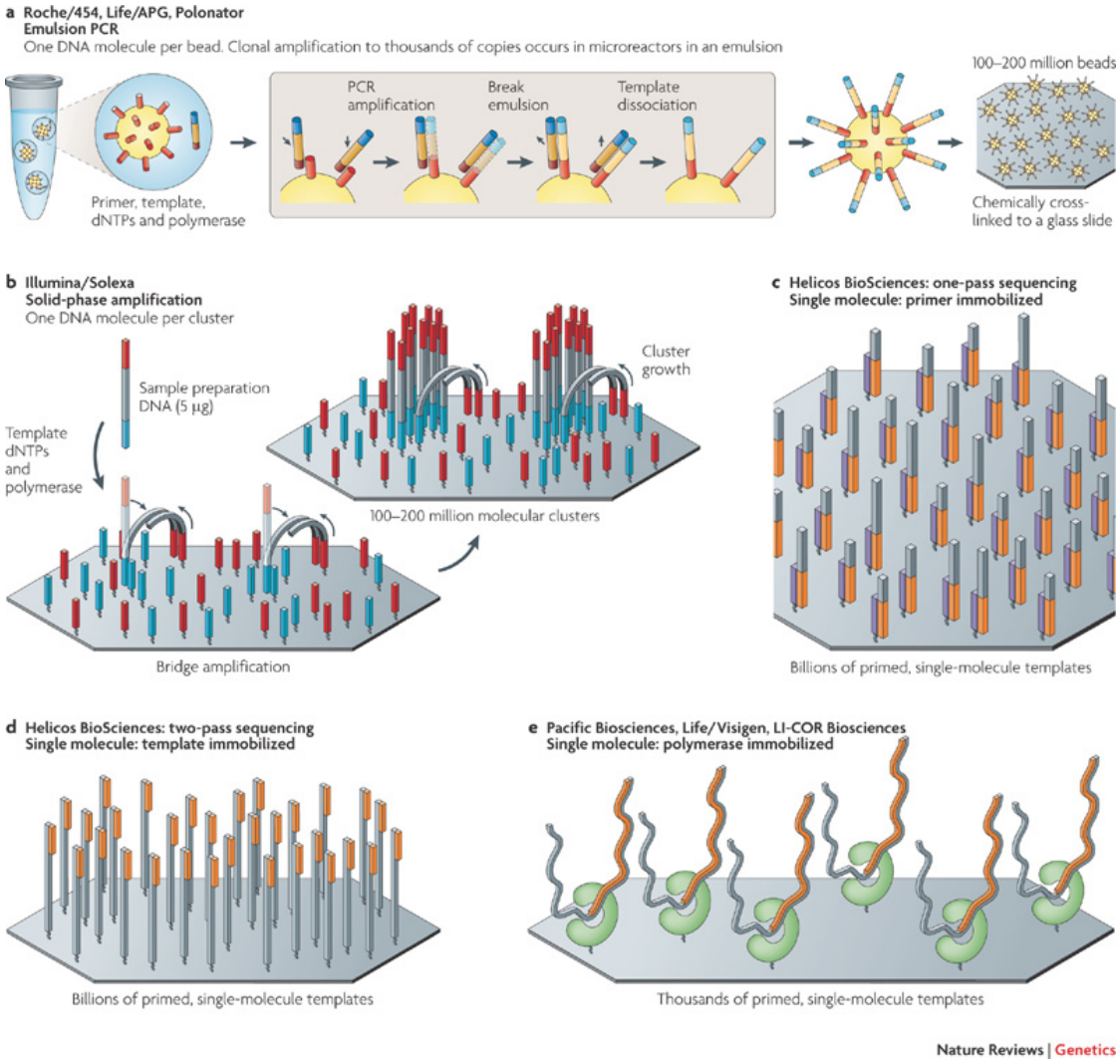


Figure 2.4 Sequencing platforms strategies, from ref. (51)

2.3 Genome Assembly

Genome assembly involves the reconstruction of a DNA sequence from a collection of randomly sampled fragments. The availability of huge amounts of data from the different next generation sequencing platform created an avenue for improved

whole-genome shotgun assembly algorithms or software (53). The pioneer method of genome assembly followed the overlap-layout-consensus paradigm; an example is the CELERA assembler, which was the first assembler, in 2000, used to accomplish the whole-genome assembly of a multi-cellular organism, Drosophila (54). Despite the success of this paradigm in assembling long fragments, the approach is unable to handle repeat regions in the overlap graph. To address this, Pevzner et al. introduced the EULER algorithm that was able to partially address the "repeat problems" using de Bruijn graphs (61) making this algorithm suitable for high-through short-read sequencing and currently, numerous genome assemblers had been developed to handle high throughput data from NGS platforms. Assemblers can be broadly grouped into three graph-based algorithms: greedy graph algorithms, Overlap-Layout-Consensus (OLC) graph methods and the Sequencing-By-Hybridization (SBH) or, commonly known as, De Bruijn graph paradigms (53). Graph-based algorithms function by first determining the pair-wise overlap information for all the sequenced reads and represent them as unweighted edges in a string graph. A collection of paths satisfying specific properties are weighted and computed to contigs (62).

The greedy algorithm is constructed based on an incremental approach by which reads with the best overlap are joined and further extended with the next highest scoring overlap until no more extensions can be done, and to avoid the incorporation of false-positive overlap, the algorithm applies a heuristic correction step in the region of overlay. The greedy paradigm is easy to implement, but has a local assembly process – unable to process long-range mate-pair links – and requires large amount of computational memory (56). Many of the first NGS assemblers applied this approach; they include SSAKE (72), SHARCGS (18) and VCAKE (34).

The OLC graph approach represents the reads as nodes and the overlaps as edges in a graph and determine the simplest path traversing all the nodes, i.e. the Hamiltonian path. The algorithm also takes into account the global relationship between reads, and other relationships such as the mate-pair links, which can be used to correctly assemble repetitive areas (55). The OLC paradigm functions in three stages; the first stage, known as the OVERLAP stage, involves the overlap discovery of all pair-wise read comparison to create an overlap graph; the second or LAYOUT stage removes all transitive edges (redundancies), and ambiguities from the graph resulting in a refined path layout; and the final stage performs a multiple sequence alignment to determine the precise layout and generate a consensus sequence for each contig, i.e. the CONSENSUS stage (Figure **2.5**B). The layout stage stores limited information about the graph, thus enables a memory efficient process (53). This approach is most useful for longer reads: such as Sanger, 454 and PacBio sequencing data (55) and has been applied in well-known assemblers such as CELERA assembler (54), ARACHNE (6) and CABOG (52).

In the de Bruijn graph approach, the graph encodes overlaps as nodes and the reads having a specific overlap with the corresponding node for that overlap as edges. The algorithm process involves breaking the reads into a collection of overlapping k - 1 k-mers (a k-mer is a substring of length k), next a de Bruijn graph is created in which each edge corresponds to a k-mer from the original sequence reads (Figure **2.5**C). With this paradigm, the reads are not directly modeled, but are implicitly represented as paths through the de Bruijn graph, and the path corresponding to all the edges only once (that is, the Eulerian path) is determined (55). The SBH approach requires approach is based on exact matches, and error correction approaches are important before and during

assembly for achieving high-quality assemblies. Also the de Bruijn graph is linear in input size and stores limited information, thus making it suitable for high coverage sequencing data and ensuring fast computation of Eulerian paths (or contigs). This approach has been applied for most modern assemblers targeted at short-read sequencing data, such as Velvet (74), SOAPdenovo (46), ALLPATHS (9) and ABySS (70).

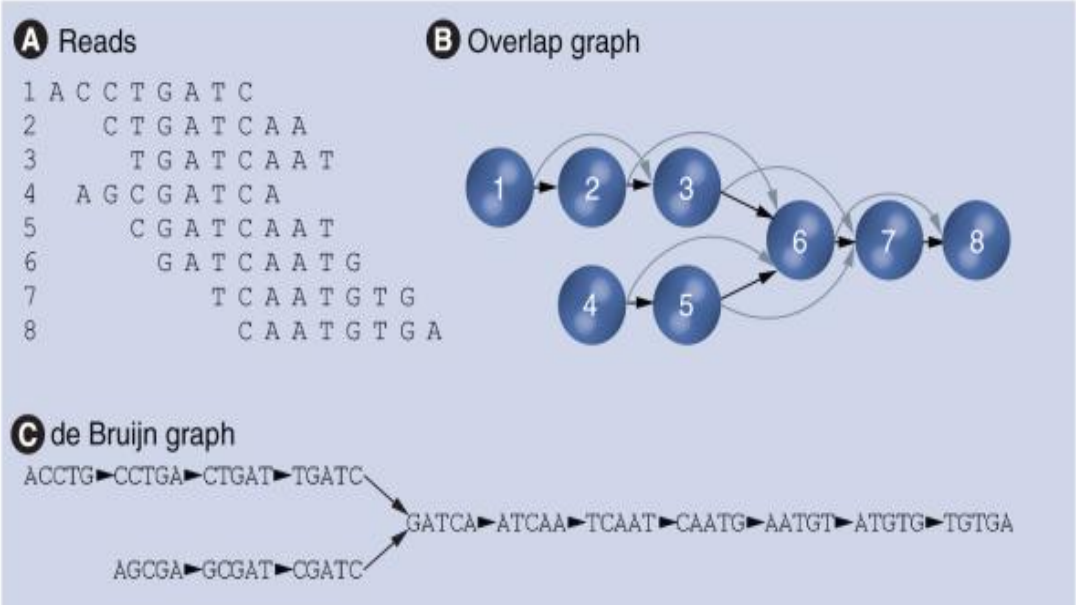


Figure 2.5 Overlap and de Bruijn graph construction on the same sets of reads, from (27).

Chapter 3

WORKFLOW OF GENOME ASSEMBLY PIPELINE

In the chapter, the whole genome assembly pipeline is discussed. We hypothesize the pipeline will ensure production of long, accurate and high quality assemblies for the fungus. This pipeline entails an extensive array of *de novo* genome assembly algorithms or tools that can efficiently leverage both second- and third-generation sequencing technology datasets.

Various genome assembly projects have typically shown successful implementation of a single next generation sequencing platform, but for smaller genomes. Given the limitations of the different generations of sequencing platforms, of either short read length, substitution and InDel errors, amplification bias, or low coverage, we hypothesized the best approach for the whole genome assembly of large eukaryotic genomes will be a combinatory approach that utilizes the advantages of both the second- and third- generation sequencing datasets.

Our approach involves the application of different *de novo* genome assembly tools that can efficiently handle the different sequencing datasets to produce high quality scaffolds, for this approach three specific aims were developed.

The first aim involves *de novo* assembly of the short reads creating contigs and scaffolds and improvement of these contigs using the PacBio reads. This strategy adopts the de Bruijn graph mapping algorithm, which involves whole genome *de novo* assembly of the short high quality reads using three different *de novo* assemblers; CLC genomics workbench (http://www.clcbio.com/), SOAPdenovo (46) and Velvet (74) to generate contigs/scaffolds. After which these contigs/scaffolds were improved using the

PacBio CLR reads to fill or reduce as many captured gaps as possible and scaffolding (i.e. joining contigs) using PBJelly (22).

The second aim involves creation of error corrected PacBio reads with the aid of the Illumina short reads. This second strategy utilizes both mapping algorithms respectively to error-correct the long PacBio reads with the higher quality Illumina short-reads using the Celera Assembler PBcR pipeline: PacBioToCA utility (38) and LSC (73).

The third aim involves scaffolding of the error-corrected PacBio reads and generated contigs or scaffolds to create genomic scaffolds. This strategy involves the final scaffolding stage; which evaluates and merges the results generated from the first strategy and the second strategy to create genomic scaffolds using the OLC (Overlap-Layout-Consensus) assembly algorithm in Celera Assembler (54). Figure **3.1** shows the schematic overview of this pipeline.

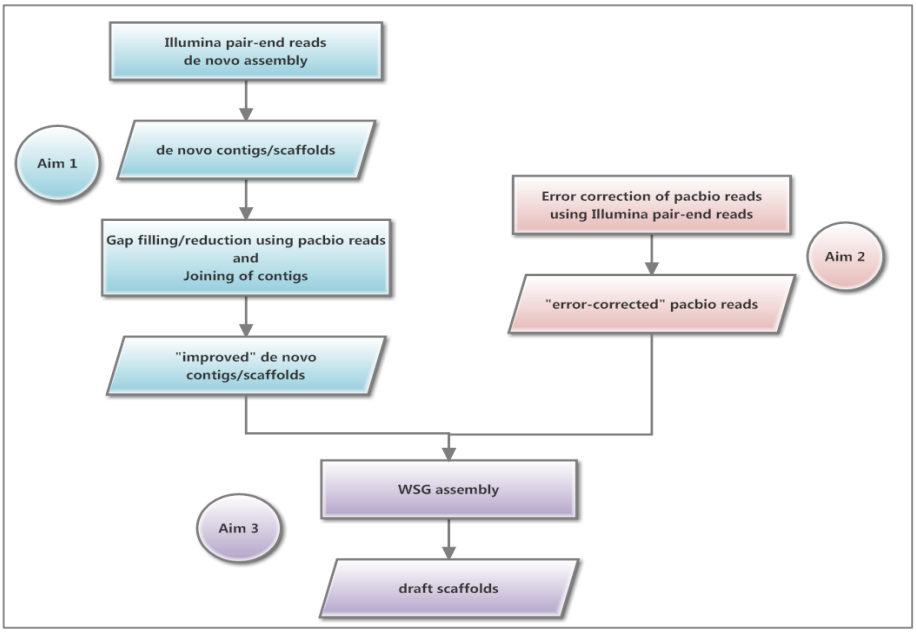


Figure 3.1 Schematic workflow of Genome Assembly Pipeline.

Chapter 4

SIMULATED SECOND- AND THIRD-GENERATION SEQUENCING DATASETS

The coffee leaf rust fungus is currently being sequenced using a second-generation sequencing technology, Illumina HiSeq, to produce high quality paired-end short read datasets, and third-generation SMRT DNA sequencing technology, PacBio RS II, to produce long-read datasets. This thesis will use simulated data to test the optimal pipeline for application to *H.* vastatrix once completed. To ensure the efficiency of the pipeline, a fully annotated and reasonably complete eukaryotic genome was used, which was *Arabidopsis thaliana* (TAIR10).

Arabidopsis thaliana has been well characterized as a flowering plant and a model organism for higher plant in plant biology. The plant has a genome size of approximately 119Mb consisting of five (5) chromosomes and two (2) plasmids, and with a GC content of 35.97%. The plant genome size was comparable to the initially expected fungal genome size of about 250Mb (16), though subsequent data have since shown that the genome is much larger, ca. 733.5Mb (12), but this does not reduce the efficacy of *A. thaliana* as a proxy, because of its similar estimated GC content and its hypothesized whole-genome duplication events attributing to the fungus' diversity and pathogenicity (24).

As actual sequence data for a comparable genome was not available, synthetic PacBio RS II continuous long reads and Illumina paired-end reads dataset were generated using respective simulation tools for next generation sequence data and these reads were generated with similar coverage as for the expected coffee leaf rust fungal

data in relation to genome size, which were; for the PacBio RS II reads at 20X coverage and the Illumina paired-end reads at 100X coverage.

$$Reads\ coverage = \frac{reads\ length * expected\ number\ of\ reads}{genome\ length}$$

Equation 4.1 Read Coverage Estimation.

Simulated data is a very important guiding tool in software development, statistical methods improvement and tool performance evaluation. There are numerous simulation software available, which are able to generate next-generation sequencing reads that have the identifying characteristics of the real data, some of which are GemSIM (50), MetaSIM (23), NeSSM (33), Grinder (2); for simulating metagenomics data, pIRS (29), wgsim from SAMTOOLS; for generating Illumina sequencing reads, ART (30), DWGSIM (https://github.com/nh13/DWGSIM); for the different Next-Generation Sequencing reads, PBSIM (59); for generating synthetic PacBio reads, and many others.

These simulation softwares are able to generate sequence reads with the appropriate sequencing errors; substitution and insertion-deletion (INDEL) errors, expected for the different sequencing platforms. GemSIM (General Error-Model based SIMulator) is a command line package written in Python. The tool creates single and paired-end reads of second generation sequencing data (Illumina and 454 Roche) using provided sample error models in SAM and FASTQ format. GemSIM can also assign quality scores to synthetic reads, and has been implemented for metagenomic data (50). Grinder is a platform-independent software package written in Perl and uses the Bioperl toolkit and Mersenne Twister algorithm to generate random numbers. It simulates in silico amplicon and shotgun datasets of next generation sequencing data. Grinder can generate single and paired-end reads with varying insert size, and introduce

experimental artifacts, like chimeras and biological biases from variations between different species. Grinder has also been applied extensive in both genomics and metagenomics (2). MetaSIM is written in Java and can be implemented via the command line or graphical user interface (GUI). The software requires a set of genome sequences and an abundance profile and can create Sanger, 454 Roche and Illumina reads based on defined error models without quality values. It is applied preferentially for metagenomic data (23). pIRS (Profile-based Illumina paired-end reads simulator) is a command line precompiled package written in C++ language. The tool generates only 100bp Illumina paired-end reads using empirical Base-Calling profiles and errors (substitution, insertion, deletion and other variations) are also introduced (29). ART is also implemented in C++ and creates single-end, paired-end and mate-pair reads of the three major next-generation sequencing platforms (Illumina 454 Roche, SOLiD) using empirical error models or quality profiles (30). DWGSIM is a modified version of wgsim from SAMTOOLS. It simulates both single- and paired-end reads of varying read lengths and insert sizes, and can assign quality scores to synthetic reads using error models. PBSIM (PacBio read simulator) is written in C language and is the only PacBio read simulator that can produce both types of PacBio libraries (continuous long reads, and circular consensus sequencing reads). PBSIM simulates reads with the characteristic features, such as the log-normal distribution, of real PacBio libraries using either a model-based or sampling-based method (59).

Given the numerous simulation software for next generation sequencing data, various simulation software that have error modeling features for Illumina reads were tested: Grinder, DWGSIM, and pIRS. Finally pIRS (Profile-based Illumina paired-end reads simulator) was chosen for generation of the high quality short paired-end

(Illumina) reads, this is because; Running Grinder took a very long processing time of more than a month to simulate Illumina paired-end reads; this could be as a result of Arabidopsis thaliana large genome size, making Grinder unsuitable. In contrary, DWGSIM simulated 150bp paired-end reads quickly (in less than a day), however the different de novo assembly tools utilized were unable to process the simulated reads, because a huge amount of temporary storage was required, leading to the termination of the assembly process before completion.

The Illumina reads are generated at 100X coverage, creating approximately 119 million reads of 100bp read length. PBSIM (PacBio reads simulator) is the only available simulator for PacBio reads and was applied for generation of the long PacBio RS II continuous-long-reads at 20X coverage generating approximately 0.96 million reads with an average read length of 2485bp.

4.1 Simulated PacBio CLR Reads

PacBio CLR (continuous long reads) for *A. thaliana* (TAIR10) were simulated using PBSIM model-based method, at 20X coverage with mean-length of 2500bp. From which 958,686 reads were produced having: maximum read length of 24,922bp; minimum read length of 100bp; average read length of 2,485bp; and N50 of 3,375bp (**Table 4.1**). The reads showed similar read length distribution as characteristic of real PacBio libraries at the time of simulation (Figure **4.1**). However, uncharacteristic of PacBio sequencing libraries is the presence of gaps, which had to be controlled for in downstream assembly, as there were 89 gaps present and this is because of the N's or gaps present in the *Arabidopsis thaliana* (TAIR10) genome sequence.

Table 4.1 Read Length distribution of simulated PacBio reads using PBSIM.

Total reads	Number of Gaps	N50 (bp)	Maximum contig length (bp)	Minimum contig length (bp)	Average contig length (bp)
958,686 reads	27,173 gaps	3,375bp	24,922bp	100 bp	2,485bp

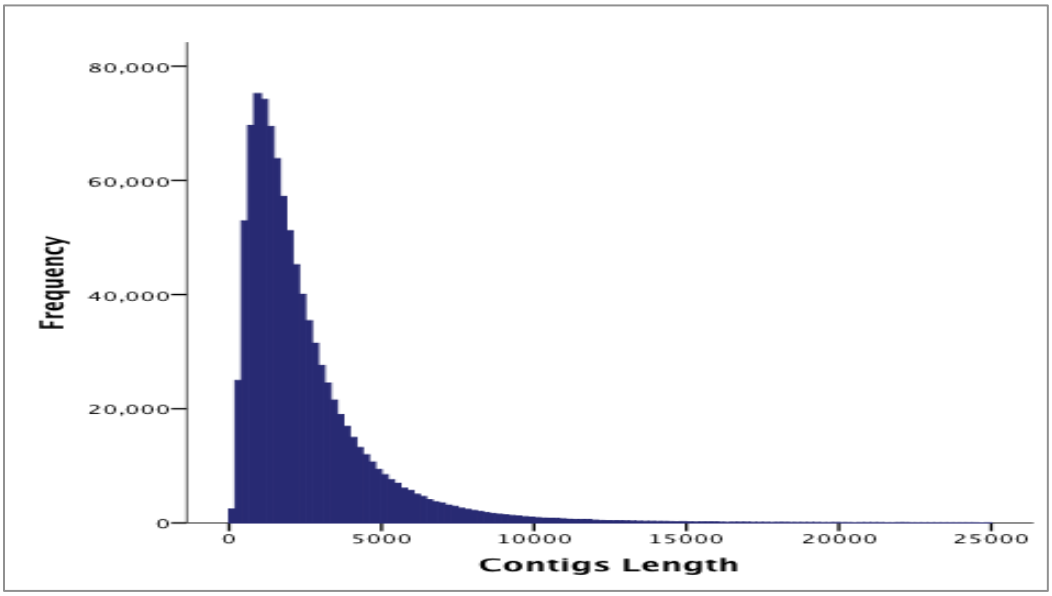


Figure 4.1 PacBio CLR read length distribution

Reference genome mapping was performed using CLC assembler at default settings for validation and assessment of reads accuracy: 87.05% of the reads mapped to the Arabidopsis reference genome, and spanned the entire length of the reference genome with an expected low average genome mapping coverage of 10.94. Over 90 million nucleotide insertions, having maximum insertion length of 36bp and over 34 million deletions, having maximum deletion length of 42bp were observed (Figure **4.2** & **4.3**). The reference alignment had high nucleotide mapping relative to reference sequence error of 8.17% (**Table 4.2**) and the error profile (Figure **4.4**) indicates

majority of the errors observed were contributed from insertions (4.3%) and from deletions or gaps in the reads (1.64%).

Table 4.2 Reference alignment statistics results of PacBio CLR reads.

Simulated dataset	PacBio CLR reads
Total reads	958,686 reads
Mapped reads	834,511 reads
Average genome mapping coverage	10.94
Fraction of reference covered	1.0
Nucleotide mapping error	8.17%

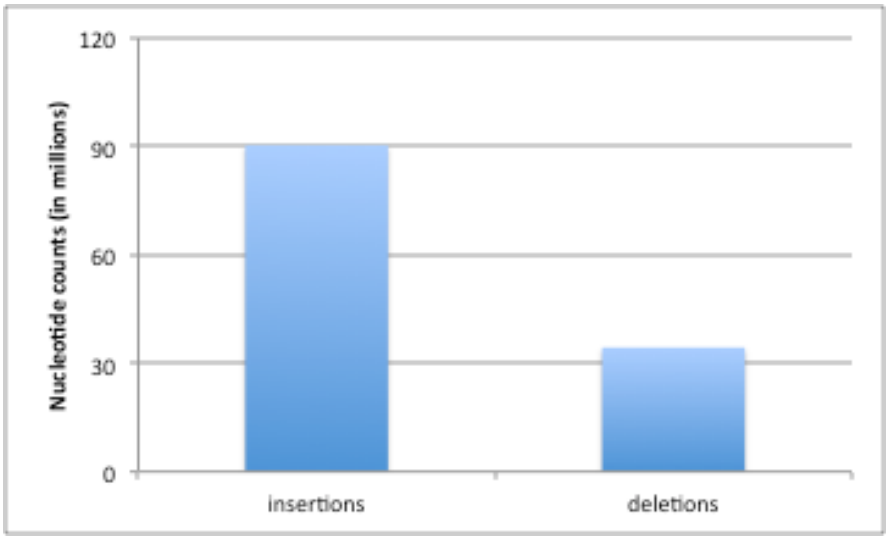


Figure 4.2 Insertion and Deletion nucleotide counts from reference mapping of PacBio reads.

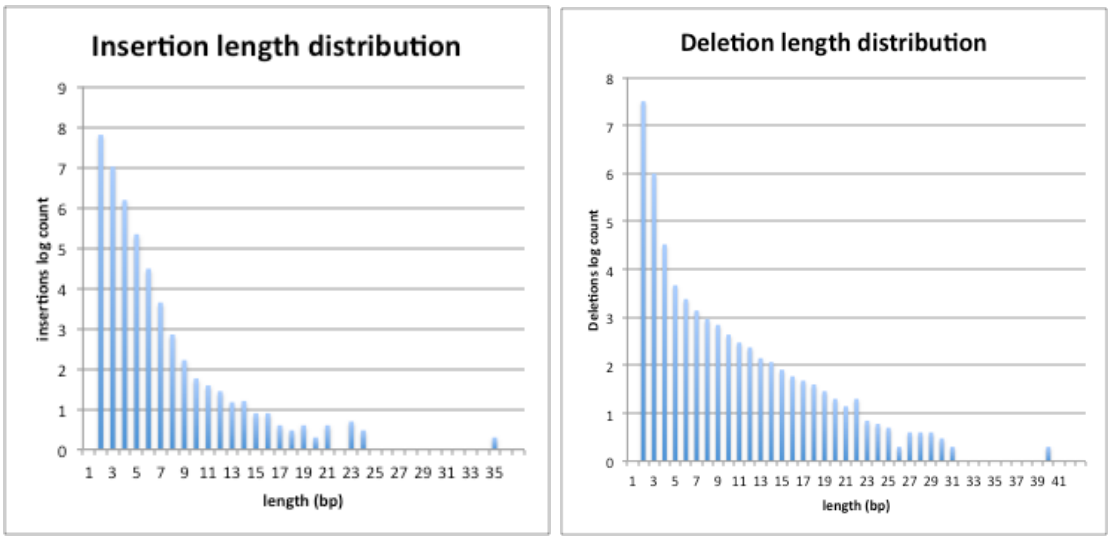


Figure 4.3 Insertion and Deletion length distribution from reference mapping of PacBio reads.

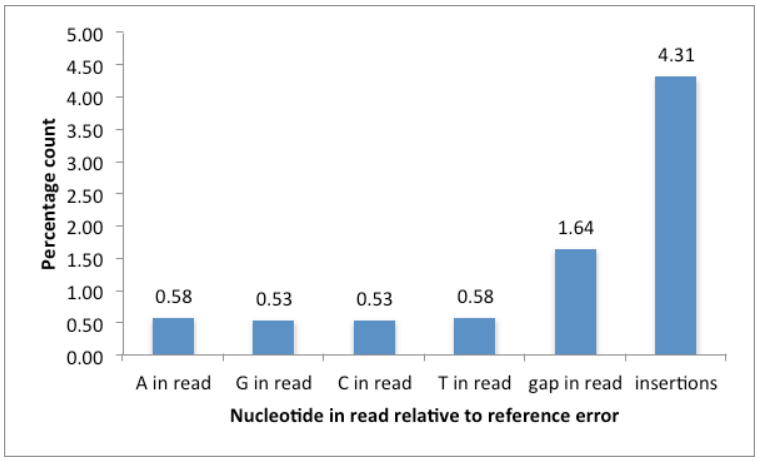


Figure 4.4 Percentage nucleotide in reads relative to reference error profile distribution from reference mapping of PacBio reads.

4.2 Simulated Illumina paired-end Reads

Illumina paired-end reads for *A. thaliana* (TAIR 10) were simulated using pIRS at 100X coverage with read length of 100bp ad mean insert size of 500bp. Over 119

million 100bp reads were simulated. Reference genome mapping was also performed using CLC assembler: 100% of the reads mapped to the reference and spanned the entire length of the reference with expected high average genome coverage of 99.94, and low nucleotide mapping error of 0.34% (**Table 4.3**). The error profile indicates the errors are due to substitutions in reference alignment (Figure **4.5**).

Table 4.3 Reference alignment statistics report of Illumina paired-end reads.

Simulated dataset	Illumina paired-end reads
Total reads	119,146,346 reads
Mapped reads	119,146,343 reads
Average genome mapping coverage	99.94
Percentage of reference covered	100%
Nucleotide mapping error	0.34%

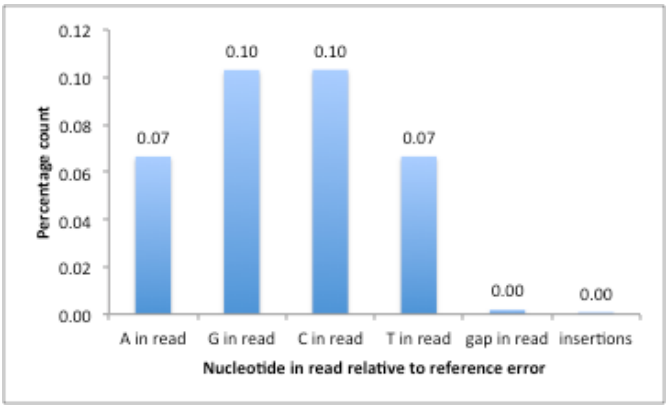


Figure 4.5 Percentage nucleotide in reads relative to reference error profile distribution from reference mapping of Illumina pair-end reads.

Chapter 5

AIM 1: GENERATION AND IMPROVEMENT OF *DE NOVO* CONTIGS WITH SMRT DNA READS

This chapter discusses the first aim for the whole genome assembly process. This approach entails two stages, the first stage involves the application of three different *de novo* assemblers that uses high throughput and high quality short-reads to create *de novo* contigs or scaffolds, and the second stage involves improvement of the *de novo* contigs or scaffolds by gap filling or reduction and joining of contigs with the long PacBio SMRT reads. A pictorial view of this approach is shown below.

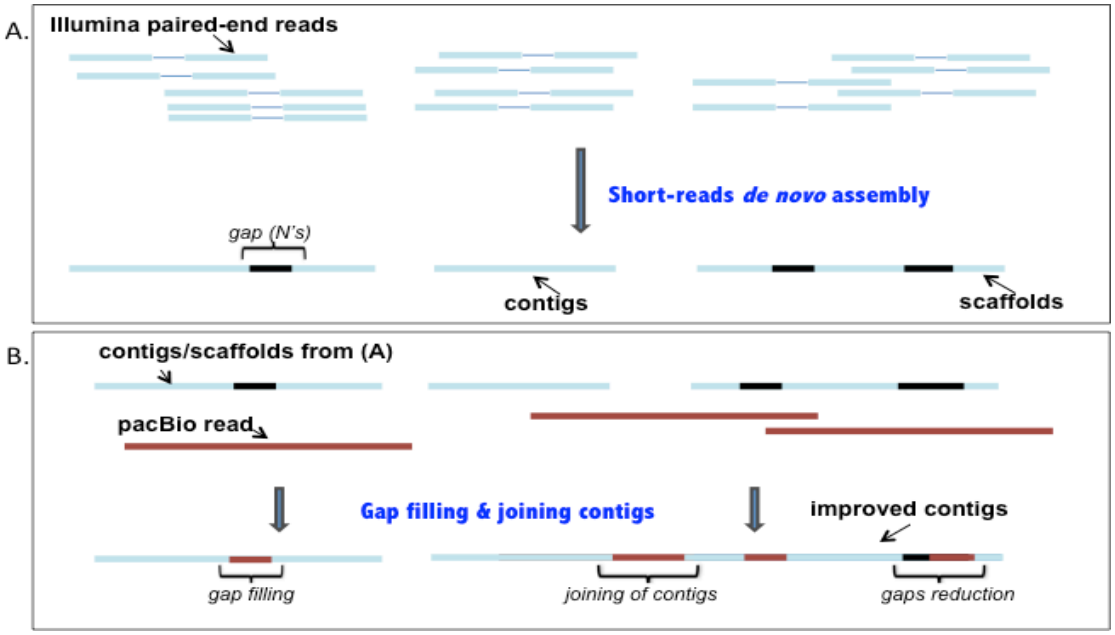


Figure 5.1 Pictorial view of the two stages for the first approach of the whole genome assembly process. (A) de novo assembly of the Illumina paired-end read using the various short-read de novo assembly creating contigs and scaffolds. (B) the scaffolds are improved by filling or reduction of captured gaps and the contigs are joined with the aid of the PacBio reads

5.1 *De novo* Assembly Stage

The first stage implements the use of three different *de Bruijn* graph *de novo* assemblers, which are Velvet (74), SOAPdenovo2 (46) and the CLC Genomics workbench (CLC Bio, version 6.0.1): *de novo* assembly feature (http://www.clcbio.com/) on the Illumina paired-end simulated reads. Velvet, developed in 2008, is an open-source compilation of algorithms that adopts the *de Bruijn* graph data structure to remove errors and resolve repeats and produce non-redundant contigs from the high-throughput, short-read assemblies (74). SOAPdenovo2 (Short Oligonucleotide Analysis Package), an improved version of SOAPdenovo, is specifically designed to create *de novo* draft assembly from Illumina GA short reads. Developed by Luo et al in 2012, it is an open-source unix command line package written in C++ language, it adopts a similar *de Bruijn* graph data structure for read error correction, paired-end read mapping, contig assembly and gap closure (46) to produce highly accurate scaffolds for *de novo* genome assembly especially for eukaryotic genomes. CLC Genomics workbench is a commercial package comprised of various features for the analysis of sequencing data platforms (Sanger, 454, Illumina and SOLiD). The *de novo* assembly feature also implements the *de Bruijn* graph data structure and is optimized for assembling high volumes of data in a very fast and memory-efficient manner.

While the open-source tools (Velvet and SOAPdenovo2) allow the option of manually optimizing the k-mer size, CLC automatically chooses the most optimal k-mer size. Assemblies were performed on all k-mer length (k+2) of both SOAPdenovo2 (version 2.04) and Velvet (version 1.2.10), and the total number of scaffolds, total number of contigs, total number of gaps (i.e. stretches of 25 or more N's within a scaffold), mean contig size, maximum contig size, minimum contig size, N50 contig

size and total bp length (shown in Appendix A) were take into consideration. With higher k-mer length, the N50 contigs size, total bp length, minimum contig size (for Velvet) and maximum contig size was increasing, while the total number of scaffolds (for SOAPdenovo2), total number of contigs and total number of gaps (for SOAPdenovo2) was decreasing. However, a significant decrease in the contigs distribution was observed with the highest k-mer lengths (97 and 99), indicating the inability to form potential branches in the assembly. Thus, the choice of size K is dependent on the dataset, where the shorter k-mers are more sensitive and can generate many more potential branches within the de Bruijn graph, while longer k-mers allow more specificity but lower coverage. The optimal k-mer size, which had the highest maximum contig size, highest N50 contig size, highest total bp length, average number of scaffolds and contigs, and lowest number of gaps, was selected, which are; kmer-87 for SOAPdenovo2 and kmer-81 for Velvet.

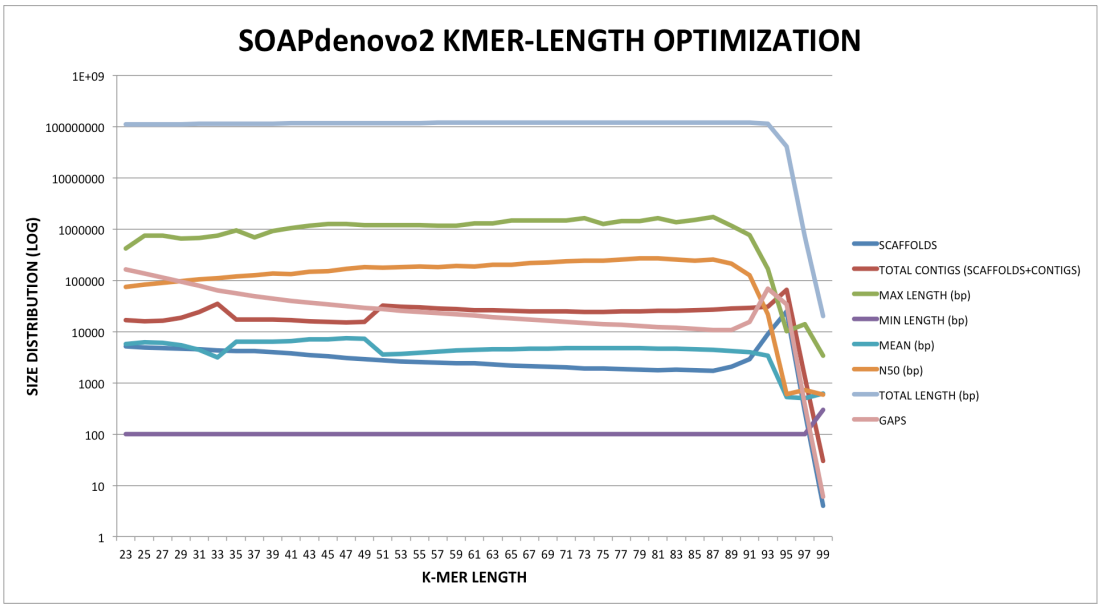


Figure 5.2 Graphical representation of all the k-mer lengths (23-99) for SOAPdenovo2, derived from Appendix A.1. Kmer-87 had the best distribution of scaffolds and contigs generated.

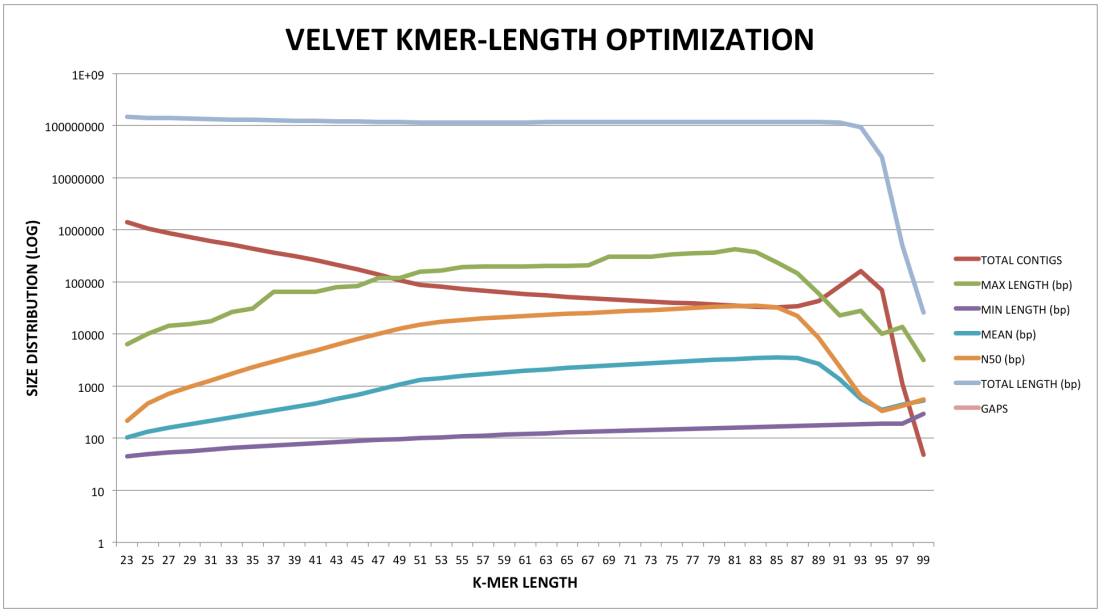


Figure 5.3 Graphical representation of the varied k-mer length (23-99) for Velvet, derived from Appendix A.2. Kmer-81 had the best distribution of scaffolds and contigs generated.

Amongst the three assemblers, SOAPdenovo2 and CLC are able to produce scaffolds and contigs while Velvet cannot produce scaffolds, only contigs. When applied in the first stage; SOAPdenovo2 assembly results had the longest contig or scaffold size of over 1.7 Mbp, longest N50 contig size, of 253,870bp, longest total base-pair length of 119,667,035 bp, as well as the highest number of gaps. Velvet assembly results had the highest amount of contigs produced (35,222 contigs), but the lowest maximum read size, of about 0.4 Mbp, and lowest N50 contig size, of 34,527 bp. While CLC assembly results had the average distribution of the prior assemblers, but with the highest mean contig size as shown below.

Table 5.1 Distribution of the different *de novo* assembler results

***de novo* assemblers**	**SOAPdenovo2**	**CLC**	**Velvet**	**Expected Genome Size**
Total Number of Scaffolds	1,736	5,863	0	5
Total Number of Contigs	25,207	n/a[1]	35,222	94
Total Number of Gaps	10,805	3,665	0	89
N50 Contig Size (bp)	253,870	111,361	34,527	23,453,993
Mean Contig size (bp)	4,391	19,750	3,333	23,792,568
Minimum Contig Size (bp)	100	98	161	18,583,056
Maximum Contig Size (bp)	1,731,725	855,222	429,438	30,263,743
Total Length (bp)	119,667,035	111,670,795	117,448,236	118,962,844

[1]n/a - not applicable. This is due to the unavailability of differentiated results between the contigs and scaffolds from the CLC assembler, thus all the results for the *de novo* assembly were grouped as scaffolds.

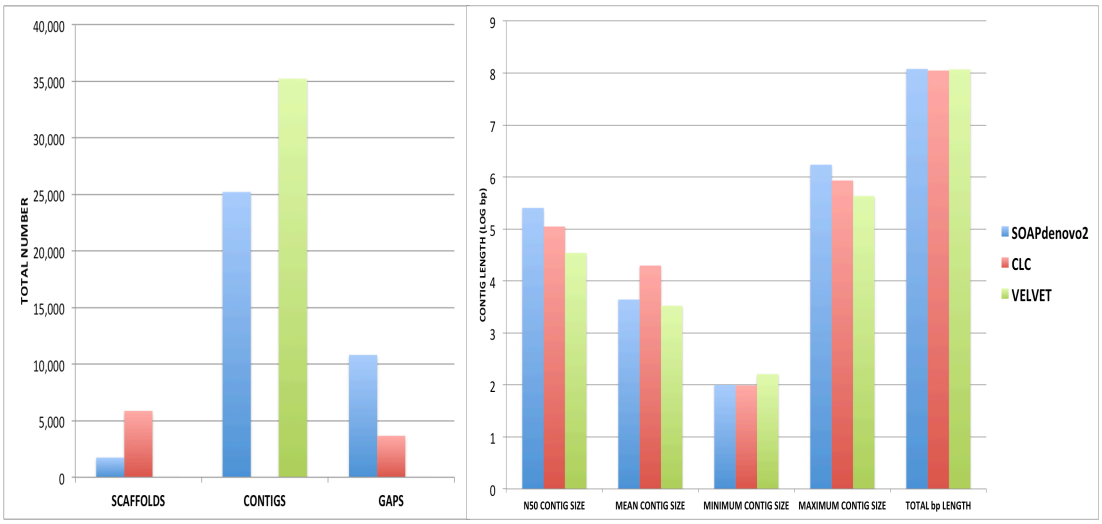


Figure 5.4 Histogram representation of the total contig number and size logarithmic distribution obtained from the three *de novo* assemblers (SOAPdenovo2, Velvet and CLC).

Given the varied distribution of results gotten from this first approach of whole genome assembly, the accuracy of the *de novo* contigs and their error profiles were assessed by reference genome mapping using two genome alignment tools: BLASRsource (13) and BWA-MEM (43), which are able to align very long scaffolds. The BWA-MEM aligner was preferentially chosen for it's fast processing speed. The mapping results were visualized using the CLC Genomics WorkBench and SeqMonk (1). As a result, all the *de novo* scaffolds mapped to the reference genome with minimal duplication and spanned a high portion of the reference (SOAPdenovo2 – 86%, Velvet – 97%, and CLC – 82%) with little to no nucleotide mapping error relative to the reference sequence (**Table 5.2**).

Table 5.2 Reference alignment results of *de novo* contigs from the different assemblers obtained from CLC

***de novo* assemblers**	**SOAPdenovo2**	**CLC**	**Velvet**
Total Number of Mapped Contigs	26,943	5,683	35,222
Total Reference Consensus Length	102,409,974 bp	103,670,588 bp	115,975,398 bp
Percentage of reference covered	86%	87%	97%
Nucleotide relative to reference mapping error	0.65%	0.05%	0%

When comparing individual assemblers against each other, the following trends are noticed from Figure **5.4**:

1. The amount of contigs/scaffolds generated does not guarantee better assembly, with CLC assembly having the fewest number of scaffolds, but the highest N50 contigs size and mean contig size.
2. Velvet contigs had the highest percentage of reference covered with no nucleotide relative to reference mapping error (**Table 5.2**). This can be attributed to the absence of gaps in velvet contigs. While due to the numerous amount of gaps in the SOAPdenovo2 scaffolds, the reference alignment had the lowest reference genome covered percentage, with the highest nucleotide relative to reference mapping error. Thus, the number of gaps present has a negating effect: leading to potential miscalls or errors during reference genome alignment.

5.2 Improvement of *de novo* Contigs/Scaffolds Stage

The second stage involves the improvement of the *de novo* contigs and scaffolds obtained from the first stage. It entails gap reduction, filling and joining of contigs/scaffolds by the implementation of the genome-upgrading tool, PBJelly (version 14.1.14).

PBJelly is a genome-improvement automated pipeline, which function by aligning long sequence reads to draft assemblies in order to fill or improve captured gaps. PBJelly relies on the input draft genome, thereby focusing only the gaps and regions with missing and/or low-quality data (22).

The PBJelly pipeline is comprised of five stages: Setup, Mapping, Support, Assembly, and Output. The Setup stage imports scaffold sequences from the input reference genome and identifies gaps (a stretch of 25 or more N's within a scaffolds), and low-quality regions (regions with consecutive N's shorter than 25bp in length). The Mapping stage maps the long reads to the reference genome using BLASR (Basic Local Alignment and Serial Refinement), and the Support stage identifies reads that addresses the gaps by comparing the aligned and un-aligned base positions within each read. After which, the reads for each gap are assembled to generate a high quality consensus sequence (Figure **5.5**). The PBJelly algorithm has provided significant improvement in filling up the gaps in the draft genome assemblies of some strains of bacteria (*Rhizobium sp., Burkholderia sp. and Pseudomonas sp.*) with PacBio reads.

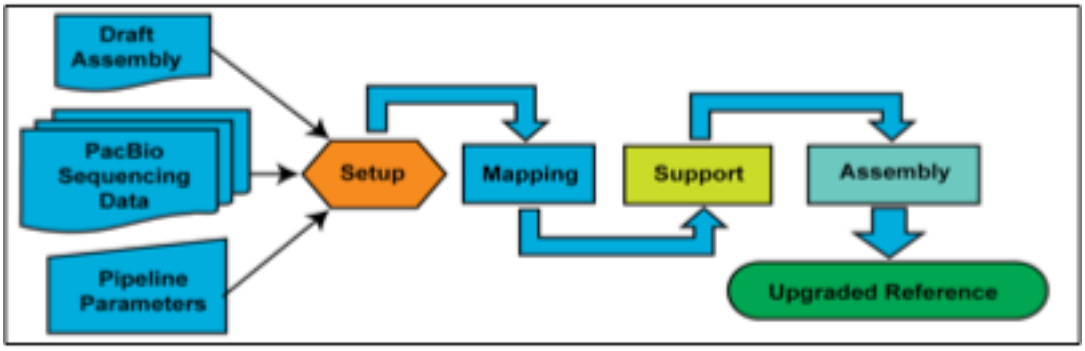


Figure 5.5 PBJelly schematic workflow. Adapted from ref. (22)

Optimization of the PBJelly parameters did not influence the results of the PBJelly process, thus the default parameters of PBJelly (version 14.1.14) was used to improve the *de novo* contigs and scaffolds, obtained in section 5.1, with the PacBio reads (**Table 5.3**).

An overall improvement in *de novo* scaffolds distribution among the three assemblers was observed, as shown in Figure **5.6**. Over two-fold decrease in number of gaps with SOAPdenovo2 and CLC results was observed (83% and 64% respectively), decrease in the total number of contigs (8%, 46% and 29% decrease with SOAPdenovo2, CLC and Velvet contigs/scaffolds respectively), increase in the maximum contig size (3%, 106% and 42% increase respectively), and also 45,119140increase in the N50 contig size (45%, 119%, 140% increase respectively) were also observed. However, with Velvet, 101 gaps were observed, this might be due to the gaps present in the synthetic PacBio reads.

Table 5.3 Distribution of the different *de novo* assembler results improved by PBJelly.

***de novo* assemblers**	**SOAPdenovo2**	**CLC**	**Velvet**	**Expected Genome Size**
Total Number of Sequences	24,799	3,137	24,868	5
Total Number of Gaps	3,932	636	101	89
N50 Contig Size (bp)	368,992	243,985	82,823	23,453,993
Mean Contig size (bp)	4,894	36,175	4,906	23,792,568
Minimum Contig Size (bp)	49	98	15	18,583,056
Maximum Contig Size (bp)	1,792,229	1,758,502	611,458	30,263,743
Total Length (bp)	121,388,312	113,482,734	122,021,636	118,962,844

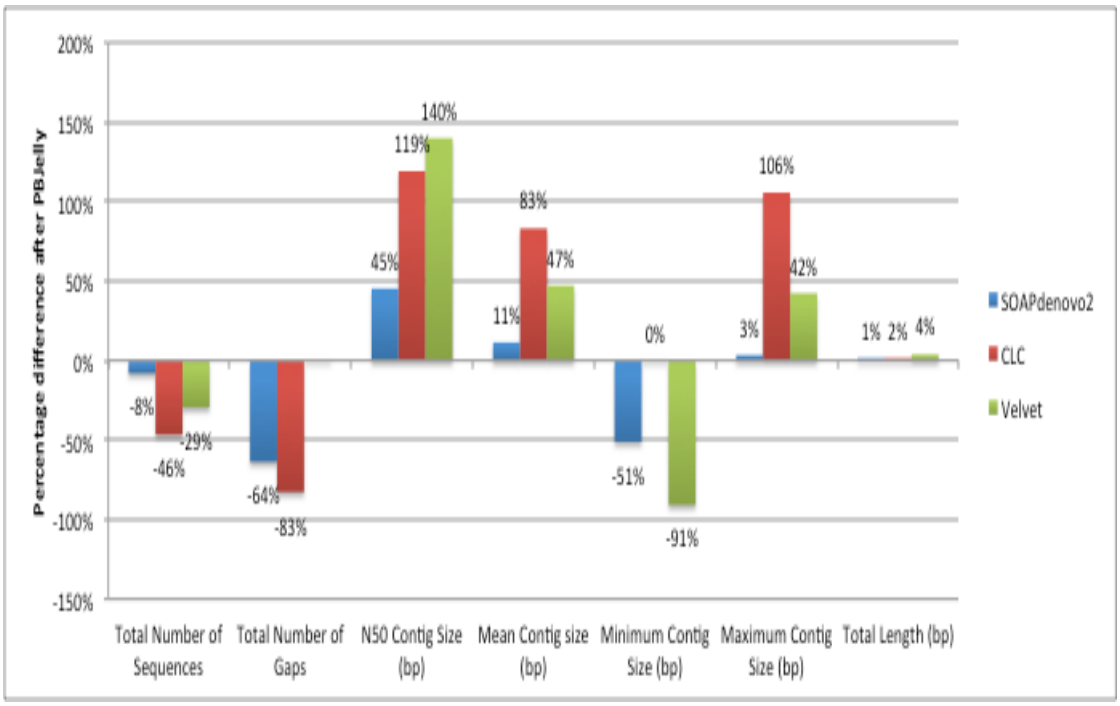


Figure 5.6 Percentage difference after PBJelly improvement. This is a bar graph representation for the percentage difference of *de novo* contigs or scaffolds distributions from the three *de novo* assemblers (SOAPdenovo2, Velvet and CLC) after improvement using PBJelly.

Validation of these "improved" *de novo* scaffolds using BWA-MEM showed that all the reads mapped to the reference genome and spanned a high portion of the reference, slightly less than what was previously observed before the second stage. But with an increased nucleotide mapping error relative to reference percentage of 6.96% for SOAPdenovo2, 0.92% for Velvet, and 0.26% for CLC.

Table 5.4 Reference alignment results of the PBJelly-improved *de novo* contigs obtained from CLC

***de novo* assemblers**	**SOAPdenovo2**	**CLC**	**Velvet**
Total Number of Mapped Contigs	24,799	3,137	24,867
Total Reference Consensus Length	95,303,495 bp	98,310,856 bp	112,706,164 bp
Percentage of reference covered	80%	82%	94%
Nucleotide relative to reference mapping error	6.96%	0.26%	0.92%

The increase in nucleotide mapping error percentage especially with SOAPdenovo2, from 0.65% to 6.96%, may be due to the large amount of gaps initially observed (10,805 gaps) with which PBJelly significantly reduced to 3,932 gaps, and may also be a result of the low quality PacBio reads. These reasons contribute to the inaccuracies observed during PBJelly's mapping stage (using BLASR) of the PacBio reads to the *de novo* scaffolds, thereby allowing a lot of false positives in this second stage.

In conclusion, as shown with the first stage, the *de novo* assemblers are able to efficiency handle the high throughput of short-read synthetic data to produce

significantly long and accurate scaffolds, and on reference genome alignment in the second stage, PBJelly was able to span regions of the reference genome, that the de *novo* scaffolds originally could not, for instance is shown in Appendix B; using SeqMonk and CLC Genomics Workbench alignment viewer to visualize Chr2: 12,250,000-12,550,000 position of the assembly. PBJelly was also able to reduce majority of the gaps present, though with inaccuracies in nucleotide mapping error. We can hypothesize that these errors are due to the high number of gaps present in the *de novo* scaffolds, in the PacBio reads and thereby would be less evident on application of real sequencing datasets that do not have the bias of N's or gaps incorporated from the genome sequence, as discussed in the previous chapter. Asides the gaps observed in the PacBio reads, another attributing factor is their low quality, allowing false mapping, thus relying solely on this approach can not guarantee high quality and long genomic scaffolds representative for large eukaryotic genomes.

Overall, the first approach assures efficiency in creating long and accurate enough contigs or scaffolds for this pipeline.

Chapter 6

AIM 2: ERROR CORRECTION OF SMRT DNA READS

This chapter discusses the second strategy implemented for the whole genome assembly process. This strategy involves the error correction of the PacBio CLR reads with the high quality short-reads, a pictorial view of this approach is shown below.

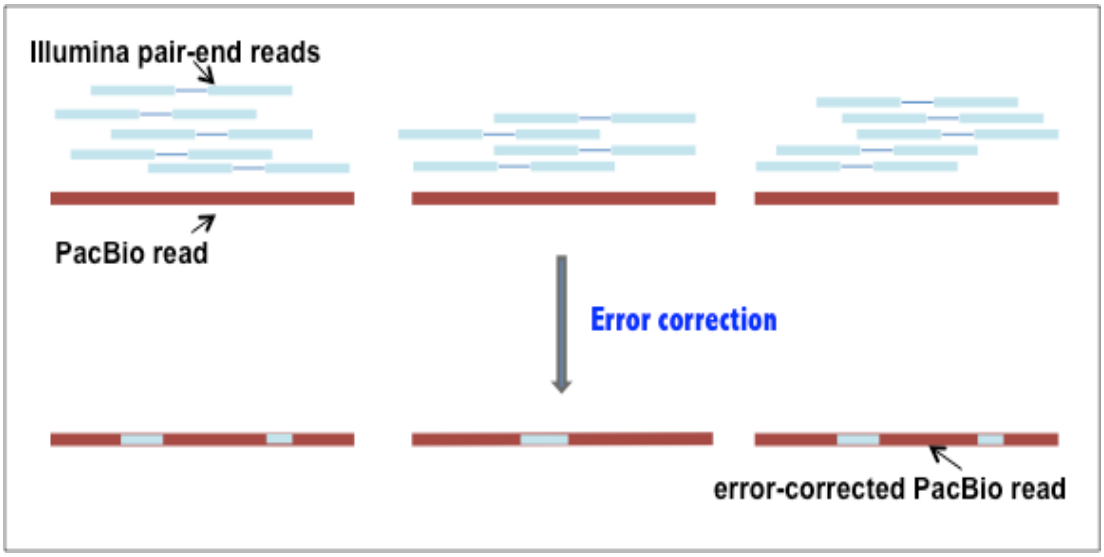


Figure 6.1 Pictorial view of the second approach of the whole genome assembly process. Each PacBio read is mapped to Illumina pair-end reads and ambiguities and low-quality regions with the read are removed or modified based on the information obtained from the short-reads.

Two open source error-correction tools were applied for this approach. The first error-correction tool implements a de Bruijn graph mapping algorithm, which is LSC (73), while the second tool implements an Overlap-Layout-Consensus graph mapping

algorithm, which is the Celera Assembler PBcR pipeline: PacBioToCA utility tool (38). LSC, developed in 2012, is a python program that performs short-reads to long-reads alignment using short-read reference genome aligners such as Bowtie2, RAzerS3, Novoalign or BWA with the aid of homopolymer compression (HC) transformation and uses the information from the alignment to modify the long-reads, thereby creating "corrected" long-reads with a much lower error rate than that of the original long-reads (73). The PacBioToCA tool is the pioneer tool for error-correction of PacBio long-reads. It is developed as a module in the Celera Assembler PBcR pipeline and functions by modification of the long-reads based on the information obtained from the long-read to short-read alignment.

In this study, Bowtie2 was used as the default aligner on LSC (version 0.3.1), Bowtie2 is a de Bruijn graph reference genome aligner, known for its speed, sensitivity and high accuracy on short-reads high throughput data (41). While for PacBioToCA, the latest version on December 2013 (version wgs-8.1) was used. **Table 6.** shows the original PacBio reads and the error-correction results using the different error-correction tools.

Table 6.1 Distribution of results obtained from application of the different error-correction tools, [B] LSC and [D] PacBioToCA. [A] The original-PacBio reads distribution (from **Table 4.1**) provides a comparison for the performance of both error-corrected tools. [C] Based on the error-correction percentage provided by LSC, the corrected reads with greater than 70% error-correction percentage was extracted for downstream analysis.

DATASET	ORIGINAL-PACBIO READS [A]	LSC-CORRECTED READS [B]	LSC-CORRECTED READS (>70%)[C]	PACBIOTOCA-CORRECTED READS [D]
Total Number of Reads	958,686	951,841	683,817	1,685,503
Total Number of Gaps	27,173	1,692	510	0
N50 Contig Size (bp)	3,375	3,257	3,045	1,998
Mean Contig size (bp)	2,486	2,264	2,066	1,324
Minimum Contig Size (bp)	100	63	63	51
Maximum Contig Size (bp)	24,922	24,323,	24,323	22,640
Total Length (bp)	2,382,925,599	2,155,327,471	1,413,176,977	2,232,739,821

LSC output showed the error-corrected reads maintained a similar read-length distribution compared to the original PacBio reads, having 951,841 corrected-reads with a comparable N50 of 3,257bp. Notably, the number of captured gaps had an over twenty-fold reduction to 1,692 gaps when compared to the original of 27,173 gaps. An interesting feature with LSC is the provision of an error-correction percentage for the short-read length coverage for each long read; majority of the reads (approximately 72%) had error-correction percentages greater than 70% (Figure **6.2**), thus were extracted as the final output in order to reduce the amount of sequencing errors in the dataset.

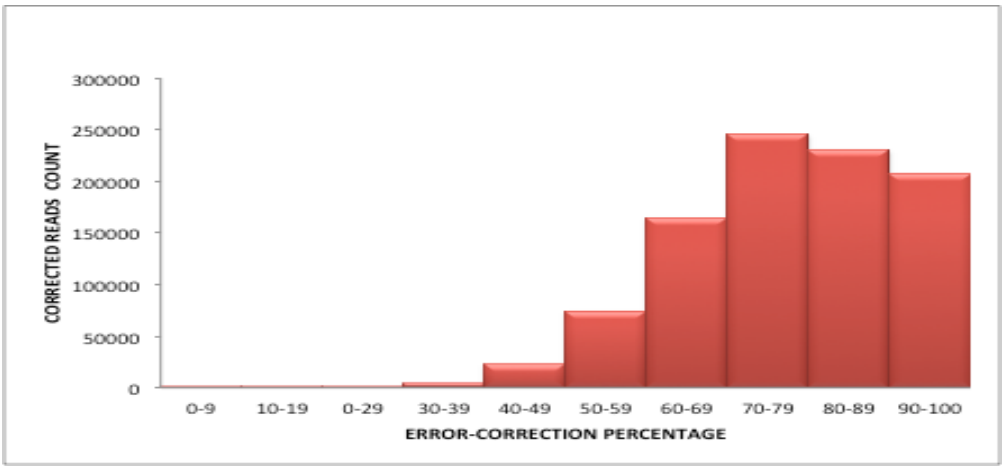


Figure 6.2 Histogram distribution of the error-correction percentage for the LSC-corrected PacBio reads. Majority of the corrected-PacBio reads have greater than 70% error-correction percentage (approximately 72%).

PacBioToCA performs a different method of error-correcting reads; the reads are split at regions of low short-read overlap coverage or at assembly errors, leading to an output of numerous corrected reads compared to the original PacBio reads, and thereby doesn't provide an error-correction percentage. PacBioToCA does not have a similar read-length distribution, as LSC, resulting in 1,685,503 corrected reads, and a smaller N50 of 1,998bp (**Table 6.**D). This is a limiting feature for PacBioToCA; with the objective of this project to create long *de novo* genomic scaffolds of large eukaryotic genomes, splitting the original PacBio reads is a setback for this process. However, a positive observation was the absence of gaps in the PacBioToCA-corrected reads, indicating all the initially captured gaps were more likely removed or error-corrected.

To evaluate the error-correction tools, the resulting corrected-PacBio reads were aligned to the reference using the CLC assembler, only corrected reads with more than

200bp read length were mapped to the reference (**Table 6.2**). Reference mapping results showed majority of the corrected-reads spanned the entire length of the reference genome (100%), and significantly reduced nucleotide relative to reference mapping error (LSC = 3.69%, PacBioToCA = 0.28%) compared to the original PacBio reads. The greatly reduced nucleotide mapping errors and the 100% reference genome covered from both error-corrected reads shows the capability of both error-correction tools to accurately modify the errors known to PacBio reads while maintaining coverage.

Table 6.2 Reference alignment results of "corrected" PacBio reads obtained from CLC.

ERROR-CORRECTION TOOLS	LSC	PACBIOTOCA
Total Number of Mapped Contigs	657,564	1,558,033
Total Reference Consensus Length	118,958,529	118,953,661
Percentage of reference covered	100%	100%
Nucleotide relative to reference mapping error	3.69%	0.28%

The insertion and deletion error profiles were also evaluated using the CLC genome alignment tool; the reference alignment showed significantly reduced number of insertions and deletions compared to those initially observed with the original PacBio reads; with the LSC scaffolds, a 61% and 66% reduction were observed in deletions and insertions respectively, and with the PacBioToCA contigs, a 98% reduction was observed with both insertions and deletions (Figure **6.3**). LSC corrected-reads had over 30 million nucleotide insertions, with maximum insertion length of

111bp, and over 13 million nucleotide deletions, with maximum deletion length of 97bp (Figure **6.4**), while for the PacBioToCA corrected-reads, over 1.7 million nucleotide insertions, with maximum insertion length of 111bp and over 0.6 million nucleotide deletions, with maximum length of 127bp (Figure **6.5**). These reductions also buttress the efficiency of both error-correction tools to correct the characteristic errors of PacBio reads.

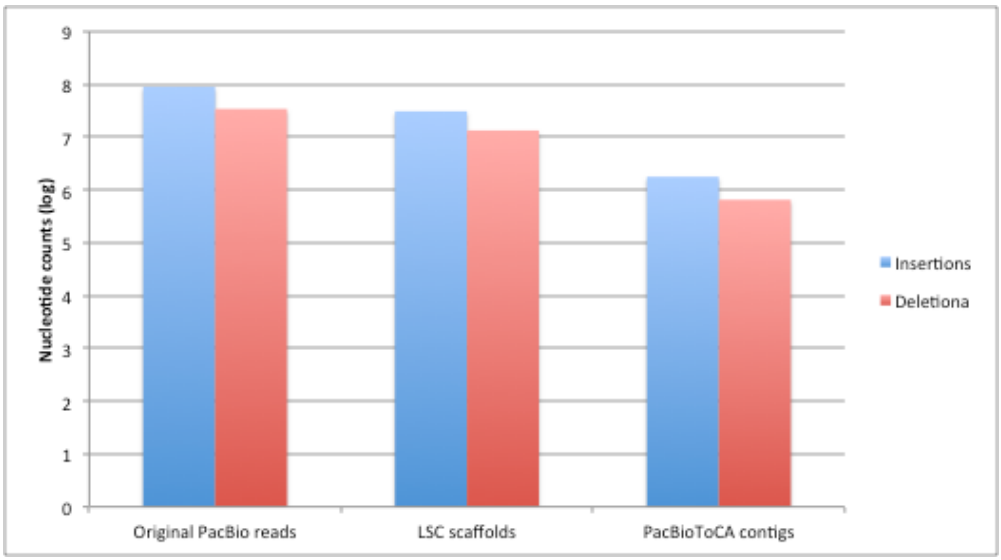


Figure 6.3 Insertion and Deletion nucleotide counts from reference mapping of Original PacBio reads, LSC scaffolds and PacBio contigs using CLC.

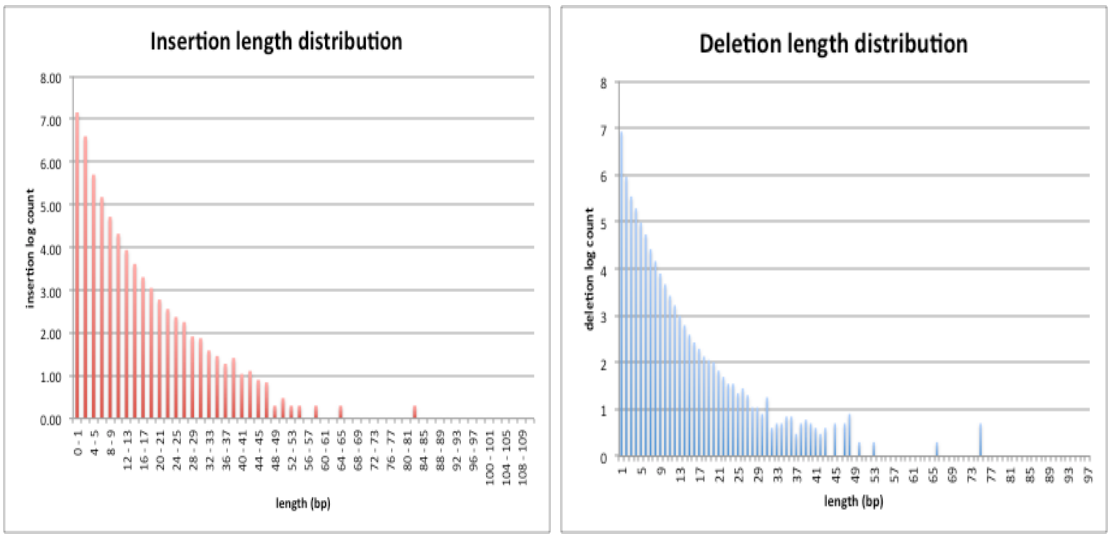


Figure 6.4 Insertion length distribution and Deletion length distribution of LSC corrected-reads from reference mapping using CLC.

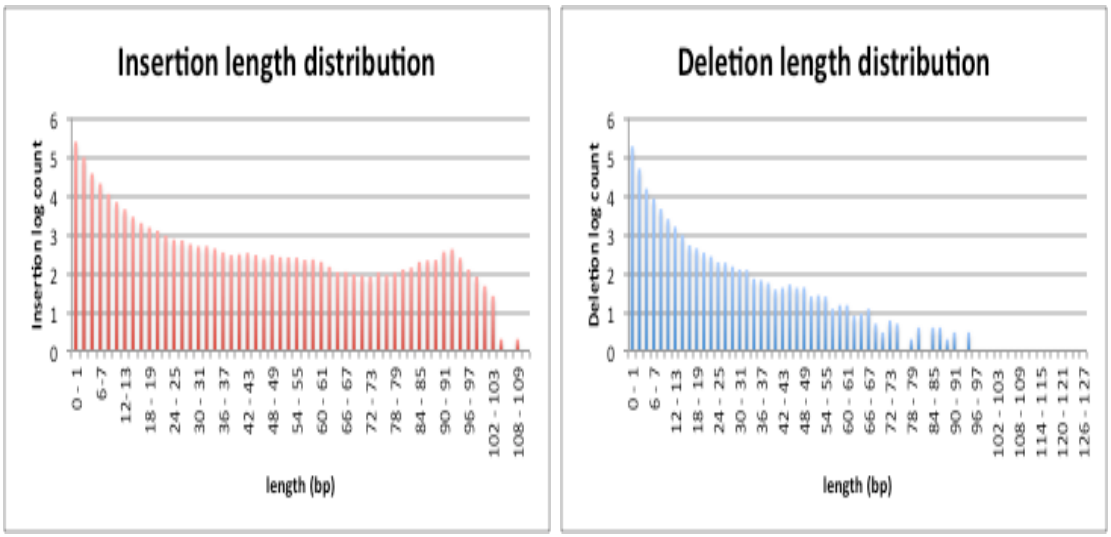


Figure 6.5 Insertion and Deletion length distribution of PacBioToCA corrected-reads from reference mapping using CLC.

Pacific Biosciences SMRT DNA sequencing reads offers the avenue for creating more accurate *de novo* assemblies of large eukaryotic genomes. PacBio read data provides much longer but noisy reads comprised of randomly distributed errors, majority of which are insertions and deletions (10). This second strategy process involves implementation of two methods (LSC and PacBioToCA) for modification of the PacBio long reads using the information obtained from the high-quality NGS short reads, in all cases to reduce sequencing errors or bias and improve the overall quality of the reads.

Comparison of both error-correction tools based on computation speed and time; the LSC took about 60 days of CPU time to complete its process, while PacBioToCA took about 9 days, thus PacBioToCA is considered to be computationally faster, however a reason for the long processing time may be the implementation of a short-read alignment process, with the use of a high-throughput short-read aligner – Bowtie2, before error-correction can commence. As earlier noted, PacBioToCA generates an output of numerous split-reads; defeating the overall aim of generating very long genomic scaffolds, while LSC maintains a similar but slightly lesser read distribution and coverage compared to the original reads. However, based on sensitivity and correctness of the reads; PacBioToCA corrected reads had a much lower nucleotide reference mapping error, as well as insertions and deletions than compared to LSC. PacBioToCA is therefore more sensitive than LSC in masking and removing captured errors based on short-read mapping.

Chapter 7

AIM 3: WHOLE GENOME ASSEMBLY PIPELINE

The final strategy for the whole genome assembly pipeline is discussed in this chapter. It involves scaffolding the improved-*de novo* contigs and corrected PacBio reads obtained from previous chapters to create long and accurate genomic scaffolds, a pictorial view of this approach is shown below.

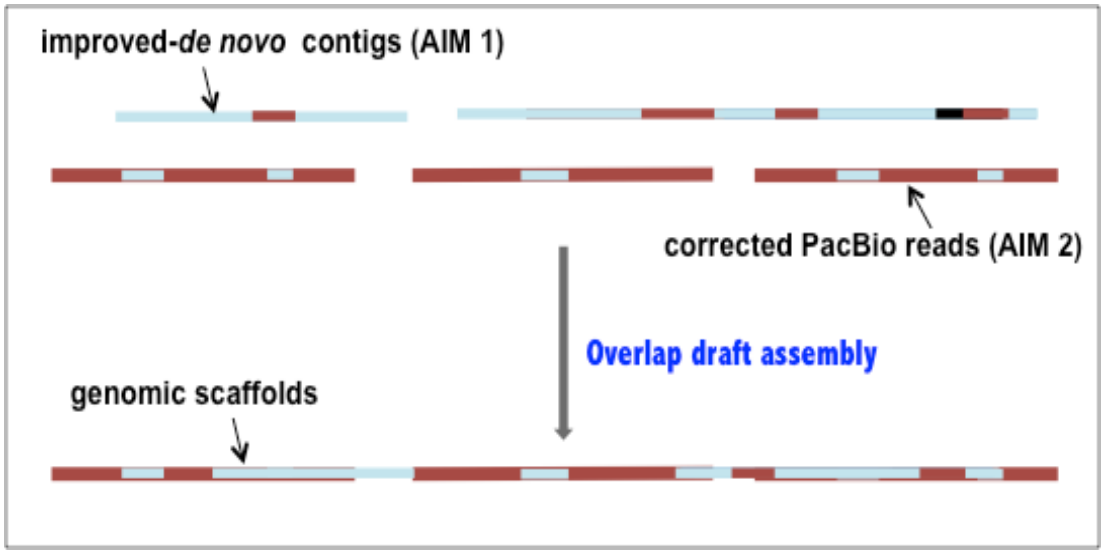


Figure 7.1 Pictorial view of the final strategy of the whole genome assembly process. Both error-corrected PacBio reads and improved-*de novo* contigs are assembled using the WGS assembler, CELERA.

The whole-genome shotgun assembly stage incorporates the overlap-layout-consensus approach to genome assembly with the aid of the whole-genome shotgun

(WGS) assembler, known as Celera Assembler. Celera Assembler is an open-source overlap-layout-consensus based *de novo* WGS DNA sequence assembler that generates long sequences of genomic DNA from whole-genome sequencing data (54). Developed by Celera Genomics from 1999, the assembler has provided the first whole-genome sequence of a multi-cellular organism, Drosophila (54), and human diploid genome sequence (42).

For this study, the Celera assembler (version wgs-8.1) was used with changes to the default gatekeeper settings of no overlap based trimming (OBT), ovlErrorRate=0.1, cnsErrorRate=0.1, and utgErrorRate=0.06. Three different approaches were implemented; the first approach applied only the improved-*de novo* contigs, the second approach applied only the error-corrected PacBio reads and the third approach applied both the improved-*de novo* contigs and error-corrected PacBio reads (Figure **7.2**).

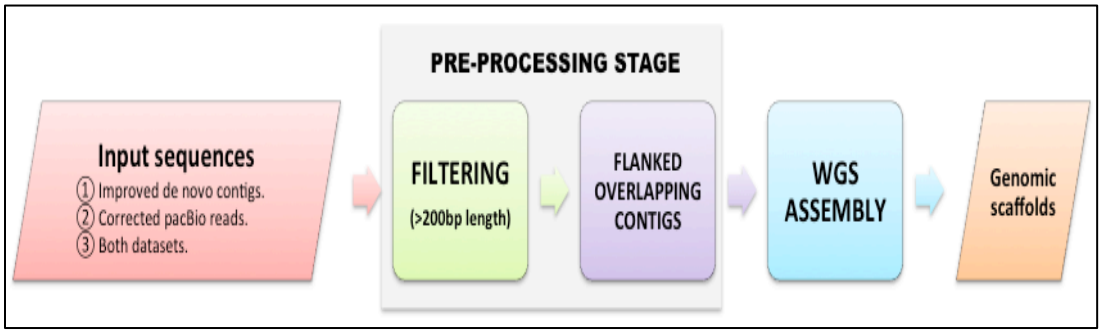


Figure 7.2 Whole-genome shotgun assembly process. Each of the three approaches are executed by; firstly, the pre-processing stage, which involves filtering out sequences of less than 200bp length, then reads with greater than 60,000bp length are split and flanked to form overlapping contigs, with the aid of a custom Perl script. After which, the processed sequences are assembled with the CELERA assembler to form genomic scaffolds and contigs.

A pre-processing stage was incorporated before the WGS assembly; the first step entails filtering out reads with less than 200bp read length in order to minimize redundancies in the assembly. An important observation with the Celera assembler is the maximum input read length limit of 65,535 bp, and having majority of the improved-*de novo* contigs above that limit, these contigs were split to overlapping (or "flanked") contigs of 60,000bp with a large overlap of 5,000bp between the split reads in order to prevent false misalignment of overlapping contigs during assembly, this is the second step in the pre-processing stage.

From the filtering step of the pre-processing stage, reductions were observed across all the improved-*de novo* contigs, especially with the improved-SOAPdenovo2 contigs whereby 85% of the contigs were below 200bp and thus removed, while the improved-CLC contigs and improved-Velvet contigs had a 4% and 27% reduction respectively (Figure **7.3**). **Table 7.1** shows the results after the pre-processing stage for the improved-*de novo* contigs.

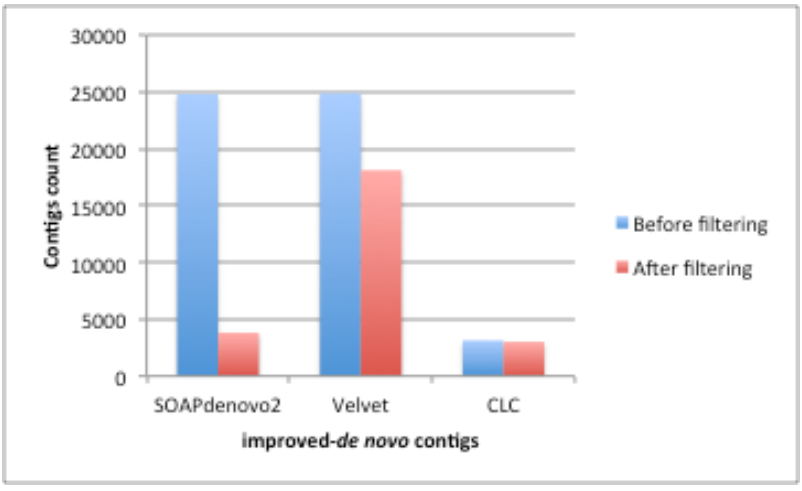


Figure 7.3 Bar graph representation of the distribution of the improved-*de* novo contigs before and after the first step (filtering step) of the pre-processing stage for the Whole-genome shotgun assembly process.

Table 7.1 Distribution of the different improved-*de novo* contigs from the pre-processing stage.

De novo assemblers	SOAPdenovo2	CLC	Velvet
Total Number of Contigs	5,607	4,505	19,148
Total Number of Gaps	4,003	640	101
N50 Contig Size (bp)*	60,000	60,000	60,000
Mean Contig size (bp)	22,749	26,835	6,583
Minimum Contig Size (bp)*	200	200	200
Maximum Contig Size (bp)*	60,000	60,000	60,000
Total Length (bp)	127,555,781	120,892,384	126,066,501

7.1 WGS – Improved *de novo* Contigs

This is the first approach of the WGS assembly process, here the flanked contigs obtained from the improved-*de novo* contigs of the different short-reads de novo assemblers (SOAPdenovo2, Velvet and CLC) were assembled through the Celera assembler, and shown below is distribution of genomic scaffolds and scaffolds reference mapping results using BWA-MEM (**Table 7.2**).

* Similar size distributions are observed across all improved-*de novo* contigs – this is due to the pre-processing stage of the whole genome shotgun assembly process.

Table 7.2 Distribution and percentage improvement of the results obtained from the WGS assembly of only the improved-*de novo* contigs and the reference mapping results using BWA-MEM and CLC.

De novo assemblers	SOAPdenovo2	CLC	Velvet
Total Number of Scaffolds	151 (-97%)	121 (-97%)	181 (-99%)
Total Number of Singletons	2,638	1,654	12,432
Total Number of Gaps	503 (-87%)	0 (-100%)	0 (-100%)
N50 Scaffold Size (bp)	667,905 (+1,013%)	479,056 (+698%)	231,108 (+285%)
Minimum Scaffold Size (bp)	201,223 (+100511%)	215,643 (+107721%)	106,562 (+53181%)
Maximum Scaffold Size (bp)	1,817,462 (+2929%)	1,758,502 (+2831%)	768,985 (+1181%)
Total Length (bp)	83,899,793 (-34%)	55,598,162 (-54%)	39,262,004 (-69%)
Percentage of reference genome covered	54%	39%	31%
Total Reference Consensus Length	64,216,542	46,835,459	36,987,362
Nucleotide relative to reference mapping error	10.58%	0.26%	0.85%

The WGS assembly, of using only the improved-*de novo* contigs obtained from the three different *de novo* assemblers; SOAPdenovo2, CLC and Velvet, generated very few scaffolds; 151, 121 and 181 respectively. But the scaffolds length distributions were on average very long with equally high N50s across the assemblers. An interesting result from this WGS assembly is the absence of gaps in the scaffolds, especially with CLC- and Velvet-WGS assembled scaffolds, except with the SOAPdenono2-WGS assembled scaffolds that had 503 gaps. This showcases the

stringency of the Celera assembler in the assembly of ambiguous nucleotides (such as gaps or N's).

Reference mapping of the assembled scaffolds using BWA-MEM showed all the scaffolds mapped to *A. thaliana* reference and given the small amount of these scaffolds, they were able to span, at most, 50% of the genome (SOAPdenovo2 = 54%, CLC = 39%, Velvet = 31%), with approximately similar nucleotide mapping error relative to reference genome for CLC- (0.26%) and Velvet-WGS assembled scaffolds (0.85%) in comparison to before WGS assembly. However for SOAPdenovo2, the nucleotide mapping error was greatly increased to 10.58% in comparison to before WGS assembly of 6.96% (in **Table 5.4**). A plausible reason for this high error rate may be due to the PBJelly-improvement stage in the first strategy of our genomic assembly pipeline; the original PacBio reads had a high error rate and, as earlier discussed, the large amount of gaps in the de novo contigs and PacBio reads contribute to the inaccuracies observed in the reference alignment of these genomic scaffolds.

To circumvent this, the second stage of the first strategy of the genome assembly pipeline (in Section 5.2) was reprocessed using the PacBioToCA error-corrected PacBio reads (in Appendix C), which were classified as the "PBEC-improved-*de novo* contigs", and were assembled through the WGS assembly process. The distribution of the WGS genomic scaffolds obtained and reference genome mapping results using BWA-MEM is shown in **Table 7.3**.

With the use of the PacBioToCA-error corrected PacBio reads, for the improvement of *de novo* contigs/ scaffolds stage (in Section 5.2) to generate the PBEC-improved-*de novo* contigs, the results are shown in Section C.1. The contigs distribution is comparable to the original improved-*de novo* contigs (**Table 5.3**), with

the exception of the number of gaps observed, whereby improved-Velvet contigs had no gaps, while improved-SOAPdenovo2 contigs and improved-CLC contigs had gaps presents; 3,341 and 664 gaps respectively. The pre-processing stage for the WGS assembly process was also performed on these contigs (Section **Error! Reference source not found.**).

Table 7.3 Distribution and percentage improvement of the WGS assembly results of the PBEC-improved *de novo* contigs and reference mapping results using BWA-MEM and CLC.

De novo assemblers	SOAPdenovo2	CLC	Velvet
Total Number of Scaffolds	124 (-98%)	123 (-97%)	183 (-99%)
Total Number of Singletons	1,549	1,880	11,851
Total Number of Gaps	0 (-100%)	0 (-100%)	0 (-100%)
N50 Scaffold Size (bp)	516,948 (+762%)	419,107 (+599%)	221,871 (+270%)
Minimum Scaffold Size (bp)	208,380 (+104090%)	205,469 (+102634%)	93,940 (+46870%)
Maximum Scaffold Size (bp)	1,276,433 (+2027%)	1,256,362 (+1994%)	524,260 (+774%)
Total Length (bp)	57,614,890 (-55%)	48,627,142 (-60%)	37,146,358 (-71%)
Percentage of reference genome covered	44%	34%	29%
Total Reference Consensus Length	52,934,921	40,591,120	34,082,570
Nucleotide relative to reference mapping error	4.28%	2.77%	0.03%

This WGS assembly of the PBEC-improved *de novo* contigs obtained from the three different *de novo* assemblers; SOAPdenovo2, CLC and Velvet, also generated very few scaffolds; 124, 123 and 183 respectively with also very long scaffold length distributions and equally high N50s. A significant improvement compared to the prior approach was the absence of gaps in the scaffolds, even with the SOAPdenono2-WGS assembled scaffolds. Though, the reference mapping results had a slightly reduced coverage (SOAPdenovo2 = 44%, CLC = 34%, Velvet = 29%); the nucleotide mapping error relative to reference for both SOAPdenovo2 and Velvet was significantly reduced even in comparison to before the WGS assembly, except for CLC-WGS assembled scaffolds, which had an increased nucleotide mapping error of 2.77%, which might be attributed to the inability of the Celera assembler to form accurate scaffolds from the PBEC-improved *de novo* scaffolds obtained from CLC, thus a lot of compromises during assembly was made. Despite the inconsistency with the CLC results, incorporation of the PBEC-improved *de novo* contigs addressed the high error rate from using just the improved-*de novo* contigs.

7.2 WGS – Corrected PacBio Reads

This is the second approach of the WGS process; the error-corrected PacBio reads obtained from the different error-correction tools (LSC, PacBioToCA) were assembled, and shown below is the distribution of genomic scaffolds and BWA-MEM reference mapping results (**Table 7.4**).

Table 7.4 Distribution and percentage improvement of the results obtained from the WGS of only the error-corrected PacBio reads and the reference mapping results using BWA-MEM and CLC.

Error-Correction Tools	LSC	PACBIOTOCA
Total Number of Scaffolds	27,852 (-96%)	13,518 (-99%)
Total Number of Singletons	24,244	12,517
Total Number of Gaps	220 (-57%)	0
N50 Scaffold Size (bp)	11,818 (+287%)	57,004 (+2730%%)
Minimum Scaffold Size (bp)	1,016 (+408%)	1,000 (+400%)
Maximum Scaffold Size (bp)	157,940 (+549%)	709,018 (+3032%)
Total Length (bp)	262,696,288 (-81%)	149,958,012 (-93%)
Percentage of reference genome covered	99%	98%
Total Reference Consensus Length	118,932,771	116,476,365
Nucleotide relative to reference mapping error	2.84%	0.95%

The pre-processing stage was not required for this approach because the error-corrected reads were already filtered and the longest read length is well below the Celera assembler's "maximum read length limit" (in Chapter 5). Thus, the WGS assembly from using only the error-corrected PacBio reads obtained from the two different error-correction tools; LSC and PacBioToCA, produced an large number of genomic scaffolds; 27,852 and 13,518 scaffolds respectively. The scaffolds length distributions were very long with equally high N50s especially from the PacBioToCA-corrected-WGS assembled scaffolds, which had a maximum contig size of over 700Kbp and N50 of 57Kbp. There was expectedly no gaps observed with the genomic scaffolds obtained from the PacBioToCA-corrected reads, but with the scaffolds obtained from the LSC-corrected reads, 220 gaps was observed. Reference mapping of

the assembled scaffolds using BWA-MEM showed all the scaffolds mapped to the reference genome and spanned almost the entire length of the genome (LSC = 99%, PacBioToCA = 98%), with significantly low nucleotide mapping relative to reference error.

Using only the error-corrected PacBio reads to create genomic scaffolds seems sufficient, because not only do the scaffolds span the entire length of the reference genome, the nucleotide mapping error was significantly low, however, a major point of consideration is with the amount of genome scaffolds generated, more scaffolds were produced than anticipated when compared to the number of chromosomes for the *Arabidopsis thaliana* genome, thus to ensure more high quality and accurate scaffolding as well as generation of longer scaffolds, a combination of both the improved-*de* novo contigs and corrected-PacBio reads for WGS assembly was explored.

7.3 WGS – Improved-*de novo* Contigs and Corrected PacBio Reads

The third approach of the WGS process can also be called a combinatory approach; this approach utilizes the different sequencing platforms; both second- and third- generation sequencing platforms for the WGS assembly. The combinatory approach takes advantage of the benefits of both sequencing platforms; such as the high quality and high throughout from the second generation sequencing data, and long read length with no amplification bias from third generation sequencing data (5), in order to create possibly highly accurate and adequate genomic scaffolds for the target genome.

For this approach, both derived datasets from the previous chapters, which are the improved-*de novo* contigs (both the improved-*de novo* contigs and the PBEC-

improved *de novo* contigs) and corrected PacBio reads are utilized for the WGS assembly. Firstly, the improved-*de novo* contigs of the different *de novo* assemblers (in **Table 7.1**) and the error-corrected PacBio reads from both error-correction tools (in **Table 6.**) are assembled through the Celera assembler; the assembly results and the BWA-MEM reference mapping results are shown in **Table 7.5**.

Table 7.5 Distribution of results obtained from the WGS assembly of the application of both the improved-*de* novo contigs and the [A] LSC-corrected PacBio reads or [B] PacBioToCA-corrected PacBio reads and their reference mapping results using BWA-MEM and CLC.

[A]	**De novo assemblers + LSC reads**	**SOAPdenovo2**	**CLC**	**Velvet**	**Expected Genome Size**
Total Number of Scaffolds		7,221	27,852	27,852	5
Total Number of Singletons		10,432	24,244	24,244	94
Total Number of Gaps		2,691	220	220	89
N50 Scaffold Size (bp)		169,849	11,818	11,818	23,453,993
Minimum Scaffold Size (bp)		908	1,016	1,016	18,583,056
Maximum Scaffold Size (bp)		1,984,474	157,940	157,940	30,263,743
Total Length (bp)		160,025,117	262,696,288	262,696,288	118,962,844
Percentage of reference genome covered		92%	99%	99%	-
Total Reference Consensus Length		109,519,212	118,932,771	118,932,771	-
Nucleotide relative to reference mapping error		1.13%	2.84%	2.84%	-

[B]	**De novo assemblers + PacBioToCA reads**	**SOAPdenovo2**	**CLC**	**Velvet**	**Expected Genome Size**

Total Number of Scaffolds	9,690	11,721	12,149	5
Total Number of Singletons	11,744	12,458	12,664	94
Total Number of Gaps	1,797	13	76	89
N50 Scaffold Size (bp)	272,277	161,175	93,986	23,453,993
Minimum Scaffold Size (bp)	527	1,000	1,000	18,583,056
Maximum Scaffold Size (bp)	2,448,708	1,763,128	954,293	30,263,743
Total Length (bp)	147,409,906	216,079,445	151,687,436	118,962,844
Percentage of reference genome covered	83%	90%	95%	-
Total Reference Consensus Length	99,099,387	106,981,379	112,913,595	-
Nucleotide relative to reference mapping error	1.53%	1.11%	1.35%	-

Table 7.5[A] shows the results of the hybrid WGS assembly approach using both the LSC error-corrected PacBio reads and the improved-*de* novo contigs. The scaffolds distribution results of this assembly for both the improved-CLC contigs and the improved-Velvet contigs are identical to the WGS assembly results of the LSC error-corrected PacBio reads, as shown in the previous approach (Section 7.2). This suggests that the improved-CLC contigs or the improved-Velvet contigs with the LSC-corrected reads were not compatible to form genomic scaffolds through the Celera assembler; a possible reason for this might be due to the OLC graph model for the WGS assembly, the potential edges in this assembly model between the different datasets can be said to be very weak, to be accepted in the final consensus stage compared to the potential edges between only the LSC-corrected contigs, hence the final results.

However, the hybrid approach of the improved-SOAPdenovo2 contigs and the LSC-corrected reads generated 7,221 genomic scaffolds; the scaffolds generated were averagely very long with a maximum read length of over 1.9Mbp, with a high N50 of about 170Kbp, and with 2,691 gaps, also the scaffolds spanned 92% of the *A. thaliana* genome from reference alignment using BWA-MEM with a 1.13% nucleotide mapping error. The combination of both the LSC-corrected PacBio reads and improved-SOAPdenovo2 contigs for the WGS assembly in comparison to the prior individual approaches; generated an average number of scaffolds, increased the maximum contig size and greatly reduced the amount of gaps compared to the LSC-corrected-WGS assembled scaffolds. From reference alignment, the nucleotide mapping relative to reference error was significantly reduced with a comparably high reference genome covered percentage. This portrays the efficiency of the combinatory approach to generate a succinct amount of scaffolds while maintaining high accuracy and span almost the entire length of the *Arabidopsis* reference.

While applying both the PacBioToCA error-corrected PacBio reads and improved *de novo* contigs (**Table 7.5**[B]), the genomic scaffolds generated were also on average very long – with the longest read size of over 2Mbp and highest N50 of 272,277bp from the improved-SOAPdenovo2 contigs assembly. However, gaps were observed from these assembled scaffolds (with SOAPdenovo2, 1,797 gaps were observed, CLC had 13 gaps, while Velvet had 76 gaps). These gaps might be due to the inability of the WGS assembler to mask all the captured gaps from the improved-*de novo* contigs using the error-corrected reads. From reference mapping, the scaffolds spanned large portions of the *Arabidopsis* genome (SOAPdenovo2 = 83%, CLC = 90% and Velvet = 95%), with low nucleotide mapping relative to reference error. Thus, the

combination of both the PacBioToCA-corrected PacBio read and improved-*de novo* contigs in comparison to their prior individual approaches also generated an average number of scaffolds, greatly increased the maximum contig size but increased the amount of gaps and from reference alignment, the scaffolds mapped a comparably high portion of the reference genome, but the nucleotide mapping error was increased.

The same approach was applied with the PBEC-improved *de novo* contigs of the different *de novo* assemblers (Appendix C) and the error-corrected PacBio reads from both error-correction tools are assembled through the Celera assembler; the assembly results and the BWA-MEM reference mapping results are shown below.

Table 7.6 Distribution of results obtained from the WGS assembly of the application of both the PBEC-improved *de* novo contigs and the [A] LSC-corrected PacBio reads or [B] PacBioToCA-corrected PacBio reads and their reference mapping results using BWA-MEM and CLC.

[A] **De novo assemblers + LSC reads**	**SOAPdenovo2**	**CLC**	**Velvet**	**Expected Genome Size**
Total Number of Scaffolds	11,076	9,174	9,304	5
Total Number of Singletons	12,916	11,263	10,033	94
Total Number of Gaps	216	207	214	89
N50 Scaffold Size (bp)	63,905	74,615	60,000	23,453,993
Minimum Scaffold Size (bp)	1,016	1,057	1,040	18,583,056
Maximum Scaffold Size (bp)	1,082,823	1,046,878	781,966	30,263,743
Total Length (bp)	181,093,831	173,459,719	174,274,716	118,962,844
Percentage of reference genome covered	97%	95%	97%	-
Total Reference Consensus Length	116,106,471	113,550,869	116,211,760	-

Nucleotide relative to reference mapping error	2.13%	1.96%	1.64%	-

[B] **De novo assemblers + PacBioToCA reads**	**SOAPdenovo2**	**CLC**	**Velvet**	**Expected Genome Size**
Total Number of Scaffolds	11,661	12,197	13,174	5
Total Number of Singletons	12,409	12,455	12,820	94
Total Number of Gaps	0	0	0	89
N50 Scaffold Size (bp)	171,275	117,059	69,701	23,453,993
Minimum Scaffold Size (bp)	1,000	1,000	1,000	18,583,056
Maximum Scaffold Size (bp)	1,848,757	1,303,827	794,242	30,263,743
Total Length (bp)	146,809,475	150,222,665	154,486,067	118,962,844
Percentage of reference genome covered	93%	90%	94%	-
Total Reference Consensus Length	111,196,489	106,995,987	111,670,386	-
Nucleotide relative to reference mapping error	1.19%	1.37%	1.02%	-

Table 7.6[A] shows the results of the hybrid WGS assembly approach using both the LSC error-corrected PacBio reads and the PBEC-improved *de* novo contigs. The resulting scaffolds from this assembly show that the WGS assembler can form unique consensus sequences from the combination of both the PBEC-improved *de novo* contigs and the LSC-corrected PacBio reads. Significantly long scaffolds were generated, with the longest size at over 1Mbp in length with a small amount of gaps. Based on reference alignment, the scaffolds spanned almost the entire length of the *A. thaliana* genome, with little nucleotide mapping error. In comparison to the previous assembly with the improved-*de novo* contigs (**Table 7.5** [A]), the different PBEC-

improved *de novo* contigs are able to generate unique consensus sequences with the LSC-corrected reads through the WGS assembler. The amount of gaps were significantly reduced, though the sequence length distribution was reduced; having a lower N50, a lower maximum contig size with a larger amount of scaffolds, the scaffolds are able to span a greater percentage of the reference sequence (the PBEC-improved SOAPdenovo2 contigs; from 92% to 97%) with slightly more nucleotide mapping relative to reference error. While, with applying both the PacBioToCA error-corrected PacBio reads and the PBEC-improved *de novo* contigs (**Table 7.6** [B]), the genomic scaffolds were also typically long, similar to the initial assembly of the improved-*de novo* contigs, but no gaps were observed from these scaffolds; thus the WGS assembly is able to mask all the captured gaps from using the PBEC-improved *de novo* contigs. The scaffolds were able to map a similar (with both the PBEC-improved CLC and Velvet contigs) or higher (with the PBEC-improved SOAPdenovo2 contigs) percentage of the reference genome in comparison to the previous assembly with the improved-*de novo* contigs (**Table 7.5** [B]), with reduced nucleotide mapping relative to reference sequence error. Overall, using the PBEC-improved *de* novo contigs, showed a slight improvement compared to the initial WGS assembly; the scaffolds distribution for each assembly was similar to the assembly with the improved-*de novo* contigs (**Table 7.5**), except with the significant reduction or removal of the captured gaps found in the results genomic scaffolds, and also the genomic scaffolds were able to map to a greater percentage of reference covered. Thus, the presence of gaps does have a negative effect to the efficiency of the assembly, and as earlier discussed in Section 4.1 from the initially-captured gaps of the simulated PacBio reads, using real time

sequencing PacBio data may have a similar or significantly better genomic scaffolds comparable to what was shown with using the PBEC-improved *de novo* contigs.

The WGS assembly of the improved-*de novo* contigs (Section 7.1), and the corrected-PacBio reads (Section 7.2) generates either millions of base pairs long but few scaffolds that spans at most 50% of the *Arabidopsis* genome based on reference alignment or shorter thousands of base pairs long but numerous scaffolds with high percentage of reference genome covered respectively. The combination of both datasets through the WGS assembler (Section 7.3), provides a unique avenue to not only create a succinct amount but very long genomic scaffolds; these scaffolds are able to accurately map to at least 90% of the reference genome with very low nucleotide mapping error of approximately <2%. Therefore making this WGS combination approach a suitable mention for creating accurate and long genomic scaffolds of large eukaryotic genomes, such as the Coffee Leaf Rust fungus.

Chapter 8

CONCLUSION

The purpose of this thesis was to implement the use of various assembly and correction tools for the whole genome assembly of large eukaryotic genomes using NGS reads from both second generation (Illumina HiSeq) and third generation (PacBio Biosciences) sequencing platforms. This thesis discusses the whole genome assembly pipeline developed and the tools applied for the different strategies, providing important insights into their performances with the aid of simulated data of the *Arabidopsis thaliana* genome.

In the first strategy, I implemented the use of three different short-read *de novo* assemblers, SOAPdenovo2, CLC and Velvet to create *de novo* contigs and scaffolds. The resulting scaffolds were improved by reducing captured gaps (> 25 N's in a single nucleotide sequence stretch) and stitching them together using the PacBio reads with the aid of an improvement tool known as PBJelly, therefore creating longer scaffolds. This strategy showcases the efficiency of the de Bruijn graph in handling high throughput of short-read data to produce significantly long and accurate scaffolds, however in comparison to the original genome sequence, numerous scaffolds (>3,000 - <25,000 scaffolds) were generated with a lot of inaccuracies in reference mapping especially with the SOAdenovo2 scaffolds. Therefore this approach is not sufficient to create high quality and long scaffolds representative of large eukaryotic genomes.

The second strategy takes a different approach of using two different error-correction tools, PacBioToCA and LSC, to improve or correct the random nucleotide, insertion and deletion errors that are characteristic of PacBio datasets. PacBioToCA showed higher sensitivity and speed in accurately correcting the PacBio reads than LSC but at the compromise of read length.

The third strategy involves the utilization of a whole-genome shotgun assembler (Celera assembler) to create long genomic scaffolds using the improved *de novo* scaffolds from the first strategy and the error-corrected PacBio reads from the second strategy. Application of the resulting datasets from the prior aims separately yielded very few (< 190 scaffolds) and long scaffolds (maximum scaffold length of 1.8mb) across the different improved-*de novo* scaffolds, but spanned less than 50% of the *A. thaliana* genome, while with the error-corrected PacBio reads yielded a large amount (over 13kb) and shorter scaffolds (maximum scaffold length of 0.7mb) but spanned the entire length of the reference genome. Combination of both data types through the whole-genome shotgun assembler gave better results, producing an average amount of scaffolds (compared to previous shotgun assemblies) and very long scaffolds (maximum scaffold length of 2.4mb).

Based on this WGS pipeline, the analysis showed the combination of both resulting datasets for the WGS assembly is the best approach for creating very long and highly accurate genomic scaffolds; with the generation of significantly few and very long genomic scaffolds, having high coverage (avg. >90%) and very low nucleotide mapping relative to reference error (avg. <2%), unlike application of individual resulting datasets. Comparison of the different whole-genome assembly results, showed the hybrid assembly of the improved-SOAPdenovo2 contigs and PacBioToCA contigs

produced the best distribution of scaffolds; having no gaps, with high scaffold size, and able to span 93% of the *A. thaliana* genome with little nucleotide error.

However, the large amount of gaps present had a negating effect on the accuracy and sensitivity of the resulting assembly, these gaps were seen to be significantly contributed from the simulated PacBio reads, thus one can assume that with using real PacBio sequencing data may result in better assemblies with less gaps present.

Finally, the overall analysis showed the efficiency of the pipeline in achieving very long genomic scaffolds for large eukaryotic genomes with the aid of different NGS tools, and further work can be done in the development or application of algorithms that can further improve these long scaffolds while maintaining read accuracy.

REFERENCES

1. Babraham bioinformatics - SeqMonk mapped sequence analysis tool [homepage on the Internet]. . Available from: http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/.

2. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* 2012 Jul; 40(12): e94.

3. Arneson PA. Coffee rust. *The Plant Health Instructor.* 2000.

4. Barzuna F. Insight special: Leaf rust. Coffee Division of ED&F MAN. 2013 March;Sect. CBS&A Coffee Business Services & Academy, a Volcafe Initiative.

5. Bashir A, Klammer AA, Robins WP, Chin C, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, Sebra R, Sorenson J, Bullard J, Yen J, Valdovino M, Mollova E, Luong K, Lin S, LaMay B, Joshi A, Rowe L, Frace M, Tarr CL, Turnsek M, Davis BM, Kasarskis A, Mekalanos JJ, Waldor MK, Schadt EE. A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol.* 2012; 30(7): 701-707.

6. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. ARACHNE: A whole-genome shotgun assembler. *Genome Res.* 2002 Jan; 12(1): 177-189.

7. Botero-Rozo DO, Botero-Rozo DO, Giraldo W, Gaitan A, Cristancho M, Riaño-Pachon D, Restrepo S. Data mining of the coffee rust genome. *Nature Precedings.* 2012.

8. Bowden J, Gregory P, Johnson C. Possible wind transport of coffee leaf rust across the atlantic ocean. 1971.

9. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res.* 2008 May; 18(5): 810-820.

10. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics.* 2012 Aug 5; 13: 375-2164-13-375.

11. Carvalho CR, Fernandes RC, Carvalho GMA, Barreto RW, Evans HC. Cryptosexuality and the genetic diversity paradox in coffee rust, *Hemileia vastatrix*. *PLoS ONE.* 2011 11/15; 6(11): e26387.

12. Carvalho GMA, Carvalho CR, Barreto RW, Evans HC. Coffee rust genome measured using flow cytometry: Does size matter? *Plant Pathol.* 2013.

13. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics.* 2012 Sep 19; 13: 238-2105-13-238.

14. Chin C, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Meth.* 2013 print; 10(6): 563-569.

15. Cressey D. Coffee rust regains foothold. *Nature.* 2013; 493(7434): 587.

16. Cristancho M, Giraldo W, Botero D, Tabima J, Ortiz D, Peralta A, GaitÃ¡n Ã, Restrepo S, RiaÃ±o D. Application of genome studies of coffee rust. In: Advances in Computational Biology. Advances in Intelligent Systems and Computing 232 ed. Switzerland 2014: Springer International Publishing; 2014. p. 133-139.

17. Deshpande V, Fung EK, Pham S, Bafna V. Cerulean: A hybrid assembly using high throughput short and long reads. In: Darling A, Stoye J, editors. Springer Berlin Heidelberg; 2013. p. 349-363.

18. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.* 2007; 17(11): 1697 - 1706.

19. Doležel J, Greilhuber J. Nuclear genome size: Are we getting closer? *Cytometry Part A.* 2010; 77A(7): 635 - 642.

20. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. *Science.* 2009; 323(5910): 133 - 138.

21. El-Metwally S, Ouda O, Helmy M. Next-generation sequencing platforms. In: Next Generation Sequencing Technologies and Challenges in Sequence Assembly. Springer New York; 2014. p. 37-44.

22. English AC. Mind the gap: Upgrading genomes with pacific biosciences RS long-read sequencing technology. *ONE Alerts.* 2012 11-21.

23. Field D, Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim—A sequencing simulator for genomics and metagenomics. *PLoS ONE.* 2008; 3(10): e3373.

24. Galagan JE. Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Res.* 2005; 15(12): 1620 - 1631.

25. Gichuru EK, Ithiru JM, Silva MC, Pereira AP, Varzea VMP. Additional physiological races of coffee leaf rust (hemileia vastatrix) identified in kenya. *Tropical Plant Pathology.* 2012; 37(6): 424 - 427.

26. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences.* 2011; 108(4): 1513 -1518.

27. Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics.* 2012 Jun; 13(8): 901-915.

28. Hoffmann M, Muruvanda T, Allard MW, Korlach J, Roberts RJ, Timme R, Payne J, McDermott PF, Evans P, Meng J, Brown EW, Zhao S. Complete genome sequence of a multidrug-resistant salmonella enterica serovar typhimurium var. 5- strain isolated from chicken breast. *Genome Announc.* 2013 Dec 19; 1(6): 10.1128/genomeA.01068-13.

29. Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N, Yue Z, Bai F, Li H, Fan W. pIRS: Profile-based illumina pair-end reads simulator. *Bioinformatics.* 2012; 28(11): 1533 - 1535.

30. Huang W, Li L, Myers JR, Marth GT. ART: A next-generation sequencing read simulator. *Bioinformatics.* 2012 Feb 15; 28(4): 593-594.

31. International Coffee Organization. Monthly coffee market report. http://dev.ico.org/documents/cy2012-13/cmr-0313-e.pdf; 2013 March 13.

32. International Coffee Organization. REPORT ON THE OUTBREAK OF COFFEE LEAF RUST IN CENTRAL AMERICA. http://dev.ico.org/documents/cy2012-13/ed-2157e-report-clr.pdf; 2013 May 13.

33. Janssen PJ, Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. NeSSM: A next-generation sequencing simulator for metagenomics. *PLoS ONE.* 2013; 8(10): e75448.

34. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD. Extending assembly of short DNA sequences to handle error. *Bioinformatics.* 2007; 23(21): 2942 - 2944.

35. Jiao X. A benchmark study on error assessment and quality control of CCS reads derived from the PacBio RS. *Journal of Data Mining in Genomics & Proteomics.* 2013; 04(03).

36. Kilambo DL, Reuben SOWM, Mamiro D. Races of hemileia vastatrix and variation in pathogenicity of colletotrichum kahawae isolates to compact coffee genotypes in tanzania. *Journal of Plant Studies.* 2013; 2(2).

37. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* 2013 Sep 13; 14(9): R101.

38. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol.* 2012; 30(7): 693 - 700.

39. Kubota L. Some insights on coffee leaf rust (hemileia vastatrix). The Specialty Coffee Chronicle. 2013 Feb 15.

40. Kullman B, Tamm H, Kullman K. Fungal genome size database. http://www.zbi.ee/fungal-genomesize; 2005.

41. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods.* 2012 Mar 4; 9(4): 357-359.

42. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AWC, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers Y, Frazier ME, Scherer SW, Strausberg RL,

Venter JC. The diploid genome sequence of an individual human. *PLoS Biology.* 2007; 5(10): e254.

43. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv org*. 2013 26 May; arXiv:1303.3997v2 [q-bio.GN]. Available from: http://arxiv.org/abs/1303.3997.

44. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. *J Biomed Biotechnol.* 2012: 251364.

45. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol.* 2012 May; 30(5): 434-439.

46. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience.* 2012 Dec 27; 1(1): 18-217X-1-18.

47. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature.* 2011 Feb 10; 470(7333): 198-203.

48. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 2008 Mar; 24(3): 133-141.

49. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences.* 1977 February 01; 74(2): 560-564.

50. McElroy KE, Luciani F, Thomas T. GemSIM: General, error-model based simulator of next-generation sequencing data. *BMC Genomics.* 2012 Feb 15; 13: 74-2164-13-74.

51. Metzker ML. Sequencing technologies — the next generation. *Nat Rev Genet.* 2010; 11(1): 31-46.

52. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics.* 2008 Dec 15; 24(24): 2818-2824.

53. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics.* 2010 Jun; 95(6): 315-327.

54. Myers EW. A whole-genome assembly of drosophila. *Science.* 2000; 287(5461): 2196 - 2204.

55. Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet.* 2013 Mar; 14(3): 157-167.

56. Narzisi G, Mishra B. Comparing de novo genome assembly: The long and short of it. *PLoS ONE.* 2011; 6(4): e19175.

57. Nederbragt L. Cod genome assembly - long reads offer unique insight. Case Study. Pacific Biosciences; 2013 6/2013.

58. Nutman F, Roberts F. Coffee leaf rust. *PANS Pest Articles & News Summaries.* 1970; 16(4): 606-624.

59. Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics.* 2012 12-20; 29(1): 119-121.

60. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet.* 2011; 52(4): 413 - 435.

61. Pevzner PA, Tang H, Waterman MS. An eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A.* 2001 Aug 14; 98(17): 9748-9753.

62. Pop M, Salzberg SL, Shumway M. Genome sequence assembly: Algorithms and issues. *Computer.* 2002; 35(7): 47-54.

63. Powers JG, Weigman VJ, Shu J, Pufky JM, Cox D, Hurban P. Efficient and accurate whole genome assembly and methylome profiling of E. coli. *BMC Genomics.* 2013 Oct 3; 14: 675-2164-14-675.

64. Rajendren RB. A new type of nuclear life cycle in hemileia vastatrix. *Mycologia.* 1967 Mar. - Apr.; 59(2): 279-285.

65. Rozo Y, Escobar C, Gaitán Á, Cristancho M. Aggressiveness and genetic diversity of hemileia vastatrixDuring an epidemic in colombia. *J Phytopathol.* 2012; 160(11-12): 732 - 740.

66. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977 Dec; 74(12): 5463-5467.

67. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet.* 2010 Oct 15; 19(R2): R227-40.

68. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: Methods and goals. *Nat Rev Genet.* 2004 May; 5(5): 335-344.

69. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008; 26(10): 1135 - 1145.

70. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: A parallel assembler for short read sequence data. *Genome Res.* 2009 Jun; 19(6): 1117-1123.

71. Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics.* 2014.

72. Warren RL, Sutton GG, Jones SJM, Holt RA. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics.* 2007; 23(4): 500-501.

73. Xing Y, Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio long read accuracy by short read alignment. *PLoS ONE.* 2012; 7(10): e46679.

74. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.* 2008; 18(5): 821 - 829.

75. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *J Genet Genomics.* 2011 Mar 20; 38(3): 95-109.

Appendix A

K-MER LENGTH OPTIMIZATION

A.1 SOAPdenovo2

K-MER length	SCAFFOLDS	TOTAL CONTIGS (SCAFFOLD +CONTIGS)	MAX LENGTH (bp)	MIN LENGTH (bp)	MEAN (bp)	N50 (bp)	TOTAL LENGTH (bp)	GAPS
23	5139	16672	425069	100	5755	74459	109683696	161920
25	4860	15875	742272	100	6179	83231	110263037	136859
27	4727	16494	752355	100	6089	89428	111090060	113647
29	4685	18833	653344	100	5453	95759	112105368	94459
31	4560	23972	670172	100	4380	103578	113329318	78350
33	4351	34727	750737	100	3105	110058	115131897	64370
35	4215	17080	949859	100	6291	120491	114080742	56250
37	4147	17016	683077	100	6381	125162	114665010	49436
39	3967	17071	922893	100	6417	134518	115151807	44160
41	3787	16709	1061016	100	6603	132477	115477535	39990
43	3485	15872	1168571	100	6994	146629	115765605	36964
45	3272	15693	1265432	100	7112	150613	115999161	33835
47	3020	15153	1266498	100	7401	168021	116166908	31622
49	2920	15596	1189276	100	7224	180057	116356945	29000
51	2746	32200	1189855	100	3563	178171	118159862	27366
53	2617	30992	1190240	100	3714	182584	118316606	25645
55	2533	29730	1190331	100	3878	185632	118328001	24180
57	2461	28342	1157646	100	4079	183045	118441578	22765
59	2423	27294	1157574	100	4240	192918	118441015	21573
61	2396	26495	1305266	100	4374	188418	118450974	20419
63	2299	25978	1305329	100	4472	200193	118571436	19200
65	2191	25436	1485982	100	4575	204029	118638358	18178
67	2131	24974	1486303	100	4666	216118	118652024	17187
69	2087	24826	1487009	100	4702	226757	118748933	16277
71	2011	24610	1487349	100	4749	233790	118775254	15448
73	1931	24418	1620769	100	4794	244140	118882535	14799
75	1914	24366	1275226	100	4811	242187	118963831	14134
77	1857	24517	1418113	100	4789	252819	119060451	13454
79	1806	24869	1418236	100	4730	265518	119207856	12889
81	1745	25149	1621381	100	4684	267622	119305431	12349
83	1792	25518	1349504	100	4623	257938	119410243	11808
85	1742	26157	1509104	100	4517	245011	119543400	11299
87	1736	26943	1721725	100	4391	253870	119667035	10805
89	2043	28063	1167094	100	4216	211867	119659308	10810
91	2885	29146	772702	100	4019	126462	119088427	15608
93	8927	30501	166695	100	3398	21549	113357881	68851
95	24673	64949	10102	100	527	608	40473015	33783
97	265	1278	14028	100	500	722	700078	363
99	4	30	3357	301	613	589	20173	6

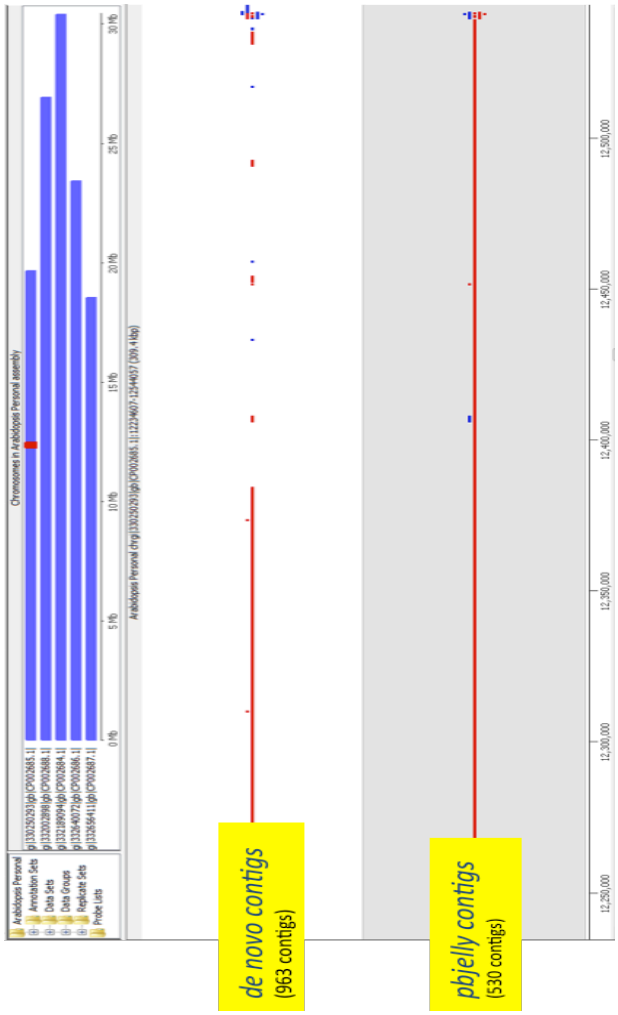
A.2 Velvet

K-MER length	TOTAL CONTIGS	MAX LENGTH (bp)	MIN LENGTH (bp)	MEAN (bp)	N50 (bp)	TOTAL LENGTH (bp)	GAPS
23	1416732	6423	45	103	215	148486162	0
25	1054613	10156	49	134	460	142593678	0
27	861116	14476	53	160	715	139345212	0
29	727180	15663	57	187	963	136914851	0
31	614946	17893	61	217	1285	134489045	0
33	517978	26763	65	253	1730	131921514	0
35	436368	31139	69	295	2285	129451444	0
37	367880	64892	73	345	2953	127312624	0
39	311008	64892	77	402	3839	125423736	0
41	262013	64892	81	470	4853	123621749	0
43	214695	79598	85	565	6295	121661675	0
45	173240	84016	89	690	8146	119811274	0
47	138876	118164	93	850	10073	118259921	0
49	107727	118757	97	1083	12561	116821295	0
51	88874	158819	101	1306	15200	116160143	0
53	80547	168300	105	1441	17208	116161452	0
55	74100	191867	109	1567	18791	116235047	0
57	67988	199962	113	1709	20220	116266464	0
59	62932	199966	117	1847	21271	116333934	0
61	58817	199970	121	1978	22422	116438630	0
63	55154	203561	125	2112	23659	116557141	0
65	51896	203563	129	2246	24485	116658701	0
67	49125	211204	133	2376	25473	116785859	0
69	46609	304167	137	2507	26702	116896132	0
71	44529	304171	141	2627	27742	117026564	0
73	42409	304175	145	2760	28879	117131057	0
75	40433	341943	149	2898	30413	117231050	0
77	38490	357004	153	3046	31978	117301577	0
79	36769	369147	157	3191	33298	117378195	0
81	**35222**	**429438**	**161**	**3333**	**34527**	**117448236**	**0**
83	33818	376981	165	3473	34852	117491728	0
85	32963	238012	169	3563	32354	117511990	0
87	34035	147036	173	3448	22149	117398189	0
89	43405	60410	177	2689	8504	116790589	0
91	82769	22986	181	1371	2380	113596897	0
93	162800	27906	185	577	647	94174492	0
95	69949	10100	189	353	338	24824483	0
97	1113	13805	193	447	419	499274	0
99	48	3170	297	533	556	25645	0

Appendix B

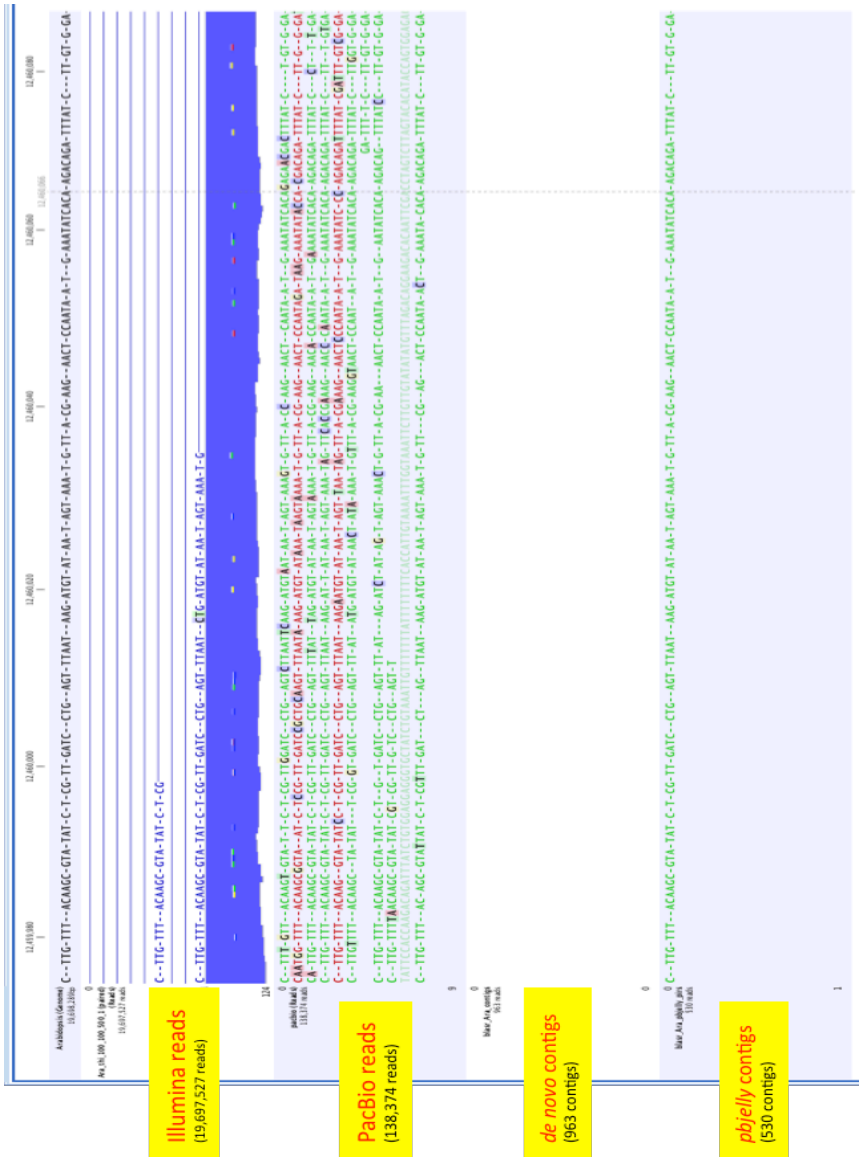
REFERENCE GENOME ALIGNMENT VISUALIZATION

B.1 Using SeqMonk



- Visualization of reference alignment results of the *de novo* contigs and improved-*de novo* (or PBJelly) contigs show the *de novo* contigs could not map to some regions of the chromosome. However, after the improvement stage using PBJelly a single "improved" contig is able to span the entire length of this chromosomal region.

B.2 Using CLC Genome Mapping Viewer



- Nucleotide-level visualization of reference alignment results of the Illumina reads, PacBio reads, *de novo* contigs and improved-*de novo* (or PBJelly) contigs. These emphasize the efficiency of PBJelly, whereby the *de novo* contigs could not be map to this region of the chromosome, which is attributed to the Illumina paired-end reads insert size. The PacBio reads mapped to this chromosomal region but with a lot of insertion errors. However, the PBJelly contigs show accurate alignment with no errors, showcasing the efficiency of the PBJelly tool to accurately improve the de novo contigs with the aid of PacBio reads.

Appendix C

DISTRIBUTION OF THE PBEC-IMPROVED *DE NOVO* CONTIGS FOR THE WGS ASSEMBLY PROCESS

C.1 PacBioTOCA (PBEC)-Improved *de novo* Contigs

De novo assemblers	SOAPdenovo2	CLC	Velvet
Total Number of Contigs	24,108	3,598	24,858
Total Number of Gaps	3,341	664	0
N50 Contig Size (bp)	318,672	180,734	57,489
Mean Contig size (bp)	5,015	31,533	4912
Minimum Contig Size (bp)	40	98	42
Maximum Contig Size (bp)	1,737,679	1,136,517	456,920
Total BP Length (bp)	120,895,399	113,455,652	122,105,709

C.2 Pre-processed PBEC-Improved *de novo* Contigs

De novo assemblers	SOAPdenovo2	CLC	Velvet
Total Number of Contigs	24,108	3,598	24,858
Total Number of Gaps	3,341	664	0
N50 Contig Size (bp)	318,672	180,734	57,489
Mean Contig size (bp)	5,015	31,533	4912
Minimum Contig Size (bp)	40	98	42
Maximum Contig Size (bp)	1,737,679	1,136,517	456,920
Total BP Length (bp)	120,895,399	113,455,652	122,105,709

Appendix D

COPYRIGHT PERMISSIONS

Reprint permission for:

Arneson PA. Coffee rust. *The Plant Health Instructor.* 2000.

Google apps @UDel.edu

Modupe Adetunji <amodupe@udel.edu>

FW: APS Contact Us form submitted by modupeore adetunji

Sue Figueroa <sfigueroa@scisoc.org> Thu, Nov 20, 2014 at 1:56 PM
To: Modupe Adetunji <amodupe@udel.edu>

Dear Modupe Adetunji,

I apologize for the delay in responding.

Permission if hereby granted for you to reproduce both of the images you describe within your master thesis provided the sources are properly credited (as listed below).

The credit for the first image (disease cycle) should appear as follows.

Courtesy V. Brewster; Reproduced, by permission, from Arneson, P. A. 2000. Coffee rust. The Plant Health Instructor. DOI: 10.1094/PHI-I-2000-0718-02

The credit for the second image (world distribution map) should appear as follows. The image was reprinted from one of our journals, so the credit is different from what you might expect.

Reproduced, by permission, from Schieber, E., and Zentmyer, G. A. 1984. Coffee rust in the western hemisphere. Plant Dis. 68:89-93.

Please let me know if I can be of further assistance. Good luck with your thesis.

Sincerely,

Sue Figueroa

==========

Sue Figueroa, Permissions Coordinator
The American Phytopathological Society

3340 Pilot Knob Road
St. Paul, MN 55121
+1.651.994.3871 (telephone)
+1.651.454.0766 (fax)
sfigueroa@scisoc.org

Reprinted by permission from Macmillan Publishers Ltd:

Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010; 11(1): 31-46.

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Dec 15, 2014

This is a License Agreement between modupe adetunji ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3530260503050
License date	Dec 15, 2014
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Genetics
Licensed content title	Sequencing technologies [mdash] the next generation
Licensed content author	Michael L. Metzker
Licensed content date	Jan 1, 2010
Volume number	11
Issue number	1
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1 | template immobilization strategies
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Improving Eukaryotic Genome Assembly Through Application of Single Molecule Real-Time Sequencing Data
Expected completion date	Dec 2014
Estimated size (number of pages)	90
Total	0.00 USD
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this

material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

Reprint Permission for:

Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics.* 2012 Jun; 13(8): 901-915.

Permission to use Future Medicine Ltd copyright material

Request from:

- Contact name: Modupeore Adetunji
- Publisher/company name: University of Delaware
- Address:
- Telephone/e-mail: amodupe@udel.edu

Request details:

- Request to use the following content: Figure 2, Pharmacogenomics, Vol. 13, No. 8, Pages 901-915
- In the following publication: As part of thesis
- In what media (print/electronic/print & electronic): Print and Electronic
- In the following languages: All

We, Future Medicine Ltd, grant permission to reuse the material specified above within the publication specified above.

Notes and conditions:

1. This permission is granted free of charge, for one-time use only.
2. Future Medicine Ltd grant the publisher non-exclusive world rights to publish the content in the publication/website specified above.
3. Future Medicine Ltd retains copyright ownership of the content.
4. Permission is granted on a one-time basis only. Separate permission is required for any further use or edition.
5. The publisher will make due acknowledgement of the original publication wherever they republish the content: citing the author, content title, publication name and Future Medicine Ltd as the original publisher.
6. The publisher will not amend, abridge, or otherwise change the content without authorization from Future Medicine Ltd.
7. Permission does not include any copyrighted material from other sources that may be incorporated within the content.
8. Failure to comply with the conditions above will result in immediate revocation of the permission here granted.

Date: 13/11/2014.....

Future Medicine Ltd, Unitec House, 2 Albert Place, London, N3 1QB, UK
T: +44 (0)20 8371 6080 **F:** +44 (0) 20 8371 6099 **E:** info@futuremedicine.com
www.futuremedicine.com